STA471 - Exam 2

Richard McCormick

2023-11-09

```r
exam.data <- readxl::read_excel( "exam2data.xlsx")
```

1. In an automobile fuel efficiency study, the following data were collected on a simple random sample of 38 cars. The variables measured are Y = Miles per gallon, X_1 = Weight (1,000lb), X_2 = Engine displacement (cubic inches), X_3 = Number of cylinders, X_4 = Horsepower, X_5 = Acceleration from 0 to 60 mph (sec).

a. (10 points) Fit the model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \epsilon$ and give the fitted equation relating Y to all five predictor variables. Interpret the estimated coefficient of X_3 in the context of the problem.

```r
model <- lm( data=exam.data, Y ~ X1 + X2 + X3 + X4 + X5 )
summary( model )
```

```
##
## Call:
## lm(formula = Y ~ X1 + X2 + X3 + X4 + X5, data = exam.data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.0121 -1.6337 -0.0557  1.3846  5.6134
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  70.37038    4.45295  15.803  < 2e-16 ***
## X1          -10.18787    2.71107  -3.758 0.000688 ***
## X2            0.05717    0.01806   3.165 0.003390 **
## X3           -0.83382    0.72155  -1.156 0.256406
## X4           -0.09648    0.04545  -2.123 0.041624 *
## X5           -0.44969    0.32442  -1.386 0.175288
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.299 on 32 degrees of freedom
## Multiple R-squared:  0.8934, Adjusted R-squared:  0.8767
## F-statistic: 53.62 on 5 and 32 DF,  p-value: 1.284e-14
```

b. (6 points) Find the predicted Y value for a car with $X_1 = 3.00$, $X_2 = 250$, $X_3 = 6$, $X_4 = 125$, $X_5 = 15$, and construct a 99% prediction inverval for the Y value.

```r
prediction <- predict( model,
                       newdata=data.frame( X1=3.00,
                                           X2=250,
                                           X3=6,
                                           X4=125,
                                           X5=15),
                       level=0.99 )

print( paste( "Predicted value for Y is: ", prediction ) )
```

```
## [1] "Predicted value for Y is:  30.2904002680581"
```

```r
pred_interval = predict( model,
                         interval="prediction",
                         level=0.99,
                         newdata=data.frame( X1=3.00,
                                           X2=250,
                                           X3=6,
                                           X4=125,
                                           X5=15)
                         )

pred_interval
```

```
##       fit      lwr      upr
## 1 30.2904 22.29121 38.28959
```

c. (6 points) Find and interpret the value of R^2 for the model that includes all five predictor variables.

```r
summary( model )
```

```
##
## Call:
## lm(formula = Y ~ X1 + X2 + X3 + X4 + X5, data = exam.data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.0121 -1.6337 -0.0557  1.3846  5.6134
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  70.37038    4.45295  15.803  < 2e-16 ***
## X1          -10.18787    2.71107  -3.758 0.000688 ***
## X2            0.05717    0.01806   3.165 0.003390 **
## X3           -0.83382    0.72155  -1.156 0.256406
## X4           -0.09648    0.04545  -2.123 0.041624 *
## X5           -0.44969    0.32442  -1.386 0.175288
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.299 on 32 degrees of freedom
## Multiple R-squared:  0.8934, Adjusted R-squared:  0.8767
## F-statistic: 53.62 on 5 and 32 DF,  p-value: 1.284e-14
```

