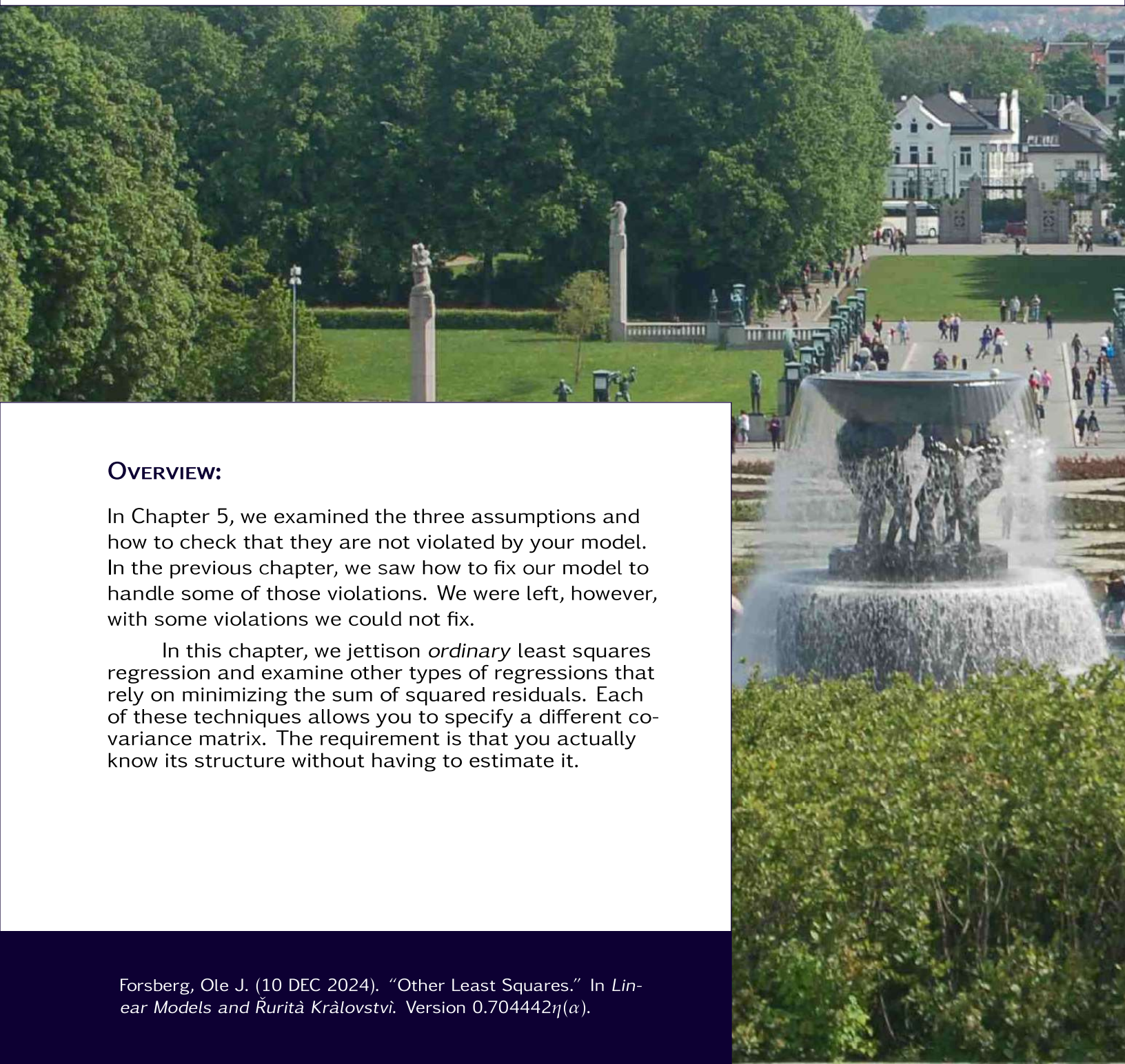
Chapter 8:
Other Least Squares



Overview:

In Chapter 5, we examined the three assumptions and how to check that they are not violated by your model. In the previous chapter, we saw how to fix our model to handle some of those violations. We were left, however, with some violations we could not fix.

In this chapter, we jettison *ordinary* least squares regression and examine other types of regressions that rely on minimizing the sum of squared residuals. Each of these techniques allows you to specify a different covariance matrix. The requirement is that you actually know its structure without having to estimate it.

Forsberg, Ole J. (10 DEC 2024). "Other Least Squares." In *Linear Models and Řuritâ Kràlovstvì*. Version 0.704442$\eta(\alpha)$.

Chapter Contents

In the past several chapters, we have examined the classical linear model (CLM) and how to estimate the parameters using ordinary least squares (OLS). That introduction came in Chapters 2, 3 and 4. In Chapter 5, we discovered how to check the requirements (assumptions) of the ordinary least squares method. Chapter 7 gave us some options for dealing with violations of the requirements.

However, it may be that those fixes do not fully succeed — or *can*not fully succeed. This chapter provides two estimation methods that offer advantages over ordinary least squares, as long as you have sufficient knowledge (science) of the structure of the problem — also known as the data-generation process.

This chapter reintroduces ordinary least squares. It then focuses on the covariance matrix of the residuals. As we reduce requirements on that matrix, we move from ordinary least squares to weighted least squares to generalized least squares.

8.1: Ordinary Least Squares

First, let us review ordinary least squares (OLS). When formulating OLS estimation of the classical linear model (CLM), we made the assumption that the residuals are independent and identically distributed Normal with constant zero expected value and variance.

In symbols, this is written as either

$$\varepsilon_i \overset{\text{iid}}{\sim} \mathcal{N}\left(0; \sigma^2\right) \tag{8.1}$$

or as

$$\mathbf{E} \sim \mathcal{N}_n\left(\mathbf{0}; \sigma^2\mathbf{I}\right) \tag{8.2}$$

The two statements are different ways of saying the exact same thing.

