Further Model Check (Statistical Modelling I)

Lubna Shaheen

Week 4, Lecture 2

Further Model Check

Outline

1. Influential Observations
2. Cook's Statistic
3. Transforming the response variable
4. Pure Error and Lack of Fit

Distinction Between Outliers and High Leverage Observations

Outlier: An outlier is a data point whose response y does not follow the general trend of the rest of the data.

Outlier is an unusual y value.

High Leverage observation: A data point has high leverage if it has "extreme" predictor x values. With a single predictor, an extreme x value is simply one that is particularly high or low.

High Leverage Observations are unusual x values.

These observations are not ones we necessary want to remove from the model, but it is good to know they are there and what effect they are having on the model output. This will become an even greater issue when we consider Multiple Linear Regression models later in the module.

For now we will look at how to detect so called **influential observation**.

Influential Observations

Definition: An **influential observation** is one which, when included in the dataset used to fit a model, alters the regression coefficients by a meaningful amount.

Remark:

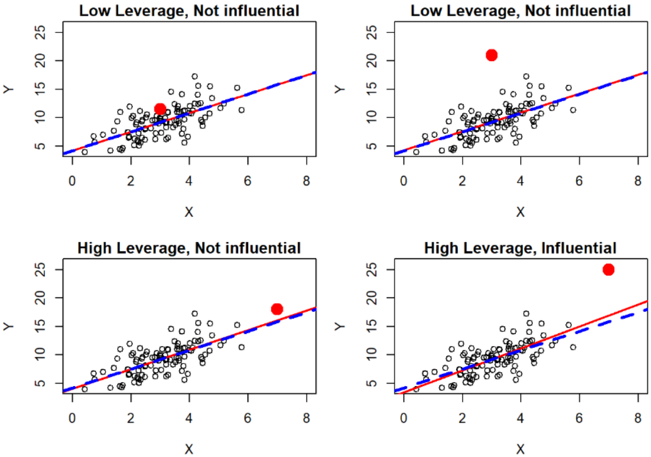
1. **High leverage values–** have the potential to influence the line greatly. If you were to take a high leverage observation and change its y value by a certain amount, the regression line would change more than if the same change were made to a low leverage observation (one near the center of the distribution of x)

2. **Not all high leverage data are truly influential-** those that are significant distance from the regression line during model fit will have a greater impact. Conversely, some less high leverage observations can be influential if their **residuals are unusually large**. The magnitude of influence is determined by the combination of **leverage and magnitude of residual**.

Unusual x_i value

To check if an observation is influential, you must fit the model with and without the observation and see how the regression coefficients change

Dashed line –without the point, **Solid line— include the point**



How the Influential observations affect the linear regression line

Influential observations pull the regression line towards themselves.

1. With and without influential observations in the analysis, the outcome predictions, parameter estimates, confidence intervals, and p-values can differ significantly.
2. While influential observations do not necessarily violate any regression assumptions, they can cast doubt on the conclusions drawn from the sample.
3. If a regression model is being used to inform real-life decisions, one would hope those decisions are not overly influenced by just one or a few observations.

Influential Observations, Diagnosis

What makes a point high leverage is how unusual it is when considering all the predictors together. Fortunately, there are diagnostics that assess leverage and influence no matter how many predictors are in the model.

1. **The hat values**: The hat value calculates the distance between an observation's predictors and those of other observations. Large hat values indicate that an observation has significant leverage and may have some influence, but not always.

   We will draw hat values table and use Leverage values formula as cutoff point to find influential observations

2. **Cook's distance**: D_i, is used to find influential outliers in a set of **predictor variables**. In other words, it's a way to identify points that negatively affect your regression model. The measurement is a combination of each observation's leverage and residual values; the higher the leverage and residuals, the higher the Cook's distance.

   A slightly technical way to interpret D_i is to find the potential outlier's percentile value using the F-distribution. A percentile of over 50 indicates a highly influential point.

Influential Observations, Possible Solutions

Common transformation of predictor: A highly skewed x distribution can lead to high leverage for points at the extreme. A transformation of predictor variable (e.g., logarithm, square root, inverse) may solve this problem.

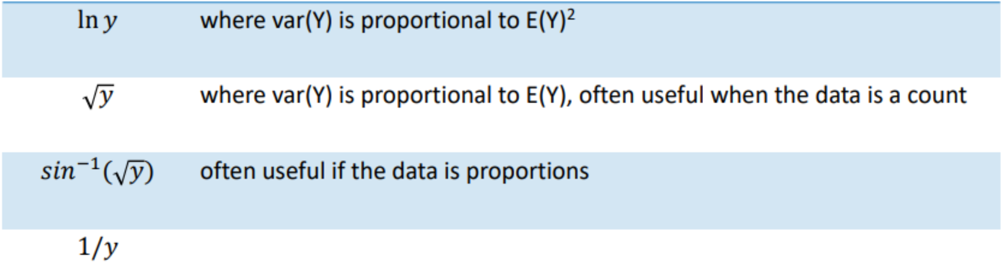
$$y = \ln(x)$$

$$y = \sqrt{x}$$

$$y = \frac{1}{x}$$

Influential Observations, Possible Solutions

Outcome transformation: If the y distribution is very skewed, that can lead to large residuals. A transformation may solve this problem.



