Creating Models for Research

Last time we saw there were two questions in evaluating our model:

1. **Is there evidence of at least one non-zero (0) coefficient?**
2. **Is there also evidence that each individual coefficient is non-zero?**

With the advertising data we saw the answer to the first question was yes, but the second was not a yes for every variable.

Let's revisit this data. Recall, we have data from a hypothetical company's advertising spending and sales for the last three years. The variables are `TV`, `Radio`, and `Newspaper` which are the amount spent on the respective ads in thousands of dollars. The final variable is `Sales` which is that month's sales in units.

Open a new R notebook and import the data using this URL: https://raw.githubusercontent.com/moravian-mspa/MGMT555/master/m5_Tutorial_FitandEffect/Advertising.csv

```
library(tidyverse)
Advertising <- read_csv("https://raw.githubusercontent.com/moravian-mspa/MGMT555/master/m5_Tutorial_Fita
```

```
Rows: 36 Columns: 5
-- Column specification --------------------------------------------------------
Delimiter: ","
dbl (5): Month, TV, Radio, Newspaper, Sales

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
head(Advertising)
```

```
# A tibble: 6 x 5
  Month    TV Radio Newspaper Sales
  <dbl> <dbl> <dbl>     <dbl> <dbl>
1     1 230.   37.8      69.2   663
2     2  44.5  39.3      45.1   312
3     3  17.2  45.9      69.3   279
4     4 152.   41.3      58.5   555
5     5 181.   10.8      58.4   387
6     6   8.7  48.9      75     216
```

We create a model of the form:

$\texttt{Sales} = b_0 + b_1\texttt{TV} + b_2\texttt{Newspaper} + b_3\texttt{Radio}$

```
summary(lm(Sales ~ TV + Newspaper + Radio, data=Advertising))
```

```
Call:
lm(formula = Sales ~ TV + Newspaper + Radio, data = Advertising)

Residuals:
    Min      1Q  Median      3Q     Max
-147.90  -27.41   16.41   37.81   74.63

Coefficients:
```

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 104.9599    24.7455   4.242 0.000177 ***
TV            1.2859     0.1038  12.390 9.39e-14 ***
Newspaper    -0.3562     0.4958  -0.718 0.477686
Radio         5.6130     0.8516   6.591 1.98e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 55.14 on 32 degrees of freedom
Multiple R-squared:  0.8687,    Adjusted R-squared:  0.8564
F-statistic: 70.56 on 3 and 32 DF,  p-value: 3.38e-14
```

At this point we saw that the model was significant, but the variable `Newspaper` was not significant. Let's try another model without that variable.

```
summary(lm(Sales ~ TV + Radio, data=Advertising))

Call:
lm(formula = Sales ~ TV + Radio, data = Advertising)

Residuals:
    Min      1Q  Median      3Q     Max
-149.53  -28.00   13.35   41.02   69.98

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  99.2844    23.2781   4.265 0.000158 ***
TV            1.3002     0.1011  12.859 2.11e-14 ***
Radio         5.2133     0.6401   8.145 2.11e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 54.73 on 33 degrees of freedom
Multiple R-squared:  0.8666,    Adjusted R-squared:  0.8585
F-statistic: 107.2 on 2 and 33 DF,  p-value: 3.689e-15
```

Notice our adjusted-R^2 decreased slightly (not good), but the overall p-value is smaller (good) and the coefficients on each of the other variables changed slightly. We are more confident in these numbers.

Backward Elimination

The process we just used is called **backward elimination**. Here's how it works.

1. Create a model with all the predictor variables that could reasonably be included.
2. Find the variable with the largest p-value. If it is larger than our threshold (we will use 0.05) then that variable is eliminated from our model.
3. Create a new model with the remaining variables and repeat until all remaining variables have p-values less than the threshold.

It is important to note that in this technique, we eliminate one variable at a time.

In fact there are several techniques that could be used to decide on which predictors are most appropriate to keep, but we will focus on using backward elimination in this course.

Another example

Imagine an antique clock dealer has collected data on recent auctions of grandfather clocks. The variables are `Price`: final selling price, `Bidders`: the number of bidders, `Age`: the age of the clock, `Temp`: the outside

temperature the day of the auction.

Here is a link to the auction data. Copy the link address and import the data to your R notebook.

```
clocks <- read_csv("https://raw.githubusercontent.com/moravian-mspa/MGMT555/master/m5_Tutorial_Creating
```

```
Rows: 32 Columns: 4
-- Column specification --------------------------------------------------------
Delimiter: ","
dbl (4): Age, Bidders, Price, Temp

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

Let's check the correlations of these variables.

```
cor(clocks)
```

```
                Age     Bidders       Price        Temp
Age      1.00000000 -0.25374910  0.73023321 -0.02776564
Bidders -0.25374910  1.00000000  0.39464036 -0.03929614
Price    0.73023321  0.39464036  1.00000000 -0.09977079
Temp    -0.02776564 -0.03929614 -0.09977079  1.00000000
```

We are most interested in what correlates well with `Price`. It looks like `Temp` has the least correlation.

We can create a model with all these variables.

```
summary(lm(Price ~ Bidders + Age + Temp, data=clocks))
```

```
Call:
lm(formula = Price ~ Bidders + Age + Temp, data = clocks)

Residuals:
     Min       1Q   Median       3Q      Max
-177.585 -119.716   -0.102   91.672  232.665

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1245.7241   205.7941  -6.053 1.59e-06 ***
Bidders        85.4659     8.7625   9.754 1.66e-10 ***
Age            12.7067     0.9079  13.996 3.64e-14 ***
Temp           -1.2855     1.5459  -0.832    0.413
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 133.9 on 28 degrees of freedom
Multiple R-squared:  0.8953,    Adjusted R-squared:  0.8841
F-statistic: 79.81 on 3 and 28 DF,  p-value: 7.833e-14
```

As we suspected, the outside temperature is the least significant, so we will remove it.

```
summary(lm(Price ~ Bidders + Age, data=clocks ))
```

```
Call:
lm(formula = Price ~ Bidders + Age, data = clocks)

Residuals:
   Min     1Q Median     3Q    Max
```

```
-207.2 -117.8   16.5  102.7  213.5

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1336.7221   173.3561  -7.711 1.67e-08 ***
Bidders        85.8151     8.7058   9.857 9.14e-11 ***
Age            12.7362     0.9024  14.114 1.60e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 133.1 on 29 degrees of freedom
Multiple R-squared:  0.8927,    Adjusted R-squared:  0.8853
F-statistic: 120.7 on 2 and 29 DF,  p-value: 8.769e-15
```

Every variable is significant, so we are done.