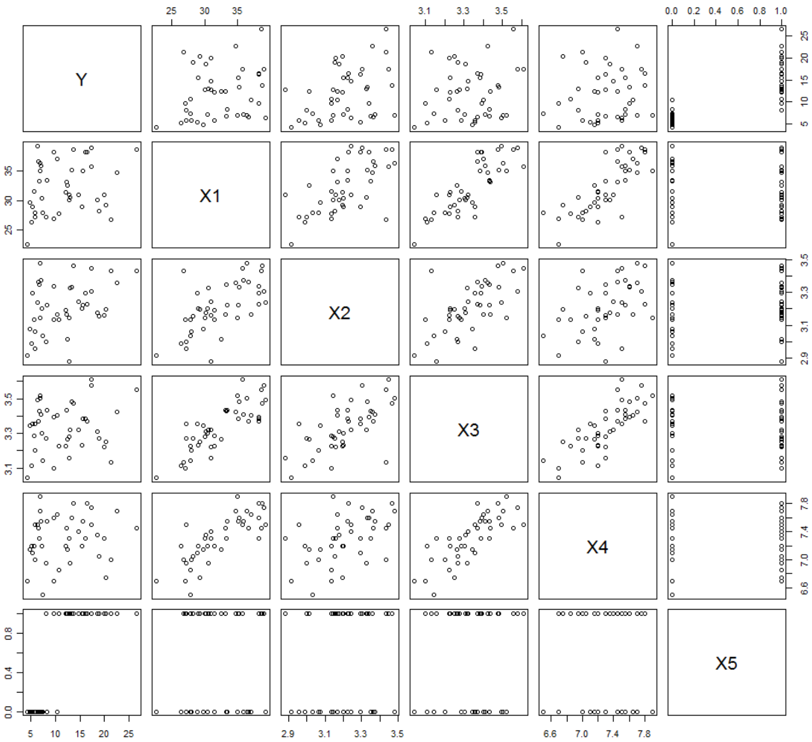
SCHOOL OF FINANCE, ACTUARIAL STUDIES AND APPLIED STATISTICS
REGRESSION MODELLING (STAT2008/STAT6038)
ASSIGNMENT 2 SOLUTIONS

(80 marks total for assignment)
(50 marks total for Q1)
1. (1) (5 marks – 2 for the plot, 3 for the comments)

```
ass2q1<-read.csv("ass2q1.csv")
attach(ass2q1)
pairs(ass2q1)
```



Y seems to have a broadly positive relationship with each variable, though it would be difficult to classify any particular relationship involving Y as either strong or linear. There seems to be reasonably strong linear relationships among the covariates – between X1 and X2, X1 and X3, X1 and X4, X2 and X3, and X3 and X4. X5 is a binary variable, so it is hard to visualise relationships, though it does appear that being an X5 does seem associated positively with Y.

(2) (13 marks – take off one mark for each error or missing plot)
The relevant R commands are:

```
> ass2q11.lm <- lm(Y ~ X1 + X2 + X3 + X4 +X5)
> summary(ass2q11.lm)
> ti <- ls.diag(ass2q11.lm)$stud.res
> plot(fitted(ass2q11.lm),ti,xlab="Fitted Values",
+       ylab="(externally) Studentized residuals")
> identify(fitted(ass2q11.lm),ti,n=2)
> abline(2,0)
> abline(-2,0)
> qqnorm(ti,ylab="(externally) Studentized residuals")
> abline(0,1,lty=2)
> install.packages("car")
> library("car")
> av.plots(ass2q11.lm)
> par(mfrow=c(2,3))
> barplot(hat(X1),xlab="lev plot for X1")
> abline(h=4/length(X1))
> barplot(hat(X2),xlab="lev plot for X2")
> abline(h=4/length(X2))
> barplot(hat(X3),xlab="lev plot for X3")
> abline(h=4/length(X3))
> barplot(hat(X4),xlab="lev plot for X4")
> abline(h=4/length(X4))
> barplot(hat(X5),xlab="lev plot for X5")
> abline(h=4/length(X5))
> barplot(hat(ass2q1[,-1]),xlab="lev plot for design matrix")
> abline(h=12/length(X5))
```

```
Call:
lm(formula = Y ~ X1 + X2 + X3 + X4 + X5)

Residuals:
    Min      1Q  Median      3Q     Max 
-6.7488 -2.1088  0.0782  1.0011  9.2442 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -29.18930   21.42112  -1.363   0.1812    
X1           -0.06182    0.25367  -0.244   0.8088    
X2            9.08199    4.85821   1.869   0.0695 .  
X3            0.69159    8.01446   0.086   0.9317    
X4            0.90056    2.88612   0.312   0.7568    
X5            8.43348    1.10401   7.639 4.06e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.402 on 37 degrees of freedom
Multiple R-squared: 0.6756,  Adjusted R-squared: 0.6317 
F-statistic: 15.41 on 5 and 37 DF,  p-value: 3.434e-08
```

