HW3

Mohammad Imtiaz Nur

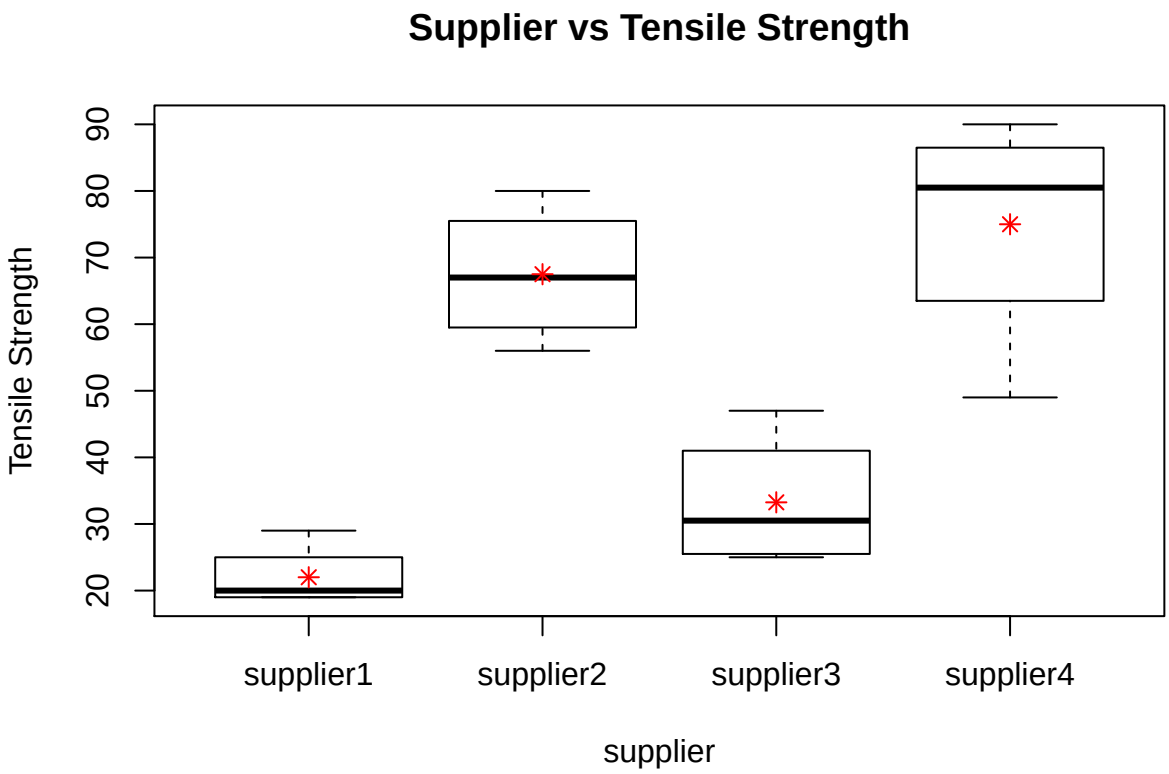
4/2/2021

TASK 1

```
data1 <- data.frame(supplier = rep(c('supplier1','supplier2','supplier3', 'supplier4'),each=4),
                    ts=c( c(19, 21, 19, 29), c(80, 71, 63, 56),
                          c(47, 26, 25, 35), c(90, 49, 83, 78)))


supplierFactor <- factor(data1$supplier)
analysis <- aov(data1$ts ~ supplierFactor)
summary(analysis)
```

```
##                Df Sum Sq Mean Sq F value  Pr(>F)
## supplierFactor  3   7978  2659.4   19.04 7.4e-05 ***
## Residuals      12   1676   139.6
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
par(mfrow=c(1,1))
boxplot(data1$ts~supplierFactor,xlab="supplier",
        ylab="Tensile Strength", main="Supplier vs Tensile Strength")
means <- tapply(data1$ts, supplierFactor, mean)
points(means,pch =8 ,col="red")
```



