

Power & multiple regression

Lada Adamic
~~November 30, 2006~~
SI 544

1 Power

1.1 t-tests

Remember that power gives the probability that we will reject the null hypothesis H_0 , given that an alternative hypothesis, H_1 is true. For example, in class we considered H_0 : the mean height of midwestern men is equal to the mean height on men in the entire US, 68 inches. Suppose H_1 is true. H_1 says that the mean height is actually 69 inches.

We are given the fact that the standard deviation in the height of a US male is 3.1 inches. If we draw a sample of 100 men, we would expect the standard error of the sample mean to be approximately

$$\frac{\sigma}{\sqrt{N}} = \frac{3.1}{\sqrt{100}} = 0.31 \quad (1)$$

We can draw two normal distributions centered around 68 and 69, corresponding to the distribution of sample means of H_0 and H_1 respectively. Where the vertical line is drawn represents the boundary of the two-sided 95% confidence interval for H_0 . The probability that a sample mean exceeds 68.6 given H_1 is 0.89, which is just the power of rejecting the null hypothesis given that H_1 is true.

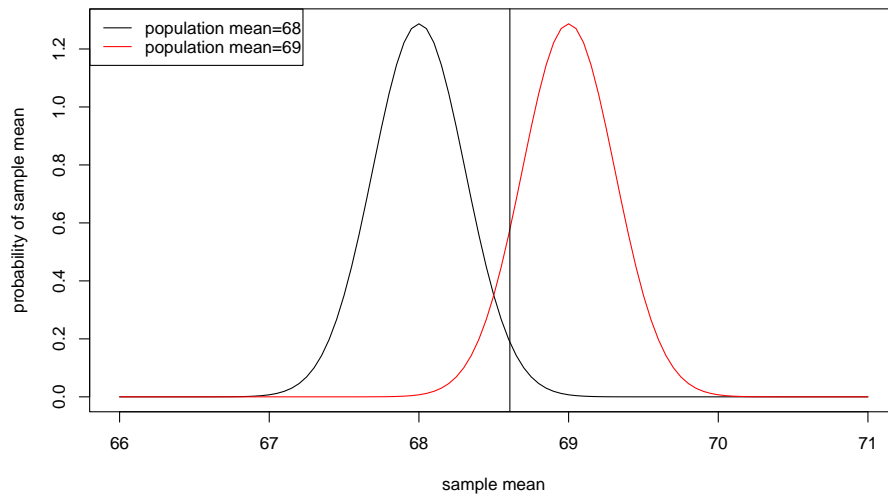
```
x = seq(66,70,by=0.05)
plot(x,dnorm(x,mean=68,sd=3.1/sqrt(100)),type='l',
     ylab="probability of sample mean",xlab="sample mean")
lines(x,dnorm(x,mean=69,sd=3.1/sqrt(100)),col="red")
legend(locator(1),legend=c("population mean=68","population mean=69"),
      lty=c(1,1),col=c("black","red"))
z196=68+1.96*3.1/sqrt(100)
abline(v=z196)
```

Visuals are well and good, but most of the time we just want the numbers. So let's try it out. We have a sample of 100, H_1 is that the underlying midwestern men population has $\mu_1 = 69$, so what is our power? For this we use R's `power.t.test()` function:

```
> power.t.test(delta=1,sd=3.1,n=100,sig.level=0.05,type="one.sample")
```

```
One-sample t test power calculation
```

```
      n = 100
  delta = 1
     sd = 3.1
sig.level = 0.05
  power = 0.8914722
alternative = two.sided
```



Right, just what we had surmised from the plot. On the other hand, we may yet have to decide how large of a sample we need to gather, but we know that we would like an 80% chance of rejecting the null if it is actually false:

```
> power.t.test(delta=1,sd=3.1,power=0.8,sig.level=0.05,type="one.sample")
```

One-sample t test power calculation

```
      n = 77.37044
  delta = 1
    sd = 3.1
sig.level = 0.05
  power = 0.8
alternative = two.sided
```

