Problems fitting multiple regression models
(Statistical Modelling I)

Lubna Shaheen

Week 10, Lecture 2

Queen Mary
University of London

Outline

Problems fitting multiple regression models

Revision

Multicollinearity: Becomes more likely to occur when we have a large number of explanatory variables

1. To get a solutions to normal equations: we need $X^T X$ to be non-singular. As if $X^T X$ is singular, then its determinant is zero, so it cannot be inverted and we cannot find a unique solution to the Normal Equations, and therefore no unique least squares beta estimates.
2. Mostly it happens when two or more variables are equal and of one variable is a linear combination if the other variable.
3. Parameters with large variances is one of the problems of multicollinearity where some of the columns of X are close to linear combination of other columns.
4. When variance is very high this can even lead to a parameter having the wrong sign.

Exams Style Question, (2019)

When the number of explanatory variables is relatively small, it may well be possible to spot multicollinearity by scanning the data.

We can calculate the VIF of each of the explanatory variable x_j against the other $p-2$ explanatory variables, so x_j is the response variables and other $p-2$ variables have their $\beta's$ parameters.

1. We calculate the co-efficients of determination of this regressor of x_j and write as a real number between 0 and 1 i.e. R_j.
2. Variance Inflation Factor $VIF_j = \dfrac{1}{1-R_j^2}$

   High R_{j^2} indicates a strong linear relationship between x_j and the other $x's$ which results in a large VIF for x_j.
   We usually take VIF $>$10 as indication of a multicollinearity problem. We would need to reduce the set of explanatory variables to remove linear combinations.
3. Another indication of Multicollinearity can be model where the overall model shows significance with an F test but none of the parameters shows siginificance with t-test

Queen Mary
University of London

Exams Style Question, (2019)

Question 3. [**32 marks**] A researcher wished to study the relationship between the annual salaries (Y in thousands of dollars) of 24 Mathematics Professors in a large American University and an index of publication quality (x_1), number of years of experience (x_2), an index of success in obtaining grants (x_3) and an index based on teaching evaluations (x_4). The data were read into R and the following commands and output were initially found.

```
salary<-lm(y~x1+x2+x3+x4)
summary(salary)

Call:
lm(formula = y ~ x1 + x2 + x3 + x4)

Residuals:
    Min      1Q  Median      3Q     Max
-6.5891 -1.6925 -0.6017  2.5454  4.7078

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 41.54908    6.66329   6.236 5.47e-06 ***
x1           2.09307    0.65199   3.210 0.004607 **
x2           0.64761    0.07387   8.767 4.19e-08 ***
x3           2.78690    0.59594   4.676 0.000164 ***
x4          -2.18893    1.82959  -1.196 0.246255
---
```

Exams Style Question, (2019)

```
Residual standard error: 3.455 on 19 degrees of freedom
Multiple R-squared:  0.917,Adjusted R-squared:  0.8995
F-statistic: 52.47 on 4 and 19 DF,  p-value: 5.234e-10
```

(a) (i) Write down the fitted model. **[2]**

(ii) What null hypothesis and alternative does the output

```
F-statistic:  52.47 on 4 and 19 DF, p-value:  5.234e-10
```

test? What is the conclusion? **[4]**

(b) The following commands were then entered.

```
stdres <- rstandard(salary)
hat<-hatvalues(salary)
i<- 1:24
plot(i,hat, main="Hat values versus i, Salary")
shapiro.test(stdres)
```

Explain briefly the meaning of each command and what output it gives. **[9]**

Exams Style Question, (2019)

(c) Look at the following output

```
> library(car)
> vif(salary)
      x1       x2       x3       x4
1.365795 1.324020 1.162684 1.052740
```

The researcher finds the vif values to investigate multicollinearity.