Note that the covariance matrix of $\mathbf{E}$ is $\sigma^2\mathbf{I}$:

$$\mathbb{V}[\mathbf{E}] = \sigma^2 \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & & 0 \\ \vdots & \vdots & & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix} = \begin{bmatrix} \sigma^2 & 0 & 0 & \cdots & 0 \\ 0 & \sigma^2 & 0 & \cdots & 0 \\ 0 & 0 & \sigma^2 & & 0 \\ \vdots & \vdots & & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \sigma^2 \end{bmatrix} \tag{8.3}$$

The values along the diagonal represent the variances of each residual *in the population*. That they are the same value, σ^2, indicates that the variance of the residuals is constant.

homoskedastic

The values off the diagonal represent the covariance between the residuals. For instance, the value at position 1,2 is the covariance between ε_1 and ε_2, which we symbolized as $\sigma_{1,2}$ in Appendix S. Since that value is 0, we are specifying that the two are linearly uncorrelated (a.k.a. independent).

independent

Thus, the covariance matrix above specifies that the variances of the residuals are constant and that the residuals are independent of each other. If this requirement is met, then we should use ordinary least squares regression. However, not always is this requirement met.

foreshadowing

8.2: Weighted Least Squares

It may be that the residuals are independent, but that their variance is known to not be constant. That is, we may have a model that leads to this assumption:

$$\varepsilon_i \overset{\text{ind}}{\sim} \mathcal{N}\left(0; \sigma_i^2\right) \tag{8.4}$$

or as

$$\mathbf{E} \sim \mathcal{N}_n\left(\mathbf{0}; \sigma^2\mathbf{D}\right) \tag{8.5}$$

Here, $\mathbf{D}$ is a diagonal matrix. Again, the two statements are different ways of saying the same thing.

Note that the covariance matrix of *this* $\mathbf{E}$ is

$$\mathbb{V}[\mathbf{E}] = \sigma^2 \begin{bmatrix} d_1 & 0 & 0 & \cdots & 0 \\ 0 & d_2 & 0 & \cdots & 0 \\ 0 & 0 & d_3 & & 0 \\ \vdots & \vdots & & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & d_n \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & 0 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & 0 & \cdots & 0 \\ 0 & 0 & \sigma_3^2 & & 0 \\ \vdots & \vdots & & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \sigma_n^2 \end{bmatrix} \tag{8.6}$$

The values along the diagonal represent the variances of each residual (in the population). That they are not necessarily the same value indicates that the variance of the residuals can differ from observation to observation.

heteroskedastic

The values off the diagonal represent the covariance between the values of the residuals. So, the value at position 1,2 is the covariance between ε_1 and ε_2, which we symbolized as $\sigma_{1,2}$ in Appendix S (note that these are population values). Since $\sigma_{1,2} = 0$, we are specifying that the two residuals are linearly uncorrelated in the population.

independent

> **Note:** Remember that Greek letters refer to the population, while Latin refer to the sample (usually).

Thus, the covariance matrix above specifies that the variances of the residuals are allowed to be different and that the residuals are independent of each other.

This assumption is the only difference between *weighted* least squares and *ordinary* least squares. But, it is a rather significant difference.

> **Note:** Remember that the value of σ^2 can indicate the variance of the population residuals *or* our uncertainty in the value of that residual.

To use weighted least squares (WLS), we need to know the structure of the **D** matrix. We do not need to know the exact values, but we need to know them up to a constant multiplier. That is, we need to know the *structure* of that heteroskedasticity. This usually comes from understanding the data-generating process.

Frequently, this is not known (thus, WLS should probably *not* be used). However, there are some cases when we *would* know this structure. For instance, if we are working with a response variable that is a proportion arising from a binomially-distributed variable, we know that the variance is

$$\sigma_i^2 = \frac{\pi(1-\pi)}{n_i} = \pi(1-\pi)\frac{1}{n_i} \tag{8.7}$$

Thus, the diagonal elements will be $d_i = 1/n_i$ and the multiplier (constant part) will be $\pi(1-\pi)$.

8.2.1 Fitting WLS: The Mathematics

Assuming we know the structure of the **D** matrix, we can determine all we need to about the WLS estimators and estimates. We just reduce this problem to a previous problem.

To clarify the similarities and differences between ordinary and weighted least squares, here is the classical linear model (CLM) for ordinary least squares:

$$\mathbf{Y} = \mathbf{XB} + \mathbf{E} \tag{8.8}$$

and for weighted least squares:

$$\mathbf{Y} = \mathbf{XB} + \mathbf{E} \tag{8.9}$$

CLM

Those are the same, whether one does OLS or WLS, because both come from the fact you are using the classical linear model. The difference comes in the assumption. Here is the assumption for OLS:

$$\mathbf{E} \sim \mathcal{N}_n\left(0; \sigma^2 \mathbf{I}\right) \tag{8.10}$$

Here is the assumption for WLS:

$$\mathbf{E} \sim \mathcal{N}_n\left(0; \sigma^2 \mathbf{D}\right) \tag{8.11}$$

Remember that $\mathbf{D}$ is a diagonal matrix.

The Transformation: There is a joke about how a mathematician solved the problem of a hose connected to a fire hydrant:

> A mathematician and a physicist were asked the following question:
>
> Suppose you walked by a burning house and saw a hydrant and a hose not connected to the hydrant. What would you do?
> **P**: I would attach the hose to the hydrant, turn on the water, and put out the fire.
> **M**: I would attach the hose to the hydrant, turn on the water, and put out the fire.
>
> Then they were asked this question:
>
> Suppose you walked by a house and saw a hose connected to a hydrant. What would you do?
> **P**: I would keep walking, as there is no problem to solve.
> **M**: I would disconnect the hose from the hydrant and set the house on fire, reducing the problem to a previously solved form.

And so, in the spirit of mathematicians, let us reduce the weighted least squares problem to that of ordinary least squares. If we can do this via a bijective transformation, then we have our confidence intervals and test statistics.

Talking Heads

If we define our weighting matrix $\mathbf{W} = \mathbf{D}^{-1/2}$, then our problem is solved, *sans* burning down the house.

Theorem 8.2.1

Let $\mathbf{E} \sim \mathcal{N}_n\left(0, \sigma^2\mathbf{D}\right)$. If we define $\mathbf{W} = \mathbf{D}^{-1/2}$, then

$$\mathbf{WE} \sim \mathcal{N}_n\left(0,\ \sigma^2\mathbf{I}\right)$$

Proof. Since $\mathbf{W}$ is a diagonal matrix and $\mathbf{E}$ has a Normal distribution, $\mathbf{WE}$ will also follow a Normal distribution. Thus, we need to calculate $\mathbb{E}[\mathbf{WE}]$ and $\mathbb{V}[\mathbf{WE}]$. In doing this, note that $\mathbf{W}$ is not a random matrix; it is known.

