Week ON4

Elmer V Villanueva

16 March 2020

SET YOUR WORKING DIRECTORY!

```
setwd("C:/Users/1/Dropbox/00 - Working Folder/Teaching/DPH112/2019-2020/Week ON4")
```

Announcements and clarifications

- From this week onward, all assignments will require the submission of RMD files. Previously, I allowed the submission of assignments in many alternate formats. However, as we progress further into the course, RMB offers the best mix of text and interactivity. There will be no exceptions.
- I have taken pains to ensure that you *understand* the results of the software that you are using. This goes to your reporting of the output of your work. For example, in question 8 of Assignment ON3, you were asked to calculate confidence intervals for three new values. The answers are below:

Table. Estimated mean and 95% confidence interval tuberculin antibody levels under the old method for specific results under the new method.

Estimate under NEW method	Mean	95% confidence interval
200	5.31	(-207.92, 218.54)
500	45.68	(-158.26, 249.62)
750	79.32	(-117.28, 275.93)

These are taken from output from R. However, is it really possible for anyone to have a tuberculin antibody level of a negative number? Of course not! Realistically, the correct answer should be

Estimate under NEW method	Mean	95% confidence interval
200	5.31	(0, 218.54)
500	45.68	(0, 249.62)
750	79.32	(0, 275.93)

However, I allowed these impossible numbers to be reported because the point of the exercise was to derive and interpret confidence intervals. This demonstrates, yet again, the need for you to know and understand the topic which you are analysing.

- I have found a new resource called Grammarly, which automatically checks your spelling and grammar. I have not explored this resource thoroughly and I have not consulted with my colleagues in the Language School about its validity. Nevertheless, resources such as this are easy ways for you to correct your spelling and grammar mistakes.
- Once the online phase ends and face-to-face classes begin, it is likely that I will *NO LONGER BE*

YOUR INSTRUCTOR. As you know, I am Director of the XJTLU Graduate School. In this role, I am in charge of almost 1,500 Master's students and 300 PhD students. Recently, I have been asked to take over the Centre for Academic Affairs, which includes five units – the Graduate School, Registry, Research Office, Academic Services Office and the Educational Quality Assurance Office. I will give up my teaching altogether to focus on this new role. Once a new instructor is assigned, I will let you know.

- Students who find new errors in the learning materials, assignment or assignment solutions should bring these to my attention via email. Genuine errors will be corrected and extra marks will be awarded.
- The final graph in the Week ON3 assignment solutions is *purposefully* incomplete. The incompleteness was apparent in the Week ON3 learning materials, too. The *first* student that provides me with appropriate R code that produces a *complete* and properly formatted graph will be awarded extra marks. Good luck!

Reading

Read and understand Vittinghoff et al., Chapter 3.3.

Review of previous learning

In the past few weeks, we introduced covariance, correlation and simple linear regression. Our discussion of the latter included the development and specification of the model and the calculation of predictions from our work. There is a big assumption here: we are assuming that the model that we have developed is appropriate to our data. In this *final* lecture on simple linear regression, we will learn to test the "fitness" of our model. In doing so, we will learn to recognise various problems and offer corrective measures.

The lectures in subsequent weeks will follow this pattern. First, we will learn to specify the model. Then, we will learn to assess the "goodness" of the model. It is often the case that novice analysts only focus on the first step. This step is the easier of the two. For example, specifying the model relating the outcome Y to the predictor X in R is a single line:

```
> DATA.LM -< lm(Y ~ X, data = DATA)
```

Novice analysts consider their work complete at this point and use this model in the (potentially mistaken) assumption that the regression equation $\hat{y_i} = b_0 + b_1 x_i + e_i$ is without any problems.

You must learn to build models *and test them.* This second step is called *model diagnostics.*

Assumptions of the simple linear model

We touched briefly on the assumptions of the simple linear model in a previous lecture. For this module, there are at least four assumptions of the simple linear model that we need to consider. For the simple linear model that arises from the regression of Y on X:

1. the *means* of the subpopulation of Y all lie on the same straight line. This is known as the *assumption of linearity.*
2. the Y values are statistically *independent.* In other words, in drawing the sample, it is assumed that the values of Y chosen at one value of X in no way depend on the values of Y chosen at another value of X.
3. the subpopulation of Y are *normally distributed.*
4. the variances of the subpopulations of Y are all equal. This is known as the *assumption of homoskedasticity.*

There is an easy mnemonic to allow students to remember these assumptions – **LINE**

- **L** - inearity

- **I** - ndependence
- **N** - ormal dsitrbution
- **E** - qual variance

The assumption of independence is often assessed non-statistically. That is, the design of the experiment will give clues as to the presence or absence of statistical independence. In the contact of paired data, for example, the assumption is violated.

For the remaining three assumptions – linearity, normality and homoskedasticity – we need to learn about *residuals*.

Residuals

Consider our general model for simple linear regression presented previously $Y = \beta_0 + \beta_1 X$ estimated by the equation $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$. In turn, this is operationalised as $y_i = b_0 + b_1 x_i + e_i$.

Solving for e_i gives the equation $e_i = y_i - (b_0 + b_1 x_i) = y_i - \hat{y_i}$. The quantity e_i is known as the ith *residual* and is the difference between the observed value y_i and the predicted or fitted value $\hat{y_i}$.

In R, we can use the `resid()` function to calculate the residuals following estimation of the simple linear regression model.

Example 1A

Let us use Example 1A from last week. Recall that the head of a lab wanted to know how long it took for 25 lab technicians to process blood samples for the presence of a particular genetic marker. Her data appear below.

