Multiple Linear Regression Example

Problem Statement

Mileage of used cars is often thought of as a good predictor of sale prices of used cars. Does this same conjecture hold for so called “luxury cars”: Porches, Jaguars, and BMWs? More precisely, do the slopes and intercepts differ when comparing mileage and price for these three brands of cars? To answer this question, data was randomly selected from an Internet car sale site. (Tweaked a bit from Cannon et al. 2013 [Chapter 1 and Chapter 4])

Competing Hypotheses

There are many hypothesis tests to run here. It’s important to first think about the model that we will fit to address these questions. We want to predict `Price` (in thousands of dollars) based on `Mileage` (in thousands of miles). A simple linear regression equation for this would be $\hat{Price} = b_0 + b_1 * Mileage$.

We are dealing with a more complicated example in this case though. We need to also include in `CarType` to our model. Since `CarType` has three levels: `BMW`, `Porche`, and `Jaguar`, we encode this as two dummy variables with `BMW` as the baseline (since it occurs first alphabetically in the list of three car types). This model would help us determine if there is a statistical difference in the intercepts of predicting `Price` based on `Mileage` for the three car types, assuming that the slope is the same for all three lines:

$$\hat{Price} = b_0 + b_1 * Mileage + b_2 * Porche + b_3 * Jaguar.$$

This is not exactly what the problem is asking for though. It wants us to see if there is also a difference in the slopes of the three fitted lines for the three car types. To do so, we need to incorporate *interaction* terms on the dummy variables of `Porche` and `Jaguar` with `Mileage`. This also creates a baseline interaction term of `BMW:Mileage`, which is not specifically included in the model but comes into play by setting `Jaguar` and `Porche` equal to 0:

$$\hat{Price} = b_0 + b_1 * Mileage + b_2 * Porche + b_3 * Jaguar + b_4 Mileage * Jaguar + b_5 Mileage * Porche.$$

In words

- Null hypothesis: The coefficients on the parameters (including interaction terms) of the least squares regression modeling price as a function of mileage and car type are zero.
- Alternative hypothesis: At least one of the coefficients on the parameters (including interaction terms) of the least squares regression modeling price as a function of mileage and car type are nonzero.

In symbols (with annotations)

- H_0 : $\beta_i = 0$, where β_i represents the population coefficient of the least squares regression modeling price as a function of mileage and car type.
- H_A : At least one $\beta_i \neq 0$