The parameter estimates (coefficients) are given in bold in the printout above; the associated standard errors are given in italics. The fitted model is:

$\hat{Y}$ = -29.1893 – 0.0618 X1 + 9.082 X2 + 0.6916 X3 + 0.9 X4 + 8.43348 I(X5), where I(X5) is an indicator function which is 0 or 1 for the X5's. The coefficient of determination is 67.56%.

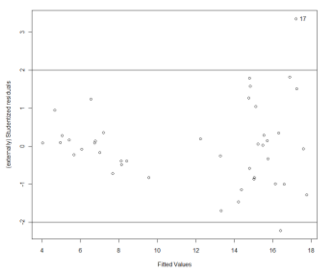
$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$

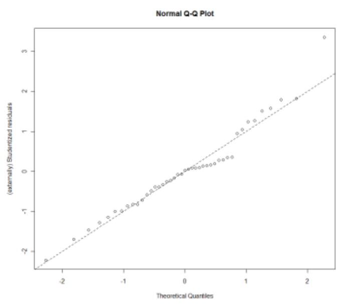
$H_A : \text{Not all slopes are zero}$

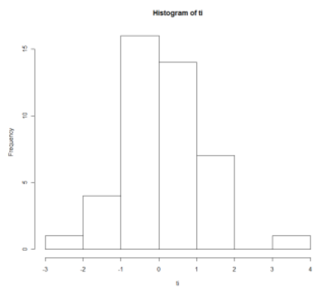
Test Statistic =F=15.41 compare to the F distribution on 5 and 37 dof.

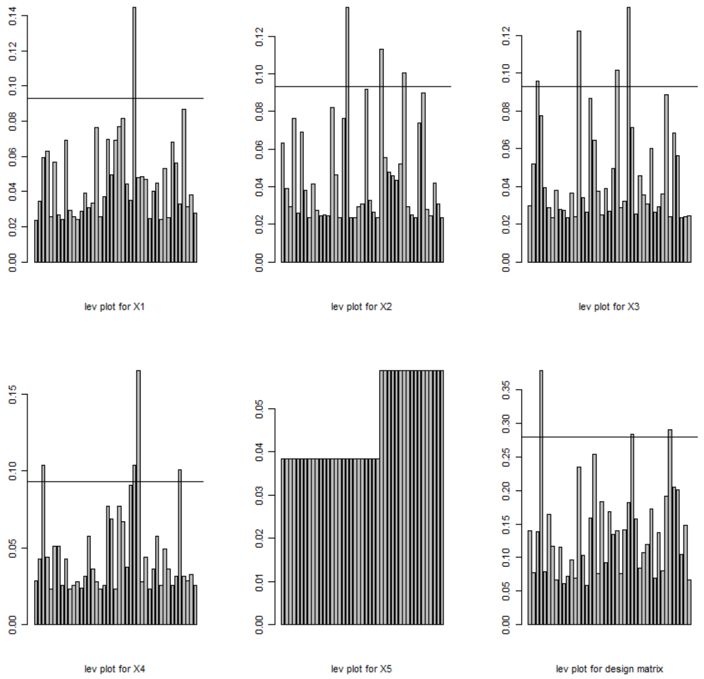
Reject the null if P(F>15.41) is small.
SKETCH