(i) What does vif stand for? [1]

(ii) What is multicollinearity and what are its effects? [5]

(iii) Is there any problem with multicollinearity here? Explain your answer. [2]

Exams Style Question, (2019)

(d) Look at the following output

```
> library(leaps)
> best.subset <- regsubsets(y~x1+x2+x3+x4, salary, nvmax=4)
> best.subset.summary <- summary(best.subset)
> best.subset.summary$outmat
          x1  x2  x3  x4
1  ( 1 ) " " "*" " " " "
2  ( 1 ) " " "*" "*" " "
3  ( 1 ) "*" "*" "*" " "
4  ( 1 ) "*" "*" "*" "*"
> best.subset.summary$adjr2
[1] 0.7182713 0.8494270 0.8973512 0.8995185
```

(i) Define **adjusted** R^2. [1]

(ii) Explain briefly what this output shows. [4]

(e) Discuss, based on all the output above, whether the variable x4 should be dropped from the model. [4]

Residuals re-cap

We have already defined residuals in multiple linear regression matrix form and stated some of their properties:

$$e = Y - \widehat{Y} = (I - H)Y$$

$$E[e] = 0$$
$$\text{var}(e) = \sigma^2(I - H)$$

where H is the hat matrix given by $H = X(X^TX)^{-1}X^T$.

Hat matrix and individual residuals

- We can use individual elements of H to tell us more about the residuals
- let h_{ij} be the $(i, j)^{th}$ elements of H
- so the diagonal elements of H are the h_{ii}

Then

$$var(e_i) = (1 - h_{ii})\sigma^2$$

$$cov(e_i, e_j) = -h_{ij}\sigma^2$$

Estimate the variance of a particular observation's residual

The elements on the diagonal of H are the important ones in many cases, because you can take, say, the 10'th observation, and you calculate the variance of the residual for that obervation:

$$var(e_{10}) = \sigma_e^2(1 - h_{10,10})$$

Queen Mary
University of London

Residuals and their plots

Notation

- In the simple linear regression model we referred to h_{ii} as ν_i
- For multiple linear regression, either notation is ok

$$var(e_i) = (1 - h_{ii})\sigma^2$$

- gives us an additional reason to standardise the residuals
  - We can see that variance of the residuals was different to the σ^2, assumed for the random error terms in the original model specification.
  - Furthermore the variation of each of the residuals might be different depending on the hat matrix.

Queen Mary University of London

Why standardise the residuals?

Simple linear

- Variance of residuals different from σ^2 assumed for the random error terms

Multiple linear

- Variance of each residual might be different depending on ***H***
- Makes detection outliers tricky

Standardised residuals

Standardised residuals are d_i where for multiple linear regression models

$$d_i = \frac{e_i}{\sqrt{S^2(1 - h_{ii})}} = \frac{e_i}{\widehat{se}(e_i)}$$

The denominator is the standard error of the estimated coefficient.

If the normal distribution assumption for the residuals is followed $d_i \sim t_{n-p}$

When n large, $h_{ij}(i \neq j)$ tends to be small.
Then asymptotically the standardised residuals d_i are iid $N(0, 1)$
This is the property we rely on most heavily in residual plots

Standardised residuals

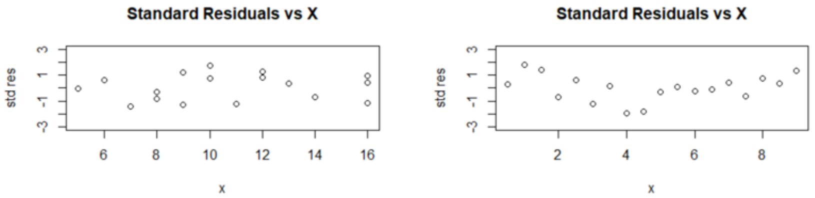
Our four most common checks using the standardised residuals are similar to those for simple linear regression models:

- Linear Relationship
- Constant Variance
- Normal Distribution
- Outliers
- Autocorrelation

