Multiple Linear Regression

IDS 702

Exercise Part 1 (complete in template)

In your group, discuss and write your answers to the following questions:

1. What is $\hat{y}$? How is it different than y ?

   $\hat{y}$ is the predicted value of y based on the model, and y is the observed outcome

2. What is $\hat{\beta}$? How is it different than β ?

   $\hat{\beta}$ is the coefficient estimate, and β is the unknown coefficient (population parameter)

3. What is a residual?

   $y - \hat{y}$, the difference between the observed y and the predicted outcome based on the model.

4. What is a confounding variable? Why are confounding variables relevant to MLR?

   A confounding variable is a variable that is associated with both the primary independent variable of interest and the outcome. We use MLR to control for confounding effects (i.e., add measured confounding variables to the model).

5. Use the Auto data given below to answer the following (note that you can use `?Auto` after loading the data to see the description and codebook):

   a. How many rows and columns are in the data? What does each row represent?

   392 rows and 9 columns, each row represents a car

   b. How many distinct years are in the data? 13

   ```
   library(tidyverse)
   library(tidymodels)
   library(ISLR2)
   n_distinct(Auto$year)
   ```

   ```
   [1] 13
   ```

   c. If I regress mpg on horsepower, weight, and year, what are the dimensions of the outcome vector, the design matrix, and the parameter vector? You do not need any code to answer this question.

   392x1, 392x4, 4x1

Load packages and data

```
library(tidyverse)
library(tidymodels)
library(ISLR2)
```

Today's data is called "Auto" and is contained in the ISLR2 library

```
data("Auto")
```

Multiple linear regression model and notation

$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p + \epsilon, \epsilon \sim N(0, \sigma^2)$

- y : the **outcome** variable. Also called the "response" or "dependent variable". In prediction problems, this is what we are interested in predicting. In linear regression, we use continuous variables for the outcome.
- x_i: the i^{th} predictor. Also commonly referred to as "regressor", "independent variable", "covariate", "feature".
- β : "constants" or **coefficients** i.e. fixed numbers. These are **population parameters**.
- ϵ : the **error**. This quantity represents observational error, i.e. the difference between our observation and the true population-level expected value

Effectively this model says our data y is linearly related to the x_1 ,...,x_p but is not perfectly observed due to some error.

Matrix Notation:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & \ldots & x_{1p} \\ 1 & x_{21} & \ldots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & \ldots & x_{np} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

where $\begin{bmatrix} 1 & x_{11} & \ldots & x_{1p} \\ 1 & x_{21} & \ldots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & \ldots & x_{np} \end{bmatrix}$ is the **design matrix**

Why does the design matrix matter? The columns must be **linearly independent** for the coefficients to be estimated.

Then, the OLS estimates are given by $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$

Fitting a multiple regression model in R

```
myModelFit <- lm(outcome ~ predictor1 + predictor2 + predictor3 + ..., data = data-set-here)
```

we can simply 'add' in new predictors! This code template will fit the model according to the ordinary least squares (OLS) objective function, i.e., we are finding the equation that minimizes the sum of squared residuals.

You can subsequently print the coefficient estimates ($\hat{\beta}$) to the screen by calling the `summary()` function on your fitted model, e.g. `summary(myModelFit)`.

Let's fit the model regressing MPG on weight and horsepower:

```
AutoModel <- lm(mpg ~ weight + horsepower +year, data = Auto)

summary(AutoModel)
```

```
Call:
lm(formula = mpg ~ weight + horsepower + year, data = Auto)

Residuals:
    Min      1Q  Median      3Q     Max
-8.7911 -2.3220 -0.1753  2.0595 14.3527

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.372e+01  4.182e+00  -3.281  0.00113 **
weight      -6.448e-03  4.089e-04 -15.768  < 2e-16 ***
horsepower  -5.000e-03  9.439e-03  -0.530  0.59663
year         7.487e-01  5.212e-02  14.365  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.43 on 388 degrees of freedom
Multiple R-squared:  0.8083,    Adjusted R-squared:  0.8068
F-statistic: 545.4 on 3 and 388 DF,  p-value: < 2.2e-16
```

