Business Analytics (110-1)

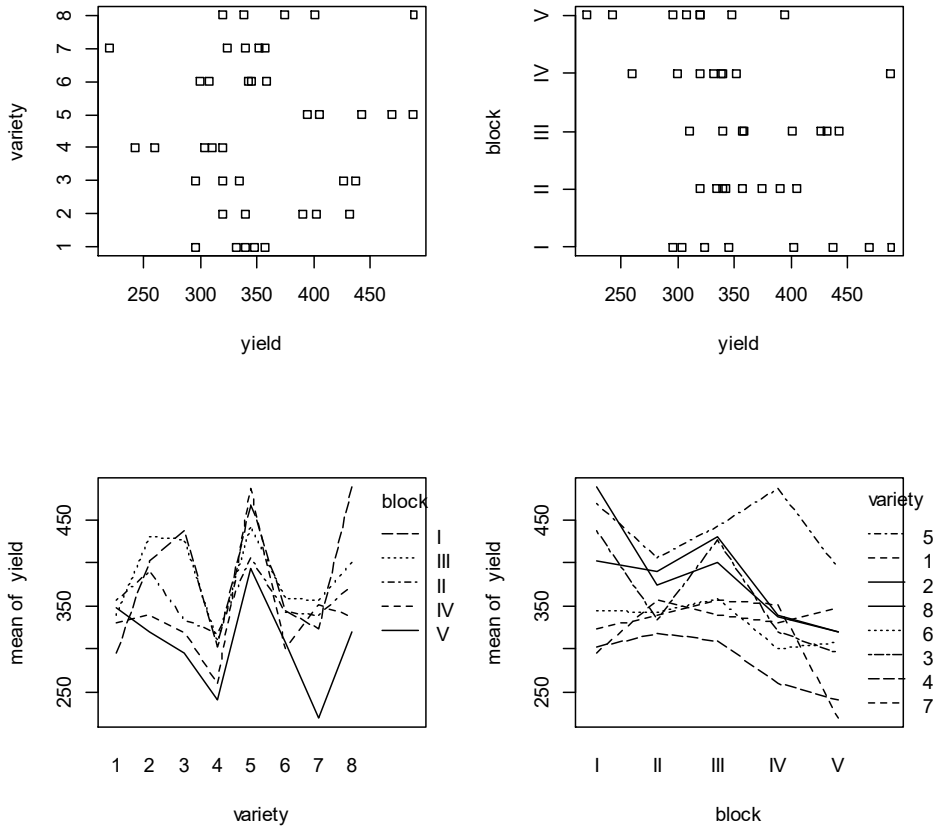
Assignment 3 – Reference Solutions

1.

(a) Randomized Block Design

(b)

```
oatvar <- read.table("oatvar.txt", header=T, sep="\t")
attach(oatvar)
xtabs(yield ~ variety + block)
par(mfrow=c(2,2))
stripchart(yield ~ variety, xlab="yield", ylab="variety")
stripchart(yield ~ block, xlab="yield", ylab="block")
interaction.plot(variety, block, yield)
interaction.plot(block, variety, yield)
```



From the plots above, interaction effects between the variety of oats and the growing area block need to be taken into account.

(c)

```
oatvar$variety <- as.factor(oatvar$variety)
ot <- lm(yield ~ block + variety, oatvar)
summary(ot); anova(ot)
```

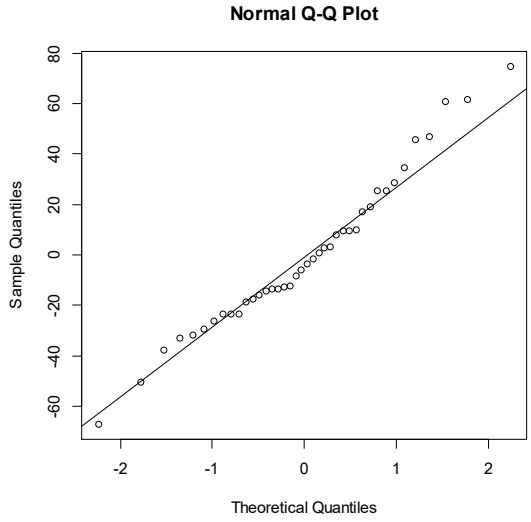
```
Analysis of Variance Table
Response: yield
          Df  Sum Sq  Mean Sq  F value  Pr(>F)
block      4   33396    8349    6.2449   0.001008 **
variety    7   77524    11075    8.2839  1.804e-05 ***
Residuals 28  37433    1337
```

