Lecture 8: Relaxing the Assumptions of the Classical Model (I)

Thomas Chadefaux

Contents

Remember that the standard regression model is:

$$y = X\beta + \varepsilon, \quad E[\varepsilon] = 0, \quad E[\varepsilon\varepsilon'] = \sigma^2 I$$

A central assumption underlying OLS and its being BLUE is that the error terms of the model have constant variance and are mutually uncorrelated. If this assumption is violated, OLS remains unbiased but it might no longer be 'best', i.e., it may no longer have minimum variance among all the other unbiased linear estimators.

1 Heteroskedasticity

First assume that the disturbances are heteroskedastic. That is,

$$E[\varepsilon\varepsilon'] = \Sigma = \begin{pmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_n^2 \end{pmatrix}$$

Note that all off-diagonal elements remain 0 because we still assume uncorrelated disturbances. BUT the diagonal elements may be different for each observation. In other words, the amount of randomness in the outcome of y_i, as measured by var(y_i)= σ^2, may differ for each observation.

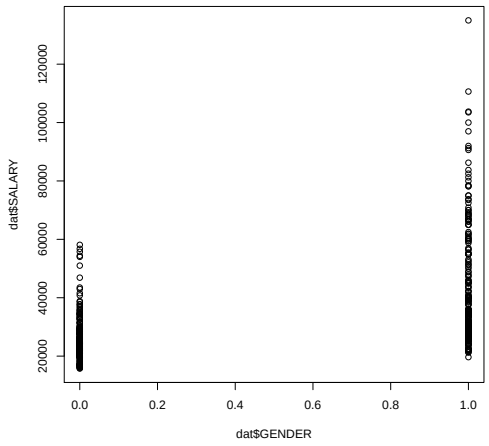
1.1 Sources of Heteroskedasticity

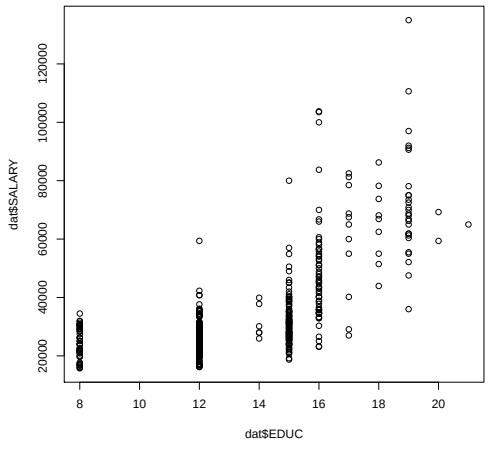
- Population heteroskedasticity. For example, the actual population error increases with level of the independent variable. For example if we regressed the proportion of income spent on food on income, then we would find that at low levels of income, people spend most of their money on food (hence low variance), whereas at high levels of income, some people spend a lot on food, whereas others spend little (hence high variance). Now, you could make the case that this is an example of omitted variable bias. If we added a third variable "taste for food", then perhaps we could explain a lot of the variance, and the high variance we observe at high levels of income would be accounted for by this variable. BUT I would respond that this variable (taste for food) is actually irrelevant. It is irrelevant, because it is not correlated with income (your taste for food is arguably independent of your income. I suppose you could make the case that it is dependent on it but anyway), and hence does not bias our inference.
- "Model" heteroskedasticity. Suppose you fit $y = \alpha + \beta_1 x + \varepsilon$ when the true model is $y = \alpha + \beta_1 x + \beta_2 x^2 + \varepsilon$. Then you have

Two Examples:

```
1  setwd('~/Documents/Academia/Teaching/TCD/2015-HT/
       P07005_Quantitative_Methods_II/Lectures/lecture7/
       ')
2  library(gdata)
3  dat <- read.xls('Data/xm501bwa.xls')
4  par(mfrow=c(1,1))
5  pdf('Figs/BoxplotSalaryGender.pdf')
6  boxplot(dat$SALARY ~ dat$GENDER, xlab='Gender', ylab=
       'Salary')
7  dev.off()
8
9  sd(dat$SALARY[dat$GENDER==1])
10 sd(dat$SALARY[dat$GENDER==0])
11
12 pdf('Figs/scatterSalaryGender.pdf')
13 plot(dat$GENDER,dat$SALARY)
14 dev.off()
15
16 pdf('Figs/scatterSalaryEduc.pdf')
17 plot(dat$EDUC,dat$SALARY)
18 dev.off()
```

Listing 1: heteroskExample1.R





high variance for high and low levels of your IV. AND it matters because clearly x^2 is related to x.

The problem we are facing here is that observations with large variances provide less information than those with low variance. When the variance of y is high, it is difficult to infer a precise b—i.e., its standard error will be high.