The fitted model equation:

$$\hat{y} = -13.7 - .006(weight) - .005(horsepower) + .7(year)$$

Interpreting the multiple linear regression coefficient estimates

- $\hat{\beta}_0$: the value of the outcome when all predictors=0
- $\hat{\beta}_i$: the amount that the outcome increases, on average, per unit increase in the i^{th} predictor, **holding all other predictors constant (or controlling for the other predictors)**
- $\hat{y}$: the predicted value of the outcome given a set of values of the predictors, according to the model

In the context of our problem:

- When a car has weight=0 and horsepower=0, its average mpg is -13.7 (not a meaningful interpretation)
- For each 1 lb increase in a vehicle's weight, the mean mpg decreases by .006, holding horsepower and year constant.
- For each increase in horsepower, the mean mpg decreases by .005, holding weight and year constant
- For a car that weighs 3500 lbs and has a horsepower of 130, the predicted mpg is 18.1.

  Note that predictions should be made within the range of the observed predictors. Making predictions on values outside of the range of the observed values is called **extrapolation**

Exercise Part 2 (complete in template)

1. Fit a model regressing mpg on weight and acceleration. Show the summary table and write interpretations for the predictor estimates (you don't need to interpret the intercept). Which predictors are statistically significant?

   ```
   mod1 <- lm(mpg~weight+acceleration, data=Auto)
   summary(mod1)
   ```

   ```
   Call:
   lm(formula = mpg ~ weight + acceleration, data = Auto)

   Residuals:
        Min       1Q   Median       3Q      Max
   -11.1371  -2.7860  -0.3355   2.4192  16.2096

   Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
   (Intercept)  41.0953288  1.8680355  21.999  < 2e-16 ***
   weight       -0.0072931  0.0002809 -25.966  < 2e-16 ***
   acceleration  0.2616504  0.0864755   3.026  0.00265 **
   ---
   Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

   Residual standard error: 4.288 on 389 degrees of freedom
   Multiple R-squared:  0.6997,    Adjusted R-squared:  0.6982
   F-statistic: 453.2 on 2 and 389 DF,  p-value: < 2.2e-16
   ```

   Per lb increase in a car's weight, mean mpg decreases by 0.007, holding acceleration constant.

   Per unit increase in acceleration, mean mpg increases by 0.26, holding weight constant.

   Both weight and acceleration are statistically significant.

2. Fit a model regressing mpg on displacement, weight, acceleration, and year. Show the summary table and write interpretations for the predictor estimates (you don't need to interpret the

intercept). Which predictors are statistically significant? Based on the t values, which predictor has the biggest impact on the outcome? Is this surprising?

```
mod2 <- lm(mpg~displacement+weight+acceleration+year, data=Auto)
summary(mod2)
```

```
Call:
lm(formula = mpg ~ displacement + weight + acceleration + year,
    data = Auto)

Residuals:
    Min      1Q  Median      3Q     Max
-8.5182 -2.3948 -0.1085  2.0405 14.2908

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1.527e+01  4.106e+00  -3.719 0.000229 ***
displacement  2.874e-03  5.310e-03   0.541 0.588651
weight       -6.852e-03  5.967e-04 -11.483  < 2e-16 ***
acceleration  8.555e-02  7.885e-02   1.085 0.278595
year          7.532e-01  5.118e-02  14.717  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.431 on 387 degrees of freedom
Multiple R-squared:  0.8088,    Adjusted R-squared:  0.8068
F-statistic: 409.2 on 4 and 387 DF,  p-value: < 2.2e-16
```

Per unit increase in displacement, mean mpg increases by .003, holding all else constant.

Per lb increase in weight, mean mpg decreases by .007, holding all else constant.

Per unit increase in acceleration, mean mpg increases by 0.09, holding all else constant.

Each year, mean mpg increases by .08, holding all else constant.

Weight and year are statistically significant.

Based on the t values, year (and weight) have the strongest association with mpg.

3. Compare your results for 1 and 2. What does the difference in the results say about the possible confounders for this problem?

   Year may confound the relationship between acceleration and mpg, because acceleration is statistically significant without year in the model, but not statistically significant when we control for year (i.e., add year to the model).