The expected value of $\mathbf{WE}$ is

$$\begin{aligned}
\mathbb{E}[\mathbf{WE}] &= \mathbf{W}\ \mathbb{E}[\mathbf{E}] && (8.12)\\
&= \mathbf{W}\ \mathbf{0} && (8.13)\\
&= \mathbf{0} && (8.14)
\end{aligned}$$

For the variance we have

$$\begin{aligned}
\mathbb{V}[\mathbf{WE}] &= \mathbf{W}\ \mathbb{V}[\mathbf{E}]\ \mathbf{W}' && (8.15)\\
&= \mathbf{W}\ \sigma^2\mathbf{D}\ \mathbf{W}' && (8.16)\\
&= \sigma^2\ \mathbf{D}\ \mathbf{W}\ \mathbf{W}' && (8.17)\\
&= \sigma^2\ \mathbf{D}\ \mathbf{D}^{-1/2}\ \left(\mathbf{D}^{-1/2}\right)' && (8.18)\\
&= \sigma^2\ \mathbf{D}\ \mathbf{D}^{-1/2}\ \mathbf{D}^{-1/2} && (8.19)\\
&= \sigma^2\ \mathbf{D}\ \mathbf{D}^{-1} && (8.20)\\
&= \sigma^2\ \mathbf{I} && (8.21)
\end{aligned}$$

In these steps, remember that matrix multiplication is commutative *if* the matrices are diagonal (Theorem M.3.3).

Thus, putting these three parts together gives our conclusion. □

How do we use this theorem? We pre-multiply the model equation by the matrix $\mathbf{W}$ to obtain the following:

$$\mathbf{Y} = \mathbf{XB} + \mathbf{E} \tag{8.22}$$

$$\mathbf{WY} = \mathbf{WXB} + \mathbf{WE} \tag{8.23}$$

Now, redefine the parts to see how useful this result is

$$\mathbf{Y}^* = \mathbf{X}^*\mathbf{B} + \mathbf{E}^* \tag{8.24}$$

with $\mathbf{E}^* \sim \mathcal{N}_n\left(\mathbf{0}, \sigma^2\mathbf{I}\right)$ from Theorem 8.2.1. Thus, we can apply all of our OLS results to WLS, as long as we speak to the transformed response variable $\mathbf{WY}$ and the transformed independent variable(s) $\mathbf{WX}$.

This quickly leads to our weighted least squares estimators of $\mathbf{B}$.

To prove this, we could proceed as we did back in Section 3.1 (page 46). Or, since we have reduced the WLS problem to an OLS problem, we can just write out the results and simplify:

$$\mathbf{b}_{WLS} = \left(\mathbf{X}^{*\prime}\mathbf{X}^*\right)^{-1}\mathbf{X}^{*\prime}\mathbf{Y}^* \tag{8.25}$$

$$= \left((\mathbf{WX})'\mathbf{WX}\right)^{-1}\mathbf{WX}'\mathbf{WY} \tag{8.26}$$

$$= \left(\mathbf{X}'\mathbf{W}'\mathbf{WX}\right)^{-1}\mathbf{X}'\mathbf{W}'\mathbf{WY} \tag{8.27}$$

$$= \left(\mathbf{X}'\mathbf{D}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{D}^{-1}\mathbf{Y} \tag{8.28}$$

It also quickly leads to showing that the WLS estimator is unbiased for $\mathbf{B}$:

Theorem 8.2.2

Under the assumptions of weighted least squares, the WLS estimator for $\mathbf{B}$ is unbiased.

Proof. I am tempted to give this to you as an exercise, but let's see how to prove it.

$$\mathbb{E}\left[\mathbf{b}_{WLS}\right] = \mathbb{E}\left[\left(\mathbf{X}'\mathbf{D}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{D}^{-1}\mathbf{Y}\right] \tag{8.29}$$

Remember that the $\mathbf{D}$ matrix is known, is *not* a random variable.

$$= \left(\mathbf{X}'\mathbf{D}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{D}^{-1}\mathbb{E}[\mathbf{Y}] \tag{8.30}$$

Since $\mathbf{WY} = \mathbf{WXB} + \mathbf{WE}$, and since $\mathbf{W}$ and $\mathbf{D}$ are invertible (why?), we have

$$= \left(\mathbf{X}'\mathbf{D}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{D}^{-1}\mathbf{XB} \tag{8.31}$$

$$= \mathbf{B} \tag{8.32}$$

Thus, the WLS estimator is unbiased if $\mathbf{D}$ is invertible and if $\mathbf{D}$ is known (non-stochastic). I will leave it as an exercise for you to prove this theorem if $\mathbf{D}$ is a random variable independent of $\mathbf{X}$. □ **exercise**

Theorem 8.2.3

Under the assumptions of weighted least squares, the variance of the WLS estimator for $\mathbf{B}$ is

$$\mathbb{V}[\mathbf{b}_{WLS}] = \sigma^2\left(\mathbf{X}'\mathbf{D}^{-1}\mathbf{X}\right)^{-1} \tag{8.33}$$

Proof. There should be no surprises with this proof. All you have to do is figure out what is a random variable and what is not. As such, I leave it as an exercise for you. **exercise**

So very generous of me. =) □

Note that the WLS estimator of **B** is a linear combination of independent Normal random variables. With that final observation, we have the distribution of the WLS estimator of **B**:

$$\mathbf{b}_{WLS} \sim \mathcal{N}\left(\mathbf{B},\ \sigma^2\left(\mathbf{X}'\mathbf{D}^{-1}\mathbf{X}\right)^{-1}\right) \tag{8.34}$$

exercises

Note: We again note that the individual estimators are not independent of each other under typical circumstances. We also note that the confidence intervals for the estimators, estimates of y, etc. can easily be determined in the WLS realm. Nothing new is here, only the mathematics is a bit more involved.

8.2.2 The Real Question Weighted least squares takes care of the problem of heteroskedasticity in our data without introducing any major change in our modeling process or understanding. It just requires that we determine **W** and transform our dependent and independent variables by premultiplying by that weighting matrix.

Question

How do we obtain that weighting matrix?

The best way of obtaining it is through **theory**. The second best way is to utilize the hat matrix, **H**.

Theory: Frequently, knowledge of the problem suggests the weighting matrix. Recall that the $\mathbb{V}[\mathbf{E}]$ covariance matrix measures our uncertainty in the residuals. If that uncertainty is known by the way the experiment is constructed, then **W** can be determined.

