Lecture 24: Weighted and Generalized Least Squares

1 Weighted Least Squares

When we use ordinary least squares to estimate linear regression, we minimize the mean squared error:

$$MSE(\mathbf{b}) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \mathbf{X}_{i\cdot}\beta)^2 \tag{1}$$

where $\mathbf{X}_{i\cdot}$ is the i^{th} row of $\mathbf{X}$. The solution is

$$\widehat{\beta}_{OLS} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}. \tag{2}$$

Suppose we minimize the weighted MSE

$$WMSE(\mathbf{b}, w_1, \ldots w_n) = \frac{1}{n}\sum_{i=1}^{n} w_i(Y_i - \mathbf{X}_{i\cdot}\mathbf{b})^2. \tag{3}$$

This includes ordinary least squares as the special case where all the weights $w_i = 1$. We can solve it by the same kind of linear algebra we used to solve the ordinary linear least squares problem. If we write $\mathbf{W}$ for the matrix with the w_i on the diagonal and zeroes everywhere else, then

$$\begin{aligned} WMSE &= n^{-1}(\mathbf{Y} - \mathbf{X}\mathbf{b})^T\mathbf{W}(\mathbf{Y} - \mathbf{X}\mathbf{b}) && (4)\\ &= \frac{1}{n}\left(\mathbf{Y}^T\mathbf{W}\mathbf{Y} - \mathbf{Y}^T\mathbf{W}\mathbf{X}\mathbf{b} - \mathbf{b}^T\mathbf{X}^T\mathbf{W}\mathbf{Y} + \mathbf{b}^T\mathbf{X}^T\mathbf{W}\mathbf{X}\mathbf{b}\right). && (5)\end{aligned}$$

Differentiating with respect to $\mathbf{b}$, we get as the gradient

$$\nabla_{\mathbf{b}} WMSE = \frac{2}{n}\left(-\mathbf{X}^T\mathbf{W}\mathbf{Y} + \mathbf{X}^T\mathbf{W}\mathbf{X}\mathbf{b}\right).$$

Setting this to zero at the optimum and solving,

$$\widehat{\beta}_{WLS} = (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\mathbf{Y}. \tag{6}$$

But why would we want to minimize Eq. 3?

1. *Focusing accuracy.* We may care very strongly about predicting the response for certain values of the input — ones we expect to see often again, ones where mistakes are especially costly or embarrassing or painful, etc. — than others. If we give the points near that region big weights, and points elsewhere smaller weights, the regression will be pulled towards matching the data in that region.

2. *Discounting imprecision.* Ordinary least squares minimizes the squared error when the variance of the noise terms ϵ is constant over all observations, so we're measuring the regression function with the same precision elsewhere. This situation, of constant noise variance, is called **homoskedasticity**. Often however the magnitude of the noise is not constant, and the data are **heteroskedastic**. Let $\sigma_i^2 = \mathbb{E}[\epsilon_i|X_i] = \sigma_i^2$. Homoskedasticity means that $\sigma_i^2 = \sigma^2$ for all i. Heteroskedasticity means that the σ_i^2 can be different.

When we have heteroskedasticity, ordinary least squares is no longer the optimal estimate — we'll see soon that other estimators can be unbiased and have smaller variance. If however we know the noise variance σ_i^2 at each measurement i, and set $w_i = 1/\sigma_i^2$, we get to minimize the variance of estimation.

3. *Doing something else.* There are a number of other optimization problems which can be transformed into, or approximated by, weighted least squares. The most important of these arises from **generalized linear models**, where the mean response is some nonlinear function of a linear predictor; we will look at them in 402.

