Announcements

1. Lab #2 Intro to data due tonight at 11:55pm
2. Start HW#3–we will go over most answers together on the 16th

Agenda

1. Guess the correlation
2. Simple Linear Regression
3. Residuals
4. Strength of Fit

Simple linear regression Linear regression can help us understand changes in a numerical response variable in terms of a numerical explanatory variable.

A simple linear regression model for y in terms of x takes the form

$$y_i = \beta_0 + \beta_1 \cdot x_i + \epsilon_i\,, \text{ for } i = 1, \ldots, n$$

- β_0 is the *intercept* and β_1 is the *slope* coefficient. The ϵ_i's are the *errors*, or *noise*.
- There is only one regression line that fits the data best using a least squares criteria. That is, the *ordinary least squares* regression line is unique.
- The true values of the unknown parameters β_0 and β_1 are estimated by b_0 and b_1 (or if you prefer, $\hat{\beta}_0$ and $\hat{\beta}_1$)
- The *fitted values*, or *predicted values* are given by

$$\hat{y}_i = b_0 + b_1 \cdot x_i$$

- The model almost never fits perfectly, but what is left over is captured by the *residuals* ($e_i = y_i - \hat{y}_i$)

Example: RailTrail Data The Pioneer Valley Planning Commission (PVPC) set up a laser sensor, with breaks in the laser beam recording when a rail-trail user passed the data collection station. The data is captured in the `RailTrail` data set.

```
library(mosaic)
data(RailTrail)
```

1. In R, create a scatterplot [`qplot()`] for the response: *volume* in terms of explanatory: *avgtemp*.

2. Describe the shape, direction, and strength of the relationship. Guess the correlation, r.

3. Compute the correlation coefficient [`cor()`]. Hint: use the same syntax that favstats() uses, that is, `cor(`$y \sim x$`, data = TheData)`. Is r what you expected?

4. Fit the linear regression model using `lm()`. Again, use the `favstats()` syntax.

5. Interpret the coefficients for the `Intercept` (b_0) and `avgtemp` (b_1) terms.

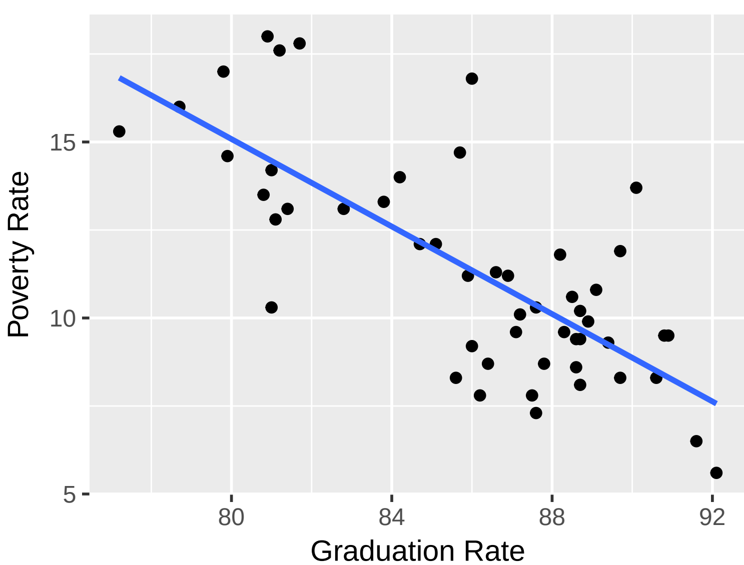
6. Using your R output, write out the fitted regression equation. Use the variable names instead of y and x.

7. Using the fitted regression equation, calculate the fitted value for the 15th case in the data set. Also calculate the residual for this case. How did the model do for this case?

8. Using the fitted regression equation, how many bikers would you predict on a day that is 72 degrees on average over the whole day?

Calculating the Regression Line by Hand Is there an association between poverty and education among states? The following plot illustrates the relationship between the *poverty rate* and the *high school graduation rate* among all 50 states and the District of Columbia.

```
library(mosaic)
poverty <- read.csv("http://math.smith.edu/~bbaumer/mth241/poverty.txt", sep = "\t")
qplot(data = poverty, x = Graduates, y = Poverty, xlab = "Graduation Rate", ylab = "Poverty Rate") +
  geom_smooth(method = "lm", se = FALSE)
```



Use the following summary statistics to calculate the least squares regression line.

```
favstats(~Poverty, data = poverty)

##  min   Q1 median   Q3 max     mean       sd  n missing
##  5.6 9.25   10.6 13.4  18 11.34902 3.099185 51       0

favstats(~Graduates, data = poverty)

##   min   Q1 median   Q3  max     mean       sd  n missing
##  77.2 83.3   86.9 88.7 92.1 86.01176 3.725998 51       0

cor(Poverty ~ Graduates, data = poverty)

## [1] -0.7468583
```

- Slope:

- Intercept:

- Interpretation:

One Categorical Explanatory Variable Suppose that instead of using temperature as our explanatory variable for ridership on the RailTrail, we just used whether it was a weekday or not. The variable *weekday* is *binary* in that it only takes on the values 0 and 1. [Such variables are also called *indicator* variables or *dummy* variables.] Such a model has the form:

$$\widehat{volume} = b_0 + b_1 \cdot weekday$$

```
## (Intercept)    weekday1
##   430.71429   -80.29493
```

1. How many riders does the model expect will visit the Rail Trail on a weekend? What about a weekday? What's the difference in predicted volume of riders bewteen weekdays and weekends? What if it's 80 degrees out, does this model tell us anything about that?

2. Draw a (tiny) scatterplot of the data and indicate this model graphically.

Measuring the Strength of Fit Just as we were able to quantify the strength of the linear relationship between two variables with the correlation coefficient, r, we can quantify the percentage of variation in the response variable (y) that is explained by the explanatory variables. This quantity is called the *coefficient of determination* and is denoted R^2.

- Like any percentage, R^2 is always between 0 and 1
- For simple linear regression (one explanatory variable), $R^2 = r^2$
- $R^2 = (s_y^2 - s_{RES}^2)/s_y^2 = 1 - s_{RES}^2/s_y^2$

```
mod <- lm(Poverty ~ Graduates, data = poverty)
n <- nrow(poverty)
varY <- var(~Poverty, data = poverty)
varE <- var(~residuals(mod), data = poverty)
1 - varE / varY

## [1] 0.5577973

rsquared(mod)

## [1] 0.5577973
```

RailTrail example Recall the RailTrail example from last time, in which we were trying to understand ridership (*volume*) in terms of temperature (*avgtemp*). We fit two models: 1) a linear regression model for *volume* as a function of *avgtemp* and 2) a linear regression model for *volume* as a function of *weekday*. The R^2 value for the second model was:

```
rsquared(lm(volume ~ weekday, data = RailTrail))

## [1] 0.08600583
```

1. What was the R^2 for the first model? Which one fit the data better?

2. Write a sentence interpretting the R^2 for the second model presented above.

3. Take a guess at the R^2 for the following model

```
lm(volume ~ avgtemp + weekday, data = RailTrail)
```