1(a): As the p-value is much less than the significance level 0.05, we can conclude that there are significant differences between the groups highlighted with "*" in the model summary.

```
TukeyHSD(analysis)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = data1$ts ~ supplierFactor)
##
## $supplierFactor
##                         diff       lwr      upr     p adj
## supplier2-supplier1  45.50  20.69183 70.30817 0.0007410
## supplier3-supplier1  11.25 -13.55817 36.05817 0.5532635
## supplier4-supplier1  53.00  28.19183 77.80817 0.0001880
## supplier3-supplier2 -34.25 -59.05817 -9.44183 0.0069607
## supplier4-supplier2   7.50 -17.30817 32.30817 0.8063290
## supplier4-supplier3  41.75  16.94183 66.55817 0.0015286
```

TASK 2

```
branch = c("branch1","branch1","branch1","branch1","branch2","branch2","branch2","branch3","branch3","b
sl = c(15, 20, 19, 14, 11, 15, 11, 18, 19, 23)
```

```r
data2 <- data.frame(branch, sl)

analysis <- aov(data2$sl ~ branch)
summary(analysis)
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## branch       2  89.83   44.92   6.206 0.0282 *
## Residuals    7  50.67    7.24
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

2(a): As the p-value is less than the significance level 0.05, we can conclude that there are significant differences between the groups.

```r
library("DescTools")
duncan_test <- PostHocTest(analysis, method = "duncan")
duncan_test
```

```
##
##   Posthoc multiple comparisons of means : Duncan's new multiple range test
##     95% family-wise confidence level
##
## $branch
##                       diff    lwr.ci     upr.ci   pval
## branch2-branch1 -4.666667 -9.525508  0.1921743 0.0574 .
## branch3-branch1  3.000000 -1.858841  7.8588409 0.1877
## branch3-branch2  7.666667  2.265647 13.0676868 0.0121 *
##
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

2(b): From the Duncan-test, we can see branch3-branch2 differ significantly at 0.05 confidence level.

Task 3

(a) The negative slope implies that, there is a negative correlation between the time spent in reading the description and the response time. Hence, with the increase of time spent in reading the description, the response time decreases.

(b)

Calculating the confidence interval:

ME = Standard Error * Critical Value

where, ME = Maximum Error; SE = Standard Error

and

Critical value is considered for (1 - alpha/ 2) and degree of freedom 18; Or the region without the tails SE = beta/t_statistics; Considering beta_0 is 0.

```r
DF <- 18
beta <- -0.03
t_stat <- -2.11
y.est <- 0 + (-0.03)*20
SE <- beta/t_stat
alpha <- 0.05
p_star <- 1 - alpha/2
critical_value <- qt(p_star,18)
ME <- SE * critical_value

print("Deviation From Beta")
```

```
## [1] "Deviation From Beta"
```

```r
ME
```

```
## [1] 0.02987093
```

```r
print("Lower interval for Beta")
```

```
## [1] "Lower interval for Beta"
```

```r
beta-ME
```

```
## [1] -0.05987093
```

```r
print("Upper interval for Beta")
```

```
## [1] "Upper interval for Beta"
```

```r
beta+ME
```

```
## [1] -0.0001290705
```

```r
y.est <- beta*20

print("Lower interval for 20")
```

```
## [1] "Lower interval for 20"
```

```r
y.est-ME
```

```
## [1] -0.6298709
```

```r
print("Prediction value for 20")
```

```
## [1] "Prediction value for 20"
```

```
y.est
```

```
## [1] -0.6
```

```
print("Upper interval for 20")
```

```
## [1] "Upper interval for 20"
```

```
y.est+ME
```

```
## [1] -0.5701291
```

#(c)

r^2 = t_stat ^ 2 / t_stat ^ 2 + Degree of freedom

```
r_squared = t_stat*t_stat/(t_stat*t_stat+18)
r_squared
```

```
## [1] 0.1982933
```

The r_squared value is 0.1983 which is only 20 percent variation of the data that is captured by our model.