Table 1. Lot size and hours to completion of DNA analysis

Lot Size	Hours
80	399
30	121
50	221
90	376
70	361
60	224
120	546
80	352
100	353
50	157
40	160
70	252
90	389
20	113
110	435
100	420
30	212
50	268
90	377
110	421
30	273
90	468
40	244
80	342

Lot Size	Hours
70	323

Let's enter the data and produce a scatterplot.

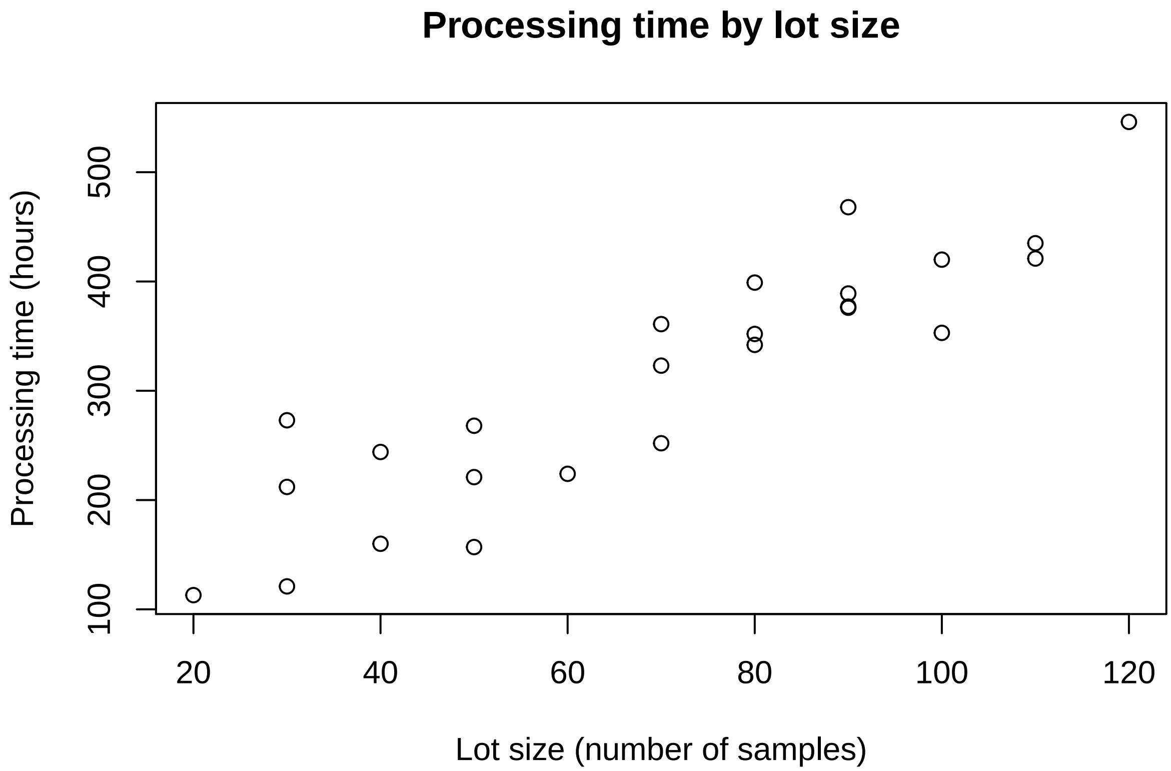
```
LOT <- c(80, 30, 50, 90, 70, 60, 120, 80, 100, 50, 40, 70, 90,
         20, 110, 100, 30, 50, 90, 110, 30, 90, 40, 80, 70)
HOURS <- c(399, 121, 221, 376, 361, 224, 546, 352, 353, 157, 160, 252, 389,
           113, 435, 420, 212, 268, 377, 421, 273, 468, 244, 342, 323)
LAB <- data.frame(LOT, HOURS)
str(LAB)
```

```
## 'data.frame':    25 obs. of  2 variables:
##  $ LOT  : num  80 30 50 90 70 60 120 80 100 50 ...
##  $ HOURS: num  399 121 221 376 361 224 546 352 353 157 ...
```

```
head(LAB)
```

```
##   LOT HOURS
## 1  80   399
## 2  30   121
## 3  50   221
## 4  90   376
## 5  70   361
## 6  60   224
```

```
plot(HOURS ~ LOT, data = LAB,
     main = "Processing time by lot size",
     ylab = "Processing time (hours)",
     xlab = "Lot size (number of samples)")
```



We can calculate the residuals in this manner

```
LAB.LM <- lm(HOURS ~ LOT, data = LAB)
LAB.RESID <- resid(LAB.LM)
summary(LAB.RESID)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -83.876 -34.088  -5.982   0.000  38.826 103.528
```

Note that the descriptive statistics of the residuals are automatically given in the output of the linear model. We have just been ignoring this information until today

```
summary(LAB.LM)
```

```
##
## Call:
## lm(formula = HOURS ~ LOT, data = LAB)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -83.876 -34.088  -5.982  38.826 103.528
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   62.366     26.177   2.382   0.0259 *
## LOT            3.570      0.347  10.290 4.45e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 48.82 on 23 degrees of freedom
## Multiple R-squared:  0.8215, Adjusted R-squared:  0.8138
## F-statistic: 105.9 on 1 and 23 DF,  p-value: 4.449e-10
```

Properties of residuals

Mean. The mean of the n residuals e_i for the simple linear regression model is $\bar{e} = \dfrac{\sum e_i}{n} = 0$.

Variance. The variance of the n residuals e_i is

$$s^2 = \frac{\sum(e_i - \bar{e})^2}{n-2} = \frac{\sum(e_i)^2}{n-2} = \frac{SSE}{n-2} = MSE$$

As noted last week, if the model is appropriate, then the MSE is an unbiased estimator of the true variance σ^2.

