Module 1 Assignment on Linear Regression

Jingwen Zhong// Graduate Student

2/18/2021

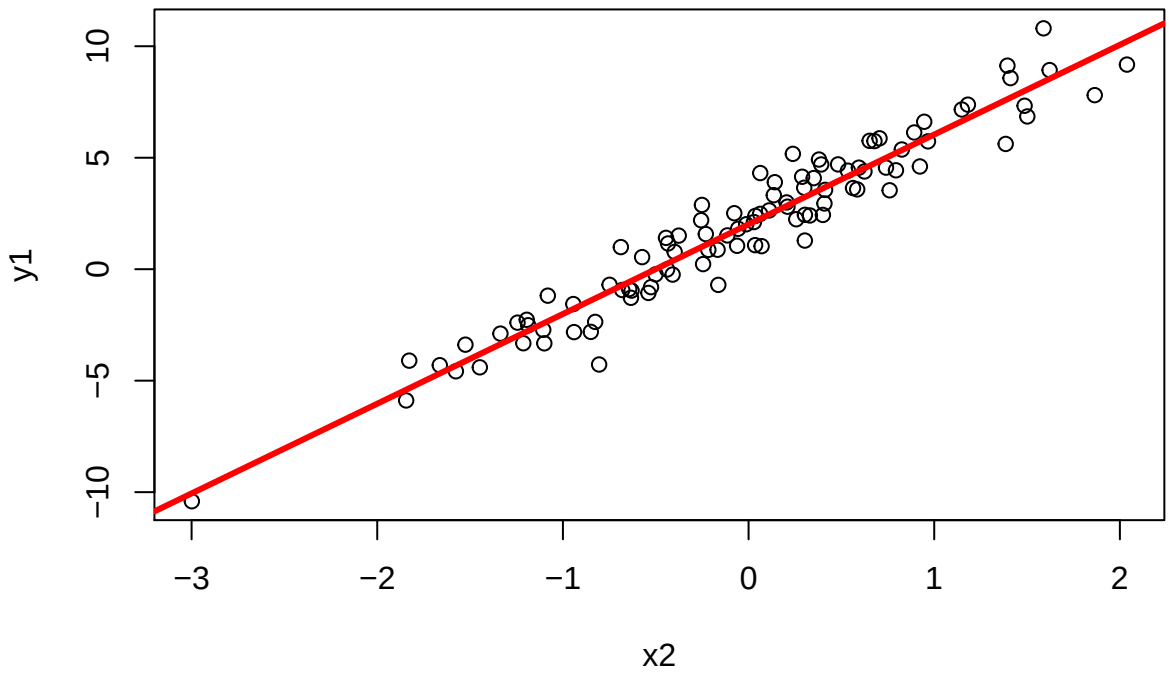
1) (*Concepts*)

Perform the following commands in R after you read the docs by running `help(runif)` and `help(rnorm)`:

The last lines correspond to creating two linear models (call Model 1 and Model 2, respectively) and their fitted results in which y1 and y2 are functions of some of the predictors x1, x2 and x3.

a. Fit a least squares (LS) regression for Model 1. Plot the response and the predictor. Use the abline() function to display the fitted line. What are the regression coefficient estimates? Report with standard errors and p-values in a table.

	coeffients	standard Error	P value
intercept	2.015434e+00	9.997083e-02	1.212630e-36
x2	4.024196e+00	1.101010e-01	6.619822e-59



---

b. Is the fitted Model 1 good? Do quality of model checck. Justify with appropriate metrics we covered.

```
##
## Call:
## lm(formula = y1 ~ x2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3.05182 -0.71368  0.00218  0.73846  2.39278
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.01543    0.09997   20.16   <2e-16 ***
## x2           4.02420    0.11010   36.55   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9992 on 98 degrees of freedom
## Multiple R-squared:  0.9317, Adjusted R-squared:  0.931
## F-statistic:  1336 on 1 and 98 DF,  p-value: < 2.2e-16
```

Yes, it is good, The R square value of this model is about 0.9317 and it is relatively large since R square

always lies between 0 and 1. An R2 statistic that is close to 1 indicates that a large proportion of the variability in the response has been explained by the regression.

---

c. Now fit a LS regression for Model 2. What are the regression coefficient estimates? Report them along with the standard errors and p-values. Are the predictors significantly contributing to the model? Explain.

