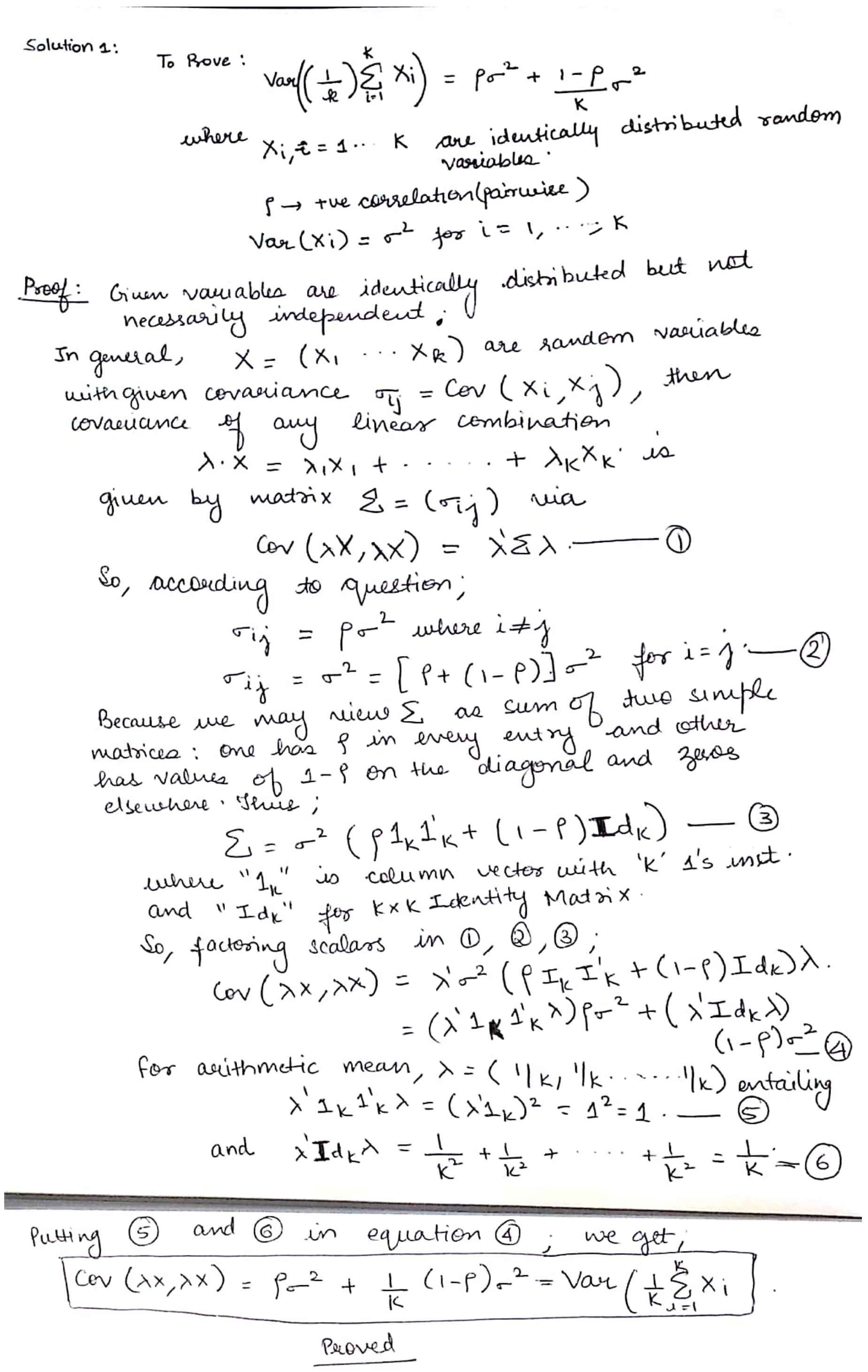
Problem Set 2 **Harshita Jhavar**

The Elements of Statistical Learning, WS 2017/18 Matriculation Number: 2566267

Prof. Dr. Dr. Thomas Lengauer s8hajhav@stud.uni-saarland.de

TA: Michael Scherer Due Date: November 15. 2017, 10:00 a.m

Solution 1 :(T, 8 Points)



Solution 1:

To Prove : $Var\left(\left(\frac{1}{k}\right)\sum_{i=1}^{k} x_i\right) = \rho\sigma^2 + \frac{1-\rho}{k}\sigma^2$

where $x_i, i = 1 \cdots K$ are identically distributed random variables.

$\rho \to$ +ve correlation(pairwise)

$Var(x_i) = \sigma^2$ for $i = 1, \cdots, K$

Proof: Given variables are identically distributed but not necessarily independent.

In general, $X = (x_1 \cdots x_k)$ are random variables with given covariance $\sigma_{ij} = Cov(x_i, x_j)$, then covariance of any linear combination

$\lambda \cdot X = \lambda_1 x_1 + \cdots + \lambda_K x_K$ is

given by matrix $\Sigma = (\sigma_{ij})$ via

$Cov(\lambda X, \lambda X) = \lambda' \Sigma \lambda$ —— ①

So, according to question;

$\sigma_{ij} = \rho\sigma^2$ where $i \neq j$

$\sigma_{ij} = \sigma^2 = [\rho + (1-\rho)]\sigma^2$ for $i = j$ —— ②

Because we may view Σ as sum of two simple matrices: one has ρ in every entry and other has values of $1-\rho$ on the diagonal and zeros elsewhere. Thus;

$\Sigma = \sigma^2(\rho 1_K 1'_K + (1-\rho) Id_K)$ —— ③

where "1_K" is column vector with 'K' 1's inst. and "Id_K" for K×K Identity Matrix.

So, factoring scalars in ①, ②, ③;

$Cov(\lambda x, \lambda x) = \lambda'\sigma^2(\rho I_K I'_K + (1-\rho)Id_K)\lambda$.