d. (10 points) How useful is the regression using X_1 alone? What does X_3 contribute, given X_1 and X_2 are already in the regression?

```r
anova( model )
```

```
## Analysis of Variance Table
##
## Response: Y
##           Df  Sum Sq Mean Sq  F value    Pr(>F)
## X1         1 1293.52 1293.52 244.7510 < 2.2e-16 ***
## X2         1   86.94   86.94  16.4510 0.0002992 ***
## X3         1   12.59   12.59   2.3816 0.1326020
## X4         1   13.77   13.77   2.6053 0.1163267
## X5         1   10.15   10.15   1.9214 0.1752882
## Residuals 32  169.12    5.29
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
RegSS <- sum( anova( model )[1:5, 2] )
writeLines( paste( round( ( anova( model )[1,2] / RegSS )*100, 3 ),
            "% of the variation in Y is explained by the regression using X1 alone." ) )
```

```
## 91.287 % of the variation in Y is explained by the regression using X1 alone.
```

```r
RegSS.X3 <- sum( anova( model )[1:2, 2] )
SS.X1 <- RegSS - RegSS.X3

writeLines( paste( round( ( SS.X1/RegSS )*100, 3 ),
              "% of the variation in Y is explained by the regression using X3,\n",
              "given that X1 and X2 are already in the model." ) )
```

```
## 2.577 % of the variation in Y is explained by the regression using X3,
##  given that X1 and X2 are already in the model.
```

e. (10 points) Test to determine whether the overall regression is significant at $\alpha = 0.05$.

```r
overall_p <- function(my_model) {
    f <- summary(my_model)$fstatistic
    p <- pf(f[1],f[2],f[3],lower.tail=F)
    attributes(p) <- NULL
    return(p)
}

summary( model )
```

```
##
## Call:
## lm(formula = Y ~ X1 + X2 + X3 + X4 + X5, data = exam.data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.0121 -1.6337 -0.0557  1.3846  5.6134
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  70.37038    4.45295  15.803  < 2e-16 ***
## X1          -10.18787    2.71107  -3.758 0.000688 ***
## X2            0.05717    0.01806   3.165 0.003390 **
## X3           -0.83382    0.72155  -1.156 0.256406
## X4           -0.09648    0.04545  -2.123 0.041624 *
## X5           -0.44969    0.32442  -1.386 0.175288
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.299 on 32 degrees of freedom
## Multiple R-squared:  0.8934, Adjusted R-squared:  0.8767
## F-statistic: 53.62 on 5 and 32 DF,  p-value: 1.284e-14
```

```r
#extract overall p-value of model
print( paste( "Model p-value is:", overall_p( model ) ) )
```

```
## [1] "Model p-value is: 1.28366664194536e-14"
```

f. (10 points) Test whether there is a linear relationship between X_4 and Y in the model that includes all the other predictor variables. Use $\alpha = 0.05$.

```
anova( model )
```

```
## Analysis of Variance Table
##
## Response: Y
##           Df  Sum Sq Mean Sq  F value    Pr(>F)
## X1         1 1293.52 1293.52 244.7510 < 2.2e-16 ***
## X2         1   86.94   86.94  16.4510 0.0002992 ***
## X3         1   12.59   12.59   2.3816 0.1326020
## X4         1   13.77   13.77   2.6053 0.1163267
## X5         1   10.15   10.15   1.9214 0.1752882
## Residuals 32  169.12    5.29
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

g. (10 points) Construct a 95% confidence interval for β_5 and interpret the confidence interval. What is your conclusion in the context of the problem based on the confidence interval?

```r
# Get the i-1th element
b5.variance <- vcov( model )[6,6]

print( paste( "Variance of b5 =", b5.variance ) )
```

```
## [1] "Variance of b5 = 0.105248723740642"
```

```r
# 95% confidence interval for b5
confint( model, level=0.95 )[6,]
```

```
##      2.5 %     97.5 %
## -1.1105152  0.2111311
```

h. (12 points) Test whether the variables X_2, X_3, and X_5 jointly have a linear relationship with Y in the model that includes all five predictor variables. Use $\alpha = 0.05$.