Set α

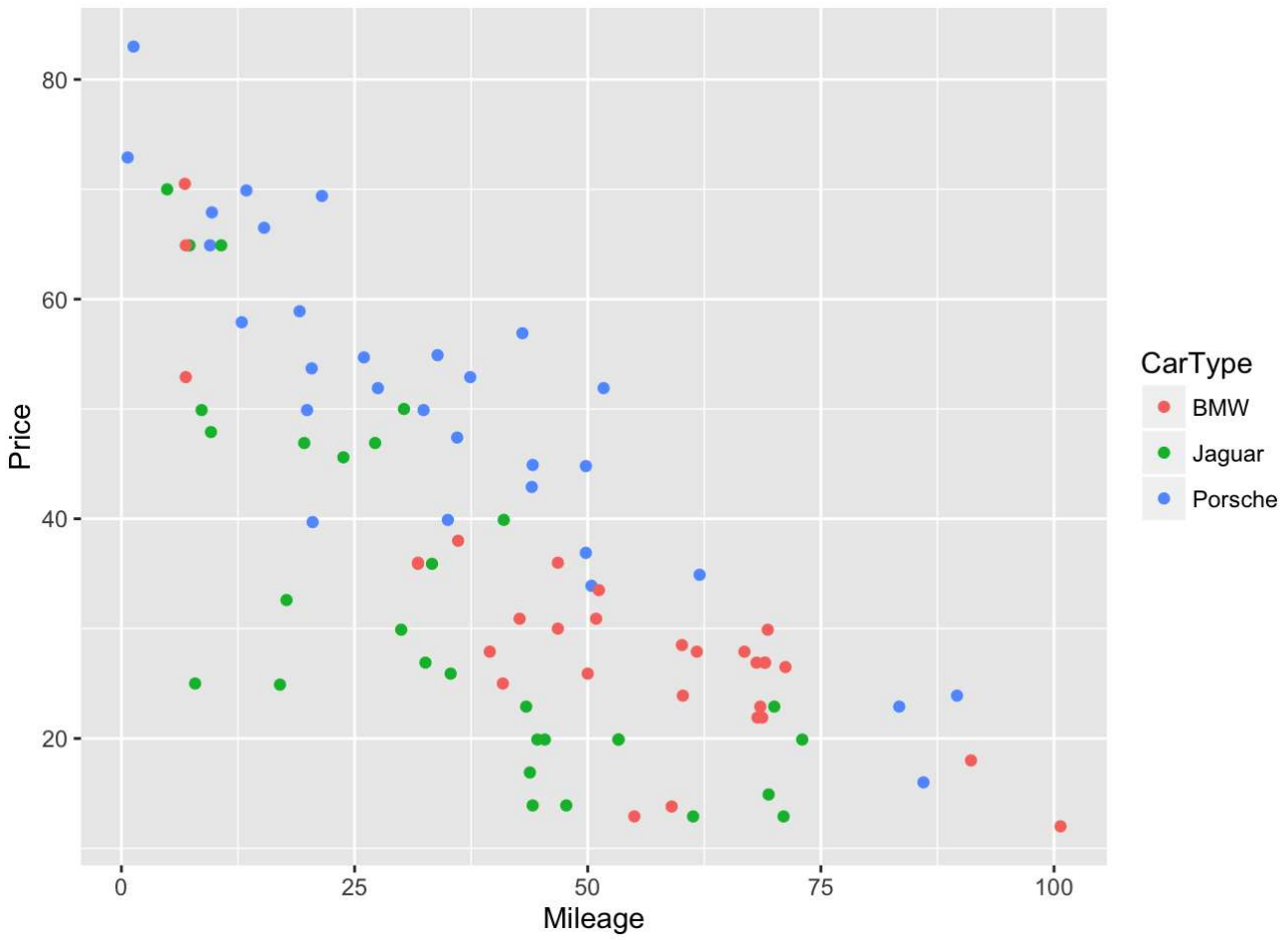
It’s important to set the significance level before starting the testing using the data. Let’s set the significance level at 5% here.

Exploring the sample data

```
library(dplyr)
library(knitr)
library(ggplot2)
library(Stat2Data)
data(ThreeCars)
ThreeCars <- ThreeCars %>%
  select(CarType, Price, Mileage) %>%
  mutate(CarType = as.character(CarType))
options(digits = 5, scipen = 20, width = 90)
```

The scatterplot below shows the relationship between mileage, price, and car type.

