Business Analytics (110-1)

Assignment 1 – Reference Solutions

1.

(a)

```
oacs <- read.table("OACs.txt", header=TRUE, sep="\t")
attach(oacs);       pairs(oacs)
oacs.1 <- lm(GPA ~ Best.6);     summary(oacs.1)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -5.35498    1.31289  -4.079 5.57e-05 ***
Best.6       0.15496    0.01458  10.632  < 2e-16 ***
---
Residual standard error: 0.8295 on 361 degrees of freedom
Multiple R-squared:  0.2385,    Adjusted R-squared:  0.2363
F-statistic:   113 on 1 and 361 DF,  p-value: < 2.2e-16
```

$t = 10.632$, p–value = 0. There is evidence of a linear relationship between the average of the best 6 OACs and university GPA. Also, $R^2 = 0.2385$, $\hat{\sigma} = 0.8295$

(b)

```
oacs.2 <- lm(GPA ~ B4.E.C);     summary(oacs.2)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -3.321975   0.854028   -3.89 0.000119 ***
B4.E.C       0.137001   0.009807   13.97  < 2e-16 ***
---
Residual standard error: 0.7658 on 361 degrees of freedom
Multiple R-squared:  0.3509,    Adjusted R-squared:  0.3491
F-statistic: 195.2 on 1 and 361 DF,  p-value: < 2.2e-16
```

$t = 13.97$, p–value = 0. There is evidence of a linear relationship between the average of the best 4 OACs plus English and calculus and university GPA. Also, $R^2 = 0.3509$, $\hat{\sigma} = 0.7658$.

(c)

The second model fits better (higher coefficient of determination and lower standard error of estimate) and as such is likely to be a better predictor of university GPA.

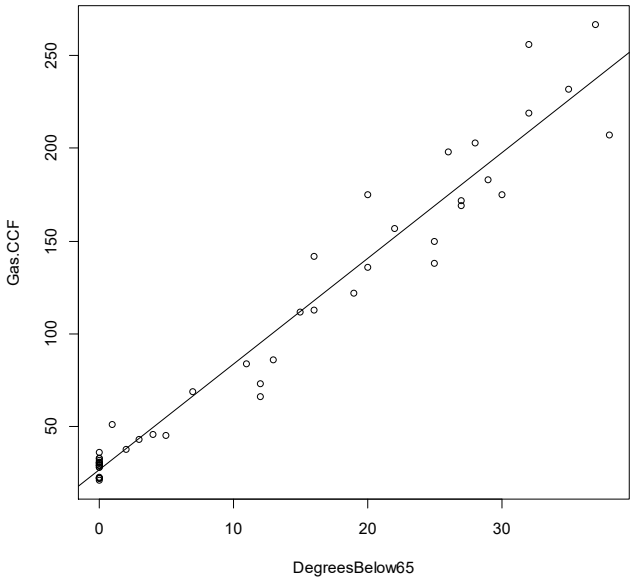
2.

