Stat 500 – Homework 4 (Solutions)

Read in data and fit the model::

```
> library(faraway)
> data(teengamb)
> g = lm(gamble ~ sex + status + income + verbal, data = teengamb)
> summary(g)
```

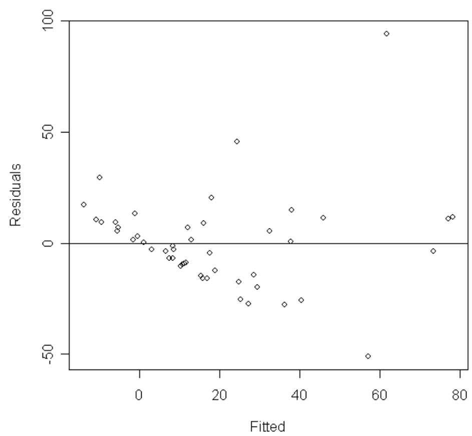
Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	22.55565	17.19680	1.312	0.1968
sex	-22.11833	8.21111	-2.694	0.0101 *
status	0.05223	0.28111	0.186	0.8535
income	4.96198	1.02539	4.839	1.79e-05 ***
verbal	-2.95949	2.17215	-1.362	0.1803

Residual standard error: 22.69 on 42 degrees of freedom
Multiple R-Squared: 0.5267, Adjusted R-squared: 0.4816
F-statistic: 11.69 on 4 and 42 DF, p-value: 1.815e-06

1) Basic diagnostic plot - residuals against fitted values::

```
plot(g$fit, g$res, xlab="Fitted", ylab="Residuals")
> abline(h=0)
```



From the above plot we observe that the variability of residuals is increasing with the increase in fitted value which indicates heteroscedasticity. We consider a square root transform of the response:

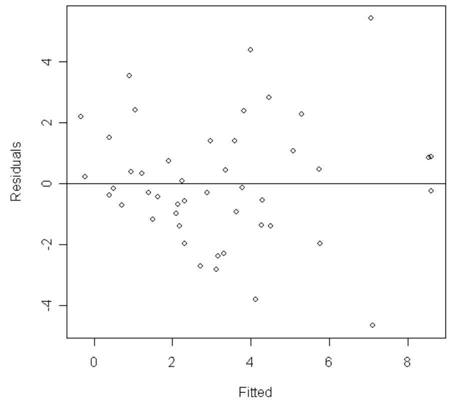
```
>gadj <- lm(sqrt(gamble)~ . , data = teengamb)
>summary(gadj)
> summary(gadj)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2.97707	1.57947	1.885	0.06638 .
sex	-2.04450	0.75416	-2.711	0.00968 **
status	0.03688	0.02582	1.428	0.16057
income	0.47938	0.09418	5.090	7.94e-06 ***
verbal	-0.42360	0.19950	-2.123	0.03967 *

Residual standard error: 2.084 on 42 degrees of freedom
Multiple R-Squared: 0.5646, Adjusted R-squared: 0.5231
F-statistic: 13.61 on 4 and 42 DF, p-value: 3.362e-07

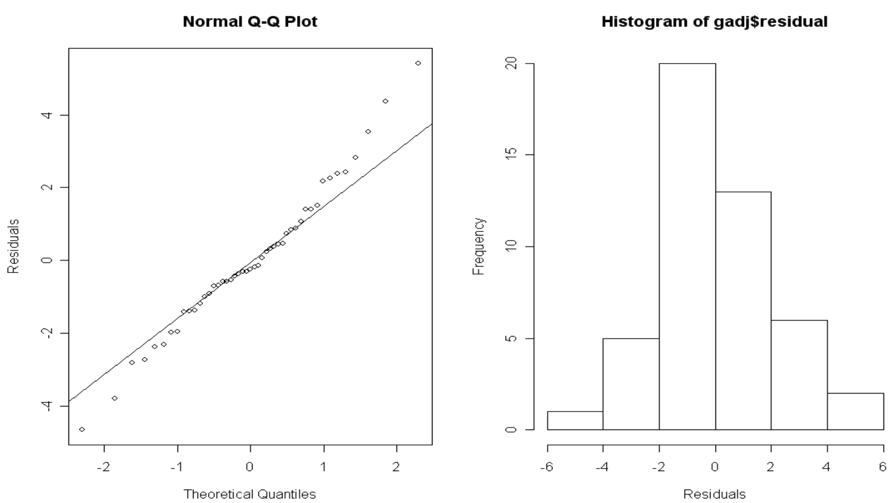
```
>plot(gadj$fit, gadj$res, xlab="Fitted", ylab="Residuals")
> abline(h=0)
```



Now the residual versus fitted plot looks much better and there does not appear to be any problems with non-constant variance. Now we perform the diagnostics on the new model.

2) Check the normality assumption::

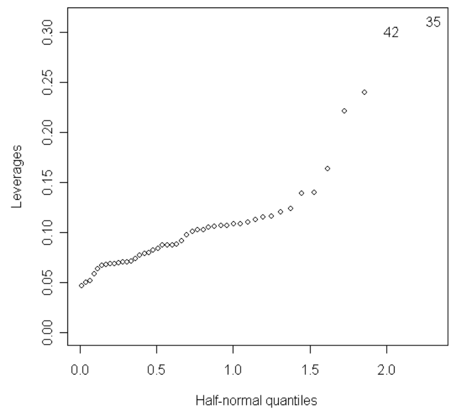
```
## QQ-plot
> qqnorm(gadj$residual, ylab="Residuals")
> qqline(gadj$residual)
## Histogram
> hist(gadj$residual, xlab="Residuals")
```



Based on the QQ-plot and the histogram we can say that there is no real issue with normality, even though the QQ-plot indicates slightly longer tails than normal.

3) Check for large leverage points::