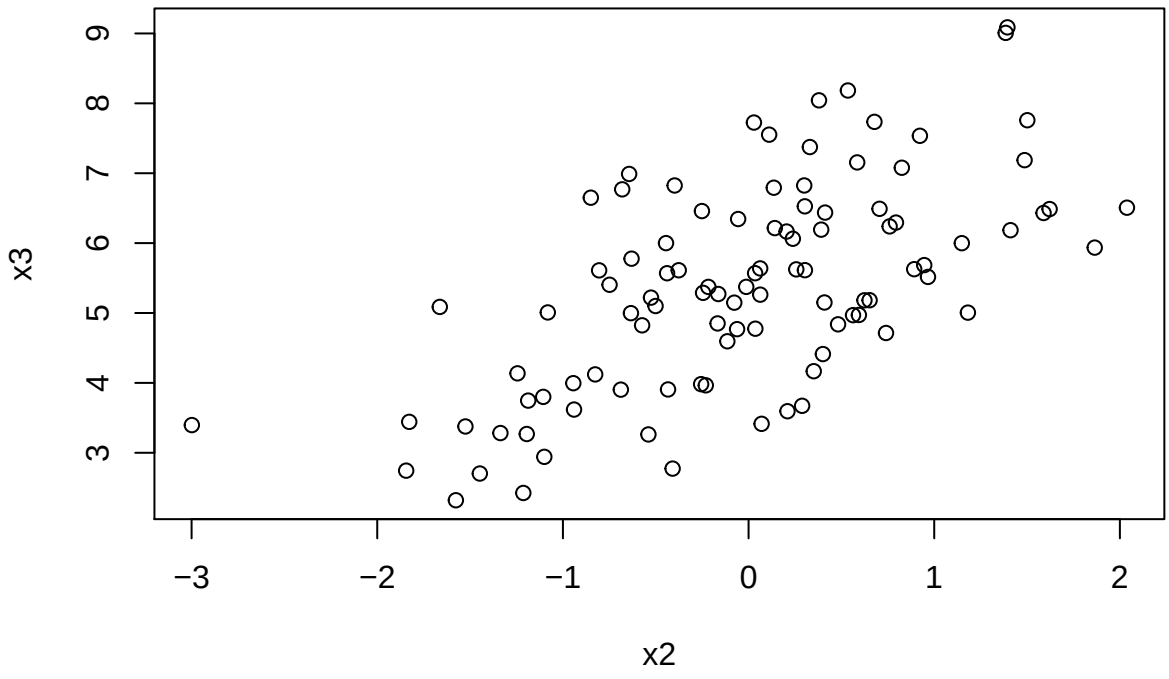
	coeffients	standard Error	P value
(intercept)	1.382320e+00	5.134321e-01	8.373220e-03
x1	3.471942e+00	3.846938e-01	1.847300e-14
x2	4.107206e+00	1.576846e-01	2.601754e-45
x3	5.063398e+00	9.601492e-02	1.087753e-72

Yes, the predictors are significantly contributing to the model, because the p value of each is small, smaller than 0.05.

---

d. What is the correlation between x2 and x3? Create a scatterplot displaying the relationship between the variables. Comment on the strength of the correlation.

```
## The Correlation between x2 and x3 is: 0.631347
```



The strength of the correlation is fairly strong, it means that when x2 increase, x3 tends to increase too.

---

e. What are the assumptions in fitted Model 2? List the four assumptions. Check each. Comment on each.

- There must be a linear relationship between the outcome variable and the independent variables.
- Multivariate Normality–Multiple regression assumes that the residuals are normally distributed.
- No Multicollinearity—Multiple regression assumes that the independent variables are not highly correlated with each other.
- Homoscedasticity–This assumption states that the variance of error terms are similar across the values of the independent variables.

It matches three of them, but not the third assumption, x_2 and x_3 is fairly highly correlated with each other.

---

f. Do you think adding the new predictors, x1 and x3, to Model 1 improved the results? Test it using ANOVA F method (use `anova(model1, model_added)`. Comment on the results.