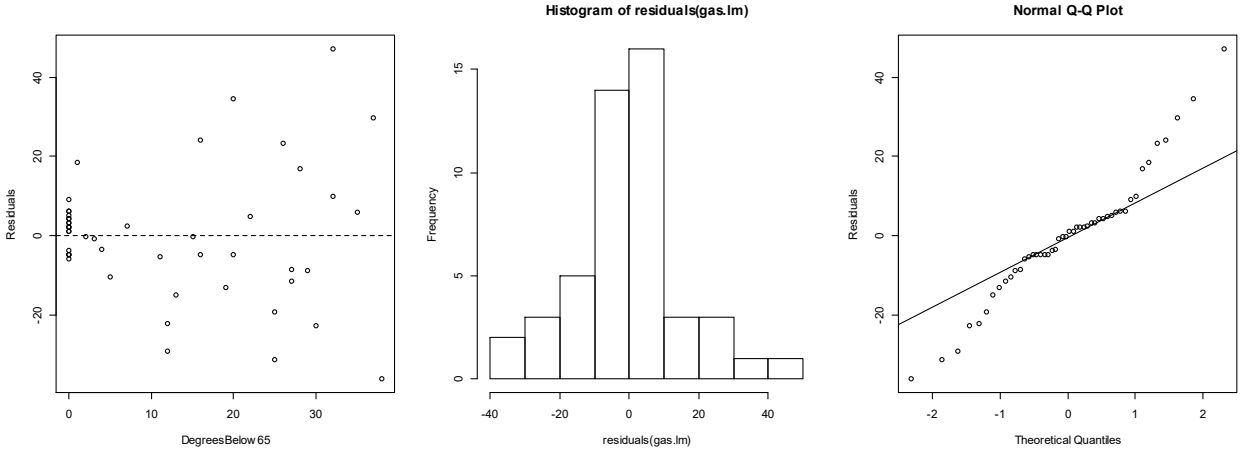
(a)

```
gas.con <- read.table("gas_consumption.txt",header=TRUE)
attach(gas.con)

plot(DegreesBelow65, Gas.CCF )

gas.lm <- lm(Gas.CCF ~ DegreesBelow65)
summary(gas.lm);    abline(gas.lm)
```






```
par(mfrow=c(1,3))
```

```
plot(DegreesBelow65, residuals(gas.lm), ylab="Residuals")
abline(h=0, lty=2)
hist(residuals(gas.lm))
qqnorm(residuals(gas.lm), ylab="Residuals");    qqline(residuals(gas.lm))
```

(b)

```
gasnew <- data.frame(DegreesBelow65 = pretty(DegreesBelow65, n=30))
yhat.ci <- predict(gas.lm, newdata=gasnew, interval="confidence")
ci <- data.frame(lower=yhat.ci[,"lwr"], upper=yhat.ci[,"upr"])
yhat.pi <- predict(gas.lm, newdata=gasnew, interval="prediction")
pi <- data.frame(lower=yhat.pi[,"lwr"], upper=yhat.pi[,"upr"])

plot(DegreesBelow65, Gas.CCF, main ="Confidence and Prediction Intervals", pch=20)
abline(gas.lm)
lines(gasnew$DegreesBelow65, ci$lower, lty=2, col="red")
lines(gasnew$DegreesBelow65, ci$upper, lty=2, col="red")
lines(gasnew$DegreesBelow65, pi$lower, lty=3, col="blue")
lines(gasnew$DegreesBelow65, pi$upper, lty=3, col="blue")
```

---

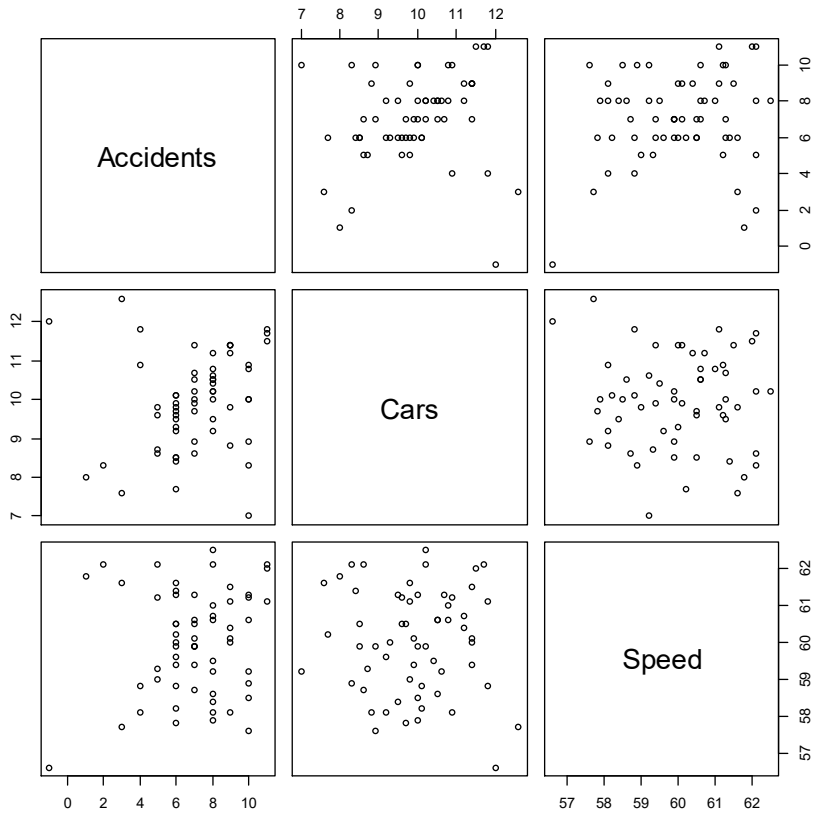
3. (Keller, 18.44 & 45)

```
acc <- read.table("car_accident.txt", header=TRUE, sep="\t")
attach(acc)
pairs(acc)   # shown on next page…
```