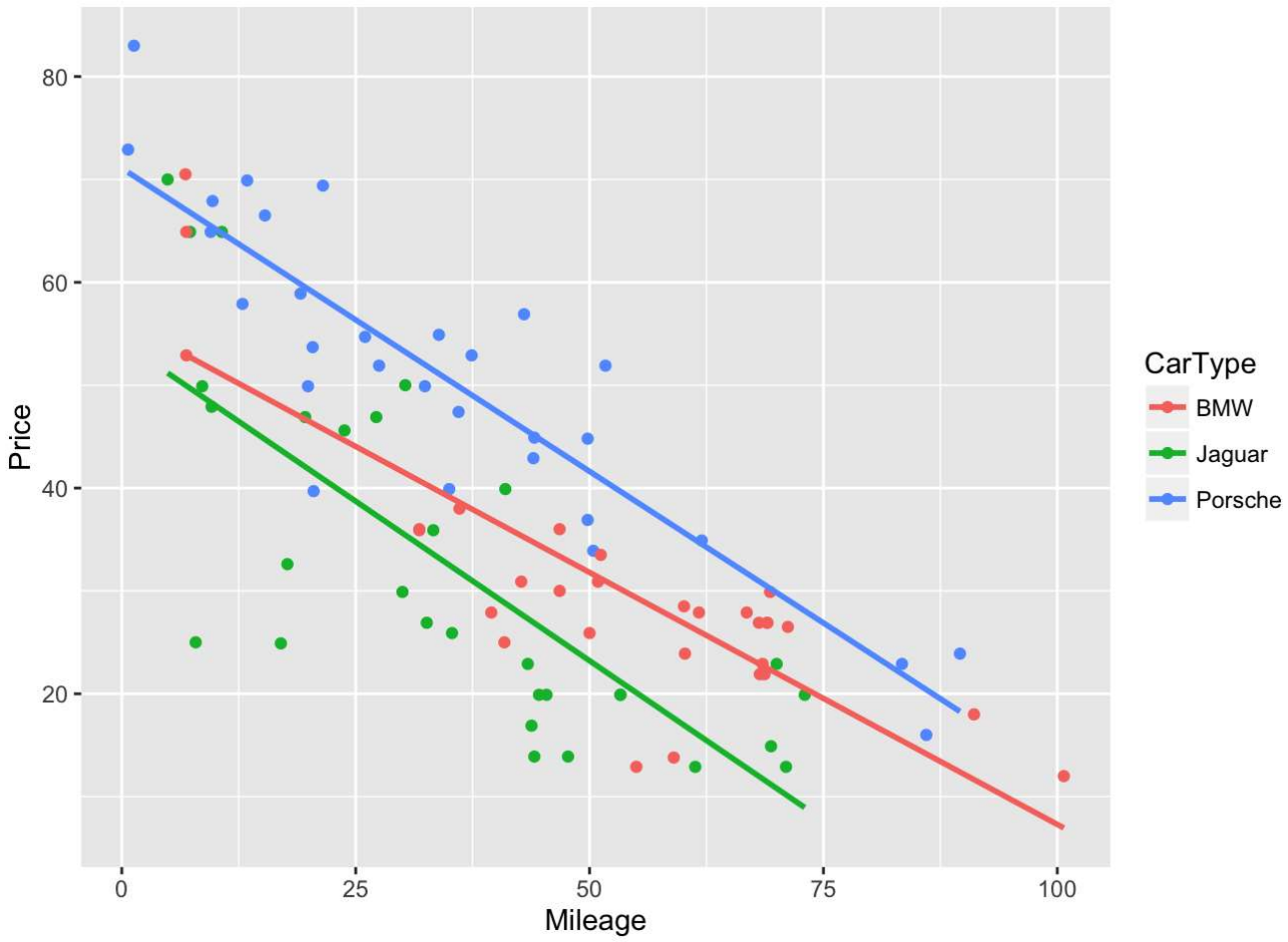
```
qplot(x = Mileage, y = Price, color = CarType, data = ThreeCars, geom = "point")
```



Guess about statistical significance

It seems that there is a difference in the intercepts of linear regression for the three car types since Porches tend to be above BMWs, which tend to be above Jaguars. BMWs and Jaguars are a bit more clustered together though. It's hard to tell exactly whether the slopes will also be statistically significantly different when looking at just the scatterplot. We add the lines below:

```
qplot(x = Mileage, y = Price, color = CarType, data = ThreeCars) +
  geom_smooth(method = "lm", se = FALSE)
```