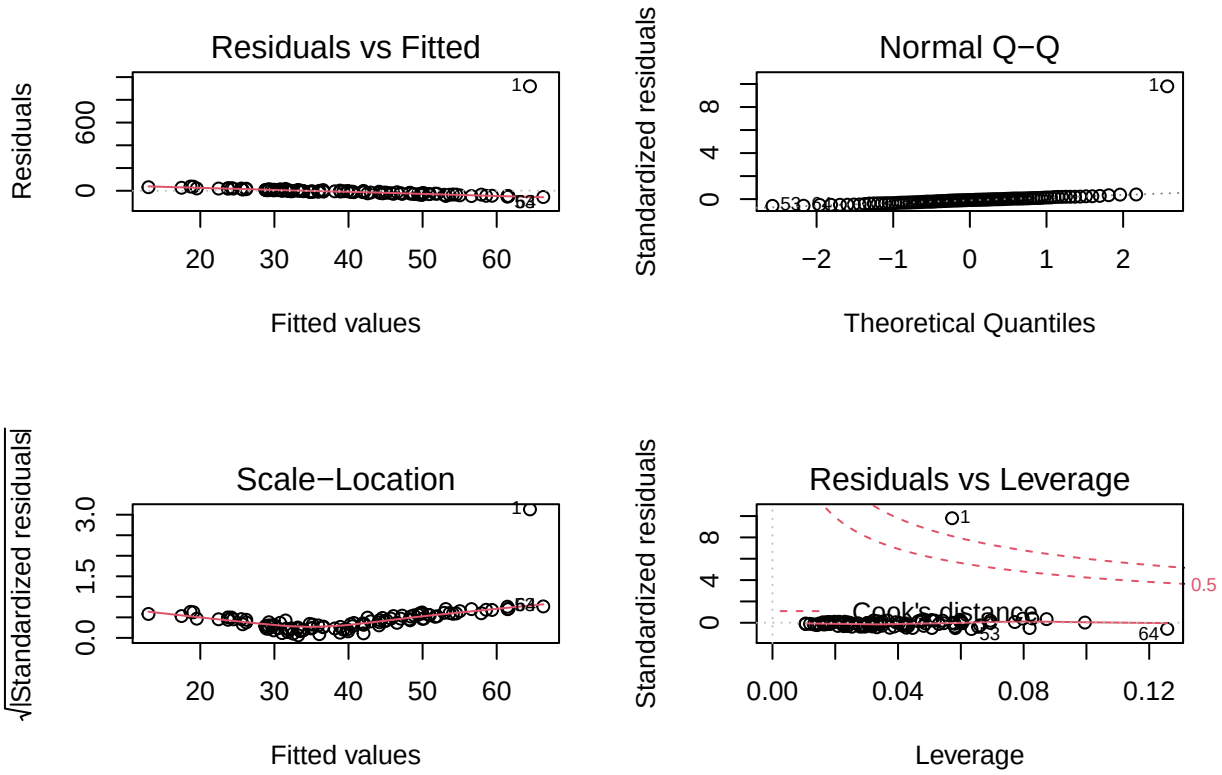
```
## Analysis of Variance Table
##
## Model 1: y1 ~ x2
## Model 2: y1 ~ x2 + x1 + x3
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     98 97.845
## 2     96 97.337  2   0.50743 0.2502 0.7791
```

The null hypothesis is that the two models fit the data equally well, and the alternative hypothesis is that the full model is superior. Here the F-statistic is 0.2502 and the associated p-value is relatively large. This provides evidence that the model containing the predictors x_1, x_2, x_3 is not superior to the model that only contains the predictor x_2.

---

g. Now suppose we corrupt one of the observations in y2: corrupt the first observation by adding 100 and then multiplying by 100 ($y2_1^* = 100 + 100 * y2_1$). Re-fit Model 2 using this new data. Address each question: What changed? What effect does this new observation have on the model? Is this observation an outlier on the fitted model? Is this observation

a high-leverage point? Explain your answers showing fully knowledge and computations.



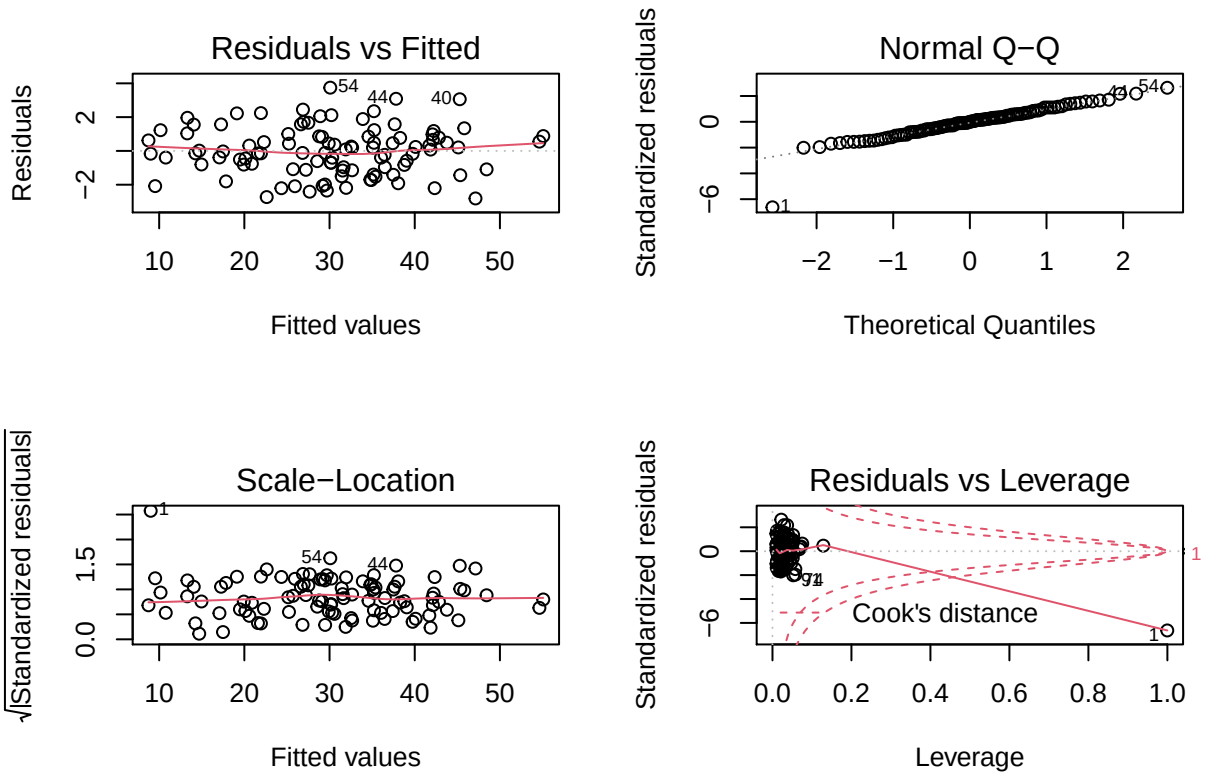
- according to the summary, almost everything has changed,
- This observation 1 is an outlier on the fitted model, because we can see that the residual plot identifies 1 as a outlier, but it has other outliers.(Outliers are observations for which the response y_i is unusual given the predictor x_i)
- This observation 1 is a high-leverage point, because observations with high leverage high leverage have an unusual value for x_i, and we can see it on the leverage plot.