$\ln y$	where var(Y) is proportional to E(Y)^2
$\sqrt{y}$	where var(Y) is proportional to E(Y), often useful when the data is a count
$sin^{-1}(\sqrt{y})$	often useful if the data is proportions
$1/y$	

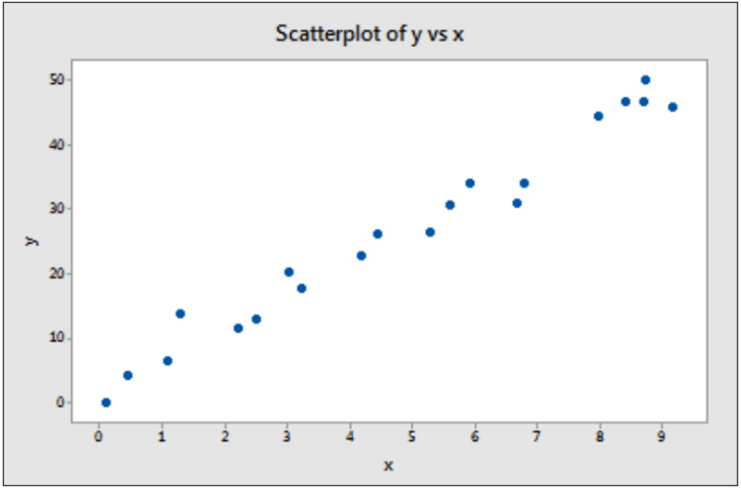
Influential Observations, Possible Solutions

Perform a sensitivity analysis

Fit the model with and without the influential observations and see what changes. If a point is influential based on the diagnostics, that alone does not justify its removal. As with outliers, never simply remove observations from the data just because they might be problematic. Instead, do the analysis with and without them and state the differences in any discussion of the results.

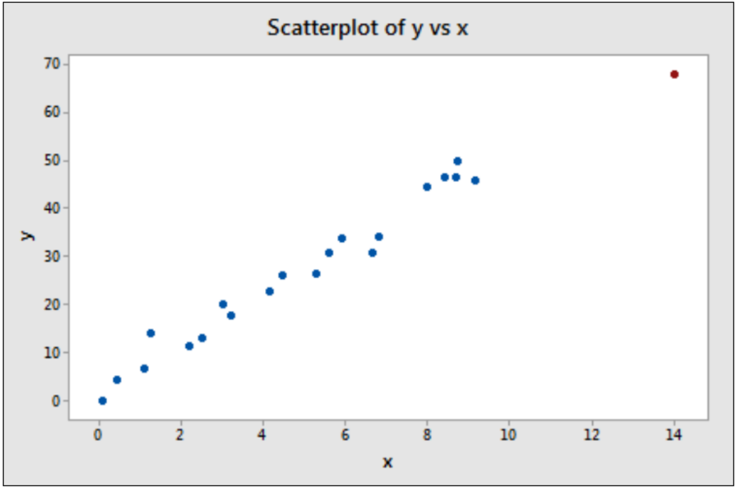
Example 1

Based on the definitions above, do you think the following data set contains any outliers? Or, any high leverage data points?



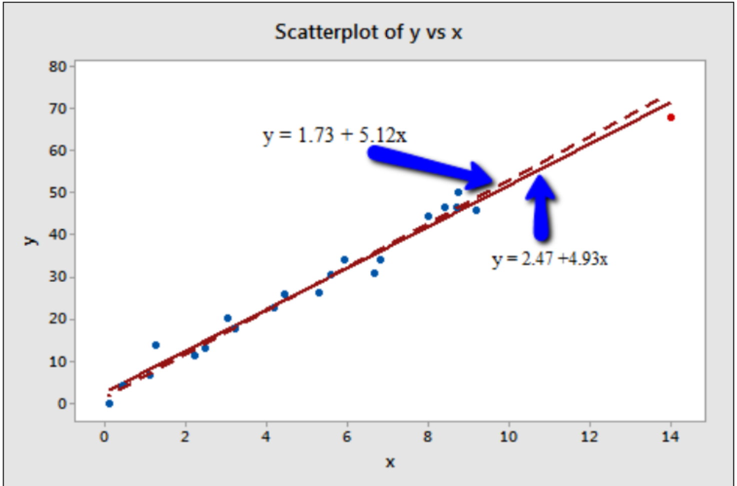
Example 2

Based on the definitions above, do you think the following data set contains any outliers? Or, any high leverage data points?



Example 2

The solid line represents the estimated regression equation with the red data point included, while the dashed line represents the estimated regression equation with the red data point taken excluded.

