Discussion 3

Rummerfield & Berman

10/20/2017

Linear Models in R

In today's lesson, one of the goals is to learn more about linear model objects in R. So far, we have glossed over the types of objects within R. Let's load a set of data, the cereal dataset,

Cereal	Brand name of cereal
Calories	Number of calories per serving
Sugar	Number of grams of sugar (per serving)
Fiber	Number of grams of fiber (per serving)

```r
## Load the data from Stat2Data
data("Cereal")
```

The goal of this study was to learn if cereals high in fiber are also high in sugar and calories. Let's build the simple linear regression models the researchers were intrested in.

```r
cereal.mod <- lm(Sugar ~ Fiber, data = Cereal)
```

In the past few weeks, we've learned how to get the residuals, fitted values, estimated coefficients, and the standard error of the coefficients

```r
## Residuals
resid( cereal.mod )
## Standardized residuals
rstandard( cereal.mod )
## Fitted values
fitted.values( cereal.mod )
## Estimated coefficients, standard errors of estimated coefficients
summary( cereal.mod )
```

```
##
## Call:
## lm(formula = Sugar ~ Fiber, data = Cereal)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.0154 -3.5808  0.5722  3.0866  7.6254
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   8.0154     0.9561   8.383 8.69e-10 ***
## Fiber        -0.6408     0.1890  -3.390  0.00179 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.039 on 34 degrees of freedom
## Multiple R-squared:  0.2526, Adjusted R-squared:  0.2306
## F-statistic: 11.49 on 1 and 34 DF,  p-value: 0.001786
```

Prediction and Confidence Intervals

We know that our model is only approximate, in that, it does not give us the truth. This means there is some uncertainty and we need to quantify that uncertainty. We'll use intervals to accomplish this.

Let's do an example using the cereal data, let's estimate how many grams of sugar a cereal with 4 grams of fiber would have:

```
predict( cereal.mod, newdata = list( Fiber = 4 ) )
```

```
##       1
## 5.45223
```

Note that you can compute this using the `summary(cereal.mod)` like so:

```
coefficients( cereal.mod )[1] + 4 * coefficients( cereal.mod )[2]
```

```
##         [,1]
## [1,] 5.45223
```

Does this mean any cereal with 4 grams of fiber will have 5.45 grams of sugar?

Let's create a 95% confidence interval for the mean grams of sugar for the average cereal with 4 grams of fiber and the mean grams of sugar for the average cereal with 8 grams of fiber:

```
predict(cereal.mod, newdata = list(Fiber = c(4,8)), interval = "confidence", level = 0.95)
```

```
##        fit      lwr      upr
## 1 5.452230 4.075226 6.829234
## 2 2.889042 0.711934 5.066150
```

Let's create a 95% prediction interval for the mean grams of sugar for a new cereal with 4 grams of fiber per serving and the mean grams of sugar for a new cereal with 8 grams of fiber per serving:

```
predict(cereal.mod, newdata = list(Fiber = c(4,8)), interval = "prediction", level = 0.95)
```

```
##        fit       lwr      upr
## 1 5.452230 -2.870707 13.77517
## 2 2.889042 -5.603009 11.38109
```

Analysis of Variance (ANOVA)

In the past few weeks of class, we've learned about analysis of variance (ANOVA), but what is ANOVA? It is analyzing how of the overall variance is explained by the linear regression. The total sum of squares of the data is calculated by $\sum_i (y_i - \bar{y})^2$. We can express the total sum of squares as the sum of two different sum of squares terms,

$$\begin{aligned}\sum_i (y_i - \bar{y})^2 &= \sum_i (y_i - \hat{y}_i + \hat{y}_i - \bar{y})^2 \\ &= \sum_i (y_i - \hat{y}_i)^2 + \sum_i (\hat{y}_i - \bar{y})^2\end{aligned}$$

The term $\sum_i (\hat{y}_i - \bar{y})^2$ is the sum of squares due to regression (SSR), while the term $\sum_i (y_i - \hat{y}_i)^2$ is the residual sum of squares (RSS). Be careful with the notation! Some statisticians (like the authors of the textbook) refer to the residual sum of squares (RSS) as the sum of squares due to error, abreviated as SSE. To avoid making mistakes, the authors of the textbook refer to the sum of squares due to regression (SSR) as the sum of squares due to the model (SS Model). Luckily for us there is a command in R which can return the sums of squares, `anova()`

```
anova( cereal.mod )
```

```
## Analysis of Variance Table
##
## Response: Sugar
##           Df Sum Sq Mean Sq F value   Pr(>F)
## Fiber      1 187.44 187.443   11.49 0.001786 **
## Residuals 34 554.66  16.314
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The command also returns the relevant degrees of freedom and the mean squared terms.