Richard Gonzalez
Psych 613
Version 3.1 (Jan 2022)

LECTURE NOTES #9: Advanced Regression Techniques II

Reading Assignment

KNNL chapters 22, 12, and 14
CCWA chapters 8, 13, & 15

1. Analysis of covariance (ANCOVA)

We have seen that regression analyses can be computed with either continuous predictors or with categorical predictors. If all the predictors are categorical and follow an experimental design, then the regression analysis is equivalent to ANOVA in terms of structural model parameter values, test statistics and p-values. Analysis of covariance (abbreviated ANCOVA) is a regression that combines both continuous and categorical predictors in the same equation. The continuous variable is usually treated as the "control" variable, analogous to a blocking factor that we saw in LN4, and is called a "covariate" in the world of regression[1]. Usually, one is not interested in the significance of the covariate itself, but the covariate is included to reduce the error variance. The covariate is analogous to the blocking variable in the randomized block design. This section shows how we mix and match ideas from regression and ANOVA to create new analyses (e.g., in this case testing differences between means in the context of a continuous blocking factor).

I'll begin with a simple model. There are two groups and you want to test the difference between the two means. In addition, there is noise in your data and you believe that a portion of the noise can be addressed by a covariate (i.e., a blocking variable). It makes sense that you want to perform a two sample t test, but simultaneously want to correct, or reduce, the variability due to the covariate. The covariate "soaks up" some of the error variance, usually leading to a more powerful test. The price of the covariate is that a degree of freedom is lost because there is one extra predictor. Further, an additional assumption—equal slopes across groups—is made, i.e., no interaction between the two predictors. This is analogous to the situation of a blocking factor where there is no interaction term, though in a randomized block design the interaction cannot be tested because there is only one observation per cell, whereas in ANCOVA the the interaction term is a testable feature of the model as we will see below.

[1]Sometimes the term covariate creates confusion because in some literatures "covariate" refers to all predictors in a regression equation, regardless of whether they are a predictor of interest or a control/blocking variable.

The regression model for this simple design with no interaction (by assumption) is

$$\mathrm{Y} = \beta_0 + \beta_1 \mathrm{C} + \beta_2 \mathrm{G} + \epsilon \quad (9\text{-}1)$$

where C is the covariate and G is the dummy variable coding the two groups (say, subjects in group A are assigned G = 1 and subjects in group B are assigned G = 0; dummy codes are ok here because we aren't creating interactions—refer to LN8 for a discussion of the pesky behavior of dummy codes in the context of interaction terms).

You are interested in testing whether there is a group difference, once the (linear) variation due to variable C has been removed. This test corresponds to testing whether $\hat{\beta}_2$ is significantly different from zero. Thus, the *t* test corresponding to the null hypothesis for β_2 is the difference between the two group means, once the linear effect of the covariate has been accounted for in the analysis.

Look at Eq 9-1 more closely. The G variable for subjects in group A is defined to be 1, thus for those subjects Eq 9-1 becomes

$$\begin{aligned} \mathrm{Y} &= \beta_0 + \beta_1 \mathrm{C} + \beta_2 \mathrm{G} + \epsilon \\ &= (\beta_0 + \beta_2) + \beta_1 \mathrm{C} + \epsilon \end{aligned}$$

For subjects in group B, G=0 and Eq 9-1 becomes

$$\begin{aligned} \mathrm{Y} &= \beta_0 + \beta_1 \mathrm{C} + \beta_2 \mathrm{G} + \epsilon \\ &= \beta_0 + \beta_1 \mathrm{C} + \epsilon \end{aligned}$$

In this model the only difference between group A and group B boils down to a difference in intercept of each group, and that difference is indexed by β_2. ANCOVA assumes that the slopes for both groups are identical—this is what allows the overall correction due to the covariate. The standard ANCOVA model does not permit an interaction because the slopes are assumed to be equal. A graphical representation of the two-group ANCOVA is shown in Figure 9-1.[2]

An interaction term may be included in the regression equation but then one loses the ability to interpret ANCOVA as a simple "adjusted ANOVA" because the difference between the two groups changes depending on the value of the covariate (i.e. the slope between DV and covariate differs across groups). In the presence of an interaction with a covariate, there is not a single estimate of the difference between the groups because the difference between the two

[2] An R command for doing these plots is presented near the end of this ANCOVA section. SPSS can produce a graph similar to the one in Figure 9-1 using the menu system. Here's what I did on an older SPSS version. The specifics for the version you use may be slightly different. Create a scatter plot in the usual way (e.g., Graphs/Scatter/Simple), define the covariate as the X-variable, the dependent variable as the Y-variable, and the grouping variable as the "set markers by" variable. When the chart comes up in the output, edit it by clicking on the edit button, then select from the chart menu "Chart/Options" and click the "subgroups" box next to the fit line option. Finally, click on the "fit options" button and then click on "linear regression." These series of clicks will produce a separate regression line for each group. I couldn't figure out a way to do all that using standard SPSS syntax.

groups will depend on the value of the covariate. The interaction implies the lines for each group are not parallel, making it difficult to interpret the main effect of the grouping variable G.

Figures 9-2 and 9-3 present several extreme cases of what ANCOVA can do. By looking at such boundary conditions, I hope you will get a better understanding of what ANCOVA does. These figures show that a covariate can be useful in interpreting results differences between groups in the context of a nonexperimental design.

Suppose the investigator didn't realize that the covariate needed to be collected and the "true" situation is Figure 9-2, then the researcher who performed a simple two-sample t test to compare the two means would fail to find a difference that is there (i.e., a difference in intercepts) once the role of the covariate has been addressed. Or, if the investigator did not collect the covariate and the true situation is Figure 9-3, then that researcher would find a significant difference on Y when performing a two-sample t test, when the difference between the means can be completely attributed to the covariate. In these two cases, the researcher would make an incorrect conclusion by conducting a two sample t-test on the means.

The implication is that a forgotten covariate can lead to situations that misinform. A model that omits a key predictor or covariate is called a *misspecified model*. This is why it is important to think through your design and the potential data patterns before you collect data. You want to be sure to collect all the data necessary to help you decipher different patterns that you may observe. Once you have an effect, you should think critically as to whether there could be a relevant covariate that could be misleading your conclusions. If you come up with a plausible alternative explanation involving a potential covariate, then that is good justification to collect those additional data and redo the study. Better that you come up with alternative explanations and address them, rather than relying on reviewers, or readers of your work, to generate plausible critiques. Keep in mind that a model that omits a key predictor simply puts that systematic variability into the error term ϵ, and consequently, the error term is systematic when it needs to be random error. We saw this in a rather simplistic manner when we considered blocking factors in Lecture Notes 4, but for correlational data, the effects of misspecified models can lead to incorrect conclusions, even incorrect interpretations of differences as Figures 9-2 to 9-4 illustrate.

These figures display an important detail that needs to be in place for ANCOVA "to work its magic." One interpretation of ANCOVA is that we are comparing group differences holding constant the effect of the covariate (recall the interpretation of regression slopes in Lecture Notes 8 in the part and partial correlation section). That is, if we could equate the two groups in terms of their values on the covariates, we could then compare the means having removed the effect of the covariate. Randomization is one way to accomplish that goal (randomized groups would be equated on the covariate), but when randomization isn't possible ANCOVA is one option. "Okay fine, equating the two groups on the covariate, I get the point" you say. However, look closely at, say, Figure 9-2. In that example, the two groups do not overlap at

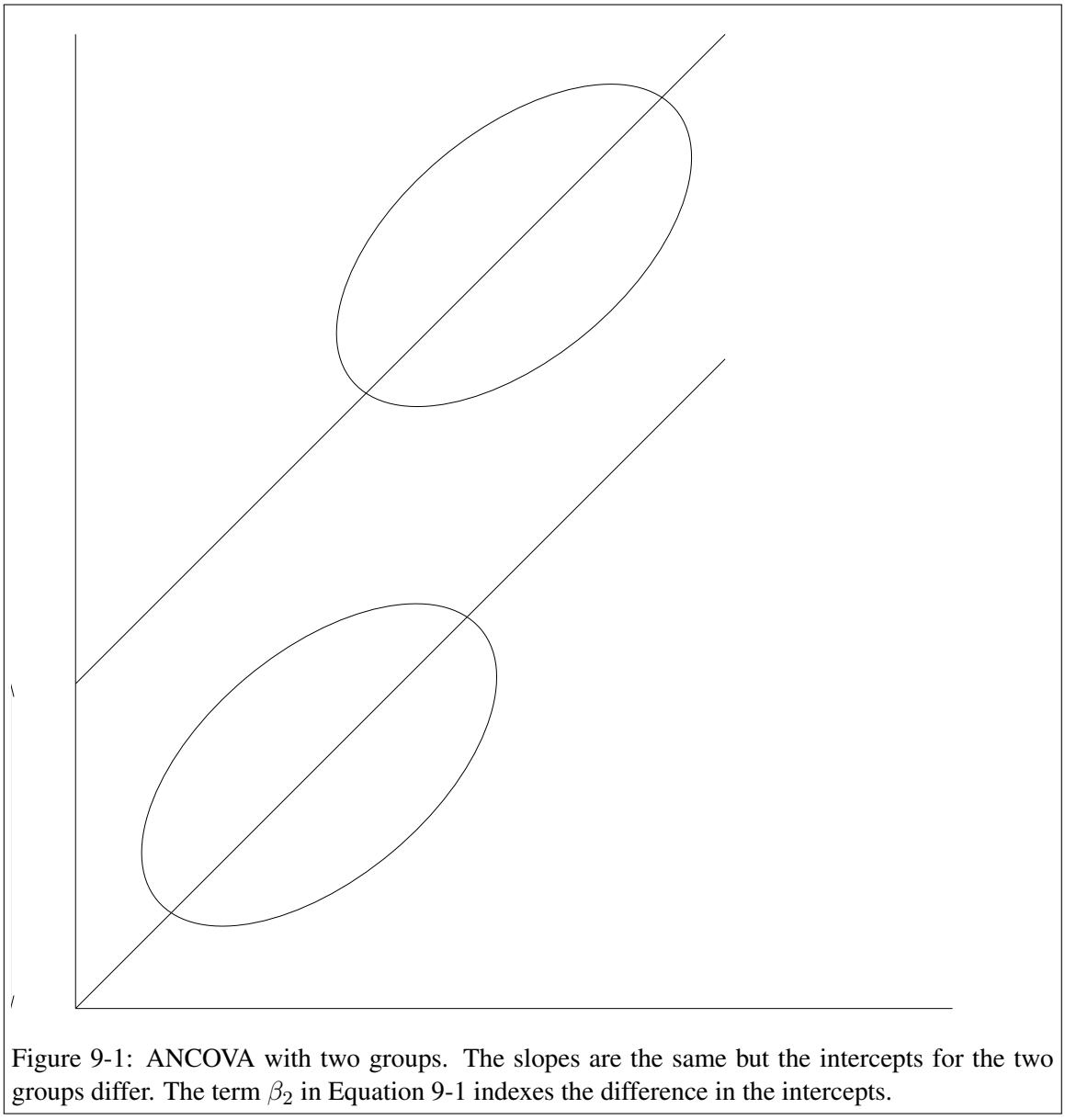


Figure 9-1: ANCOVA with two groups. The slopes are the same but the intercepts for the two groups differ. The term β_2 in Equation 9-1 indexes the difference in the intercepts.

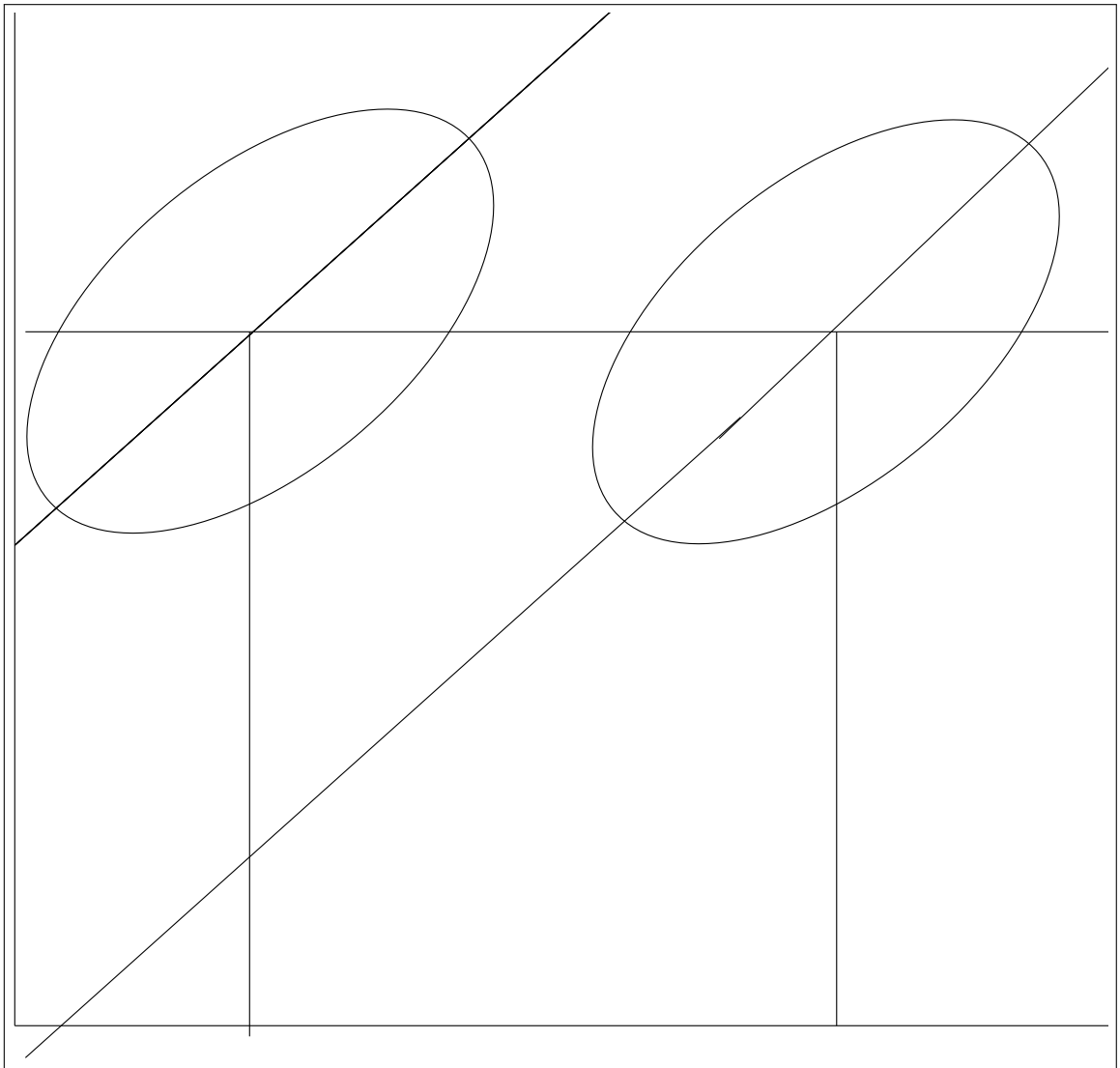


Figure 9-2: ANCOVA with two groups. The slopes are the same for the two groups but the intercepts differ. In this example, there is no difference between the two unadjusted group means (as indicated by the horizontal line going through the two ovals). That is, both groups have the same unadjusted mean value on Y, so a classic two-sample t test would show no difference. However, performing an ANCOVA yields a group difference in adjusted means, as indicated by the difference in the intercept between the two lines. The reason is that both groups differ on the value of the covariate.