```r
SSreg.reduced <- sum( anova( model )[2:3, 2] ) + sum( anova( model )[5, 2] )
print( paste( "SSreg for X2, X3, and X5:", SSreg.reduced ) )
```

```
## [1] "SSreg for X2, X3, and X5: 109.685312548133"
```

```r
anova( model )
```

```
## Analysis of Variance Table
##
## Response: Y
##           Df  Sum Sq Mean Sq  F value    Pr(>F)
## X1         1 1293.52 1293.52 244.7510 < 2.2e-16 ***
## X2         1   86.94   86.94  16.4510 0.0002992 ***
## X3         1   12.59   12.59   2.3816 0.1326020
## X4         1   13.77   13.77   2.6053 0.1163267
## X5         1   10.15   10.15   1.9214 0.1752882
## Residuals 32  169.12    5.29
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
SSreg <- sum( anova( model )[1:5, 2] )
print( paste( "SSreg for full model:", SSreg ) )
```

```
## [1] "SSreg for full model: 1416.96995205853"
```

```r
SSresid <- sum( anova( model )[6, 2] )

f.val <- ((SSreg - SSreg.reduced)/2) / ( ( SSresid )/4 )
print( paste( "Observed F-value:", f.val ) )
```

```
## [1] "Observed F-value: 15.4597701796056"
```

i. (10 points) Use the backward elimination procedure to find an appropriate model for the data at $\alpha = 0.05$. What is the fitted equation for the model?

```r
summary( model )
```

```
##
## Call:
## lm(formula = Y ~ X1 + X2 + X3 + X4 + X5, data = exam.data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.0121 -1.6337 -0.0557  1.3846  5.6134
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  70.37038    4.45295  15.803  < 2e-16 ***
## X1          -10.18787    2.71107  -3.758 0.000688 ***
## X2            0.05717    0.01806   3.165 0.003390 **
## X3           -0.83382    0.72155  -1.156 0.256406
## X4           -0.09648    0.04545  -2.123 0.041624 *
## X5           -0.44969    0.32442  -1.386 0.175288
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.299 on 32 degrees of freedom
## Multiple R-squared:  0.8934, Adjusted R-squared:  0.8767
## F-statistic: 53.62 on 5 and 32 DF,  p-value: 1.284e-14
```

```r
model.2 <- lm( data=exam.data, Y ~ X1 + X2 + X4 + X5 )
summary( model.2 )
```

```
##
## Call:
## lm(formula = Y ~ X1 + X2 + X4 + X5, data = exam.data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.7781 -1.7242 -0.1874  1.0363  5.4468
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  69.34194    4.38522  15.813  < 2e-16 ***
## X1          -10.26937    2.72389  -3.770 0.000643 ***
## X2            0.04596    0.01532   3.001 0.005091 **
## X4           -0.10564    0.04499  -2.348 0.025008 *
## X5           -0.47092    0.32554  -1.447 0.157446
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.311 on 33 degrees of freedom
## Multiple R-squared:  0.8889, Adjusted R-squared:  0.8755
## F-statistic: 66.02 on 4 and 33 DF,  p-value: 2.804e-15
```