1.2 Properties of OLS under heteroskedasticity

First, remember that

$$\begin{aligned} b &= (X'X)^{-1}X'y \\ &= (X'X)^{-1}X'(X\beta + \varepsilon) \\ &= \beta + (X'X)^{-1}X'\varepsilon \\ \Rightarrow E[b] &= \beta \end{aligned} \tag{1}$$

i.e., b remains unbiased since our proof has never relied on homoskedasticity. However, let us now calculate $var[b]$, given that $E[\varepsilon\varepsilon'] = \Sigma \neq \sigma^2 I$:

$$\begin{aligned} var[b] &= E[(b - E[b])(b - E[b])'] \\ &= E[(b-\beta)(b-\beta)'] \\ &= E[(X'X)^{-1}X'\varepsilon)(X'X)^{-1}X'\varepsilon)'] \\ &= E[(X'X)^{-1}X'\varepsilon\varepsilon'X(X'X)^{-1}] \\ &= (X'X)^{-1}X'\Sigma X(X'X)^{-1} \end{aligned}$$

Note that this is no longer $\sigma^2(X'X)^{-1}$. Therefore the standard errors of our coefficients are wrong if we use the standard OLS method, and hence our inferences will no longer be valid. I.e., the p-value that R or Stata report when you run lm or reg will be wrong. Moreover, the OLS estimate is no longer best (i.e., it no longer has minimum variance), and hence there are other linear unbiased estimators that have smaller variance.

A review of APSR articles shows that 2/3 of them correct for this problem. So this is not an obscure detail. This is something you should know.

1.3 What to do if you have heteroskedasticity?

- **Option 1: Transform your data** Make sure you have the right model!

- **Option 2: White (Sandwich) Standard Errors**

The idea here is that instead of relying on $Var[b] = \sigma^2(X'X)^{-1}$, we should use $(X'X)^{-1}X'\Sigma X(X'X)^{-1}$. One problem is that we usually do not know Σ, but as before we can estimate it using our sample data: $Var[b] = (X'X)^{-1}(\sum_{i=1}^{N} e_i^2 x_i x_i')(X'X)^{-1}$.[1]

Intuition for the white standard errors: before, we estimated $var[b]$ using $(X'X)^{-1}X'\sigma^2 IX(X'X)^{-1} = \frac{e'e}{n-k}(X'X)^{-1}$. This implies that each observation affected the variance of b equally. With the robust SE, however, we are estimating $var[b]$ using $(X'X)^{-1}X'\Sigma X(X'X)^{-1}$, which implies that each observation gets a different weight. Observations with a large error increase the variance a lot, whereas those with low e don't increase it much. If our data is homoskedastic, this makes no difference since all our observations have the same expected e_i, and hence it does not hurt to just multiply them all by σ^2. With heteroskedasticity, however, we should assign different weight to different observations, which is what the White SE do.

- **Option 3: Weighted Least Squares**. Not covered here. The basic idea is to assign lower weight to those observations that have high variance, and a higher weight to those with low variance. So instead of minimising the sum of squared residuals, as we did in regular OLS:

$$OLS : min \sum_i (y_i - x_i'\beta)^2,$$

we could minimise a function of the form

$$\sum_i w_i^2 (y_i - x_i'\beta)^2,$$

with smaller weights for larger values of σ_i^2. The question is: how to choose these weights, and we do not cover this here.

1.4 Tests for homoskedasticity

[1] Note that the only reason we are using $Var[b] = (X'X)^{-1}(\sum_{i=1}^{N} e_i^2 x_i x_i')(X'X)^{-1}$ instead of $Var[b] = (X'X)^{-1}X'ee'X(X'X)^{-1}$ is that ee' might have non-zero elements on the off-diagonal

```
n <- 3000
x <- rnorm(n, mean=3, sd=1)
X <- cbind(1, x)
y <- x  + x*rnorm(n, mean=1)
plot(x,y)

#--- First estimate the model using regular OLS
lm1 <- lm(y ~ x)
summary(lm1)
e <- resid(lm1)

# now calculate robust SE manually
bread <- solve(t(X)%*%X)
eprimee <- e%*%t(e)
# we need to remove the non-diagonal elements, as we
    assumed they are 0
eprimeediag <- diag(nrow(eprimee))
diag(eprimeediag) <- diag(eprimee)
meat <- t(X)%*%eprimeediag%*%X
# degrees of freedom correction:
dfc <- n/(n-ncol(X))
se.b <- sqrt(dfc*bread %*% meat %*% bread)
se.b

#--- Now calculate robust SE and associated t and p-
    values using canned function
library("sandwich")
library("lmtest")
se.b.canned <- sqrt(vcovHC(lm1)); se.b.canned
```