$= (\lambda' 1_K 1'_K \lambda)\rho\sigma^2 + (\lambda' Id_K \lambda)(1-\rho)\sigma^2$ ④

for arithmetic mean, $\lambda = (1/K, 1/K \cdots 1/K)$ entailing

$\lambda' 1_K 1'_K \lambda = (\lambda' 1_K)^2 = 1^2 = 1$. —— ⑤

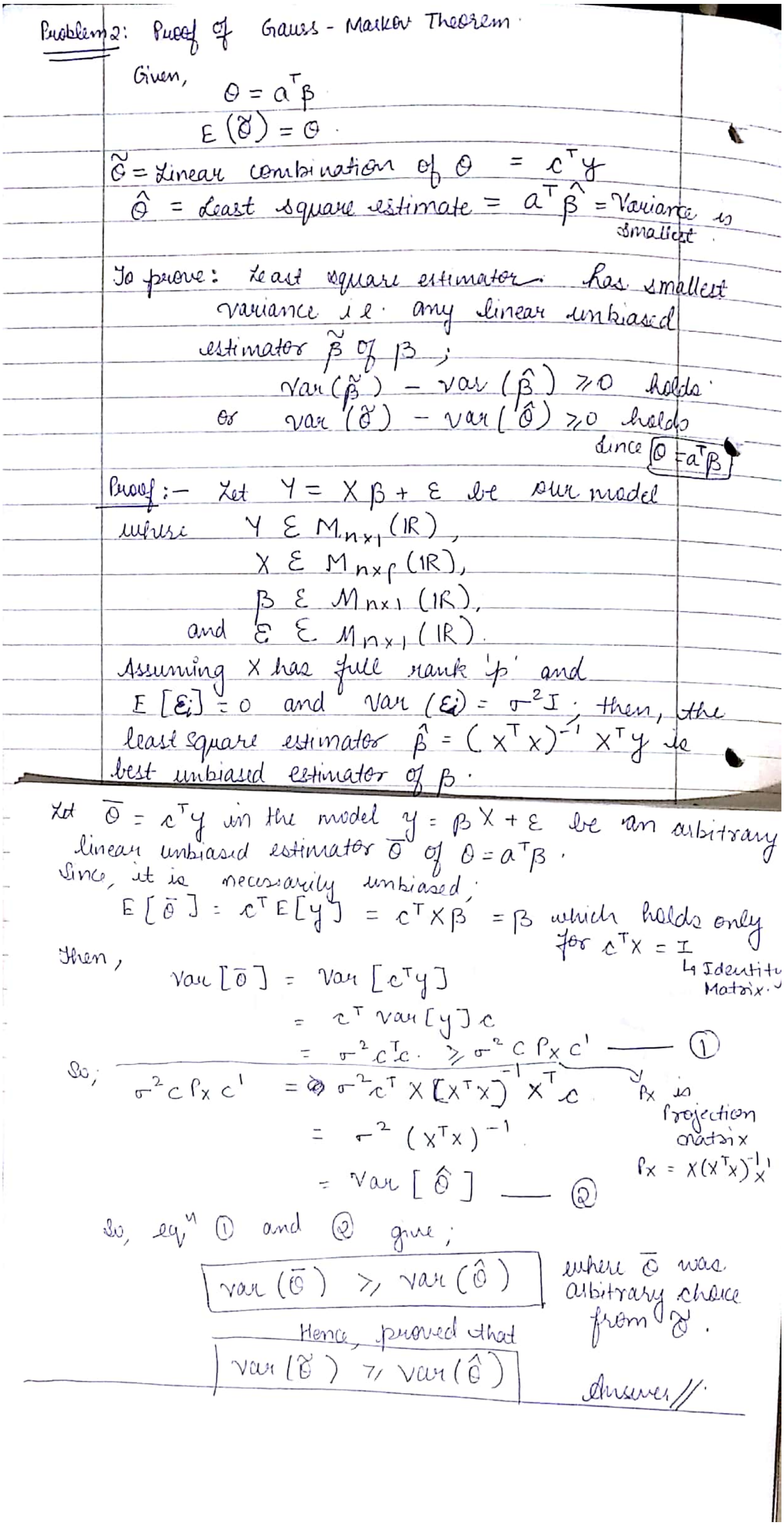
and $\lambda' Id_K \lambda = \frac{1}{K^2} + \frac{1}{K^2} + \cdots + \frac{1}{K^2} = \frac{1}{K}$ — ⑥

Putting ⑤ and ⑥ in equation ④ ; we get,

$$\boxed{Cov(\lambda x, \lambda x) = \rho\sigma^2 + \frac{1}{K}(1-\rho)\sigma^2 = Var\left(\frac{1}{K}\sum_{i=1}^{K} x_i\right)}.$$

Proved

Solution 2 :(T, 12 Points)



Problem 2: Proof of Gauss - Markov Theorem.

Given,

$\theta = a^T\beta$

$E(\tilde{\theta}) = \theta$.

$\tilde{\theta}$ = Linear combination of θ $= c^T y$

$\hat{\theta}$ = Least square estimate $= a^T\hat{\beta}$ = Variance is smallest.

To prove: Least square estimator has smallest variance i.e. any linear unbiased estimator $\tilde{\beta}$ of β;

$var(\tilde{\beta}) - var(\hat{\beta}) \geq 0$ holds.

or $var(\tilde{\theta}) - var(\hat{\theta}) \geq 0$ holds

since $\boxed{\theta = a^T\beta}$

Proof:- Let $Y = X\beta + \varepsilon$ be our model

where $Y \in M_{n\times 1}(\mathbb{R})$,

$X \in M_{n\times p}(\mathbb{R})$,

$\beta \in M_{n\times 1}(\mathbb{R})$,

and $\varepsilon \in M_{n\times 1}(\mathbb{R})$.

Assuming X has full rank 'p' and $E[\varepsilon_i] = 0$ and $Var(\varepsilon_i) = \sigma^2 I$. then, the least square estimator $\hat{\beta} = (X^T X)^{-1} X^T y$ is best unbiased estimator of β.

