Agenda

1. Leverage, influence, and outliers
2. Parallel Slopes Models

Warmup: Regression

1. In 1966 Cyril Burt published a paper called "The genetic determination of differences in intelligence: A study of monozygotic twins reared apart." The data consist of IQ scores for [an assumed random sample of] 27 identical twins, one raised by foster parents, the other by the biological parents.

   Here is the regression output for using *Biological* IQ to predict *Foster* IQ:

```
require(mosaic)
require(faraway)
mod <- lm(Foster ~ Biological, data = twins)
coef(mod)

## (Intercept)  Biological
##    9.207599    0.901436

rsquared(mod)

## [1] 0.7779022
```

   Which of the following is **FALSE**? Justify your answers.

   (a) Alice and Beth were raised by their biological parents. If Beth's IQ is 10 points higher than Alice's, then we would expect that her foster twin Bernice's IQ is 9 points higher than the IQ of Alice's foster twin Ashley.

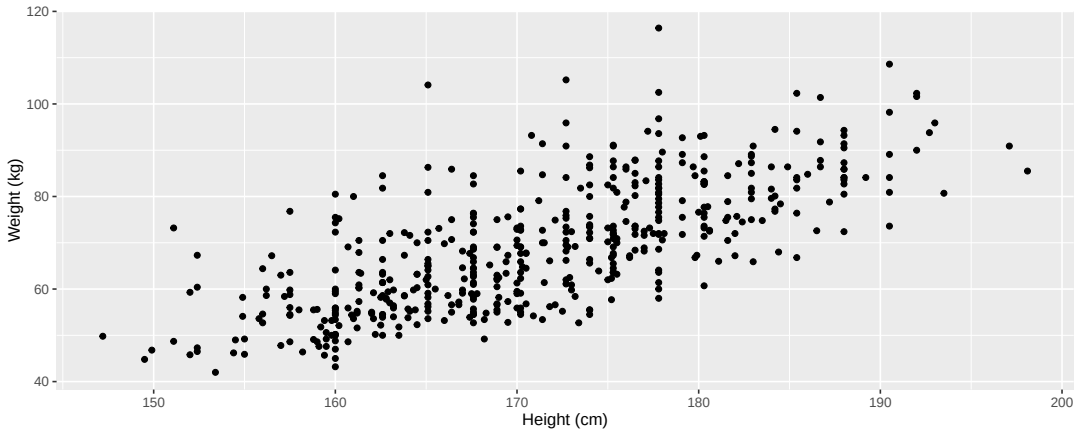
   (b) Roughly 78% of the foster twins' IQs can be accurately predicted by the model.

   (c) The linear model is $\widehat{Foster} = 9.2 + 0.9 \times Biological$.

   (d) Foster twins with IQs higher than average are expected to have biological twins with higher than average IQs as well.

2. The scatterplot and least squares summary below show the relationship between weight measured in kilograms and height measured in centimeters of 507 physically active individuals.

```
require(openintro)
qplot(data = bdims, x = hgt, y = wgt, xlab = "Height (cm)", ylab = "Weight (kg)")
```




```
coef(lm(wgt ~ hgt, data = bdims))

## (Intercept)         hgt
## -105.011254    1.017617
```

(a) Describe the relationship between height and weight.

(b) Write the equation of the regression line. Interpret the slope and intercept in context.

(c) The correlation coefficient for height and weight is 0.72. Calculate R^2 and interpret it in context.

Outliers, Leverage, and Influence It is important to identify the outliers and understand their role in determing the regression line.

- An *outlier* is an observation that doesn't seem to fit the general pattern of the data
- An observation with an extreme value of the explanatory variable is a point of high *leverage*
- A high leverage point that exerts disproportionate influence on the slope of the regression line is an *influential point*

Quick True or False

1. Influential points always change the intercept of the regression line.
2. Influential points always reduce R^2.
3. It is much more likely for a low leverage point to be influential, than a high leverage point.