So we'll need to measure the heights of about 77-78 men before we reach 80% chance of rejecting the null (assuming of course that H_1 is true). OK, finally, assuming that we have a sample of 100 men and we want a power of .80 to detect a difference between the midwest and the rest of the US, how different do the mean heights in the US and midwestern population actually have to be?

```
> power.t.test(power=0.8,sd=3.1,n=100,sig.level=0.05,type="one.sample")
```

One-sample t test power calculation

```
      n = 100
  delta = 0.8770312
    sd = 3.1
sig.level = 0.05
  power = 0.8
alternative = two.sided
```

So the midwestern men would need to be at least 0.877" taller or shorter on average in order for our t-test to be able to pick out that there's a difference with 0.8 probability for a sample of 100.

1.2 proportions

It may seem just like last week... oh, wait, it was last week, that we were doing tests of proportion. R's `power.prop.test()` will tell us how many trials we need to do in each sample in order to tell if the proportion of successes is different between the samples. Suppose we are testing two different UI's on our website. For each UI we check whether the user returns for another visit.

If 100 users are exposed to each UI, and the proportion of returning users for UI is 0.20 for the first, and 0.22 for the other, what is the power?

```
> power.prop.test(n=100,p1=0.2,p2=0.22,sig.level=0.05)
```

```
Two-sample comparison of proportions power calculation
```

```
      n = 100
     p1 = 0.2
     p2 = 0.22
sig.level = 0.05
  power = 0.05334612
alternative = two.sided
```

NOTE: n is number in *each* group

Gosh, 0.05 is not that much power. In order to have an 80% chance of rejecting the null, how many users would need to see each UI?

```
> power.prop.test(p1=0.2,p2=0.22,sig.level=0.05,power=0.8)
```

```
Two-sample comparison of proportions power calculation
```

```
      n = 6509.467
     p1 = 0.2
     p2 = 0.22
sig.level = 0.05
  power = 0.8
alternative = two.sided
```

NOTE: n is number in *each* group

Wow, 6,500 users. I guess this is something you could do if you are a really large retailer and getting 10% more repeat visitors was a big deal. So how big of an actual difference in proportion would we have needed in order to be able to already detect a significant improvement after trying things out on two sets of 100 users?

```
> power.prop.test(p1=0.2,n=100,sig.level=0.05,power=0.8)
```

```
Two-sample comparison of proportions power calculation
```

```
      n = 100
     p1 = 0.2
     p2 = 0.3785940
sig.level = 0.05
  power = 0.8
alternative = two.sided
```

NOTE: n is number in *each* group

So the return rate would have to be nearly twice as high, in order for us to be able to reject the null at the 0.05 significance level with a power of 0.8 with only 100 users observed with the new UI.

1.3 power and anova

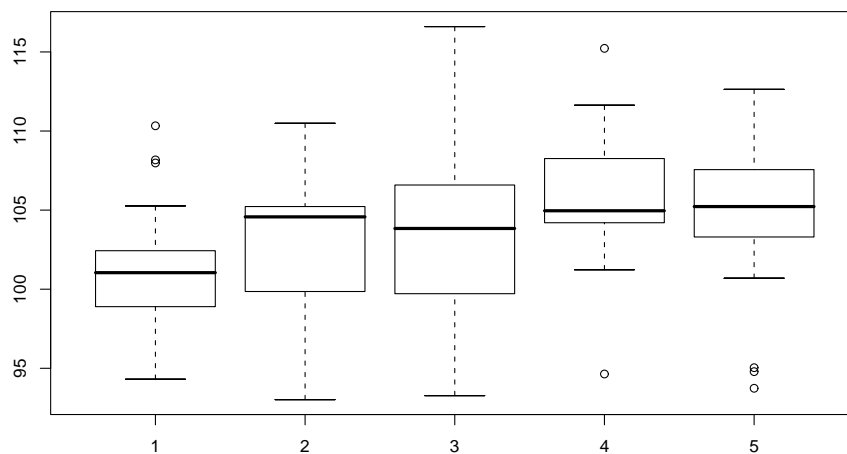
For the anova, we can calculate the power given the within group variance, the between group variance, the number of groups, the number of observations per group and the desired confidence level. That is a lot to know and to put in. But say you run your anova, it is going to give you your within and between group variance, and so you can run the power test to see how much power you had in rejecting the null hypothesis in the first place.