2 Heteroskedasticity

Suppose that

$$Y_i = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_p X_{pi} + \epsilon_i$$

where $\mathbb{E}[\epsilon_i] = 0$ and $\mathrm{Var}[\epsilon_i] = \sigma_i^2$. (As usual, we are treating the X_i's as fixed.) This is called the *Heteroskedastic linear regression model.* **For now, assume we know $\sigma_1, \ldots, \sigma_p$.**

The model is

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon$$

where now (treating $\mathbf{X}$ as fixed), $\mathbf{E}[\epsilon] = \mathbf{0}$ and $\mathrm{Var}(\epsilon) = \Sigma$ and Σ is no longer of the form $\sigma^2 I$. The weighted least squares estimator is

$$\widehat{\beta} = KY$$

where $K = (\mathbf{X}^T W \mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}$. So,

$$\widehat{\beta} = KY = K(\mathbf{X}\beta + \epsilon) = \beta + K\epsilon.$$

Hence,

$$\mathbb{E}[\widehat{\beta}] = \beta, \quad \mathrm{Var}(\widehat{\beta}) = K\Sigma K^T. \tag{7}$$

Note that the estimator is unbiased.

3 The Gauss-Markov Theorem

We've seen that when we do weighted least squares, our estimates of β are linear in $\mathbf{Y}$, and unbiased:

$$\widehat{\beta} = (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\mathbf{Y}$$

and $\mathbf{E}[\widehat{\beta}] = \beta$.

Let us consider a special case: suppose we take $\mathbf{W} = \Sigma^{-1}$. Then, from (7)

$$\mathrm{Var}(\widehat{\beta}) = K\Sigma K^T = \left[(\mathbf{X}^T\Sigma^{-1}\mathbf{X})^{-1}\mathbf{X}^T\Sigma^{-1}\right]\Sigma\left[\Sigma^{-1}\mathbf{X}(\mathbf{X}^T\Sigma^{-1}\mathbf{X})^{-1}\right] = (\mathbf{X}^T\Sigma^{-1}\mathbf{X})^{-1}. \tag{8}$$

We will now show that $\widehat{\beta}$ is, in a certain sense, optimal.

Like any optimality result, it is crucial to lay out carefully the range of possible alternatives, and the criterion by which those alternatives will be compared. The classical optimality result for estimating linear models is the **Gauss-Markov theorem**, which takes the range of possibilities to be *linear, unbiased estimators of* β, and the criterion to be *variance of the estimator*.

Any linear estimator, say $\widetilde{\beta}$, could be written as

$$\widetilde{\beta} = \mathbf{QY}$$

where $\mathbf{Q}$ would be a $(p+1) \times n$ matrix. We will show that if $\widetilde{\beta}$ is unbiased, then it has larger variance than $\widehat{\beta}_{\mathrm{WLS}}$.

For $\widetilde{\beta}$ to be an unbiased estimator, we must have

$$\mathbb{E}\left[\mathbf{QY}|\mathbf{X}\right] = \mathbf{QX}\beta = \beta.$$

Since this must hold for all β and all $\mathbf{X}$, we have to have $\mathbf{QX} = \mathbf{I}$. The variance is then

$$\mathrm{Var}\left[\mathbf{QY}|\mathbf{X}\right] = \mathbf{Q}\mathrm{Var}\left[\epsilon|\mathbf{X}\right]\mathbf{Q}^T = \mathbf{Q\Sigma Q} \tag{9}$$

where $\mathbf{\Sigma} = \mathrm{Var}\left[\epsilon|\mathbf{X}\right]$. Let

$$\mathbf{K} \equiv (\mathbf{X}^T\mathbf{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{\Sigma}^{-1}.$$

Now, whatever $\mathbf{Q}$ might be, we can always write

$$\mathbf{Q} = \mathbf{K} + \mathbf{R} \tag{10}$$

for some matrix $\mathbf{R}$. The unbiasedness constraint on $\mathbf{Q}$ implies that

$$\mathbf{RX} = \mathbf{0}$$

because $\mathbf{KX} = \mathbf{I}$. Now we substitute Eq. 10 into Eq. 9:

$$\begin{aligned}
\mathrm{Var}\left[\widetilde{\beta}\right] &= (\mathbf{K}+\mathbf{R})\mathbf{\Sigma}(\mathbf{K}+\mathbf{R})^T & (11)\\
&= \mathbf{K\Sigma K}^T + \mathbf{R\Sigma K}^T + \mathbf{K\Sigma R}^T + \mathbf{R\Sigma R}^T & (12)\\
&= (\mathbf{X}^T\mathbf{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{\Sigma}^{-1}\mathbf{\Sigma}\mathbf{\Sigma}^{-1}\mathbf{X}(\mathbf{X}^T\mathbf{\Sigma}^{-1}\mathbf{X})^{-1} & (13)\\
&\quad +\mathbf{R\Sigma\Sigma}^{-1}\mathbf{X}(\mathbf{X}^T\mathbf{\Sigma}^{-1}\mathbf{X})^{-1}\\
&\quad +(\mathbf{X}^T\mathbf{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{\Sigma}^{-1}\mathbf{\Sigma}\mathbf{R}^T\\
&\quad +\mathbf{R\Sigma R}^T\\
&= (\mathbf{X}^T\mathbf{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{\Sigma}^{-1}\mathbf{X}(\mathbf{X}^T\mathbf{\Sigma}^{-1}\mathbf{X})^{-1} & (14)\\
&\quad +\mathbf{RX}(\mathbf{X}^T\mathbf{\Sigma}^{-1}\mathbf{X})^{-1} + (\mathbf{X}^T\mathbf{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{R}^T\\
&\quad +\mathbf{R\Sigma R}^T\\
&= (\mathbf{X}^T\mathbf{\Sigma}^{-1}\mathbf{X})^{-1} + \mathbf{R\Sigma R}^T & (15)
\end{aligned}$$

where the last step uses the fact that $\mathbf{RX} = \mathbf{0}$ (and so $\mathbf{X}^T\mathbf{R}^T = \mathbf{0}^T$).

Hence,

$$\mathrm{Var}\left[\tilde{\beta}_i\right] = (\mathbf{X}^T\mathbf{\Sigma}^{-1}\mathbf{X})^{-1}_{ii} + R_i^T\Sigma R_i = \mathrm{Var}\left[\widehat{\beta}_i\right] + R_i^T\Sigma R_i$$

where R_i denotes the i^{th} row of $\mathbf{R}$. Since $\mathbf{\Sigma}$ is a covariance matrix, it's positive definite, meaning that $a\mathbf{\Sigma}a^T \geq 0$ for any vector a. Hence,

$$\mathrm{Var}\left[\tilde{\beta}_i\right] \geq \mathrm{Var}\left[\widehat{\beta}_i\right].$$

We conclude that WLS, with $\mathbf{W} = \mathbf{\Sigma}^{-1}$, has the least variance among all possible linear, unbiased estimators of the regression coefficients.

Notes:

1. If all the noise variances are equal, then we've proved the optimality of OLS.
2. The theorem doesn't rule out linear, biased estimators with smaller variance. As an example, albeit a trivial one, $\mathbf{0Y}$ is linear and has variance $\mathbf{0}$, but is (generally) very biased.
3. The theorem also doesn't rule out non-linear unbiased estimators of smaller variance. Or indeed non-linear biased estimators of even smaller variance.
4. The proof actually doesn't require the variance matrix to be diagonal.

4 Finding the Variance and Weights

So far we have assumed that we know $\sigma_1, \ldots, \sigma_p$. Here are some cases where this might be true.

Multiple measurements. The easiest case is when our measurements of the response are actually averages over individual measurements, each with some variance σ^2. If some Y_i are based on averaging more individual measurements than others, there will be heteroskedasticity. The variance of the average of n_i uncorrelated measurements will be σ^2/n_i, so in this situation we could take $w_i \propto n_i$.

Binomial counts Suppose our response variable is a count, derived from a binomial distribution, i.e., $Y_i \sim \text{Binom}(n_i, p_i)$. We would usually model p_i as a function of the predictor variables — at this level of statistical knowledge, a linear function. This would imply that Y_i had expectation $n_i p_i$, and variance $n_i p_i(1 - p_i)$. We would be well-advised to use this formula for the variance, rather than pretending that all observations had equal variance.

Proportions based on binomials If our response variable is a proportion based on a binomial, we'd see an expectation value of p_i and a variance of $\frac{p_i(1-p_i)}{n_i}$. Again, this is not equal across different values of n_i, or for that matter different values of p_i.

