Maria Varga

Problem Set 1 AQM

Linear Regression

Problem 1 - Least Squares

The general regression model with 'n' observations and 'k' explanatories variables is given by

$$y = X\beta + \varepsilon$$

The **sum of squared residual function** is given by

$$\boldsymbol{RSS} = \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} \boldsymbol{\varepsilon^2} = \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} (\boldsymbol{y_i} - \boldsymbol{x_i}'\boldsymbol{\beta})^{\boldsymbol{2}}$$

Using matrix notation, we have

$$\begin{gathered} RSS = (y - X\beta)'(y - X\beta) \\ = y'y - \beta'X'y - y'X\beta + \beta'X'X\beta \\ = y'y - 2\beta'X'y + \beta'X'X\beta \end{gathered}$$

The ordinary least squares (OLS) is obtained by minimizing the sum of squared residual in respect to beta, that is minimizing the vertical distance between the observed and the predicted.

To find $\beta = b$ that minimizes the sum of squared residuals, we take the derivative and set it equal to zero

$$\frac{dRSS}{db} = -2X'y + 2X'Xb = 0$$

as long as X has full rank, this is a positive definite matrix and hence a minimum.

We then have

$$(X'X)b = X'y$$

if the inverse of (X’X) exists, we have

$$(X'X)^{-1}(X'X)b = (X'X)^{-1}X'y$$

As $(X'X)^{-1}(X'X) = I$, the identity matrix, the **estimator b** that minimize the sum of square error, the OLS, is given by

$$\boldsymbol{b} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$$

Problem 3 - Interpretation

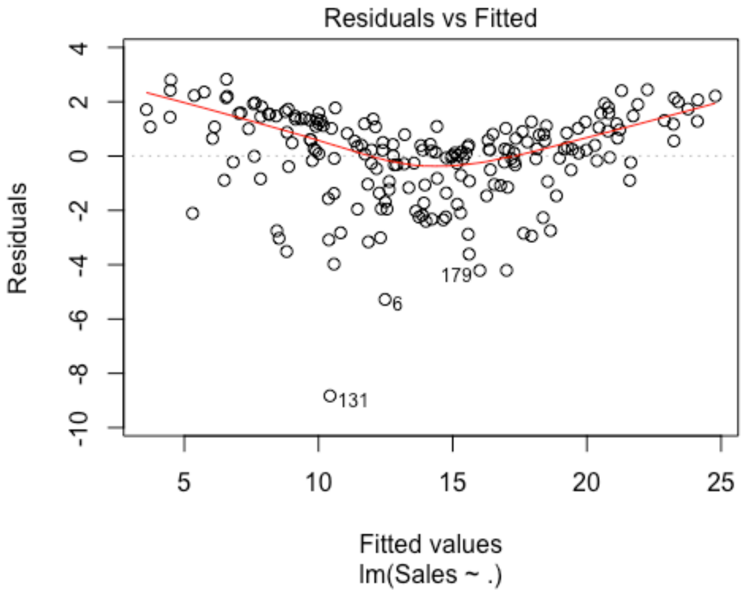
PS: After some analysis, the entry number 131 was excluded from the dataset advertising.csv because it was considered an outlier that could have substantial effect on the estimated linear model.

1. **Is the relationship approximately linear?**

Using the dataset advertising.csv we can build a linear model that is a linear combination of each particular product given by

$$Sales \sim TV + Radio + Newspaper$$

After using lm function to fit the model, we can see the **Residuals vs Fitted** graph given below:



This graph shows the residuals on the vertical axis and the fitted value on the horizontal axis.

Randomly distributed dots is an evidence of goodness. In our case the graph might indicate a non-linear relationship between predictor and response variable as we have a sort of curved line.

2. **What other assumptions discussed in class should be considered? If an assumption is violated, correct for it the best you can.**

Some of the assumption in regression model are:

Linear relationship: it is assumed a linear relationship between predictor and response variable. It can be overcome adding functions of the predictor and/or response to correct for linearity.

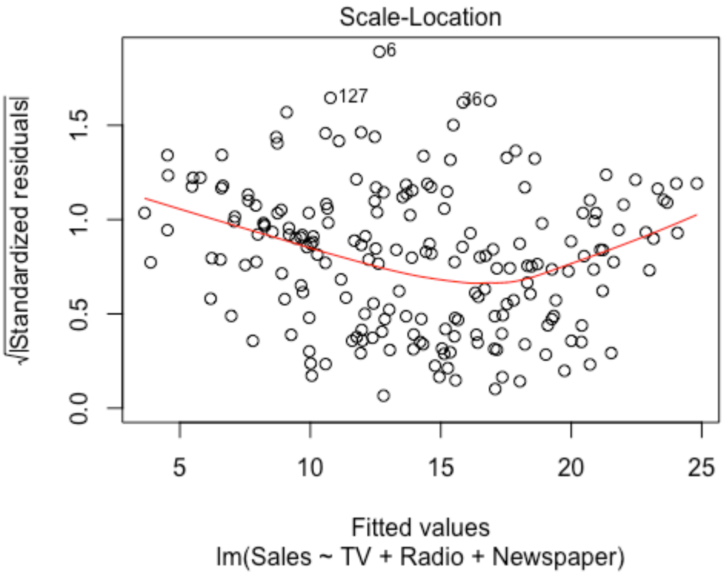
No or little multicollinearity: linear regression assumes that there is little or no multicollinearity in the data.
The matrix below shows the correlation between variables