Standardised residuals

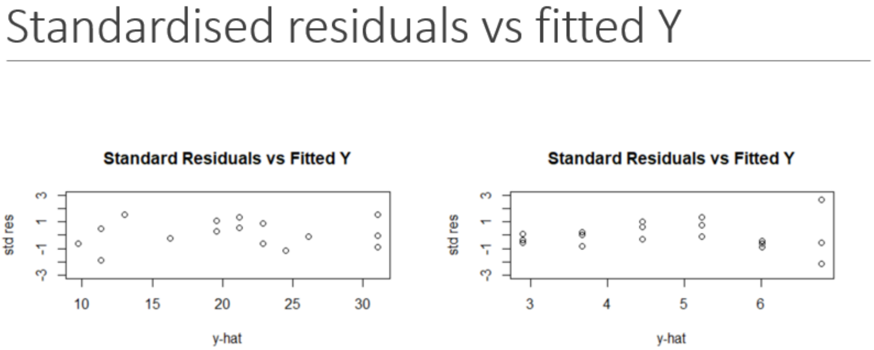
We plot d_i against each of the explanatory variables x_i

Standardised residuals vs each x_i



Standardised residuals

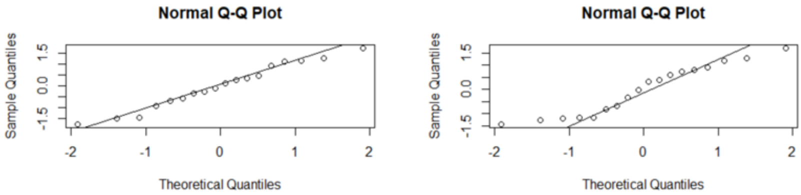
We plot d_i against the fitted values $\widehat{y}_i$



Standardised residuals

The Q-Q plot is used to check the assumption of normally distributed residuals.

Q-Q Plot



Further checks with residuals plots

Any of the three plots above can be checked for outliers

- Large absolute value of d_i
- If we record observations with a measure of time it can be useful to plot the standardised residuals against time t
- Even if time is not an explanatory variable
- Used to check for " autocorrelation"

Influential observations and leverage

Influential observations and leverage

We previously discussed this with simple linear regression models

Previously we calculated leverage v_i

Now we can relate leverage to the hat matrix

$$v_i = h_{ii} \text{ is the } i^{th} \text{ diagonal elements of H}$$

$$\begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_{N-1} \\ \hat{y}_N \end{bmatrix} = \begin{bmatrix} h_{11} & h_{12} & h_{13} & & & h_{1N} \\ h_{21} & & & & & h_{2N} \\ h_{31} & & & \ddots & & \vdots \\ \vdots & & & & & \\ & & & & & h_{N-1N} \\ h_{N1} & & & & h_{NN-1} & h_{NN} \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_{N-1} \\ y_N \end{bmatrix}$$

Queen Mary University of London

Viewed through the fitted model

Fitted models in $\widehat{Y} = \widehat{\mu} = X\widehat{\beta} = HY$.
So the i^{th} fitted value is

$$\widehat{y_i} = \widehat{u_i} = \sum_{i=1}^{n} h_{ij} y_j = h_{ii} y_i + \sum_{i \neq j} h_{ij} y_j$$

So h_{ii} indicates the extent to which the observation with y_i contributes to the fitted value $\widehat{\mu}_i$. This is what leverage is

$$= \begin{bmatrix} y_1 h_{11} & +y_2 h_{12} & +y_3 h_{13} & + & \cdots & & +y_N h_{1N} \\ y_1 h_{21} & + & & & & & +y_N h_{2N} \\ y_1 h_{31} & + & & \ddots & & & \vdots \\ \vdots & & & & & & \\ & & & & & & +y_N h_{N-1N} \\ y_1 h_{N1} & + & \cdots & & & +y_{N-1} h_{NN-1} & +y_N h_{NN} \end{bmatrix}$$

Viewed through the fitted model

If you look at this for a while, it becomes apparent that the element, hij gives the influence of the j'th observation on the i'th predicted value, $y\hat{}i$. If you compare across row i in the hat matrix, and some values are huge, it means that some observations are exercising a disproportionate influence on the prediction for the i'th observation. If you concentrate on the diagonal elements, h_{ii}, you are focusing on the effects that observations have on their own predicted values. If a model estimated without observation i offers a grossly different predicted value for y_i than a model that includes i, then you know that observation i is having a pretty dramatic effect on the fitted model.