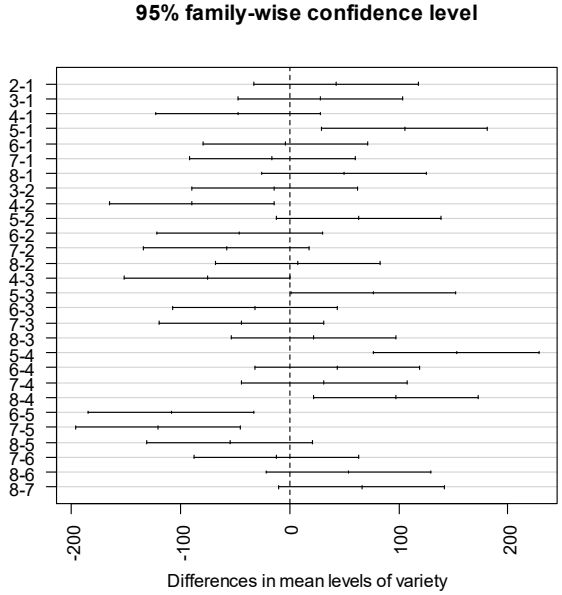
H_0: There is no difference in population mean yield of oats based on varieties
P-value is `1.804e-05` < 0.05. The data suggests to reject H_0.
We conclude that yield of oats is affected by different varieties. Further details are provided by model summary `summary(ot)`.

(d)

```
plot(fitted(ot),residuals(ot),xlab="Fitted",ylab="Residuals")
abline(h=0)
qqnorm(residuals(ot));  qqline(residuals(ot))
```

By and large, the QQ plot looks fine.



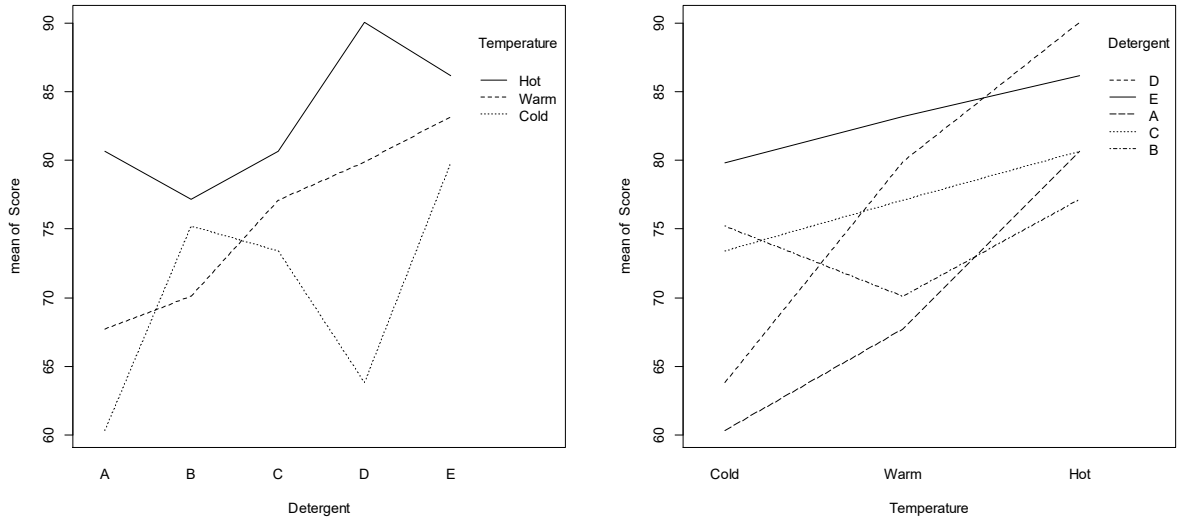


(e)

```
othsd <- TukeyHSD(aov(yield ~ block+variety, oatvar), "variety")
plot(othsd, las=2)
```

2.

```
detergent <- read.table("detergent.txt", header=TRUE, sep="\t")
attach(detergent)
interaction.plot(Detergent, Temperature, Score)
interaction.plot(Temperature, Detergent, Score)
```



From the plots above, interaction effects between the detergent and the water temperature need to be considered.