Let's start with a fake data set. We'll have the independent variable be the number of years with the firm, ranging from 1..5. Well draw 100 samples with replacement from this vector, getting approximately but not exactly 2 in each group:

```
> x = sample(seq(1,5),100,replace=T)
> summary(factor(x))
 1  2  3  4  5
21 18 17 17 27
```

We then set the salary to be $100+x+(\text{normally distributed noise with standard deviation of } 5)$:

```
> y = x+rnorm(100,100,5);
> boxplot(y~x)
```



From the boxplots we can see that we've successfully added quite a bit of noise. Can an ANOVA still tell the difference? Let's see:

```
> anova(lm(y~factor(x)))
Analysis of Variance Table
```

```
Response: y
      Df Sum Sq Mean Sq F value Pr(>F)
factor(x) 4  227.77   56.94  2.4424 0.05193 .
Residuals 95 2214.82   23.31
```

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Shoot, we just missed it. What would our power be?

```
> power.anova.test(groups=5,n=20,between.var=227.77,within.var=2214.82,sig.level=0.05)
```

Balanced one-way analysis of variance power calculation

```
groups = 5
n = 20
between.var = 227.77
within.var = 2214.82
sig.level = 0.05
power = 0.5940449
```

NOTE: n is number in each group

So our power is only about 60%. Could we have derived the between and within variance just from knowing our fake data setup? The between variance should be something like $(2^2 + 1^2 + 0 + 1^2 + 2^2)/5 * 100 = 200$. The within group variance should be $5^2 * 100 = 2500$. This is pretty close to what we actually observed. So how large of a sample would we need in order to have 80% power?

```
> power.anova.test(groups=5,n=40,between.var=betvar,within.var=withinvar,sig.level=0.05)
```

Balanced one-way analysis of variance power calculation

```
groups = 5
n = 40
between.var = 400
within.var = 5000
sig.level = 0.05
power = 0.8201283
```

NOTE: n is number in each group

We need about 40 observations in each group. Let's go back and do that:

```
> # REDO with larger sample
> x = sample(seq(1,5),200,replace=T)
> summary(factor(x))
 1  2  3  4  5
44 31 43 39 43
>
> #create a noisy salary variable that depends on x
> y = x+rnorm(200,100,5);
> boxplot(y~x)
>
> #an anova, remembering to treat x as a factor
> anova(lm(y~factor(x)))
Analysis of Variance Table

Response: y
      Df Sum Sq Mean Sq F value    Pr(>F)
factor(x)  4  562.8   140.7   5.6425 0.0002561 ***
Residuals 195 4862.1    24.9
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

So it worked out with the larger samples. We can tell that different years correspond to different (fake) salaries. But 20% of the time we still would not have been able to reject the null with this sample size. Enough of power, let's move on to multiple regression.

2 Multiple regression

This is just a repeat of the powerpoint slides. First let's set up and get our data and models in place. We are loading the libraries data, but keeping just Maine in order to not have too much data that would give us significant results even for very small effects. We will also normalize all the variables (circulation, attendance) by the population served by the library. We are also log-transforming the variables because of the large variance in library size.

```
> maine = libraries[libraries$STABR=="ME",]
>
> attach(maine)
> kiddata = data.frame(log((KIDATTEND+0.001)/POPU), log((TOTCIR-KIDCIRCL+0.001)/POPU),
+ log((KIDCIRCL+0.001)/POPU), log((BKVOL+0.001)/POPU))
> detach(maine)
> colnames(kiddata)=c("kid_attendance", "other_circulation", "kid_circulation", "bookvolume")
>
> pairs(kiddata)
>
> m1 = lm(kid_circulation ~ kid_attendance+other_circulation+bookvolume, data=kiddata)
> m1b = lm(kid_circulation ~ other_circulation+kid_attendance+bookvolume, data=kiddata)
>
> m2 = lm(kid_circulation ~ other_circulation, data=kiddata)
>
> m3 = lm(kid_circulation ~ other_circulation+kid_attendance, data=kiddata)
```

The first thing we can do is to visualize: Already we see that other circulation (total - kid) is quite highly correlated with kid circulation. The other variables are visibly less correlated. But until we do our anovas and multiple regressions, we won't know if they still might come in handy when modeling the amount of circulation of kids materials.