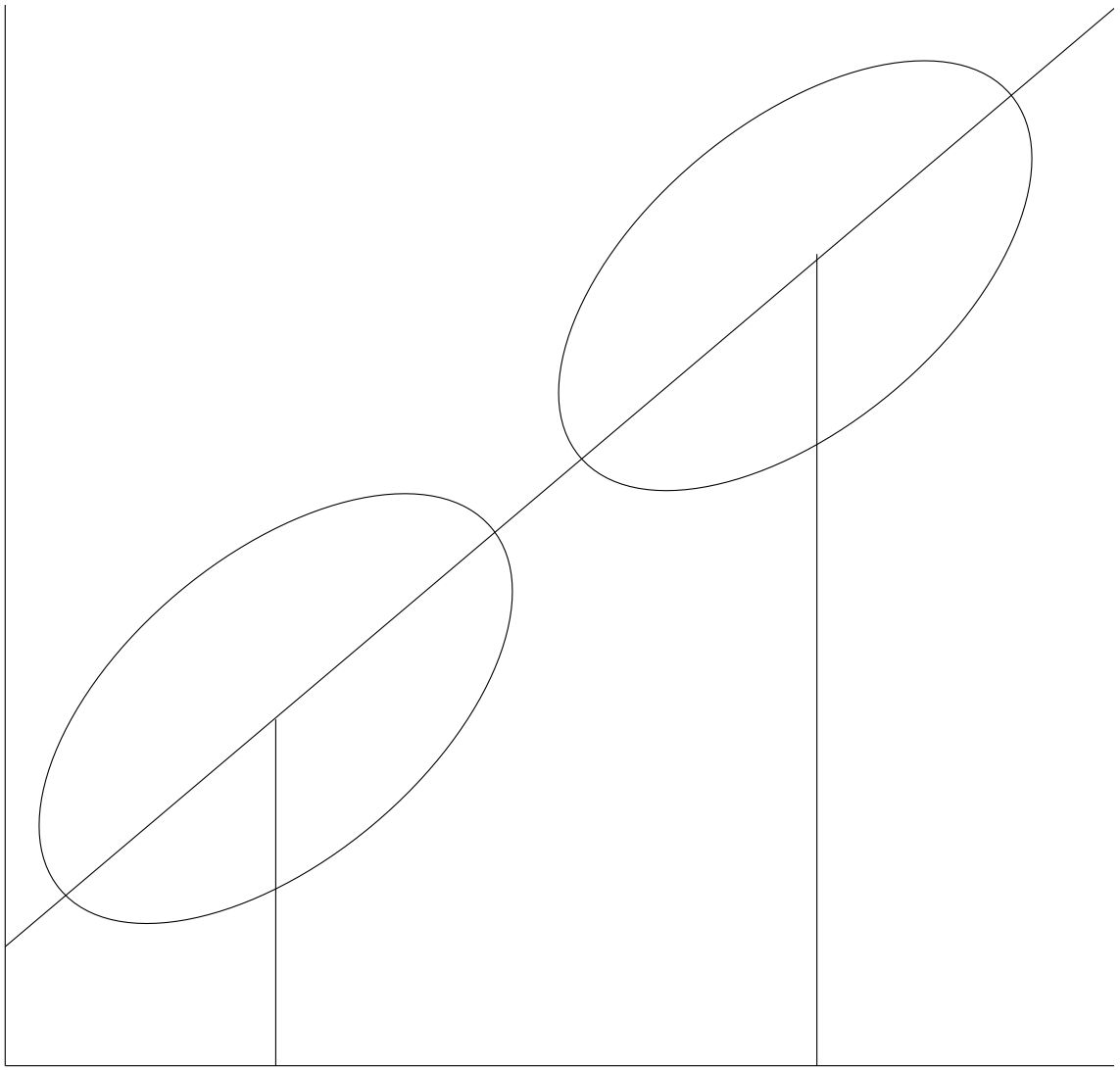


Figure 9-3: ANCOVA with two groups. The slopes and intercepts are the same for the two groups. In this example, the ANCOVA reveals no difference in adjusted group means because the intercept is the same for each group. Thus, any initial difference on Y can be explained by the covariate. Once the effect of the covariate is removed, then no group difference remains. However, the investigator who didn't realize the dependency of the covariate and simply performed a two-sample t test would find a difference between the two group means, even though that difference can be accounted for by different values on the covariate.

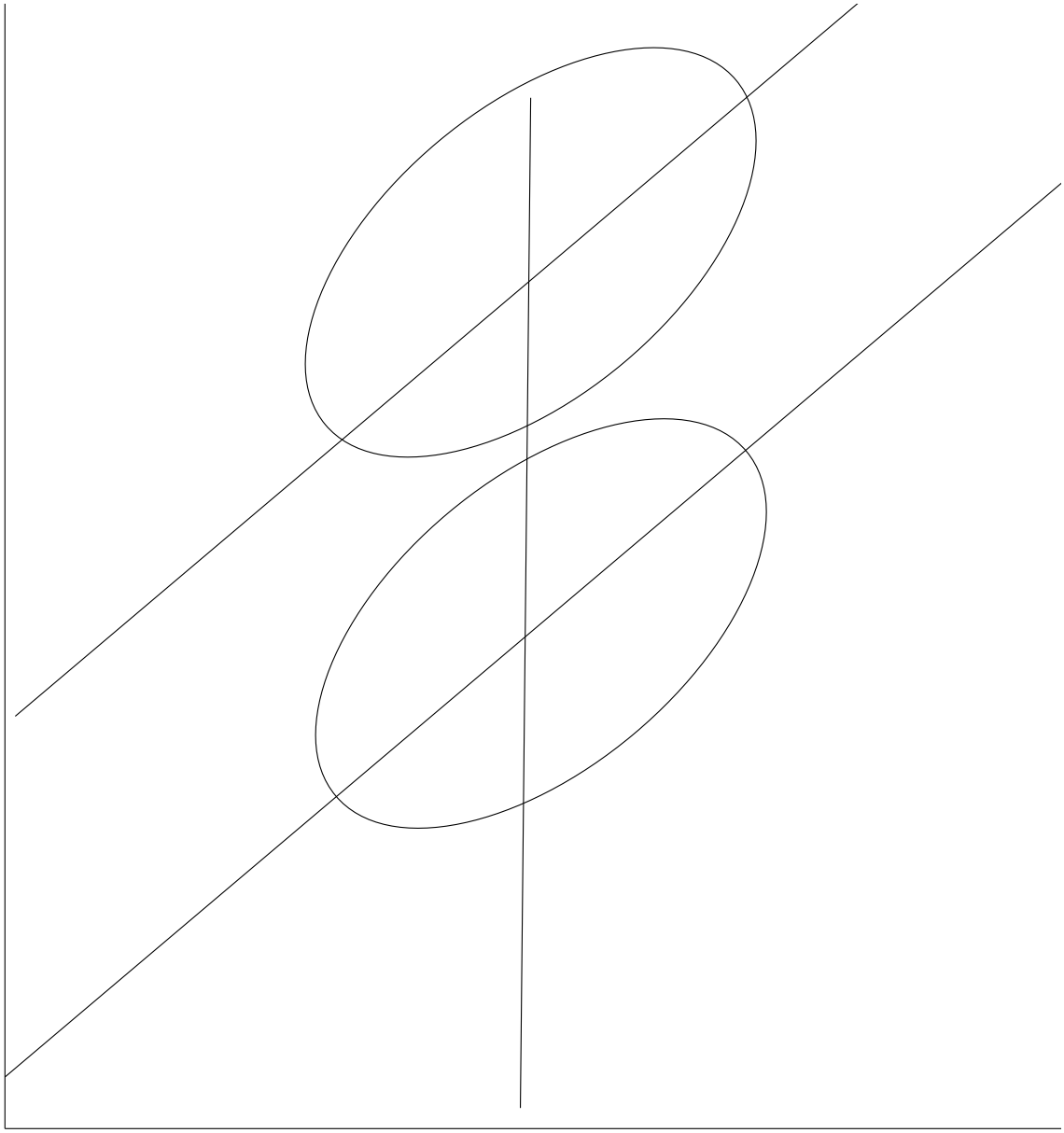


Figure 9-4: ANCOVA with two groups. The slopes are the same but the intercepts differ. In this example, the ANCOVA reveals a group difference in adjusted means even though the mean value of the covariate is the same for each group.

all on the values of the covariate so the concept of "equating the two groups on the covariate" is meaningless because there is no X value for which the two groups coincide. For this adjustment concept to have much meaning, there must be some "overlap" in the values of the covariate between the two groups so that the correction introduced by ANCOVA could make sense (e.g., there needs to be two subjects with the same score on the covariate but one of those subjects is in group 1 and the other subject is in group 2, and there needs to be several such pairs with different values on the covariate). I'll discuss this point in more detail when reviewing the technique of *propensity scores*, but we need to get logistic regression under our belt before tackling that procedure.

2. Example of ANCOVA: Rich goes to Lamaze class

There are three suggested breathing exercises for pain management during labor (slow paced, accelerated, patterned). Which method works the best? How would one answer this question empirically? What would be an appropriate dependent variable? What covariates would make sense?

Obviously, the best experimental test of these breathing techniques would involve random assignment (i.e., the gold standard clinical trial). However, random assignment in such a domain is difficult (and may be unethical) so a study would probably allow the pregnant mother to "self-select" one breathing exercise. But, there may be systematic reasons why some individuals select one breathing exercise over another, and these may relate to the success of the breathing exercise (the intervention). In such a case, careful selection of covariates (such as pain tolerance) may help improve the nature of the inferences that are possible from such a study. A covariate is, in a sense, a way to control statistically for external variables rather than control for them experimentally as in a randomized block design. In some domains, statistical control is the only option one has. We will discuss in subsequent lecture notes other approaches to address causality when experiments are not possible.

3. Testing for Parallelism in ANCOVA

parallelism assumption

The assumptions for ANCOVA are the same as in the usual regression. There is one additional assumption: the slope corresponding to the covariate is the same across all groups. Look at Eq 9-1–there is nothing there that allows each group to have a different slope for the covariate. Hence, the assumption that the slopes for each group are equal.

The equality of slopes assumption is easy to test. Just run a second regression that includes the interaction between the covariate and the dummy code. For example, to test parallelism for a model having one covariate and two groups, one would run the following regression

$$\mathrm{Y} = \beta_0 + \beta_1 \mathrm{C} + \beta_2 \mathrm{G} + \beta_3 \mathrm{CG} + \epsilon \qquad (9\text{-}2)$$

where CG is simply the product of C and G (i.e., the interaction). The only test you care

about in Eq 9-2 is whether or not $\hat{\beta}_3$ is statistically different from zero. The interaction slope β_3 is the only test that is interpreted in this second regression because the covariate is likely correlated with the grouping variable. Recall that sometimes the presence of an interaction term in the model makes the main effects uninterpretable due to multicolinearity (see Lecture Notes 8) and that is why the sequential approach is often used where the main effects are interpreted and tested in a main-effects only regression. Centering the predictor variables will help make the main effect interpretable. But keep in mind that the regression in Equation 9-2 is for testing the parallel slope assumption, so all we care about is the interaction. Hence, for this regression centering isn't really needed, but no harm if you choose to center the predictors.

If the interaction in Equation 9-2 is not significant, then parallelism is okay; but if $\hat{\beta}_3$ is statistically significant, then parallelism is violated because there is an interaction between the covariate and the grouping variable.[3]

SPSS ANCOVA issues

Most of my lecture notes on ANCOVA make use of the REGRESSION command in SPSS or the corresponding lm() command in R. Within SPSS, MANOVA and GLM are two other SPSS commands that also handle covariates. I noticed a strange difference between the MANOVA and GLM commands: they each handle covariates differently. The MANOVA command behaves nicely and automatically centers all predictors (regardless of whether they are expressed as factors through BY or expressed as covariates through WITH). The GLM command though doesn't center the covariates (those variables after the WITH), but does for factors specified through the BY subcommand. So, while GLM does automatically test the parallelism assumption by fitting interactions of covariate and predictor, it doesn't center the variables so the analyses for the main effects are off. GLM does center the factors in the usual factorial (those using the BY command)—basically everything after BY is centered but everything after WITH is not[4]. Of course, only discrete factors make sense to include as part of the BY subcommand, not continuous predictors. I have had several cases in collaborative work where a couple of us rerun the analyses and we find differences in our results usually due to subtle differences like one of us used BY and the other WITH, or one of us used one command in SPSS/R and the other used another command in SPSS/R that should have yielded the same answer but the commands differ in small ways such as how they handle order of entry in a regression, centering, etc.

The basic lesson here is that if you want to test any interaction, and in the case of ANCOVA you do want to test the parallelism assumption by testing the interaction term, then center or you could get incorrect results for the main effects just by some weird detail of the particular program you are using (such as MANOVA or GLM in SPSS but also with R commands) and, in SPSS, subtle differences between whether you use the BY or the WITH command. Don't trust the stat package to do the right thing. Get in the habit of centering all your variables

[3] Of course, issues of power come into play when testing the parallelism assumption—with enough subjects you will typically reject the assumption even if the magnitude of the parallelism violation is relatively small.

[4] The MIXED command in SPSS suffers from the same issue of how it handles BY and WITH, be careful with that command too.

whenever you include interaction terms of any kind. This way you can interpret the main effects even in the presence of a statistically significant interaction.

Johnson-Neyman

If you **do** find a violation of the parallelism assumption, then alternative procedures such as the Johnson-Neyman technique can be used (see Rogosa, 1980, *Psychological Bulletin*, for an excellent discussion). Maxwell and Delaney also discuss this in the ANCOVA chapters of their experimental design textbook. Obviously, if the slopes differ, then it isn't appropriate to talk about a single effect of the categorical variable (e.g., the difference between the means of two groups) because the effect depends on the level, or value, of the covariate. The more complicated techniques amount to finding tests of significance for the categorical predictor conditional on particular values of the covariate. They are similar to what we saw in Lecture Notes #8 for testing interactions in regression.

To understand Equation 9-2 that includes the interaction term, it is helpful to code the groups denoted by variable G as either 0 or 1 (i.e., dummy codes) and write out the equation. When G = 0, Equation 9-2 reduces to a simple linear regression of the dependent variable on the covariate with intercept β_0 and slope β_1C. However, when G = 1, Equation 9-2 is also a simple linear regression of the dependent variable on the covariate but the intercept is $(\beta_0 + \beta_2)$ and the slope is $(\beta_1 + \beta_3)$C. Thus, β_3 is a test of whether the slope is the same in both groups and β_2 is a test of whether the intercept is the same for both groups. For instance, if β_3 = 0, then the slope in each group is the same but if $\beta_3 \neq 0$, then the slopes differ across the two groups. Here, I used dummy codes but as usual be cautious of interpreting main effects in the presence of interactions when the variables were not centered—dummy codes of 0s and 1s are not centered. If you use the sequential approach, and interpret the main effects in the main-effects only regression model and interpret only the interaction term(s) in the full regression Equation 9-2, you will be ok.

More complicated ANCOVA designs can have more than one covariate and more than two groups. To get more than one covariate just include each covariate as a separate predictor. To get more than two groups just do the usual dummy coding (or contrast coding or effects coding) on the categorical factor. Here is an example with two covariates (C_1 and C_2) and four groups (needing three dummy codes or contrast codes—G_1, G_2, and G_3; see LN8 for a review of coding factors in regression):

$$Y = \beta_0 + \beta_1 C_1 + \beta_2 C_2 + \beta_3 G_1 + \beta_4 G_2 + \beta_5 G_3 + \epsilon \quad (9\text{-}3)$$

This model tests whether there is a difference between the four groups once the variability due to the two covariates has been removed. Thus, we would be interested in comparing the full model (with all five predictors) to the reduced model (omitting the three dummy codes). This "increment in R^2" test corresponds to the research question: do the grouping variables add a significant amount to the R^2 over and above what the covariates are already accounting for. If dummy codes are used, the increment in R^2 test described above tests the omnibus hypothesis correcting for the two covariates. It is also possible to replace the three dummy codes with a set of three orthogonal contrasts. The contrast codings would allow you to test the unique contribution of the contrasts while at the same time correcting for the variability

of the covariates, thus reducing the error term. The *t*-tests corresponding to the βs for each contrast correspond to the test of the contrast having accounted for the covariate (i.e., reducing the error variance in the presence of the covariates). That is, the individual $\hat{\beta}$'s correspond the tests of significance for that particular contrast (having corrected for the covariate). Again, the usual regression/ANOVA assumptions of independence, equal variances and normality hold as well as the equality of slopes assumption needed for ANCOVA, which permits one to interpret the results of ANCOVA in terms of adjusted means based on intercept differences. Further, there should be good overlap across both the covariates for each group. If some covariates are completely nonoverlapping across the groups, then there isn't much meaning to "controlling" for the covariates, as we saw earlier in these lecture notes in the case of one covariate. This example shows that the contrast material in Lecture Notes 3 can be generalized to include covariates by recasting an original contrast/ANOVA problem (Lecture Notes 3 material) into a predictor/ANCOVA problem. We now have a procedure that tests contrasts in the context of a model that adjusts for one or more covariates.