Listing 2: robustSE.R

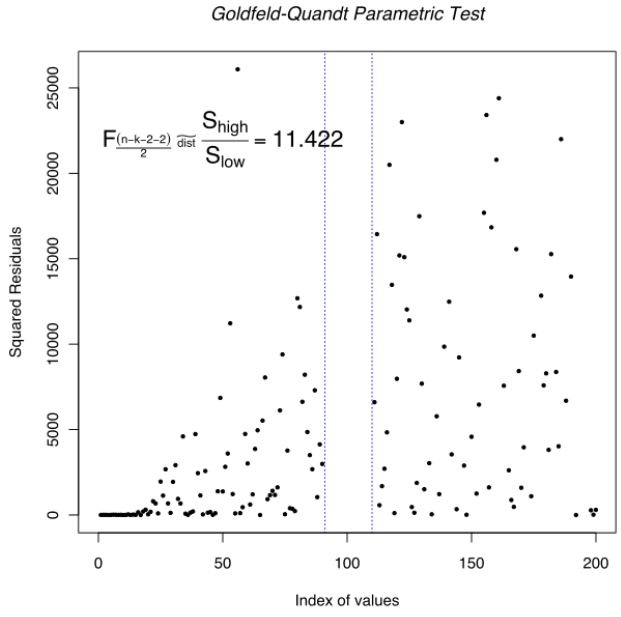
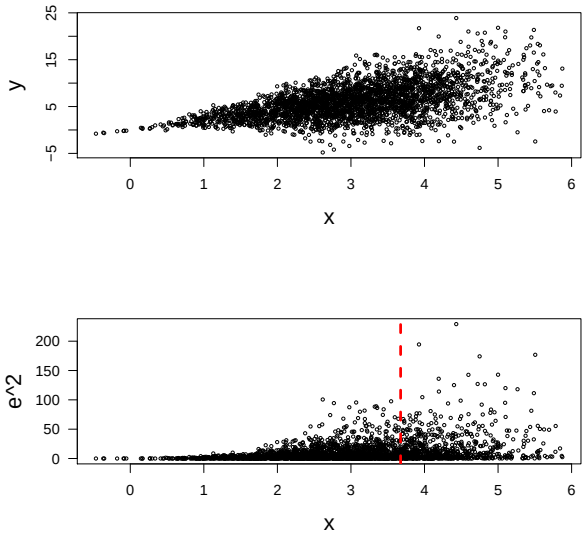


Figure 1: source: wikipedia

Figure 2: source: GoldfeldQuandtTest.R



1. The Goldfeld-Quandt test. We have some population process, say $Y = \alpha + \beta_1 X_1 + \ldots + \beta_k X_k + \varepsilon$, and we want to test the hypothesis that $Var(u) = \sigma^2$ vs the hypothesis that $Var(u) = \sigma^2 f(x_i, \ldots)$. In theory $f(\cdot)$ could be a function of any of the IVs, but with the Goldfeld-Quant test we will test it for a specific variable, say x_1. The test is very simple. We cannot observe $sigma^2$, but we can observe e^2. In this test, we plot e^2 against x_1. Homoskedasticity would then dictate that there is no relationship between the two. One way to test this is to divide the plot into two samples, and then to see whether there is a difference in the sum of squared residuals on both sides (after taking into account the number of observations on each side). If there is a difference, then that my be indicative of the fact that the sample error is increasing (or decreasing), i.e., we have heteroskedasticity. This test takes the form of an F-test:

$$F = \frac{\sum_i^{N_2} e_i^2/(N_2 - k)}{\sum_i^{N_1} e_i^2/(N_1 - k)} \sim F_{N_2, N_1}$$

The null hypothesis here is that we have homoskedastic errors, i.e., the numerator and denominator should be very similar, so F will be close to 1. Significant deviation from 1 indicate heteroskedasticity.

- Why is this a good test? It is simple and visual. Visual is good, because relying on statistics alone might lead us to reject the null hypothesis without really understanding what

```
#------ Goldfeld-Quandt test ------#

#--- Generate Data
n <- 3000
x <- rnorm(n, mean=3, sd=1)
y <- x  + x*rnorm(n, mean=1)

#--- Estimate the model
formula1 <- y~x
lm1 <- lm(formula1)

#--- order the residuals according to x and plot
df1 <- data.frame(x = x, e = resid(lm1))
df1 <- df1[order(x),]
e <- df1$e

#--- split the residuals in 2
e1 <- e[1:(length(e)/2)]
e2 <- e[(length(e)/2+1):length(e)]

#--- Calculate the F statistic
Ftest1 <- sum( e2^2/length(e2) ) / sum( e1^2/length(
    e1) )

#--- verify with the canned version:
library(lmtest)
gqtest(formula1, point = 0.5, fraction = 0,
    alternative = c("greater"),
       order.by = x, data = list())

#--- Plot
library(fields)
pdf('Figs/gqtestR.pdf')
par(mfrow=c(2,1))
plot(x,y, cex=0.5, cex.lab=1.5)
plot(df1$x, df1$e^2, cex=0.5, xlab='x', ylab='e^2',
    las=1, cex.lab=1.5)
xline(x[length(e)/2], col=2, lty=2, lwd=3)
dev.off()
```

Listing 3: GoldfeldQuandtTest.R

is happening. Suppose for example that we have an outlier. This might lead to an increase in F, and hence a rejection of the null. But an outlier is probably not what we want to evaluate when we test for heteroskedasticity.