---

h. (BONUS) Now suppose we corrupt one of the observations in x1: corrupt the first observation by adding 100 and then multiplying by 100 ($x1_1^* = 100 + 100 * x1_1$). Re-fit Model 2 using this new data. Address each question: What changed? What effect does this new observation have on the model? Is this observation an outlier on the fitted model? Is this observation a high-leverage point? Are the affects of corrupted data on model estimates same as in the part g?

```
## Warning in sqrt(crit * p * (1 - hh)/hh): NaNs produced

## Warning in sqrt(crit * p * (1 - hh)/hh): NaNs produced
```



- β_1 become smaller, means x_1 has less effects on y. The P value of the β_1 changed, it becomes larger than 0.05.
- This observation 1 is not an outlier on the fitted model, because we can see that the residual plot does not identify 1 as a outlier, but it has other outliers.(Outliers are observations for which the response y_i is unusual given the predictor x_i)
- This observation 1 is a high-leverage point, because observations with high leverage high leverage have an unusual value for x_i, and we can see it on the leverage plot.

2) (*Application*)

This question involves the use of multiple linear regression on the `Auto` data set on 9 variables (one response, six numerical, and two categorical) and 392 vehicles with a dependent (target) variable `mpg`.

Variable names:

- `mpg`: miles per gallon
- `cylinders`: Number of cylinders between 4 and 8
- `displacement`: Engine displacement (cu. inches)
- `horsepower`: Engine horsepower
- `weight`: Vehicle weight (lbs.)
- `acceleration`: Time to accelerate from 0 to 60 mph (sec.)
- `year`: Model year (modulo 100)
- `origin`: Origin of car (1. American, 2. European, 3. Japanese)
- `name`: Vehicle name

Before doing a model fit, do exploratory data analysis (EDA) by getting numerical or graph summaries. For example, the sample mean and sd of mpg is 23.45 and 7.81. Determine types of data: If predictors are numerical, lm() will work directly; if categorical, you need to make dummy or factor() will do it.

In the SLR fitted model, the R^2 of the fit is 0.6059, meaning 60.59% of the variance in mpg is explained by horsepower in the linear model.

In this part, you will fit multiple linear regression (MLR) models using the lm() with mpg as the response and all the other features as the predictor. Use the summary() function to print the results. Use the plot() function to produce diagnostic plots of the least squares regression fit. Include and comment on the output.

I used the RMarkdown to produce the previous paragraph so use this feature when needed:

In this part, you will fit multiple linear regression (MLR) models using the lm() with mpg as the response and all the other features as the predictor. Use the summary() function to print the results. Use the plot() function to produce diagnostic plots of the least squares regression fit. Include and comment on the output.