4. Two ANCOVA examples

I show the effects of a covariate reducing noise (MSE) as compared to an analysis of the group differences without the covariate (i.e., using, for example, the ONEWAY command in SPSS and later the lm() command in R). The first column is the grouping code, the second column is the covariate, and the third column is the dependent variable.

```
1 5 6
1 3 4
1 1 0
1 4 3
1 6 7
2 2 1
2 6 7
2 4 3
2 7 8
2 3 2

data list file='data' free/ grp covar dv.

plot format = regression
 /plot= dv with covar by grp.

regression variables = grp covar dv
  /statistics r anova coef ci
  /dependent dv
  /method=enter covar grp.

oneway dv by grp(1,2).
```

```
Multiple R           .97188          Analysis of Variance
R Square             .94456                          DF      Sum of Squares      Mean Square
Adjusted R Square    .92872          Regression           2            65.08000         32.54000
Standard Error       .73872          Residual             7             3.82000           .54571
```

```
                                F =      59.62827       Signif F =  .0000

Variable             B        SE B     95% Confdnce Intrvl B       Beta         T  Sig T

COVAR          1.425000     .130589     1.116206     1.733794    .984698    10.912  .0000
GRP            -.655000     .473735    -1.775203      .465203   -.124768    -1.383  .2093
(Constant)     -.760000     .848730    -2.766923     1.246923                -.895  .4003


      Variable  DV
   By Variable  GRP

                              ANALYSIS OF VARIANCE

                              SUM OF         MEAN            F      F
       SOURCE           D.F.  SQUARES       SQUARES         RATIO  PROB.

BETWEEN GROUPS             1      .1000         .1000        .0116  .9168

WITHIN GROUPS              8    68.8000        8.6000

TOTAL                      9    68.9000
```

This version illustrates how a slight change in the arrangement of the plot changes the ANCOVA result. It is important to check the scatterplot before performing the ANCOVA so you know exactly what the covariate is doing. In these examples parallelism holds. One should always check whether parallelism holds (i.e., there is no interaction between the grouping variable G and the covariate C) before proceeding with the analysis of covariance.

```
DATA SET #2 (JUST CHANGED THE THIRD COLUMN FOR TREATMENT #1)
1 5 11
1 3 9
1 1 5
1 4 8
1 6 12
2 2 1
2 6 7
2 4 3
2 7 8
2 3 2

data list file='data2' free/ grp covar dv.

plot format = regression
 /plot= dv with covar by grp.

regression variables = grp covar dv
  /statistics r anova coef ci
  /dependent dv
  /method=enter covar grp.

oneway dv by grp(1,2).
```

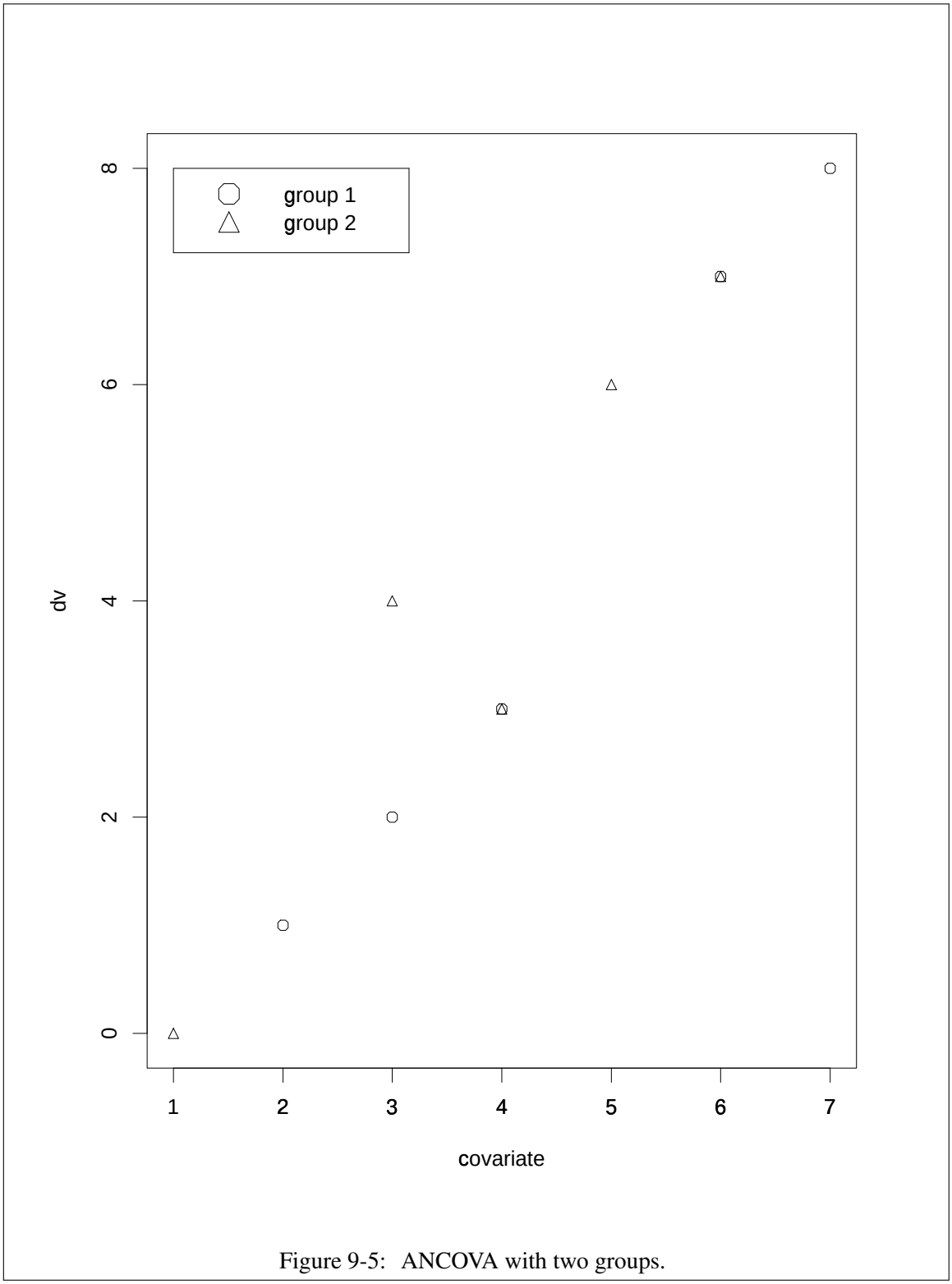


Figure 9-5: ANCOVA with two groups.

```
Multiple R           .98477          Analysis of Variance
R Square             .96978                          DF      Sum of Squares      Mean Square
Adjusted R Square    .96114          Regression       2           122.58000         61.29000
Standard Error       .73872          Residual         7             3.82000           .54571

                                     F =     112.31152       Signif F =  .0000

Variable              B        SE B     95% Confdnce Intrvl B       Beta         T  Sig T

COVAR          1.425000     .130589     1.116206     1.733794    .727008    10.912  .0000
GRP           -5.655000     .473735    -6.775203    -4.534797   -.795297   -11.937  .0000
(Constant)     9.240000     .848730     7.233077    11.246923               10.887  .0000
```

ANALYSIS OF VARIANCE

SOURCE	D.F.	SUM OF SQUARES	MEAN SQUARES	F RATIO	F PROB.
BETWEEN GROUPS	1	57.6000	57.6000	6.6977	.0322
WITHIN GROUPS	8	68.8000	8.6000		
TOTAL	9	126.4000			

The identical analysis can be done within the MANOVA command. The covariate is added at the first line after the keyword "with". The PMEANS subcommand is helpful because it gives the adjusted means (i.e., what the ANCOVA predicts in the regression) when the value of the covariate is set equal to the mean of the covariate. For example, the mean of the covariate C (collapsing over all groups) is 4.1. The predicted mean on the DV for group 1 using the regression estimates (i.e,. plug in G = 1) from the regression output is

$$\begin{aligned} Y &= \beta_0 + \beta_1 C + \beta_2 G + \epsilon && \text{(9-4)} \\ &= 9.24 + (1.425)(4.1) + (-5.665)(1) && \text{(9-5)} \\ &= 9.4275 && \text{(9-6)} \end{aligned}$$

For group 2, it's the same equation except that G=2 rather than G=1 is entered into the regression, so we have

$$\begin{aligned} Y &= \beta_0 + \beta_1 C + \beta_2 G + \epsilon && \text{(9-7)} \\ &= 9.24 + (1.425)(4.1) + (-5.665)(2) && \text{(9-8)} \\ &= 3.7725 && \text{(9-9)} \end{aligned}$$

In the following output, I verify that the *F* tests for the covariate and the grouping variable in the source table from a MANOVA command are identical to the square of the t-tests given in the previous regression run (i.e., $10.912^2 = 119.07$ and $-11.937^2 = 142.49$ for the two main effects). The PMEANS subcommand of MANOVA produces the predicted means of 9.4 and 3.77 for the two groups as I computed manually from the regression equation above.

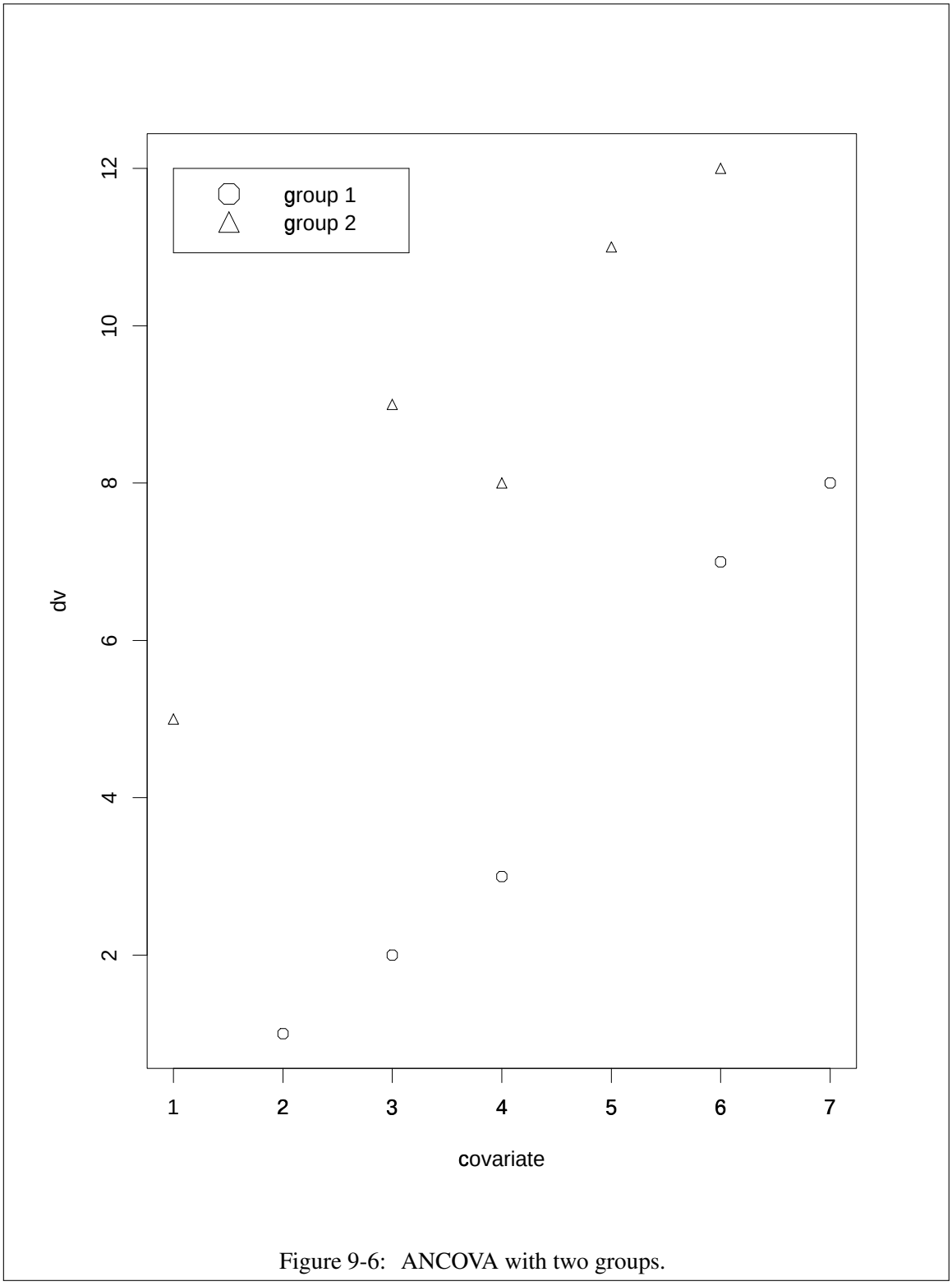


Figure 9-6: ANCOVA with two groups.

```
manova  dv by grp(1,2) with covar
 /pmeans tables(grp)
 /design grp .


 Tests of Significance for DV using UNIQUE sums of squares
 Source of Variation          SS      DF        MS         F  Sig of F

 WITHIN+RESIDUAL            3.82       7       .55
 REGRESSION                64.98       1     64.98    119.07      .000
 GRP                       77.76       1     77.76    142.49      .000

 (Model)                  122.58       2     61.29    112.31      .000
 (Total)                  126.40       9     14.04

 R-Squared =           .970
 Adjusted R-Squared =  .961

 Regression analysis for WITHIN+RESIDUAL error term
--- Individual Univariate .9500 confidence intervals
 Dependent variable .. DV

 COVARIATE          B           Beta      Std. Err.    t-Value   Sig. of t  Lower -95%  CL- Upper

 COVAR   1.4250000000    .7270081448         .13059   10.91207        .000     1.11621    1.73379


 Adjusted and Estimated Means
 Variable .. DV
      Factor          Code        Obs. Mean     Adj. Mean     Est. Mean    Raw Resid.   Std. Resid.

  GRP                   1         9.00000       9.42750       9.00000        .00000        .00000
  GRP                   2         4.20000       3.77250       4.20000        .00000        .00000


 Combined Adjusted Means for GRP
 Variable .. DV
           GRP
             1      UNWGT.     9.42750
             2      UNWGT.     3.77250
```

It is possible to test the equality of slope assumption (i.e., the interaction) in the MANOVA command by this trick. The design line enters the two main effects and the interaction, and there is this new subcommand ANALYSIS, which I don't want to get into here because it is somewhat esoteric. The test of the interaction term appears in the ANOVA source table. The only thing you care about in this output is the significance test for the interaction. Here is my preferred syntax for testing the interaction between the covariate and the grouping variable when using the MANOVA command:

```
manova dv by grp(1,2) with covar
 /pmeans tables(grp)
 /analysis dv
 /design grp covar grp by covar.
```

If the equality of slope assumption is rejected, then you can fit an ANCOVA-like model that permits the two groups to have different slopes using the following syntax (the covariate is "nested" within the grouping code). But my own preference is to stick with the previous analysis that uses an interaction. The following approach, though, is more commonly accepted. The F test for the grouping variable is identical for both approaches and the MSE term is also identical (as are the "adjusted means," which are identical to the observed means). The two approaches differ in that the interaction approach keeps the main effect for covariate and the interaction separate, whereas the following approach lumps the sum of squares for interaction and covariate main effect into a single omnibus test (and you know how I feel about omnibus tests). Here is the more common syntax, though not my favorite way of testing the interaction:

```
manova dv by grp(1,2) with covar
 /pmeans tables(grp)
 /analysis dv
 /design covar within grp, grp.
```

5. ANCOVA and R

You can use the lm() command for analysis of covariance given that ANCOVA is a regression problem. I'll repeat the simple example in the lecture notes (data set 2) with grouping variable G, covariate C and dependent variable Y.

```
data2 <- read.table("data2")
colnames(data2) <- c("G", "C", "Y")
data2
```

```
##    G C  Y
## 1  1 5 11
## 2  1 3  9
## 3  1 1  5
## 4  1 4  8
## 5  1 6 12
## 6  2 2  1
## 7  2 6  7
## 8  2 4  3
## 9  2 7  8
## 10 2 3  2
```

```
out.lm <- lm(Y ~ C + G, data = data2)
summary(out.lm)
```

```
##
## Call:
## lm(formula = Y ~ C + G, data = data2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.2850 -0.1875  0.0425  0.2725  1.1400
##
## Coefficients:
##             Estimate Std. Error t value
## (Intercept)   9.2400     0.8487   10.89
## C             1.4250     0.1306   10.91
## G            -5.6550     0.4737  -11.94
##             Pr(>|t|)
## (Intercept) 1.22e-05 ***
## C           1.20e-05 ***
## G           6.59e-06 ***
## ---
## Signif. codes:
##   0 '***' 0.001 '**' 0.01 '*' 0.05
##   '.' 0.1 ' ' 1
##
## Residual standard error: 0.7387 on 7 degrees of freedom
## Multiple R-squared:  0.9698,Adjusted R-squared:  0.9611
## F-statistic: 112.3 on 2 and 7 DF,  p-value: 4.799e-06
```

Test the parallel slopes assumption through the interaction of the grouping factor and the covariate, focusing just on the interaction term as the test of the assumption. If there are more than two groups, then one tests the interaction of every predictor associated with the grouping variable and the covariate (i.e., if there are four groups, then there are three predictors such as dummy codes or contrasts, and there will be three interaction terms for the three ways to multiply a predictor corresponding to factor G with the covariate C).

```
out.lm.int <- lm(Y ~ C * G, data = data2)
summary(out.lm.int)
```

```
##
## Call:
## lm(formula = Y ~ C * G, data = data2)
##
## Residuals:
##      Min       1Q   Median       3Q
```

```
## -1.27027 -0.19123 -0.02137  0.37681
##      Max
##  1.08108
##
## Coefficients:
##             Estimate Std. Error t value
## (Intercept)  10.0786     1.9204   5.248
## C             1.2143     0.4482   2.709
## G            -6.2137     1.2370  -5.023
## C:G           0.1370     0.2773   0.494
##             Pr(>|t|)
## (Intercept)  0.00192 **
## C            0.03515 *
## G            0.00240 **
## C:G          0.63880
## ---
## Signif. codes:
##   0 '***' 0.001 '**' 0.01 '*' 0.05
##   '.' 0.1 ' ' 1
##
## Residual standard error: 0.7822 on 6 degrees of freedom
## Multiple R-squared:  0.971,Adjusted R-squared:  0.9564
## F-statistic: 66.87 on 3 and 6 DF,  p-value: 5.298e-05
```

To get the predicted (aka adjusted means) for the groups with the covariate set to its mean, you can do something like this. I define the predictor G values at 1 and 2, the covariate mean as the value for C, put everything into a list as required by the predict command, and then execute the predict command. This reproduces the manual computation I presented earlier. I also get confidence intervals around the adjusted group means.

```
newgroup <- c(1, 2)
newcovariate <- rep(mean(data2$C), 2)
newpredictor <- list(G = newgroup, C = newcovariate)
predict(out.lm, newdata = newpredictor, se.fit = T,
    interval = "confidence")
```

```
## $fit
##      fit      lwr       upr
## 1 9.4275 8.640831 10.214169
## 2 3.7725 2.985831  4.559169
##
## $se.fit
```

```
##         1         2
## 0.3326825 0.3326825
##
## $df
## [1] 7
##
## $residual.scale
## [1] 0.7387248
```