- Why is this a limited test?
  - First, it only tests for one predetermined IV. But we might want to test in general
  - It assumes that the error variance is a monotonic function of the specific explavatory variable.

2. The Breusch Pagan test. Here we want to test for more general forms of heteroskedasticity than in the Goldfeld-Quandt test. The basic idea is to regress the squared residuals e^2 on our variables. I.e.,

$$e^2 = \delta + \delta_1 x_1 + \delta_2 x_2 + \ldots + \delta_k x_k + \varepsilon \tag{2}$$

Then we can either test the statistical significance of the δs. If all are zero, then we have homoskedasticity, otherwise we have heteroskedasticity. But typically we don't really care where exactly heteroskedasticity is coming from; we just want to know whether we have it or not. So we can just run an F-test that all coefficients are equal to 0, i.e., $\delta_1 = \delta_2 = \ldots = \delta_k = 0$. A related way of testing the same thing is to run what is called a Lagrange Multiplier test, where $LM = nR^2$, where R^2 is the R^2 of the auxiliary regression (2). Under the null hypothesis, this test follows a chi-squared distribution with k degrees of freedom. The underlying idea here is that if R^2 is pretty high, then it means that I can explain variation in my squared residuals pretty well using my IVs, which is a sign of heteroskedasticity.

3. An even more general test for heteroskedasticity is the *White Test*. The idea is very similar to the Breusch-Pagan test. We want to test whether there is any relationship between e^2 and my independent variables. But this time I want to be more general and allow squared and cubic variables, and perhaps also interaction terms. So what I want to estimate is:

$$\begin{aligned} e^2 = {} & \delta + \delta_1 x_1 + \delta_2 x_2 + \ldots + \delta_k x_k + \varepsilon \\ & + \gamma_1 x_1^2 + \gamma_2 x_2^2 + \ldots + \gamma_k x_k^2 + \varepsilon \\ & + \nu_1 x_1 x^2 + \nu_2 x_1 x_3 + \ldots + \nu_k x_k x_i + u \end{aligned}$$

But obviously, this is a mess and will use a lot of degrees of freedom, and hence have low power. White therefore came up with a clever way around this. The idea is that $\hat{y}$ is defined as

$$\hat{y} = \hat{\alpha} + \hat{\beta}_1 x_1 + \ldots + \hat{\beta}_k x_k \tag{3}$$

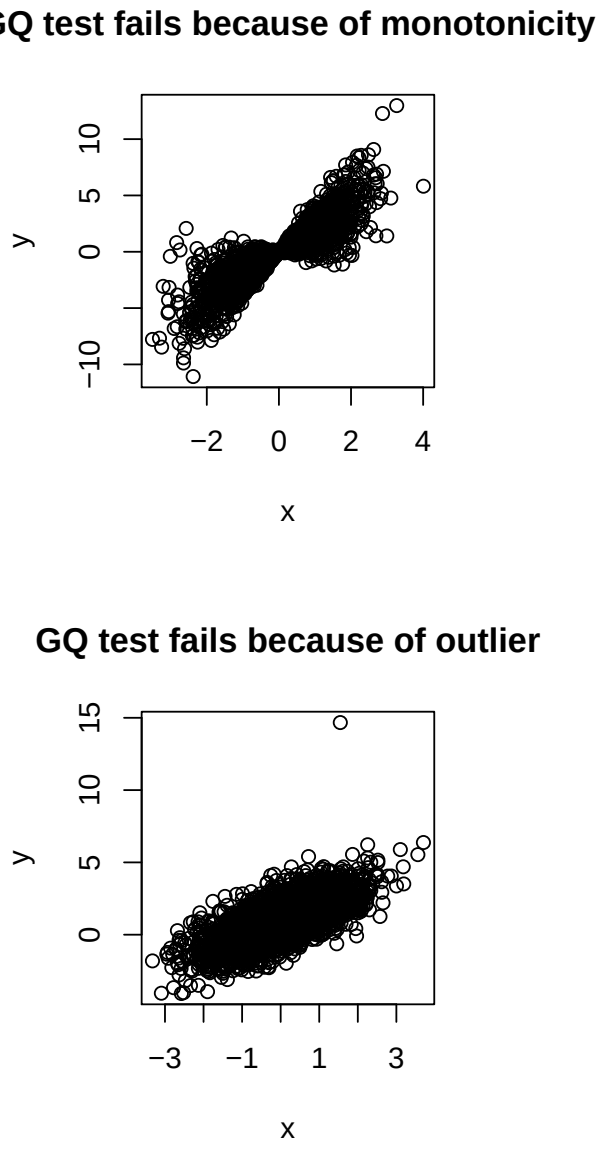


Figure 3: source: GoldfeldQuandtTest.R

So instead of regressing e^2 on all these variables, I can just regress it on $\hat{y}$ and $\hat{y}^2$, which will include not only all my main terms, but also cross terms and squares.

$$e^2 = \lambda_0 + \lambda_1 \hat{y} + \lambda_2 \hat{y}^2 + u$$

So I am checking whether there is any relationship between my IVs or combinations or squares thereof, without actually having to include all these terms. So we run the regression, and then again run an F-test or LM test that all coefficients are equal to 0.