Correlation

	TV	**Radio**	**Newspaper**	**Sales**
TV	1	0.055	0.057	0.782
Radio	0.055	1	0.354	0.576
Newspaper	0.057	0.354	1	0.228
Sales	0.782	0.576	0.228	1

As we can see, there is a correlation between Newspaper and Radio which might suggest that Newspaper should be excluded from the model (As Radio is mode correlated to the response variable). However, in some situations the variable might be required to be in the model regardless of its significance or correlation with other response variable.

Homoscedasticity: it is assumed that the variance of the error term is constant.
One way to check the assumption of equal variance (homoscedasticity) is to check the **Scale-Location** graph that shows the residuals spread along the ranges of predictors. It's good if you see a horizontal line with equally (randomly) spread points. The graph below shows a slight violation of this when using the model $Sales \sim TV + Radio + Newspaper$



3. **Is there a relationship between advertising sales and budget?**

Analyzing the results from the lm model given by $Sales \sim TV + Radio + Newspaper$, we can see the coefficient table

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	2.938889	0.311908	9.422	<2e-16	***
TV	0.045765	0.001395	32.809	<2e-16	***
Radio	0.18853	0.008611	21.893	<2e-16	***
Newspaper	-0.001037	0.005871	-0.177	0.86	

This table shows that there is a significant relationship between TV budget and Sales as well as between Radio budget and Sales. The table also shows that the relationship between Newspaper budget and Sales is not significant as the p-value is greater than the usual threshold 0.05.

4. **How strong is the relationship?**

For both TV and Radio, the relationship is strong enough to be considered in the model as shown by the p value in the table above. In this case, the probability that the null hypothesis is true (that the true coefficient is zero) is very low, suggesting that the relationship between Sales and TV and between Sales and Radio are significant.

The R-squared value equal to 0.9096 also shows that a strong relationship exists as it shows that 90% of the response variable variation is explained by the linear model.

5. **How large is the effect of each medium on sales?**

The **coefficient**s tell us how much the dependent variable is expected to increase when that independent variable increases by one, holding all the other independent variables constant.

That means that if we increase TV budget by 1, we expect Sales to increase by 0.045765 and if we increase Radio budget by 1, we expect sales to increase by 0.18853. In the case of Newspaper, it appears as having a negative impact on sales. However, as the p-value is not significant, this impact is irrelevant.

6. **Could collinearity be the reason that the confidence interval associated with newspaper is so wide?**

In our case Collinarity means that the part of the variability explained by Newspaper is already incorporated by the others explanatory variables. This could lead to a non-significance of the variable Newspaper in the model which has a larger confidence interval that must contains the value 0. That is, the parameters beta associated to this variable is not significant enough to be considered in the model.

7. **As there is risk involved in driving sales, provide a lower bound estimate involved with al- locating the following budget: TV=149, Radio=22, Newspaper=25. How about TV=149, Radio=60, Newspaper=25?**

Using the same model $Sales \sim TV + Radio + Newspaper$ we can predict the sales for the two following budgets

Case 1: TV=149, Radio=22, Newspaper=25

fit	lwr	upr
13.9	13.6	14.1

Case 2: TV=149, Radio=60, Newspaper=25

fit	lwr	upr
21.0	20.4	21.7

We can clearly see that the budget for the second case gives a much higher sales prediction. With the first case we expect a lower bound estimation of 13.6 thousands of units while with the second case we expect a lower bound estimation of 20.4.

8. **Once you feel your model is suitable, suggest your marketing plan.**

After some tests and residual analysis, the two best model I came across are:

$$Sales \sim TV^2 + TV + TV * Radio$$

and

$$Sales \sim TV + Radio + TV * Radio$$

Both gave higher Adjusted R-squared, 0.9867 and 0.9744 respectively, indicating that the TV-Radio interaction is an important factor to consider. (However, you must be aware that the model might be overfitting the data).

Analyzing the model and taking in consideration other factors as well, I would suggest having the TV budged between 50%-60% of total, the Radio 30%-40% and only 10% on Newspaper.