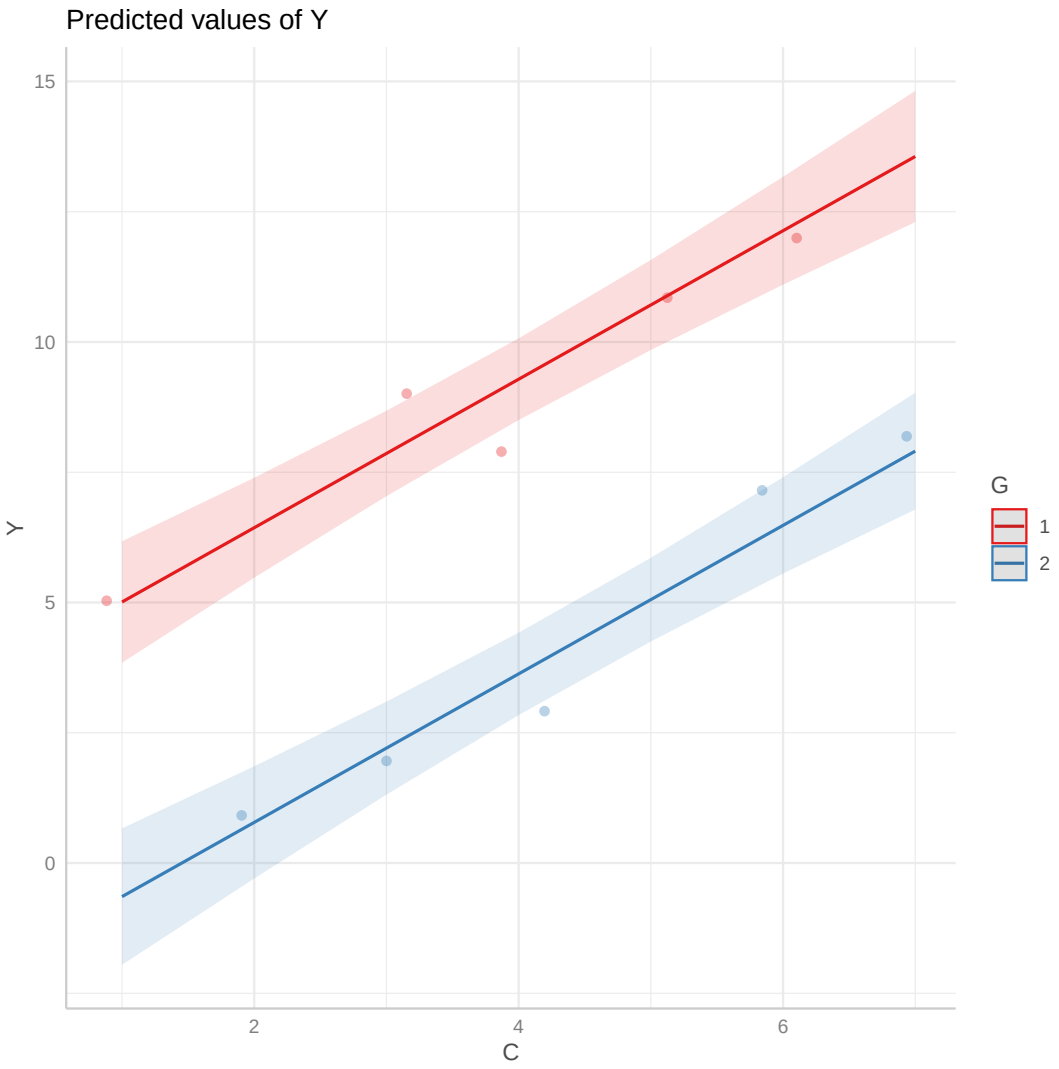
If you want to get the predicted values through specific R commands, you can use the ggeffects package, which is a wrapper to lots of interesting plots and different ways of computing predicted scores after controlling for other variables. Here is the ggeffect version that replicates what I did in SPSS. There are other functions in the ggeffects package such as ggemmeans() and ggpredict() that have different defaults and different ways of approaching these prediction problems yielding potentially different results—read the manual and vignettes for more information.

```
# print to 4 digits to compare with SPSS and
# previous R
library(ggeffects)
print(ggeffect(out.lm, terms = ~G), digits = 4)
```

```
## # Predicted values of Y
##
## G | Predicted |            95% CI
## --------------------------------
## 1 |    9.4275 | [8.6408, 10.2142]
## 2 |    3.7725 | [2.9858,  4.5592]
```

You can plot the ANCOVA lines by using this command along with superimposing the raw data using the add.data=T argument. See the ggeffects package for various plotting options.

```
plot(ggeffect(out.lm, terms = ~C + G), add.data = T)
```



6. A Standard ANCOVA Misunderstanding

In Lecture Notes 8 I reviewed the definitions of part and partial correlation. I also showed how one could estimate part and partial correlations through regressions on residuals. Recall that the key issue was that the other predictors are removed from the predictor of interest, and then the residuals of that predictor are correlated with the dependent variable.

ANCOVA follows that same logic. The variance of the covariate C is removed from the grouping variable G and the residuals from that regression are used as the predictor of the dependent variable Y. Thus, one can interpret ANCOVA as purging the linear effect of the covariate C from the **grouping** variable G. Many users though have the misconception that ANCOVA purges the covariate from the dependent variable, that is not completely the correct

intuition however.

Model notation should make all this clearer. ANCOVA compares the reduced model to the full model (here, C is the covariate and G is the grouping variable):

$$\begin{aligned} \text{reduced } R^2 \text{ from: } \quad Y &= \beta_0 + \beta_1 C + \epsilon \\ \text{full } R^2 \text{ from: } \quad Y &= \beta_0 + \beta_1 C + \beta_2 G + \epsilon \end{aligned}$$

with the focus being on the *t*-test of β_2, the slope of the grouping variable in the context of the full regression. This comparison of these two R^2s is identical, as we saw in LN8 in the part correlation section, to first running a regression with G as the dependent variable and the covariate C as the sole predictor, storing the residuals, and then correlating those residuals with the dependent variable Y. This result provides the interpretation to the slope β_2 that it is capturing the unique part of predictor G after the linear effects of the other predictors have been removed. So, we use the t-test of β_2 from the full regression to test the null hypothesis and we use the part correlation interpretation of β_2 to understand what these two regressions are doing. The computation of the part correlation can be done either through the square root of the difference in the two R^2s or through the correlation of residuals—both will yield identical values. This leads to a potential source of confusion that I will turn to soon.

Remember this important fact: ANCOVA purges the covariate from the grouping code predictor. This is the same logic of any regression—we always interpret a slope in a multiple regression in terms of the part correlation logic. This can be worded as the unique contribution of that variable added last in the sequence, or equivalently, in terms of saving residuals from a regression where the grouping variable plays the role of the dependent variable and using those residuals (the part of G that has been purged of the linear relation of covariate C) to predict the actual dependent variable.

Maxwell et al (1985) have a nice paper explaining some common misunderstandings, with equations, a numerical example and a tutorial figure, which I reproduce in Figure 9-7.

The correct interpretation of ANCOVA follows directly from the development of the part and partial correlations I introduced in Lecture Notes #8, where the inclusion of additional predictors modifies the other predictors already in the regression equation rather than modifying the dependent variable, which is the misunderstanding commonly found in the literature. Basically, ANCOVA and blocking factors control for other predictors and do not exclusively "adjust" the dependent variable.

Here is some R code to illustrate this correct interpretation. First, I run a regression that appears strange at first because it has the grouping variable G as the DV and the covariate C as the predictor. Second, I correlate the residuals from that regression (the unique part of the grouping variable having removed the linear effect of the covariate) with the actual dependent

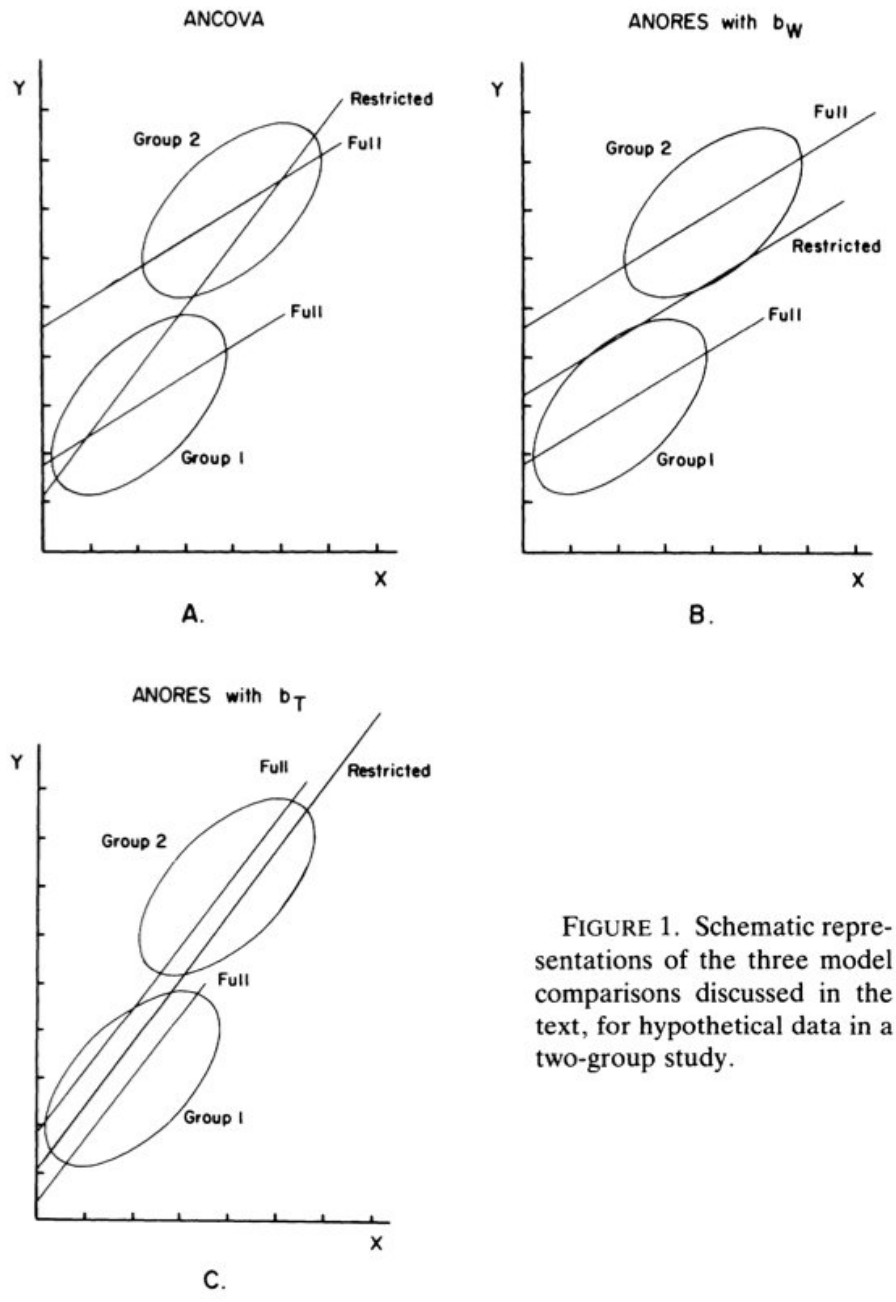


FIGURE 1. Schematic representations of the three model comparisons discussed in the text, for hypothetical data in a two-group study.

Figure 9-7: Figure from Maxwell, Delaney & Mannheimer (1985). Panel A is the correct ANCOVA representation with the restricted model being the slope that emerges from the model without group membership and the full model having different intercepts and a common slope. Panel C depicts the incorrect model of saving the residuals from the reduced model (where the residuals are taken from an incorrect reduced regression) and using those residuals as the dependent variable with the grouping as the predictor. Panel B is a variant of the regression on residuals from the dependent variable approach that has appeared in the literature. See their paper for more details.

variable V3.

```
out.lm1 <- lm(G ~ C, data = data2)
cor(data2$Y, resid(out.lm1))
```

```
## [1] -0.7843432
```

This correlation of -.784 has only removed the covariate from the predictor and has not touched the dependent variable. The claim I am making is that this correlation should be equivalent to the part correlation, which can be computed as the square root of the difference between the R^2 of the full regression that includes the grouping variable G and the covariate C as two predicitors and R^2 of the reduced regression that omits the grouping variable.

```
sqrt(summary(lm(Y ~ G + C, data2))$r.squared -
     summary(lm(Y ~ C, data2))$r.squared)
```

```
## [1] 0.7843432
```

```
#or done in pieces so you can see what is going on;
#R2 for full regression with both grouping and covariate
#as predictors
summary(lm(Y ~ G + C, data2))$r.squared
```

```
## [1] 0.9697785
```

```
#R2 for reduced regression with only the covariate
#as the predictor
summary(lm(Y ~ C, data2))$r.squared
```

```
## [1] 0.3545843
```

```
#sqrt of the difference of the two r squares
sqrt(0.9697785 - 0.3545843)
```

```
## [1] 0.7843432
```

```
#equals corr above; interpretation is verified
```

Those two correlations are equal.

Finally, to nail this, let me outright do what many people incorrectly think ANCOVA is doing and show it doesn't equal what ANCOVA is actually doing. Some people believe that if you run a regression with covariate predicting the DV (rather than the grouping variable G), save the residuals from that regression, and then perform a two-sample t-test using those residuals as the dependent variable that you are doing the same thing as ANCOVA. But this is false as I show. This incorrect approach correctly runs the first regression with the covariate as the sole predictor (wonderful so far), but then (this is where it falls apart) uses residuals from the first regression as the dependent variable in the second regression. As shown below, this is not equivalent to the ANCOVA, which if you want to use a "residuals logic" you need to run a different first regression where grouping code G is the dependent variable and covariate C is the covariate as described earlier in these notes followinng the Lecture Notes #8 logic for the part correlation.

```
# run regression with covariate predicting DV,
# save residual
wrong.resid <- resid(lm(Y ~ C, data2))

# residual is the part of the DV unrelated to the
# covariate; now perform a two-sample t test on
# the residuals
t.test(wrong.resid ~ data2$G, var.equal = TRUE)
```

```
##
##  Two Sample t-test
##
## data:  wrong.resid by data2$G
## t = 10.087, df = 8, p-value =
## 7.959e-06
## alternative hypothesis: true difference in means between group 1 and grou
## 95 percent confidence interval:
##  4.242824 6.757783
## sample estimates:
## mean in group 1 mean in group 2
##        2.750152       -2.750152
```

```
# equivalent version conducted as a regression
summary(lm(wrong.resid ~ data2$G))
```

```
##
## Call:
## lm(formula = wrong.resid ~ data2$G)
##
## Residuals:
##      Min       1Q   Median       3Q
## -1.23343 -0.69043  0.01672  0.72386
##      Max
##  0.93374
##
## Coefficients:
##             Estimate Std. Error t value
## (Intercept)   8.2505     0.8622   9.569
## data2$G      -5.5003     0.5453 -10.087
##             Pr(>|t|)
## (Intercept) 1.18e-05 ***
## data2$G     7.96e-06 ***
## ---
## Signif. codes:
##   0 '***' 0.001 '**' 0.01 '*' 0.05
##   '.' 0.1 ' ' 1
##
## Residual standard error: 0.8622 on 8 degrees of freedom
## Multiple R-squared:  0.9271,Adjusted R-squared:  0.918
## F-statistic: 101.7 on 1 and 8 DF,  p-value: 7.959e-06
```

The two sample t test on the residuals from the first regression that removes the covariate C from the dependent variable Y is 10.087. However, that is not the same as the t for the grouping code G from the ANCOVA, i.e., the t test for the grouping variable in the regression that includes both the grouping variable and the covariate as predictors is 11.94 (ignore the sign difference as that is just because of a difference of how contrasts are defined in the t.test() command versus how I coded the two groups in the regression).

```
# correct model we saw earlier reprinted here
summary(out.lm)
```

```
##
## Call:
## lm(formula = Y ~ C + G, data = data2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -1.2850 -0.1875  0.0425  0.2725  1.1400
##
## Coefficients:
##             Estimate Std. Error t value
## (Intercept)   9.2400     0.8487   10.89
## C             1.4250     0.1306   10.91
## G            -5.6550     0.4737  -11.94
##             Pr(>|t|)
## (Intercept) 1.22e-05 ***
## C           1.20e-05 ***
## G           6.59e-06 ***
## ---
## Signif. codes:
##   0 '***' 0.001 '**' 0.01 '*' 0.05
##   '.' 0.1 ' ' 1
##
## Residual standard error: 0.7387 on 7 degrees of freedom
## Multiple R-squared:  0.9698,Adjusted R-squared:  0.9611
## F-statistic: 112.3 on 2 and 7 DF,  p-value: 4.799e-06
```

And as you'd expect, the correlation of the residuals from this incorrect regression Y on C does not yield the same value for the part correlation we saw earlier of 0.7843432, not even close (ignore sign because of how I coded groups as 1 and 2).

```
cor(wrong.resid, data2$G)
```

```
## [1] -0.9628605
```

Key takeaway: always go back to basics and interpret the role of a covariate in terms of the usual part correlation logic used throughout multiple regression. The covariate is removed from the grouping variable, and then the part of the grouping variable that has been purged of the covariate is then correlated/regressed with the actual dependent variable.

7. General observations about ANCOVA

ANCOVA is a way to remove the linear effect of a nuisance variable through a statistical model. These nuisance variables are not manipulated. This is the main difference between how ANCOVA and randomized block designs are usually implemented. The former controls through statistical analysis. The latter design controls for blocking variables through experimental control, and that experimental control was used to reduce the error term. Most

ANCOVA applications do not offer the luxury of experimental manipulation and adjusting for nuisance through ANCOVA is about the only way to deal with nuisance variables.