Standardised residuals. We can standardise the residuals for better analysis. Since the standard deviation of the residuals is $s = \sqrt{s^2} = \sqrt{MSE}$, then we can consider the following as a typical extension of our standardisation techniue of subtracting the mean and dividing by the standard deviation:

$$e_i^* = \frac{e_i - \bar{e}}{\sqrt{MSE}} = \frac{e_i}{\sqrt{MSE}}$$

We call the statistic e_i^* a *standardised residual* or *internally standardised residual.*

Standardised residuals are derived using the function `rstandard()`.

Studentised residuals. We can calculate another form of standardised residual that is more useful in the detection of outliers. This residual excludes the current data point and calculates the standardised residual as usual. In this form, the residual is known as a *studentised residual* or *jackknifed residual* or *externally standardised residuals.*

Studentised residuals are derived using the function `rstudent()`.

Diagnostics for residuals

Residuals allow us to evaluate whether the assumptions of the model are met. In particular, we will test the assumptions of normality, linearity and equal variance using plots derived from residuals.

IMPORTANT It is important that you understand that these tests are largely visual and your interpretation of the results may vary from my interpretation of results.

IMPORTANT We will demonstrate using ordinary residuals and leave to you the application of these techniques on standardised and studentised residuals.

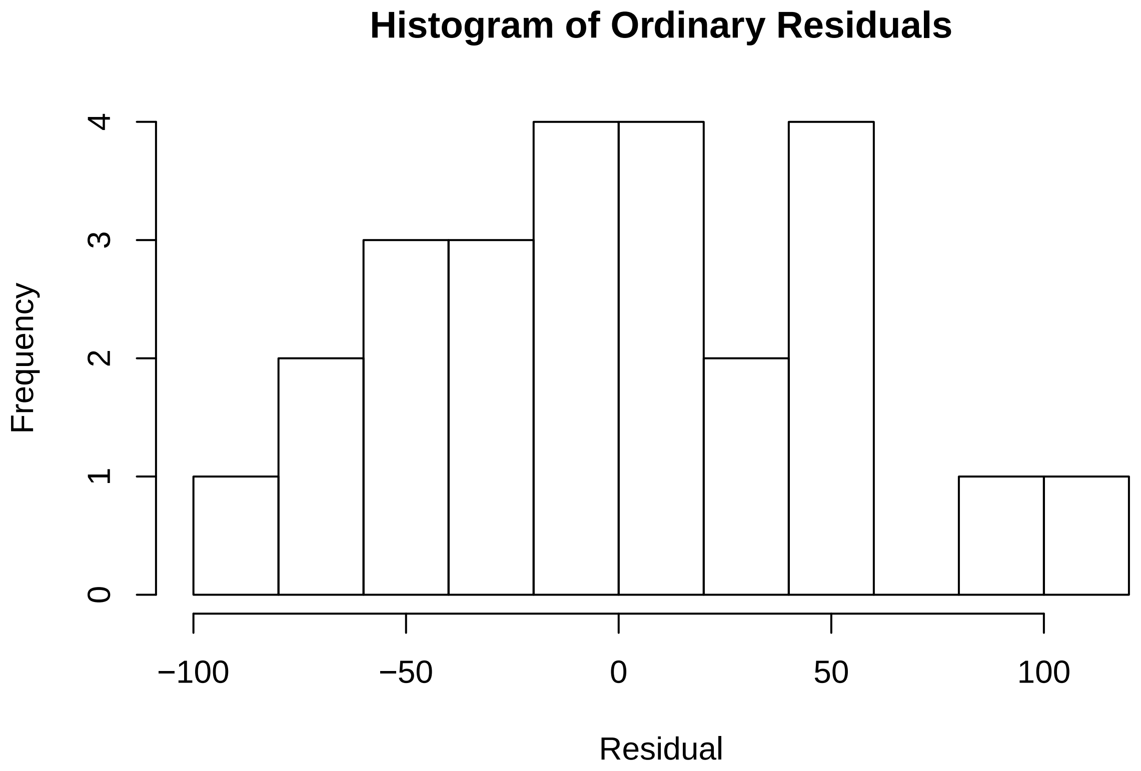
Assessing normality

The assumption of normality is tested using histograms and distributional plots such as the quantile-quantile plot.