Leverage as diagonal element of H

When we think of leverage as coming from the hat matrix rather than as an independent calculations, a number of properties emerge

$$var(e_i) = \sigma^2(1 - h_{ii})$$

- Now $h_{ii} < 1$ but h_{ii} close to 1 will give $var(e_i)$ close to zero
- that is a fitted value close to the observed value
- In general h_{ii} is small when x_{ij} is close to its mean $\overline{x_i}$ and gets larger the further x_{ij} is from its mean

Large leverage observations

$$\frac{1}{n} < h_{ii} < 1 \text{ and } \sum_{i=1}^{n} h_{ii} = p$$

So average leverage is $\frac{p}{n}$

- This is the general case of average $= \frac{2}{n}$ in simple linear regression

We usually consider leverage

- $> \frac{2p}{n}$ as "high leverage"
- $> \frac{3p}{n}$ as " very high leverage"

Number of potential causes of high leverage
data collection, unique observations

Cook's Statistic

We check leverage because we are concerned if a single observation exerts influence over the regression result

Unusually large Cook's Statistic is one indicator of this influence

We can generalise the formula for Cook's Statistics in multiple linear regression

$$D_i = \frac{(\widehat{\beta} - \widehat{\beta}_i)^T (X^T X)(\widehat{\beta} - \widehat{\beta}_i)}{ps^2}$$

$\widehat{\beta}$ is the vector of least squares parameters

$\widehat{\beta}_i$ is the estimates of parameters found when the i^{th} observation is omitted

Once again, an unusually large value for D_i can be taken as evidence of an influential observation.

Cook's distance can be calculated as:

$$D_j = \frac{r_j^2}{p} \frac{h_{jj}}{(1 - h_{jj})}$$

Exams Style Questions (2020)

Question 6 [15 marks].

In a study of the efficiency of a plant which oxidises ammonia to nitric acid the dependent variable is stack loss and the independent variables are Airflow (flow of cooling air), Water.Temp (cooling water inlet temperature) and Acid.Conc (concentration of acid). The data were read into R and the commands and output are shown below.

```
> stack <- lm(stack.loss ~ Airflow + Water.Temp + Acid.Conc)

> summary(stack)

Call:
lm(formula = stack.loss ~ Airflow + Water.Temp + Acid.Conc)

Residuals:
    Min      1Q  Median      3Q     Max
-7.2377 -1.7117 -0.4551  2.3614  5.6978

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       -39.9197    11.8960  -3.356  0.00375 **
Airflow             0.7156     0.1349   5.307  5.8e-05 ***
Water.Temp          1.2953     0.3680   3.520  0.00263 **
Acid.Conc          -0.1521     0.1563  -0.973  0.34405
---
```

Exams Style Questions (2020)

```
Residual standard error: 3.243 on 17 degrees of freedom
Multiple R-squared:  0.9136,Adjusted R-squared:  0.8983
F-statistic:  59.9 on 3 and 17 DF,  p-value: 3.016e-09

> stdres<-rstandard(stack)
> hat<- hatvalues(stack)
> cook<-cooks.distance(stack)
> i<- 1:21
> plot(i,stdres, main="Standardised residual values")
> plot(i,hat, main="Leverage values")
> plot(i,cook, main="Cooks distance values")
> qf(0.5, 4, 17)
[1] 0.8735735
```

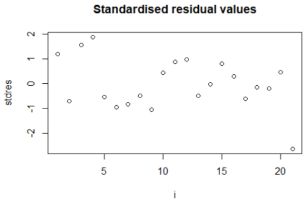
(a) Write down the fitted model. **[2]**

