STATS191 Winter 2018 Homework 3

SUNet ID: avati
Name: Anand Avati

By turning in this assignment, I agree by the Stanford honor code and declare that all of this is my own work.

Problem 1

(a) Let the given model be

$$Y = \gamma_0 X^{\gamma_1}$$

We apply the following transformations:

$$Y' = \log Y$$

$$X' = \log X$$

Taking logarithms on both sides

$$\log Y = \log \gamma_0 X^{\gamma_1}$$

$$\Rightarrow \log Y = \log \gamma_0 + \gamma_1 \log X$$

$$\Rightarrow \boxed{Y' = \beta_0 + \beta_1 X'}$$

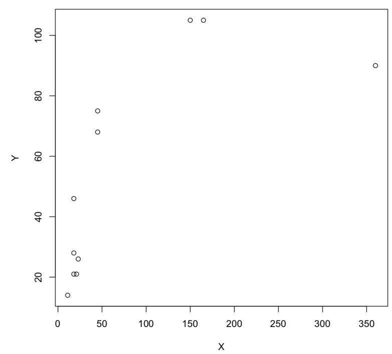
where

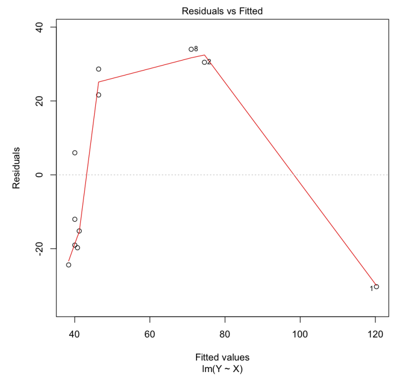
$$\boxed{\beta_0 = \log \gamma_0, \qquad \beta_1 = \gamma_1}$$

(b) • Is a linear model relating X and Y appropriate?

A linear model is NOT appropriate. A simple scatter plot does not show a linear relationship. The residual plot shows a pattern that should not be present in a linear model.

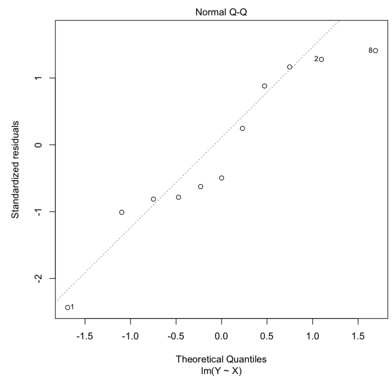
```
> X = read.table('Locomotion.txt')
> Y = read.table('Play.txt')
> d = data.frame(X, Y)
> colnames(d) <- c('X', 'Y')
> plot(d)
> reg = lm(Y ~ X, d)
> plot(reg)
```





- Does the data appear normal?

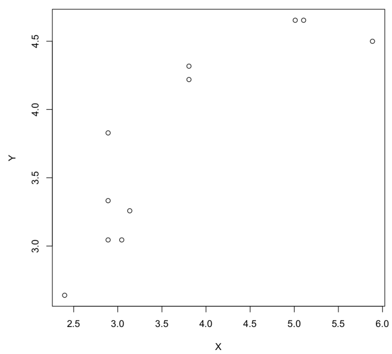
  The data does NOT appear normal. A visual inspection of the qq-norm plot confirms this.

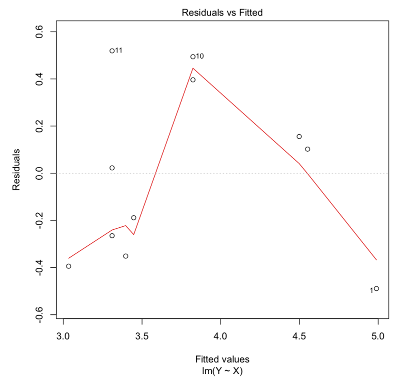


- Is a transformation along the lines of that studied in (a) appropriate?

```
> logd = log(d)
> plot(logd)
> reglog = lm(Y ~ X, logd)
> plot(reglog)
```

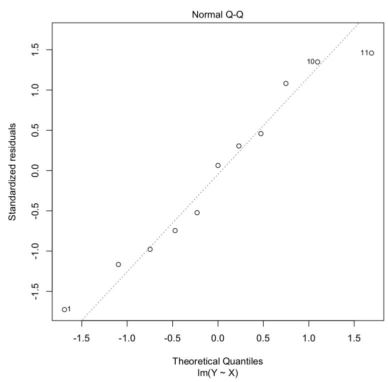
The transformation along the lines of (a) is appropriate. After transformation the scatter plot and residual plot appear more linear.