Let $\bar{\theta} = c^T y$ in the model $y = \beta X + \varepsilon$ be an arbitrary linear unbiased estimator $\bar{\theta}$ of $\theta = a^T\beta$.

Since, it is necessarily unbiased;

$E[\bar{\theta}] = c^T E[y] = c^T X\beta = \beta$ which holds only for $c^T X = I$

↳ Identity Matrix.

Then,

$$Var[\bar{\theta}] = Var[c^T y]$$

$$= c^T var[y] c$$

$$= \sigma^2 c^T c. \geq \sigma^2 c P_x c' \quad \text{——} \; ①$$

So;

$$\sigma^2 c P_x c' = \sigma^2 c^T X [X^T X]^{-1} X^T c$$

P_x is Projection matrix

$P_x = X(X^T X)^{-1} X'$

$$= \sigma^2 (X^T X)^{-1}.$$

$$= Var[\hat{\theta}] \quad \text{——} \; ②$$

So, eq^n ① and ② give;

$$\boxed{var(\bar{\theta}) \geq var(\hat{\theta})}$$

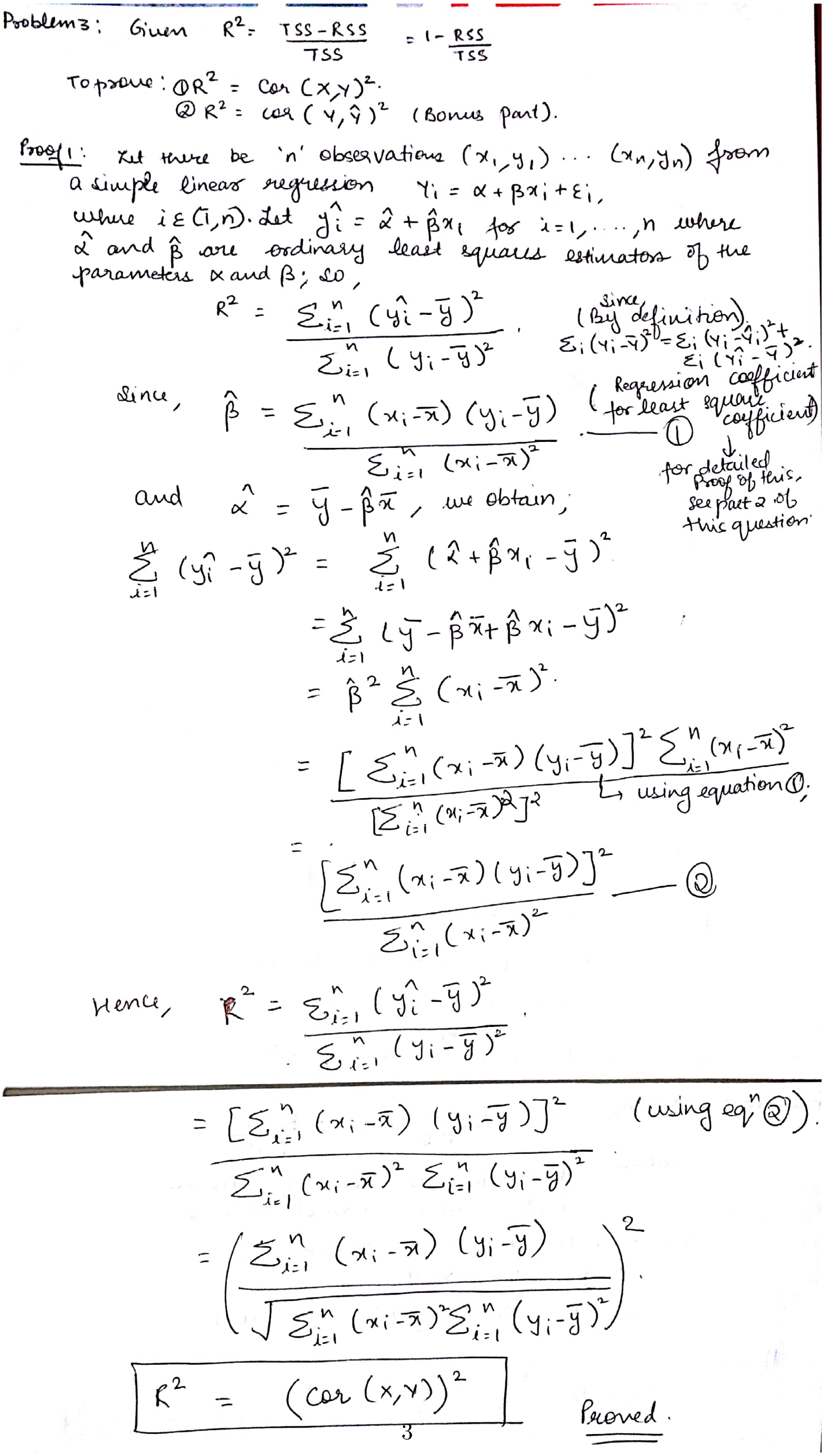
where $\bar{\theta}$ was arbitrary choice from $\tilde{\theta}$.

Hence, proved that

$$\boxed{var(\tilde{\theta}) \geq var(\hat{\theta})}$$

Answer//.

Solution 3:(T, 10 Points+Bonus)



Problem3: Given $R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS}$

To prove: ① $R^2 = Cor(X,Y)^2$.

② $R^2 = cor(Y, \hat{Y})^2$ (Bonus part).

Proof 1: Let there be 'n' observations $(x_1, y_1) \cdots (x_n, y_n)$ from a simple linear regression $Y_i = \alpha + \beta x_i + \varepsilon_i$, where $i \in (1, n)$. Let $\hat{y_i} = \hat{\alpha} + \hat{\beta} x_i$ for $i = 1, \dots, n$ where $\hat{\alpha}$ and $\hat{\beta}$ are ordinary least squares estimators of the parameters α and β; so,

$$R^2 = \frac{\Sigma_{i=1}^{n} (\hat{y_i} - \bar{y})^2}{\Sigma_{i=1}^{n} (y_i - \bar{y})^2}.$$

(Since By definition), $\Sigma_i (Y_i - \bar{Y})^2 = \Sigma_i (Y_i - \hat{Y}_i)^2 + \Sigma_i (\hat{Y}_i - \bar{Y})^2$.