2 Autocorrelation, aka serial correlation

Our error terms might be correlated, perhaps in space or time. I.e., it means that $Cov(\epsilon_i, \epsilon_j) \neq 0$.This violates the assumption that $E[\varepsilon\varepsilon'] = \sigma^2 I$, since it implies that the off-diagonal elements of the variance-covariance matrix of the error terms are not zero:

$$E[\varepsilon\varepsilon'] = \Sigma = \begin{pmatrix} E(\varepsilon_1^2) & E(\varepsilon_1\varepsilon_2) & \cdots & E(\varepsilon_1\varepsilon_N) \\ E(\varepsilon_2\varepsilon_1) & E(\varepsilon_1^2) & \cdots & E(\varepsilon_2\varepsilon_N) \\ \vdots & \vdots & \ddots & \vdots \\ E(\varepsilon_N\varepsilon_1) & E(\varepsilon_N\varepsilon_2) & \cdots & E(\varepsilon_1^2) \end{pmatrix} \tag{4}$$

If you have serial correlation, then OLS is no longer BLUE. Why? Because there are other estimators that are better, while still unbiased and linear (remember that "better" means lower variance). But the main problem is that serial correlation tends to be indicative of more serious problems:

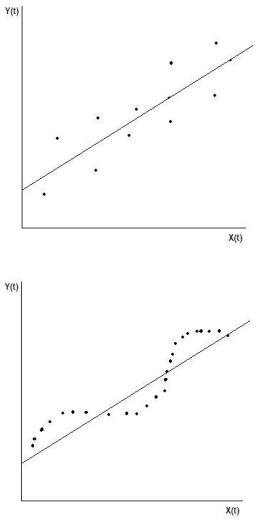


Figure 4: positive and negative autocorrelation

2.1 How serial correlation arises

- Inertia/time to adjust. Especially in time series data
- Omit an important variable. That means that the effect of that variable is instead relegated to the error term. Suppose that that variable is persistent through time. Then there will also be persistence in the error term over time. Example: I am interested in the evolution of GDP over time, but forgot to include seasonal components.
- Functional misspecification. Suppose the true relationship is $y = \alpha + \beta_1 x_1 + \beta_2 x_1^2 + \varepsilon$ but we fit $y = \alpha + \beta_1 x_1 + \varepsilon$. Then positive errors will follow positive errors, and negative ones will follow negative ones.

- Measurement error in independent variable. If the measurement error persists through time, then it will be present in ε over time.
- clustering. More on this in the lecture on panel data.

2.2 Consequences of autocorrelation

The problem is very similar to the one of heteroskedasticity:

- The coefficients are still unbiased. See (1)
- However, as with heteroskedasticity, we will underestimate the variance of b. Intuitively, that is because we are assuming that we have more independent information than we really have. As a result, hypothesis testing is invalid.

2.3 Tests for serial correlation

- The Durbin-Watson test. The test is defined as

$$d = \frac{\sum_{i=2}^{n}(e_t - e_{t-1})^2}{\sum_{i=2}^{n} e_i^2}$$

  The basic idea is that if successive error terms are positively (negatively) correlated, then the differences $\varepsilon_t - \varepsilon_{t-1}$ tend to be relatively small (large). The statistic is such that $0 \geq d \geq 4$. Values of d close to 0 indicate positive serial correlation, whereas values close to 4 indicate negative serial correlation.

- The Breusch-Godfrey test (the preferred choice): the idea here is simple: regress e_t on e_{t-1}:

$$\varepsilon = \delta_0 + \delta_1 \varepsilon_{t-1} + \delta_2 \varepsilon_{t-2} + u$$

  Then we construct a standard t-statistic for δ_1 and/or δ_2.

2.4 What to do?

- Add lagged variables. [from Heij, p. 368:] If the residuals of an estimated equation are serially correlated, this indicates that the model is not correctly specified. For (ordered) cross section data this may be caused by non-linearities in the functional form. See lecture 6 for possible adjustments of the model. For time series data, serial correlation means that some of the dynamic properties of the data are not captured by the model. In this case one can adjust the model?for instance, by including lagged values of the explanatory variables and of the explained variable as additional regressors. As an example, suppose that the model

$$y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$$

is estimated by OLS and that the residuals are serially correlated. This suggests that $\varepsilon_t = y_t - b_1 - b_2 x_t$ is correlated with $\varepsilon_{t-1} = y_{t-1} - b_1 - b_2 x_{t-1}$ This may be caused by correlation of y_t with y_{t-1} and/or x_{t-1}, which can be expressed by the model

$$y_i = \lambda_1 + \lambda_2 x_{t-1} + \lambda_2 y_{t-1} + u_t$$

When the disturbances XX of this model are identically and independently distributed (IID), then the model is said to have a correct dynamic specification.