- Do the transformed variables appear normal?

  The transformed variables appear normal.



- Do any of the data points look suspicious (in either model)?

  We find data point 1 to be an outlier (strong outlier in the original model, weak outlier in the transformed model). We use the cooks.distance to measure outliers.

```
> cooks.distance(reg)
         1          2          3          4          5          6          7 
9.88102898 0.14866317 0.04509596 0.02620999 0.07703064 0.01731612 0.04332179 
         8          9         10         11 
0.15334123 0.04358641 0.07636143 0.00425722 
> cooks.distance(reglog)
           1            2            3            4            5            6 
1.2494752572 0.0147627285 0.0687879147 0.0181190575 0.1977176044 0.0003361866 
           7            8            9           10           11 
0.0465674939 0.0298354009 0.0588949720 0.0916315903 0.1783889681 
```

- Fit both a linear and nonlinear model (i.e. using transformations) and carry out appropriate tests of significance in answering this question. Be careful to check all assumptions.

Assumptions of linearity, homoscedastic errors and normality of errors are verified in the plots above.

The linear model has a p-value of 0.013 (for the null hypothesis $\gamma_1 = 0$).

```
> summary(reg)

Call:
lm(formula = Y ~ X, data = d)

Residuals:
   Min     1Q Median     3Q    Max 
-30.29 -19.39 -12.03  25.13  33.99 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)   
(Intercept) 35.80946    9.88193   3.624  0.00554 **
X            0.23466    0.07611   3.083  0.01307 * 
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1 

Residual standard error: 25.92 on 9 degrees of freedom
Multiple R-squared:  0.5137,Adjusted R-squared:  0.4596 
F-statistic: 9.506 on 1 and 9 DF,  p-value: 0.01307 
```

The non-linear model (transformed variables) has a p-value of 0.0005361 (for the null hypothesis $\gamma_1 = 0$).

```
> summary(reglog)

Call:
lm(formula = Y ~ X, data = logd)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.48933 -0.30836  0.02253  0.27588  0.51897 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   1.6893     0.4142   4.079 0.002763 ** 
X             0.5606     0.1070   5.238 0.000536 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1 

Residual standard error: 0.3845 on 9 degrees of freedom
```

```
Multiple R-squared:  0.753,Adjusted R-squared:  0.7256
F-statistic: 27.44 on 1 and 9 DF,  p-value: 0.0005361
```