It is possible to mix ANCOVA with everything we discussed this term. For instance, it is possible to add a naturally occurring covariate to a randomized block design (or even a latin square design). In the case of a randomized design with a covariate, there are two nuisance, or blocking variables—one that is manipulated and one that isn't. There could be random or nested effects; there could be repeated measures where the covariate is also repeated. Anything we've done so far can be extended to include covariates.

If a variable is relevant in your data but you ignore it as a covariate and just run a regular ANOVA or regression, you have what is called a misspecified model. A model that is misspecified could seem to be violating some assumptions.

For example, it is possible that what looks to you as a violation of the equality of variance assumption could really be due to a misspecified model. The presence of a covariate that isn't included in the model could influence the cell variances and make it appear that there are unequal variances. The methods we discussed last semester for dealing with unequal variances, such as spread-and-level plots and transformations, will typically not be able to handle unequal variances due to omitted covariates.

Another example is that there could be heterogeneity in the treatment effect. Different people respond differently to the treatment effect α, so that the treatment isn't a constant within a group as assumed by ANOVA (recall that the treatment effect α from Lecture Notes #2 is constant for all subjects in the same condition). If you happen to have measured the covariate that is related to this heterogeneity of treatment response, then the inclusion of that covariate in the ANOVA/regression model would eliminate the source of unequal variances. This is a completely different way of addressing the equality of variance assumption than we saw last semester (no transformation, no Welch, no nonparametric tests; just include the right covariate in the model).

The basic idea is that ANOVA assumes the treatment effect is a constant for every case in a cell. However, ANCOVA allows the treatment effect to vary as a function of the covariate; in other words, ANCOVA allows cases within the same treatment group to have different effects and so is one way to model individual differences. Two excellent papers by Byrk and Raudenbush (1988, *Psychological Bulletin, 104,* 396-04) and Porter and Raudenbush (1987, *Journal of Counseling Psychology, 34*, 383-92) develop this point more generally. Extensions to random effects regression slopes (where each subject has his/her own slope and intercept) allows for a more complete description of individual differences in regression. This is handled through multilevel models with many applications, including growth curve analysis as used in developmental psychology. We touched on multilevel models briefly last semester (e.g., nested effects ANOVA) and we will cover this in more detail later this semester.

8. The Big Picture: Why Do We Care About the General Linear Model?

benefits of the general linear model

We have now seen four advantages to performing hypothesis tests in terms of the general linear model. The general linear model

(a) Connects regression and ANOVA, providing a deeper understanding of ANOVA designs

(b) Provides the ability to perform residual analysis (examine how the model is going wrong, gives hints about how to correct misspecified models and violations of assumptions, allows one to examine the effects of outliers, allows one to see whether additional variables will be useful to include in the model)

(c) Allows one to combine what has traditionally been in the domain of ANOVA (categorical predictor variables) with what has traditionally been in the domain of regression analysis (continuous predictor variables) to create new tests such as ANCOVA

(d) Provides a foundation on which (almost) all advanced statistical techniques in the behavioral sciences are based (e.g., principal components, factor analysis, multivariate analysis of variance, discriminant analysis, canonical correlation, time series analysis, log-linear analysis, classical multidimensional scaling, structural equation modeling, multilevel modeling)

The first three points will have tremendous impact on the quality of your data analysis (deeper understanding of ANOVA, ability to talk intelligently about your model and its shortcomings, and, whenever possible, include covariates to reduce the random error).

Yes, it is more work in SPSS or R to get the dummy codes (or effect codes or contrast codes) into a regression analysis compared to the relatively simple ANOVA-style commands, but the benefits are worth it. It won't be long before statistics programs provide easy routines for creating dummy, effects, and contrast codes (competitors such as SAS already have built-in functions that automatically create the proper codes in a regression), but I have been waiting for several years for this to appear in SPSS and in a better way than is currently implemented in R (recall the issues with contrasts on factors that we had to address in the ANOVA section in R).

9. Binary data

The general linear model can be extended to distributions that are nonnormal. A commonly occurring nonnormal distribution is the binomial. There are situations when the dependent

variable consists of 0's and 1's and one is interested in the number of 1s per group (or equivalently, the proportion). Note that this differs from dummy codes that are predictors—we are now talking about the dependent variable being 0 or 1. This means that the dependent variable is nonnormal. The dummy codes we encountered when connecting ANOVA and regression were predictors not dependent variables and no distributional assumptions were made about those predictors. Computer programs let you run a regression with the usual syntax but change the distribution of the dependent variable (e.g., normal, binomial, Poisson). SPSS accomplishes this through the GENLIN command, and R accomplishes this through the command glm().

variance of a proportion

Here I'll cover some basics about data that follow a binomial distribution. I'll denote the sample proportion as p and the population proportion that p estimates as ρ (Greek letter rho). The variance of the sampling distribution of p is

$$\frac{\text{p (1-p)}}{\text{N}} \tag{9-10}$$

where N is the sample size that went into the computation of that sample estimate p. The square root of this variance is the standard error of p. Note something very important about the standard error of p: it is related to the value of p. Thus, if two groups have different proportions, then it is possible for the two groups to have different variances. I say possible because the standard error is symmetric around the proportion .5. So, the proportions .2 and .8 have the same variance when their sample sizes are equal, but the proportions .1 and .7 (which also have a difference of .6) each have different variances. This creates a major problem for our standard approach because if the standard error depends on p *we cannot make the equality of variance assumption*. If groups have different proportions, then they likely have different variances and the equality of variance assumption is violated before we do anything. So other procedures are needed and we turn to a few of them.

can't make the equal variance assumption with proportions

Another complication when dealing with proportions is that the regression model needs to be constrained to not produce out of bounds values. That is, it wouldn't make much sense to have $\hat{Y}$ from the structural model be -23 or 1.3 because proportions must be between 0 and 1, inclusive. Also, observed proportions at the boundaries 0 or 1 can wreak havoc if one blindly applies Equation 9-10 because the standard error of those proportions is 0. The problem of "division by zero" occurs when you divide an estimate by it's standard errors of zero.

(a) Comparing two binomial proportions—Chi-square test

You've probably heard about the χ^2 contingency table test. It appears in almost every introductory statistics textbook. It can be used to test two proportions. The test can be used, for instance, to test whether the proportion in group A differs from the proportion in group B.

The setup for the χ^2 contingency table test to compare two proportions involves creating

a 2×2 table of counts, such as the case of 100 males and 100 females distributed as follows

	B	
A	yes	no
male	20	80
female	90	10

Such a table results, for instance, when the dependent variable B is some yes or no question (each subject produces one binary observation), and the grouping variable A is gender. In this example, the data show that 20% of the men said yes and 90% of the women said yes. We want to compare these two proportions, 0.20 versus 0.90. The χ^2 test on contingency tables involves computing the expected frequencies in each of the cells under a null hypothesis that the rows and columns are independent. The test then compares the observed frequencies (the numbers in the table above) with those computed expected frequencies, and constructs a test of significance using the χ^2 distribution. For computational details see an introductory statistics text.

The SPSS command CROSSTABS provides such analyses. You select the row and column variables, under statistics select Chi-square test, and off you go. The syntax is

```
CROSSTABS
  /TABLES=sex BY DV
  /FORMAT=AVALUE TABLES
  /STATISTICS=CHISQ
  /CELLS=COUNT EXPECTED
  /COUNT ROUND CELL
```

which produces a table of both the observed counts and the expected counts. The value for the Pearson χ^2 and the corresponding *p* value are printed as well.

The CROSSTABS command is also useful for different measures of correlation that are applicable when one has ordinal or nominal data. For instance, it includes correlations such as gamma and lambda. Gamma is one of my favorite measures that isn't used much—it measures the extent to which "as one variable goes up, the other variable goes up." That is, it measures basic ordinal relations between pairs of variables—what most people incorrectly think the correlation coefficient of Lecture Notes #6 measures.

In R the basic Chi square command is chisq.test(). In addition to the hypothesis test it provides observed and expected counts.

The main limitation of the χ^2 test on contingency tables in both SPSS and R is that

it cannot be generalized to perform tests on several proportions. For example, if you have a factorial design, with a proportion as the cell mean and you want to test for main effects, interactions and contrasts, the standard χ^2 test on contingency tables won't help you. At best, the χ^2 contingency table test gives you an omnibus test saying that at least one of the observed proportions is different from the others, much like the omnibus F in a oneway ANOVA. (You have to do some tricks, which aren't implemented in most computer programs, to get anything useful out of the χ^2 contingency table test). Also, χ^2 contingency table analysis doesn't generalize readily to include things like covariates. So, instead I'll present a different approach—logistic regression—that pretty much allows you accomplish all the machinery of regression that we have covered so far but using binary data (i.e., assuming binary data rather than normally distributed data).

(b) Comparing two binomial proportions—logistic regression

There are several alternative formulations for logistic regression. For a complete review see a textbook on the Generalized Linear Model such as McCullagh and Nelder's book.

I'll present logistic regression for the simple problem of comparing two proportions. First, I need to review some of the building blocks. Logistic regression operates on the "logit" (aka log odds) rather than the actual proportion. A logit is defined as follows:

logit

$$\text{logit} \quad = \quad \log \frac{p}{1-p} \tag{9-11}$$

In other words, divide p by (1-p), this ratio is known as "odds", and take the natural log (ln button on a calculator).

variance of the logit

The asymptotic variance of a logit is

$$\text{variance(logit)} \quad = \quad \frac{1}{Np(1-p)} \tag{9-12}$$

where N is the sample size and p is the proportion. The square root of this variance is the standard error of the logit. The standard error of the logit suffers the same problem as the standard error of the proportion because it is related to the value of the proportion and, consequently, we cannot in general make the equal variance assumption.

Throughout these lecture notes, I mean "natural log" when I say "log". On most calculators this is the "ln" button. Being consistent with the base is important because the formula for the variance given above (Equation 9-12) depends on the base. If log base 10 is used instead of the natural log, then the variance of the "logit" is $\frac{1}{\ln(10)^2 Np(1-p)}$, where ln is the natural log. Of course, the value of the logit also depends on the base of the log and, luckily, the value of the logit and the standard error of the logit are each influenced by the base in the same way so when one divides the two to form a Z test, the effect of the base cancels giving the identical test result. To keep things simple, just

stick with the natural log, the "ln" key on your calculator, and all the formulas presented in these lecture notes will work correctly.

What do we do when we can't make the equal variance assumption? We can use a basic result from Wald that approximates a normal distribution. In words, here is a sketch of the result: a test on a contrast can be approximated by dividing the contrast $\hat{\text{I}}$ with the square root of the variance of the contrast. The difference here compared to what we did in Lecture Notes #1 is that we won't make the assumption of equal variances and we won't correct using a Welch-like procedure. We rely on Wald's proof that as sample size gets large, things behave reasonably well, and the estimate divided by the standard error of the estimate follows a normal distribution[5].

Let's look at Wald's test for the special case of comparing two logits. This test turns out to be asymptotically equivalent to the familiar χ^2 test on contingency tables applied to testing two proportions. The null hypothesis is

$$\text{population one logit - population two logit = 0} \tag{9-13}$$

which is the same as the (1 -1) contrast on the logits.

The contrast $\hat{\text{I}}$ is

$$(1)\log\frac{p_1}{1-p_1} + (-1)\log\frac{p_2}{1-p_2} \tag{9-14}$$

where the subscripts on the proportion refer to group number. Under the assumption of independence (justified here because I have a between-subjects design and the two proportions are independent), the standard error of this contrast is given by

$$\text{se(contrast)} \quad = \quad \sqrt{\frac{1}{N_1 p_1(1-p_1)} + \frac{1}{N_2 p_2(1-p_2)}} \tag{9-15}$$

that is, the square root of the sum of the two logit variances.

Finally, the test of this contrast is given by

$$Z \quad = \quad \frac{\hat{\text{I}}}{\text{se(contrast)}} \tag{9-16}$$

where Z follows the standard, normal distribution. If Z is greater than 1.96 then you reject the null hypothesis using $\alpha = .05$ (two-tailed).

[5]I once checked this for Type I error rates with simulations and the as "sample size gets large" starts to kick in when sample sizes are about 25 or 30 for proportions between .1 and .9. For proportions closer to 0 or 1, then the sample size needs to be more like 100 or more. I checked Type I error rates and not other aspects of the statistical test such as power. It may be possible that to achieve adequate power, sample sizes need to be even greater than they do for adequate Type I error protection. A question to be answered another day..., but a tractable one. This could be good project for someone with a good calculus and linear algebra background interested in learning more about mathematical statistics and simulations. Takers?

(c) Testing contrasts with more than two proportions

This Wald test can be generalized to more than two proportions. Take any number of sample proportions $p_1, \ldots, p_T$ and a contrast λ over those T proportions. Here are the ingredients needed for the Wald Z test. The value of $\hat{\text{I}}$ is (contrast weights applied to each logit, then sum the products)

$$\hat{\text{I}} = \sum_{i=1}^{T} \lambda_i \log \frac{p_i}{1-p_i} \tag{9-17}$$

The standard error of the contrast is

$$\text{se(contrast)} = \sqrt{\sum_{i=1}^{T} \lambda_i^2 \frac{1}{N_i p_i (1-p_i)}} \tag{9-18}$$

Finally, the Wald Z test for the null hypothesis that population I = 0 is simply

$$\frac{\hat{\text{I}}}{\text{se(contrast)}} \tag{9-19}$$

Again, the critical value for this Z is 1.96. If the observed Z exceeds 1.96, then the null hypothesis that the population I is 0 can be rejected using $\alpha = .05$ (two-tailed).

This test is the proportion version of the test I presented in Lecture Notes #6 for testing contrasts on correlations. Note the analogous structure: instead of transforming the correlation with Fisher's r-to-Z we take the log odds of the proportion. Aside from differences in the specific transformation and the resulting standard errors, everything else is pretty much the same as with the test on correlations.

(d) Example

Table 1 shows an example taken from Langer and Abelson (1972), who studied compliance to a favor. They used a different analysis than logistic regression but I'll illustrate with logistic regression. Behavior was coded in terms of whether or not the participant complied with the request. Two factors were manipulated: the legitimacy of the favor and the orientation of the appeal (i.e., either victim- or target-oriented), leading to a 2×2 factorial design. The summary statistic in each condition was the proportion of participants complying to the favor. Langer and Abelson were interested in testing whether the interaction between the two factors was statistically significant. Recall that

Table 1: Data from Langer and Abelson, 1972 (Study 1). The first table shows raw proportions of subjects who complied with the favor, the second shows proportions that have been transformed by the logit, and the third shows the interaction contrast.

Raw Proportions

	Orientation		
Legitimacy	Victim	Target	Marginal Mean
Legitimate	0.700	0.300	0.500
Illegitimate	0.350	0.500	0.425
Marginal Mean	0.525	0.400	grand mean = 0.462

Logit Transformation

	Orientation		
Legitimacy	Victim	Target	Marginal Mean
Legitimate	0.85	-0.85	0.00
Illegitimate	-0.62	0.00	-0.31
Marginal Mean	0.11	-0.42	grand mean = -0.15

Interaction Contrast

	Orientation	
Legitimacy	Victim	Target
Legitimate	1	-1
Illegitimate	-1	1

Note: Twenty subjects per cell. The 2×2 table of proportions is equivalent to a $2 \times 2 \times 2$ contingency table where the third variable is the dependent variable.

in the case of a 2 $\times$ 2 factorial design the interaction is identical to the (1, -1, -1, 1) contrast.

Plugging the numbers from Table 1 into the formula for the Wald test, yields an $\hat{\text{I}}$ = 2.31, a se(contrast) = .9466, and a Z = 2.44, which is statistically significant at $\alpha = .05$ (two-tailed).

(e) General Comments on Logistic Regression and Examples

It is relatively easy to perform these logit contrasts by hand or within a spreadsheet program like Excel. You may prefer doing this in SPSS or R, but it requires more setup because you have to convert the problem into a regression problem (e.g., define the appropriate contrasts as predictors)[6]. The benefit of SPSS or R logistic regression commands though is that they are more general and can incorporate more bells and whistles such as ANCOVA-like analyses, within-subjects factors, random and nested effects, etc.

Here are the Langer & Abelson data coded for the SPSS logistic regression. Data need to be entered subject by subject. The first column is the grouping code (1-4), one main effect contrast for victim/target, the second main effect contrast for the legitimacy manipulation, the interaction contrast and the dependent variable (1=compliance, 0=no compliance).

The syntax of the LOGISTIC command is similar to that of the REGRESSION command. The dependent variable is specified in the first line, then the /method=enter lists all the predictor variables. The structural model being tested by this syntax is

$$\text{logit(p)} = \beta_0 + \beta_1 \text{me.vt} + \beta_2 \text{me.legit} + \beta_3 \text{interact} \quad (9\text{-}20)$$

Note that the structural model is modelling the logit rather than the raw proportion, so its structural model is additive in the log odds scale.