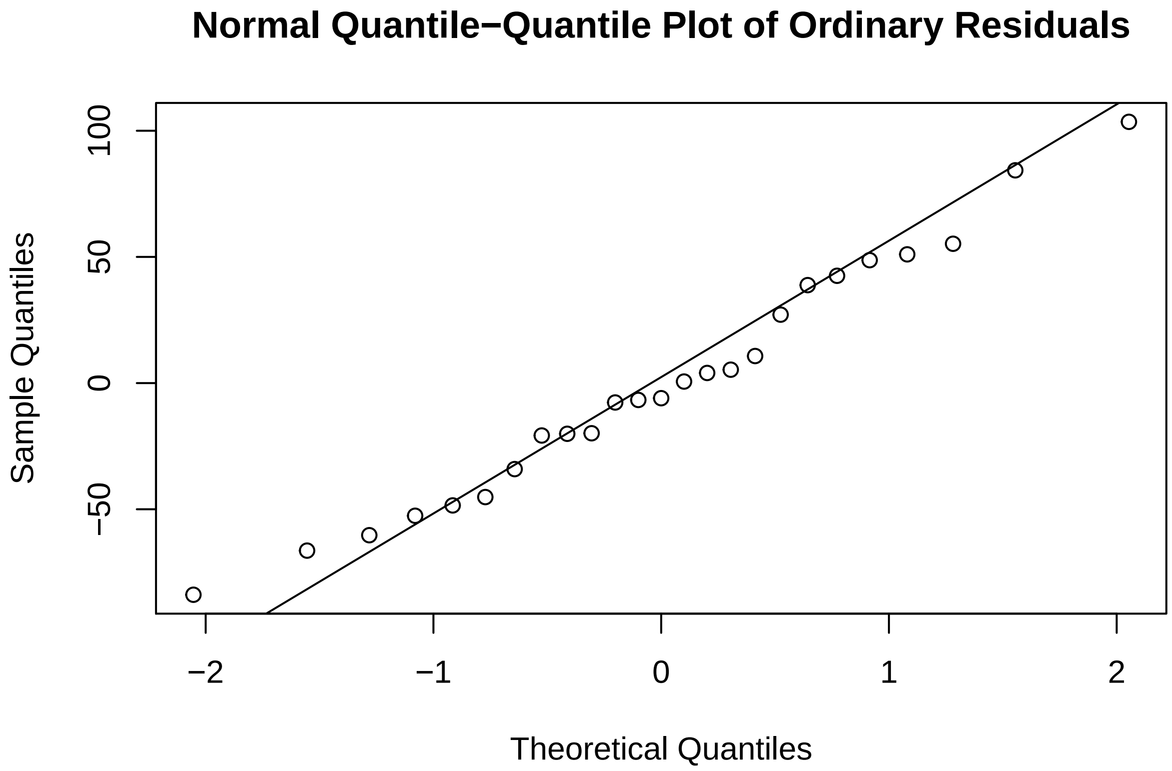
Example 1B

Assess the normality of the regression of processing time on lot size estimated above.

```r
hist(LAB.RESID, breaks = 8,
     main = "Histogram of Ordinary Residuals",
     xlab = "Residual")
```



```
qqnorm(LAB.RESID,
       main = "Normal Quantile-Quantile Plot of Ordinary Residuals")
qqline(LAB.RESID)
```



Assessment. The residuals do not deviate substantially from normality.

Conclusion. The assumption of normality is met.

Assessing linearity

The appropriateness of a *linear* function as a summary of the data can be studied from a residual plot against the fitted values. Such a plot is called the *RvF plot* or the *residuals versus fitted plot*.

If a linear model is appropriate, then the RvF plot will show a random scatter of points around zero (Figure 1). If a linear model is inappropriate, then some pattern (such as that shown in Figure 2) will be clear.

Example 1C

Assess the linearity assumption of the lab model estimated previously.

Let us calculate the fitted values and residuals for our model.

```
FITTED <- predict(LAB.LM, type = "response")
RESID <- resid(LAB.LM)
LAB2 <- cbind(LAB, FITTED, RESID)
str(LAB2)
```

```
## 'data.frame':    25 obs. of  4 variables:
##  $ LOT   : num  80 30 50 90 70 60 120 80 100 50 ...
##  $ HOURS : num  399 121 221 376 361 224 546 352 353 157 ...
##  $ FITTED: num  348 169 241 384 312 ...
##  $ RESID : num  51.02 -48.47 -19.88 -7.68 48.72 ...
```

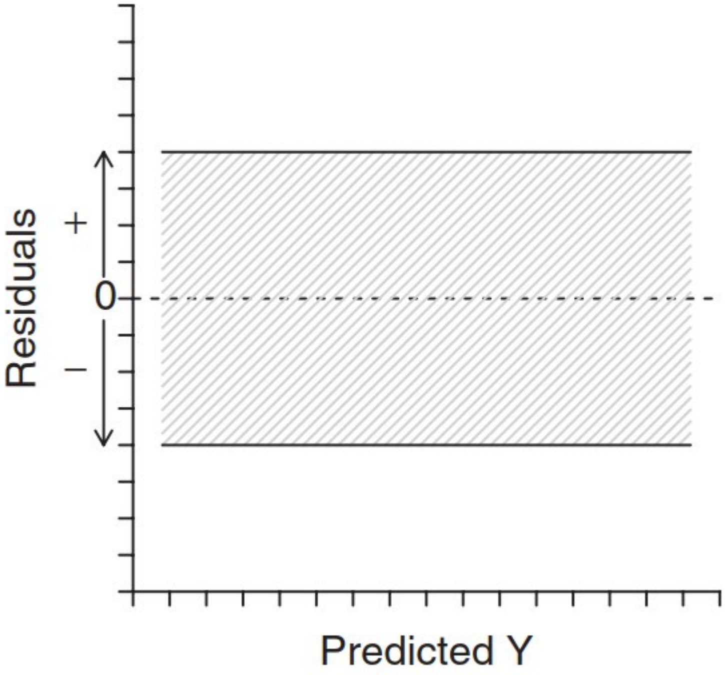


Figure 1: The spread of residuals in an RvF plot will show no clear pattern.