```
## To find leverage
>halfnorm(lm.influence(gadj)$hat,labs=row.names(teengamb),ylab="Leverages")
```



From the above plot we see that 42nd and 35th data points have high leverage.

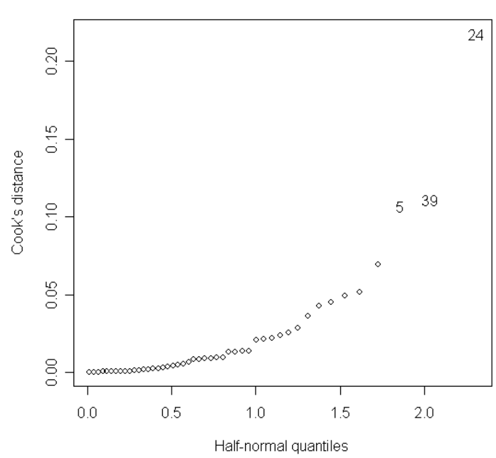
4) Check for outliers ::

```
## To find outliers
> jack <- rstudent(gadj)
> jack[order(abs(jack),decreasing=TRUE)][1:5]
      24        39        36        23         5
 3.037005 -2.486949  2.249705 -1.953221  1.877841

> ## To compute p=value
> 2*(1-pt(max(abs(jack)),df=47-5-1))
[1] 0.00414277

> ## To compare to alpha/n
> 0.05/47
[1] 0.001063830
```

There are no major outliers in the data since there are no significant residuals according to the test.

5) Check for influential points::

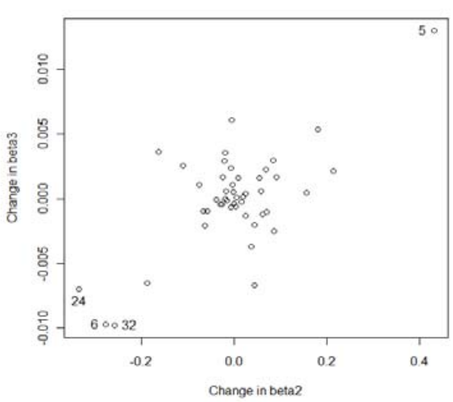
```
## To find influential points
> cook = cooks.distance(gadj)
> halfnorm(cook, nlab = 3, ylab="Cook's distance")
```

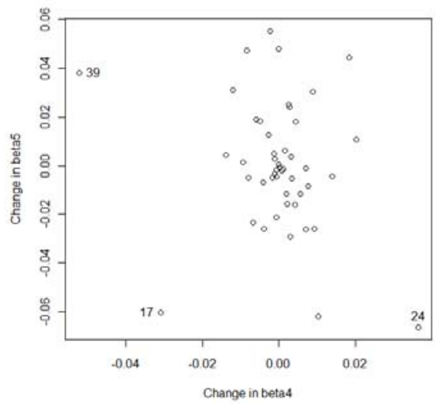


The data having case no. 24, 39 and 5 have high cook's distance and as well as high residuals but they don't have high leverage.

Now we consider the following graphs of changes in the β's by dropping one observation:

```
## Compute changes in coefficients
 result.inf <- lm.influence(gadj)

 plot(result.inf$coef[,2], result.inf$coef[,3],xlab="Change in beta2",ylab="Change in beta3")
## interactive tool to identify points by clicking
 identify(result.inf$coef[, 2], result.inf$coef[, 3])

plot(result.inf$coef[,4], result.inf$coef[,5],xlab="Change in beta4",ylab="Change in beta5")
## interactive tool to identify points by clicking
 identify(result.inf$coef[, 4], result.inf$coef[, 5])
```





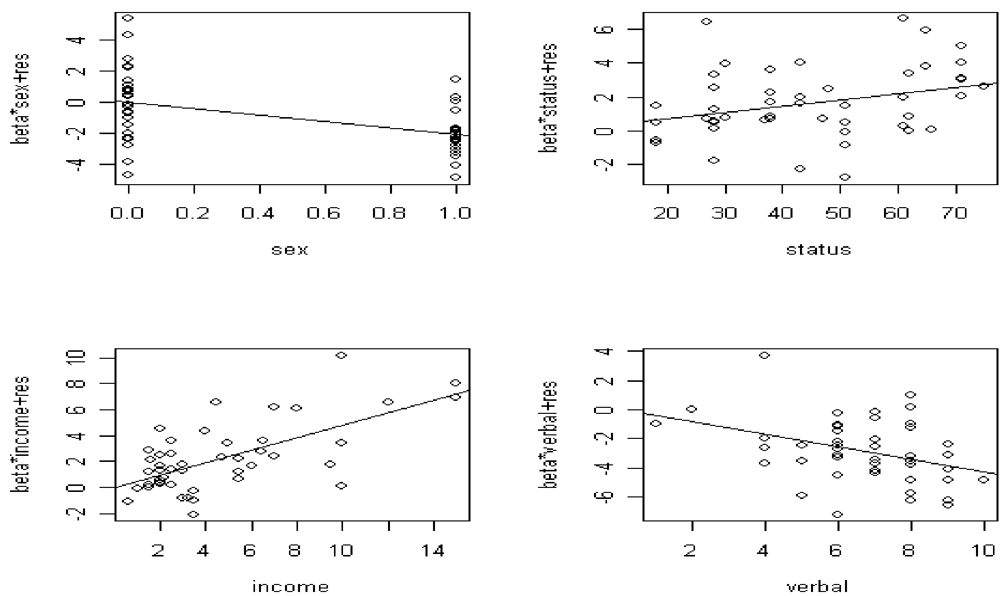
From the above plots 6th, 32nd and 17th observations are identified as influential points along with 5th, 24th and 39th observations. Now we construct regression models by dropping each of these observations. It seems that the observations other than 39th and 24th does not make much difference. The models constructed by dropping 39th and 24th observations are given below:

```
> g1 <- lm(sqrt(gamble)~ ., data=teengamb , subset=(cook < max(cook)))
> summary(g1)
Model excluding 24th obs:
Coefficients:
             Estimate   Std. Error   t value   Pr(>|t|)
(Intercept)   2.11915    1.47175      1.440    0.1575
sex          -1.70997    0.69840     -2.448    0.0187 *
status        0.04387    0.02372      1.849    0.0716 .
income        0.44312    0.08695      5.096    8.22e-06 ***
verbal       -0.35706    0.18375     -1.943    0.0589 .
Residual standard error: 1.906 on 41 degrees of freedom
Multiple R-Squared: 0.5503,     Adjusted R-squared: 0.5065
F-statistic: 12.55 on 4 and 41 DF,  p-value: 9.403e-07

> g2 <- lm(sqrt(gamble)~ ., data=teengamb[-39,])
> summary(g2)
Model excluding 39th obs:
Coefficients:
             Estimate   Std. Error   t value   Pr(>|t|)
(Intercept)   2.81342    1.49162      1.886    0.06637 .
sex          -2.08803    0.71174     -2.934    0.00546 **
status        0.04357    0.02451      1.778    0.08288 .
income        0.53150    0.09129      5.822    7.75e-07 ***
verbal       -0.46173    0.18885     -2.445    0.01887 *

Residual standard error: 1.966 on 41 degrees of freedom
Multiple R-Squared: 0.6211,     Adjusted R-squared: 0.5841
F-statistic:  16.8 on 4 and 41 DF,  p-value: 3.147e-08
```

We can see that dropping the 24th observation does not change the estimates much or the R^2. The p-values are still close even though some change formal significance levels (sex , status and verbal). On the other hand dropping the 39th observation improves the fit (since R^2 increases) but the estimates and corresponding p-values are almost same. Overall , we can say that the models are fairly stable.

6) Check the structure of the relationship between the predictors and the Response:: [not graded]

```
## Partial residual plots::
>par(mfrow=c(2,2))
> prplot(gadj, 1)
> prplot(gadj, 2)
> prplot(gadj, 3)
> prplot(gadj, 4)
```



Income seems to have stronger linear relationship with response in comparison to other predictors**.** There is no indication of a nonlinear structure.