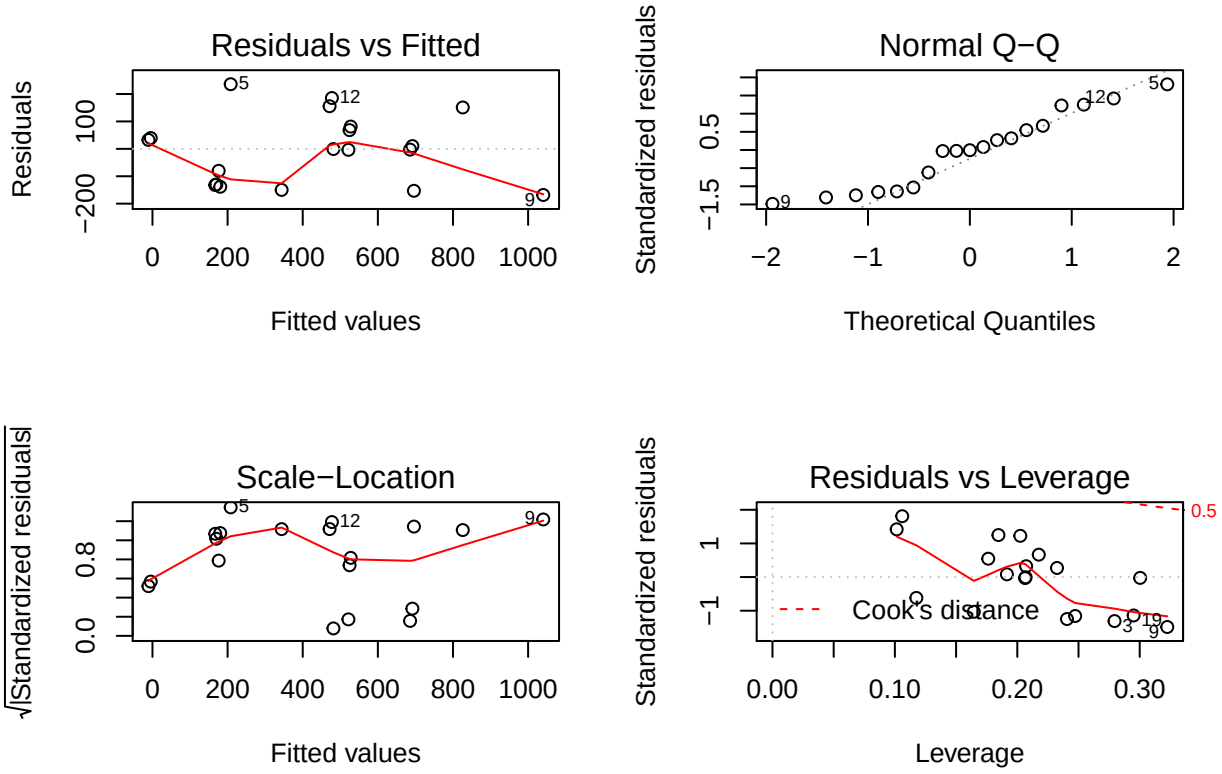
Task 4

```
percent_binder = c(5.3, 5.3, 5.3, 6.0, 7.8, 8.0, 8.0, 8.0, 8.0, 8.0, 8.0, 8.0, 8.0, 10.0, 12.0, 12.0, 1
loading_rate = c(0.02, 0.02, 0.02, 2.00, 0.20, 2.0, 2.0, 2.0, 2.0, 0.02, 0.02, 0.02, 0.02, 2.0, 0.02, 0
ambient_temperature = c(77, 32, 0, 77, 77, 104, 77, 32, 0, 104, 77, 32, 0, 77, 77, 32, 0, 104, 77)
stress  = c(42, 481, 543, 609, 444, 194, 593, 977, 872, 35, 96, 663, 702, 518, 40, 627, 683, 22, 35)
strain  = c(3.20, 0.73, 0.16, 1.44, 3.68, 3.11, 3.07, 0.19, 0.00, 5.86, 5.97, 0.29, 0.04, 2.72, 7.35, 1

df<- data.frame(percent_binder, loading_rate, ambient_temperature, stress, strain)

stress_model <- lm(stress~percent_binder+loading_rate+ambient_temperature, data = df)
summary(stress_model)
```

```
##
## Call:
## lm(formula = stress ~ percent_binder + loading_rate + ambient_temperature,
##     data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -168.380 -131.124   -0.743   74.773  235.765
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)         700.6180   125.8722   5.566 5.40e-05 ***
```

```
## percent_binder       -1.5257    13.0242  -0.117 0.908302
## loading_rate        175.9839    35.6550   4.936 0.000179 ***
## ambient_temperature  -6.6971     0.8847  -7.570 1.69e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 137.9 on 15 degrees of freedom
## Multiple R-squared:  0.8376, Adjusted R-squared:  0.8051
## F-statistic: 25.79 on 3 and 15 DF,  p-value: 3.599e-06
```

```
par(mfrow=c(2, 2))
plot(stress_model)
```