- Call the sample mean of `mpg`, `Model Baseline`.
- Perform a SLR with `mpg` as the response and `horsepower` as the predictor. Call this model, `Model 1`.
- Perform a MLR with `mpg` as the response and `horsepower` and `year` as the predictors. Call this model, `Model 2`.
- Perform a MLR with `mpg` as the response and all other variables except the categorical variables as the predictors. Call this model, `Model 3`.
- Perform a MLR with `mpg` as the response and all variables including the categorical variables as the predictors. Call this model, `Model Full`.

```
## The following objects are masked from Auto (pos = 3):
##
##     acceleration, cylinders, displacement, horsepower, mpg, name,
##     origin, weight, year
```

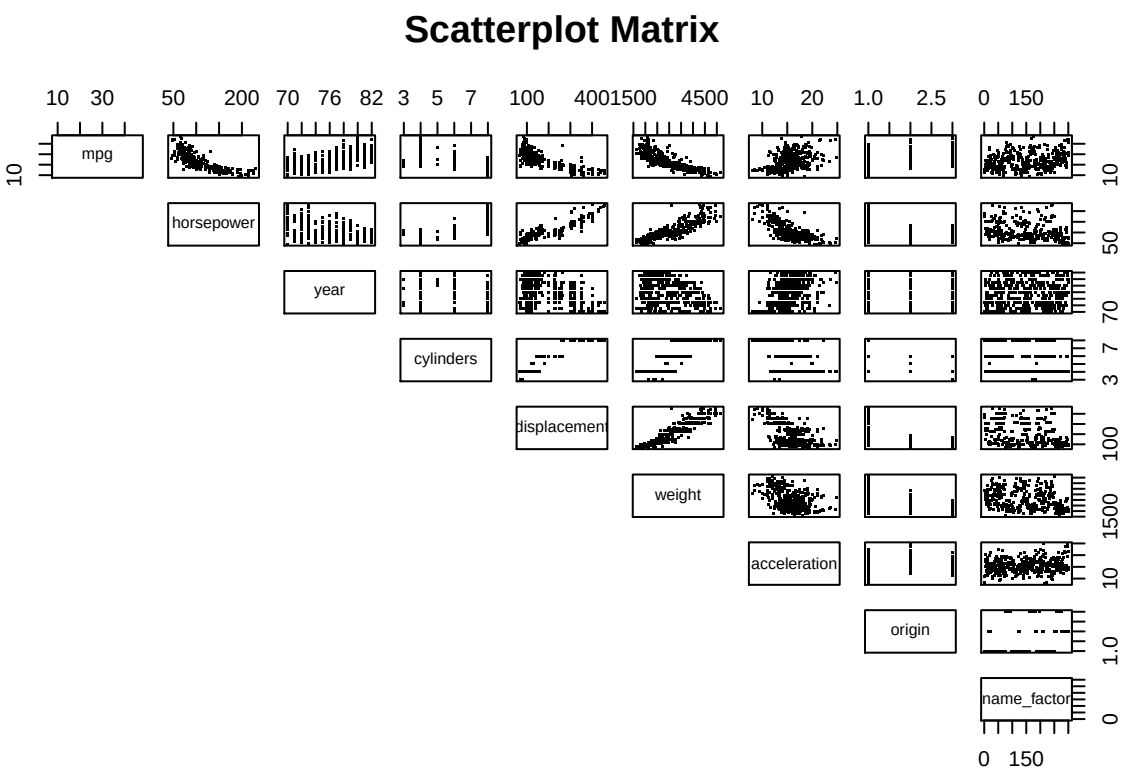
```
## Warning in model.matrix.default(mt, mf, contrasts): the response appeared on the
## right-hand side and was dropped
```

```
## Warning in model.matrix.default(mt, mf, contrasts): problem with term 3 in
## model.matrix: no columns are assigned
```

```
## Warning in model.matrix.default(mt, mf, contrasts): the response appeared on the
## right-hand side and was dropped

## Warning in model.matrix.default(mt, mf, contrasts): problem with term 3 in
## model.matrix: no columns are assigned
```

a. Produce a scatterplot matrix which includes all of the variables in the data set.



---

b. Compute the matrix of correlations between the variables using the function cor(). You will need to exclude the qualitative variables.

Table 3: Correlation Matrix

	mpg	cylinders	displacement	horsepower	weight	acceleration	year	origin
mpg	1.0000000	-0.7776175	-0.8051269	-0.7784268	-0.8322442	0.4233285	0.5805410	0.5652088
cylinders	-0.7776175	1.0000000	0.9508233	0.8429834	0.8975273	-0.5046834	-0.3456474	-0.5689316
displacement	-0.8051269	0.9508233	1.0000000	0.8972570	0.9329944	-0.5438005	-0.3698552	-0.6145351
horsepower	-0.7784268	0.8429834	0.8972570	1.0000000	0.8645377	-0.6891955	-0.4163615	-0.4551715

	mpg	cylinders	displacement	horsepower	weight	acceleration	year	origin
weight	-0.8322442	0.8975273	0.9329944	0.8645377	1.0000000	-0.4168392	-0.3091199	-0.5850054
acceleration	0.4233285	-0.5046834	-0.5438005	-0.6891955	-0.4168392	1.0000000	0.2903161	0.2127458
year	0.5805410	-0.3456474	-0.3698552	-0.4163615	-0.3091199	0.2903161	1.0000000	0.1815277
origin	0.5652088	-0.5689316	-0.6145351	-0.4551715	-0.5850054	0.2127458	0.1815277	1.0000000

---

c. What does the coefficient for the `horsepower` variable suggest in Model 1? Does it change in other models?