Problem 2

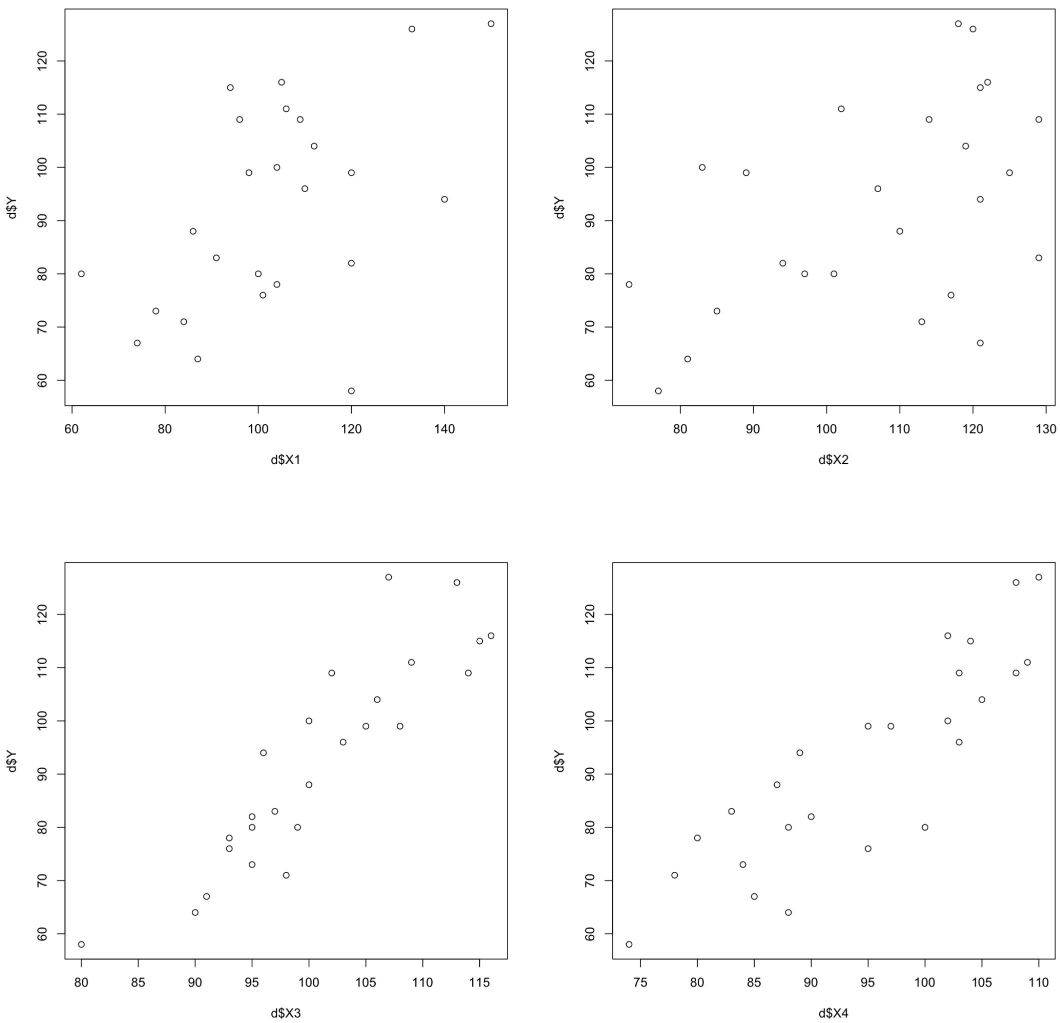
(a) Stem and leaf plots for each of the test scores

```
> d = read.table('Job.txt', col.names=c('X1', 'X2', 'X3', 'X4', 'Y'))
> stem(d$X1)

  The decimal point is 1 digit(s) to the right of the |

   6 | 248
   8 | 4671468
  10 | 014456902
  12 | 0003
  14 | 00

> stem(d$X2)

  The decimal point is 1 digit(s) to the right of the |

   6 | 37
   8 | 135947
  10 | 127034789
  12 | 01112599

> stem(d$X3)

  The decimal point is 1 digit(s) to the right of the |

   8 | 0
   9 | 01335556789
  10 | 002356789
  11 | 3456

> stem(d$X4)

  The decimal point is 1 digit(s) to the right of the |

   7 | 48
   8 | 03457889
```

```
 9 | 0557
10 | 0223345889
11 | 0
```

We see that $X4$ appears bimodal, and $X2$ and $X3$ appear skewed.

(b) Scatter plots of each text against Proficiency:



The variable $X1$ and $X2$ appear to be weakly correlated with Y. $X3$ and $X4$ appear to have a stronger linear (positive) relationship with Y.

(c) The correlation matrix of X

```
> cor(d)
```

```
          X1        X2        X3        X4
X1 1.0000000 0.1022689 0.1807692 0.3266632
X2 0.1022689 1.0000000 0.5190448 0.3967101
X3 0.1807692 0.5190448 1.0000000 0.7820385
X4 0.3266632 0.3967101 0.7820385 1.0000000
```

The variable pairs $X2$ and $X3$, and the pairs $X3$ and $X4$ appear to be strongly correlated and contributing towards multicollinearity in the data set.