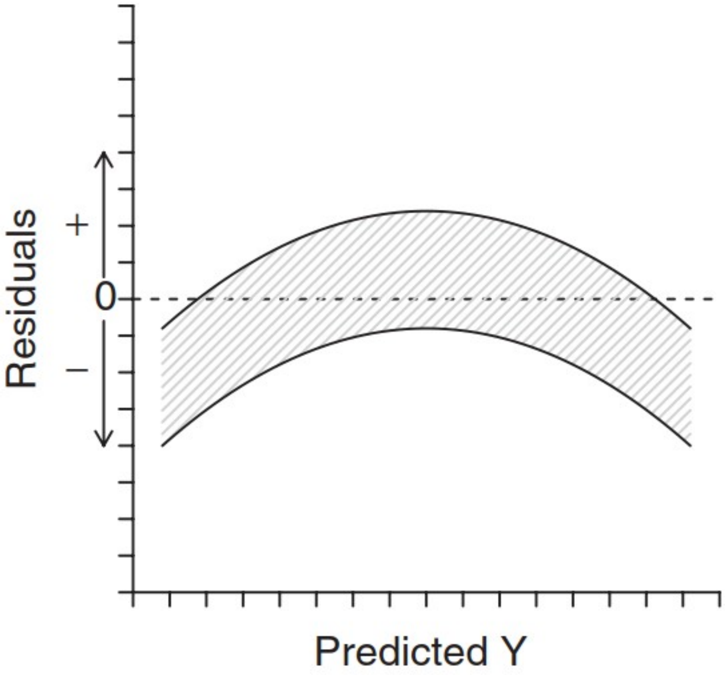


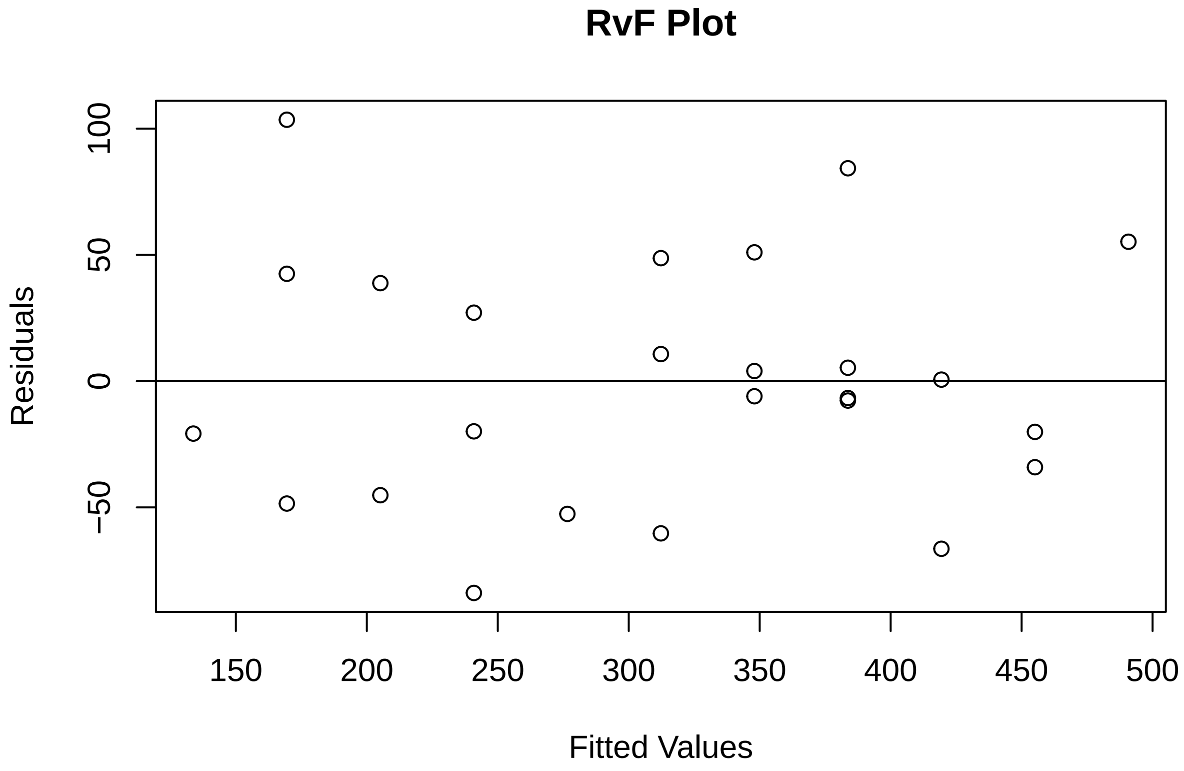
Figure 2: The spread of residuals show a clear non-linear pattern

```
head(LAB2)
```

```
##   LOT HOURS   FITTED     RESID
## 1  80   399 347.9820  51.01798
## 2  30   121 169.4719 -48.47192
## 3  50   221 240.8760 -19.87596
## 4  90   376 383.6840  -7.68404
## 5  70   361 312.2800  48.72000
## 6  60   224 276.5780 -52.57798
```

We can now produce the RvF plot.

```
plot(RESID ~ FITTED, data = LAB2,
     main = "RvF Plot",
     ylab = "Residuals",
     xlab = "Fitted Values")
abline(h = 0)
```



Assessment. The spread of residuals does not show any obvious pattern.

Conclusion. The assumption of linearity is met.

Assessing homoskedasticity

Plots of the residuals against the fitted values are not only helpful to study whether a linear regression function is appropriate but also to examine whether the variance of the error terms is constant. When the variance is equal, the vertical width of the band of residuals is the same throughout the horizontal axis. When this is not the case, then you will observe “trumpeting” (Figure 3) of the graph or some other pattern.

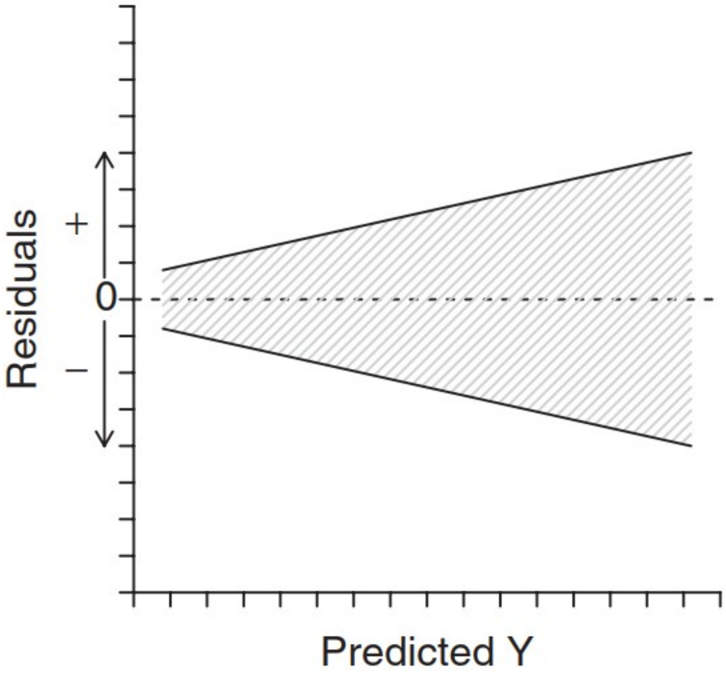


Figure 3: The "trumpeting" effect of unequal variance is clear with smaller values having smaller variance than larger values.