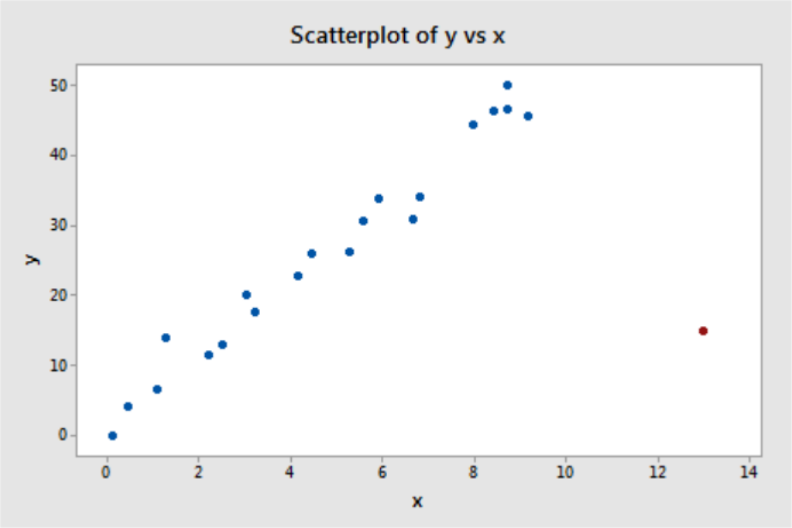


Example 2

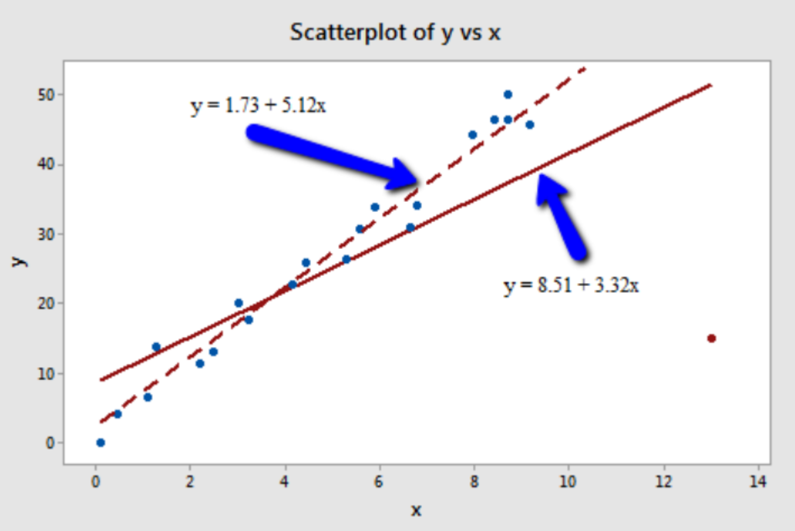
By looking at the R out puts we have obtained when the red data point included or excluded.

1. The R^2 value has hardly changed at all, increasing only slightly from 97.3% to 97.7%
2. In either case, the relationship between y and x is strong.
3. The standard error of β_1 is about the same in each case — 0.172 when the red data point is included, and 0.200 when the red data point is excluded.
4. Therefore, the width of the confidence intervals for β_1 would largely remain unaffected by the existence of the red data point. You might take note that this is because the data point is not an outlier heavily impacting MSE.
5. In each case, the p-value for testing $H_0 : \beta_1 = 0$ is less than 0.001. In either case, we can conclude that there is sufficient evidence at the 0.05 level to conclude that, in the population, x is related to y.

Example 3



Example 3


Scatterplot of y vs x
y = 1.73 + 5.12x
y = 8.51 + 3.32x
y
x
0
10
20
30
40
50
0
2
4
6
8
10
12
14

Example 3

By looking at the R out puts we have obtained when the red data point included or excluded.

1. Here the R^2 value has increased substantially from 55.19% to 97.32%. Therefore, if we include the red data point, we conclude that the relationship between y and x is only moderately strong, whereas if we exclude the red data point, we conclude that the relationship between y and x is very strong.
2. The standard error also is almost 3.5 times larger when the red data point is included i.e. increasing from 0.20 to 0.686. This increase would have a substantial effect on the width of our confidence intervals too. Again, the increase is because the red data point is an outlier in the y direction.
3. In each case, the p value for testing $H_0 : \beta_1 = 0$ is less than 0.001. In both the cases, we can conclude that there is sufficient evidence present at the 0.05 level to conclude that, in the population, x is related to y as largely the data points are in favor of it.

Standardised residuals and Leverage

The standardised residuals are given by

$$d_i = \frac{e_i}{[s^2(1-\nu_i)]^{\frac{1}{2}}}$$

where

$$\nu_i = \frac{1}{n} + \frac{(x_i - \overline{x})^2}{S_{xx}}$$

ν_i us known as the leverage of an observation

$$\nu_i = \frac{1}{n} + \frac{(x_i - \overline{x})^2}{S_{xx}}$$

Now $\sum_i \nu_i = 2$.

Because each of the 2 terms in ν_i sum to 1 over the n observations. Which means that the average leverage for an observation is $\frac{2}{n}$.

Rule for identifying leverage points ?

Average leverage for an observation is $\frac{2}{n}$.

- Leverage $> \frac{4}{n}$ (twice average) is "large leverage"
- Leverage $> \frac{6}{n}$ (three times average) is "very large leverage"

If we have more than two parameters

- Large leverage values are above $2 * \frac{k}{n}$
- Very high leverage $> 3 * \frac{k}{n}$ (where k is the number of estimated parameters).