For instance, if the dependent variable is the result of a Binomial experiment, perhaps it is the number of successes out of a given number of trials (which may change), then the weighting matrix is just a diagonal matrix of the square root of trial sizes.

Why? Recall that the variance of a binomially-distributed random variable is $\sigma^2 = \frac{\pi(1-\pi)}{N}$. The π are the unknown (constant) population proportion. The N_i is the (known) size within group i. The population parameter is assumed constant. The sample size is measurable.

This leads to the **D** matrix being of the form

$$\mathbf{D} = \begin{bmatrix} 1/N_1 & 0 & 0 & \cdots & 0 \\ 0 & 1/N_2 & 0 & \cdots & 0 \\ 0 & 0 & 1/N_3 & & 0 \\ \vdots & \vdots & & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1/N_n \end{bmatrix} \tag{8.35}$$

The Hat Matrix: When we do not have the theory to know the structure of the **D** matrix, one may want to use the hat matrix to give us a hint about its structure.

Warning: *Make no mistake. This process is not mathematically correct and perfect... but what statistics procedure is? Statistics stands astride the real and the ideal, trying to get as much information about the real while acknowledging its limitations.*

Remembering Chapter 5, not all violations affect inferences the same. Perhaps a good thing for you to do is to use the processes of Chapter 5 to see how much using the hat matrix in lieu of a theoretically-driven **D** *matrix affects the estimates, confidence intervals, and p-values.*

Let us use the symbol $\mathbf{e}$ to represent the *observed* residuals. Until this point, we have only been working with the *theoretical* residuals, $\mathbf{E}$. The conceptual difference between the two is really just the difference between the population (theoretical) and the sample (observations). In effect, the difference is in terms of the variances.

Question

What is the variance of $\mathbf{e}$?

Theorem 8.2.4

The variance of $\mathbf{e}$ is $\mathbb{V}[\mathbf{e}] = \sigma^2(\mathbf{I} - \mathbf{H})$.

Proof.

$$\mathbb{V}[\mathbf{e}] = \mathbb{V}[(\mathbf{I} - \mathbf{H})\mathbf{Y}] \tag{8.36}$$

$$= (\mathbf{I} - \mathbf{H})\mathbb{V}[\mathbf{Y}](\mathbf{I} - \mathbf{H})' \tag{8.37}$$

$$= (\mathbf{I} - \mathbf{H})\sigma^2\mathbf{I}(\mathbf{I} - \mathbf{H}) \tag{8.38}$$

$$= \sigma^2(\mathbf{I} - \mathbf{H}) \tag{8.39}$$

□

So, that was totes cool. What was its purpose? What does it mean?

Remember how we can interpret that variance. It is either the variance of a gazillion observed residuals, *or* we can see it as the uncertainty inherent in the measured residual.

uncertainty

For example, the inherent uncertainty in the first residual can be estimated as

$$s_{1,1} = \text{MSE}(1 - h_{1,1}) \tag{8.40}$$

Here, $h_{1,1}$ is the first element of the diagonal of the hat matrix.

That means, those diagonal elements of $\mathbf{I} - \mathbf{H}$ indicate (or are estimates of) the precision of the y estimate for a given value of x. An estimate of the structure of the $\mathbf{D}$ matrix is just the diagonal of the $\mathbf{I} - \mathbf{H}$ matrix.

precision

Note: The problem is that weighted least squares requires us to *know* the **D** matrix, not that we estimate it from the data. This explains why the hat matrix technique is used only until something better comes along.

It does work nicely, but we statisticians like to "see the math" — sometimes. Also, if we are trying to draw important conclusions, using approximate methods tends to undercut the conclusions for many, especially for those who do not really understand statistics.

Question

What do the off-diagonal elements of $\mathbf{I} - \mathbf{H}$ estimate?

8.3: Generalized Least Squares

Both ordinary least squares and weighted least squares requires the errors be independent. Reality does not always meet this requirement. If the dependent variable consists of repeated measures on one unit over time, such as in modeling stock prices, it is quite likely that the residuals will be correlated. Also, if the dependent variable is measured on geographic structure, such as states in a country or trees in a forest, it is also likely that errors of near units are correlated.

In such examples, the covariance matrix of $\mathbf{E}$ will *not* be diagonal. Thankfully, it is a covariance matrix, and therefore positive definite under the usual assumption of no multi-colinearity (Appendix M, Section M.5.1). Since it is positive definite, it is invertible. Thus, we can do a trick not unlike what we did for weighted least squares.

For a reminder, here are the model equations for ordinary, weighted, and general least squares:

Ordinary Least Squares	$\mathbf{Y} = \mathbf{XB} + \mathbf{E}$
Weighted Least Squares	$\mathbf{Y} = \mathbf{XB} + \mathbf{E}$
General Least Squares	$\mathbf{Y} = \mathbf{XB} + \mathbf{E}$

They sure do look similar. That's because this is the classical linear model (CLM). The requirements on the residuals differs, however:

Ordinary Least Squares	$\mathbf{E} \sim \mathcal{N}\left(\mathbf{0};\, \sigma^2\mathbf{I}\right)$
Weighted Least Squares	$\mathbf{E} \sim \mathcal{N}\left(\mathbf{0};\, \sigma^2\mathbf{D}\right)$
General Least Squares	$\mathbf{E} \sim \mathcal{N}\left(\mathbf{0};\, \boldsymbol{\Sigma}\right)$

For ordinary least squares, the covariance matrix of the residuals is a constant multiple of the identity matrix, $\mathbf{I}$. This indicates the residuals are independent and have the same variance (uncertainty). For weighted least squares, the covariance matrix of the residuals is a constant multiple of a diagonal matrix, $\mathbf{D}$. This indicates the residuals are independent, but possibly with unequal variances.

For generalized least squares (GLS), the covariance matrix is (a constant multiple of) a symmetric, positive definite matrix, $\boldsymbol{\Sigma}$. This indicates the residuals are possibly correlated and with possibly unequal variances.

As with weighted least squares, you *do* need to know the structure of the covariance matrix. This requirement is sometimes met by the structure of the problem. The following are two examples showing how one *can* determine the $\boldsymbol{\Sigma}$ matrix.

An understanding of these examples is not needed. They are here only to illustrate that there are times Σ_i can be determined from the problem.