- Newey-West estimator of the covariance matrix. The idea is similar to White's standard error correction for heteroskedasticity, except that the 'meat' is estimated differently, using a function of the maximum number of lags you believe could be correlated with the current residuals. More formally, we can no longer assume that

$$Var[b] = \sigma^2 (X'X)^{-1},$$

but rather must use

$$Var[b] = (X'X)^{-1} X'\Sigma X (X'X)^{-1}, \tag{5}$$

just like we did for the heteroskedastic case. Remember that in the heteroskedastic case, we had

$$E[\varepsilon\varepsilon'] = \Sigma = \begin{pmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_n^2 \end{pmatrix}$$

and to estimate Σ, we used

$$\hat{\Sigma} = \begin{pmatrix} e_1^2 & 0 & \cdots & 0 \\ 0 & e_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & e_n^2 \end{pmatrix}$$

But in the case of auto-correlated residuals, the off-diagonal elements of Σ are not 0 (see (4)).[2] So now we need an estimate of a lot of values (N^2 values to be precise). Now we could apply the same strategy as in the White SE used for heteroskedasticity, namely to use ee':

$$\hat{\Sigma} = \begin{pmatrix} e_1^2 & e_1 e_2 & \cdots & e_1 e_n \\ e_2 e_1 & e_2^2 & \cdots & e_2 e_n \\ \vdots & \vdots & \ddots & \vdots \\ e_n e_1 & e_n e_2 & \cdots & e_n^2 \end{pmatrix}$$

[2] Note that the off-diagonal elements represent the covariance between error terms. For example, $\Sigma_{1,2} = Cov(\varepsilon_1, \varepsilon_2)$.

But here is the problem: this is exactly the same as ee'. Why is this a problem? Plug this back into (5). Then we get

$$Var[b] = (X'X)^{-1}X'ee'X(X'X)^{-1},$$

But remember that $X'e = 0$, so we would get $Var[b] = 0$. That is of course nonsense, and hence we cannot use this method.

Instead, Newey-West came up with a clever workaround, namely to use the following way to estimate the variance covariance matrix of the residuals:

$$\hat{\Sigma}_{NW} = \begin{pmatrix} e_1^2 & w_1 e_1 e_2 & w_2 e_1 e_3 & 0 & \cdots & 0 \\ w_1 e_2 e_1 & e_2^2 & w_1 e_2 e_3 & w_2 e_2 e_4 & \cdots & 0 \\ w_2 e_3 e_1 & w_1 e_3 e_2 & e_3^3 & w_1 e_2 e_4 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & w_2 e_n e_{n-2} & w_1 e_n e_{n-1} & e_n^2 \end{pmatrix},$$

$w_i = 1 - \frac{i}{N+1}$. For example, the weight w_1 in the example above would be $w_1 = 1 - \frac{1}{N+1}$. So for $N = 100$, say, we'd have $w_1 = 0.99$ Note two things: first, we have set some values to 0. And then we have assigned some weights to some of the off-diagonal elements (here 2 off-diagonals, but it could be more or less). By doing this, we have removed the problem of $\hat{\Sigma} = ee' = 0$. Second, we have assigned some weights to some elements. In particular, we include positive weights for the covariances between error terms that are close in time (or space).

What does this all mean. The i, j element of Σ tells you how correlated two more or less consecutive error terms are. If they are highly correlated, then in essence these observations add little information, and hence the standard error of the coefficient should be increased to reflect this fact. So if $e_i e_j$ is high, then the variance of b will increase.

Now that we have fixed the standard errors, we can again use our regular t-test to test hypotheses about coefficients:

$$t_\beta = \frac{\beta - \beta_0}{SE_{NW}}$$

```
1 vcovHAC(lm1)
```

Listing 4: command to estimate Newey-West standard error (heteroskedasticity + autocorrelation corrected) in R

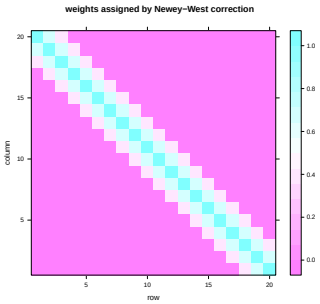


Figure 5: Newey-West scheme for Σ. Source: newey-west.R

3 Disturbance Distribution

3.1 Normality

- Main method: look at the histogram or qq-plot of your residuals (aka Shapiro-Wilk test).
- Jarque-Bera test: not covered here
- Kolmogorov-Smirnov: not covered here