Table 4: coefficient of mpv vs. horsepower of 3 models

	coeffients
Model1	-0.1578
Model2	-0.1317
Model3	-0.0170
Model Full	-0.0160

The coefficient for the mpg variable in model 1 suggest that horsepower and mpg has negative relationship and when horsepower increase 1, mpg will decrease 0.16. The coeffient between mpg and horsepower changed smaller in other model,but still negative.

---

d. Make a table and report the measures of $SSTO$, MSE, R^2, R^2_{adj}, BIC, F-ts and F-pvalue for each model (4 models + 1 baseline), if applicable. (you may write an r function that calculates all these, so you can use it in other tasks)

Table 5: summary of four models

	SSTO	MSE	R^2	R^2adj	BIC	F-ts	F-pvalue	Comments
Model Baseline	0	0	0	0	0	0	0.000000e+00	no fitness data
Model1	23818.99	24.07	0.61	0.6	2375.24	599.72	7.031989e-81	Fits ok
Model2	23818.99	19.26	0.69	0.68	2292.82	423.94	9.699789e-93	Fits better
Model3	23818.99	1246.9	0.82	0.82	2100.69	252.43	1.703467e-125	Fits well
Model Full	23818.99	11.01	0.82	0.82	2103.53	222.45	1.092733e-125	overfitting

---

e. Comment briefly on the quality of the fit for each model. Do this in the table you created in part d. See above

---

f. Which predictors appear `most important` in the `Model Full` fit in terms of relationship to the response? How do you justify?

```
##
## Call:
## lm(formula = mpg ~ horsepower + year + mpg + cylinders + displacement +
##     weight + acceleration + origin + name_factor)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.4493 -2.2781 -0.1094  1.9409 13.1484
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -1.789e+01  4.648e+00  -3.850 0.000139 ***
## horsepower   -1.604e-02  1.376e-02  -1.166 0.244366
## year          7.540e-01  5.087e-02  14.822  < 2e-16 ***
## cylinders    -4.774e-01  3.225e-01  -1.480 0.139650
## displacement  2.048e-02  7.502e-03   2.730 0.006622 **
## weight       -6.568e-03  6.525e-04 -10.066  < 2e-16 ***
## acceleration  8.499e-02  9.861e-02   0.862 0.389312
## origin        1.308e+00  2.854e-01   4.583 6.22e-06 ***
## name_factor   3.584e-03  2.043e-03   1.754 0.080230 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.319 on 383 degrees of freedom
## Multiple R-squared:  0.8229, Adjusted R-squared:  0.8192
## F-statistic: 222.5 on 8 and 383 DF,  p-value: < 2.2e-16
```

I think origin is important in the model full in terms of relationship to the prediction, I justify it by how larger is the absolute value of coefficient. The larger the coefficient is, the larger the impact the predictor has, when origin increase 1, mpg will increase 1.308.

---

g. Using Model 2, predict the mpg at `c(horsepower, year)=c(200, 80)`. Report the 95% confidence interval for the prediction.

```
## Predicted mpg at horsepower=200, year=80 is 13.51137
```