```
> summary(m1)
```

Call:

```
lm(formula = kid_circulation ~ kid_attendance + other_circulation +
    bookvolume, data = kiddata)
```

Residuals:

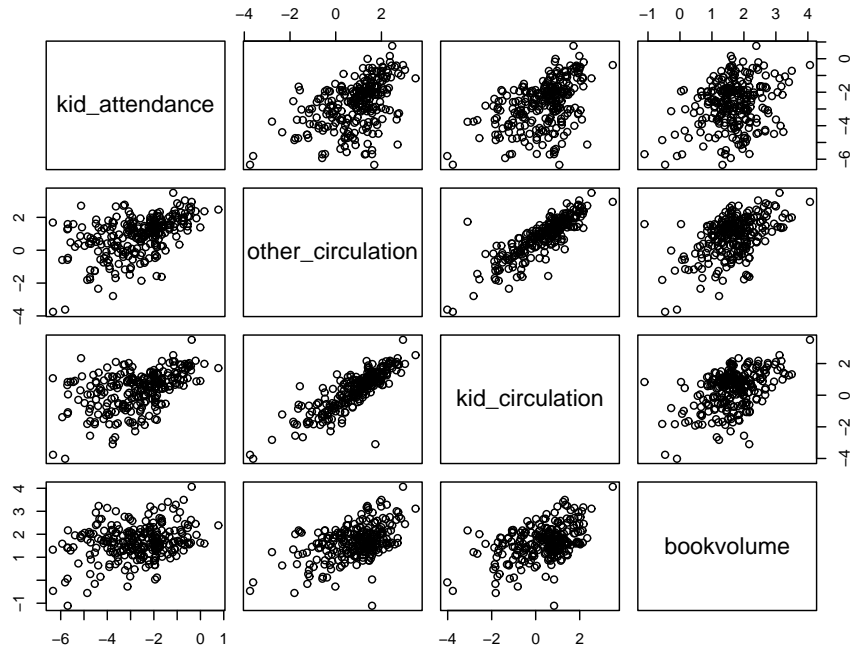
Min	1Q	Median	3Q	Max
-4.13904	-0.30004	-0.01340	0.33643	2.06613

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-0.33925	0.13697	-2.477	0.0139 *
kid_attendance	0.06019	0.02964	2.031	0.0433 *
other_circulation	0.79608	0.04069	19.566	<2e-16 ***
bookvolume	0.09728	0.05911	1.646	0.1010

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6031 on 261 degrees of freedom



Multiple R-Squared: 0.7386, Adjusted R-squared: 0.7356
 F-statistic: 245.9 on 3 and 261 DF, p-value: < 2.2e-16

Note that kid_attendance and other_circulation are both significant (as shown by the t-statistics), as well as the entire model (according to the F statistic).

The order matters for the anova:

```
> anova(m1)
Analysis of Variance Table

Response: kid_circulation
      Df Sum Sq Mean Sq  F value Pr(>F)
kid_attendance    1  82.273   82.273  226.1626 <2e-16 ***
other_circulation  1 185.049  185.049  508.6864 <2e-16 ***
bookvolume        1   0.985    0.985   2.7085 0.1010
Residuals        261  94.946    0.364
```

```
> anova(m1b)
Analysis of Variance Table

Response: kid_circulation
      Df Sum Sq Mean Sq  F value  Pr(>F)
other_circulation  1 265.740  265.740  730.5004 < 2e-16 ***
kid_attendance    1   1.582    1.582   4.3485 0.03801 *
bookvolume        1   0.985    0.985   2.7085 0.10102
Residuals        261  94.946    0.364
```

Note that the only difference between models m1 and m1b is the order of the variables. It does not make a difference to the multiple regression, since it considers the contribution of each variable keeping the others fixed. But for the anova, it only considers the amount of variance explained by each variable that was

not already explained by the variables above it. So we can see that although *kid_attendance* at first seems highly significant ($F=226$ and $p < 10^{-16}$), once *other_circulation* is taken into account, it is only marginally significant ($F=4.34$, $p = 0.038$).

This brings us to the idea of trying to drop some variables. We can do this manually using the `anova` function to compare a submodel with the full model (it will give us the difference in the sum of squared errors and the F statistic). Another option is to use the function `step()`.

