Lesson 6. The Simple Linear Regression Model – Part 1

Note. In Part 2 of this lesson, you can run the R code that generates the plots and outputs in here Part 1.

1 Choosing a simple linear model

- We need:
  1. One quantitative explanatory variable and one quantitative response variable
  2. A consistent linear trend

Example 1. We want to predict the price of a used Porsche from its mileage. The data for 30 used Porsches is in the `PorschePrice` data frame in the `Stat2Data` package. For each Porsche in the dataset, we have its price (in thousands of \$), age (in years), and mileage (in thousands of miles).

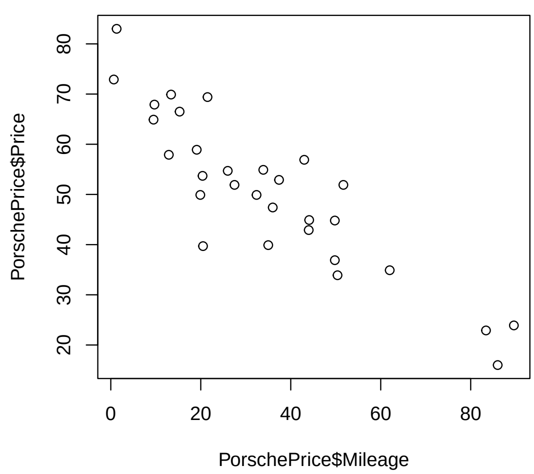
a. Identify the explanatory and response variables. Are they both quantitative?

b. Suppose we run the following R code:

```
library(Stat2Data)                                  # Access the Stat2Data package
data(PorschePrice)                                  # Import the PorschePrice data frame

plot(PorschePrice$Mileage, PorschePrice$Price)     # Make scatterplot
```

The code above imports the data frame and makes the following scatterplot:



Does the data exhibit a linear trend?

1.1 Writing the model

Example 2. State the model for used Porsche prices. Note that this is the population-level model. It should not include any numerical estimates from the sample.

2 Fitting a simple linear model

- How do we find the "best" estimates of β_0 and β_1?
- We use a technique called **least squares regression** to obtain estimates $\hat{\beta}_0$ and $\hat{\beta}_1$
- The **residual** for observation i is:
- Least squares regression works by minimizing the **sum of squared errors (SSE)**:
- The fitted model, or the **least squares line** is

Example 3. Let's continue with Examples 1 and 2. Suppose we run the following R code:

```
fit <- lm(Price ~ Mileage, data=PorschePrice)     # Fit the model

plot(PorschePrice$Mileage, PorschePrice$Price)    # Make scatterplot
abline(fit)                                       # Add fitted line to the scatterplot

summary(fit)                                      # Output a lot of useful info
anova(fit)                                        # Get SSE, among other things
```

The resulting output is on page 4.

a. The least squares line is:

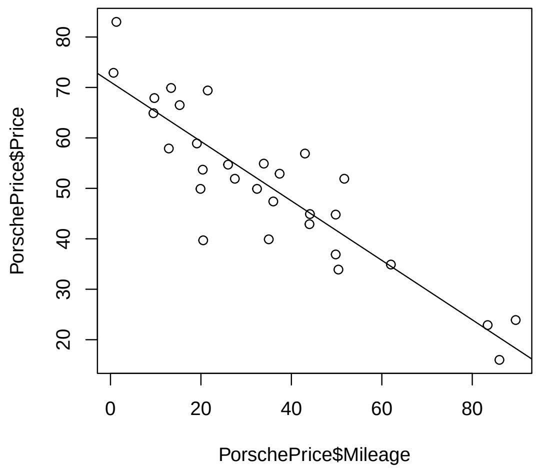
b. What can we learn from the estimated slope? Be careful about units.

c. If a used Porsche has 0 miles on it, what would we predict the price to be?

Note that this question doesn't really make sense. We are rarely interested in interpreting the estimated intercept.

d. Calculate the residual for the first car in the dataset, which has a price of \$69,400 and 21,500 miles.

e. What is the SSE?




```
Call:
lm(formula = Price ~ Mileage, data = PorschePrice)

Residuals:
     Min       1Q   Median       3Q      Max 
-19.3077  -4.0470  -0.3945   3.8374  12.6758 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 71.09045    2.36986    30.0  < 2e-16 ***
Mileage     -0.58940    0.05665   -10.4 3.98e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.17 on 28 degrees of freedom
Multiple R-squared:  0.7945,  Adjusted R-squared:  0.7872 
F-statistic: 108.3 on 1 and 28 DF,  p-value: 3.982e-11
```

A anova: 2 × 5

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
	<int>	<dbl>	<dbl>	<dbl>	<dbl>
Mileage	1	5565.685	5565.68453	108.2543	3.981734e-11
Residuals	28	1439.565	51.41304	NA	NA