I illustrate by presenting the entire dataset with codes entered directly into the file. The first column is the group code (useful for a later demonstration); columns 2, 3 and 4 are main effects and interaction contrast codes; column 5 lists the dependent variable.

```
data list free/ group me.vt me.legit interact dv.
```

[6]The SPSS logistic regression command and corresponding glm() command in R are quite general. The commands will accept predictors that are "continuous", unlike the simple "by-hand" method I present here that only tests contrasts so in regression parlance uses categorical predictors rather than continuous predictors. This feature could be useful in some applications, e.g., regressing teenage pregnancy (coded as "yes" or "no") on GPA, parent's salary, and the interaction between GPA and parent's salary. If you are interested more in how to perform regression-like analyses on binary data, I suggest that you take a more detailed course on categorical data analysis.

```
begin data
     1     1     1     1     1
     1     1     1     1     1
     1     1     1     1     1
     1     1     1     1     1
     1     1     1     1     1
     1     1     1     1     1
     1     1     1     1     1
     1     1     1     1     1
     1     1     1     1     1
     1     1     1     1     1
     1     1     1     1     1
     1     1     1     1     1
     1     1     1     1     1
     1     1     1     1     1
     1     1     1     1     0
     1     1     1     1     0
     1     1     1     1     0
     1     1     1     1     0
     1     1     1     1     0
     1     1     1     1     0
     2     1    -1    -1     1
     2     1    -1    -1     1
     2     1    -1    -1     1
     2     1    -1    -1     1
     2     1    -1    -1     1
     2     1    -1    -1     1
     2     1    -1    -1     0
     2     1    -1    -1     0
     2     1    -1    -1     0
     2     1    -1    -1     0
     2     1    -1    -1     0
     2     1    -1    -1     0
     2     1    -1    -1     0
     2     1    -1    -1     0
     2     1    -1    -1     0
     2     1    -1    -1     0
     2     1    -1    -1     0
     3    -1     1    -1     1
     3    -1     1    -1     1
     3    -1     1    -1     1
     3    -1     1    -1     1
     3    -1     1    -1     1
     3    -1     1    -1     1
     3    -1     1    -1     1
     3    -1     1    -1     0
     3    -1     1    -1     0
     3    -1     1    -1     0
     3    -1     1    -1     0
     3    -1     1    -1     0
     3    -1     1    -1     0
     3    -1     1    -1     0
     3    -1     1    -1     0
     3    -1     1    -1     0
     3    -1     1    -1     0
     3    -1     1    -1     0
     3    -1     1    -1     0
     3    -1     1    -1     0
     3    -1     1    -1     0
     4    -1    -1     1     1
     4    -1    -1     1     1
```

```
    4   -1   -1    1    1
    4   -1   -1    1    1
    4   -1   -1    1    1
    4   -1   -1    1    1
    4   -1   -1    1    1
    4   -1   -1    1    1
    4   -1   -1    1    1
    4   -1   -1    1    1
    4   -1   -1    1    0
    4   -1   -1    1    0
    4   -1   -1    1    0
    4   -1   -1    1    0
    4   -1   -1    1    0
    4   -1   -1    1    0
    4   -1   -1    1    0
    4   -1   -1    1    0
    4   -1   -1    1    0

logistic dv
 /method = enter me.vt me.legit interact
 /print all.
```

This syntax produces a χ^2 for the omnibus test of all predictors (much like the F test of total R^2 in a regression) and a table of coefficients and *t*-tests (much like βs in a regression). Recall that we computed a $Z = 2.44$ using the hand computation on the logit, and SPSS produces a "Wald test" for the interaction. Take the square root of the Wald test in the print out and you get the same result within roundoff error and convergence (the SPSS program is iterative rather than exact). That is, $\sqrt{5.9744} = 2.444$.

```
                    Chi-Square    df Significance

 Model Chi-Square         7.960     3        .0468
 Improvement              7.960     3        .0468


Variable            B    S.E.     Wald    df      Sig       R   Exp(B)

ME.VT           .1548   .2366    .4277     1    .5131   .0000   1.1674
ME.LEGIT        .2689   .2366   1.2911     1    .2558   .0000   1.3085
INTERACT        .5784   .2366   5.9744     1    .0145   .1897   1.7832
Constant       -.1548   .2366    .4277     1    .5131
```

The syntax for the LOGISTIC command has more bells and whistles than the REGRESSION command. For instance, it can define contrasts in the same way as the MANOVA command does, saving you time in creating dummy, contrast, or effect codes. I could have used this syntax in the previous example and I would have gotten the identical output. Note that only the grouping variable is used here; the individual contrast predictors are not used but are created directly in the syntax just like in the MANOVA command.

```
logistic dv
 /contrast(group) = special( 1 1 1 1
                             1 1 -1 -1
                             1 -1 1 -1
                             1 -1 -1 1)
 /method = enter group
 /print all.
```

Boy, ANOVA in the context of regression would be much easier in SPSS if the REGRESSION command also had this feature of being able to define contrasts on predictor variables!

It it a good idea to use the contrast=special command on ALL categorical variables (even ones with only two levels). This tells the logistic command to treat the variable correctly and helps guard against forgetting to treat categorical variables as categorical. Recall that when discussing anova in the context of regression (lecture notes 8), I brought up the issue that there should be as many predictors for each variable as there are levels of that variable minus one.

The logistic command allows one to call a variable categorical, which further helps clarify communication with the program. Below I've added the categorical subcommand. The output will be identical as above. The categorical subcommand is useful when you have lots of variables and you need to keep track of what is categorical and what isn't (you can write `/categorical = LIST_OF_VARIABLES`).

```
logistic dv
 /categorical group
 /contrast(group) = special( 1 1 1 1
                             1 1 -1 -1
                             1 -1 1 -1
                             1 -1 -1 1)
 /method = enter group
 /print all.
```

On some versions of SPSS you may get a warning message saying that the first row in contrast=special was ignored (i.e., SPSS is ignoring the unit vector–no big deal).

You can also use this simple syntax to specify ANOVAs as well (be sure to specify all the main effects and interactions). Here is the syntax for a three way ANOVA with factors A (four levels), B (two levels), and C (three levels):

```
logistic dv
 /categorical A B C
 /contrast(A) = special( 1 1 1 1
```

```
                          1 1 -1 -1
                          1 -1 1 -1
                          1 -1 -1 1)
/contrast(B) = special(1 1
                       1 -1)
/contrast(C) = special(1 1 1
                       0 1 -1
                       -2 1 1)
/method = enter A, B, C, A by B, A by C, B by C, A by B by C
/print all.
```

(f) Logistic Regression and R

This is a relatively easy transition in R. We use the glm command instead of the lm command. When the DV is a binary variable you use the glm (generalized linear model) command and indicate the data are distributed binomial. Here is a sketch syntax with three predictors.

```
output <- glm(DV ~ Predictor1 + Predictor2 + Predictor3,
                   family=binomial)
summary(output)
```

Everything from the lm() command carries over naturally to the glm() command. For example, if a predictor is a factor you can attach contrasts to the factor just like you've been doing with the lm() command. There is an analogue in the lme4() package to allow random and nested factors in the context of a generalized linear model; this command is glmer().

Here is the helping example I showed.

```
out.glm <- glm(dv~ me.vt + me.legit + int, data=data,
               family=binomial)
summary(out.glm)
```

```
##
## Call:
## glm(formula = dv ~ me.vt + me.legit + int, family = binomial,
##     data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q
## -1.5518  -0.9282  -0.8446   1.1774
##     Max
##  1.5518
```

```
##
## Coefficients:
##             Estimate Std. Error z value
## (Intercept)  -0.1548     0.2366  -0.654
## me.vt         0.1548     0.2366   0.654
## me.legit      0.2689     0.2366   1.136
## int           0.5784     0.2366   2.444
##             Pr(>|z|)
## (Intercept)   0.5131
## me.vt         0.5131
## me.legit      0.2558
## int           0.0145 *
## ---
## Signif. codes:
##   0 '***' 0.001 '**' 0.01 '*' 0.05
##   '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 110.45  on 79  degrees of freedom
## Residual deviance: 102.49  on 76  degrees of freedom
## AIC: 110.49
##
## Number of Fisher Scoring iterations: 4
```

The p-values are the same as in SPSS. If you square the Z value in R you get the identical test statistic as SPSS calls "Wald" (example, $1.136^2 = 1.29$).

(g) Caveat: important issue of error variance.

One needs to be careful when comparing coefficients across logistic regression models, or even coefficients across groups. The reason is that the error variance in logistic regression is a constant. Every logistic regression has the same error variance equal to $\frac{\pi^2}{3} = 3.29$. This means that as one adds more variables to a logistic regression, the scale changes and a β in one regression is on a different scale than one in a different regression. This could make it look like a slope changed, or that slopes across groups differ, when in reality the difference may be due to scale. This is not at all intuitive or obvious, and it has created headaches across many literatures where papers, for example, claim differences between groups when it could just be an issue of different scales (an oranges and apples comparison). Traditional normally distributed regression does not have this problem because the error sum of squares and the explained sum of squares are both estimated from the data and must sum to the same amount—the sum of squares of the dependent variable.

10. Propensity scores

propsensity score

The concept of propensity scores is relatively new and quite powerful. It works well when there are many covariates, it isn't as restrictive as ANCOVA, it makes clear when there is or isn't overlap in the covariate across the groups, and it makes interesting use of logistic regression as a first step to remove the effect of covariates. I present it now because we just learned about logistic regression and ANCOVA's goal of removing covariation due to variables outside of our control.

Let's assume we have two groups and two covariates for this simple example. We want to examine if there are differences in the group means on the dependent variable controlling for the two covariates, but want something better than ANCOVA. What if we could match participants in group 1 with counterparts in group 2 that have similar values on the covariates, in this way "controlling for" the covariates.

One runs two regressions in propensity score analysis. The first regression is a logistic regression where the binary code for group membership (0 and 1) representing the two groups is the dependent variable and all covariates are the predictors. The purpose of this regression is to get predicted values (the $\hat{Y}$s) of group membership based on only using information from the covariates. That is, use the covariates to predict actual group membership. Two subjects with the same or similar $\hat{Y}$ have the same "propensity" of being assigned to a particular group. The same propensity can occur with multiple combinations of covariates. More deeply, if two people have the same propensity score, they have the same "probability" of being assigned by the covariates to the group. If one of those individuals is actually in one group and the other is in the second group, it is as though two individuals "matched" on the covariates by virtue of having the same propensity score were assigned to different groups. Not quite the same as random assignment but an interesting way to conceptualize matching because we can then claim that participants in different groups but having the same propensity score are in some ways "equated" across the possible values of the covariates, something that the approach of random assignment tries to accomplish through a design rather than statistical approach.

Second, you conduct a different regression that controls for the propensity scores. There are several ways one can do this. For building intuition I'll first mention one method. Conduct a regular ANCOVA with the dependent variable of interest and the grouping variable as the predictor, but don't use the actual covariate(s) in this regression, instead use the propensity score (the $\hat{Y}$s) from the first regression as the sole covariate. This allows one "to equate" nonexperimental groups in a way that resembles random assignment. But I mentioned that method just for intuition as it isn't necessarily the optimal approach; this particular version is not currently in favor and propensity score proponents would argue that it misses the point of using propensity scores.

Instead, there are more complicated procedures for the second regression where the propensity scores are used to create blocks (or strata), and then a randomized block design is used

rather than an ANCOVA with the grouped propensity scores playing the role of the blocking factor. There are many, many more sophisticated ways in which one can match on propensity scores and this is currently a major topic of research.

There is a nice paper by D. Rubin on propensity scores in 1997, *Annals of Internal Medicine, 127,* 757-63, and a complete book on the topic, Guo & Fraser (2010), *Propensity Score Analysis*. A paper by Shadish et al (2008, JASA) conducts an experiment where subjects are either assigned into an experiment (randomly assigned to one of two treatments) or a nonrandomized experiment (they choose one of two treatments). Overall results suggest that propensity score methods do not work as well as one would have hoped in correcting for bias in the nonexperimental portion of the study (i.e., the self-selection bias that was studied in this paper); there are too many ways to compute, calibrate, etc., propensity scores with little consensus on the optimal approach. Propensity score approaches, do have a role at the study design stage if one wants to match subjects. If you can't randomly assign in your research domain keep an eye out for new developments in the propensity score literature as this is one of the more promising ideas in a long time, though as Shadish et al point out it just isn't there yet.

11. Difference-in-Difference Estimation

ANCOVA is one of the earliest methods of applying statistical control to arrive at better estimates, there has been much progress the last few decades at developing newer approaches to address confounds in data and hopefully arrive at better estimates of the effects of interest. Some researchers refer to these methods as causal analysis, but that may be a stretch because the bottom line is we are still working with correlations and regressions, just in fancier ways. One such new method is the propensity score approach I just covered. Propensity score analysis attempts to accomplish statistically what we were unable to accomplish through the design of the study because random assignment could not be done. It is an open question whether the specific method of propensity score analysis did an adequate job of matching in order to mimic random assignment.

One common component to all these approaches is that they try to make clear mathematically the role that random assignment plays in the ability to make causal inferences. We see this intuitively in the case of propensity score analysis, where random assignment creates groups that are in some sense equated on all the variables that we cannot control. Propensity score analysis is a step in the direction of implementing this logic through a statistical model. If we have all the right covariates, the logistic regresison part computes a propensity score, and then we can match participants on that score, and compare across participants who are in different groups but have the same propensity score. Of course, we have to assume we have all the right covariates included in the logistic regression, they are measures appropriately, etc. Not perfect, and not necessarily the clear road to causality, but a reasonable start to improve the conclusions we draw from data and nonexperimental designs.

SUTVA

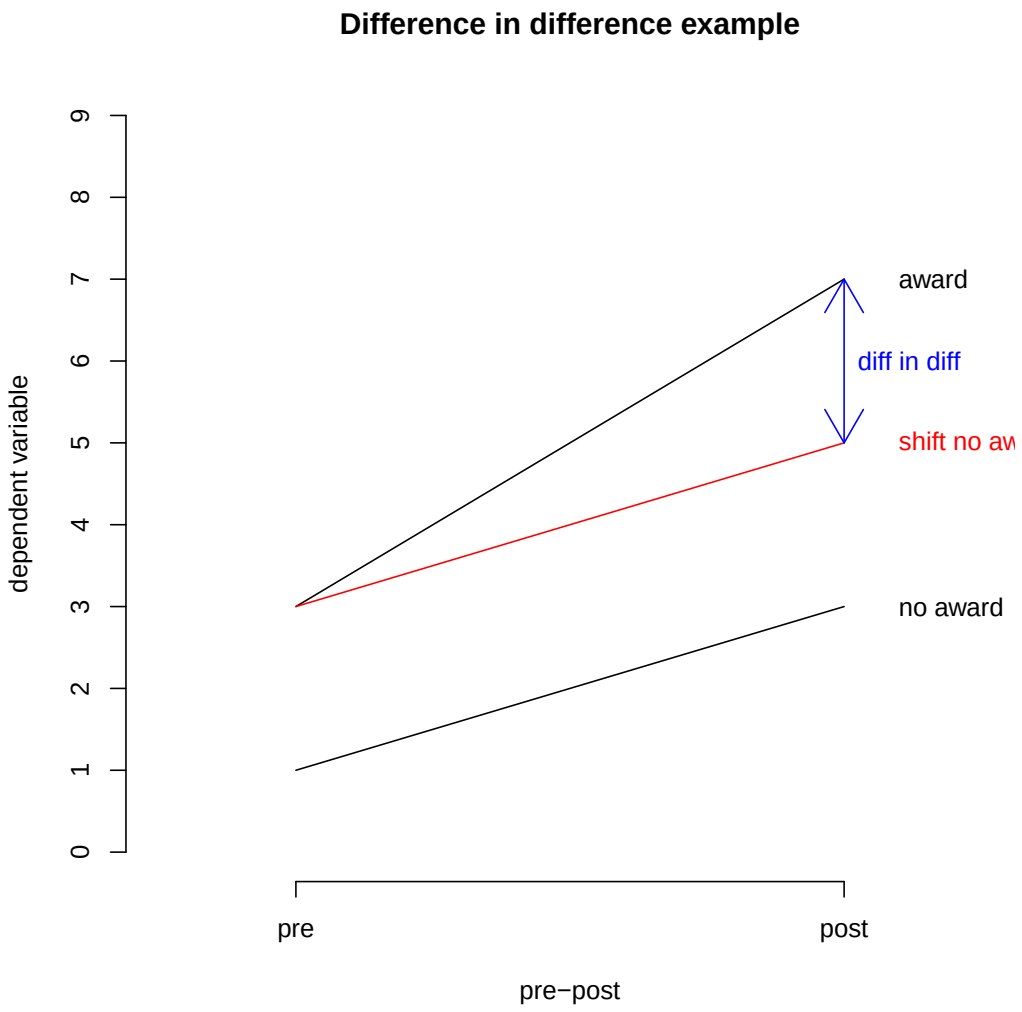
If we are willing to make more assumptions, then we can do more than basic ANCOVA. One such assumption in many of these approaches to causal analysis on nonexperimental data is the Stable Unit Treatment Value Analysis (SUTVA), which basically puts a restriction on "potential outcomes." A potential outcome is the counterfactual of the score the subject would have if assigned to treatment A and the score they would have if assigned to treatment B. Once the subject is assigned to a condition in an experiment, then their score for that treatment is revealed/observed (and the score for the treatment they were not assigned to is not revealed/observed). SUTVA says that these potential outcomes are not affected by the treatment conditions other participants were assigned. SUTVA, along with other assumptions, is what gives random assignment "teeth." The logic of applying this approach to nonexperimental data is that one can take their specific context, see which of the assumptions implied by random assignment does not hold in their setting and then use a technique that addresses those assumptions that do not hold.