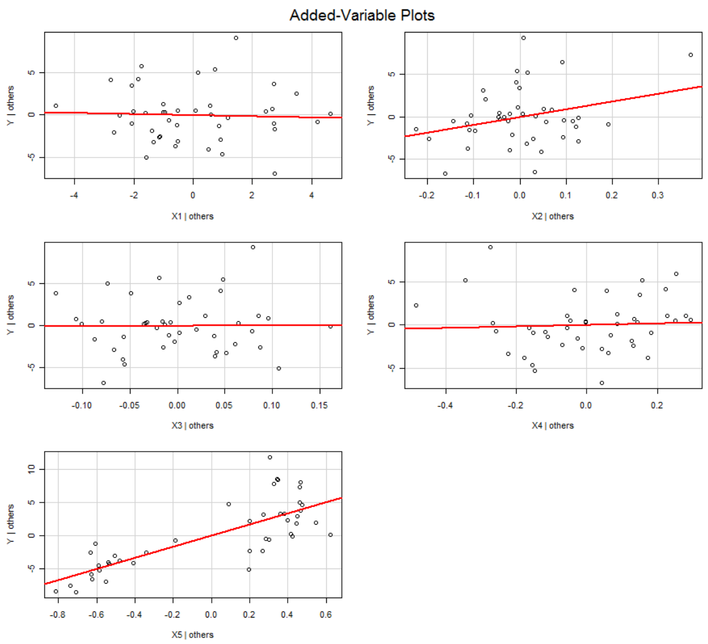
The model is highly significant (*p* value is near zero).











The errors are assumed to be Normally distributed with zero mean, independent with constant variance. A linear relationship between the predictors and the response is also assumed.

The externally studentised residuals shows clear evidence of heteroscedasticity among the errors in the vs fits plot. There appears to be increasing variance for increasing fitted values. The plot doesn't show any clear signs of the violation of the independence assumption. The NQ plot and histogram show no serious departures from the assumption of normality. The added variable plots indicate that there may be a positive linear relationship between X2 and Y and X5 and Y. The other predictors don't appear to have a relationship with the response. The leverage plot for the full design matrix indicates that the fourth observation has the potential to be influential.

(3) **(13 marks – same breakdown as for (b))** The relevant R commands are:

```
> ass2q13.lm <- lm(log(Y) ~ X1 + X2 + X3 + X4 + X5)
> summary(ass2q13.lm)
> ti <- ls.diag(ass2q13.lm)$stud.res
> plot(fitted(ass2q13.lm),ti,xlab="Fitted Values",    ylab="(externally) Studentized residuals")
```

```
> identify(fitted(ass2q13.lm),ti,n=3)


Call:
lm(formula = log(Y) ~ X1 + X2 + X3 + X4 + X5)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.55070 -0.15643 -0.02498  0.09328  0.43105 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -0.677777   1.507806  -0.450   0.6557    
X1           0.005047   0.017855   0.283   0.7790    
X2           0.619338   0.341963   1.811   0.0782 .  
X3          -0.087295   0.564128  -0.155   0.8779    
X4           0.096454   0.203151   0.475   0.6377    
X5           0.797930   0.077710  10.268 2.22e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1 

Residual standard error: 0.2395 on 37 degrees of freedom
Multiple R-squared: 0.7857,  Adjusted R-squared: 0.7568 
F-statistic: 27.13 on 5 and 37 DF,  p-value: 1.973e-11
```

The parameter estimates (coefficients) are rendered as bold in the above printout, and their associated standard errors are rendered in italics. The fitted model is:

$$\hat{Y} = \exp(\mathbf{-0.6777} + \mathbf{0.0051}\ X1 + \mathbf{0.6193}\ X2 - \mathbf{0.0873}\ X3 + \mathbf{0.0964}\ X4 + \mathbf{0.7979}\ I(X5)).$$

