Math 4780 - Homework 2

Liam Fruzyna

#3.1 Consider the NFL data in table B.1

Load in dataset

```
library(MPV)
nfl = table.b1
```

a) Fit a multiple linear regression model relating y to x2, x7, and x8

```
model <- lm(y ~ x2 + x7 + x8, data=nfl)
summary <- summary(model)
model$coefficients
```

```
##  (Intercept)           x2           x7           x8
## -1.808372059  0.003598070  0.193960210 -0.004815494
```

$\hat{y} = -1.8 + 0.0036x_2 + 0.196x_7 - 0.0048x_8$

b) Construct the anova table and test for significance

```
anova(model)
```

```
## Analysis of Variance Table
##
## Response: y
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## x2         1  76.193  76.193  26.172 3.100e-05 ***
## x7         1 139.501 139.501  47.918 3.698e-07 ***
## x8         1  41.400  41.400  14.221 0.0009378 ***
## Residuals 24  69.870   2.911
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

All 3 variables are significant, their p-values are very small.

c) Calculate t statistics for testing the hypotheses

```
summary
```

```
##
## Call:
## lm(formula = y ~ x2 + x7 + x8, data = nfl)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.0370 -0.7129 -0.2043  1.1101  3.7049
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.808372   7.900859  -0.229 0.820899
## x2           0.003598   0.000695   5.177 2.66e-05 ***
```

```
## x7           0.193960   0.088233   2.198 0.037815 *
## x8          -0.004816   0.001277  -3.771 0.000938 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.706 on 24 degrees of freedom
## Multiple R-squared:  0.7863, Adjusted R-squared:  0.7596
## F-statistic: 29.44 on 3 and 24 DF,  p-value: 3.273e-08
```

```
Coefficient  T-Val P-Value
x2           5.177 2.66e-05
x7           2.198 0.037815
x8          -3.771 0.000938
```

d) Calculate R^2 and adjusted R^2 values

```r
summary$r.squared
```

```
## [1] 0.7863069
```

```r
summary$adj.r.squared
```

```
## [1] 0.7595953
```

$R^2 = 79\%$ and Adjusted $R^2 = 76\%$

e) Using the partial F test determine the contribution of x7 to the model

```r
reduced <- lm(y ~ x2 + x8, data=nfl)
anova(reduced, model)
```

```
## Analysis of Variance Table
##
## Model 1: y ~ x2 + x8
## Model 2: y ~ x2 + x7 + x8
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1     25 83.938
## 2     24 69.870  1    14.068 4.8324 0.03782 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

F-Value is 4.83, it is significant at 95% bounds. This test statistic is the square of the t-statistic.

#3.3 Using 3.1

a) 95% CI on B7

```r
confint(model, 'x7', level=0.95)
```

```
##         2.5 %    97.5 %
## x7 0.01185532 0.3760651
```

The 95% confidence interval of β_7 is 0.012 to 0.376

b) 95% CI on mean number of games won when x2=2300 x7=56 and x8=2100

```
predict(model, data.frame(x2=2300, x7=56, x8=2100), interval='confidence')
```

```
##        fit      lwr      upr
## 1 7.216424 6.436203 7.996645
```

The 95% confidence interval of y is 6.436 to 7.997

#3.9 Consider the data in table B.6

Load in dataset

```
library(MPV)
NbOCl <- table.b6
```

a) Fit a multiple linear regression model relating y to x1 and x4

```
model <- lm(y ~ x1 + x4, data=NbOCl)
summary <- summary(model)
model$coefficients
```

```
##   (Intercept)            x1            x4
##  0.0048332893 -0.3449837404 -0.0001430047
```

$\hat{y} = 0.005 - 0.345x_1 - 0.0001x_4$

b) Test for significance of regression

```
anova <- anova(model)
anova
```

```
## Analysis of Variance Table
##
## Response: y
##           Df     Sum Sq    Mean Sq F value   Pr(>F)
## x1         1 1.6615e-05 1.6615e-05 49.3177 2.32e-07 ***
## x4         1 1.0000e-10 1.0000e-10  0.0003   0.9855
## Residuals 25 8.4222e-06 3.3690e-07
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
mean(anova$`F value`[1:2])
```

```
## [1] 24.65903
```

The F value is 24.66 which is significant to 95%

c) Calculate R^2 and adjusted R^2

```
summary$r.squared
```

```
## [1] 0.663608
```

```
summary$adj.r.squared
```

```
## [1] 0.6366966
```

$R^2 = 66\%$ and Adjusted $R^2 = 64\%$

d) Determine the contribution of x1 and x4 to the model with t-tests. Are they both necessary?

```r
summary
```

```
##
## Call:
## lm(formula = y ~ x1 + x4, data = NbOCl)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -0.0009015 -0.0003526 -0.0001538  0.0003847  0.0010874
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.0048333  0.0008142   5.936 3.39e-06 ***
## x1          -0.3449837  0.0673963  -5.119 2.74e-05 ***
## x4          -0.0001430  0.0078151  -0.018    0.986
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0005804 on 25 degrees of freedom
## Multiple R-squared:  0.6636, Adjusted R-squared:  0.6367
## F-statistic: 24.66 on 2 and 25 DF,  p-value: 1.218e-06
```

x_1 is definitely significant at its p-value is very low (significant to 99.9%), however, x_4 is not significant as its p-value is very high.

e) Is multicollinearity a concern?

```r
library(car)
```

```
## Loading required package: carData
```

```r
vif(model)
```

```
##       x1       x4
## 1.891525 1.891525
```

No multicollinearity is not a concern because VIF is low for both variables.

#9 Prove H and I-H are idempotent

$HH = X(X^TX)^{-1}X^TX(X^TX)^{-1}X^T$

$HH = X(X^TX)^{-1}X^T$

$HH = H$

$(I-H)(I-H) = I - H - H + HH$

$(I-H)(I-H) = I - H - H + H$

$(I-H)(I-H) = (I-H)$