Lesson 7. 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 here in Part 1.

1 Choosing a simple linear regression model

- We need:
  1. One quantitative explanatory variable and one quantitative response variable
  2. A consistent linear trend
- Suppose we have n observations of the explanatory variable X and response variable Y
- The **simple linear regression model** is

  - This is the **population-level model**
  - The slope coefficient β_1 describes the relationship between response and explanatory variables

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

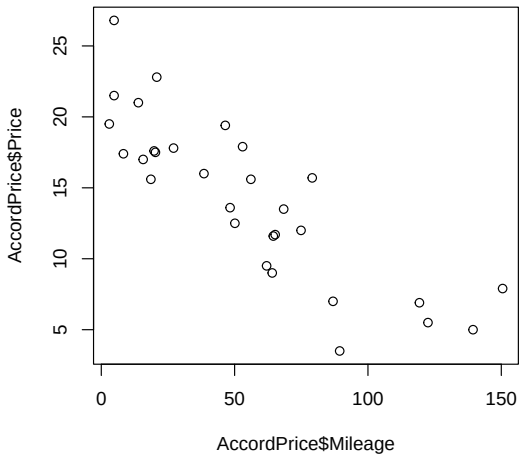
a. Identify the explanatory and response variables. Are they both quantitative?

b. Suppose we run the following R code:

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

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

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



Does the data exhibit a linear trend?

c. State the population-level model for used Accord prices.

2 Fitting a simple linear regression model

- The **fitted model** (or prediction equation) is:

  - The estimated slope $\hat{\beta}_1$ describes the estimated average relationship between the response and explanatory variables
  - Interpretation:

    On average, an increase/decrease of 1 unit in the explanatory variable is associated with an increase/decrease of $|\hat{\beta}_1|$ in the response variable.

    The underlined parts above should be rephrased to correspond to the context of the problem

2.1 Finding the fitted model

- How do we find the estimates $\hat{\beta}_0$ and $\hat{\beta}_1$?
- We use a technique called **least squares regression** to estimate the best fit

- Let y_i be the observed value of the response variable Y for observation i
- The predicted value of the response variable Y for observation i is:

- The **residual** for observation i is:

- Least squares regression works by minimizing the **sum of squared errors (SSE)**:

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

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

plot(AccordPrice$Mileage, AccordPrice$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 fitted model is:

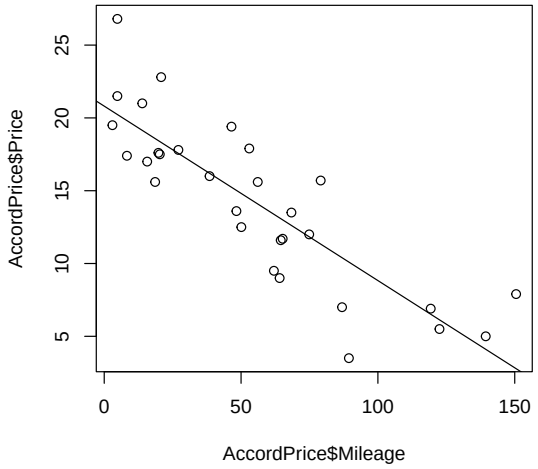
b. What can we learn from the estimated slope? Be careful about units.

c. If a used Accord 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 $12,000 and 74,900 miles.

e. What is the SSE?




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

Residuals:
    Min      1Q  Median      3Q     Max
-6.5984 -1.8169 -0.4148  1.4502  6.5655

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  20.8096     0.9529   21.84  < 2e-16 ***
Mileage      -0.1198     0.0141   -8.50 3.06e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.085 on 28 degrees of freedom
Multiple R-squared:  0.7207,  Adjusted R-squared:  0.7107
F-statistic: 72.25 on 1 and 28 DF,  p-value: 3.055e-09
```

A anova: 2 × 5

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
	<int>	<dbl>	<dbl>	<dbl>	<dbl>
Mileage	1	687.6644	687.664433	72.25284	3.055011e-09
Residuals	28	266.4892	9.517473	NA	NA

2.2 Standard deviation of the error term

- The simple linear regression model has three unknown parameters:
  - the coefficients β_0 and β_1
  - the standard deviation of the error term σ_ε
- For a simple linear regression model, the **estimated standard deviation of the error term** $\hat{\sigma}_\varepsilon$ based on the least squares fit to a sample of n observations is

- $\hat{\sigma}_\varepsilon$ is sometimes called the **regression standard error** or **residual standard error**
  - It is interpreted as
  - It gives us a feel for how far individual cases might lie above or below the regression line

Example 3. Continuing Examples 1 and 2...

a. Use the R output to calculate the regression standard error.
b. Using mileage to predict the price of a used Accord, the typical error will be around what size? Your answer should have dollars as the units.