Strategies for dealing with ”leverage points”

(i) **Remove invalid data points** (Cook’s Distance)
(ii) **Fit a different regression model** (Transformations)

What does this mean for our model? Large (or very large) leverage observations are called influential observations above. We discussed how they effect the linearity of the model.

- Are influential
- whether they are included or not causes a large change in the β parameters
- **we can measure this influence using Cook’s Statistic**
- which is usually designated D_i
- this compares the linear regression results with and without the influential observation

Cook's Statistic

We discussed here the first (i) Remove invalid data points.

Cook's Statistic: For observation i where $i = 1, 2, \cdots, n$ from our (x_i, y_i) observations

- first complete the linear regression as usual to obtain $\widehat{\beta}_0$, $\widehat{\beta}_1$ and hence the fitted $\widehat{y}$ values
- then take out the one i^{th} observation
- repeat the linear regression to get new $\widehat{\beta}_0, \widehat{\beta}_1$ and hence new fitted values which we will call $\widehat{y}^{(i)}$

Cook's Statistic

The Cook's Statistics for this i^{th} observation is

$$D_i = \frac{1}{2S^2} \sum_{j=1}^{n} (\widehat{y_j}^{(i)} - \widehat{y_j})^2$$

Where there will be a separate value for D_i for each of our n observations.

Now it can be shown that this statistic is related to the leverage v_i of the same observation.

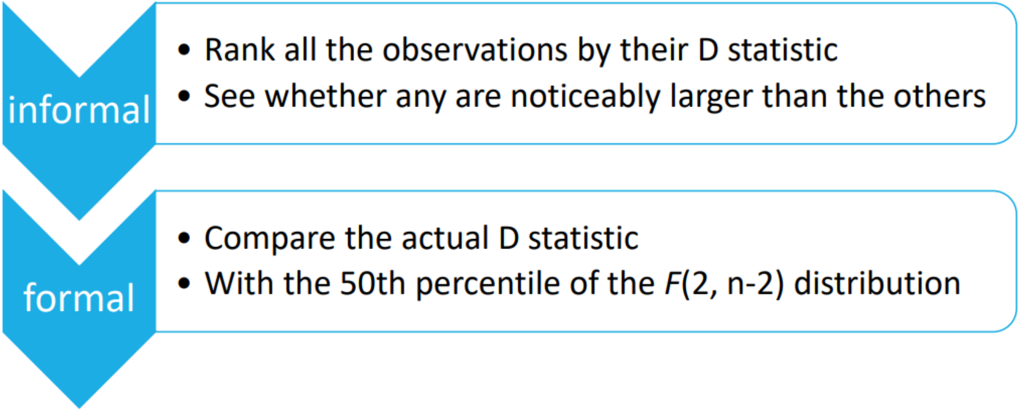
Cook and Leverage

$$D_i = \frac{1}{2} d_i^2 \frac{v_i}{1 - v_i}$$

So Cook's statistic depends on

- the standardised residual for an observation
- and its leverage

Using Cook's Statistic



informal

- Rank all the observations by their D statistic
- See whether any are noticeably larger than the others

formal

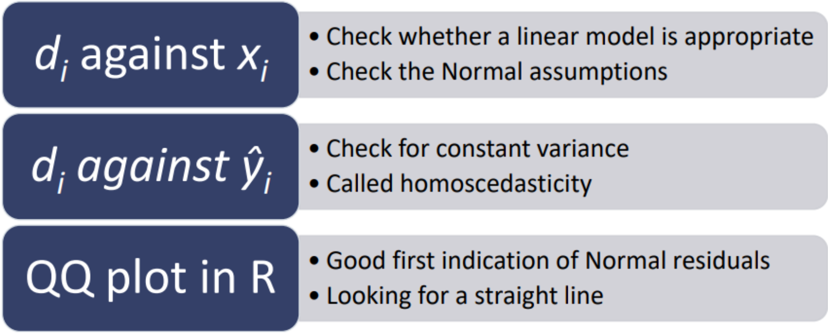
- Compare the actual D statistic
- With the 50th percentile of the *F*(2, n-2) distribution

What to do with influential observations

We don't need to remove influential observations in same way as outliers. But when we present the results of a modelling study that includes influential observations we should

- highlight the observation(s)
- indicate how much they have affected the model output and conclusions

Remember the residual plots in weeks 2 & 3



d_i against x_i

- Check whether a linear model is appropriate
- Check the Normal assumptions

d_i against $\hat{y}_i$

- Check for constant variance
- Called homoscedasticity

QQ plot in R

- Good first indication of Normal residuals
- Looking for a straight line

Transforming the response variable

What should be do if one or more of these plots shows an issue?

