Regression

• Trying to predict a response variable (y) using a predictor/explanatory variable (x)

• Can use a scatterplot to examine this relationship before using regression to obtain a mathematical relationship – a least squares regression line (when you see a linear relationship)

Variable Types

• Response Variable – Y – also called the dependent variable

• Explanatory Variable – X – also called the independent or predictor variable

• In regression, we assume that X can convey some information to us about Y through a linear relationship

Formula for Regression Line

• Familiar form: $y = mx+b = b + mx$

• Our form: $E(y) = \beta_0 + \beta_1x$
  where $E(y)$ is the mean response based on the single predictor $x$

• We need to find estimates of $\beta_0$ and $\beta_1$
  (these are parameters, and we need parameter estimates)

Population Parameters

• $\beta_0$ is the population y-intercept; It’s where the true regression line crosses the y-axis (i.e. when $x = 0$)

• $\beta_1$ is the population slope; It’s the change in the mean response $y$ for a 1-unit increase in $x$
Estimating the Parameters

- $\beta_0$ and $\beta_1$ are estimated with $b_0$ and $b_1$ respectively.
- The resulting line is $\hat{y} = b_0 + b_1 x$.
- The estimates ($b_0$ and $b_1$) are called least squares estimates.

Least Squares Line

- The least squares regression line is the line that minimizes the sum of the squared vertical distances from the data points to the line.
- Fitting the least squares line will give you the estimates $b_0$ and $b_1$ of $\beta_0$ and $\beta_1$.

Regression Terms: $r$

- $r$ is the correlation (or coefficient of correlation) between $x$ and $y$. It measures the strength and direction of the linear relationship between $x$ and $y$.
- $r$ must be between $-1$ and $1$.
- A positive correlation means that as the value of the predictor increases, so does the value of the response.
- A negative correlation means that as the value of the predictor increases, the value of the response decreases.

Terms (continued): $r^2$

- AKA $r$-squared or coefficient of determination.
- Tells you how much (the proportion) of the observed variation in the response can be explained through the regression relationship (i.e. by the predictor).
  - Based on a linear relationship between $x$ and $y$.
- $r$-squared must be between 0 and 1.

Use $r$ and $r^2$?

- Should you use $r$ and $r^2$ here to describe the relationship between the $x$ and $y$ variables?

Guess the Correlation
Module 10: Linear Regression

Variables for Poverty Data Set
- Location
- Poverty rate (PovPct): % of the population living in households with income below poverty level for each state and the DC
- Teen birth rate (TeenBrth): Birth rate for females 15 to 19 years old, recorded as number of births per 1,000 females in that age group
- Violent crime rate (ViolCrime)
- Birth rate for females 15 to 17 years old (Brth15to17)
- Birth rate for females 18 to 19 years old (Brth18to19)

Activity 1: Page 66
- Problem: Is there a relationship between the poverty rate (PovPct) and the teen birth rate (TeenBrth)?
- Task: Produce a scatterplot to help examine the plausibility of a linear relationship between the poverty rate and the teen birth rate
  Data: poverty.sav

Module 10: Activity 1
- Which variable should be the response and which should be the explanatory?

Yes or No
- Based on the scatterplot, does there appear to be a linear relationship between teen birth rate and poverty rate?

Module 10: Activity 1
- Are there any unusual observations or outliers present?
Module 10: Activity 1

- Which states have the highest poverty rates? Highest teen birth rates? Do states with low poverty rates tend to have low teen birth rates also? What does this tell you about the direction of the association between these two variables?

Module 10: Activity 1

- Does there appear to be a strong relationship between poverty rates and teen birth rates?

Module 10: Linear Regression

Activity 2: Page 68

Task:
- Fit a linear model to the poverty data.
  
  Data: poverty.sav

Module 10: Linear Regression

To perform a Linear Regression:

Analyze> Regression> Linear
– Make sure you select “save” and select “Unstandardized” under “Residuals”, this automatically creates a new column of residuals in your data set. You will need them for checking assumptions!

Module 10: Activity 2

- Regression Line and Slope interpretation:

  The estimated regression line is:
  Interpret the estimated slope ($b_1$):

  Note: It is YOUR job to determine the sign of r!!
Module 10: Activity 2

• Report $r^2$ and interpret it:

<table>
<thead>
<tr>
<th>Model</th>
<th>R</th>
<th>R Square</th>
<th>Adjusted R Square</th>
<th>Std. Error of the Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.703</td>
<td>.495</td>
<td>.484</td>
<td>.846</td>
</tr>
</tbody>
</table>

- Predictors: (Constant), PovPct.
- Dependent Variable: TeenBrth.

• Use the predicted line to predict the number of births per 1,000 females ages 15-19 for New Mexico (with a poverty rate of 25.3%) and for Michigan (with a poverty rate of 12.2%). How do they compare to the observed TeenBrth values for New Mexico and Michigan?

Aside - Think about it

• Would you use this model to predict the teen birth rate for a state that has a poverty rate of 35%? How about 2%? Why or why not?

Module 10: Linear Regression

Activity 3: Page 69

• Is there a Significant (non-zero) Linear Relationship between Teen Birth Rate and Poverty Rate?
  • M1: Using regression output
    – Look at p-value associated with t-statistic
    – $H_0$: $\beta_1 = 0$
  • M2: Using ANOVA output
    – Look at p-value associated with F-statistic
    – $H_0$: Slope term is not necessary

Module 10: Linear Regression

Solutions to activities 3 and 4 will be posted on C-tools

Make sure to look at summaries on pp. 73-75. They have useful info; not all has been covered in lab

Create regression output for HW11 in lab

Was there a difference between your first and second exam scores on average?

Comparison of exam scores for Section 19

Comparison of exam scores for Section 49