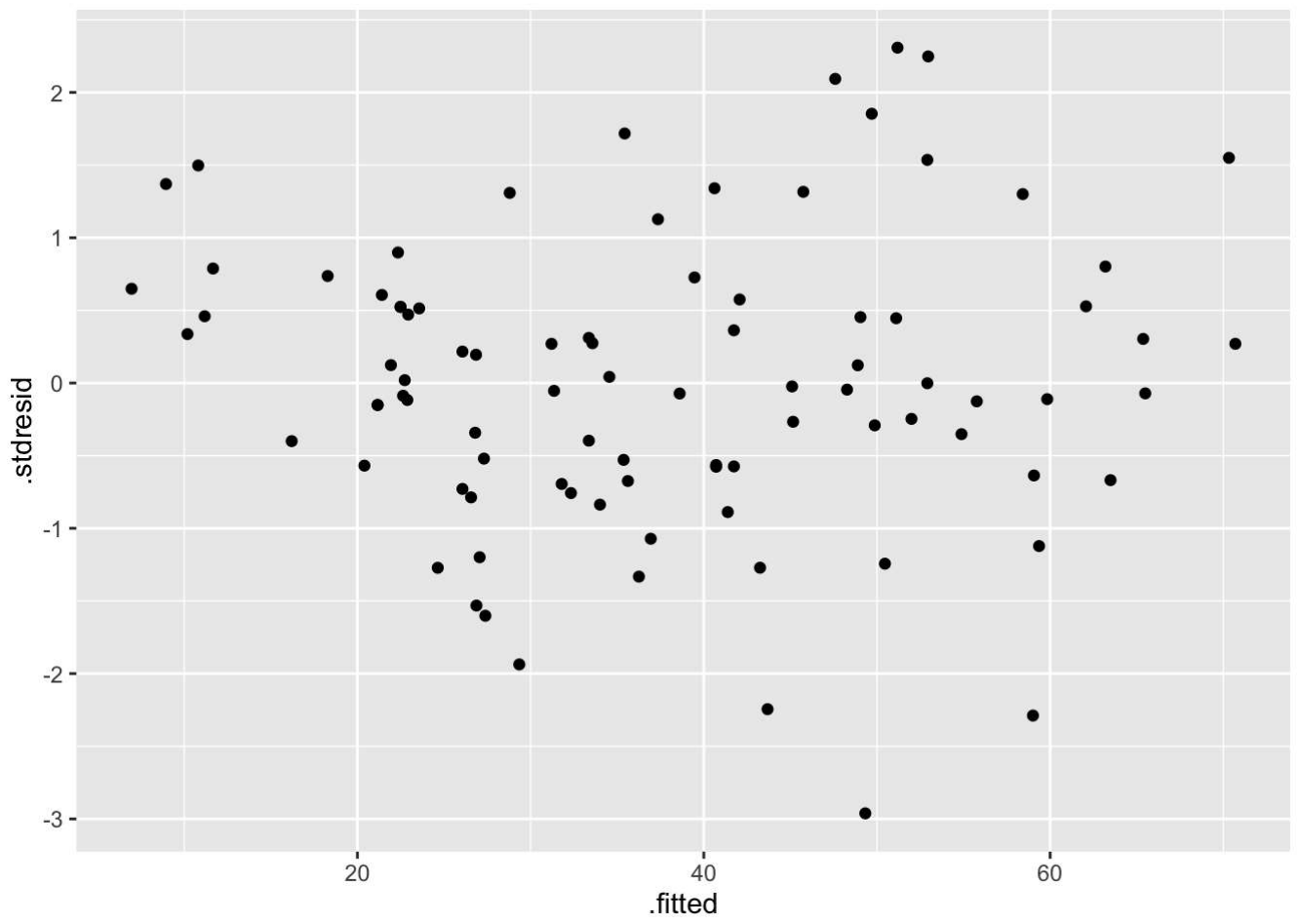
Based on the plot, we might guess that at least one of the coefficients will be statistically different since the BMW line does appear to not be parallel with the others.

Check conditions

Remember that in order to use the shortcut (formula-based, theoretical) approach, we need to check that some conditions are met.

1. *Linear relationship between between response and predictors*: You can check the scatterplots above to get a feel for a linear relationship between reasonable. The preferred methodology is to look in the residual plot to see if the standardized residuals (errors) from the model fit are randomly distributed:

```
car_mult_lm <- lm(Price ~ Mileage + CarType + Mileage:CarType, data = ThreeCars)
qplot(x = .fitted, y = .stdresid, data = car_mult_lm)
```