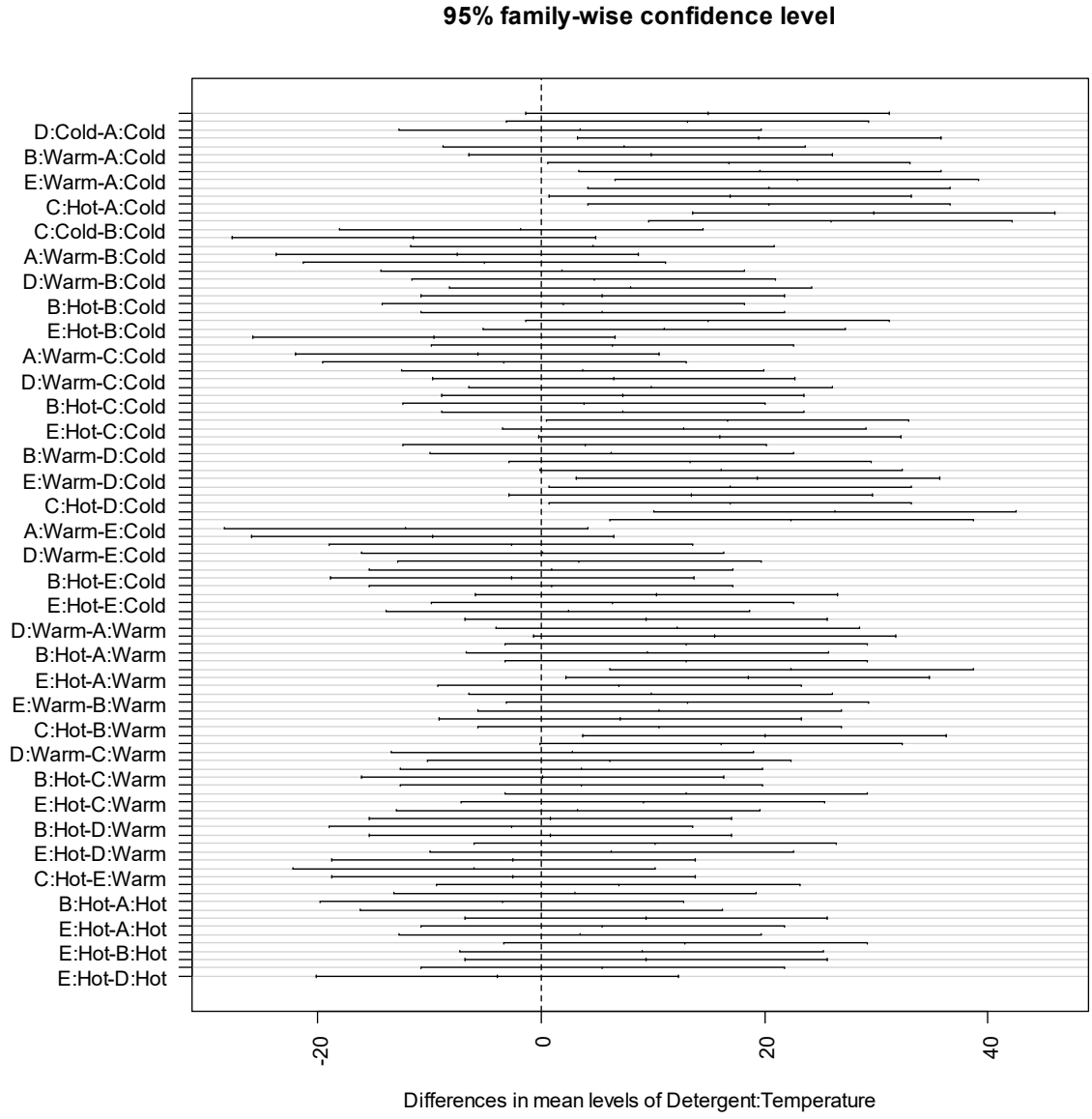
```
dt.m0 <- lm(Score ~ Detergent + Temperature)
summary(dt.m0);     anova(dt.m0)
dt.m1 <- lm(Score ~ Detergent * Temperature)
summary(dt.m1);     anova(dt.m1)

anova(dt.m0, dt.m1)

Analysis of Variance Table
Model 1: Score ~ Detergent + Temperature
Model 2: Score ~ Detergent * Temperature
  Res.Df    RSS  Df    Sum of Sq        F    Pr(>F)
1    143  17362
2    135  14910   8       2452.1    2.7752   0.00714 **
```

The full model is preferred, implying that the interaction effect is statistically significant.

```
par(fig=c(0.1,1,0,1))
plot( TukeyHSD( aov(Score ~ Detergent * Temperature), "Detergent:Temperature" ),
las=2 )
```



The Tukey's HSD test also reveals clearly that the interaction between Temperature and Detergent is significant as many of the 95% CIs do not include 0.