In the "Residual vs Fitted" plot, there is non-linearity in the data, as the red line has irregular curve and the residual samples are not equally distributed two sides of the red line.

In the "Normal Q-Q" plot, residuals deviation from normality throughout the sample.

In the "Scale-Location" plot, there is an irragular trend as the red line is not flat which implies the errors are non-constant and proves the presence of heteroscedasticity.

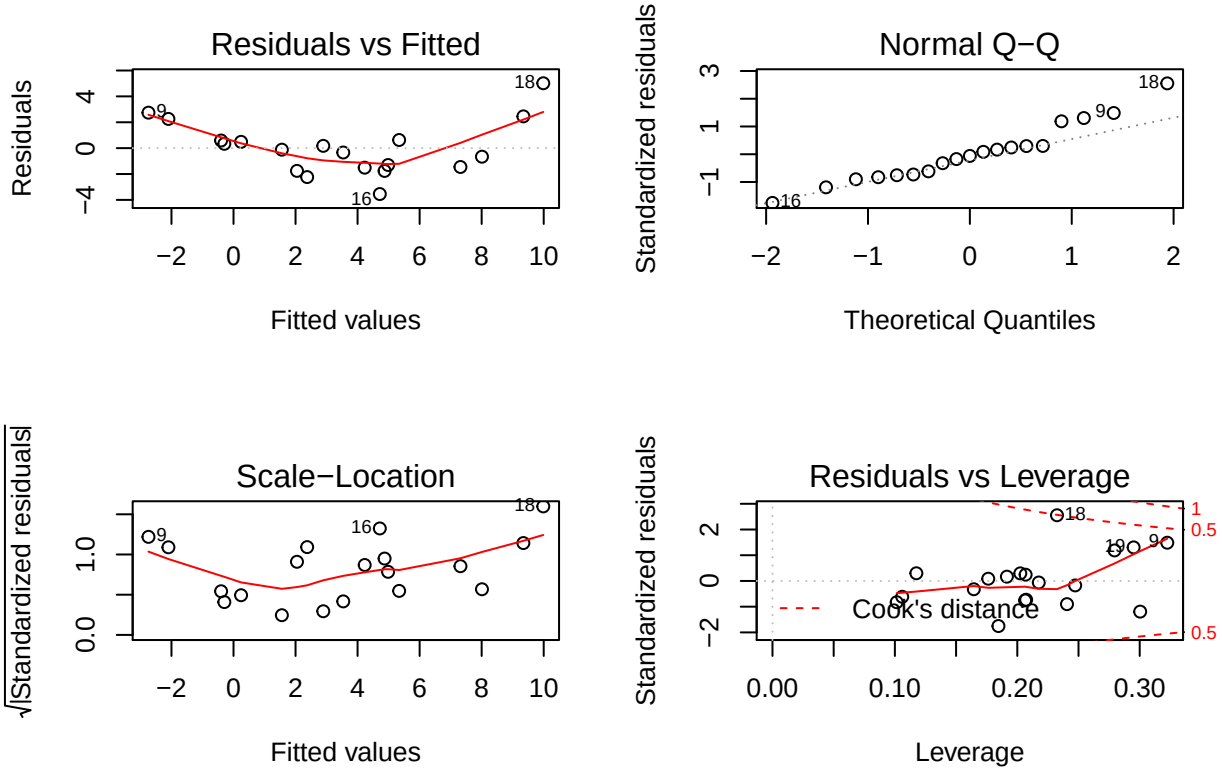
In the "Residuals vs Leverage" plot, we can see most points are clustered on a high leverage point. Though there are some outliers, the outliers are not that influential.