(d)
```
> reg = lm(Y ~ X1 + X2 + X3 + X4, d)
> summary(reg)

Call:
lm(formula = Y ~ X1 + X2 + X3 + X4, data = d)

Residuals:
    Min      1Q  Median      3Q     Max
-5.9779 -3.4506  0.0941  2.4749  5.9959

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -124.38182    9.94106 -12.512 6.48e-11 ***
X1             0.29573    0.04397   6.725 1.52e-06 ***
X2             0.04829    0.05662   0.853  0.40383
X3             1.30601    0.16409   7.959 1.26e-07 ***
X4             0.51982    0.13194   3.940  0.00081 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 4.099 on 20 degrees of freedom
Multiple R-squared:  0.9629,Adjusted R-squared:  0.9555
F-statistic: 129.7 on 4 and 20 DF,  p-value: 5.262e-14
```

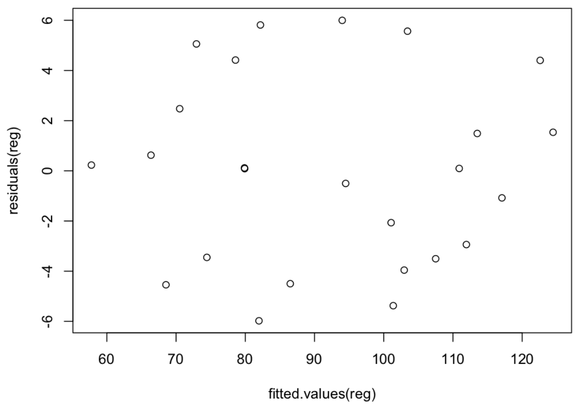
(e) Variance inflation factors

```
> library(car)
> vif(reg)
      X1       X2       X3       X4
1.138043 1.369512 3.016549 2.834776
```

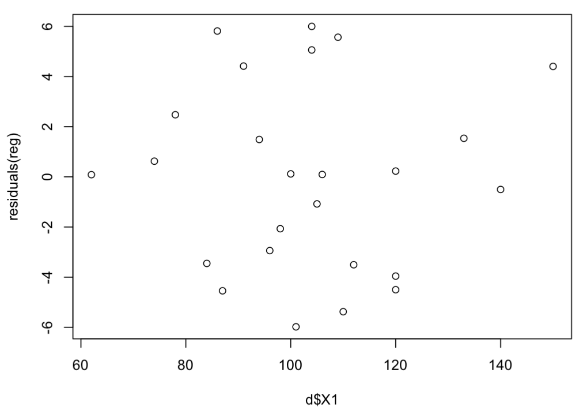
The values suggest there is some multicollinearity (but not very serious since all values are < 10).

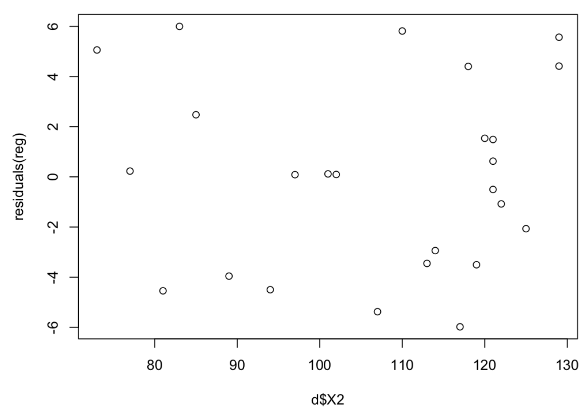
(f)
```
> plot(residuals(reg), fitted.values(reg))
> plot(residuals(reg), d$X1)
> plot(residuals(reg), d$X2)
> plot(residuals(reg), d$X3)
> plot(residuals(reg), d$X4)
```

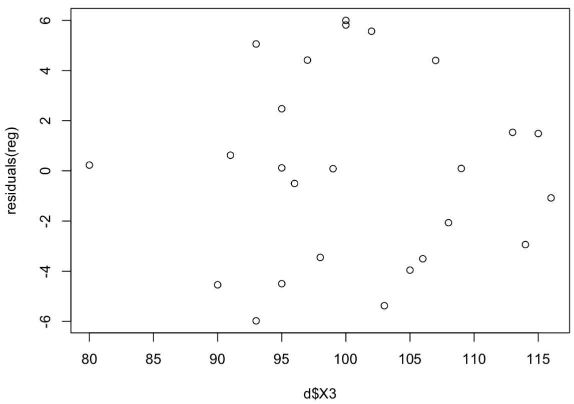
Plots of residuals vs fitted values:

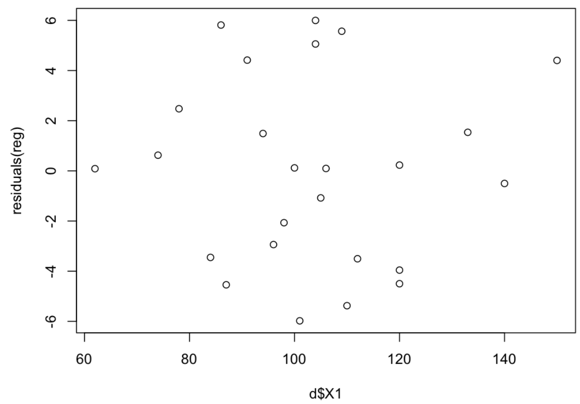


Plots of residuals vs independent variables:





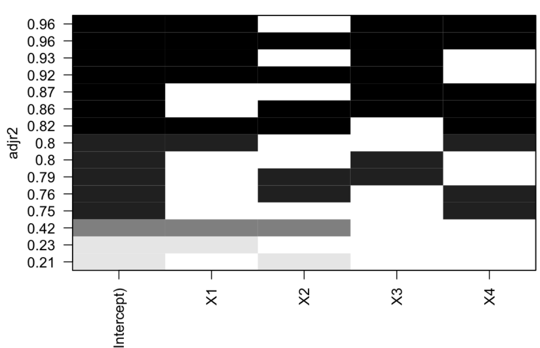




The above plots do not suggest any changes to the model.

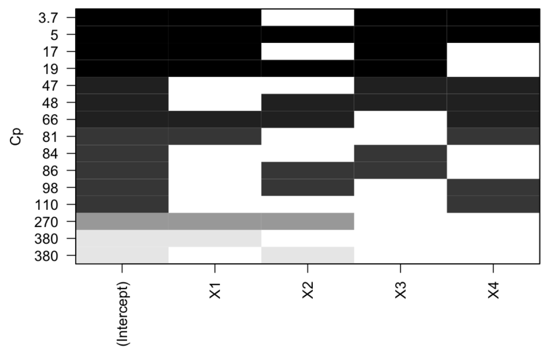
(g)
```
> library(leaps)
> subsets = regsubsets(Y~., data, nbest=16)
> plot(subsets, 'adjr2')
> plot(subsets, 'Cp')
> plot(subsets, 'r2')
```

Four best models by adjusted R^2:



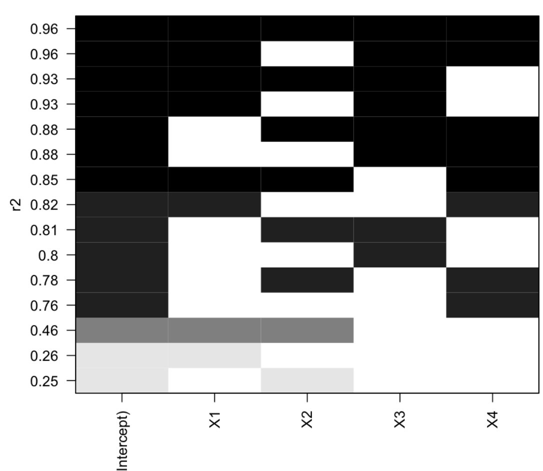
$$Y \sim X_1 + X_3 + X4$$
$$Y \sim X_1 + X_2 + X_3 + X_4$$
$$Y \sim X_1 + X_3$$
$$Y \sim X_1 + X_2 + X_3$$

Four best models by C_p:



$$
\begin{aligned}
&Y \sim X_1 + X_3 + X4 \\
&Y \sim X_1 + X_2 + X_3 + X_4 \\
&Y \sim X_1 + X_3 \\
&Y \sim X_1 + X_2 + X_3
\end{aligned}
$$

Four best models by unadjusted R^2:



$$Y \sim X_1 + X_2 + X_3 + X4$$
$$Y \sim X_1 + X_3 + X_4$$
$$Y \sim X_1 + X_2 + X_3$$
$$Y \sim X_1 + X_3$$

The top four choices based on adjusted R^2 and C_p criteria are exactly the same.

With unadjusted R^2, the best choice is always the model that includes all the variables since adding variables only improves the R^2. We also observe that a strict subset of a model is always ranked below it.

(h) Since the best subset models have very little difference in adjusted R^2, we select the model in that set that has the smallest p.