Problem 4 - Weighted Regression

1. **Please solve the minimization and derive the weighted least square estimators analytically.**

Using similar deduction from exercise 1, we are now interested in minimize the weighted RSS

$$WRSS = (1/n)(y - X\beta)'W(y - X\beta)$$

Similarly as before, we have that

$$= 1/n(y'Wy - \beta'X'Wy - y'WX\beta + \beta'X'WX\beta)$$
$$= \frac{1}{n}y'Wy - \frac{1}{n}2\beta'X'Wy + \frac{1}{n}\beta'X'WX\beta$$

To find $\beta = b$ that minimizes the sum of squared residuals, we take the derivative and set it equal to zero

$$\frac{dWRSS}{db} = \frac{1}{n}(-2X'Wy + 2X'WXb) = 0$$

With similar arguments as before, we can then have

$$(X'WX)b = X'Wy$$

and find that the **estimator** b that minimize the weighted RSS is given by

$$\boldsymbol{b} = (\boldsymbol{X'WX})^{-1}\boldsymbol{X'Wy}$$

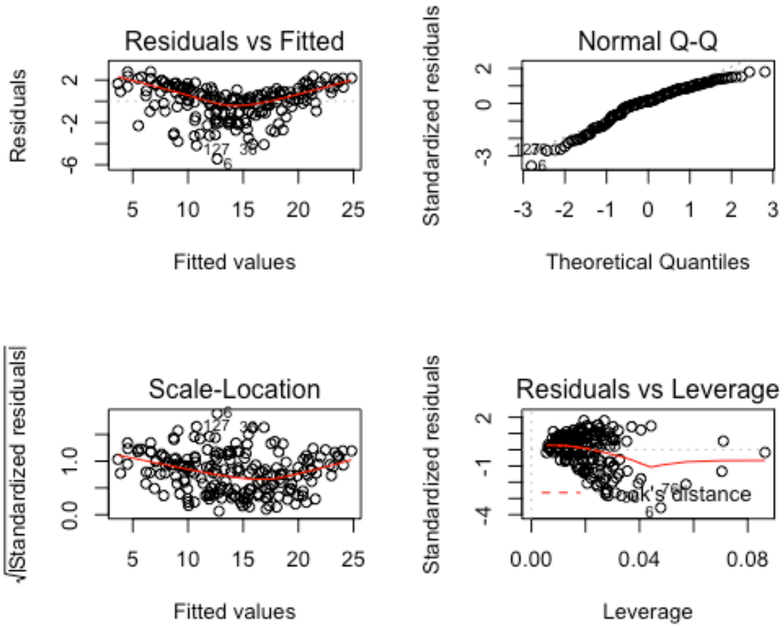
2. **Use this solution to fit a weighted least squares to the advertising data. How does the fit differ from your original OLS?**

The structure of **W** is unknown, so we have to perform an ordinary least squares (OLS) regression first. The residuals however are much variable to be used directly in estimating the variance so we regress the squared residuals against that predictor and use the resulting fitted values as our weights in our weighted least squares regression model.

Model lm(Sales~TV+Radio+Newspaper,data=adv):

Coefficients:

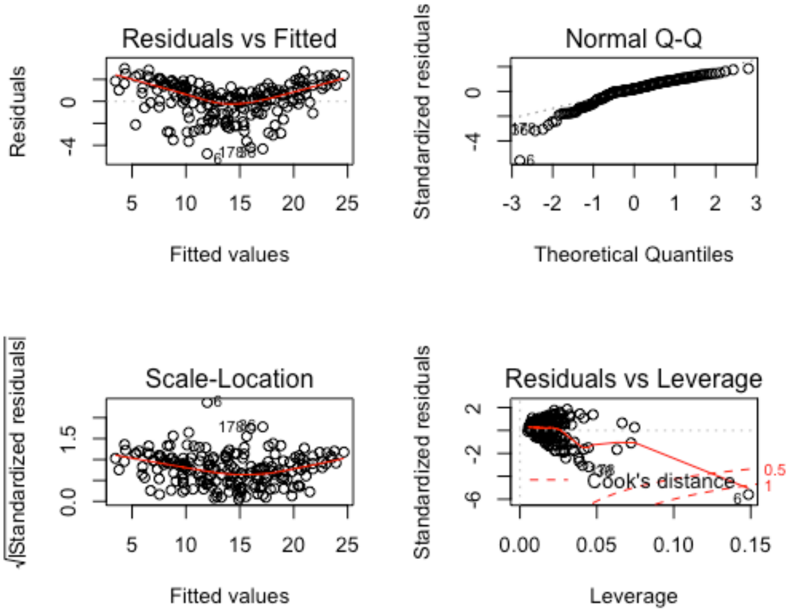
	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	2.938889	0.311908	9.422	<2e-16	***
TV	0.045765	0.001395	32.809	<2e-16	***
Radio	0.18853	0.008611	21.893	<2e-16	***
Newspaper	-0.001037	0.005871	-0.177	0.86	



Model lm(Sales~TV+Radio+Newspaper,data=adv, weights = 1/sig):

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	2.68051	0.29734	9.015	<2e-16	***
TV	0.045438	0.001279	35.529	<2e-16	***
Radio	0.207063	0.007795	26.565	<2e-16	***
Newspaper	-0.003602	0.005195	-0.693	0.489	



This weighted least squares model does not differ much from the original ordinary least squares. But analyzing the Scale-Location graph we can see that it slightly improved homoscedasticity as expected. There are, however, other weights estimative that could be used to see if homoscedasticity could be further improved.

3. Why would we weight by the inverse of the variance?

The model is simply treating each observation as more or less informative about the underlying relationship. Since each weight is inversely proportional to the error variance, it reflects the information in that observation. So, an observation with small error variance has a large weight since it contains relatively more information than an observation with large error variance (small weight).