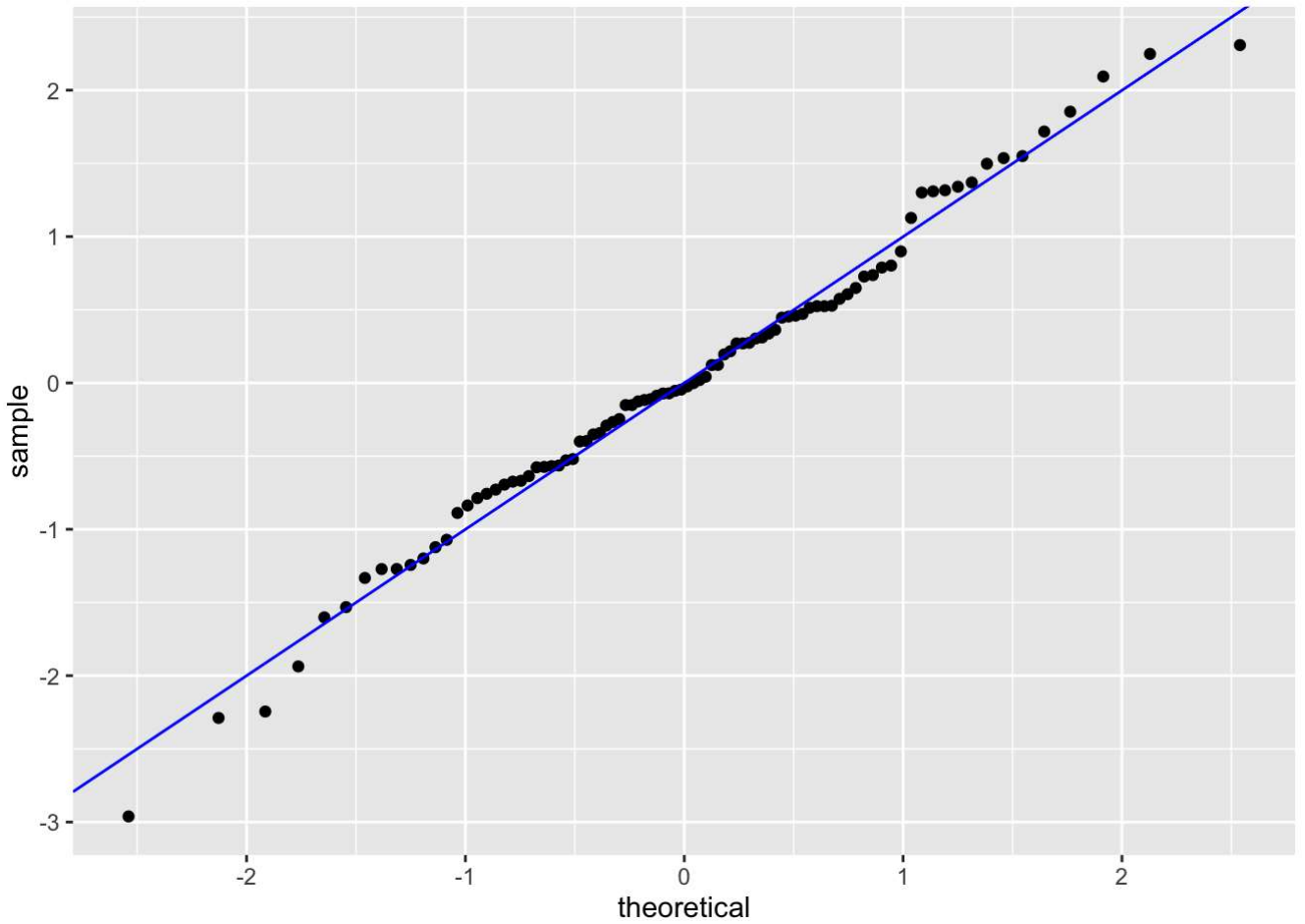
There does not appear to be any pattern (quadratic, sinusoidal, exponential, etc.) in the residuals so this condition is met.

2. *Independent observations and errors*: If cases are selected at random, the independent observations condition is met. If no time series-like patterns emerge in the residuals plot, the independent errors condition is met.

   The cars were selected at random here so the independent observations condition is met. We do not see any time series-like patterns in the residual plot above so that condition is met as well.

3. *Nearly normal residuals*: Check a Q-Q plot on the standardized residuals to see if they are approximately normally distributed.

```
qplot(sample = .stdresid, data = car_mult_lm) +
  geom_abline(color = "blue")
```



   There are some small deviations from normality but this is a pretty good fit for normality of residuals.

4. *Equal variances across explanatory variable*: Check the residuals plot for fan-shaped patterns.

   The residual plot does show a bit of a fan-shaped patter from left to right, but it is not drastic.

Test statistic

The test statistics are random variables based on the sample data. Here, we want to look at a way to estimate the population coefficients β_i. A good guess is the sample coefficients B_i. Recall that these sample coefficients are actually random variables that will vary as different samples are (theoretically, would be) collected.