3.2 Influence, leverage and outliers

- An outlier is a point that has an extreme x value, an extreme y value, or both. Or it is far away from the main trend of the points. is far from the fitted values. I.e., it has a large residual. Why do we care? Two main reasons:
  - An outlier might signal a misspecification of the model. Perhaps you are missing an important variable, or your functional form is wrong.
  - An outlier might indicate a data error. Perhaps a comma was misplaced, etc.
  - Though not a problem per se, outliers can change the fit and alter the coefficients significantly.
- How to detect them? If you have no outlier, then your residuals should be spread evenly around your predicted value. I.e., there should be no discernible pattern between $\hat{y}_i$ and e_i.
  - One easy way to check is to just plot your fitted values against your residuals. Note that the outlier might not be visible by simply looking at your x1 against y plot, hence the need to look at residuals. In Fig. 10, for example, the outlier cannot be detected, even though it is obvious once we look at residuals (Fig. 11). Therefore we typically we plot fitted values (i.e., $\hat{y}$) against residuals (or standardised residuals).
  - But looking at residuals might not be enough. The problem with regular residuals is that influential observations will just pull the regression toward themselves, and hence reduce the variance of that residual. To correct for this, we calculate the studentized residuals as the residual divided by its variance:

$$\text{Studentized } e = \frac{e_i}{s\sqrt{1-h_i}},$$

where h_i is a measure of leverage (see below).

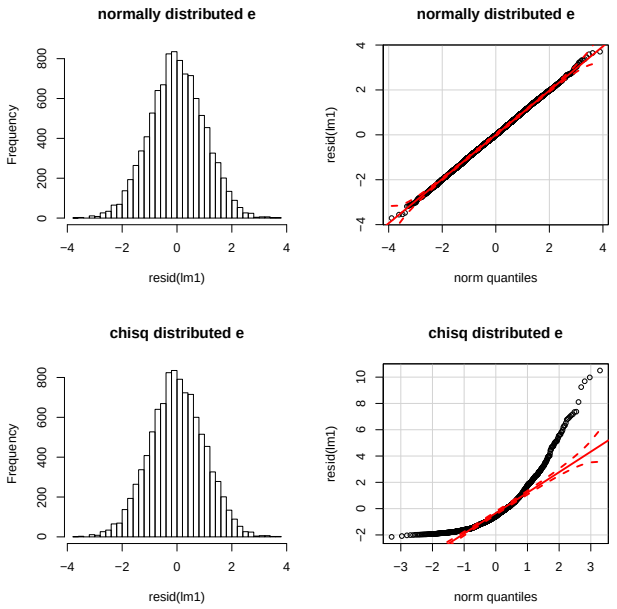


Figure 6: Normality of outliers. Source: normality.R

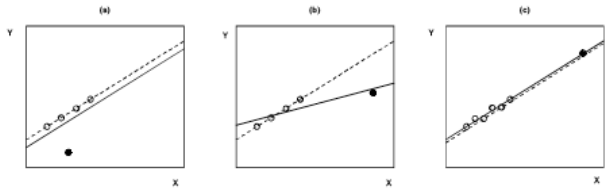


Figure 1. Unusual data in regression: (a) a low-leverage and hence uninfluential outlier; (b) a high-leverage and hence influential outlier; (c) a high-leverage in-line observation. In each case, the solid line is the least-squares line for all of the data; the broken line is the least-squares line with the unusual observation omitted.

Figure 7: Types of outliers. Source: Fox, Applied regression analysis, p. 242

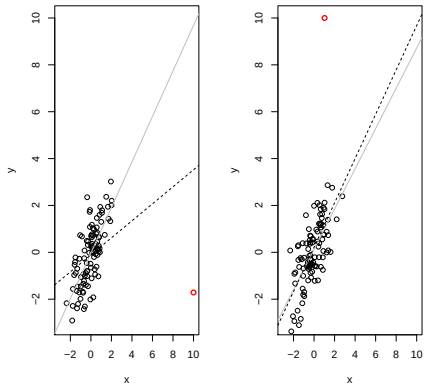


Figure 8: Types of outliers. Source: typesOfOutliers.R

- But the standardised residuals don't tell the whole story: we can have a large residual, but not much influence on our regression. So we also want a measure of *leverage* In particular, we are interested in the effect of a change in y_i on our estimate $\hat{y}_i$. Observations with little influence can be changed without affecting the slope and intercept, and hence $\hat{y}$. So what we want is a measure of how $\hat{y}$ will change as a function of a change in y_i. Remember that