What have you observed from this scatterplot matrix? Let's build the first linear model with the given explanatory variables in the dataset.

```
acc.lm1 <- lm(Accidents ~ Cars + Speed);    summary(acc.lm1)
```

```
Coefficients:
            Estimate   Std. Error   t value   Pr(>|t|)
(Intercept) -12.8719    13.7901  -0.933    0.355
Cars          0.3733     0.2587   1.443    0.155
Speed         0.2699     0.2232   1.209    0.232
---
Residual standard error: 2.408 on 57 degrees of freedom
Multiple R-squared:  0.05548,   Adjusted R-squared:  0.02234
F-statistic: 1.674 on 2 and 57 DF,  p-value: 0.1965
```



The first model fits poorly… Let's improve the model by adding the interaction.

```
acc.lm2 <- lm(Accidents ~ Cars * Speed);     summary(acc.lm2)
```

```
Coefficients:
            Estimate   Std. Error   t value   Pr(>|t|)
(Intercept) 640.76174   53.80478    11.91    <2e-16 ***
Cars        -64.16571    5.26594   -12.19    <2e-16 ***
Speed       -10.62885    0.89668   -11.85    <2e-16 ***
Cars:Speed    1.07632    0.08779    12.26    <2e-16 ***
---
Residual standard error: 1.266 on 56 degrees of freedom
Multiple R-squared:  0.7436,    Adjusted R-squared:  0.7299
F-statistic: 54.14 on 3 and 56 DF,  p-value: < 2.2e-16
```

The interaction term creates significant improvement to the model, implying that the variance within the dependent variable can be well explained by this model.
Let's see if the model can be further improved by considering the second-order effect.

```
acc.lm3 <- lm(Accidents ~ Cars * Speed + I(Cars^2) + I(Speed^2))
```

```
summary(acc.lm3)

Coefficients:
            Estimate   Std. Error   t value   Pr(>|t|)
(Intercept)  404.49882  327.00467   1.237    0.221
Cars         -66.56803    6.53512 -10.186    3.54e-14 ***
Speed         -2.34571   10.53501  -0.223    0.825
I(Cars^2)      0.10696    0.09681   1.105    0.274
I(Speed^2)    -0.06960    0.08528  -0.816    0.418
Cars:Speed     1.08154    0.09648  11.210    1.03e-15 ***
---
Residual standard error: 1.269 on 54 degrees of freedom
Multiple R-squared:  0.7514,    Adjusted R-squared:  0.7284
F-statistic: 32.65 on 5 and 54 DF,  p-value: 3.564e-15
```

The second-order effects are not significant. The nested model test delivers the same message.

```
anova(acc.lm2, acc.lm3)
Model 1: Accidents ~ Cars * Speed
Model 2: Accidents ~ Cars * Speed + I(Cars^2) + I(Speed^2)
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     56 89.715
2     54 86.987  2    2.7286 0.8469 0.4343
```

Let's diagnose the second model to see if the regression assumptions are all satisfied.