8.3.1 Time Series Issues When data are collected on a single unit over time, the measurements will tend to be correlated. For instance, the unemployment rate in Ruritania over the past 20 years is 11.35, 11.41, 11.12, 11.08, 10.93, 10.86, 10.96, 11.05, 11.10, 10.87, 10.79, 10.76, 10.94, 10.94, 10.92, 11.01, 11.04, 11.16, 11.13, and 11.14.

Solution: Let us fit this using ordinary least squares regression, then examine the residuals for autocorrelation (correlation between subsequent values).

```
unemp = c(11.35, 11.41, 11.12, 11.08, 10.93, 10.86, 10.96,
  11.05, 11.10, 10.87, 10.79, 10.76, 10.94, 10.94, 10.92,
  11.01, 11.04, 11.16, 11.13, 11.14)
year = 1:20

mod = lm(unemp ~ year)
E = residuals(mod)

autocor.test(E)
```

Note the sample autocorrelation is 0.719 with a p-value of 0.0005 and a 95% confidence interval from 0.393 to 0.884. The p-value indicates the autocorrelation is not 0. The confidence interval indicates that the residuals are moderately-to-highly correlated.

In other words, adjacent observations are not independent, as both ordinary and weighted least squares require. Really, this makes sense because next year's unemployment rate will be heavily influenced by this year's rate.

There are many ways of modeling such a situation. One is called "Autoregressive-1" or AR(1) or ARIMA(1,0,0). This model assumes that the primary correlation is only directly between adjacent years. The covariance matrix, $\boldsymbol{\Sigma}_i$, would have this structure if the correlation between those adjacent years is $\rho = 0.500$:

$$\boldsymbol{\Sigma}_i = \sigma^2 \begin{bmatrix} 1 & 0.5 & 0.25 & 0.125 & \cdots \\ 0.5 & 1 & 0.5 & 0.25 & \cdots \\ 0.25 & 0.5 & 1 & 0.5 & \cdots \\ 0.125 & 0.25 & 0.5 & 1 & \\ \vdots & \vdots & \vdots & \vdots & \ddots \\ & & & & 1 \end{bmatrix} \tag{8.41}$$

You can get this particular matrix using this R code:

```
Sigma = diag(20)
Sigma = 0.5^abs(row(Sigma)-col(Sigma))
```

Note that the matrix has 1s along the diagonal and higher powers of 0.5 farther from the diagonal. The zeroes arise from the fact that the matrix is 20×20; that is, e.g., the entry in cell $(1, 20)$ is actually $0.50^{19} \approx 0$. ♦

Note: Again, this was just an example to show that the structure of $\boldsymbol{\Sigma}_i$ can be determined from some problems. There are entire sub-disciplines of statistics that examine such serial correlation. This sub-discipline is called "time series."

Figure 8.1: *A map showing the administrative divisions (kraj) or the Kingdom of Ruritania. For this example, note that no kraj abuts all other kraj.*

8.3.2 Geographic Issues When data are collected from geographical units, such as neighborhoods, counties, or states, the residuals may be spatially correlated. This is a violation of the independence assumption of ordinary least squares.

How that geographic correlation is modeled is up to the expert (researcher). The subject of spatial modeling is extensive and quite interesting… and important. It can, with appropriate matrices, be extended to modeling three-dimensional spatial correlation over time. If you have the opportunity, I suggest studying this topic (Bivand, Pebesma, and Gómez-Rubio 2013, Blangiardo and Cameletti 2013, Sen 2016). If nothing else, it leads to fun maps!

Figure 8.1 is a map of Ruritania showing the nine Kraj. Note that some kraj abut some kraj but not others. For instance, region CS does not touch region CC.

If we are trying to model the spread of something (disease, unemployment, wealth), we *may* decide to take into consideration the fact that some units neighbor others. Thus, from the map above, we know there is a first-level transmission between CS and CD but not between CS and CF.

Example 1

Geographical Data Let us determine a matrix describing the adjacencies for the nine kraj.

Solution: Check that the following is the adjacency matrix for Ruritania[1]

$$\Sigma_{\not\prec} = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 & 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 1 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 1 & 1 & 0 & 0 & 1 & 0 & 0 & 1 \end{bmatrix} \tag{8.42}$$

It is important to ensure the kraj order for the columns is the same as for the rows. The kraj ordering is: CA, CB, CC, CD, CF, CI, CM, CS, CSM. ♦

symmetry

Note that the adjacency matrix is symmetric. Why will this matrix always be symmetric, regardless of the map? Now, that we know $\Sigma_{\not\prec}$ is symmetric, we can use the discussion following Lemma M.10 to conclude that $\Sigma_{\not\prec}$ is positive definite, which indicates it is allowable as a covariance matrix.

> **Note:** We do not know the constant multiplier, σ^2. No probs. We only need to know the structure of the covariance matrix. We use the data to estimate the constant multiplier σ^2.

Also, note that the analysis based on this covariance structure is only as good as our assumption that the contagion spreads through touch. If it spreads based on some sort of distance, then the $\Sigma_{\not\prec}$ is not correct and we will need to create an appropriate covariance matrix *given our scientific understanding*... if such exists.

Finally, let me reiterate a point I made above. The purpose of this example is *only* to illustrate that these covariance structures *can* be determined from the problem without resorting to estimating them from the data.

> **Note:** However, there is a lesson for all of us here. If we do not know the correct structure of the correlation matrix, then we should use several and see how sensitive our estimates, confidence intervals, and p-values are to that matrix. The results may be very sensitive, which is not a good position to be in, especially if we do not know the right mode of transmission.

[1] One area of **geographical analysis** tries to decide what adjacency rules are appropriate for a given research question. This example uses a simple 0-1 scheme. Other schemes include distances (measured in some manner).

If the estimates, etc. are not sensitive to our choice of covariance matrix, then we need not be as concerned.

The rule is to explore all models that make sense and see how important our assumptions are to our results.

8.3.3 One More Note

Above, we have only focused on being able to determine the structure of the covariance matrix (at least to a scalar multiple). There is one more thing that we need to pay attention to: What is the square root of $\boldsymbol{\Sigma}_{\boldsymbol{\gamma}}$?

When we introduced $\mathbf{D}^{1/2}$ in Section 8.2, we knew we could calculate it. After all, one square root for a diagonal matrix would be

$$\left[\mathbf{D}^{1/2}\right]_{i,j} = d_{i,j}^{1/2} \tag{8.43}$$

That is, the elements of the square root matrix are the square root of the entries of the matrix. This shortcut works because $\mathbf{D}$ is diagonal.

