SSC 442 Class Ex 1

Team Awesome (19) - Sayem Lincoln, Joshua Schwimmer, John Townshend.

1/27/2020

# Initial regression model

summary(lm(balance~ age+day+duration+campaign+previous, data = data))

##   
## Call:  
## lm(formula = balance ~ age + day + duration + campaign + previous,   
## data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -5136 -1307 -912 67 69299   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 517.8536 209.1859 2.476 0.0133 \*   
## age 23.8309 4.2186 5.649 1.71e-08 \*\*\*  
## day -1.7331 5.4878 -0.316 0.7522   
## duration -0.1957 0.1721 -1.137 0.2555   
## campaign -7.9129 14.5894 -0.542 0.5876   
## previous 46.1345 26.4352 1.745 0.0810 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2999 on 4515 degrees of freedom  
## Multiple R-squared: 0.008092, Adjusted R-squared: 0.006993   
## F-statistic: 7.367 on 5 and 4515 DF, p-value: 6.893e-07

# F-test

res.ftest1

##   
## F test to compare two variances  
##   
## data: age by y  
## F = 0.60343, num df = 3999, denom df = 520, p-value = 2.682e-16  
## alternative hypothesis: true ratio of variances is not equal to 1  
## 95 percent confidence interval:  
## 0.5284958 0.6845010  
## sample estimates:  
## ratio of variances   
## 0.603426

The p-value of F-test is p = 2.682e-16 which is lesser than the significance level 0.05. In conclusion, there is a significant difference between the two variances.

res.ftest2

##   
## F test to compare two variances  
##   
## data: balance by y  
## F = 1.5829, num df = 3999, denom df = 520, p-value = 5.734e-11  
## alternative hypothesis: true ratio of variances is not equal to 1  
## 95 percent confidence interval:  
## 1.386316 1.795539  
## sample estimates:  
## ratio of variances   
## 1.582868

The p-value of F-test is p = 5.734e-11 which is greater than the significance level 0.05. In conclusion, there is no significant difference between the two variances.

res.ftest3

##   
## F test to compare two variances  
##   
## data: day by y  
## F = 1.0035, num df = 3999, denom df = 520, p-value = 0.9707  
## alternative hypothesis: true ratio of variances is not equal to 1  
## 95 percent confidence interval:  
## 0.8789309 1.1383800  
## sample estimates:  
## ratio of variances   
## 1.003546

The p-value of F-test is p = 0.9707 which is greater than the significance level 0.05. In conclusion, there is no significant difference between the two variances.

res.ftest4

##   
## F test to compare two variances  
##   
## data: duration by y  
## F = 0.29032, num df = 3999, denom df = 520, p-value < 2.2e-16  
## alternative hypothesis: true ratio of variances is not equal to 1  
## 95 percent confidence interval:  
## 0.2542716 0.3293293  
## sample estimates:  
## ratio of variances   
## 0.2903222

The p-value of F-test is p < 2.2e-16 which is lesser than the significance level 0.05. In conclusion, there is a significant difference between the two variances.

res.ftest5

##   
## F test to compare two variances  
##   
## data: campaign by y  
## F = 2.3581, num df = 3999, denom df = 520, p-value < 2.2e-16  
## alternative hypothesis: true ratio of variances is not equal to 1  
## 95 percent confidence interval:  
## 2.065287 2.674933  
## sample estimates:  
## ratio of variances   
## 2.358104

The p-value of F-test is p < 2.2e-16 which is lesser than the significance level 0.05. In conclusion, there is a significant difference between the two variances.

res.ftest6

##   
## F test to compare two variances  
##   
## data: previous by y  
## F = 0.62689, num df = 3999, denom df = 520, p-value = 4.938e-14  
## alternative hypothesis: true ratio of variances is not equal to 1  
## 95 percent confidence interval:  
## 0.5490491 0.7111214  
## sample estimates:  
## ratio of variances   
## 0.6268934

The p-value of F-test is p = 4.938e-14 which is lesser than the significance level 0.05. In conclusion, there is a significant difference between the two variances.

# Resulting model

summary(lm(balance~day, data = data1))

##   
## Call:  
## lm(formula = balance ~ day, data = data1)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -4758 -1351 -977 62 69734   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1473.051 97.298 15.140 <2e-16 \*\*\*  
## day -3.166 5.428 -0.583 0.56   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3010 on 4519 degrees of freedom  
## Multiple R-squared: 7.529e-05, Adjusted R-squared: -0.000146   
## F-statistic: 0.3403 on 1 and 4519 DF, p-value: 0.5597

The resulting model ouputs the following- Residual standard error: 3010 on 4519 degrees of freedom Multiple R-squared: 7.529e-05, Adjusted R-squared: -0.000146 F-statistic: 0.3403 on 1 and 4519 DF, p-value: 0.5597

Comparing it to the previous model’s output - Residual standard error: 2999 on 4515 degrees of freedom Multiple R-squared: 0.008092, Adjusted R-squared: 0.006993 F-statistic: 7.367 on 5 and 4515 DF, p-value: 6.893e-07

We can see that the F-statistic ,Adjusted R-squared, and p-value went down and the Residual standard error value, Multiple R-squared went up.

As the In general, a F-statistic is a ratio of two quantities that are expected to be roughly equal under the null hypothesis, which produces an F-statistic of approximately 1. So, the new model’s F statistic is closer to 1 thus the new model has a better F statistic and the new model is an improvement on the previous model.