Example 1D

Assess the equal variance assumption of the lab model estimated previously.

Assessment. The RvF plot did not show evidence of trumpeting or other patterns.

Conclusion. The assumption of homoskedasticity is met.

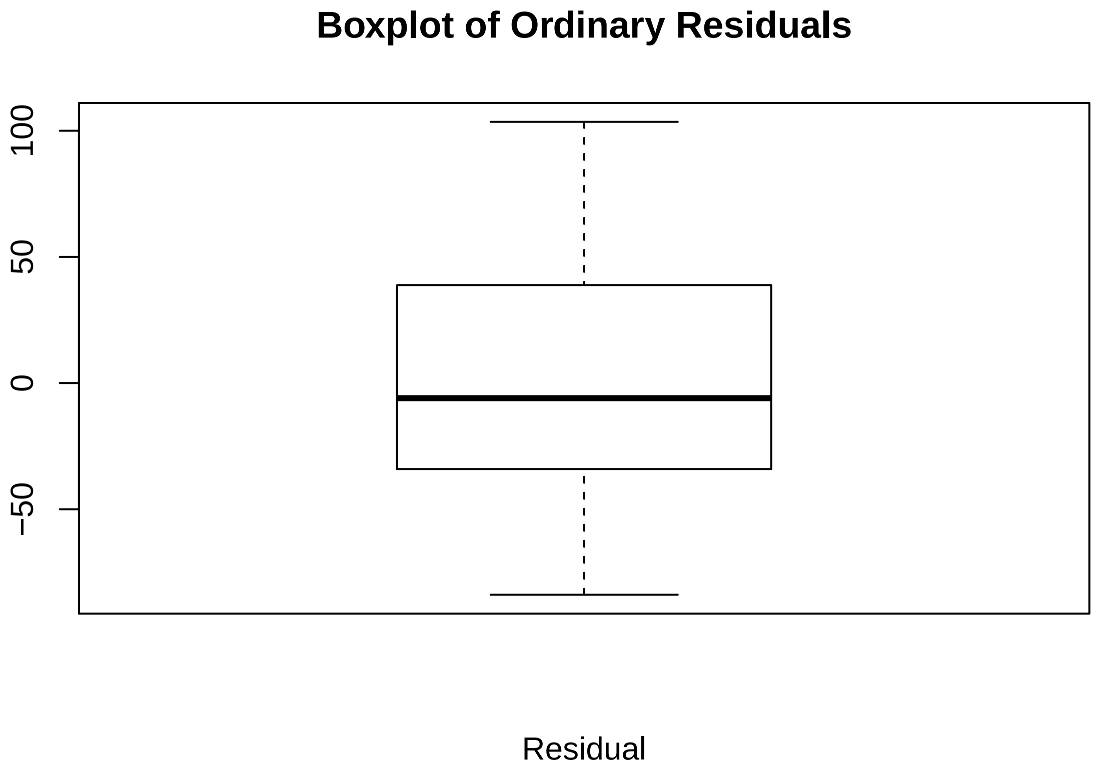
Presence of outliers

The model is sometimes sensitive to influential points such as outliers. Identifying outliers involves examining a boxplot of the residuals.

Example 1E

Assess whether there are outliers in the lab model estimated previously.

```
boxplot(LAB.RESID,
        main = "Boxplot of Ordinary Residuals",
        xlab = "Residual")
```



Assessment. The boxplot did not show evidence of outliers.

Conclusion. There are no outliers present.

Techniques to use when assumptions are not met

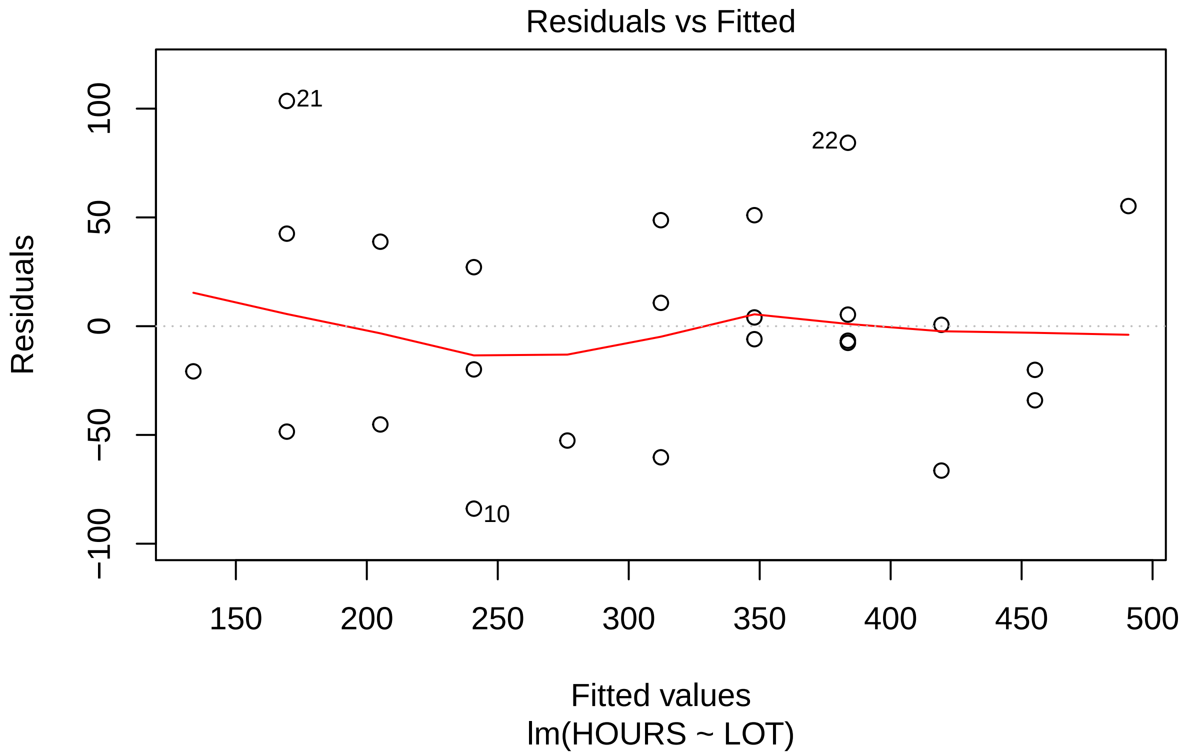
When you detect that the assumptions of the simple linear regresion model have been violated, there are a few techniques that you can try. Violations of linearity, normality and homoskedasticity may be corrected by transforming the data. You can try taking the natural logarithm of the outcome, predictor or both variables. You can try taking the square root or the reciprocal, too. More complicated transformations are possible.