Transforming the response variable

If we doubt the $x \longrightarrow y$ relationship is linear

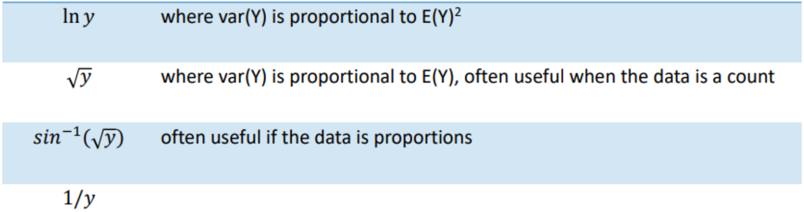
Or we doubt the variance of y is constant

Or we doubt the data is from a Normal distribution

Then good first thing to try is a simple transformation of the y_i

The most usual transformation (if no negative data) is $\ln(y)$

Common transformations



$\ln y$	where var(Y) is proportional to E(Y)^2
$\sqrt{y}$	where var(Y) is proportional to E(Y), often useful when the data is a count
$sin^{-1}(\sqrt{y})$	often useful if the data is proportions
$1/y$	

Transformation Questions

Question 1

We will use the Janka dataset described during the Practical in R sessions of this week. In the Practical, we have seen that the Shapiro-Wilk test had a small p-value, thus implying that the assumption of normality is not supported by the data. Moreover, the variance may be increasing suggests we should transform the dependent variable hardness. The first transformation we usually try is to take the logarithm of y.We aim today to transform the data to the following transformations.

(a) The first transformation we usually try is to take the **logarithm of y.**

`modly<- lm(ly∼x)`. Look at the Summary and anova.

(b) It may happen that the plot of standardized residuals vs x suggested a quadratic model may be needed. To fit a polynomial model of degree 2, i.e. a quadratic model we use the command

```
modlyq <- lm(ly ∼ poly(x,2,raw=TRUE))
```

Have a look at the summary and anova. The summary shows the t tests for the intercept, linear and quadratic terms. The anova table is testing whether at least one of the linear and quadratic terms are needed.

Transformation Questions

(c) Save the standardised residuals and fitted values for this model and check the residuals versus fitted values plot and the QQ-plot. Does this model seem to fit ok?

(d) Check the leverage values for this model.

```
hat <- hatvalues(modlyq)
cook<-cooks.distance(modlyq)
i<-1:36
plot(i,hat, main="Leverage values")
plot(i,cook, main="Cooks distance values")
```

Note that as there are now three parameters in the model the relevant figures for high and very high leverage are $\frac{6}{n} = \frac{2p}{n}$ and $\frac{9}{n} = \frac{3p}{n}$. Similarly the cut-off for Cook's distance is the 50% point of an F on 3 and 33 degrees of freedom.

Transformation Questions

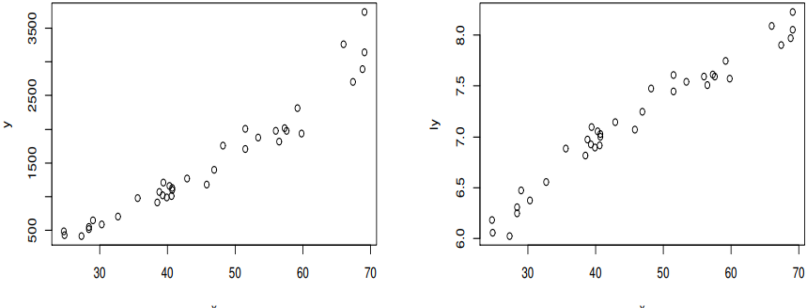
Solution: The first transformation we usually try is to take the logarithm of y.

```
ly <- log(y)
```

Now fit a simple linear regression model with ly as the dependent variable.

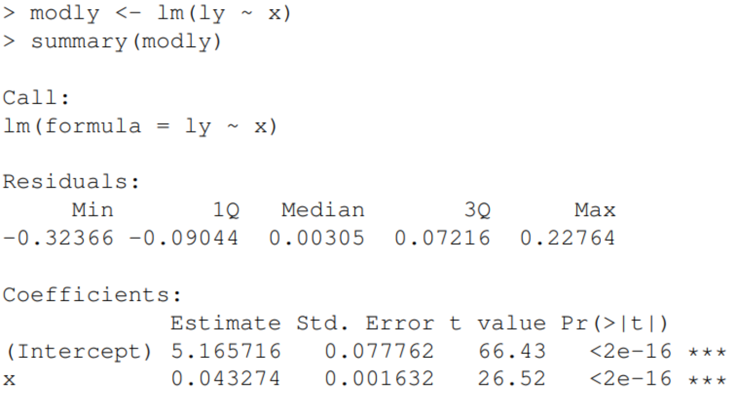
```
modly<- lm(ly~x)
```

In the previous Practical, we have seen that the Shapiro-Wilk test had a small p-value, thus implying that the assumption of normality is not supported by the data.



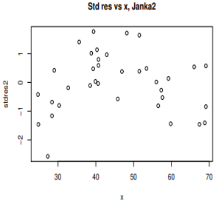
Transformation Questions

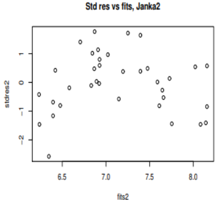
The model with log(y) is an improvement. The fitted model is log(yi) = 5.166 + 0.00433xi with an R^2 equal to 95.39% thus explaining a bit more variation than the original model.



