Exercises04

Dennis Perrett

1/14/2022

Exercise 16. Exercise 16. In continuation of the previous exercise, consider the data for predicting the overall clinical impression based on the dose of an antipsychotic X_1 and days since hospitalization X_2 for four patients in a psychiatric hospital. Check the null hypothesis ($\alpha = 0.05$)

$$H_0 : \beta_1 = 8 \text{ and } \beta_2 = 0$$

for the parameter estimates resulting for the regression model

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i. \quad (i = 1, ..., 4)$$

Perform the calculations by hand and using R.

$$F = \frac{\frac{Q}{r}}{\frac{SS_{res}}{n-p-1}} = \frac{\frac{Q}{r}}{\hat{\sigma}^2}$$

where

$$\begin{aligned} r &= rank \\ Q &= ((\beta - \beta_{h0)})' \cdot [c \cdot (X'X)^{-1} \cdot c] \cdot (\beta - \beta_{h0})) \\ n &= \text{number of observations} \\ p &= \text{number of parameters X (being tested or total?)} \end{aligned}$$

Therefore

$$\begin{aligned} Q &= ((\beta - \beta_{h0)})' \cdot [c \cdot (X'X)^{-1} \cdot c]^{-1} \cdot (\beta - \beta_{h0})) = 0.4819277 \\ SS_{res} &= (\hat{y} - y)'(\hat{y} - y) = \\ F &= \frac{\frac{Q}{r}}{\frac{SS_{res}}{n-p-1}} = \frac{\frac{Q}{r}}{\hat{\sigma}^2} = \frac{\frac{0.4819277}{2}}{\frac{29.51807}{4-2-1}} = 0.05502063 < F_{crit} = 199.5 \end{aligned}$$

"By Hand" with R

```
beta_h0 <- matrix(c(8,0),2,1)
c <- matrix(c(0,0,1,0,0,1),2,3)
beta_test = c%*%betas
XX_inv <-solve(t(X)%*%X)
q <- t(beta_test-beta_h0) %*% solve(c%*%XX_inv%*%t(c)) %*% (beta_test-beta_h0)
q
```

```
##           [,1]
## [1,] 0.4819277
```

```
yhat <- X%*%betas
ss.res <-  t(yhat-y)%*%(yhat-y)
f <- (q/2)/(ss.res/4-2-1)
qf(0.95,2,1)
```

```
## [1] 199.5
```

Exercise 17. The data in file orthography.txt are taken from a diagnostic test of spel- ling difficulties in 5th and 6th graders from secondary schools in Baden-Württemberg. It includes information on the following variables:

X_1	*CFT*	Culture Fair Intelligence Test (Subtest 1)
X_2	*WM*	Phonological working memory performance (subtest of VLMT)
X_3	*sex*	Gender (dummy coded: 0 male, 1 female)
X_4	*school*	School type (dummy coded: 0 Hauptschule, 1 Realschule)
X_5	*class*	Grade level (dummy coded: 0 grade 5, 1 grade 6)
Y	*score*	Number of correct spellings for 20 words

Answer the following questions in the context of linear regression analyses performed using R. For statistical tests assume a Type 1 error rate of $\alpha = 0.05$.

a. Do the general cognitive abilities captured by X_1 and X_2 contribute to predicting spelling performance, and how large is the proportion of explained variance?

```
m1 <- lm(score ~ CFT + WM, dt)
summary(m1)
```

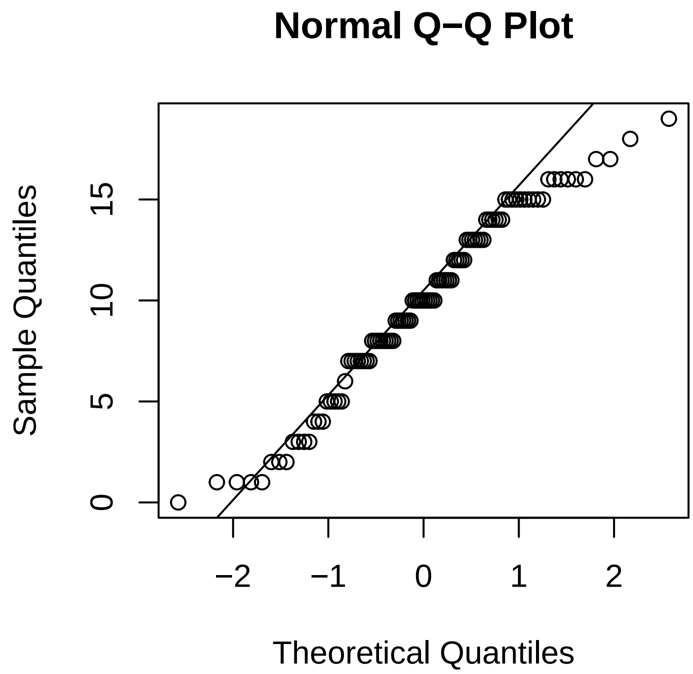
```
##
## Call:
## lm(formula = score ~ CFT + WM, data = dt)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.0929  -3.1309   0.4071   3.2332   9.3202
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.5271     2.0572   0.742 0.459679
## CFT           0.3478     0.1588   2.190 0.030921 *
## WM            0.7175     0.2000   3.587 0.000525 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.247 on 97 degrees of freedom
## Multiple R-squared:  0.1544, Adjusted R-squared:  0.1369
## F-statistic: 8.855 on 2 and 97 DF,  p-value: 0.0002937
```

