Agenda

1. Linear regression and alternative facts
2. Strength of Fit
3. Parallel Slopes Models

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:

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), $R^2 = r^2$
- $R^2 = (s_y^2 - s_{RES}^2)/s_y^2 = 1 - s_{RES}^2/s_y^2$

```
poverty <- read.csv("http://math.smith.edu/~bbaumer/mth241/poverty.txt", sep = "\t")
mod <- lm(Poverty ~ Graduates, data = 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)
```

Multiple Regression Multiple regression is a natural extension of simple linear regression.

- SLR: one response variable, one explanatory variable

$$Y = \beta_0 + \beta_1 \cdot X + \epsilon$$

- MLR: one response variable, *more than one* explanatory variable

$$Y = \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \cdots + \beta_p \cdot X_p + \epsilon$$

- Estimated coefficients (e.g. $\hat{\beta}_i$'s or b_i's) now are interpreted in relation to (or "conditional on") the other variables
- b_i reflects the *predicted* change in Y associated with a one unit increase in X_i, conditional upon the rest of the X_i's.
- R^2 has the same interpretation (proportion of variability explained by the model)

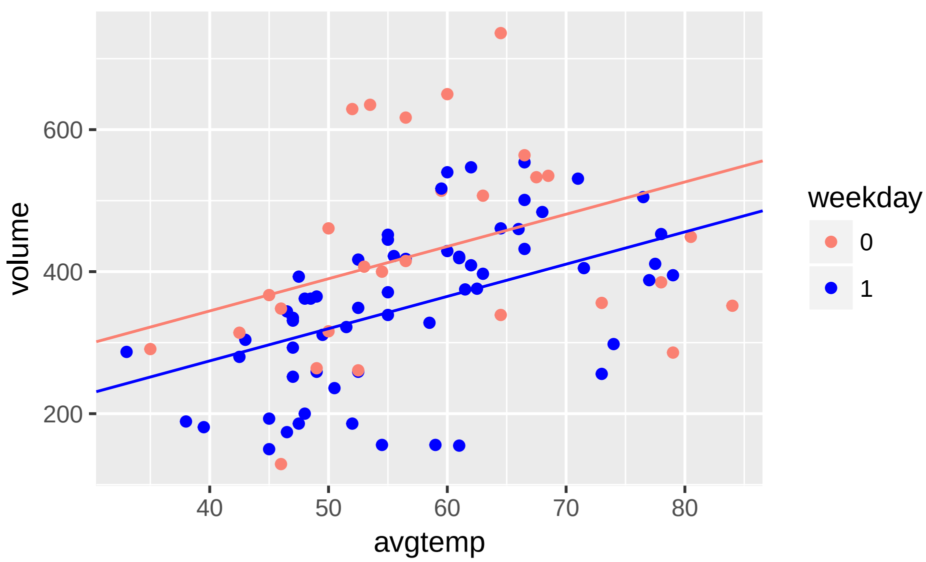
Multiple Regression with a Categorical Variable Consider the case where X_1 is quantitative, but X_2 is an *indicator* variable that can only be 0 or 1 (e.g. $isWeekday$). Then,

$$\hat{Y} = b_0 + b_1 \cdot X_1 + b_2 \cdot X_2$$

So then,

$$\begin{aligned} \text{For weekend,} \quad & \hat{Y}|_{X_1, X_2=0} = b_0 + b_1 \cdot X_1 \\ \text{For weekdays,} \quad & \hat{Y}|_{X_1, X_2=1} = b_0 + b_1 \cdot X_1 + b_2 \cdot 1 \\ & \qquad\qquad\quad = (b_0 + b_2) + b_1 \cdot X_1 \end{aligned}$$

This is called a *parallel slopes* model. [Why?]



Example: Italian Restaurants The Zagat guide contains restaurant ratings and reviews for many major world cities. We want to understand variation in the average *Price* of a dinner in Italian restaurants in New York City. Specifically, we want to know how customer ratings (measured on a scale of 0 to 30) of the *Food*, *Decor*, and *Service*, as well as whether the restaurant is located to the *East* or west of 5th Avenue, are associated with the average *Price* of a meal. The data contains ratings and prices for 168 Italian restaurants in 2001.

```
library(mosaic)
NYC <- read.csv("http://www.math.smith.edu/~bbaumer/mth241/nyc.csv")

qplot(data = NYC, x = Food, y = Price, geom = "jitter") +
  geom_smooth(method = "lm", se = 0)
lm(Price ~ Food, data = NYC)

##
## Call:
## lm(formula = Price ~ Food, data = NYC)
##
## Coefficients:
## (Intercept)         Food
##     -17.832        2.939
```

In-Class Activity

1. Use `qplot()` to examine the bivariate relationships between *Price*, *Food* and *Service*.

2. What do you observe? Describe the form, direction, and strength of the relationships.

3. Use `lm()` to build a SLR model for *Price* as a function of *Food*. (See code above). Interpret the coefficients of this model. How is the quality of the food at these restaurants associated with its price? Calculate the R^2 for this model and interpret it in a sentence.

4. Build a parallel slopes model by conditioning on the *East* variable. (Hint: formula = `Price Food + East`)

5. Interpret the coefficients of this model. What is the value of being on the East Side of Fifth Avenue? What is the R^2 for this model?

6. Calculate the expected *Price* of a restaurant in the East Village with a *Food* rating of 23.

7. Add `geom_abline()`'s to a `qplot()` visualize your model in the data space. How would you add color to the points to differentiate East from West?

```
qplot(data = NYC, x = Food, y = Price, geom = "jitter") +
geom_abline(intercept = -17.430, slope = 2.875) +
geom_abline(intercept = -17.430 + 1.459, slope = 2.875)
```