$$\hat{y} = Xb = X(X'X)^{-1}X'y = Hy,$$

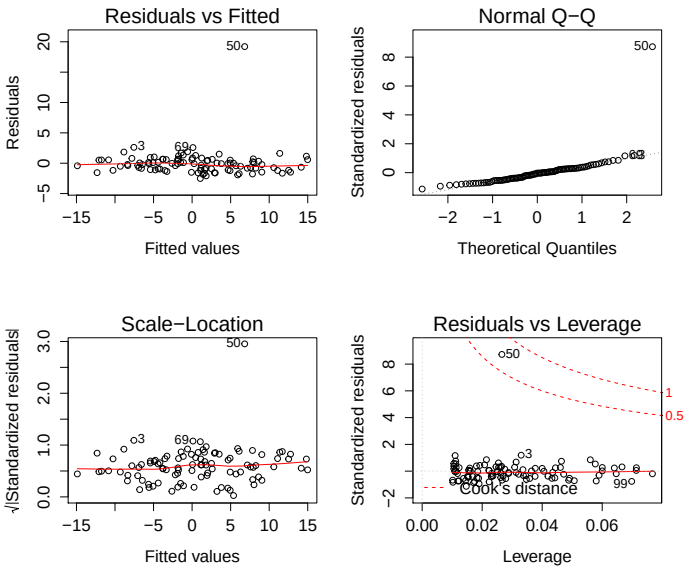
where $H = X(X'X)^{-1}X'$. Note that a change in y changes $\hat{y}$ by H, so H is indeed the measure of leverage we are interested in.

- Finally, we want to put these two things together: (studentized) residuals and leverage. One way to think about it is that

$$\text{Influence on coefficients} = \text{Leverage} \times \text{Discrepancy}$$

So we plot the hat values vs the studentized residuals (fig. 12). A more formal way of measuring this is to calculate the impact on each coefficient of deleting each observation in turn. Cook has proposed measuring the distance between the coefficients by calculating the F-statistic for the hypothesis that

Finally, note the default diagnostic plots returned by R:



Clockwise:

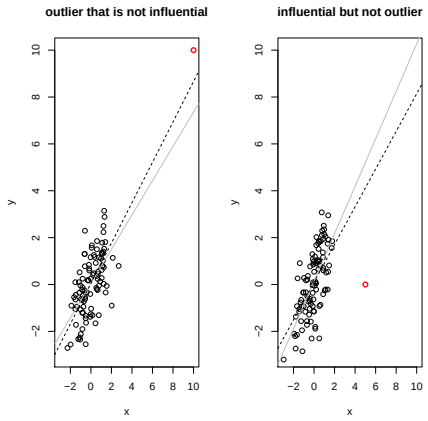


Figure 9: Outliers vs influential obs. Source: outliersAndInfluentials.R

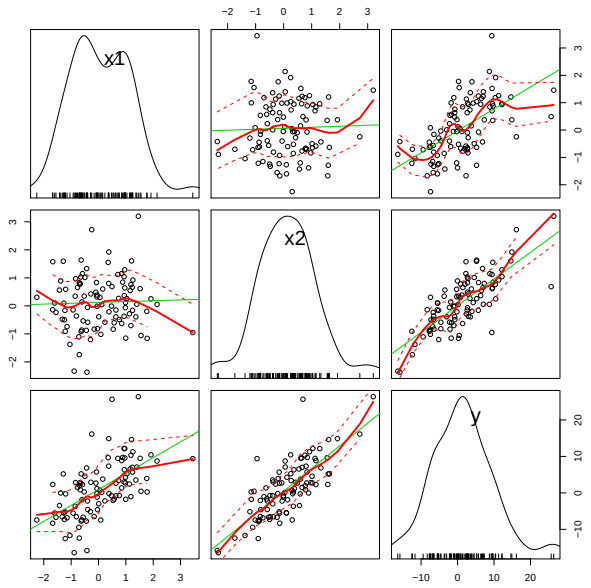


Figure 10: Scatterplot of x1,x2 and y. Source: FittedVsResid.R

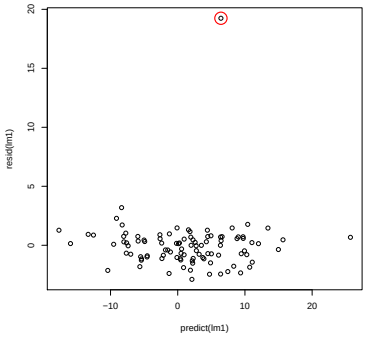


Figure 11: Plot of fitted values and residuals for y =x1+x2. Source: FittedVsResid.R

1. Residuals versus fitted (predicted) values. Under the usual assumptions for the linear regression model, we don't expect the variability of the residuals to change over the range of the dependent variable, so there shouldn't be any discernable pattern to this plot.

2. A normal quantile-quantile plot of the standardized residuals. We'd expect a normal quantile-quantile plot of the residuals to follow a straight line.

3. A scale-location plot. This plot is similar to the residuals versus fitted values plot, but it uses the square root of the standardized residuals. Like the first plot, there should be no discernable pattern to the plot.

4. A Cook's distance plot

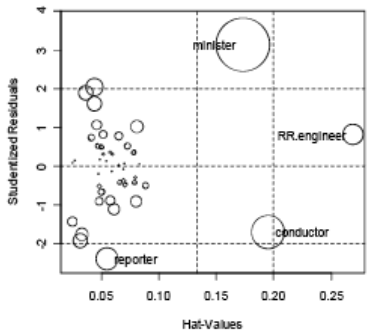


Figure 12: hat values vs residuals. Source: Fox