This leads us to selecting the model

$$Y \sim X_1 + X_3$$

(i) The best model using forward stepwise regression is $Y \sim X_1 + X_2 + X_3$

```
> step(lm(Y~1, data=d), direction="forward", scope=~X1 + X2 + X3 + X4)
Start:  AIC=149.3
Y ~ 1

       Df Sum of Sq    RSS    AIC
+ X3    1    7286.0 1768.0 110.47
+ X4    1    6843.3 2210.7 116.06
+ X1    1    2395.9 6658.1 143.62
+ X2    1    2236.5 6817.5 144.21
<none>              9054.0 149.30

Step:  AIC=110.47
Y ~ X3

       Df Sum of Sq     RSS     AIC
+ X1    1   1161.37  606.66  85.727
+ X4    1    656.71 1111.31 100.861
<none>              1768.02 110.469
+ X2    1     12.21 1755.81 112.295

Step:  AIC=85.73
Y ~ X3 + X1
```

```
       Df Sum of Sq    RSS    AIC
+ X4    1   258.460 348.20 73.847
<none>              606.66 85.727
+ X2    1     9.937 596.72 87.314

Step:  AIC=73.85
Y ~ X3 + X1 + X4

       Df Sum of Sq    RSS    AIC
<none>              348.20 73.847
+ X2    1     12.22 335.98 74.954

Call:
lm(formula = Y ~ X3 + X1 + X4, data = d)

Coefficients:
(Intercept)           X3           X1           X4
  -124.2000       1.3570       0.2963       0.5174
```

(j) The model selected by adjusted R^2 and forward stepwise regression are the same: $Y \sim X_1 + X_3 + X_4$

(k) The PRESS statistic for the reduced model is 760.1, whereas SSE is 606.65. This suggests that MSE is a slight under-estimator of the predictive error of the fitted model, and can be used as an indicator of it's predictive ability.

```
> regn = lm(Y~X1 + X3, d)
> SSE = sum(regn$residuals ^ 2)
> SSE
[1] 606.6574
> press(regn)
[1] 760.9744

>
```

(l) Correlation matrix of the validation data suggests a strong correlation between X_3 and X_4, just as in the original data, but a weaker correlation between X_2 and X_3 than the original data. X_1 and X_2 remain having low correlation with other independent variables as in the original data.

```
> dval = read.table('Jobval.txt', col.names=c('X1', 'X2', 'X3', 'X4', 'Y'))
> cor(dval)
```

```
           X1         X2         X3        X4         Y
X1 1.00000000 0.01100676 0.1817488 0.3176931 0.5429631
X2 0.01100676 1.00000000 0.3350669 0.2192434 0.3407173
X3 0.18174882 0.33506692 1.0000000 0.8562381 0.8775224
X4 0.31769314 0.21924340 0.8562381 1.0000000 0.8884018
Y  0.54296308 0.34071728 0.8775224 0.8884018 1.0000000
```

(m) Coefficients, standard error, mean squared error, and coefficient of multiple determination from original (model building) data set:

```
Coefficients:
              Estimate Std. Error
(Intercept) -127.59569   12.68526
X1             0.34846    0.05369
X3             1.82321    0.12307

> mean(regn$residuals ^ 2)
[1] 24.2663
> summary(regn)$r.squared
[1] 0.9329956
```

Coefficients, standard error, mean squared error, and coefficient of multiple determination from the validation data set:

```
Coefficients:
              Estimate Std. Error
(Intercept) -127.58441   13.53919
X1             0.34847    0.05317
X3             1.81852    0.13661

> mean(regval$residuals ^ 2)
[1] 23.07889
> summary(regval)$r.squared
[1] 0.9221217
```

The estimates from both the models appear reasonably similar.

(n)
```
> mean( (predict(regn, dval) - dval$Y) ^ 2)
[1] 23.28036
> mean(regn$residuals ^ 2)
[1] 24.2663
```

The mean squared predictive error and mean squared error are very similar. This does not suggest a high bias problem, and is consistent with the observations in (k).

(o)

```
> dall = rbind(d, dval)
> regall = lm(Y~X1 + X3, dall)
> summary(regall)
...
Coefficients:
              Estimate Std. Error
(Intercept) -128.08172    8.81751
X1             0.34900    0.03655
X3             1.82529    0.08747
...
```

We observe that the standard errors have reduced significantly upon using the entire data, while the estimates are approximately the same.