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

- SLR: one response variable, one explanatory variable

$$Y = \beta_0 + \beta_1 \cdot X + \epsilon, \text{ where } \epsilon \sim N(0, \sigma_\epsilon)$$

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

$$Y = \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \cdots + \beta_p \cdot X_p + \epsilon, \text{ where } \epsilon \sim N(0, \sigma_\epsilon)$$

- Estimated coefficients (e.g. $\hat{\beta}_i$'s) now are interpreted in relation to (or "conditional on") the other variables
- β_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. $isFemale$). Then,

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 \cdot X_1 + \hat{\beta}_2 \cdot X_2$$

So then,

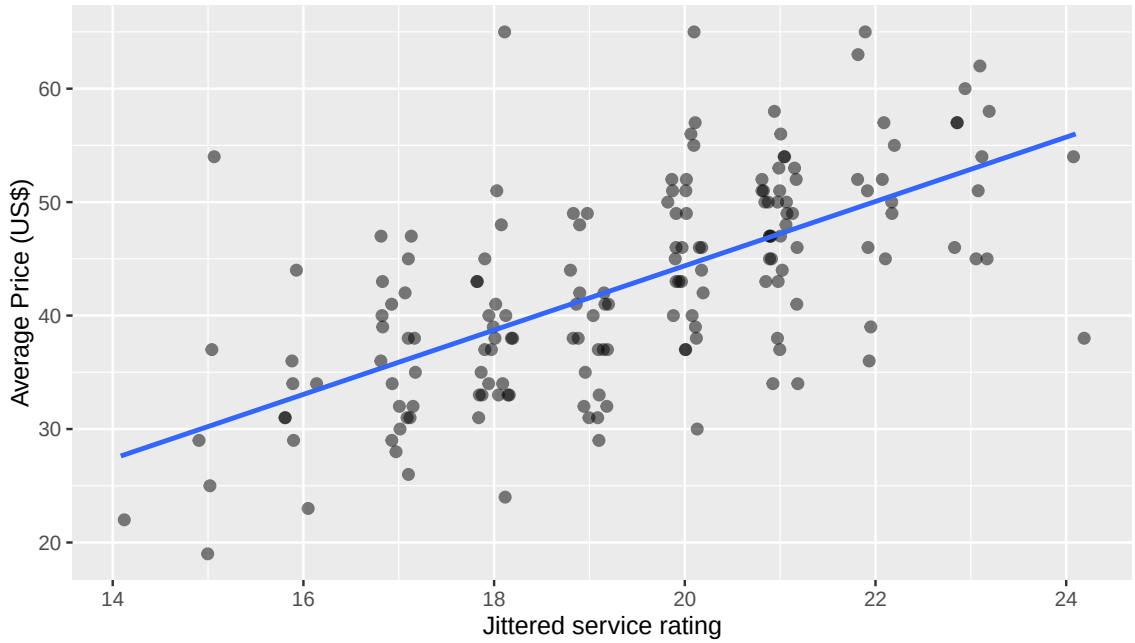
$$\begin{aligned} \text{For men,} \quad \hat{Y}|_{X_1, X_2=0} &= \hat{\beta}_0 + \hat{\beta}_1 \cdot X_1 \\ \text{For women,} \quad \hat{Y}|_{X_1, X_2=1} &= \hat{\beta}_0 + \hat{\beta}_1 \cdot X_1 + \hat{\beta}_2 \cdot 1 \\ &= \left(\hat{\beta}_0 + \hat{\beta}_2\right) + \hat{\beta}_1 \cdot X_1 \end{aligned}$$

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.

```
NYC <- read.csv("http://www.math.smith.edu/~bbaumer/mth241/nyc.csv")
ggplot(data = NYC, aes(x = jitter(Service), y = Price)) +
  geom_point(alpha = 0.5, size = 2) + geom_smooth(method = "lm", se = 0) +
  xlab("Jittered service rating") + ylab("Average Price (US$)")
lm(Price ~ Service, data = NYC)

##
## Call:
## lm(formula = Price ~ Service, data = NYC)
##
## Coefficients:
## (Intercept)      Service
##     -11.978        2.818
```



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*. Interpret the coefficients of this model. How is the quality of the food at these restaurants associated with its price?

4. Build a parallel slopes model by conditioning on the *East* variable. Interpret the coefficients of this model. What is the value of being on the East Side of Fifth Avenue?