Since, $$\hat{\beta} = \frac{\Sigma_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\Sigma_{i=1}^{n} (x_i - \bar{x})^2} \quad \text{—— ①}$$

(Regression coefficient for least square coefficient)

↓ for detailed Proof of this, see part 2 of this question.

and $\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$, we obtain,

$$\sum_{i=1}^{n} (\hat{y_i} - \bar{y})^2 = \sum_{i=1}^{n} (\hat{\alpha} + \hat{\beta} x_i - \bar{y})^2$$

$$= \sum_{i=1}^{n} (\bar{y} - \hat{\beta}\bar{x} + \hat{\beta} x_i - \bar{y})^2$$

$$= \hat{\beta}^2 \sum_{i=1}^{n} (x_i - \bar{x})^2.$$

$$= \frac{[\Sigma_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})]^2 \Sigma_{i=1}^{n} (x_i - \bar{x})^2}{[\Sigma_{i=1}^{n} (x_i - \bar{x})^2]^2}$$

↳ using equation ①,

$$= \frac{[\Sigma_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})]^2}{\Sigma_{i=1}^{n} (x_i - \bar{x})^2} \quad \text{—— ②}$$

Hence, $$R^2 = \frac{\Sigma_{i=1}^{n} (\hat{y_i} - \bar{y})^2}{\Sigma_{i=1}^{n} (y_i - \bar{y})^2}.$$

$$= \frac{[\Sigma_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})]^2}{\Sigma_{i=1}^{n} (x_i - \bar{x})^2 \, \Sigma_{i=1}^{n} (y_i - \bar{y})^2} \quad (\text{using eq}^n \text{ ②}).$$

$$= \left(\frac{\Sigma_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\Sigma_{i=1}^{n} (x_i - \bar{x})^2 \Sigma_{i=1}^{n} (y_i - \bar{y})^2}} \right)^2.$$

$$\boxed{R^2 = (Cor(X,Y))^2}$$

Proved.

Proof :- ② To prove: $R^2 = cor(Y, \hat{Y})^2$.

Since, $(y_i - \bar{y}) = (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})$.

Squaring both sides and sum over all i;

$$\Rightarrow \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n} 2(\hat{y}_i - \bar{y})(y_i - \hat{y}_i).$$

Considering only last term and proving it zero;

$$\hat{y}_i = \hat{\alpha} + \hat{b}x_i$$

$$\bar{y} = \hat{\alpha} + \hat{b}\bar{x}.$$

$$\boxed{\hat{b} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}} \quad \text{— ①}$$

So;

$$\overline{y_i} - \bar{y} = \hat{b}(x_i - \bar{x})$$

$$\Rightarrow y_i - \hat{y}_i = (y_i - \bar{y}) - (\hat{y}_i - \bar{y})$$

$$= (y_i - \bar{y}) - \hat{b}(x_i - \bar{x}).$$

Therefore;

$$\sum_{i=1}^{n} 2(\hat{y}_i - \bar{y})(y_i - \hat{y}_i)$$

$$= 2\hat{b}\sum_{i=1}^{n}(x_i - \bar{x})\left((y_i - \bar{y}_i) - \hat{b}(x_i - \bar{x})\right)$$

$$= 2\hat{b}\left(\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) - \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{j=1}^{n}(x_j - \bar{x})(y_j - \bar{y})}{\sum_{j=1}^{n}(x_j - \bar{x})^2}\right)$$

$$= 2\hat{b}(0)$$

$$= 0.$$

Thus;

$$\boxed{\underbrace{\sum_{i=1}^{n}(y_i - \bar{y})^2}_{TSS} = \overbrace{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}^{RSS} + \overbrace{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}^{ESS}.} \quad \text{— ②}$$

Since; $RSS = \|y - \hat{y}\|_2^2$, $TSS = \|y - \bar{y}\|_2^2$ and $ESS = \|\hat{y} - \bar{y}\|$

So, $\hat{y}^T\hat{y} = y^T X(X^TX)^{-1}X^TX(X^TX)^{-1}X^Ty$
$= y^TX(X^TX)^{-1}X^Ty = y^T\hat{y}$. (written in form of matrix to ease calculation)

So; $y^T\bar{y} = \hat{y}^T\bar{y}$.

Thus; ~~RSS~~ R^2~~co~~= we conclude here that for least square regression model, the sample covariance between $\hat{Y_i}$ and $Y_i - \hat{Y_i}$ is zero resulting in

$$R(Y,\hat{Y}) = \frac{\Sigma_i (Y_i - \bar{Y})(\hat{Y_i} - \bar{Y})}{\sqrt{\Sigma_i (Y_i - \bar{Y})^2 \, \Sigma_i (\hat{Y_i} - \bar{Y})^2}}.$$

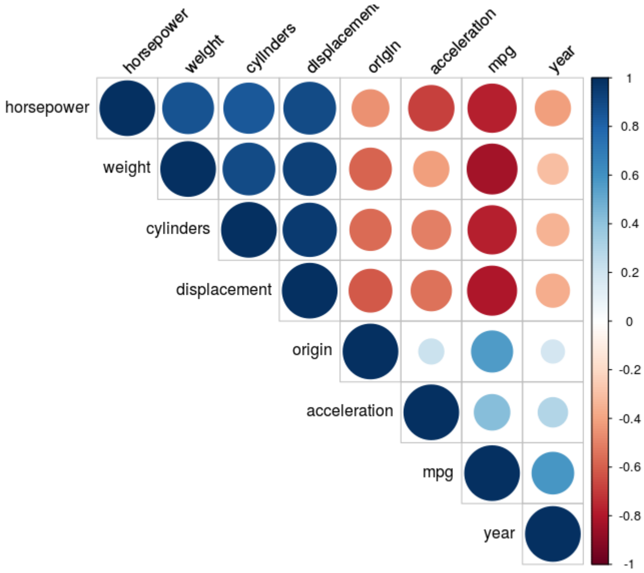
$$\Rightarrow \boxed{R^2 = cov(Y,\hat{Y})^2}.$$

Hence proved.

