The F-test in R

In this lesson, we will perform both the full and partial F-tests in R.

Recall again, the Amazon book data. The data consists of data on $n = 325$ books and includes measurements of:

- `aprice` : The price listed on Amazon (dollars)
- `lprice` : The book's list price (dollars)
- `weight` : The book's weight (ounces)
- `pages` : The number of pages in the book
- `height` : The book's height (inches)
- `width` : The book's width (inches)
- `thick` : The thickness of the book (inches)
- `cover` : Whether the book is a hard cover of paperback.
- And other variables...

We'll explore model using `lprice` , `pages` , and `width` to predict `aprice` . But first, we'll repeat the data cleaning from our lesson on t-tests. For all tests in this lesson, let $\alpha = 0.05$.

In [6]:

```
library(RCurl) #a package that includes the function getURL(), which allows for
library(ggplot2)
url = getURL(paste0("https://raw.githubusercontent.com/bzaharatos/",
                    "-Statistical-Modeling-for-Data-Science-Applications/",
                    "master/Modern%20Regression%20Analysis%20/Datasets/amazon.tx
amazon = read.csv(text = url, sep = "\t")
names(amazon)
df = data.frame(aprice = amazon$Amazon.Price, lprice = as.numeric(amazon$List.Pr
                pages = amazon$NumPages, width = amazon$Width, weight = amazon$W
                height = amazon$Height, thick = amazon$Thick, cover = amazon$Har

#cleaning the data, as was done in our lesson on t-tests
df$weight[which(is.na(df$weight))] = mean(df$weight, na.rm = TRUE)
df$pages[which(is.na(df$pages))] = mean(df$pages, na.rm = TRUE)
df$height[which(is.na(df$height))] = mean(df$height, na.rm = TRUE)
df$width[which(is.na(df$width))] = mean(df$width, na.rm = TRUE)
df$thick[which(is.na(df$thick))] = mean(df$thick, na.rm = TRUE)
df = df[-205,]
summary(df)
```

1. 'Title'
2. 'Author'

3. 'List.Price'
4. 'Amazon.Price'
5. 'Hard..Paper'
6. 'NumPages'
7. 'Publisher'
8. 'Pub.year'
9. 'ISBN.10'
10. 'Height'
11. 'Width'
12. 'Thick'
13. 'Weight..oz.'

```
     aprice           lprice            pages            width      
 Min.   :  0.770   Min.   :  1.50   Min.   : 24.0   Min.   :4.100  
 1st Qu.:  8.598   1st Qu.: 13.95   1st Qu.:208.0   1st Qu.:5.200  
 Median : 10.200   Median : 15.00   Median :320.0   Median :5.400  
 Mean   : 13.010   Mean   : 18.58   Mean   :335.8   Mean   :5.584  
 3rd Qu.: 13.033   3rd Qu.: 19.95   3rd Qu.:416.0   3rd Qu.:5.900  
 Max.   :139.950   Max.   :139.95   Max.   :896.0   Max.   :9.500  
     weight          height           thick        cover  
 Min.   : 1.20   Min.   : 5.100   Min.   :0.100   H: 89  
 1st Qu.: 7.80   1st Qu.: 7.900   1st Qu.:0.600   P:235  
 Median :11.20   Median : 8.100   Median :0.900          
 Mean   :12.48   Mean   : 8.161   Mean   :0.908          
 3rd Qu.:16.00   3rd Qu.: 8.500   3rd Qu.:1.100          
 Max.   :35.20   Max.   :12.100   Max.   :2.100          
```

Let's fit the "full" model from our lesson on t-tests, namely, the model that includes `lprice`, `pages`, and `width` as predictors.

In [7]:

```r
lm_amazon = lm(aprice ~ lprice + pages + width, data = df)
summary(lm_amazon)
```

```
Call:
lm(formula = aprice ~ lprice + pages + width, data = df)

Residuals:
     Min       1Q   Median       3Q      Max 
-19.3092  -1.7824  -0.0695   1.3374  22.9248 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.862994   1.573723   0.548    0.584    
lprice       0.854834   0.017848  47.895  < 2e-16 ***
pages       -0.006044   0.001348  -4.482 1.03e-05 ***
width       -0.305456   0.285426  -1.070    0.285    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.774 on 320 degrees of freedom
Multiple R-squared:  0.9089,	Adjusted R-squared:  0.908 
F-statistic:  1064 on 3 and 320 DF,  p-value: < 2.2e-16
```

First, note that the full F-test has a very large F-statistic (1064), and very small p-value (2.2×10^{-16}, effectively zero). Typically, we should look at the full F-test first, to see if there is any evidence that any of the predictors are necesary in the model. Only after a significant full F-test should we look at an individual t-test.

