Agenda

1. Simple Linear Regression
2. Residuals

Warmup: Correlation An article reported that there was a 0.42 correlation between alcohol consumption and income among adults with a four-year college degree. Is it reasonable to conclude that increasing one's alcohol consumption will increase one's income? Explain why or why not.

A college newspaper interviews a psychologist about student ratings of the teaching of faculty members. The psychologist says, "The evidence indicates that the correlation between the research productivity and teaching rating of faculty members is close to zero." The paper reports this as "Prof. McDaniel said that good researchers tend to be poor teachers, and vice versa." Explain why the paper's report is wrong. Write a statement in plain language (don't use the word *correlation*) to explain the psychologist's meaning.

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* 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 The Pioneer Valley Planning Commission (PVPC) collected data north of Chestnut Street in Florence, MA for ninety days from April 5, 2005 to November 15, 2005. Data collectors 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.

```
require(mosaic)
```

```
## Warning:  Installed Rcpp (0.12.12) different from Rcpp used to build dplyr (0.12.10).
## Please reinstall dplyr to avoid random crashes or undefined behavior.
```

```
data(RailTrail)
```

1. Write the code to create a scatterplot [`qplot()`] for the *volume* in terms of *avgtemp*

2. Describe the form, direction, and strength of the relationship

```
m1 <- lm(volume~avgtemp, data=RailTrail)
coef(m1)

## (Intercept)     avgtemp
##    99.60227     4.80205
```

1. Interpret the coefficients for the `Intercept` and `avgtemp` terms

Model visualization Compare the least squares regression line (right) with the average (left).