Solution 4: (P, 20 Points)

Part A:

A representation of the correlation matrix can be seen in the plot below:

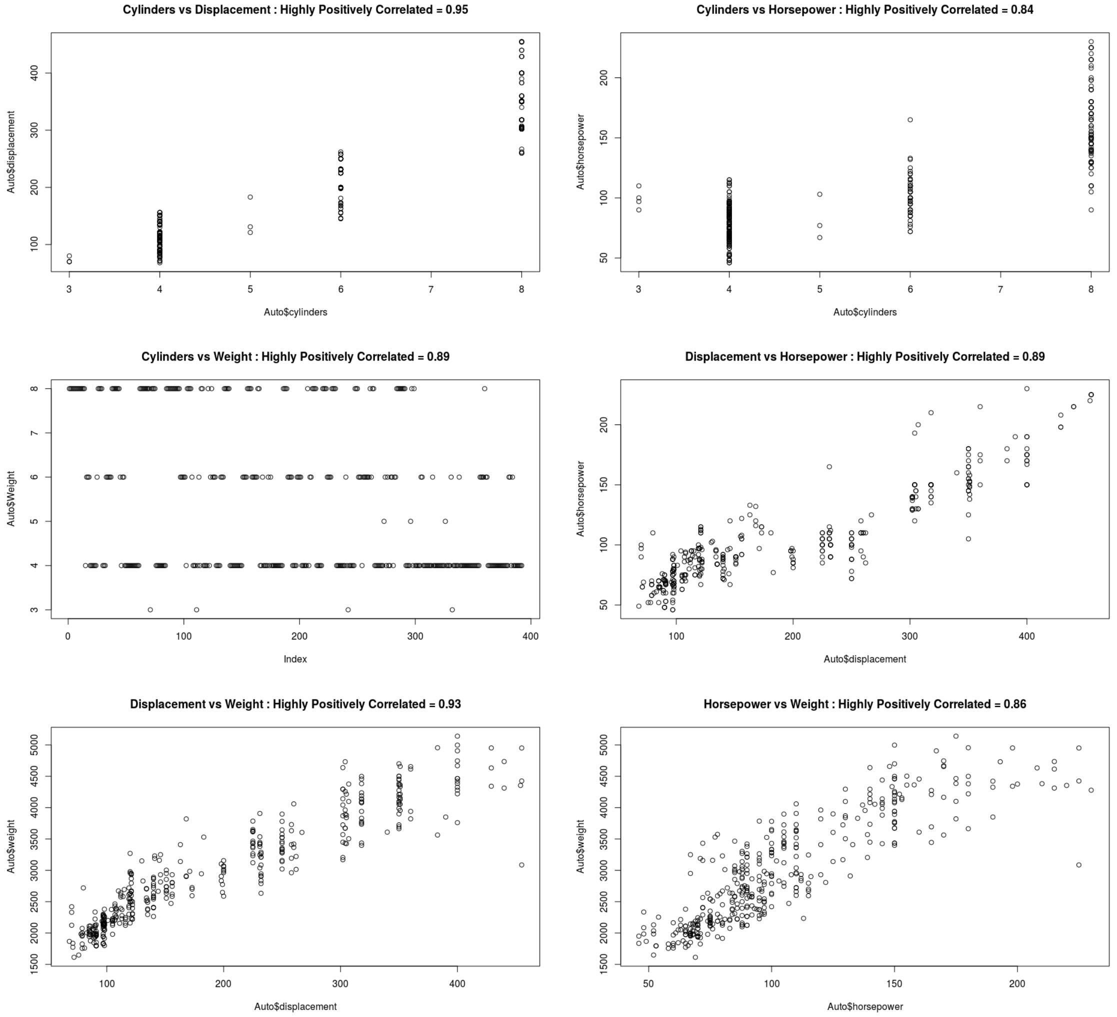


Observations: The observations from the correlation matrix values and the plot above are as follows:

- The value of horsepower and value of weight, cylinders and displacement will increase together strongly while for mpg and accelerator, it will decrease. This is because Horsepower is highly positively correlated with weight, cylinder and displacement while extremely negatively correlated with mpg.

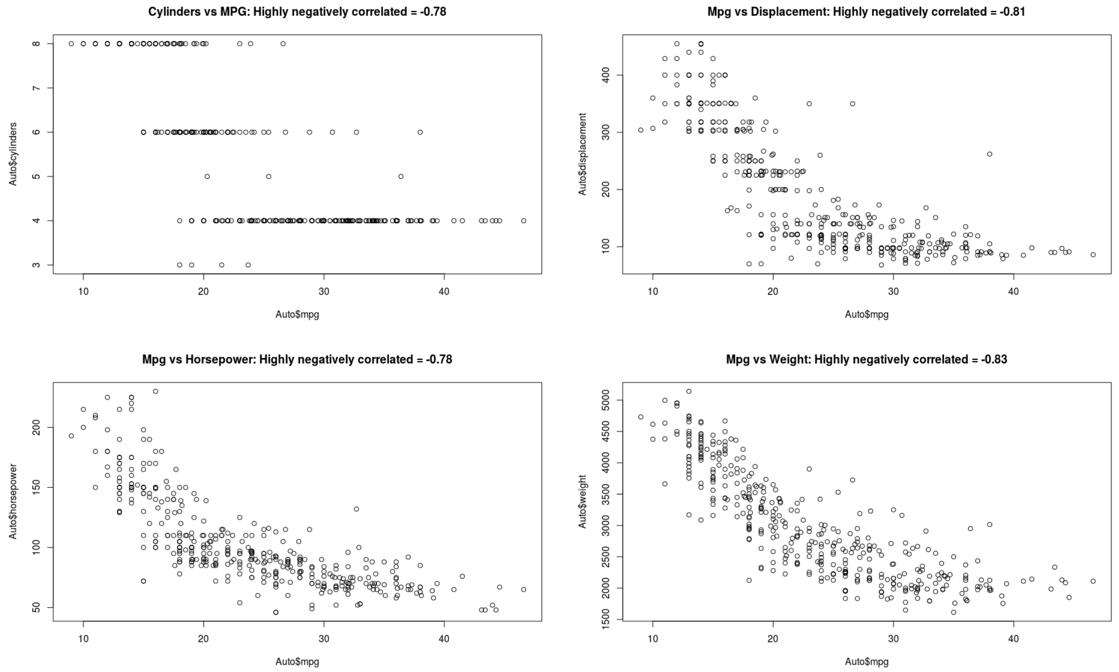
- Weight is highly positively correlated with cylinders and displacement but is highly negatively correlated with mpg and very less correlation with year. So, the weight of the automobile increases together with the value of cylinder and displacement with a very close association while mpg decreases with a close association again.
- Value of displacement increase and mpg's value decreases together with very close association.
- Origin has very less association with year and origin.
- Acceleration has very less association with year.