```r
model.3 <- lm( data=exam.data, Y ~ X1 + X2 + X4 )
summary( model.3 )
```

```
##
## Call:
## lm(formula = Y ~ X1 + X2 + X4, data = exam.data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.7052 -1.6079 -0.1802  1.2018  5.8293
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  64.44543    2.83242  22.753  < 2e-16 ***
## X1          -12.72155    2.16613  -5.873 1.26e-06 ***
## X2            0.05560    0.01401   3.968 0.000355 ***
## X4           -0.06672    0.03663  -1.821 0.077349 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.347 on 34 degrees of freedom
## Multiple R-squared:  0.8819, Adjusted R-squared:  0.8715
## F-statistic: 84.61 on 3 and 34 DF,  p-value: 7.602e-16
```

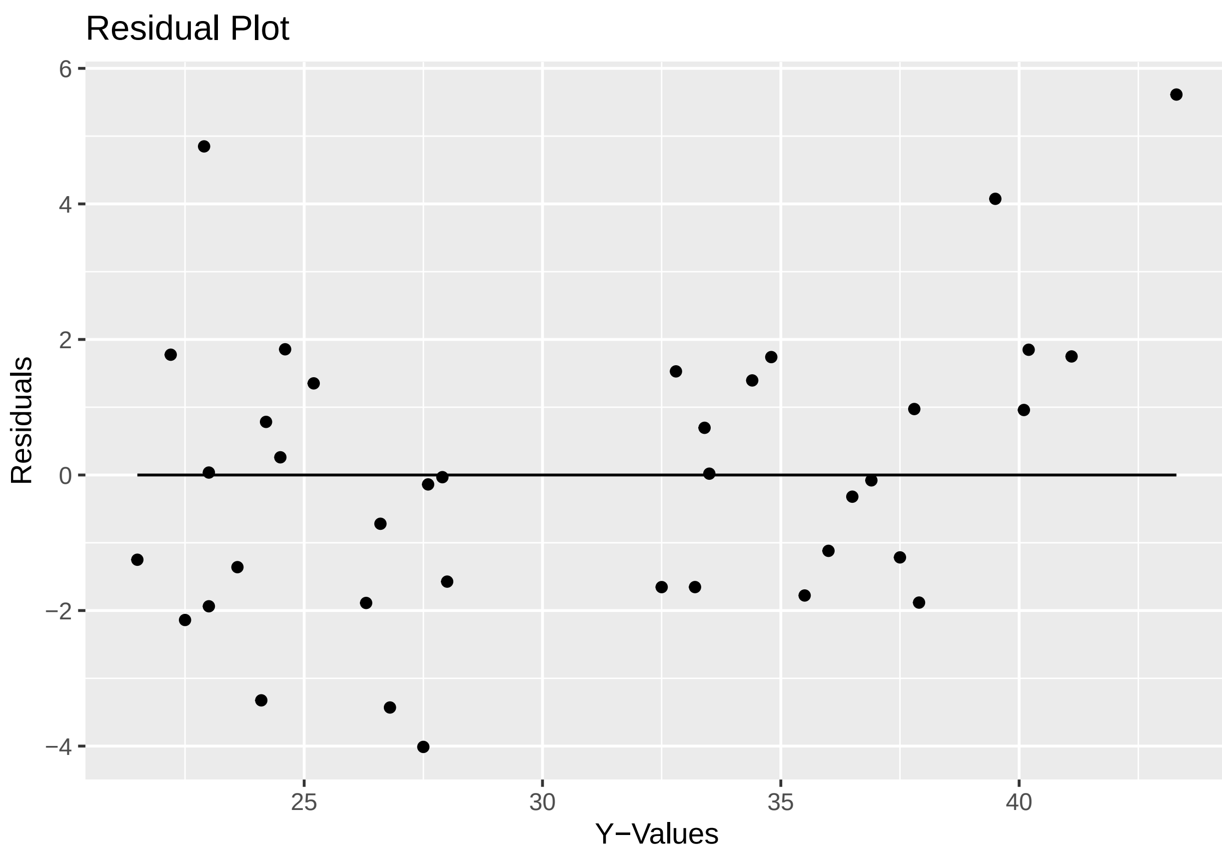
```r
model.4 <- lm( data=exam.data, Y ~ X1 + X2 )
summary( model.4 )
```