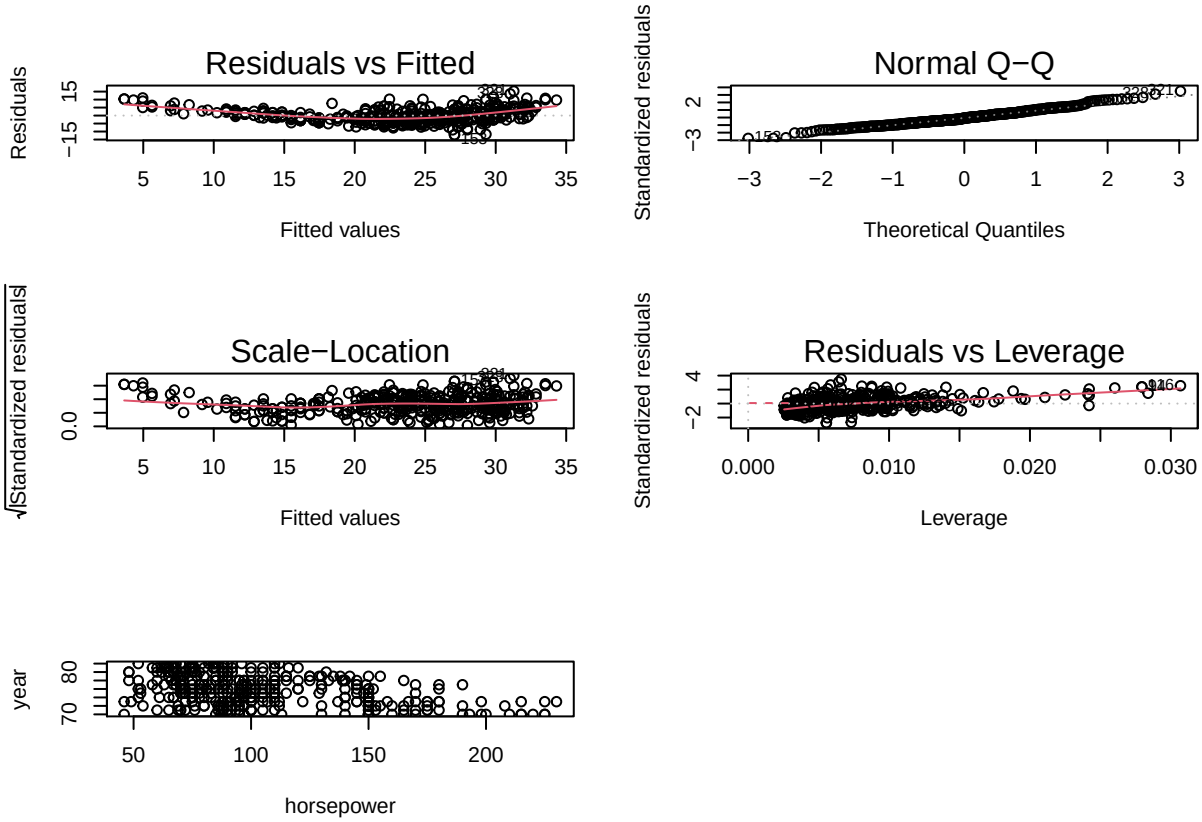
```
## Confidence interval is : ( 4.745447 , 22.2773 )
```

---

h. Do the fit diagnostics for the Model 2 fit by doing:

- Check some assumptions. Include necessary plots. Avoid including uncommented outputs. Comment on any problems you see with the fit.
- Do the residual plots suggest any unusually large outliers?
- Does the leverage plot identify any observations with unusually high leverage?

- Do any interactions between horsepower and year appear to be statistically significant?
- Try a transformation of the mpg variable, such as log(X), in order to improve the R^2_{adj}. Comment on your findings.




```
## orignal adjusted r square is 0.68388
```

```
## new adjusted r square is 0.7447858
```

- Ideally, the residual plot will show no fitted pattern and the red line should be approximately horizontal at zero. Here, it shows that fitted value has some relationship with fiited value, which is not good. Also, residuals are not normally distributed in Q-Q plot. ALso. scale-location tells us our model did not match the assumption of equal variance (homoscedasticity), althoght there is a horizontal line, but it doesnt' have equally (randomly) spread points.
- There are unusual large outliers:321,328, adn small one:153.
- There are unusual high leverages: 94, 116, they are far beyond the Cook's distance lines.
- When horsepower increase, year tends to be decrease.
- I use log(horsepower), and R^2_{adj} increase a lot, and a large value of adjusted R2 indicates a model with a small test error.

---

i. (BONUS) Using Model 2, estimate the mpg at `c(horsepower, year)=c(200, 80)`. Report the 95% confidence interval for the estimation. Does this differ from the prediction interval in part g? Explain the differences.

```
## Predicted mpg at horsepower=200, year=80 is 13.51137
```

```
## Confidence interval is : ( 11.96143 , 15.06132 )
```

3) (*Theory*)

In SLR, model errors are defined as

$$e_i = y_i - y_i = y_i - (\beta_0 + \beta_1 x_i).$$

The ordinary LS estimation argument with cost function notation can be expressed as

$$\hat{\beta}_{LS} : argmin J(\beta) = argmin \frac{1}{n} \sum_1^n e_i^2.$$

a. Obtain the estimating equation for the model parameter β_1 (using differentiation). If you prefer matrix notation way to obtain the equation in a LR model, this would be great. Then, express the $\hat{\beta}_1$.

$$Q = \sum_1^n e_i^2 = \sum_1^n (y_i - (\beta_0 + \beta_1 x_i))^2$$

$$\frac{dQ}{d\beta_1} = -2 \sum_1^n x_i (y_i - (\beta_0 + \beta_1 x_i)) = -2(\sum_1^n x_i y_i - \sum_1^n \beta_0 x_i + \sum_1^n \beta_1 x_i^2) = 0$$

$$\frac{dQ}{d\beta_0} = -2 \sum_1^n (y_i - (\beta_0 + \beta_1 x_i)) = -2(\sum_1^n y_i - \sum_1^n \beta_0 + \sum_1^n \beta_1 x_i) = 0$$

$$\sum_1^n x_i y_i - \beta_0 \sum_1^n x_i - \beta_1 \sum_1^n x_i^2 = 0$$

$$\sum_1^n y_i - \sum_1^n \beta_0 + \sum_1^n \beta_1 x_i = \sum_1^n y_i - n\beta_0 - \beta_1 \sum_1^n x_i = 0$$

$$\beta_1 = \frac{\sum_1^n x_i y_i - \sum_1^n x_i \sum_1^n y_i}{(\sum_1^n x_i)^2 - \sum_1^n x_i^2}$$