In general, $\boldsymbol{\Sigma}_{\boldsymbol{\gamma}}$ does not have to have a well-defined square root. Some do, but some do not. Without $\boldsymbol{\Sigma}_{\boldsymbol{\gamma}}^{1/2}$, calculating the GLS estimates is not possible using this method.

Sadness abounds.

8.4: Full Example: May the (Strong) Force be with You

Ruritania is a patron of high-energy physics — and of Star Wars enthusiasts. King Rudolph donated several million crowns to Switzerland to aid in researching the strong force.

That money was used at CERN (the *Conseil Européen pour la Recherche Nucléaire*) for several experiments. Each experiment consisted of a beam of protons crashing into a target. That beam had a constant energy level. What changed were the target sizes and the energy level of the proton after the collision. Many experiments were run at each energy level, and the standard deviation of the energies was measured.

In a theory proposed by Ruritanian scientists that is not entirely clear to His Majesty (or to your author), there should be a linear relationship between the cross sectional area and the inverse of the energy. The data are given in Table 8.1.

The first column is the value of the independent variable. The second column is the mean of the energy level of the photon after the collision. The third column is the standard deviation in those energy levels. Note that the variability at each cross-section differs. This is based on both the number of experiments *and* the inherent variability at that area.

Cross Section [b]	Energy [MeV]	St. Dev. [MeV]
1	848.9	7.8
2	476.9	9.2
3	350.9	9.4
4	289.2	10.2
5	251.7	7.4
6	225.8	9.3
7	209.7	7.2
8	193.9	5.3

Table 8.1: *Data for the example regarding the strong nuclear force. Units are given in brackets.*

8.4.1 Ordinary Least Squares

Let us ignore the different uncertainties in each energy level (the standard deviations). That is, let us just fit this as an OLS model.

Here is the code:

```
barns  = c(1,2,3,4,5,6,7,8)
energy = c(848.9,476.9,350.9,289.2,251.7,225.8,209.7,193.9)
Ibarns = 1/barns

modOLS = lm(energy ~ Ibarns)
summary(modOLS)
confint(modOLS)
```

The output suggests that the relationship between the cross sectional area of the target and the inverse of the resulting energy of the photon is statistically significant.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 101.9148     0.5464   186.5 1.60e-12 ***
Ibarns      747.5306     1.2505   597.8 1.48e-15 ***
---

Residual standard error: 0.9719 on 6 degrees of freedom
Multiple R-squared:      1, Adjusted R-squared:      1
F-statistic: 3.574e+05 on 1 and 6 DF,  p-value: 1.479e-15
```

A 95% confidence interval for the relationship is from 744.5 to 750.6 (with units of MeV·barn).

8.4.2 Weighted Least Squares

Note, however, that the uncertainty in the measurements varies. We are more uncertain with some of our estimated energy than with others. If we do not take this uncertainty into consideration, we may be biasing our results. To include this information, we can use weighted least squares regression.

The code to fit with weighted least regression is as follows:

```
barns  = c(1,2,3,4,5,6,7,8)
energy = c(848.9,476.9,350.9,289.2,251.7,225.8,209.7,193.9)
stdev  = c(7.8,9.2,9.4,10.2,7.4,9.3,7.2,5.3)
v      = stdev^2

Ibarns = 1/barns

modWLS = lm(energy ~ Ibarns, weight=1/v)
summary(modWLS)
confint(modWLS)
```

Note that we are 95% confident the effect of the target's cross section on the resulting energy is from 744.5 to 751.2 MeV·barn.

> **Note:** In R, as in many statistical programs, the weights you provide in the function call are inversely proportional to the variances. This is why we used `weight=1/v` in the function call.
>
> With that being said, **always check the documentation** to make doubly sure. Frequently, this information is difficult to find and an error will not be thrown to let you know.

How do we know this?

Or, an even better question:

Question

How do we check that the weights really are proportional to the inverse of the variance?

The short answer is to check your work "by hand." As we once did OLS by hand using the matrix functions in R, we can also do WLS by hand.

As always, make sure you know what each line does and why they are put together as they are:

```
## The data
barns  = c(1,2,3,4,5,6,7,8)
energy = c(848.9,476.9,350.9,289.2,251.7,225.8,209.7,193.9)
stdev  = c(7.8,9.2,9.4,10.2,7.4,9.3,7.2,5.3)

## A few minor calculations
v = stdev^2
Ibarns = 1/barns
n = length(Ibarns)

## The needed matrices
Y = matrix(energy, ncol=1)
X = matrix( c(rep(1,n), Ibarns), ncol=2 )
colnames(X) = c("b0","b1")
D = diag(v)

## The estimate vector
solve(t(X) %*% solve(D) %*% X) %*% t(X) %*% solve(D) %*% Y
```

When I run this, I get the following output