***Question**: How many pairs of comparisons are there shown in this TukeyHSD plot?*

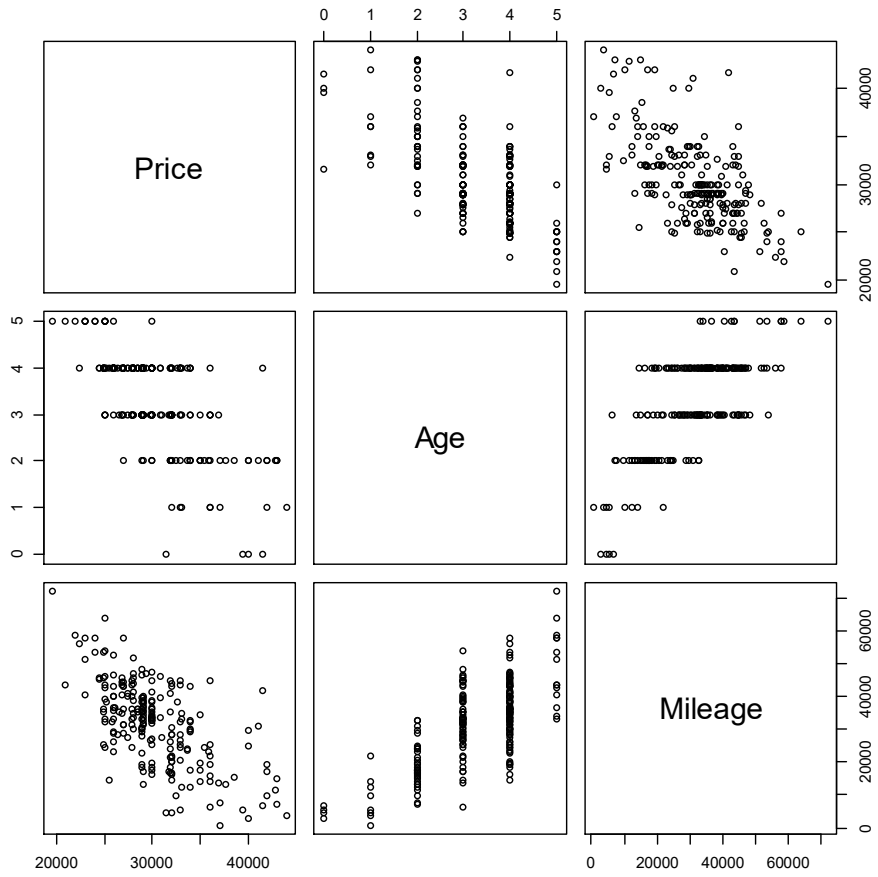
```
plot(fitted(dt.m1), residuals(dt.m1), xlab="Fitted", ylab="Residuals")
abline(h=0, lty=2)
qqnorm(residuals(dt.m1));   qqline(residuals(dt.m1))
```

Residual diagnostic plots indicate that the four assumptions for multiple linear regression are all held.

3.

(a)

```
bmw <- read.table("used_bmw.txt", header=TRUE, sep="\t")
pairs(bmw[,c(1:3)])
```



The plots appear straight enough, and we can see the collinearity between the two proposed explanatory variables. A few outliers appear in the plots, but none of them seem extreme.