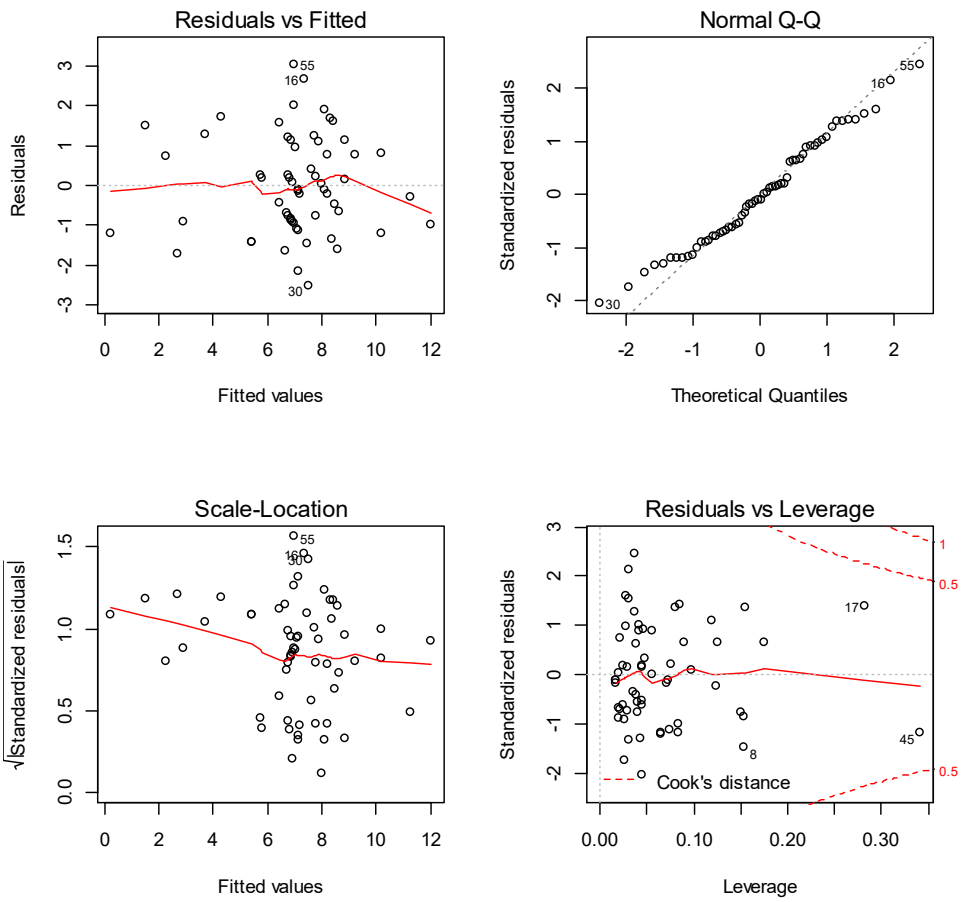
```
par(mfrow=c(2,2));      plot(acc.lm2)       # shown on next page…
```

What have you observed from these diagnostic plots?

How would you interpret the results of model 2?
By comparing models 1 and 2, it is clear that the interaction between the two predictors is the key. The regression equation for model 2 is

$$\hat{y} = 640.762 - 64.166\ Cars - 10.629\ Speed + 1.076\ Cars * Speed$$

With the average speed at 60, the equation becomes $\hat{y} = 3.022 + 0.394\ Cars$. It suggests that reducing one car may lower the number of accidents by 0.394, assuming that the average number of cars driving through is between 7 and 12.6 and that their average driving speed is 60. The effect of *Speed* can be explained in the similar way.


Residuals vs Fitted
Residuals
Fitted values
Normal Q-Q
Standardized residuals
Theoretical Quantiles
Scale-Location
√|Standardized residuals|
Fitted values
Residuals vs Leverage
Standardized residuals
Cook's distance
Leverage