- Yes, X_1 and X_2 contribute to predicting Y. CFT significant at the 10% level, so for this analysis we would consider this not significant. WM is significant at the 1% level.

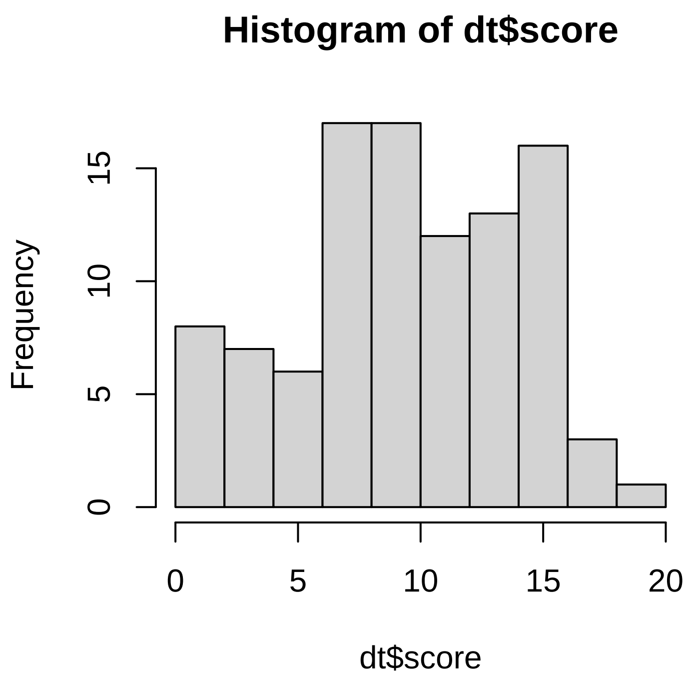
b. Do the data meet the assumptions for conducting the regression analysis and the corresponding statistical tests? Perform appropriate graphical tests.

The assumptions, among others are normally distributed Y values.

```
qqnorm(dt$score)
qqline(dt$score)
```




```
hist(dt$score)
```



Both plots indicate the data could pass as normal. Given that n>30, we can assume normality based on the CLT.

c. In order to decide whether there are gender-related differences, include variable X_3 as an additional predictor into the model.

- Specify the general gender-specific regression equations for male and female students.

$$score_i = \beta_0 + \beta_1 CFT_i + \beta_2 WM_i + \beta_3 sex_i + \epsilon_i$$