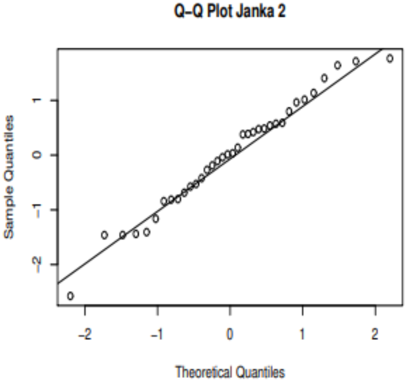
```
> modly <- lm(ly ~ x)
> summary(modly)

Call:
lm(formula = ly ~ x)

Residuals:
     Min        1Q    Median        3Q       Max
-0.32366 -0.09044  0.00305  0.07216  0.22764

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 5.165716   0.077762   66.43   <2e-16 ***
x           0.043274   0.001632   26.52   <2e-16 ***
```

Transformation Questions



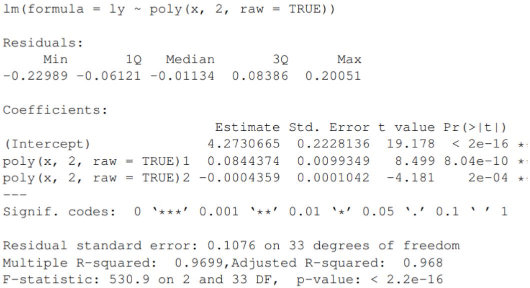


From the Shapiro-Wilk test there is no evidence against the normality (p-value of 0.7159), which can be see also from the Q-Q plot. However, looking at the plot of the standardized residuals against x shows curvature and suggests that fitting a quadratic term in x might be of value.

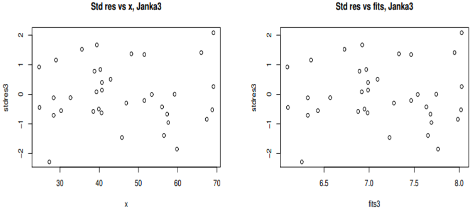


Transformation Questions

We run a model with the logarithm of y and a quadratic model for the X. Thus the model is of the form $log(y_i) = \alpha + \beta_0 x_i + \beta_1 x_i^2 + \varepsilon_i$.



```
lm(formula = ly ~ poly(x, 2, raw = TRUE))

Residuals:
     Min       1Q   Median       3Q      Max 
-0.22989 -0.06121 -0.01134  0.08386  0.20051 

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)              4.2730665  0.2228136  19.178  < 2e-16 *
poly(x, 2, raw = TRUE)1  0.0844374  0.0099349   8.499 8.04e-10 *
poly(x, 2, raw = TRUE)2 -0.0004359  0.0001042  -4.181    2e-04 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1076 on 33 degrees of freedom
Multiple R-squared:  0.9699,Adjusted R-squared:  0.968 
F-statistic: 530.9 on 2 and 33 DF,  p-value: < 2.2e-16
```



The fitted model is $log(y) = 4.27 + 0.08x - 0.00044x^2$. All the parameters are highly significant. Moreover looking at the R^2 we have that the variation is explained mostly by this model, since the R^2 is equal to 96.99%.

Transformation Questions

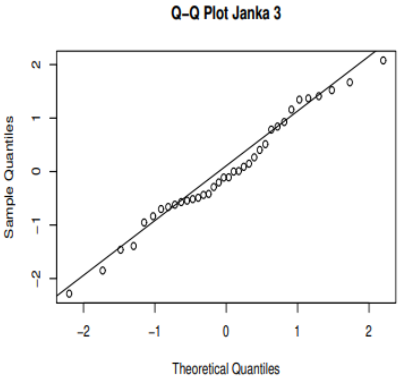
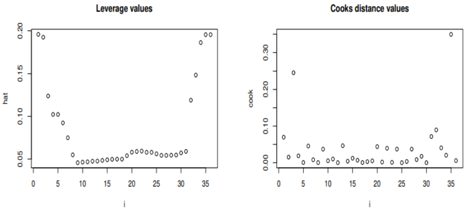


Figure 1.5: Plot of Q-Q plot.

Looking at the Shapiro-Wilk test, the assumption of the normality is not rejected. In fact the p-value is around 0.71 and this is confirmed by the Q-Q plot. Looking at the standardized residuals versus x and versus the fitted values, it seems random and it looks okay.

Transformation Questions



Plot of the Leverage values (left) and of the Cook's distance (right).

We have the largest leverage values are above $2 * k/n$ but not $3 * k/n$, where k is the number of estimated parameters. In our case, we have k equal to 3, the total number of observation n equal to 36, thus $6/n$ and $9/n$.

In numbers, we have high leverage if it is greater than $6/36 = 0.16$ and very high leverage if greater than $9/36 = 0.25$. The largest Cook's distance is for observation 35 but at about 0.35 it is well below the value of 0.805 (which is the F value with 3 and 33 degrees of freedom), which would indicate a highly influential observation.

In conclusion, we can see that the model with quadratic term for x fits very well the data.