```
         [,1]
b0 101.6074
b1 747.8364
```

These are the matrix calculations according to Equation 8.28 on page 234. Thus, this is the *correct* answer. Double-check that this known correct answer matches the answer given in `modWLS`. If it does not, then you will need to change the expression in the `weights` part of the function call. In R, it will match. In other pieces of software, it may not. **Be aware.**

> **Note:** The difference between the effects estimated from using ordinary least squares and using weighted least squares is rather minor in this example. It need not be, as the next example shows.

8.5: Full Example: Elections in Ruritania

Even though it is an absolute monarchy, national elections are held in Ruritania to elect members of the Ruritanian parliament, the *Národní Shromáždění* (National Assembly). There are many parties represented in the parliament, but the party that consistently receives a majority of the seats and votes is the monarchist *Pohyb pro Ruritánii* (PR; Movement for Ruritania).

The main opposition party is the *Demokratické Hnutí* (DH; Democratic Movement) party, but votes are also usually received by the *Socialistická* (SP; Socialist Party), *Křstanská Demokratická* (KD, Christian Democratic), and *Republikánská* (RS, Republican) parties.

It is fortuitous that Ruritania does not use computerized ballots. They use ballot papers for the parliamentary election that consist of the party names, symbols, and abbreviations... and a box for the voter to place their inked fingerprint next to the party. After voting, the ballot is placed in a ballot box to await counting.

At the end of the evening, the ballot papers at each precinct are securely transported to the division headquarters, where they are counted by electoral officials. Each ballot is checked by that official to ensure that it was lawfully cast and that the "will of the voter" can be discerned.

When the division is finished counting the ballots, the totals are then telephoned to the Independent Electoral Commission (*Nezávislá Volební Komise*, NVK) in the capital. With much pomp and circumstance, and not a little fanfare, the division totals are added and reported to the people.

After the last election, the exiles in Denmark claimed that the ballot boxes were stuffed. That is, the ballot boxes had votes for the PR party in them before voting began. Because guarantees of the secret ballot are built into the Ruritanian Constitution, the ballot boxes are opaque.

In other words, direct evidence of ballot box stuffing does not exist, only claims by those who live in exile in another country (Denmark). However, if ballot box stuffing existed in this election to any great extent, it would leave evidence. Why/how?

Question

What do stuffing ballots have that naturally cast ballots do not?

Easy: The stuffing ballots are all for the ruling PR party *and* they are all completed (filled in) correctly. The naturally cast ballots will consist of votes for all parties and will include ballots not filled-in correctly.

And so, in the presence of systematic and significant ballot box stuffing, there will be a relationship between the invalidation rate and the level of support for the ruling party.

That is the theory. The exiles are paying for this analysis. We like the money, so we need to be confident — and clear — in our conclusions. The NVK is providing the official counts in the `rur2013parl` data file, so we need to ensure that the statistical analysis is clean. That is, it is up to us to do the analysis correctly, neither concluding too much nor too little.

And, as always, being clear in our reasoning.

8.5.1 Ordinary Least Squares

The first analysis we will do is ordinary least squares. The dependent variable is the invalidation rate; the independent variable is the support for the Movement for Ruritania (PR) party. Why them? Since they were in charge before the election, they were in position to stuff the ballot boxes.

Here is the code to load the data, create the variables, fit the model, and determine if a relationship between the invalidation rate and PR support rate can be detected.

```
votes=read.csv("http://rur.kvasaheim.com/data/rur2013parl.csv")
attach(votes)

Valid  = Total-Invalid
pPR    = PR/Valid
pInv   = Invalid/Total

modOLS = lm(pInv ~ pPR)
summary(modOLS)
```

These results indicate that we did not detect a relationship... at the $\alpha = 0.05$ level ($p = 0.0668$). Thus, ordinary least squares *did not detect* unfairness in the vote.

> **Note:** It is important to emphasize here that the correct terminology is that we did not detect unfairness. We cannot say there was no fairness. We can only say we didn't detect it.

Remember to check the assumptions. This point cannot be over-emphasized. If the assumptions are not met, then the model is not correct. Well, not *perfectly* correct. See Chapter 5 for a discussion of this point.

8.5.2 Weighted Least Squares

Note that ordinary least squares is *not* be the best option here. The invalidation rate has greater inherent variability in smaller divisions than in larger. We know this because of the distribution of the invalidation rate. Invalidation counts follow something akin to a Binomial distribution. Its two parameters are sample size (number of votes cast) and success probability (invalidation rate). The variance of a Binomial random variable is $n\pi(1-\pi)$.

Dividing the invalidation count by the number of votes cast gives the invalidation rate. The distribution of the invalidation rate can be approximated with a Normal distribution (see the Central Limit Theorem, Section S.6.4). The expected value of the observed invalidation rate is π, the inherent invalidation rate. The variance is $\pi(1-\pi)/n \propto 1/n$.

Because the data are heteroskedastic in nature, and because the structure of the heteroskedasticity is known, weighted least squares will be more appropriate here.

Here is the code. Compare it to the ordinary least squares code from above.

```
Valid = Total-Invalid
pPR   = PR/Valid
pInv  = Invalid/Total

modWLS = lm(pInv ~ pPR, weights=Total)
summary(modWLS)
```

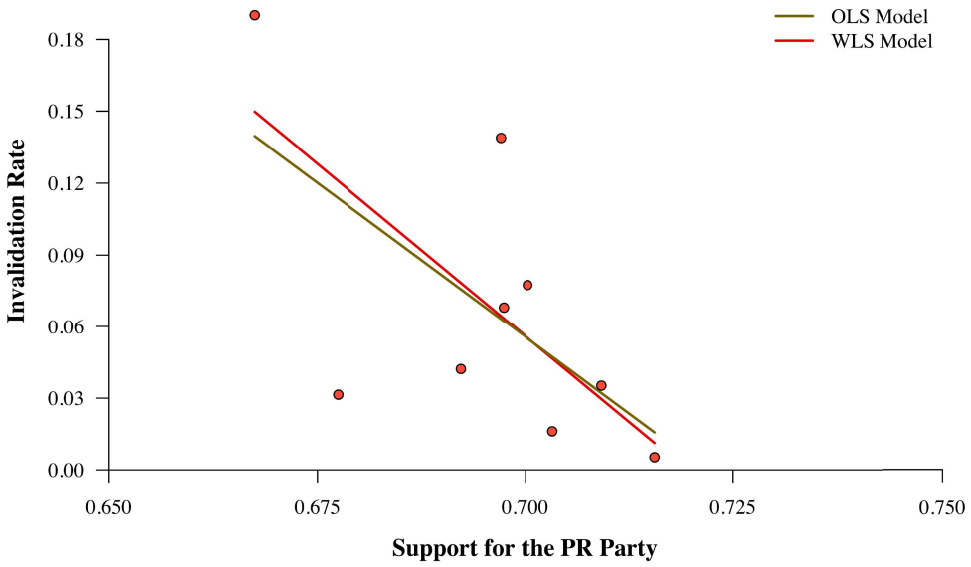


Figure 8.2: *An invalidation plot for the 2013 Ruritanian parliamentary election. The lines of best fit are provided. The OLS fit is in brown and the WLS is in red.*

This produces the following (abbreviated) output:

```
Coefficients:
             Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)   2.0726      0.7109    2.915    0.0225 *
pPR          -2.8808      1.0124   -2.845    0.0249 *
```

Note that this method *did* detect a relationship between the invalidation rate and the PR party support rate (p-value $< 0.05 = \alpha$). Furthermore, that differential invalidation helped the ruling party. The negative coefficient indicates that those *kraj* with higher PR support also tended to count more of the votes (reject fewer). Thus, we can conclude that the data are consistent with the exile claim of ballot box stuffing. Figure 8.2 illustrates this.

Question

How much is the differential invalidation?

Well, from the regression output, we know that when the support for PR increases by 10 percentage points (from 65 to 75%, for instance), the invalidation drops by an average of 27pp.

That seems rather substantial to me.

proof

Note: Nothing in statistics ever constitutes *proof*. Nothing. Ever. Period. End of thought. *Žádné další*!

Statistics only provides evidence in favor of — or against — the null hypothesis. In this case, the p-value is 0.0249. If the null hypothesis is correct, then we would observed results this extreme or more so 2.49% of the time. This is not too rare, especially when you realize you are claiming the government cheated. Cheating is a more serious claim than just that someone was mistaken.

It is always better to report the results, interpret the results, be explicit that there is no proof and that the null hypothesis has a non-zero probability of being reality.

Do not live your statistical life ruled by $\alpha = 0.05$. Realize — and accept — that the p-value is a measure of how well the data support the null hypothesis, the hypothesis of no relationship/difference/effect/evidence.

8.6: Conclusion

In the previous chapters, we focused on ordinary least squares. This method required that the residuals were independent and identically distributed. From that assumption, we were able to generate a series of rich conclusions.

However, it is not true that residuals are always identically distributed or independent. While we did find a way of "fixing" the problem of heteroskedasticity, it is frequently better to use a modeling scheme that uses that heteroskedasticity instead of merely finding a way of ignoring it. This is what weighted least squares does. If you have theory behind how the variances should vary for each record, you can use this method. If not, then you are reduced to the "fixes" of Chapter 7.

Similarly, if your data are not independent, but you understand the structure of that dependence, you can use generalized least squares to model the relationship better... as long as that covariance matrix has an inverse square root matrix. And, there is no guarantee that it does.

8.7: End-of-Chapter Materials

8.7.1 R Functions

In this chapter, we were introduced to a few R functions that will be useful in the future. These are listed here.

Packages:

nlme This package gives R the functionality to fit generalized least squares using the `gls` function. It actually has many other useful functions that allow us to fit non-linear models and random-effects models. Those are beyond the scope of this book, however.

RFS This package does not yet exist. It is a package that adds much general functionality to R. In lieu of using `library(RFS)` to access these functions, run the following line in R:
`source("http://rfs.kvasaheim.com/rfs.R")`

Statistics:

autocor.test(e) This function calculates the auto-correlation, which is just the correlation between sequential values in the vector. It is a part of the `RFS` package.

gls(formula) This function performs generalized least squares regression. It even allows you to specify the correlation structure via the `correlation` parameter.

lm(formula) This function performs linear regression on the data, with the supplied formula. If you specify the `weights`, then they are applied and you are fitting the model using weighted least squares. As there is much information contained in this function, you will want to save the results in a variable.

residuals(mod) This calculates the simple residuals in a model, the observed values minus the predicted values.

MATHEMATICS:

`%*%` This multiplies two matrices in R. Thus, running the command `A%*%B` will return the matrix product **AB**.

`abs(x)` This returns the absolute value of the real number x, a.k.a. $|x|$.

`column(A)` This returns the column number of the matrix **A**.

`diag(n)` If n is an integer, then this returns the $\mathbf{I}_n$ identity matrix.

`diag(v)` This returns a diagonal matrix with the elements of the vector **v** along the diagonal.

`diag(A)` This returns the diagonal entries of the matrix **A**.

`rep(n,x)` This returns a vector of the number x repeated n times.

`row(A)` This returns the row number of the matrix **A**.

`solve(A)` This returns the inverse of the matrix **A**.

`t(A)` This returns the transpose of the matrix **A**.

8.7.2 Exercises

1. Let $\mathbf{E} \sim \mathcal{N}\left(\mathbf{0};\ \sigma^2\mathbf{D}\right)$ be the residuals. Prove that if $\mathbf{D}$ is a diagonal covariance matrix, then it is invertible.
2. Let $\mathbf{E} \sim \mathcal{N}\left(\mathbf{0};\ \sigma^2\mathbf{D}\right)$ be the residuals. Here, $\mathbf{D}$ is a diagonal covariance matrix. Determine a matrix $\mathbf{W}$ such that $\mathbf{WW} = \mathbf{D}$.
3. Prove Theorem 8.2.2.
4. Under the assumptions of weighted least squares, determine the formula for a confidence interval for β_1.
5. What is the difference between $\mathbf{e}$ and $\mathbf{E}$?
6. Under the assumptions of generalized least squares, determine the formula for the estimator of $\mathbf{B}$.
7. Under the assumptions of generalized least squares, determine the formula for a confidence interval for $\mathbf{b}$.
8. Determine if Theorem 8.2.1 holds if the weights matrix $\mathbf{D}$ is a random matrix independent of $\mathbf{X}$. If it does not, what is the distribution of $\mathbf{WE}$?
9. Prove Theorem 8.2.2 if $\mathbf{D}$ is independent of $\mathbf{X}$.
10. Theorem 8.2.3 requires $\mathbf{D}$ is non-random. Determine the variance of $\mathbf{b}_{\text{wls}}$ if $\mathbf{D}$ is random, but independent of $\mathbf{X}$.
11. In Example 8.3.2, I state that the adjacency matrix is symmetric. Explain why this is so.

8.7.3 Theory Readings

- Adrian Baddeley, Ege Rubak, and Rolf Turner. (2015) *Spatial Point Patterns: Methodology and Applications with R*. Boca Raton, FL: Chapman & Hall/CRC.
- Roger S. Bivand, Edzer Pebesma, and Virgilio Gómez-Rubio. (2013) *Applied Spatial Data Analysis with R*. New York: Springer-Verlag.
- Marta Blangiardo and Michela Cameletti. (2015) *Spatial and Spatio-temporal Bayesian Models with R*. Hoboken, NJ: John Wiley & Sons.
- Chris Brunsdon and Lex Comber. (2015) *An Introduction to R for Spatial Analysis and Mapping*. Thousand Oaks, CA: SAGE Publications.
- Robert Haining. (2003) *Spatial Data Analysis: Theory and Practice*. Cambridge, UK: Cambridge University Press.
- Tonny J. Oyana and Florence Margai. (2013) *Spatial Analysis: Statistics, Visualization, and Computational Methods*. Boca Raton, FL: Chapman & Hall/CRC.
- Zekai Sen. (2016) *Spatial Modeling Principles in Earth Sciences*. New York: Springer-Verlag.
- Thorsten Wiegand and Kirk A. Moloney. (2013) *Handbook of Spatial Point-Pattern Analysis in Ecology*. Boca Raton, FL: Chapman & Hall/CRC.