$H_0: \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$

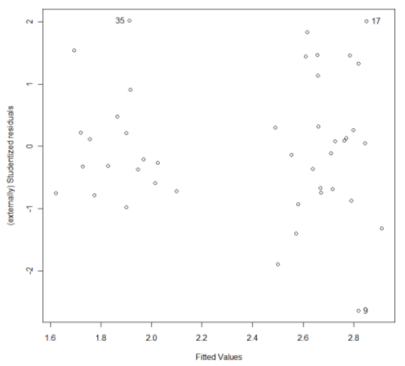
$H_A: \textit{Not all slopes are zero}$

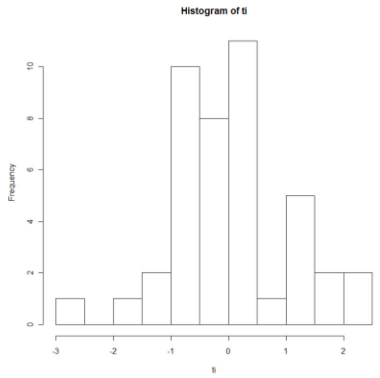
Test Statistic =F=27.13 compare to the F distribution on 5 and 37 dof.

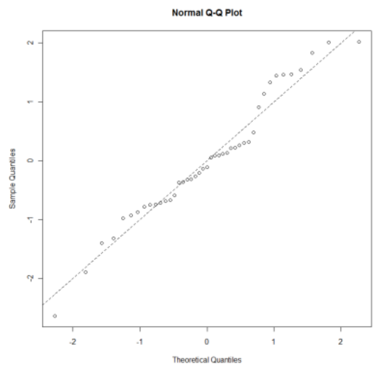
Reject the null if P(F>27.13) is small.
SKETCH

The model is highly significant (*p* value is near zero).

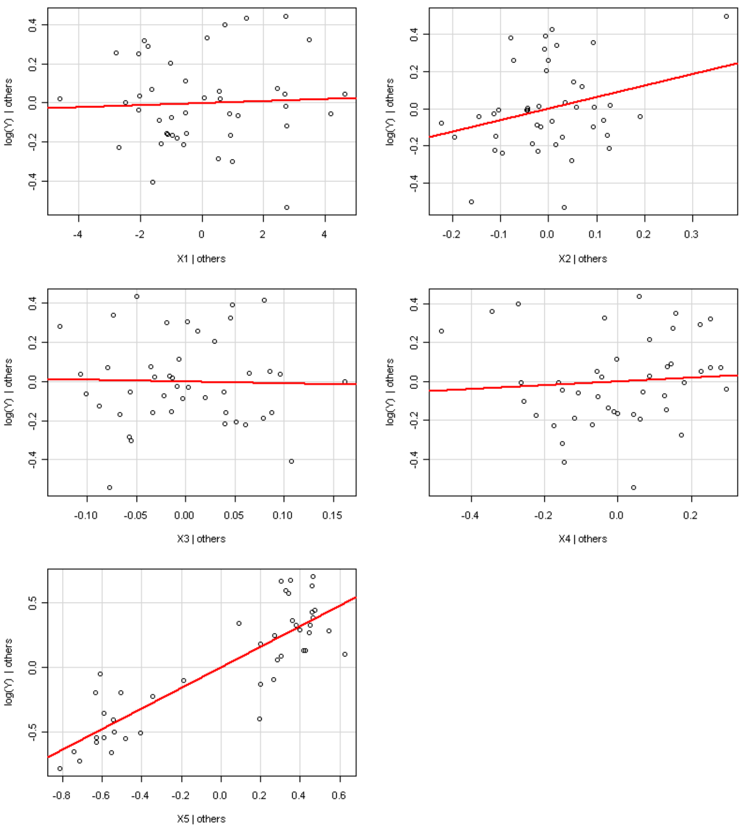
The coefficient of determination is higher than in the preceding model, at 78.57%. The diagnostic plots are below, and show no remaining evidence of heteroscedasticity nor non-normality. The independence assumption appears to be satisfied. The model now appears reasonably acceptable. X2 and X5 still show a positive linear relationship with the transformed response on the added variable plots.