Here, H0 = Model with no independent variables fits the data as well as our model, H1 = Model fits the data better than the intercept-only model.

As the p-value is less than 0.05, we can reject the null hypothesis. Our regression model fits the data better than the model with no independent variables.

From the t- statistics we can see loading_rate and ambient_temperature is significant for the stress, but percent_binder is non-significant.

```
##
## Call:
## lm(formula = strain ~ percent_binder + loading_rate + ambient_temperature,
##     data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5466 -1.4827 -0.1190  0.6097  5.0135
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)         -5.61130    2.04575  -2.743 0.015100 *
## percent_binder       0.66754    0.21168   3.154 0.006558 **
## loading_rate        -1.23535    0.57949  -2.132 0.049966 *
## ambient_temperature  0.07319    0.01438   5.090 0.000133 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.241 on 15 degrees of freedom
## Multiple R-squared:  0.7601, Adjusted R-squared:  0.7121
## F-statistic: 15.84 on 3 and 15 DF,  p-value: 6.447e-05
```



In the "Residual vs Fitted" plot, there is non-linearity in the data, as the red line has irregular curve and

the residual samples are not equally distributed two sides of the red line.

In the "Normal Q-Q" plot, residuals deviation from normality throughout the sample.

In the "Scale-Location" plot, there is an irragular trend as the red line is not flat which implies the errors are non-constant and proves the presence of heteroscedasticity.

In the "Residuals vs Leverage" plot, we can see most points are clustered on a high leverage point. Though there are some outliers, the outliers are not that influential.

Here, H0 = Model with no independent variables fits the data as well as our model, H1 = Model fits the data better than the intercept-only model.

As the p-value is less than 0.05, we can reject the null hypothesis. Our regression model fits the data better than the model with no independent variables.

From the t- statistics we can see all three variables are significant for the strain.

Task 5

Lets assume our first model where m = 2 is the restricted version of our second model. So our unrestricted model has five variables and in the restricted version we have restricted three burnout variables. Apparently our unrestricted model has better R squared value, so we can say it predicts better. Now, to do the significance test, I have calculated F statistics for the Restricted vs Unrestricted model. The equation is:

F = ((URSS - RRSS)/(1-URSS))*((N-K)/q)

Where,

URSS = Unrestricted Residual Sum of Squares

RRSS = Restricted Residual Sum of Squares

N = Sample size

K = The number of parameters estimated in the unrestricted model

q = The number of restrictions imposed

```
URSS = 0.34
RRSS = 0.05
N = 220
K = 5
q = 3

F = ((URSS - RRSS)/(1-URSS))*((N-K)/q)

F
```

```
## [1] 31.4899
```

```
p = 0.05

# number of x variables in the model
df1 = 2

# sample size - number of x variables - 1
df2 = N - df1 - 1
```

```
# Critical value from table for p, df1 and df2
critical_value = 3.307
```

Here, we found F value=31.4899. But for our two variables from the model and alpha 0.05, our F value need to be greater than critical value (3.307). So, we can reject the null hypothesis, that is the regression component of burnout variables are zero. So, the alternate hypothesis is true, that is, they are not zero and we can say at least one of the burnout scores is related to psychosomatic complaints.