Agenda

1. Mathematics behind linear regression
2. Strength of Fit

Mathematics behind linear regression Looking at the relationship between the y_is, the $\hat{y}_i$s, and $\bar{y}$, we can see that

$$(y_i - \bar{y}) = (\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i)$$

We can use the relationship between those quantities to gain some intuition for this:

$$\begin{aligned} \sum_{i=1}^{n}(y_i - \bar{y})^2 &= \sum_{i=1}^{n}(\hat{y}_i - \bar{y}) + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \\ SST &= SSM + SSE \end{aligned}$$

r is the correlation between two variables

$$r = \frac{1}{1-n} \sum_{i=1}^{n} \frac{x_i - \bar{x}}{s_x} \frac{y_i - \bar{y}}{s_y}$$

and conveniently,

$$\hat{\beta}_1 = \frac{s_y}{s_x} \cdot r$$

Once you know $\hat{\beta}_1$, you can find the intercept ($\hat{\beta}_0$) by plugging in $(\bar{x}, \bar{y})$– a point that is always on the line.

$$y - \bar{y} = \hat{\beta}_1(x - \bar{x})$$

Example: Poverty and Education 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.

```
require(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 = 0)
```

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:

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), $R^2 = r^2$
- $R^2 = 1 - SSE/SST = SSM/SST$

```
qplot(data = poverty, x = Graduates, y = Poverty, xlab = "Graduation Rate", ylab = "Poverty Rate") +
  geom_smooth(method = "lm", se = 0, size = 3)
mod <- lm(Poverty ~ Graduates, data = poverty)
n <- nrow(poverty)
SST <- var(~Poverty, data = poverty) * (n - 1)
SSE <- var(residuals(mod)) * (n - 1)
1 - SSE / SST

## [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: a simple model in based strictly on the average volume, and a linear regression model for *volume* as a function of *avgtemp*. The R^2 value for the second model was:

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

## [1] 0.1822039

# rsquared(lm(volume ~ 1, data=RailTrail))
```

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.