Let's apply this setup to a different type of setting where we can make stronger claims than mere correlations, even though we do not have an experimental design with random assignment. This method is called difference-in-difference and it amounts to a 2 x 2 ANOVA where one factor is between-subjects and the other factor is within-subjects (aka repeated measures). Suppose we have two groups of subjects each measured at two points in time, such as in a pre-post design. One group of subjects had something happen to them (somewhat analogous to receiving a treatment) and another group of subjects did not have that thing happen to them, such as one group of academics received an award and another group of academics did not receive an award. Award was not randomly assigned yet the researach question is to examine the role of the award on some measure such as the total grant dollars they generated or the number of citations generated. There are many ways in which these two groups of academics could differ. We could measure those things and use those variables as covariates in an ANCOVA or use propsensity score matching to control for them and hopefully get a cleaner answer to the role an award has on the outcome measure.

This pre-post structure though offers a different way than ANCOVA or propensity score analysis to compare the effect of those who received an award and those who did not receive an award. We have two "groups" each measured twice, hence a 2 x 2 structure. If we are also willing to make additional assumptions, such as a parallelism assumption (aka parallel trends assumption) that the trend among those who did not receive an award is the same trend (parallel lines) as we would have seen (the counterfactual) among had those who received an award had not received an award. Under this assumption and SUTVA and other assumptions, the group that did not receive an award serves as a type of control for those who did receive an award. The interaction in that 2 x 2 design, which like all ANOVA interactions, is a "difference between differences" as we saw last semester. So, if we take the difference in means on the dependent variable between pre and post for those who did not receive an award and treat that as a type of "this is the degree of change expected even in the absence of an award" we can subtract that from the difference in means on the dependent variable between pre and post for those who did receive an award (hence the difference in difference idea that gives this method its name). We compare the degree of change in those who received

an award to the degree of change in those who did not receive an award to yield a measure of the "effect" of receiving an award on the dependent variable (such as the effect of receiving an award on grant dollars generated or number of citations or number of publications). This interaction term, the same term we would have computed last semester when learning about 2 x 2 ANOVA designs with one between-subjects factor and one within-subjects factor, can be given an interpretation that resembles a causal interpretation even on nonexperimental data (award status was not manipulated) as long as these additional assumptions hold. This approach is not so much about introducing new methods, but highlighting what additional assumptions need to be made to make stronger inferences in a nonexperimental setting.

Here is an illustration of the differenec-in-difference approach using the hypothetical award example. If the same process in the no award group (acting as a type of control condition) applies to the award group, then the line for the award group should be the red line, which is just the no award line shifted up to the pre-level of the award group maintaining the same slope. The difference between the actual post mean and the expected post mean for the award group is the difference-in-difference estimate (denoted in blue), and the interaction term provides a convenient estimate.



Difference-in-difference has been extended to more complex designs than simply 2 x 2s, and is a hot area of research that is progressing rapidly.

To learn more about these types of reasoning and the associated analyses that go along with them, check out some general causal inference books such as one by Imbens and Rubin (Causal Inference for Statistics, Social and Biomedical Sciences) or one by Morgan and Winship (Counterfactuals and Causal Inferences). These general references review many of these techniques, such as propensity score analysis, difference-in-difference, and others such as instrumental variables, graphical analysis, regression discontinuity and inverse probability weighting. Of course, there are also book length treatments on each of these methods, such as I mentioned above in the propensity score section. I also suggest the classic paper on strong inference by Platt (1964, *Science*). One of the points of Platt's paper is how "good science" progresses by rigorously testing alternative explanations, the methods discussed in this

subsection provide several approaches in the context of nonexperimental designs to testing alternative explanations and exluding explanations that are not supported by data.

12. Cross-validation

A criterion that has received more attention lately is the ability for models to predict out of sample. That is, how well can a model predict new data it has not seen. Most of traditional statistics has focused on analyzing the entire data set in the spirit of using all the relevant information to estimate parameters, build confidence intervals and test hypotheses. However, a reasonable criterion of a model is that it predict new data points it has not seen. This criterion has been made popular recently by various advances in Bayesian statistics as well as big data/data science methods, where there are additional parameters in the model (such as tuning parameters) that can't always be directly estimated from data but can be evaluated in their ability to predict out of sample.

I'll go into cross-validation in more detail later in the year, but here I'll show how to use cross-validation in the context of general linear model with lm in R. This can easily be applied to any other type of regression model we learned like generalized linear model.

The basic idea of cross-validation is that you split the sample in some way (such as 80/20 where 80% of the sample becomes the training sample where the model is fit and the remaining 20% becomes the test sample on which the predictions of the model are tested), fit the model on the training part, then see how well the model does on the hold out sample. We can evaluate predictive accuracy on the test sample by using R^2 or other measures such as the square root of the mean residuals in the test sample (aka RMSE, which stands for "Root Mean Squared Error") and the mean absolute error (MAE).

There are several approaches to doing the cross validation, such K-fold cross-validation and repeated K-fold cross-validation. The former involves randomly splitting the data into K subsets and using one of the subsets as the testing data set (e.g., if K is 10, then the model is fit on 9 of the subsets and tested on the 10th), then you repeat this process until each of the K subsets play the role of the testing data set and then average the predictive accuracy measure over the K runs. In other words, you hold out one of the K subsets, fit the model on the remaining K-1 subsets, test on the hold out subset, and repeat until you've fit the model on all ways to form K-1 subsets. The repeated K-fold cross-validation involves doing K-fold cross-validation multiple times and averaging over all the repeats of the K folds.

There is nice R package, caret, that makes it easy to accomplish cross-validation. I'll illustrate with a simple regression with one predictor. I'll use the iris data set in R to predict Sepal.Length from Petal.Length of three species of flowers. I'll set seed of the random number generator in order to get reproducible results. The caret package works with many of the functions available in R and we'll see a variety of uses later in the term. While my example

uses the lm regression command, the caret package can currently estimate over 230 different type of models including glm, Bayesian models, and various clustering programs. The argument savePredictions=T allows the prediction in the holdout Fold to be saved. You can view the predictions by examining the object model$pred.

```
library(caret)
set.seed(12345)
#k-fold cross-validation; k=10
train.control <- trainControl(method = "cv", number = 10,
                              savePredictions=T)
model <- train(Sepal.Length ~ Petal.Length, data = iris,
               method = "lm", trControl = train.control)
print(model)
```

```
## Linear Regression
##
## 150 samples
##   1 predictor
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 135, 135, 136, 134, 134, 135, ...
## Resampling results:
##
##   RMSE       Rsquared  MAE
##   0.4057148  0.766142  0.3298002
##
## Tuning parameter 'intercept' was
##  held constant at a value of TRUE
```

```
#repeated k-fold cross-validation; k=10 with 5 repeats
#similar to above but repeated 5 times and averaged
train.control <- trainControl(method = "repeatedcv",
                              number = 10, repeats=5,
                              savePredictions=T)
model <- train(Sepal.Length ~ Petal.Length, data = iris,
               method = "lm",
               trControl = train.control)
print(model)
```

```
## Linear Regression
```

```
##
## 150 samples
##   1 predictor
##
## No pre-processing
## Resampling: Cross-Validated (10 fold, repeated 5 times)
## Summary of sample sizes: 136, 134, 135, 135, 135, 135, ...
## Resampling results:
##
##   RMSE      Rsquared   MAE
##   0.402308  0.7730008  0.3293196
##
## Tuning parameter 'intercept' was
##  held constant at a value of TRUE
```

13. Time series

If time permits.... Okay, not funny if you are a time series aficionado.

We have already covered a major type of time series analyses when we did repeated measures ANOVA in LN #6 and a little when I showed how to do a paired t test in the context of regression (LN #8). There is much more to time series; I'll cover some in class and will return later in the semester. The basic ideas:

(a) Smoothing

Sometimes data analysts like to smooth a time series to make it easier to understand and find underlying patterns. There are many ways of smoothing time series data. Figure 13a shows some blood pressure data. The raw data (points) look jagged and don't have much of a visible pattern, but a smoothed version suggests a systematic, underlying pattern. Smoothed data can make it easier to see patterns, but can also hide noisy data. I like to plot both data and model fits.

(b) Correlated error

A key issue with time series is that the residuals are correlated. For example, an error at time 1 could be related to the error at time 2, etc. There are several ways of dealing with correlated error, or the violation of the independence assumption. We saw a few of them when doing repeated measures. At one extreme we just ignore the violation and pretend the data are independent. This amounts to a covariance matrix over time where

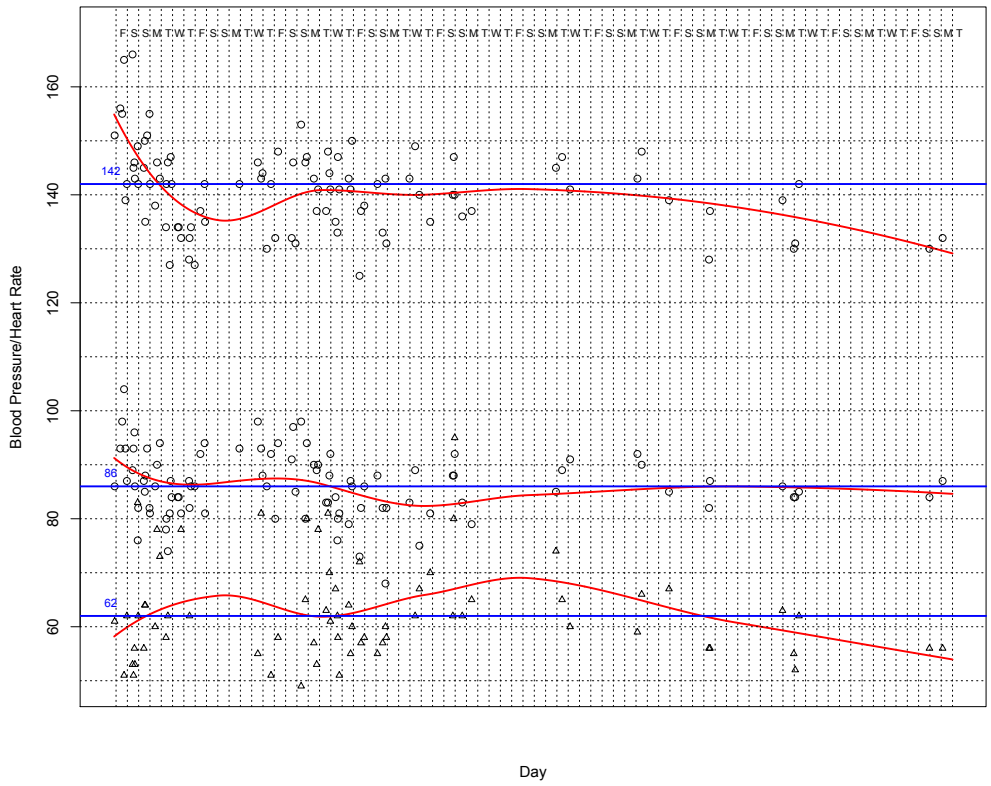


Figure 9-8: Smoothing time series data. Blood pressure and heart rate of one subject. Horizontal lines are the means; curves are smoothed prediction models.

all the off diagonals (covariance of time i with time j) are zero. At the other extreme we allow all covariance terms to be different and estimate them as parameters. This is what we did in repeated measures under the "multivariate" form, also known as unstructured. This is the method I advocated in the repeated measures section of the course because it bypasses pooling error terms so no need to make sphericity or compound symmetry assumptions. By focusing on contrasts in that setting we created special error terms for each contrast and we didn't have to make assumptions about the structure of the error covariance matrix.

Somewhere between those two extremes are several other options. In repeated measures ANOVA we covered one, where we assumed compound symmetry, but there are others. For example, one can assume a structure where adjacent times are correlated but all other correlations are zero. There are many, many different error structures that one can assume and/or test in a data set; you can look at the SPSS syntax help file for MANOVA or GLM for different assumptions one can make on the correlated error term. The standard approach is to test these approaches against each other, settling on one approach that balances the fit of the data with the increase in extra parameters that need to be estimated. There is much work that can be done in this area and opportunity for developing new analytic approaches.

(c) Dealing with correlated error

One approach to dealing with correlated error is to transform the problem to reduce the issue of a correlated error. The Kutner et al book goes into this detail. Essentially one takes a difference score of the Y variable, a difference score of the X variable (i.e., both are lag 1) and uses those "transformed" variables in the regression instead of the raw variables. The difference score of lag 1 refers to using the score minus its previous time's value (such as Y_2 - Y_1 where the subscript denotes time). There are also variants of this approach where instead of taking differences, one takes a weighted differences as in $Y' = Y_{i+1} - \rho Y_i$. A difference is still taken but the earlier time is weighted by ρ. If $\rho = 1$, then this is the same as a lagged difference score[7].

Another approach is to model the correlations directly. This can be done in several ways such as using random effects multilevel models, by using a generalized least squares approach, by moving to different estimation procedures such as GEE, or using some classic approaches involving variance components (generalizations of the ANOVA concepts). Lots of options here. I'd say consensus is building up in using multilevel models (aka random effect models like we saw in ANOVA), which can be estimated with either traditional methods like we have seen in this class or Bayesian approaches, and perhaps the GEE technique coming in second in some literatures like in public health.

[7]The idea of taking such lagged differences can begin to approximate derivatives. The difference of two lagged differences approximates a second derivative, etc. Such tools can come in handy when modeling nonlinear time series and can provide new insight into the interpretation of the parameters.

(d) Indices of correlated error

As we saw in the repeated measures world with the ϵ measure, there are various approaches to estimating the interdependence over time. The Durbin-Watson approach is relatively common in time series. The D-W measure is the ratio of sum of squared residual differences over the sum of squared residuals ($\frac{\sum_{t=2}^{n}(\epsilon_t - \epsilon_{t-1})^2}{\sum \epsilon^2}$). Kutner et al provide a table on which to compare statistical significance of the D-W measure, which is based on the sample size and the number of predictors in the regression equation. The D-W measure can then be used to make adjustments to the data to deal with the correlated temporal structure in an analogous way that a Welch or Greenhouse-Geiser makes adjustments. I don't find this approach useful because correlated error comes with the territory of time series. It isn't a problem with the data but a feature. I feel better modeling it directly and trying to understand it.

(e) General frameworks

A general framework for working with time series data is called the ARIMA approach. This allows correlated errors, lagged predictors, the ability to model "shocks", and a whole host of other processes. There are entire textbooks and semester long courses on ARIMA models, so this topic is beyond the scope of this course.

As I said earlier, a complete treatment of time series requires at least a semester course. I suggest taking such a course if you think you'll need to do time series analyses in your research. A key issue is how many time points you typically encounter. If you analyze 3-8 time points per subject, then probably repeated-measures style procedures, including modern multilevel modeling approaches that we'll cover later in the term, will suit you well. If you collect 100s or 1000s of data points per subject as one does in EEG, fMRI or with sensor data (fitbits, heart rate sensors, etc), then a more sophisticated time series course would be relevant for you.

In addition to statistical ways of analyzing time series data, there are also various techniques that are commonly used in engineering and in physics such as signal processing procedures like filtering, decomposition, convolutions, and other approaches that have also been useful in social science data.

14. Even more general models

Over the last decade there has been a proliferation of models more general than the general linear model. I list two here:

(a) Generalized Linear Model: allows distributions other than the normal; the user specifies the underlying distribution (e.g., normal, binomial, Poisson, beta, etc.); the logic is the same as multiple regression but without the constraint that the data be normal; the user needs to know, however, the distribution of the data; by generalizing the normal-normal plot to work with any theoretical distribution, the user can get a good handle on what the underlying error distribution is for the data under study. For more detail see McCullough & Nelder, *Generalized Linear Models*.

Logistic regression is one example of a generalized linear model as is the usual regression with a normally distributed error term, but there are many more distributions that fall under this family of techniques.

SPSS now implements the generalized linear model in GENLIM. R has the generalized linear model in the glm() command.

There is even a more general model that allows random effects in the context of the generalized linear model, sometimes called the generalized linear mixed model (GLMM). This allows you to do everything you already know how to do with regression, extend it to nonnormal distributions like proportions and count data, and also deal with random effects. An example where random effects is useful is when you want each subject to have their own slope and intercept, as in latent growth models. The generality of the GLMM is that you can do this for any type of distribution in the family. This isn't implemented in SPSS yet, though it is implemented in SAS, R, Stata, and a few other specialized programs. The R packages nlme and lme4 both handled GLMM models; there are also some specialty packages in R that offer variations on the GLMM framework.

Finally, a different type of generalization is to let the program search for different types of subjects, cluster subjects and then analyze the data based on those empirically derived clusters. This is useful, for instance, if you think there are different classes, or clusters, of subjects such that within a cluster the slopes and intercept are the same but across clusters they differ. But you don't know the clusters in advance. This technique, known as a mixture model, can help you find those clusters. This type of analysis is commonly used in marketing research to find empirically-defined market segments (groups of people who share common slopes and intercept). It is also used in developmental psychology to cluster different growth curves, or trajectories (i.e., kids with similar growth patterns). Once a cluster is in hand, then the cluster can be studied to see if there are variables that identify cluster membership. This technique can also include random effects so that not all members of the same cluster have the identical slopes and intercepts. We'll first need to cover clustering (LN #10), and then if there is time we can return to this technique at the end of the term. The computational aspects of this type of technique are complex and require specialized algorithms that are very different from those based on the central limit theorem that we have used in this class. Plus we need

to learn how to work with algorithms that offer numerical approximations (LN#10 and LN#11) before delving too deeply into mixture models.

(b) Generalized Additive Model: these are my favorite "regression-like" models because they make so few assumptions. They are fairly new and statistical tests for these procedures have not been completely worked out yet. An excellent reference is Hastie & Tibshirani, *Generalized Additive Models* (GAM). Even though they are relatively new it turns out that there are several special cases of this more general approach that have been around for decades, such as monotonic regression that we will cover in LN10. This GAM framework provides a scientific justification for various techniques that have been lingering around without much theoretical meat until now.