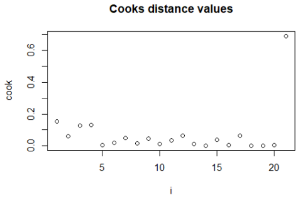
(b) Explain what is meant by an **outlier**, **leverage** and an **influential observation**. Include the relationship between these concepts and how they can be detected. **[8]**

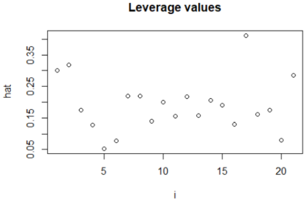
(c) Comment on what the three plots on page 7 tell us about possible outliers, high leverage values and influential observations in these data. **[5]**

Queen Mary
University of London

Exams Style Questions (2020)







Exams Style Questions (2020)

Solution: (a) Stackloss = -39.9187+0.7156 Airflow + 1.2953 Water.temp-0.1521AcidConic

(b) An outlier is an observation with large standardised residual. This means the observation lies well away from the fitted line. Leverage measures how unusual the combinations of regressor values is. It can measure by h_{ii} where $H = X(X^TX)^{-1}X^T$ is the hat matrix.
An influential observation has a large value of Cook's Statistics, which measures the fitted line without this observation has changed.

Detection: How large a stat residual has to be depend on n. High leverage is $h_{ii} > \frac{2p}{n}$ very high $> \dfrac{3p}{n}$
Cook's $D_i > F^p_{n-p}(0.50)$.
Outliers or high leverage values may be influential.
Outlier + High leverage very likely to be influential.

Exams Style Questions (2020)

(c) $p = 4$, $n = 21$, From table
$|d_i| > 2.8$, $h_{ii} > \frac{8}{21} = 0.38$ or $h_{ii} > \frac{12}{21} = 0.57$.
Cook's Statistics $D_i > 0.874$ from output
From graphs there are no outliers, also it has high leverage, also 21 is most influential but not highly so.

What is a linear model?

We've covered a lot of modelling ground

- Least squares estimation
- Properties of estimators
- Interpretation of model results
- Analysis of Variance
- Tests of Significance
- Confidence & Prediction Intervals
- Matrix approaches
- Maximum Likelihood
- Model Building
- Outliers & Leverage
- Automated Methods

What is a linear model?)

An unanswered question

Simple *Linear* Regression Model

Multiple *Linear* Regression Model

What makes a model linear ?

What is a linear model?)

Definition

β A linear model is one that is linear in the parameters

not necessarily one linear in the explanatory variables

What is a linear model?)

Examples of linear models

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$$

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 \sqrt{x_{2i}} + \varepsilon_i$$

$$y_i = \beta_0 + \beta_1 \sin(x_{1i}) + \beta_2 x_{2i} + \varepsilon_i$$

What is a linear model?)

Linearising a model

Sometimes (not always) a non-linear model can be converted into a linear one through a transformation of the response

For example

$y_i = \varepsilon_i \exp(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i})$ is not linear

But taking natural logarithms

$\ln(y_i) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ln(\varepsilon_i)$ is now linear

However we need to be careful

Care needed on the assumption we make about the residuals

To use the techniques we have developed in this module we need

$\ln(\varepsilon_i) \sim N(0, \sigma^2)$ for some constant variance σ^2

not the usual, $\varepsilon_i \sim N(0, \sigma^2)$

Other variations of this model can be linearised by a log transformation

$$y_i = \varepsilon_i \exp(\beta_0 + \beta_1 x_{1i} + \frac{\beta_2}{x_{2i}})$$

Queen Mary
University of London

What is a linear model?)

Further examples

$y_i = \frac{1}{\beta_0+\beta_1 x_{1i}+\beta_2 x_{2i}+\varepsilon_i}$ is not linear in its parameters

This can be linearised by inverting the response as long as we are prepared to accept the condition that $y_i \neq 0$

$$\frac{1}{y_i} = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$$

Queen Mary University of London