**Agenda**

1. Categorical explanatory variable
2. Leverage, influence, and outliers

**Warmup: Regression** 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:

```
library(mosaic)
library(faraway)
m1 <- lm(Foster ~ Biological, data = twins)
coef(m1)

## (Intercept)  Biological
##    9.207599    0.901436

rsquared(m1)

## [1] 0.7779022
```

1. 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.

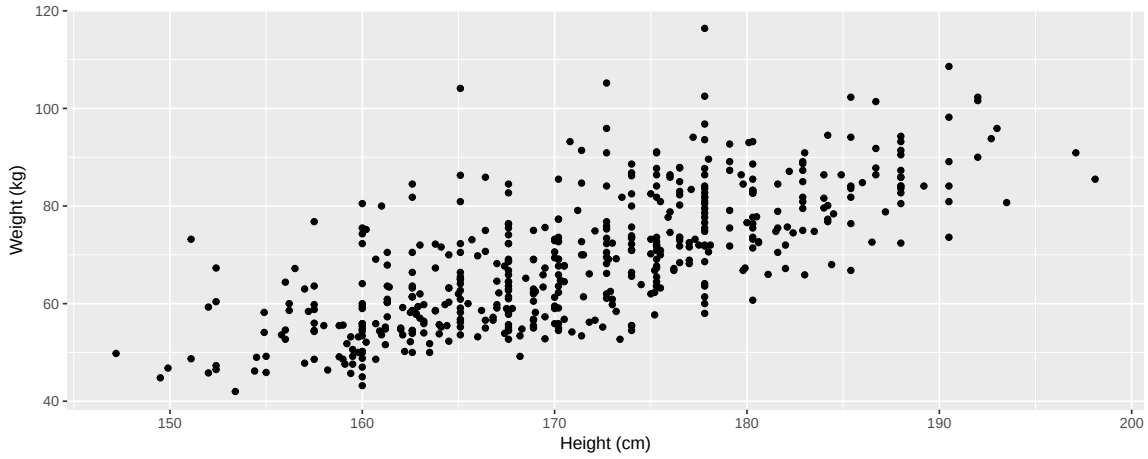
   (c) The linear model is $\widehat{Foster} = 9.2 + 0.9 \times Biological$.

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

2. Interpret the coefficients of the model.

**Height and weight** 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.

```
library(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
```

1. Describe the relationship between height and weight.

2. Write the equation of the regression line. Interpret the slope and intercept in context.

3. The correlation coefficient for height and weight is 0.72. Calculate $R^2$ and interpret it in context.

**One Categorical Explanatory Variable** Recall our Rail Trail data example, and 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 (by mathematicians) or *dummy* variables (by economists).] Such a model has the form:

$$\widehat{volume} = \hat{\beta}_0 + \hat{\beta}_1 \cdot weekday$$

```
## (Intercept)    weekday1
##   430.71429   -80.29493
```

1. How many riders does the model expect will visit the Rail Trail on a weekday?

2. How many riders does the model expect on a weekend?

3. What if it's a weekend and it's 80 degrees out?

4. Draw a scatterplot of the data and indicate this model graphically.

5. *Estimate* the $R^2$ for this model. Is it greater or less than the $R^2$ for the model with temperature as an explanatory variable?

**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.