We note again that the t-test associated with `width` is not significant, suggesting that there is no evidence that the parameter associated with `width` is different from zero.

But even though `pages` is significant, it seems clear that `lprice` is most strongly associated with `aprice` (pages predictor value is very close to 0. So, we might look at an F-test comparing the models:

$$H_0 : Y_i = \beta_0 + \beta_{lprice}\,(lprice) + \varepsilon_i$$

with

$$H_1 : \text{number of pages or width (or both) should be included in the model.}$$

In [14]:

```
lm_amazon_reduced = lm(aprice ~ lprice, data = df)
anova(lm_amazon_reduced, lm_amazon)
```

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
322	4846.160	NA	NA	NA	NA
320	4557.841	2	288.3194	10.12126	5.46791e-05

Note that the p-value associated with this partial F-test is small ($5.46791 \times 10^{-5} < \alpha = 0.05$). This, we conclude that there is evidence that the reduced model is insufficient, and that we need at least one of the other predictors. We know that `width` is not statistically significant, and so we will only add back `pages` . This would leave us with the model

$$Y_i = \beta_0 + \beta_{lprice}\,(lprice) + \beta_{pages}\,(pages) + \varepsilon_i.$$

Interestingly, F-tests can be used when comparing two models that differ only by one predictor. For example, comparing

$$\omega : Y_i = \beta_0 + \beta_{lprice}\,(lprice) + \beta_{pages}\,(pages) + \varepsilon_i$$

with

$$\Omega : Y_i = \beta_0 + \beta_{lprice}\,(lprice) + \beta_{pages}\,(pages) + \beta_{width}\,(width) + \varepsilon_i.$$

Does the individual t-test and the F-test give consistent results? Let's check!

In [16]:

```
lm_amazon_reduced2 = lm(aprice ~ lprice + pages, data = df)
anova(lm_amazon_reduced2, lm_amazon)
summary(lm_amazon)
```

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
321	4574.153	NA	NA	NA	NA
320	4557.841	1	16.31249	1.145279	0.2853462

```
Call:
lm(formula = aprice ~ lprice + pages + width, data = df)

Residuals:
     Min       1Q   Median       3Q      Max 
-19.3092  -1.7824  -0.0695   1.3374  22.9248 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.862994   1.573723   0.548    0.584    
lprice       0.854834   0.017848  47.895  < 2e-16 ***
pages       -0.006044   0.001348  -4.482 1.03e-05 ***
width       -0.305456   0.285426  -1.070    0.285    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.774 on 320 degrees of freedom
Multiple R-squared:  0.9089,	Adjusted R-squared:  0.908 
F-statistic:  1064 on 3 and 320 DF,  p-value: < 2.2e-16
```

Notice that the p-value for the individual t-test for the parameter associated with `width`, and the p-value for this partial F-test are the same! This is not an accident, but a consequence of the relationship between the t-distribution and the F-distribution: if $X \sim t(n)$ then $X^2 \sim F_{1,n}$.

In [20]:

```
summary(lm_amazon_reduced2)
```

```
Call:
lm(formula = aprice ~ lprice + pages, data = df)

Residuals:
     Min       1Q   Median       3Q      Max 
-19.0969  -1.8256  -0.0329   1.4436  23.3954 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -0.727973   0.516361  -1.410     0.16    
lprice       0.844690   0.015127  55.841  < 2e-16 ***
pages       -0.005824   0.001333  -4.369 1.69e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.775 on 321 degrees of freedom
Multiple R-squared:  0.9086,	Adjusted R-squared:  0.908 
F-statistic:  1595 on 2 and 321 DF,  p-value: < 2.2e-16
```

If wanting to do a confident interval for the mean response in R, we do the following: `predict(lm_data, new=x*, interval="confidence")`. Where x* is the new data points we're implementing. The `level` parameter sets the confidence level.

For a CI of the parameters we would have `confint(lm_data)`

In [ ]: