Agenda

1. Exam 1 due on Wed!
2. Class projects reminder
3. Multiple Regression
4. Inference through Randomization

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)

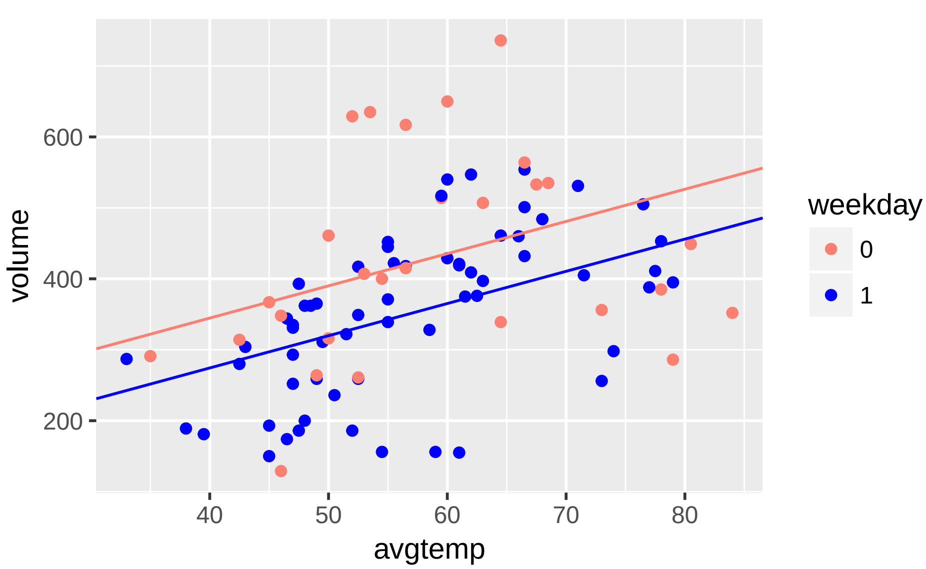
Multiple Regression with a Categorical Variable Consider the case where X_1 is quantitative, but X_2 is an *indicator* variable that can only be 0 or 1 (e.g. $isWeekday$). Then,

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

So then,

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

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)
```

Multiple Regression with a Second Quantitative Variable If X_2 is a quantitative variable, then we have

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

Example: SAT scores The `SAT` data was assembled for a statistics education journal article on the link between SAT scores and measures of educational expenditures.

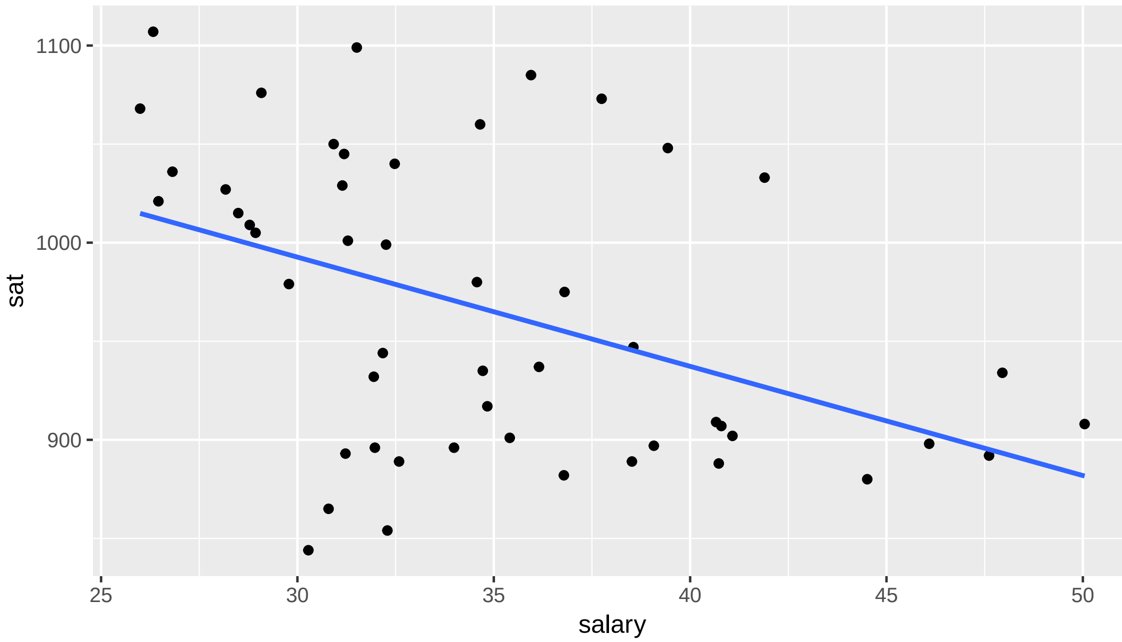
```
library(mosaic)
data(SAT)
cor(sat ~ salary, data = SAT)

## [1] -0.4398834

lm(sat ~ salary, data = SAT)

##
## Call:
## lm(formula = sat ~ salary, data = SAT)
##
## Coefficients:
## (Intercept)       salary
##     1158.86        -5.54

ggplot(data = SAT, aes(salary, sat)) +
  geom_point() +
  geom_smooth(method = "lm", se = 0)
```



Now suppose that we want to improve our model by considering not only the teachers' average `salary` in thousands, but also the precentage of students taking the exam, `frac`. We can do this in `R` by simply adding another variable to our regression model.

```
cor(sat ~ frac, data = SAT)

## [1] -0.8871187

cor(frac ~ salary, data = SAT)

## [1] 0.6167799

lm(sat ~ salary + frac, data = SAT)

##
## Call:
## lm(formula = sat ~ salary + frac, data = SAT)
##
## Coefficients:
## (Intercept)       salary         frac
##     987.900        2.180       -2.779
```

Our model is no longer a line, rather it is a *plane* that lives in three dimensions!

1. Interpret the coefficients of this model. What does the coefficient of *salary* mean in the real-world context of the problem? *frac*?

2. How important is *salary* relative to $frac$? Is it fair to compare the two coefficients?

3. Find the expected *sat* score of a state with an average teacher *salary* of $43,000, and 82% of eligible students taking the exam.

4. Find the percent of variation in states' *sat* scores explained by this model.

5. How would you add the state's expenditure per pupil, *expend*, to the model? Give it a try! Does the R-squared go up when you add *expend*?

6. Do you think the coefficent for *salary* in this model is meaningful or could it be due to chance?