Part b:
Scatter plot between the variables that are highly positively correlated ((assumed)correlation value ≥ 0.7) is as follows:



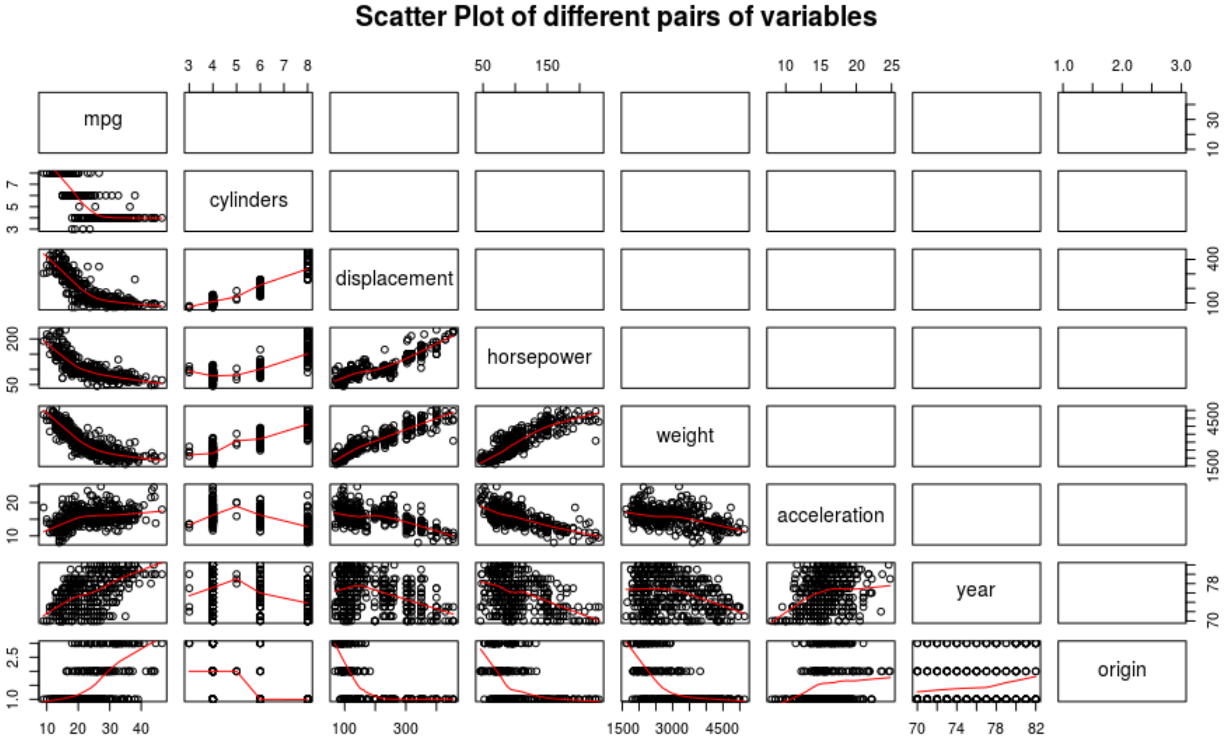
From the above plots it is clear that among highly positively correlated variable pairs, only 3 variable pairs which are Displacement vs Horsepower, Displacement vs Weight and Horsepower vs weight follow an almost linear relationship.

Scatter plot between the variables that are highly negatively correlated ((assumed)correlation value $\leq$ -0.7) is as follows:



From the above plots it is clear that among highly negatively correlated variable pairs, only 3 variable pairs which are MPG vs Displacement, MPG vs Horsepower and MPG vs Weight follow an almost negative linear relationship.

Scatter Plotting all the variables together in the plot below to get an overall understanding of the association of the relationship of the different variables in one plot. We can see that based on our analysis above, there are 6 variable pairs which are highly positively correlated and 4 variable pairs which are highly negativvely correlated. So, while choosing features, I would play around these variable pairs so my model can incorporate these relationships in a similar fashion.



Part c:

Model 1: formula = mpg ∼ cylinders
Observation from the summary of the model: MPG and cylinder have a statistically significant relationship because the F-statistic value is much larger than 1, thus, there is a dependence in the two variables. The p value is very much low for the intercept and the predictor 'cylinder' with three stars alongside which signifies that the null hypothesis of assuming the two as independent can be rejected with strong evidence. The adjusted Rsquared value is 0.60 which implies a strong linear relationship between the cylinder and the mpg.

Model 2: formula = mpg ∼ displacement
Observation from the model summary: Mpg and displacemnt has a statistically significant relationship because the f-statistic value is much greater than 1 which is 718.70. Also, the p-value is again much far from zero, and is very small, so the null hypothesis can be rejected and there is a dependence in the two variables. The adjusted Rsquared value is 0.64 implies a strong linear relationship between the displacement and the mpg.