We next look at our fitted regression coefficients from our sample of data:

```
summary(car_mult_lm)
```

```
##
## Call:
## lm(formula = Price ~ Mileage + CarType + Mileage:CarType, data = ThreeCars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -24.327  -4.832  -0.285   4.423  18.812
##
## Coefficients:
##                         Estimate Std. Error t value             Pr(>|t|)
## (Intercept)              56.2901     4.1551   13.55 < 0.0000000000000002 ***
## Mileage                  -0.4899     0.0723   -6.78         0.0000000016 ***
## CarTypeJaguar            -2.0626     5.2358   -0.39                0.6946
## CarTypePorsche           14.8004     5.0415    2.94                0.0043 **
## Mileage:CarTypeJaguar    -0.1304     0.1057   -1.23                0.2206
## Mileage:CarTypePorsche   -0.0995     0.0994   -1.00                0.3196
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.64 on 84 degrees of freedom
## Multiple R-squared:  0.774,  Adjusted R-squared:  0.76
## F-statistic: 57.4 on 5 and 84 DF,  p-value: <0.0000000000000002
```

We are looking to see how likely is it for us to have observed sample coefficients $b_{i,obs}$ or more extreme assuming that the population coefficients are 0 (assuming the null hypothesis is true). If the conditions are met and assuming H_0 is true, we can "standardize" this original test statistic of B_i into T statistics that follow a t distribution with degrees of freedom equal to $df = n - k$ where k is the number of parameters in the model:

$$T = \frac{B_i - 0}{SE_i} \sim t(df = n - k)$$

where SE_i represents the standard deviation of the distribution of the sample coefficients.

Observed test statistic

While one could compute these observed test statistics by "hand", the focus here is on the set-up of the problem and in understanding which formula for the test statistic applies. We can use the `lm` function here to fit a line and conduct the hypothesis test. (We've already run this code earlier in the analysis, but it is shown here again for clarity.)

```
car_mult_lm <- lm(Price ~ Mileage + CarType + Mileage:CarType, data = ThreeCars)
summary(car_mult_lm)$coefficients
```

```
##                         Estimate Std. Error  t value                          Pr(>|t|)
## (Intercept)             56.29007   4.155120 13.54716 0.00000000000000000000077348
## Mileage                 -0.48988   0.072272 -6.77826 0.00000000157516917179816623 8
## CarTypeJaguar           -2.06261   5.235754 -0.39395 0.69461832640811038697137291 8
## CarTypePorsche          14.80038   5.041489  2.93572 0.00429165979604010771469857 5
## Mileage:CarTypeJaguar   -0.13042   0.105670 -1.23420 0.22056881252210527866175482 4
## Mileage:CarTypePorsche -0.09952   0.099403 -1.00118 0.31961562148914629544194099 3
```

Interpretations of the coefficients here need to also incorporate in the other terms in the model. We will address a couple of the b_i value interpretations below:

> For every one thousand mile increase in `Mileage` for a BMW car (holding all other variables constant), we expect `Price` to decrease by 0.48988 thousands of dollars ($489.88).

> We predict Jaguars to cost $2062.61 less than BMWs and Porches to cost $14,800.37 more than BMWs (holding mileage and interaction terms fixed).

> For every one thousand mile increase in `Mileage` for a Jaguar car, we expect `Price` will decrease by 0.6203 (0.48988 + 0.13042) thousands of dollars ($620.30) (holding all other variables constant).

> For every one thousand mile increase in `Mileage` for a Porche car, we expect `Price` will decrease by 0.5894 (0.48988 + 0.09952) thousands of dollars ($589.40) (holding all other variables constant).

Note that an interpretation of the observed intercept can also be done:

> we expect a BMW car with zero miles to have a price of $56,290.07.

We should be a little cautious of this prediction though since there are no cars in our sample of used cars that have zero mileage.

Compute p-values

The p-values correspond to the probability of observing a t_{90-6} value of $b_{i,obs}$ or more extreme in our null distribution. We see that the `(Intercept)`, `Mileage` and `CarTypePorche` are statistically significant at the 5% level, while the others are not.

We show below how we can obtain one of these p-values (for `CarTypeJaguar`) in R directly:

```
2 * pt(-0.3939471629, df = 84, lower.tail = TRUE)
```

```
## [1] 0.69462
```

State conclusion

We, therefore, have sufficient evidence to reject the null hypothesis for `Mileage` and the intercept on `Porche` compared to the intercept on `BMW` (which is also significant), assuming the other terms are in the model. Our initial guess that the slopes would differ on the lines for at least one of the three fitted lines based on car type was not validated by our statistical analyses here though.

Cannon, Ann R., George W. Cobb, Bradley A. Hartlaub, Julie M. Legler, Robin H. Lock, Thomas L. Moore, Allan J. Rossman, and Jeffrey A. Witmer. 2013. *STAT2 - Building Models for a World of Data*.