Added-Variable Plots



(4) **(10 marks)** Assuming all other variables are equal, then the predicted difference in log(Y)'s is:

$$\log(Y)_{X5=0} - \log(Y)_{X5=1} = \hat{\beta}_{X5=1} = -0.797930$$

and $e^{-0.797930} = 0.45026$

The predicted ratio in Y's is 0.45026

(5) **(9 marks – 3 for each test)** The relevant order fits X4 last, X3 second-last and X1 third last. The relevant R commands are:

```
> ass2q13a.lm <- lm(log(Y) ~ X2+X5+X1+X3+X4)
> summary(ass2q13a.lm)

Call:
lm(formula = log(Y) ~ X2 + X5 + X1 + X3 + X4)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.55070 -0.15643 -0.02498  0.09328  0.43105 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -0.677777   1.507806  -0.450   0.6557    
X2           0.619338   0.341963   1.811   0.0782 .  
X5           0.797930   0.077710  10.268 2.22e-12 ***
X1           0.005047   0.017855   0.283   0.7790    
X3          -0.087295   0.564128  -0.155   0.8779    
X4           0.096454   0.203151   0.475   0.6377    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 

Residual standard error: 0.2395 on 37 degrees of freedom
Multiple R-squared: 0.7857,  Adjusted R-squared: 0.7568 
F-statistic: 27.13 on 5 and 37 DF,  p-value: 1.973e-11
```

```
> anova(ass2q13a.lm)
Analysis of Variance Table

Response: log(Y)
          Df Sum Sq Mean Sq  F value    Pr(>F)    
X2         1 1.4262  1.4262  24.8711 1.468e-05 ***
X5         1 6.3003  6.3003 109.8661 1.253e-12 ***
X1         1 0.0402  0.0402   0.7016    0.4076    
X3         1 0.0001  0.0001   0.0014    0.9702    
X4         1 0.0129  0.0129   0.2254    0.6377    
Residuals 37 2.1218  0.0573                       
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Test One

$H_0: \beta_4 = 0$

$H_A: \beta_4 \neq 0$

Test Statistic = F=0.2254

Compare to the F distribution on 1 numerator and 37 denominator degrees of freedom. We will reject the Null if P(F>0.2254) is less than alpha.

SKETCH

Since the p-value is large, (p-value = 0.6377) then we cannot reject the null. The coefficient for X4 is non-significant.

Test Two

$H_0: \beta_3 = \beta_4 = 0$

$H_A: \textit{Not all slopes are zero}$

Test Statistic = $F = \dfrac{(.0001 + .0129)/2}{.0573}$ =.1134

Compare to the F distribution on 2 numerator and 37 denominator degrees of freedom. We will reject the Null if P(F>0.1134) is less than alpha or F>5.229

SKETCH

