Week 6 Option 11

Nosky, Christopher

1/23/2021

This week I am going to attempt to kill two birds with one stone, that is to say that in this weeks discussion while I will be looking into Question 11:

***“Compare the models albumSales.2 and albumSales.3 using anova(). How can you say that the model albumSales.3 is a better than the model albumSales.2?”***.

In additon to answering this question I’m also attempting to do this using Rmarkdown, I’m doing this second part primarily because I want to gain increased comfort in using Rmarkdown and as always, practice makes perfect.

So to start off, what did we learn this week? We learned that regression is used to test a model’s fit of the data being presented. Through the course of reading this week (or in my case reading, re-reading and banging my head against the wall) we were taken from performing a simple regression with the output variable and a single predictor variable through performing a multiple regression with the output with multiple predictor variables. The output looked like this:

##   
## Call:  
## lm(formula = sales ~ adverts, data = album2)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -152.949 -43.796 -0.393 37.040 211.866   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.341e+02 7.537e+00 17.799 <2e-16 \*\*\*  
## adverts 9.612e-02 9.632e-03 9.979 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 65.99 on 198 degrees of freedom  
## Multiple R-squared: 0.3346, Adjusted R-squared: 0.3313   
## F-statistic: 99.59 on 1 and 198 DF, p-value: < 2.2e-16

As presented in the reading, this output from albumSales.2 shows us quite a lot. Starting with the function used to call this output at the top (oddly enough I didn’t notice this until it was pointed out) followed by the residuals or the difference between the observed value and the mean value of the model, the coefficients giving the intercept and slope of the regression and finally the last section Residual standard error, Multiple R-squared and F-statistic. The values I’m most interested in for answering this week’s topic are from Multiple R-squared: 0.3346 and Adjusted R-squared: 0.3313 these values describe the overall model. These values tell us how much of the variability in album sales can be accounted for by advertising budget, which the book converts to a percentage of 33.5%. Next we will look at the summary of albumSales.3.

##   
## Call:  
## lm(formula = sales ~ adverts + airplay + attract, data = album2)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -121.324 -28.336 -0.451 28.967 144.132   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -26.612958 17.350001 -1.534 0.127   
## adverts 0.084885 0.006923 12.261 < 2e-16 \*\*\*  
## airplay 3.367425 0.277771 12.123 < 2e-16 \*\*\*  
## attract 11.086335 2.437849 4.548 9.49e-06 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 47.09 on 196 degrees of freedom  
## Multiple R-squared: 0.6647, Adjusted R-squared: 0.6595   
## F-statistic: 129.5 on 3 and 196 DF, p-value: < 2.2e-16

Here we see the same summary output as before but with the additional predictor variables of airplay and attract added to the model. Again for the purpose of the question we will look at the R-squared values and see that now this model shows that about 66% of the variability in album sales can be accounted for by advertising, airplay and attractiveness. With these regression summaries we might at this point be able to make an educated guess that albumSales.3 is a better model than albumSales.2.

To know which model is a better fit we need to compare them to see if the

***‘ value is significantly higher in the second model than in the first’***

The significance can be tested using an F-ratio which the formula used is for a single predictor variable or if we have multiple predictors . Fortunatly, for those of us who look at those equations and scratch our heads while saying “huh?!?” there is an easier way thanks to R. To do this we will use the anova() function which if I remember correctly from the reading is ‘Analysis of Variation’ which looks like this:

anova(albumSales.2, albumSales.3)

## Analysis of Variance Table  
##   
## Model 1: sales ~ adverts  
## Model 2: sales ~ adverts + airplay + attract  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 198 862264   
## 2 196 434575 2 427690 96.447 < 2.2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

What does this all mean? Well I’m not an expert, but based on what I’ve read this week the anova() function shows us degrees of freedom, sum of squares, the F-ratio and the p value. The p-value can help provide a significance interpretation of our F-ratio (or F-statistic) which is showing us that the F-ratio 96.447 < 2.2e-16 or as the book describes it a *“very small value and we can say that albumSales.3 significantly improved the fit of the model to the data compared to albumSales.2.”*

While my conclusion resulted in the same outcome as the book, I did make the previous statement that we could make the educated guess that the model albumSales.3 was a better fit. At the time, my thought was that we could perhaps save ourselves the step of doing the anova() test but upon thinking about it further I can say that it would be a mistake to do so. First, it would call into question my entire analysis causing questions about where else I may have cut corners, secondly imagine if I was looking at this later and decided to run the analysis and found that I was originally wrong? How long have we been using the wrong model? and How much could that have cost us? Moving forward I would do well to remember this.