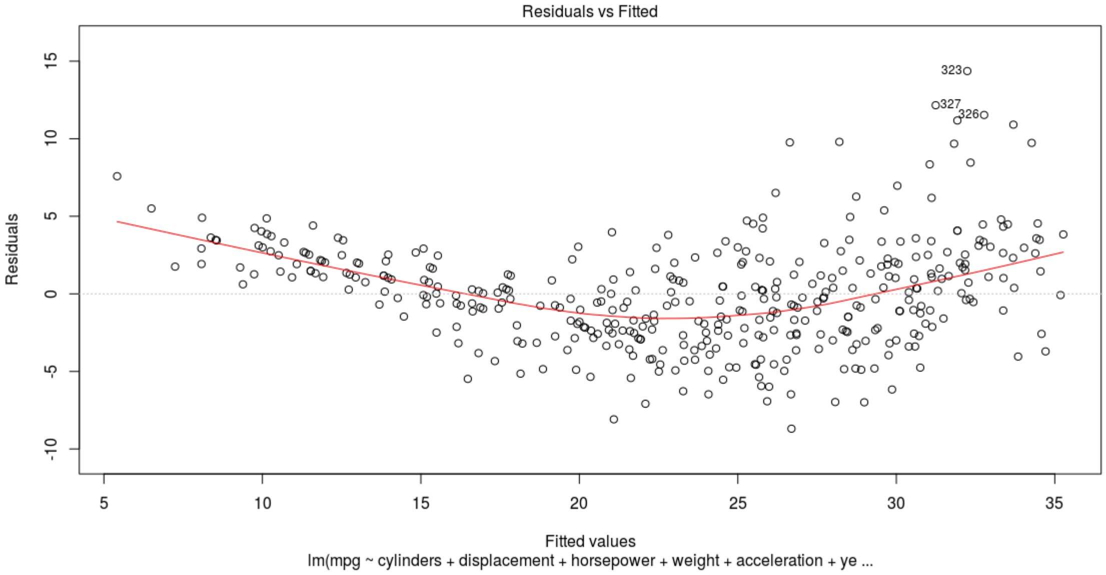
Model 3: formula = mpg ∼ horsepower Observation from the model summary: Again, MPG and horsepower have significant dependency between them as the F-statistic value is 599.7 which is much larger than 1 and p-value is very small, so null hypothesis can be easily rejected. The adjuested R-squared value is 0.6 which shows that there is strong linear relationship between the two variables.

Model 4: formul = mpg ∼ year
Observation from the model summary: Here, again the p-value is very small, so the null value can be rejected. However, the value of F-statistic observed here is the least among all the pre-

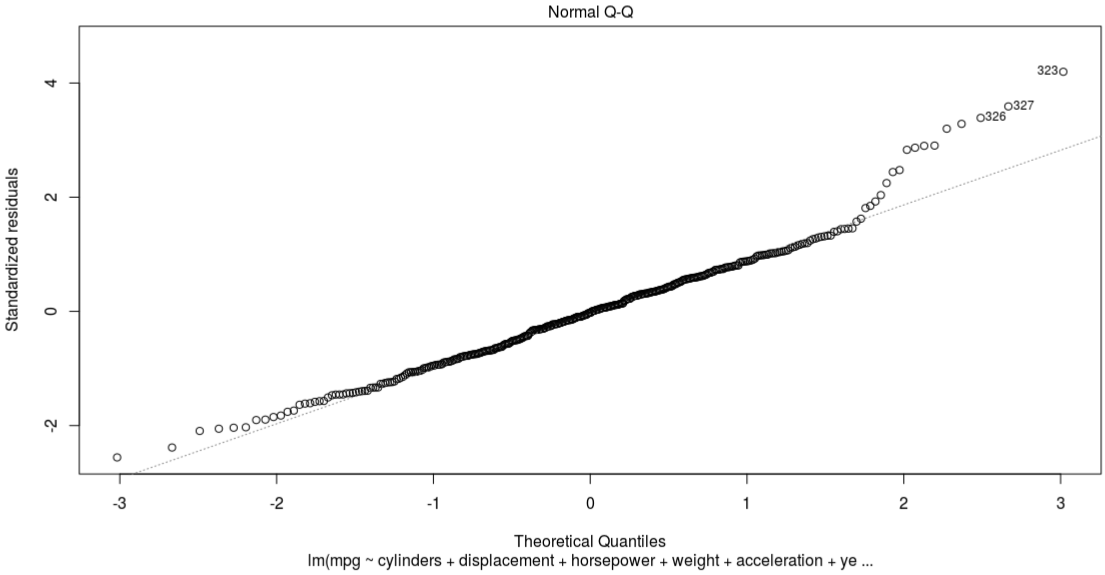
vious model, but still it is much greater than 1 also, the adjusted R-squared value is 0.33 which tells us that there is no strong linear relationship between the two models. So, this model is not a good fit as compared to the above three models.

Part d:
The full model is formula = mpg ∼ cylinders + displacement + horsepower + weight + acceleration + year. Here, in comparison with the summary generated with the above models, the adjusted Rsquared value is highest (0.8) which implies a very strong relationship between the model and the variables and make it a good fit. However, if we consider the p-value of the variables, we can see in summary that only intercept, weight and year have smaller p-value hwere intercept p-value is almost zero. However, overall p-value of the model is much smaller than 0 and thus, the null hypothesis is rejected like that of the model in previous part.
Interestingly, the value of the standard error for each of the variables has become very small as compared to the models in the previous part. Thus, this model is better than the previous model.
The -ve sign of the coefficient tells us that as the value of the predictor increases, there is a drop in the value of the response while =ve implies that with the increase in the value of the predictor, there is an increase in the value of the response.