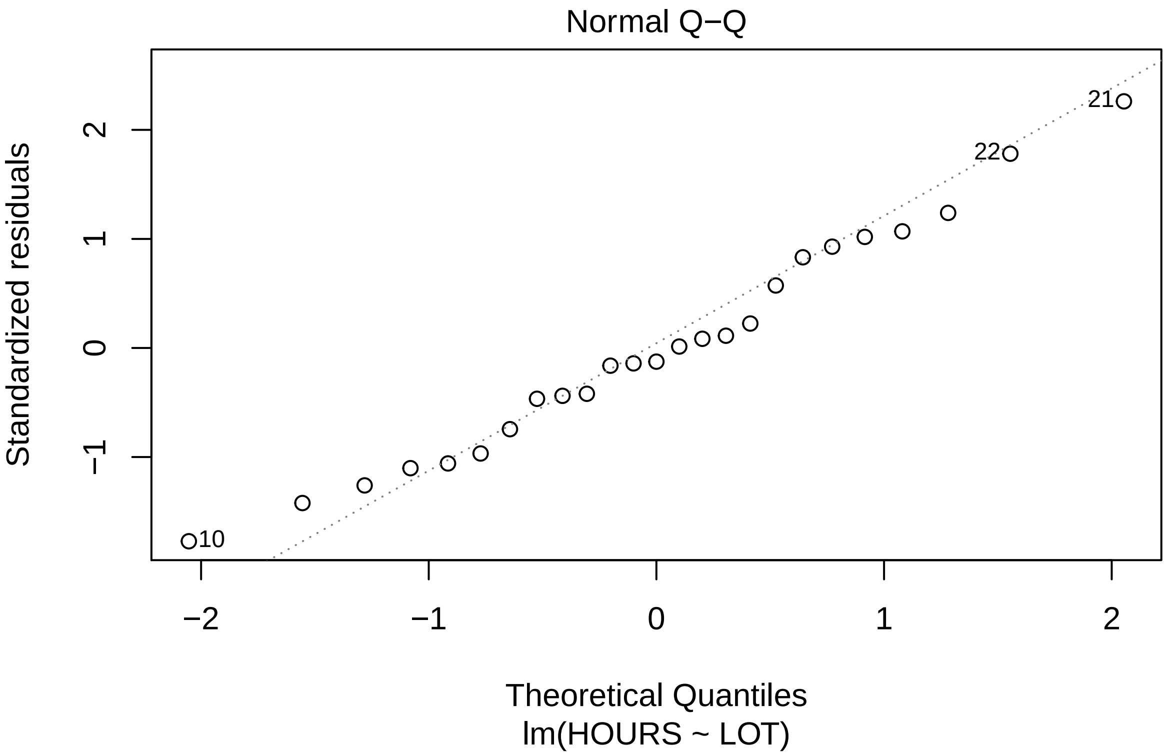
You can also consider the use of nonlinear models, but that is beyond the scope of this module.