Pure Error and Lack of Fit

We have listed the scenarios where a simple linear regression model might not be appropriate:

- residuals not from Normal distribution
- variance not constant

But so far we have relied on looking at residual plots to assess this

- Is there some more formal way to show this lack of fit?
- We will now look at one type of case of poorly fitting model

Replications

Replications are where we have more than one observation with the same x value but they have different y values
We use $y_{ij}$to be the j^{th} observation at x_i

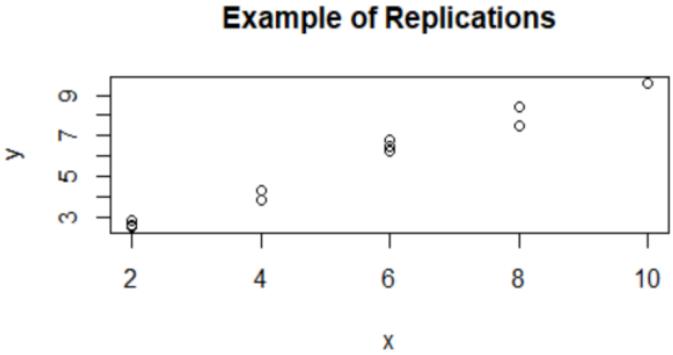
where $i = 1, 2, \cdots m$ and $j = 1, 2, \cdots , n_i$

In our linear regression model, although each of the y_{ij} observations might be different at a certain x_i, the fitted value will be the same $\hat{y}_i$ for all

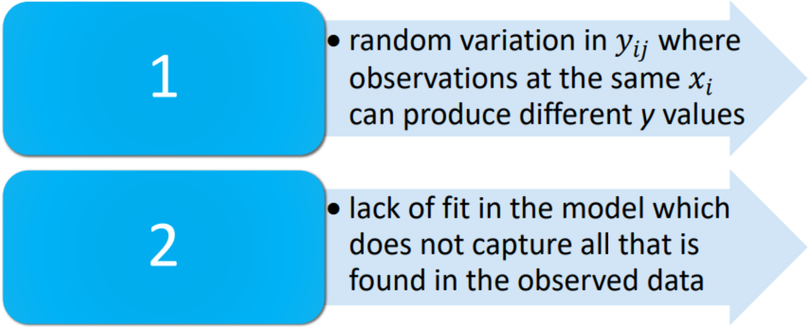
Residuals

The residuals are now

$$e_{ij} = y_{ij} - \hat{y}_i$$



Two sources of Residual Error



1

- random variation in y_{ij} where observations at the same x_i can produce different y values

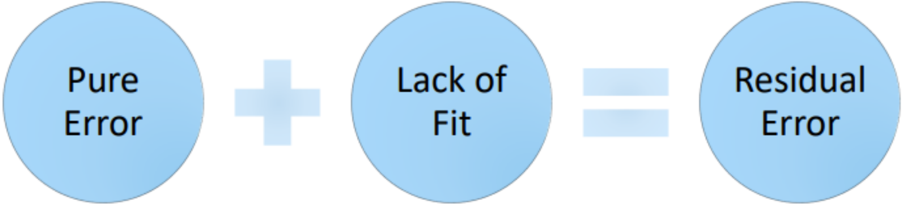
2

- lack of fit in the model which does not capture all that is found in the observed data

Two sources of Residual Error

Pure Error = $y_{ij} - \bar{y}_i$

Lack of Fit = $\bar{y}_i - \hat{y}_i$



Pure Error + Lack of Fit = Residual Error

Residual Sum of Squares

We can split the residual sum of squares SS_E (from our ANOVA table) into:

- Pure Error sum of squares SS_{PE}
  —— measures overall random variation
- Lack of Fit sum of squares SS_{LoF}
  —— measures overall model lack of fit

Residual Sum of Squares

With the i, j notation SS_E becomes

$$SS_E = \sum_{i=1}^{m} \sum_{j=1}^{n_i} (y_{ij} - \hat{y}_i)^2$$

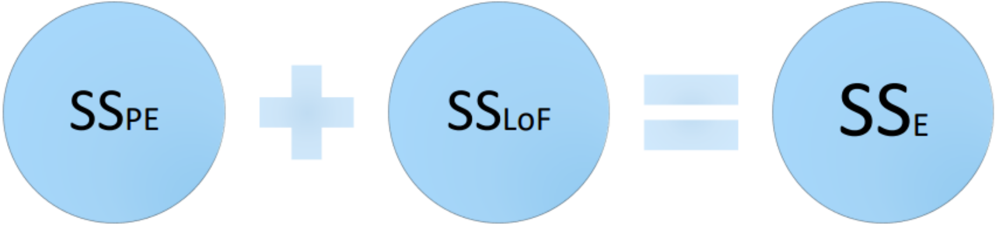
And this can be split between:

$$SS_{PE} = \sum_{i=1}^{m} \sum_{j=1}^{n_i} (y_{ij} - \overline{y_i})^2$$

$$SS_{LOF} = \sum_{i=1}^{m} \sum_{j=1}^{n_i} (\overline{y_i} - \hat{y}_i)^2 = \sum_{i=1}^{m} n_i (\overline{y_i} - \hat{y}_i)^2$$