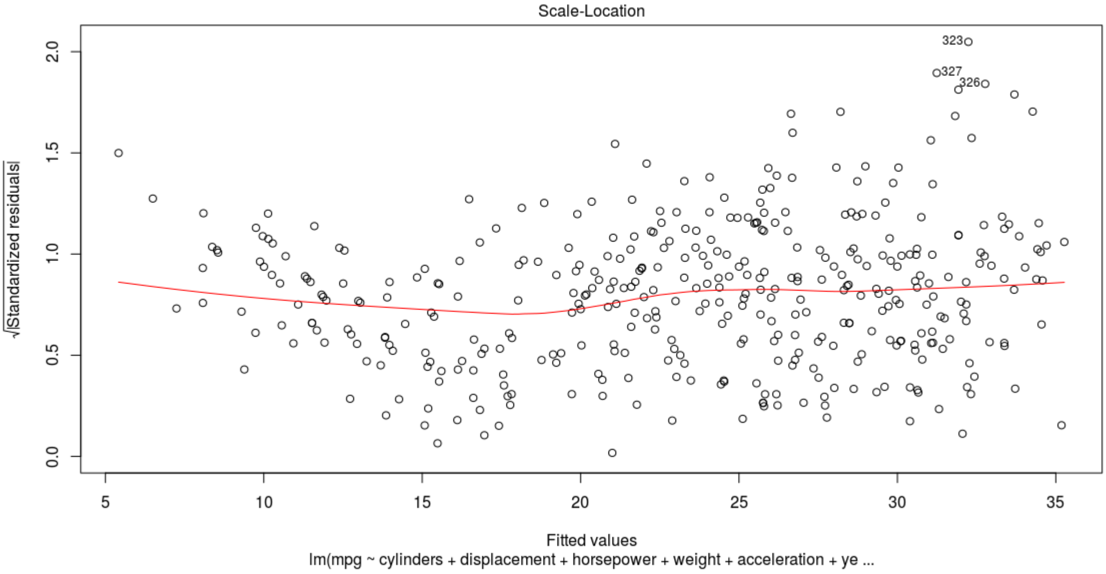
Part e:
The diagnostic plots of the model mpg ∼ cylinders + displacement + horsepower + weight + acceleration + year is as follows:



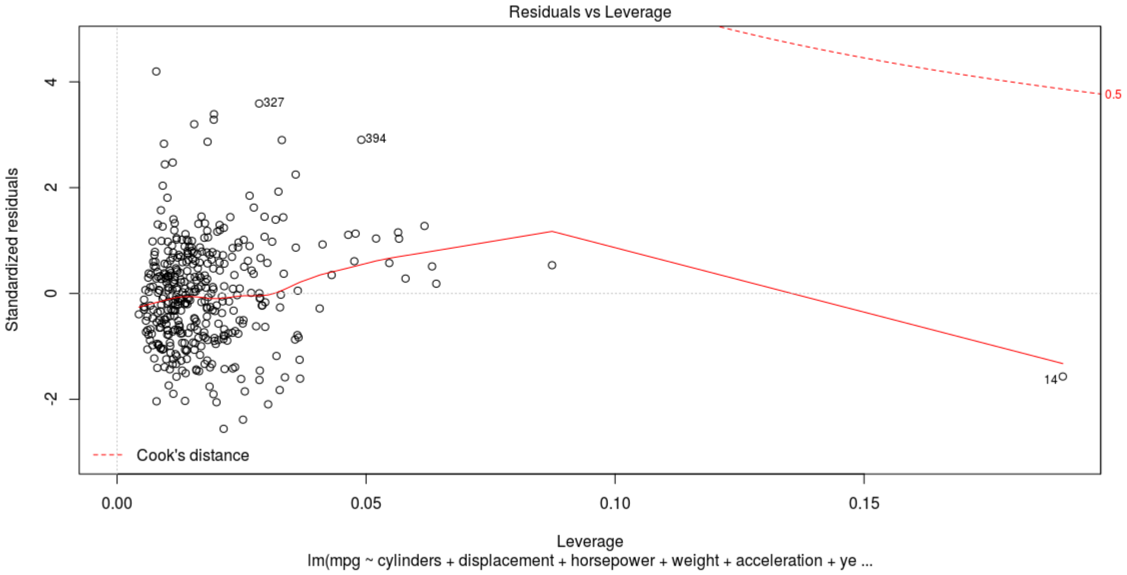
The above plot is the residual part. Ideally, the red plot should have been a straight line to support our assumption that regression fits with this model. However, here the red plot is a curve which leads us to conclude that some non-linearity has not been explained by this model. Also, there are many values away from the red plot especially for 30-35 which is also an indication of the outliers.



The above plot is the q-q plot which is almost a straight line except in the values after the second quantile in both the directions. This clearly is an indicative that there can be some outliers in the higher quantiles as most of the data which is there in the first two quantiles on either side of the mean has straight line q-q plot and thus, normalized resiual error values.



In the above plot,the residuals appear to spread wider along the x-axis which questions our assumptions that the residuals are equally spread along the range of the predictor if we think of this model as a good model.



In the above plot, almost all the observations are in the Cook's distance range as we can barely see the red plot in the lower right corner. However, 394, 14 and 327 have been identified as leverages. There is a very high chance that these are mere outliers as there are not many values in the vicinity of these points. Thus, overall, not a significant amount of leverage cases are obtained.

Part f:

Interaction linear model 1: mpg ∼ cylinders:weight + weight:year + year:cylinders + log(displacement)
Interaction linear model 2: mpg ∼ cylinders:weight + weight:year + year:cylinders + sqrt(displacement)
Interaction linear model 3: mpg ∼ cylinders:weight + weight:year + year:cylinders + (displacement^2)

Observations: When we compare the summary of each of these models, we conclude that the standard error in each of these models is low, the residual standard error is almost 4 for all three models. The adjusted R-squared value is almost similar and approximated to around 0.7 for each of these models which clearly indicates that there is a strong dependence in the model predictors. The p-value for each of these models is much far from 0 and are very low, so the null hypothesiscan be rejected.

However, for the different factor of the displacement variable, there is different level of significance for the different variables. For example: for log(displacement), weight:year is least significant with only one star against it while for other two models, it is not significant at all. Also, the residual standard error is least for interaction model 1 than model 2 or 3. Also, when I ran annova test, again model 1 had the least variance and proves to be a best model among the three.