- Does gender provide an additional contribution to the prediction of spelling performance? Convince yourself that the test of the regression coefficient β_3 as appearing in the output of summary() gives the same result as computing an incremental F-test (via function anova()) to the previously considered regression model with predictors X_1 and X_2.

```
m2 <- lm(score ~ CFT + WM + sex,dt)
summary(m2)
```

```
##
## Call:
## lm(formula = score ~ CFT + WM + sex, data = dt)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.4695 -2.9641  0.2696  2.8253  8.9874
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.8452     1.9919   0.926  0.35660
## CFT           0.2674     0.1562   1.712  0.09020 .
## WM            0.5854     0.1990   2.941  0.00409 **
## sex           2.4175     0.8655   2.793  0.00630 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.105 on 96 degrees of freedom
## Multiple R-squared:  0.2179, Adjusted R-squared:  0.1935
## F-statistic: 8.917 on 3 and 96 DF,  p-value: 2.86e-05
```

```
anova(m1,m2)
```

```
## Analysis of Variance Table
##
## Model 1: score ~ CFT + WM
## Model 2: score ~ CFT + WM + sex
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
## 1     97 1749.4
## 2     96 1617.9  1    131.48 7.8016 0.006301 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p-value for $sex = 0.00630$. The P-value for the incremental f-test is also 0.00630. We can conclude that gender does contribute to the score of the spelling test.

c. - How is the prediction of the regression model based on the predictors X_1, X_2, and X_3 to be interpreted geometrically?
     - Each variable can be described as a different dimension (or axes). First, the value creeps along the first dimension until it reaches the appropriate X1 value, then it along second axis until it reaches the finaly point in space aka the prediciton.

d. Is it possible to identify the most significant predictor for spelling ability in a model including all the predictors $X_1, ..., X_5$? Evaluate multicollinearity by considering bivariate intercorrelations and variance inflation factors.

$$VIF = \frac{1}{1 - R^2_{model}}$$

where

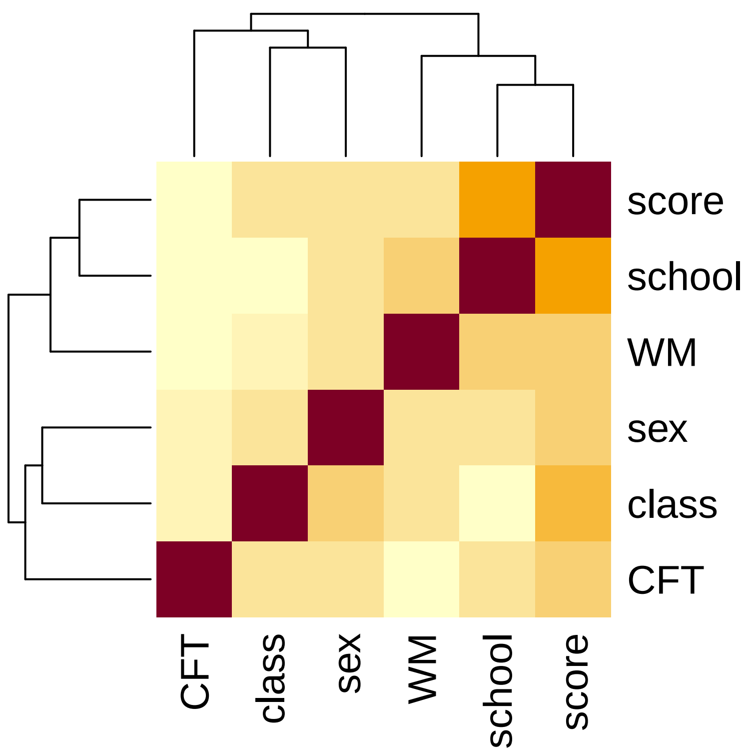
$$model = X_1 = \beta X$$

where X does not include the independent X variable.

```r
m3 <- lm(score ~ CFT + WM + sex + school + class, dt)
summary(m3)
```

```
##
## Call:
## lm(formula = score ~ CFT + WM + sex + school + class, data = dt)
##
## Residuals:
##    Min      1Q Median     3Q    Max
## -8.018 -2.524 -0.035  2.372  8.925
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.9854     1.7261   1.150  0.25297
## CFT           0.1688     0.1364   1.238  0.21894
## WM            0.2710     0.1807   1.500  0.13698
## sex           1.3641     0.7746   1.761  0.08150 .
## school        4.1262     0.7926   5.206 1.13e-06 ***
## class         2.0782     0.7397   2.809  0.00604 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.554 on 94 degrees of freedom
## Multiple R-squared:  0.4259, Adjusted R-squared:  0.3954
## F-statistic: 13.95 on 5 and 94 DF,  p-value: 3.43e-10
```

```r
heatmap(cor(dt))
```




```r
cor(dt)
```

```
##                 CFT          WM       sex     school      class     score
## CFT    1.000000000 0.002747507 0.1799174 0.11805189 0.11967656 0.2053986
## WM     0.002747507 1.000000000 0.2341423 0.31593292 0.15540644 0.3355159
## sex    0.179917372 0.234142295 1.0000000 0.23867369 0.24174689 0.3561080
## school 0.118051894 0.315932917 0.2386737 1.00000000 0.06213698 0.5393987
## class  0.119676562 0.155406438 0.2417469 0.06213698 1.00000000 0.3231940
## score  0.205398636 0.335515885 0.3561080 0.53939873 0.32319403 1.0000000
```

```r
# VIF
1/(1-summary(lm(school ~ WM + sex + class + CFT,data = dt))$r.squared)
```

```
## [1] 1.158918
```

```r
1/(1-summary(lm(WM ~ school + sex + class + CFT,data = dt))$r.squared)
```

```
## [1] 1.164793
```

```r
1/(1-summary(lm(sex ~ WM + school + class + CFT,data = dt))$r.squared)
```

```
## [1] 1.170292
```

```r
1/(1-summary(lm(class ~ WM + sex + school + CFT,data = dt))$r.squared)
```

```
## [1] 1.082812
```

```r
1/(1-summary(lm(CFT ~ WM + sex + class + school,data = dt))$r.squared)
```

```
## [1] 1.052418
```

Both correlations and VIF indicate no multicollinearity. VIFs are all very small < 5. Correlations are almost all below .. School correlations somewhat highly with score, but with no other predictor variables.