These models are "nonparametric" in the sense that they try to find an optimal scaling for each variable in the equation for an additive equation (under a variety of error models). Thus, one is scaling psychological constructs and running regression simultaneously, so doing scaling, prediction, estimation and testing all in one framework. We won't cover generalized additive models as a complete topic in this course. We will cover a cousin of this technique, monotonic regression, when we take up the topic of multidimensional scaling in lecture notes #10. R has the function gam() and a growing list of packages that can fit algorithms in this general approach.

15. Statistics and Uncertainty—Deep Thoughts

There are different viewpoints about the central problem in statistical inference. They differ primarily on where the locus of uncertainty is placed. Fisher advocated a view where all uncertainty is in the data. The unknown parameter (such as μ or ρ or β) is held fixed and the data are observed conditional on that fixed parameter. His framework amounted to comparing the observed data to all the other data sets that could have been observed. That is, he computed the probability that the data sets that could be observed would be more extreme than the data you actually observed. An example of this test is "Fisher's exact test" for a 2×2 contingency table. Fisher did not have anything like power or Type I/Type II error rates in his framework. For him, the p-value refers to the proportion of possible data sets that are more extreme than the one actually observed given the null hypothesis (i.e., the unknown parameter held fixed). It turns out that Fisher derived ANOVA as an approximation to the permutation test, which gives the probability of all results more extreme than the one observed and was too difficult to compute by hand in the early 1900s. So Fisher figured out that by making a few assumptions (independence, equal variances and normally distributed error) one could approximate the permutation test. If computers were available in the early 1900s we probably would not have ANOVA and the general linear model because the computers would have been able to compute the probability of data more extreme than the actual data. This view also leads to the admittedly weird wording of confidence intervals we saw in LN1 (95% of possible intervals include the true value).

A different viewpoint, the Bayesian approach, suggests that there is relatively little uncertainty in the observed data. Everyone can see the data you observed, and analyses should not be based on datasets you could have observed but didn't, which is what Fisher advocated. For Bayesians, the uncertainty arises because we don't have knowledge about the unknown parameter (such as μ or ρ); that is, our uncertainty is about the hypothesis we are testing. Bayesians impose distributions over possible values of the parameter and they use the observed data to "revise" that distribution. When the initial knowledge is "uniform" in the sense that no single parameter value is more likely than any other parameter value (aka noninformative prior), then for most situations Fisher's approach and the Bayesian approach lead to identical p-values and intervals (the difference is in how those p-values and intervals are interpreted). One advantage of the Bayesian approach is that it formalizes the process of accumulating information from several sources. For instance, how to incorporate the information from a single study into an existing body of literature. There is a sense in which we are all Bayesians when we write a discussion section and we use words to integrate our findings with the existing literature. There are many different "flavors" of Bayesian thinking so the phrase "I used a Bayesian approach" is not very informative; one needs to explain the prior, the likelihood, the numerical algorithm, convergence criteria, etc., for that statement to have any meaning. A few people claim we should completely move away from hypothesis testing and use Bayesian approaches, but this is an empty statement unless more details are provided.

The Fisherian and Bayesian approaches are merely two ways of addressing the problem of uncertainty. Fisherians place it on possible data we could have observed and Bayesians place it on our hypotheses.

A third viewpoint, that of Neyman & Pearson, can be viewed as a compromise position between Fisher and the Bayesians. This viewpoint acknowledges uncertainty in both data and hypothesis. They are the ones who introduced the idea of power and Type II error rates. For them, they take the view that there are two hypotheses to consider (the null and the alternative; Fisher only had the former) and both are equally likely before you run the study (so they have uncertainty between two hypotheses, but the Bayesians allow for more complex priors than the basic Neyman-Pearson definition). Further, N&P take Fisher's view separately for the null and for the alternative hypothesis (i.e., run through Fisher's argument first fixing under the null then a second time fixing under the alternative). They propose doing something that amounts to comparing a full model (the alternative hypothesis) to a reduced model (the null hypothesis). For those of you who read Maxwell and Delaney for the ANOVA portion of the course, you will recognize this type of model comparison.

I'll summarize this discussion with notation, though abusing symbols a little bit. Let X be the observed data and θ be the vector of parameters in the structural model (e.g., all the βs in a regression structural model could be organized into a column of parameters called θ). The function f(X$|\theta$) denotes the likelihood that data X are observed given the parameters in θ.

Fisher thought everything needed for statistical inference is in likelihood f(X|θ). All we have to do is find the values of θ that maximize the likelihood (hence the term "maximum likelihood"), and usual method from calculus can be used to estimate the parameters in θ and their standard errors.

The Bayesians, however, argue for something proportional to the likelihood: f(X|θ)g(θ), where the g function is the prior distribution over the parameters θ. The Bayesians have to deal with what they mean by the function g, and this is hotly debated with many considering this problem the "Achilles heel" of the Bayesian approach. A common "g" is that possible values of θ are equally likely. Of course, as more data come in, it is possible to use that information to hone in on possible values of the parameters θ. At that point, the statement "all values of θ are equally likely" wouldn't make much sense (because you are honing in on the value) so then the function g would need to be altered to account for the accumulated information. With simple rules of probability it can be shown that the Bayesians are really dealing with the converse of Fisher: the Bayesians are interested in the likelihood of the parameter θ given the observed data (f(θ|X)) whereas Fisher was interested in the likelihood of the data given the parameter θ (f(X|θ)). A subtle but important distinction; it makes a huge difference what is to the left and to the right of that vertical bar.

Neyman/Pearson argued for a comparison of two hypothesis, which can be summarized as

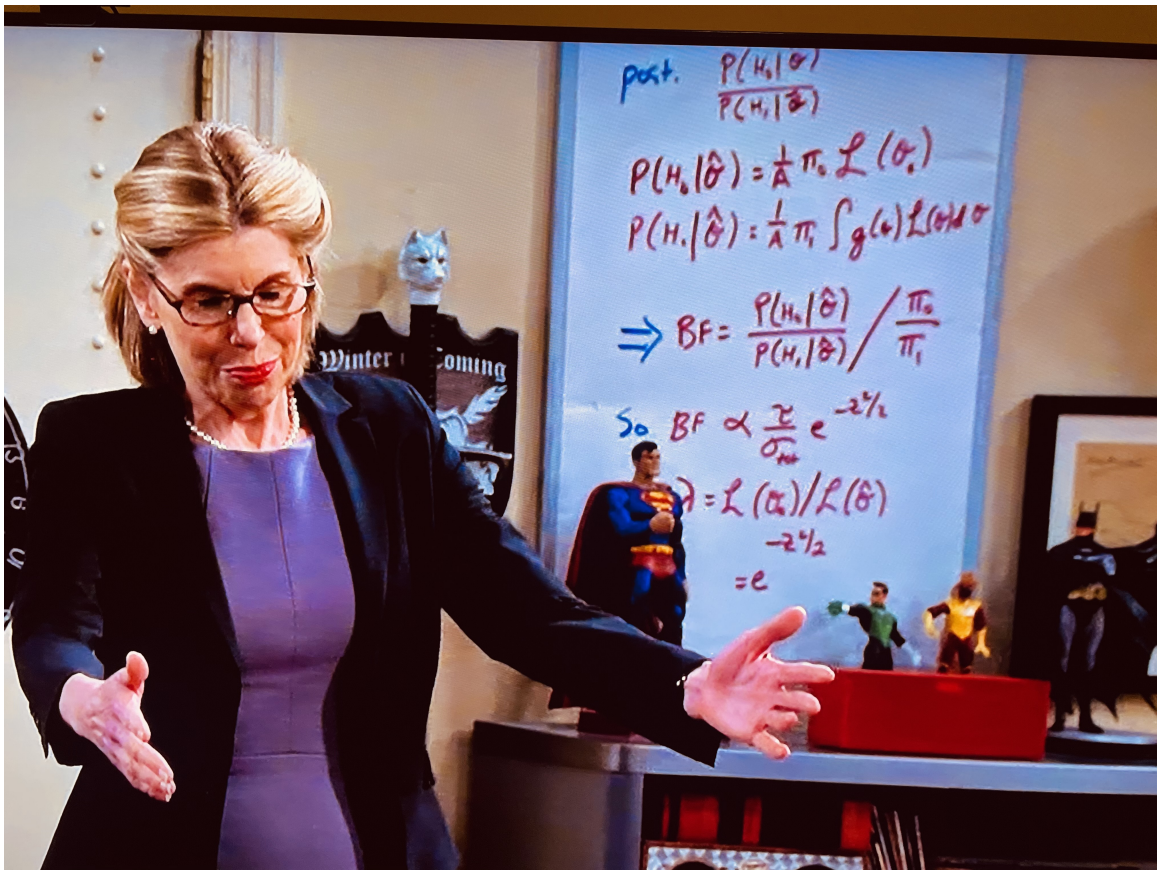
$$\frac{f(X|\theta_A)g(\theta_A)}{f(X|\theta_0)g(\theta_0)} \tag{9-21}$$

where the subscripts A refers to alternative hypothesis and 0 refers to the null hypothesis. This looks like a ratio of two Bayesian statements. For the special case where we assume that g(θ_A) = g(θ_0) the two "g" functions drop out of their framework and we have a ratio of likelihoods (known as a likelihood ratio).

As you can see, these three methods are not as different as some people claim they are, but there are huge differences in how everything is interpreted. Recall the convoluted language we had to use the first lecture in September when we reviewed confidence intervals (95% of such intervals will contain ...); the Bayesian interpretation is a more natural one and consistent with what we typically want to know from our data.

When we limit the Bayesians to a uniform prior, the three approaches lead to identical p-values and intervals in most ANOVA and regression problems. But for more complicated situations the three approaches may diverge. There are deep issues here in what the statistical test means and what information we as researchers can take from a statistical test. If this type of discussion about foundations interests you, then you might consider taking a course in theoretical statistics.

Bayesian ideas have even entered pop culture. I spotted the Bayesian definition of Bayes factor on the whiteboard in Sheldon and Leonard's apartment (the TV series Big Bang Theory).


post.
Winter
oming

Appendix 1: R Syntax

Various R tidbits.

Bayesian Logistic Regression

Here is the basic sketch of running logistic regression using the default settings such as the default prior, number of chains, and burnin. You can, of course, make relevant changes to those defaults.

```
library(brms)
out.glm <- brm(dv ~ me.vt + me.legit + int, data = data,
    family = "bernoulli")
```

```
summary(out.glm)
plot(out.glm)
```

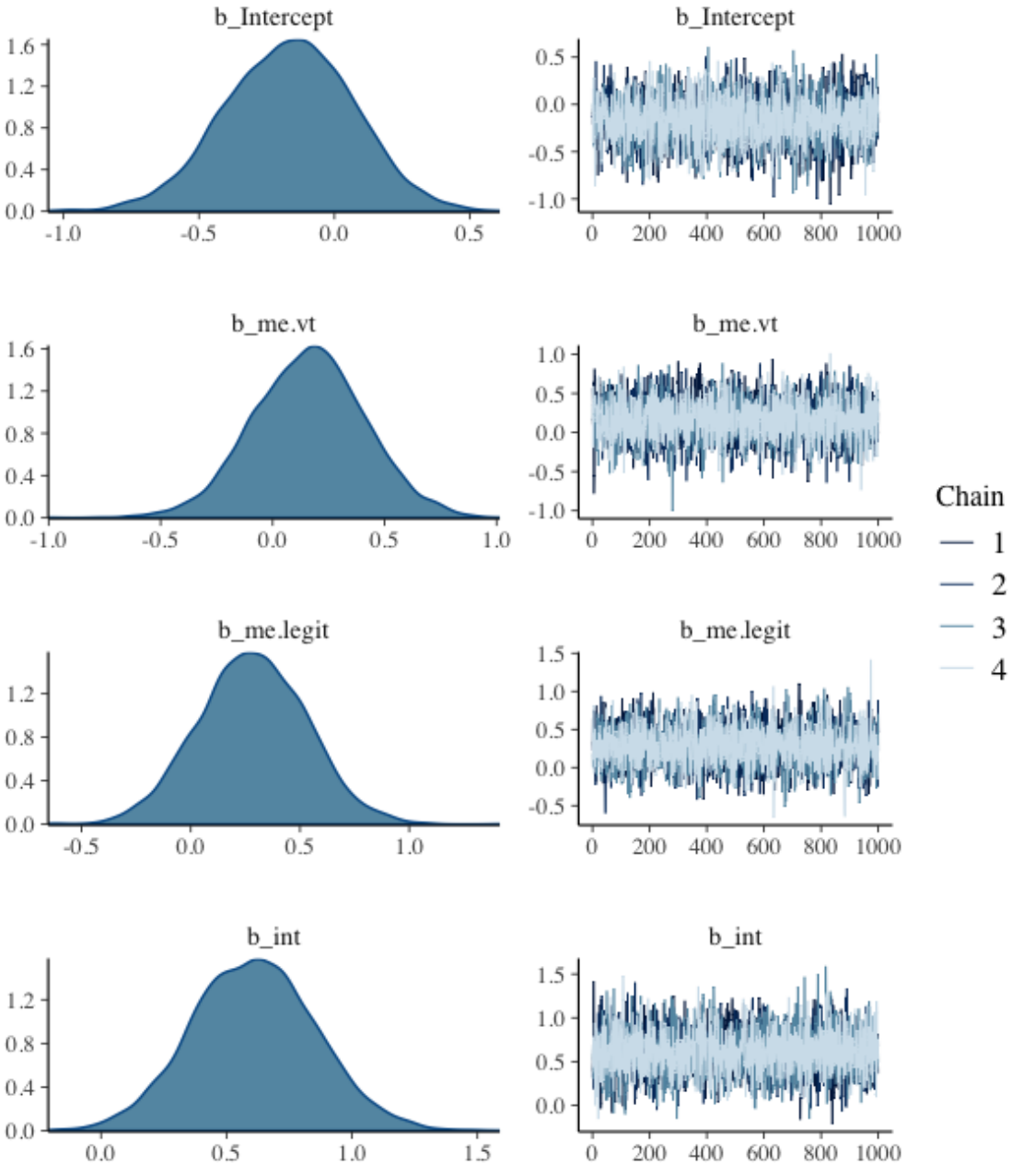
The snippets of the output include

```
 Family: bernoulli
  Links: mu = logit
Formula: dv ~ me.vt + me.legit + int
   Data: data (Number of observations: 80)
Samples: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
         total post-warmup samples = 4000

Population-Level Effects:
          Estimate Est.Error l-95% CI u-95% CI Eff.Sample Rhat
Intercept    -0.17      0.24    -0.63     0.29       3458 1.00
me.vt         0.16      0.25    -0.33     0.65       4228 1.00
me.legit      0.28      0.25    -0.20     0.77       4036 1.00
int           0.61      0.25     0.13     1.10       3823 1.00

Samples were drawn using sampling(NUTS). For each parameter, Eff.Sample
is a crude measure of effective sample size, and Rhat is the potential
scale reduction factor on split chains (at convergence, Rhat = 1).
```

But make sure you understand the role of priors in Bayesian statistics because the choice of prior affects the estimate, especially in small samples. The default prior used in common logistic regression analyses (and implemented in brm) invokes a noninformative prior over the logit scale rather than a nonuniform prior over the proportion scale (or noninformative prior on the arcsin scale as would be the case with the arcsin transformation).

Here is a simple example showing that brm estimates a term that is relatively close to the standard estimate of the proportion. Let's take a simple case of 10 coin tosses with 6 heads and 4 tails.

Figure 9-9: Density plots for each of the three contrasts.



```
data <- data.frame(choice = c(1, 1, 1, 1, 1, 1, 0,
    0, 0, 0))
# maximum likelihood estimate as expected .6
mean(data$choice)
# bayes estimate with noninformative prior on
# logit scale is close
out.ch <- brm(choice ~ 1, data = data, family = "bernoulli")
summary(out.ch)
plogis(fixef(out.ch)[1])
```

The snippets of the output include

```
 Family: bernoulli
  Links: mu = logit
Formula: choice ~ 1
   Data: data (Number of observations: 10)
Samples: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
         total post-warmup samples = 4000

Population-Level Effects:
          Estimate Est.Error l-95% CI u-95% CI Eff.Sample Rhat
Intercept     0.45      0.67    -0.84     1.86       1483 1.00

### need to convert logit scale back to proportion to compare to maximum
### likelihood estimate of .6 (slightly off due to simulation)
> plogis(fixef(out.ch)[1])
[1] 0.6098792
```

Propensity Scores

There are several packages in R that handle propensity scores directly. I suggest looking at MatchIt, Matching, optmatch, twang and PSAgraphics. These packages provide additional features in the analyses of propensity scores than the two-step procedure I introduced in these lecture notes.

Time Series

There are dozens of packages for dealing with time series data and relevant issues. There is a "task view" page at `http://cran.r-project.org/web/views/TimeSeries.html` that lists relevant R packages and provides short descriptions.

Special attention needs to be given to how you specific time factors and if you are importing data from peripherals that time stamp data. Examples include data from sensors such as eyetrackers or

survey software that provide time stamped data of when each participant entered the survey, how long they took on each question and when they exited.

Generalized Additive Models

The gam and mgcv packages in R are excellent. Other packages extend the features of these packages, for example, gamm4 adds random effects (aka mixed models; recall LN5 where we covered random effects in the context of ANOVA) to generalized additive models.