Residual Sum of Squares

In the simple linear regression model with replications



$SS_{PE} + SS_{LoF} = SS_E$

Expanded ANOVA table with replications

Using Pure Error and Lack of Fit we can expand the ANOVA table

Where there are replications

Splitting the Residual Sum of Squares SS_E

Note the Regression Sum of Squares entry is unchanged

Degrees of freedom

To calculate pure error we need m means for the $\overline{y_i}$ $(i = 1, 2, \cdots, m)$

Each of these mean calculations takes a degree of freedom

Therefore degrees of freedom for Pure Error $= n - m$

Previously Residuals had $n - 2$ d.f.

Therefore degrees of freedom for Lack of Fit $= (n - 2) - (n - m) = m - 2$

Mean Squares

We will see later that

$E[SS_{PE}] = (n - m)\sigma^2$ whether the model assumptions are true or not
$E[SS_{LoF}] = (m - 2)\sigma^2$ if the model assumption are true

Which means that:

- MS_{PE} gives an unbiased estimator of σ^2
- MS_{LoF} gives an unbiased estimator of σ^2 if the model assumptions are true

Variance Ratio

Therefore in all cases

$$\frac{(n-m)MS_{PE}}{\sigma^2} \sim \mathcal{X}^2_{n-m}$$

and if the model assumptions are true

$$\frac{(m-2)MS_{LoF}}{\sigma^2} \sim \mathcal{X}^2_{m-2}$$

We can now use the ratio of these two divided by their respective d.f. to calculate another Variance Ratio

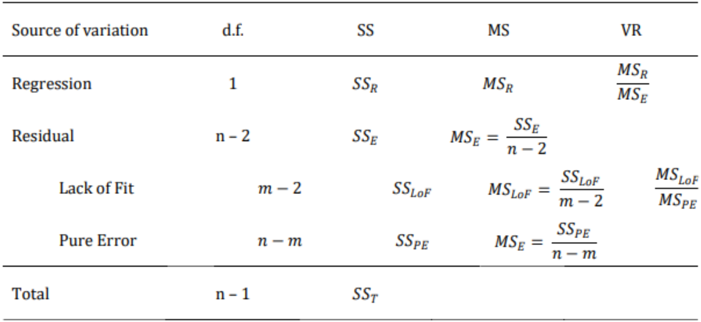
Variance Ratio for residuals

If the regression model assumption are true

$$\frac{MS_{LoF}}{MS_{PE}} \sim F^{m-2}_{n-m}$$

We are now able to construct an expanded ANOVA table for the case where there are replications

ANOVA

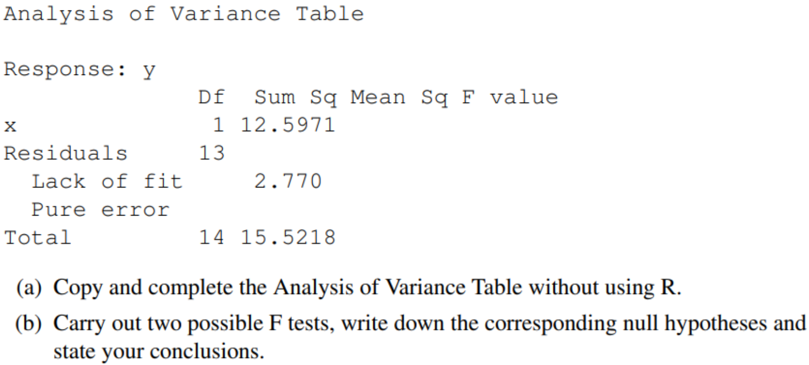


Source of variation	d.f.	SS	MS	VR
Regression	1	SS_R	MS_R	$\frac{MS_R}{MS_E}$
Residual	n – 2	SS_E	$MS_E = \frac{SS_E}{n-2}$	
Lack of Fit	$m-2$	SS_{LoF}	$MS_{LoF} = \frac{SS_{LoF}}{m-2}$	$\frac{MS_{LoF}}{MS_{PE}}$
Pure Error	$n-m$	SS_{PE}	$MS_E = \frac{SS_{PE}}{n-m}$	
Total	n – 1	SS_T		

Expanded ANOVA table

Example

A chemist studied the concentration of a solution (Y) over time (x). Fifteen identical solutions were prepared. The solutions were randomly divided into five sets of three, and the five sets were measured, respectively after 1, 3, 5, 7, and 9 hours. Without making any plots the chemist entered the data into R, fitted a simple linear regression model and then carried out a goodness of fit test. The following is the Analysis of Variance table she produced but with some figures missing.



```
Analysis of Variance Table

Response: y
              Df  Sum Sq Mean Sq F value
x              1 12.5971
Residuals     13
  Lack of fit     2.770
  Pure error
Total         14 15.5218
```

(a) Copy and complete the Analysis of Variance Table without using R.

(b) Carry out two possible F tests, write down the corresponding null hypotheses and state your conclusions.

ANOVA

ANOVA