```
> 1-pf(.1138,2,37)
[1] 0.8927476
> qf(.99,2,37)
[1] 5.229022
```

We can see that the test statistic does not lie in the rejection region and the p-value is large so we cannot reject the null hypothesis. It appears that the coefficients for these two predictors are zero.

Test Three

$H_0: \beta_1 = \beta_3 = \beta_4 = 0$

$H_A: \textit{Slopes not all zero}$

Test Statistic = $F = \dfrac{(.0001 + .0129 + .0402)/3}{.0573}$ =.3095

Compare to the F distribution on 3 numerator and 37 denominator degrees of freedom. We will reject the Null if P(F>0..3095) is less than alpha or F>4.35954

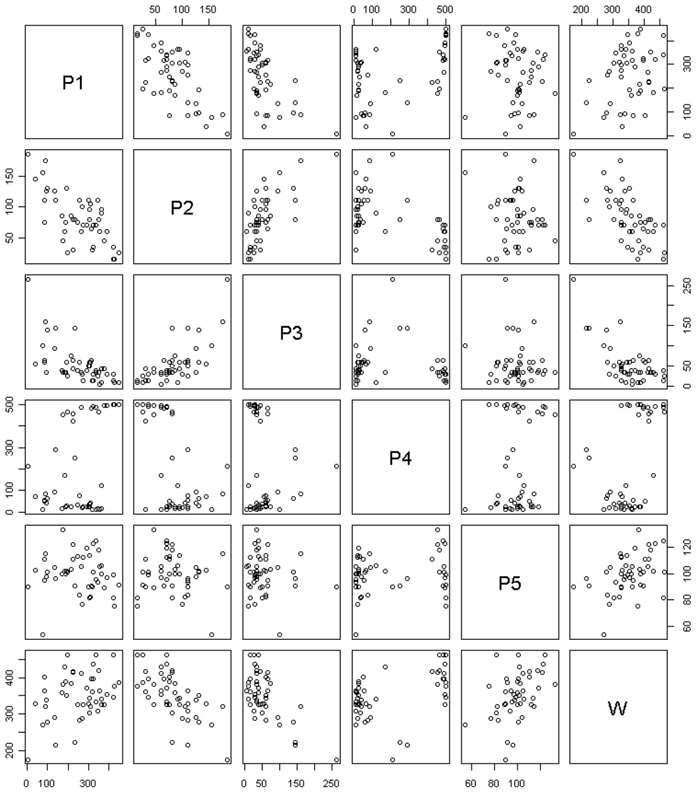
SKETCH

```
> qf(.99,3,37)
[1] 4.35954
> 1-pf(0.3095,3,37)
[1] 0.8183591
```

We can see that the test statistic does not lie in the rejection region and the p-value is large Overall, it seems that the variables X1, X3 and X4 are not needed in the model.

Note that this is consistent with what we saw in the added variable plots as well.

(30 marks total for Q2)
2. (1) Take off 1 mark wherever a mistake is made or part of the analysis is missing.




```
> ass2q21.lm <- lm(W ~ P1 + P2 + P3 + P4 + P5)
> summary(ass2q21.lm)

Call:
lm(formula = W ~ P1 + P2 + P3 + P4 + P5)

Residuals:
    Min      1Q  Median      3Q     Max
-76.361 -26.321   2.507  20.588  76.590