(b)

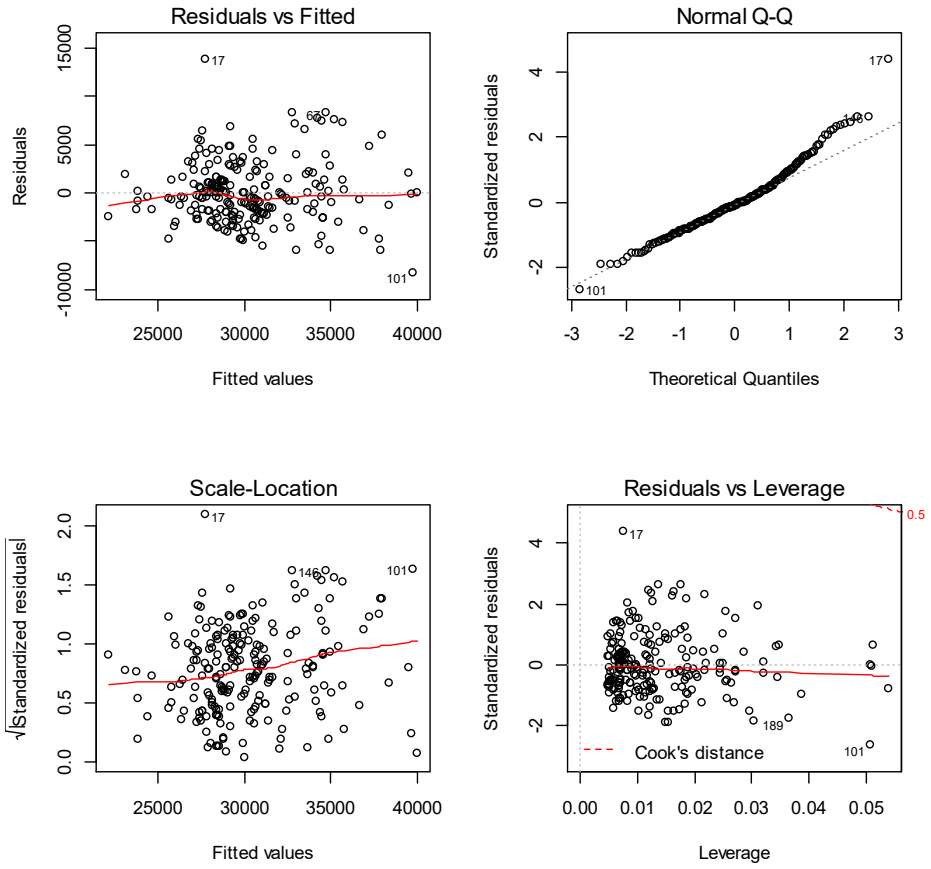
```
bmw.resprice.1  <-  lm(Price ~ Age + Mileage, data = bmw)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.032e+04  7.218e+02  55.862  < 2e-16 ***
Age         -1.854e+03  2.889e+02  -6.417 8.72e-10 ***
Mileage     -1.240e-01  2.375e-02  -5.222 4.17e-07 ***
---
Residual standard error: 3179 on 215 degrees of freedom
Multiple R-squared:  0.5104,    Adjusted R-squared:  0.5058
F-statistic: 112.1 on 2 and 215 DF,  p-value: < 2.2e-16
```

```
bmw.resprice.2  <-  lm(Price ~ Age * Mileage, data = bmw)
anova(bmw.resprice.1, bmw.resprice.2)
Analysis of Variance Table

Model 1: Price ~ Age + Mileage
Model 2: Price ~ Age * Mileage
  Res.Df        RSS   Df   Sum of Sq      F      Pr(>F)
1    215  2172633226
2    214  2170336629    1    2296597  0.2264    0.6347
```

The reduced model is preferred.

The residuals have similar variances and are nearly normal. There is one unusually expensive care, # 17, $ 13,800, but otherwise nothing stands out as particularly troublesome.



(c)

```
confint(bmw.resprice.1)
                  2.5 %       97.5 %
(Intercept) 38901.1326609  4.174674e+04
Age        -2423.2011508 -1.284405e+03
Mileage       -0.1708354 -7.720998e-02
```

The 95% confidence intervals for the effects of age and mileage are (-2423, -1284) and (-0.171, -0.077), respectively.

(d)

To cover the loss in value of the car over the term of the lease, we recommend structuring the lease for a 3-series BMW to cost $2,400 per year and additional $0.18 per mile. These estimates on average will cover the costs due to aging with 95% confidence.

(e)

Note that the R^2 of this model is just 0.51; namely, about half of the variance of the residual price has not been explained. First of all, this analysis ignores the fact that these cars cost different amounts at the time of purchase. We have not observed the actual loss in value; we have only seen how time and mileage have affected their value. They did not all start from the same initial cost. Also, we have not identified other differences among these cars, such as special options that might increase the value of car further.