```
##
## Call:
## lm(formula = Y ~ X1 + X2, data = exam.data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.0759 -1.6857 -0.2239  0.6119  6.3355
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  63.88733    2.90753  21.973  < 2e-16 ***
## X1          -15.01750    1.81899  -8.256  9.9e-10 ***
## X2            0.05565    0.01447   3.847 0.000485 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.424 on 35 degrees of freedom
## Multiple R-squared:  0.8704, Adjusted R-squared:  0.8629
## F-statistic: 117.5 on 2 and 35 DF,  p-value: 2.974e-16
```

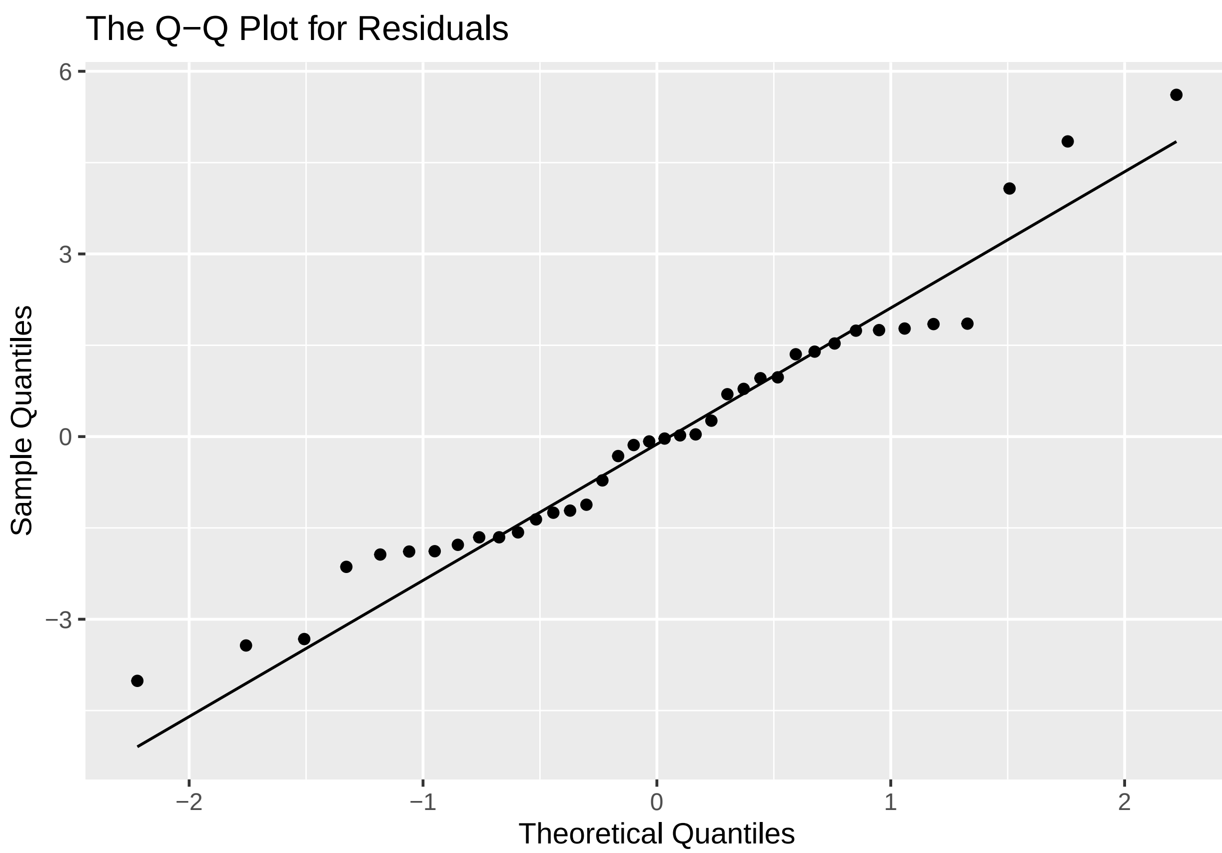
j. (16 points) What are the assumptions for the model? Check the assumptions by a residual plot and a $Q-Q$ plot of the residuals. In addition, conduct Shapiro and Wilk test for normality based on residuals.

```
exam.data$resid <- resid(model)

ggplot( data=exam.data, aes( x=Y, y=resid ) ) +
  geom_point() +
  geom_line( aes( y=0 ) ) +
  labs( title="Residual Plot", y="Residuals", x="Y-Values")
```




```
ggplot( data = exam.data, aes( sample=resid ) ) +
  stat_qq() +
  geom_qq_line() +
  labs( title="The Q-Q Plot for Residuals", x="Theoretical Quantiles",
        y="Sample Quantiles" )
```




```r
shapiro.test( resid( model ) )
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resid(model)
## W = 0.95876, p-value = 0.173
```