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 334.55199   53.52589   6.250 1.91e-07 ***
P1           -0.17210    0.07030  -2.448  0.01873 *
P2           -0.25778    0.25387  -1.015  0.31587
P3           -0.87095    0.18300  -4.759 2.42e-05 ***
P4            0.10414    0.03525   2.954  0.00517 **
P5            1.07699    0.38168   2.822  0.00733 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 35.82 on 41 degrees of freedom
Multiple R-squared: 0.7068,  Adjusted R-squared: 0.671
F-statistic: 19.77 on 5 and 41 DF,  p-value: 5.574e-10
```

Test Three

$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$

$H_A : Not\ all\ slopes\ are\ zero$

Test Statistic =F=19.77 compare to the F distribution on 5 and 41 dof.

Reject the null if P(F>19.77) is small.
SKETCH

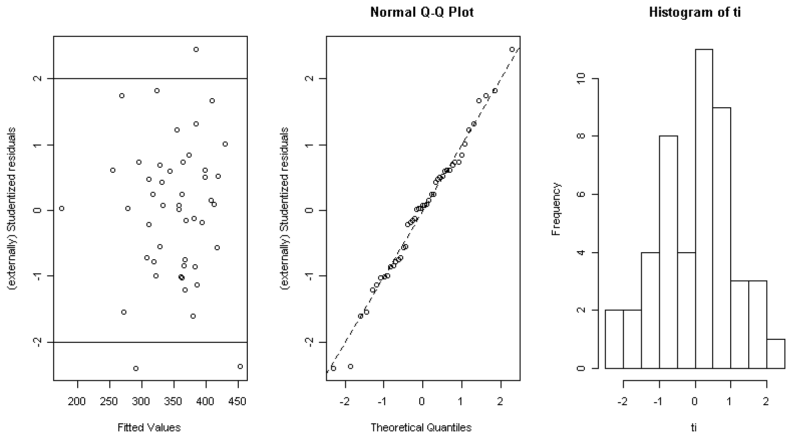
The model is highly significant (p value is near zero).

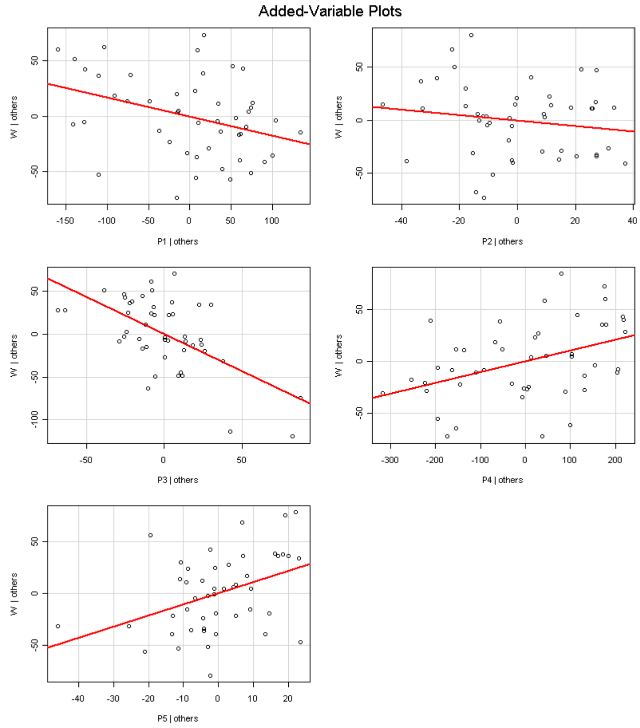
The coefficient of determination is 70.68%. The model coefficients and standard errors are given in the table above, and the fitted model is

$$\hat{W} = 334.55199 - 0.172\ P1 - 0.258\ P2 - 0.871\ P3 + 0.104\ P4 + 1.077\ P5.$$

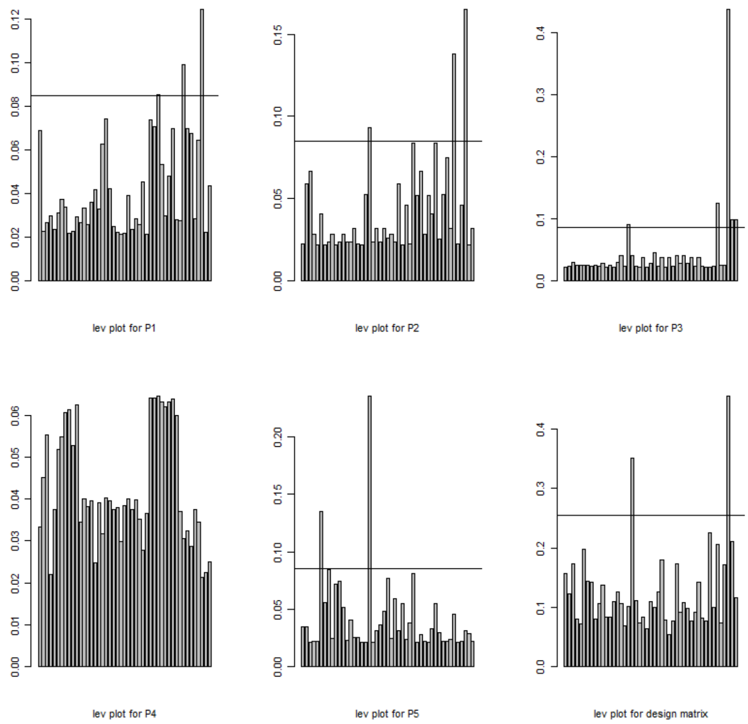
(2) **(8 marks)**