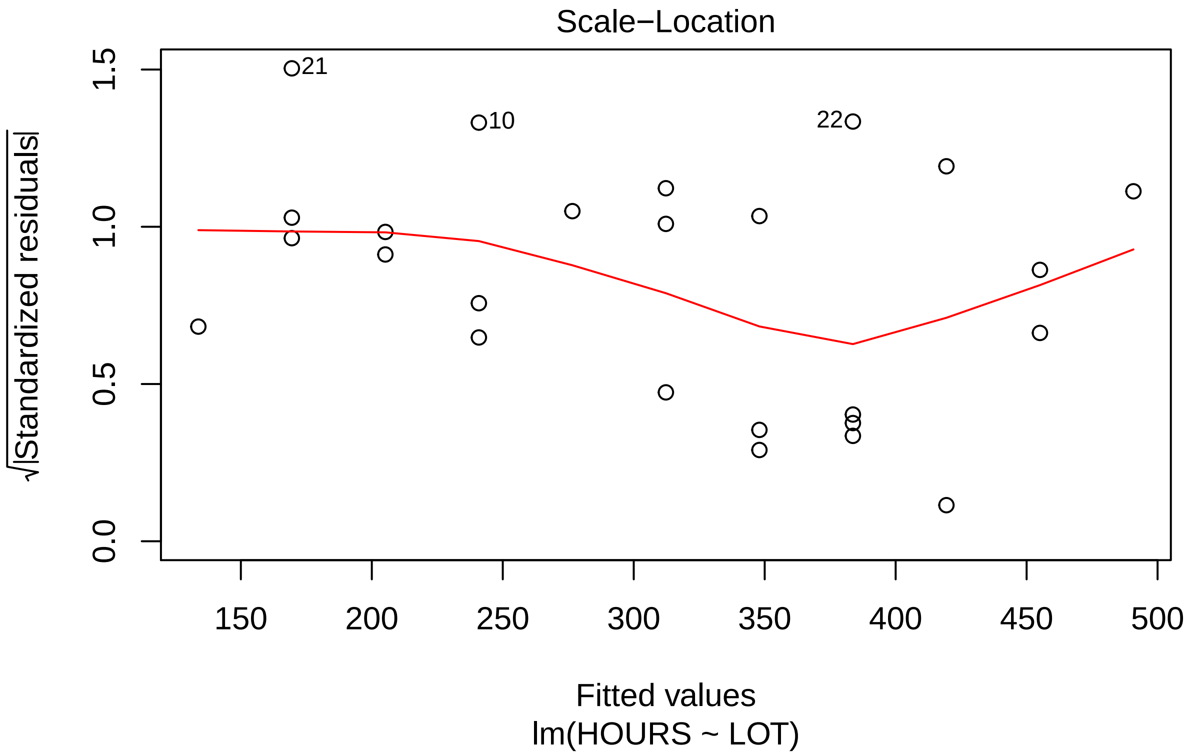
Diagnostic plots in R

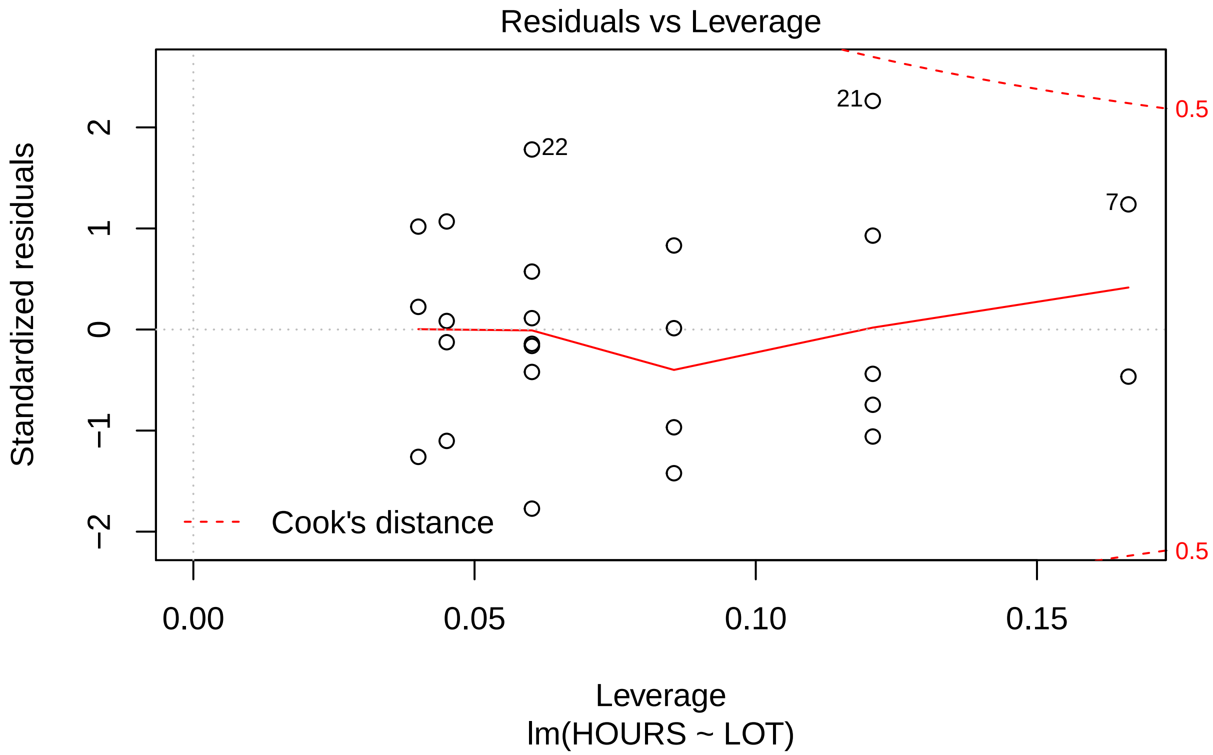
In the discussion thus far, we calculated and produces diagnostic plots as we needed them. In R, there is a simpler way to do this, but you need to be very careful about what you are doing.

```
plot(LAB.LM)
```


Residuals vs Fitted
Residuals
100
50
0
−50
−100
150
200
250
300
350
400
450
500
Fitted values
lm(HOURS ~ LOT)
21
22
10


Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(HOURS ~ LOT)
−2
−1
0
1
2
−1
0
1
2
10
21
22


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
150
200
250
300
350
400
450
500
21
10
22
Fitted values
lm(HOURS ~ LOT)



This command will produce four graphs at once. The first two were presented in the previous sections. The last two are more relevant to more complicated regression techniques.

IMPORTANT For the purposes of this class, the above command will NOT be accepted as an appropriate demonstration of your production of appropriate diagnostsic plots.

Concluding remarks

This ends the discussion of simple linear regression. In the next session, we will extend our understsanding to models that have more than one independent variable. That is, we will try to fit the model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$ where p is the number of predictor variables. These models are called *multiple* linear regression models.

The time we spent developing our understanding of simple linear models will help us in multiple linear models. We will divide the topic into (1) specification of the model and (2) model diagnostics.

THE END