*** #### b. In SLR, is there any difference between $var(\hat{\mu}_{y_i|x_i})$ and $var(\hat{y}_{x_0})$, where $\hat{\mu}_{y_i|x_i}$ is estimation at x_i and $\hat{y}_{x_0}$ is prediction at a future value x_0? Explain.

No different between them.

$$var(\hat{\mu}_{y_i|x_i}) = var(\hat{\beta}_0 + \hat{\beta}_1 x_i) = \hat{\beta}_1^{\,2} var(x_i)$$

$$var(\hat{y}_{x_0}) = var(\hat{\beta}_0 + \hat{\beta}_1 x_0) = \hat{\beta}_1^{\,2} var(x_0)$$

---

c. `Leverage statistic` of observation x_i on $\hat{y}$ in a LS regression model is $h_i = H_{ii}$, which describes the degree by which the $i-$th measured value influences the ith fitted value. In the slides, we reviewed:

$$X \cdot \hat{\beta} = X \cdot (X^t \cdot X)^{-1} \cdot X^t \cdot y = H \cdot y = \hat{y}$$

Also, some mathematical properties are expressed as these two arguments: $1/n \leq h_i \leq 1$, $\bar{h} = (p+1)/n$. Verify these two formulas numerically using the Model 2 fit in Q2, Auto dataset. Report the calculations. Comment on the calculations whether or not these are verified. First argument $\frac{1}{n} \leq h_i < 1$ is true with Model 2 fit in Q2 Second argument $\bar{h} = \frac{p+1}{n}$ is true with Model 2 fit in Q2

```
## 1/n<min_hii<max_hii<1:  0.00255102 < 0.0025519 < 0.03068558 < 1
```

```
## (p+1)/n = ( 2 + 1 ) / 392 = 0.007653061
```

```
## mean(hii)= 0.007653061
```

---

d. (BONUS) R^2 in SLR has two expressions:

$$R^2 = \frac{\left[\sum(x_i - \bar{x})(y_i - \bar{y})\right]^2}{\sum(x_j - \bar{x})^2 \sum(y_k - \bar{y})^2}$$

and

$$R^2 = \frac{\sum(y_i - \bar{y})^2 - \sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2} = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}.$$

Prove that these are equivalent.

$$R^2 = \frac{\left[\sum(x_i - \bar{x})(y_i - \bar{y})\right]^2}{\sum(x_j - \bar{x})^2 \sum(y_i - \bar{y})^2}$$

$$R^2 = \frac{\sum(y_i - \bar{y})^2 - \sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2} = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2} = \frac{\sum(x_j - \bar{x})^2}{\sum(x_j - \bar{x})^2} - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}$$

$$\frac{\sum(x_j - \bar{x})^2}{\sum(x_j - \bar{x})^2} - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2} = \frac{\sum(x_j - \bar{x})^2 \sum(y_i - \bar{y})^2 - \sum(x_j - \bar{x})^2 \sum(y_i - \hat{y}_i)^2}{\sum(x_j - \bar{x})^2 \sum(y_i - \bar{y})^2} = \frac{\sum(x_j - \bar{x})^2 (\sum(y_i - \bar{y})^2 - \sum(y_i - \hat{y}_i)^2)}{\sum(x_j - \bar{x})^2 \sum(y_i - \bar{y})^2}$$

$$\frac{\sum(x_j - \bar{x})^2 (\sum(y_i - \bar{y})^2 - \sum(y_i - \hat{y}_i)^2)}{\sum(x_j - \bar{x})^2 \sum(y_i - \bar{y})^2} = \frac{\sum(x_j - \bar{x})^2 \sum(\hat{y}_i - \bar{y})^2}{\sum(x_j - \bar{x})^2 \sum(y_i - \bar{y})^2}$$

---

e. (BONUS) Ask a challenging question and answer (under the assignment context).

I hereby write and submit my solutions without violating the academic honesty and integrity. If not, I accept the consequences.

How long did the assignment solutions take?: 15+ hrs

References

Retrieved 16 Feb.2021 https://www.statisticssolutions.com/assumptions-of-multiple-linear-regression/

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning (Vol. 112, p. 18). New York: springer.