```
>plot(fitted(ass2q21.lm),ti,xlab="Fitted Values",
+    ylab="(externally) Studentized residuals")
> abline(2,0)
> abline(-2,0)
> qqnorm(ti,ylab="(externally) Studentized residuals")
> abline(0,1,lty=2)
> hist(ti)
> par(mfrow=c(2,3))
> barplot(hat(P1),xlab="lev plot for P1")
> abline(h=4/length(P1))
> barplot(hat(P2),xlab="lev plot for P2")
> abline(h=4/length(P2))
> barplot(hat(P3),xlab="lev plot for P3")
> abline(h=4/length(P3))
> barplot(hat(P4),xlab="lev plot for P4")
> abline(h=4/length(P4))
> barplot(hat(P5),xlab="lev plot for P5")
> abline(h=4/length(P5))
> barplot(hat(ass2q2[,-6]),xlab="lev plot for design matrix")
> abline(h=12/47)
```





The added variable plots show that P1 and P3 have a possible weak negative linear relationship with W. P4 and P5 show a possible positive linear relationship with W. P2, shows a very weak negative linear relationship. There doesn't seem to be any real problems with outliers.



There appear to be no problems with the assumptions of linearity,independence, homoscedasticity or normality, and no outliers. The leverage plot shows two points with high leverage, and these might be investigated further for influence.

(3) (5 marks – 3 for the prediction, 2 for the interval (0 if they use a CI instead of a PI)) The prediction is below. The predicted value for W is 545.7109, and the 90% prediction interval is (423.7157, 667.706).

```
> P1<-150
```

```
> P2<-130
> P3<-65
> P4<-390
> P5<-266
> predict(ass2q21.lm,as.data.frame(cbind(P1, P2, P3, P4, P5)), interval="prediction",level=.90)
       fit      lwr     upr
1 545.7109 423.7157 667.706
```

(4) **(9 marks – 3 for writing the hypothesis, 3 for writing down full and reduced models, 3 for doing the test).**

The reduced model in this case, fit below, is:

$$W = \beta_0 + \beta_1 P_1 + \beta_2 P_2 + \beta_3 (P_3 - P_5) + \beta_4 P_4 + \varepsilon$$

```
> rm(P1,P2,P3,P4,P5)
> newVar <- P3 - P5
> ass2q21.red <- lm(W ~ P1 + P2 + P4 + newVar)
> anova(ass2q21.lm)
Analysis of Variance Table

Response: W
          Df Sum Sq Mean Sq F value    Pr(>F)    
P1         1  22371   22371  17.432 0.0001514 ***
P2         1  55259   55259  43.059 6.872e-08 ***
P3         1  22295   22295  17.373 0.0001547 ***
P4         1  16688   16688  13.004 0.0008353 ***
P5         1  10218   10218   7.962 0.0073333 ** 
Residuals 41  52617    1283                      
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> anova(ass2q21.red)
Analysis of Variance Table

Response: W
          Df Sum Sq Mean Sq F value    Pr(>F)    
P1         1  22371   22371 17.7667 0.0001297 ***
P2         1  55259   55259 43.8863 5.016e-08 ***
P4         1   2422    2422  1.9238 0.1727521    
newVar     1  46512   46512 36.9390 3.070e-07 ***
Residuals 42  52884    1259                      
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The test proceeds by comparing full and reduced models:

$$H_0 : \beta_3 = -\beta_5$$
$$H_A : \beta_3 \neq -\beta_5$$

$F = \dfrac{(52884 - 52617)/1}{1283}$=.2081 Compare to the F distribution on 1 and 41 dof. We will reject if P(F>.2081) is less than alpha or if F>4.078546

```
> 1-pf(.2081,1,41)
[1] 0.650668
> qf(0.95,1,41)
[1] 4.078546
```

SKETCH

We conclude that the null hypothesis is plausible, at the 5% level.