```
> anova(m1,m2)
Analysis of Variance Table

Model 1: kid_circulation ~ kid_attendance + other_circulation + bookvolume
Model 2: kid_circulation ~ other_circulation
  Res.Df    RSS Df Sum of Sq    F Pr(>F)
1     261 94.946
2     263 97.513  -2    -2.567 3.5285 0.03076 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
> anova(m1,m3)
Analysis of Variance Table

Model 1: kid_circulation ~ kid_attendance + other_circulation + bookvolume
Model 2: kid_circulation ~ other_circulation + kid_attendance
  Res.Df    RSS Df Sum of Sq    F Pr(>F)
1     261 94.946
2     262 95.931  -1    -0.985 2.7085 0.1010
```

What we can see from the above is that we shouldn't reduce our model to just *other_circulation* (there is a significant benefit to keeping more variables than that). The second anova tells us that it is sufficient to keep just *kid_attendance*, but having the *book_volume* as well does not explain significantly more of the variation.

We can also automate this process with the `step()` function, which will order the variables by their ability to explain the variance in the dependent variable:

```
> step(m1)
Start:  AIC= -264
kid_circulation ~ kid_attendance + other_circulation + bookvolume

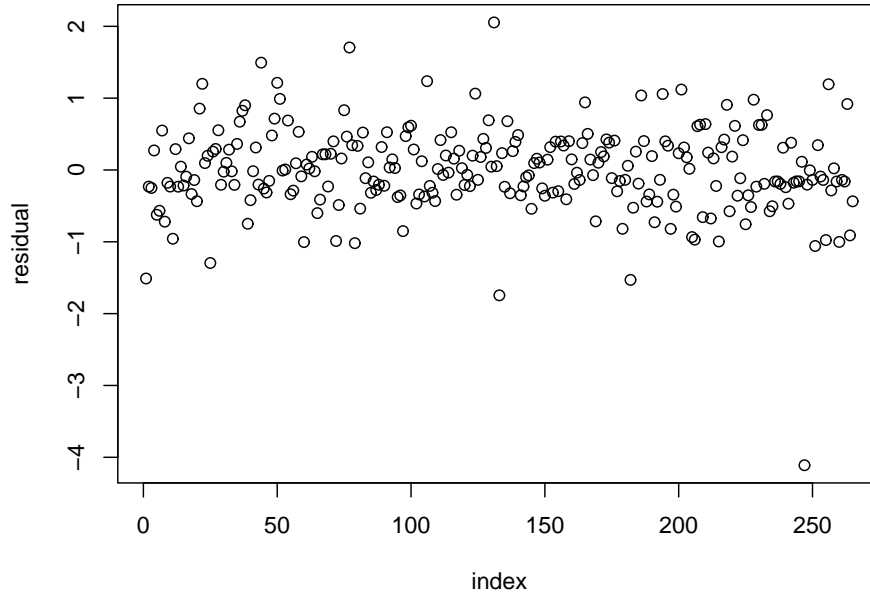
              Df Sum of Sq      RSS      AIC
<none>                94.946 -264.002
- bookvolume           1    0.985  95.931 -263.266
- kid_attendance       1    1.500  96.446 -261.847
- other_circulation    1  139.265 234.211 -26.729

Call:
lm(formula = kid_circulation ~ kid_attendance + other_circulation +
    bookvolume, data = kiddata)

Coefficients:
(Intercept)    kid_attendance  other_circulation    bookvolume
   -0.33925         0.06019         0.79608         0.09728
```

`step()` confirms the same order of importance for the variables.

Finally, let's check that the residuals are normally distributed so that we know that we are justified in doing a linear regression in the first place:



We're fine. The residuals look approximately normally distributed and there is no visible trend.