Poisson counts Binomial counts have a hard upper limit, n_i; if the upper limit is immense or even (theoretically) infinite, we may be better off using a Poisson distribution. In such situations, the mean of the Poisson λ_i will be a (possibly-linear) function of the predictors, and the variance will *also* be equal to λ_i.

Other counts The binomial and Poisson distributions rest on independence across "trials" (whatever those might be). There are a range of discrete probability models which allow for correlation across trials (leadings to more or less variance). These may, in particular situations, be more appropriate.

5 Conditional Variance Function Estimation

If we don't know the variances, we might be able to estimate them. There are two common ways to estimate conditional variances, which differ slightly in how they use non-parametric smoothing. (We will discuss non-parametric smoothing in more detail later and in 402).

Method 1:

1. Estimate $m(x)$ with your favorite regression method, getting $\widehat{m}(x)$.
2. Construct the **squared residuals**, $u_i = (Y_i - \widehat{m}(x_i))^2$.
3. Use your favorite *non-parametric* method to estimate the conditional mean of the u_i, call it $\widehat{q}(x)$.
4. Estimate the variance using $\widehat{\sigma}^2_x = \widehat{q}(x)$.

Here is method 2:

1. Estimate $m(x)$ with your favorite regression method, getting $\widehat{m}(x)$.
2. Construct the **log squared residuals**, $z_i = \log{(Y_i - \widehat{m}(x_i))}^2$.
3. Use your favorite *non-parametric* method to estimate the conditional mean of the z_i, call it $\widehat{s}(x)$.
4. Predict the variance using $\widehat{\sigma}^2_x = \exp \widehat{s}(x)$.

The second method ensures that the estimates variances are positive.

We are estimating the variance function to do weighted least squares, but these methods can be used more generally. It's often important to understand variance in its own right, and this is a general method for estimating it. Our estimate of the variance function depends on first having a good estimate of the regression function

5.1 Example

```
> plot(x,y)
> out = lm(y ~ x)
> abline(out)
> summary(out)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.02934    0.03935   0.745    0.458
x            1.89866    0.07092  26.772   <2e-16 ***

> plot(x,rstudent(out))
> abline(h=0)
>
> u = log((resid(out))^2)
> tmp = loess(u ~ x)
> s2 = exp(tmp$fitted)
> plot(x,s2)
> w = 1/s2
> out2 = lm(y ˜ x,weights=w)
> summary(out2)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.002967   0.008252    0.36     0.72
x           2.046929   0.043872   46.66   <2e-16 ***
```

6 Correlated Noise and Generalized Least Squares

Sometimes, we might believe the right model is (in matrix form)

$$\begin{aligned} \mathbf{Y} &= \mathbf{X}\beta + \epsilon & (16)\\ \mathbb{E}\left[\epsilon|\mathbf{X}\right] &= \mathbf{0} & (17)\\ \mathrm{Var}\left[\epsilon|\mathbf{X}\right] &= \mathbf{\Sigma} & (18) \end{aligned}$$

where the matrix $\mathbf{\Sigma}$ is *not* diagonal. The off-diagonal entries represent covariance in the noise terms, $\mathrm{Cov}\left[\epsilon_i, \epsilon_j\right] = \Sigma_{ij}$. How should we estimate β? Here are two approaches. We assume that Σ is known.

Approach 1. Because $\mathbf{\Sigma}$ is a variance matrix, we know it is square, symmetric, and positive-definite. This implies that there exists a square root matrix $\mathbf{S}$ such that $\mathbf{SS} = \Sigma$. Now we multiply our model by $\mathbf{S}^{-1}$:

$$\mathbf{S}^{-1}\mathbf{Y} = \mathbf{S}^{-1}\mathbf{X}\beta + \mathbf{S}^{-1}\epsilon$$

This os a linear regression of $\mathbf{S}^{-1}\mathbf{Y}$ on $\mathbf{S}^{-1}\mathbf{X}$, with the same coefficients β as our original regression. However, we have improved the properties of the noise. The noise is still zero in expectation,

$$\mathbb{E}\left[\mathbf{S}^{-1}\epsilon|\mathbf{X}\right] = \mathbf{S}^{-1}\mathbf{0} = \mathbf{0}$$

but

$$\begin{aligned} \mathrm{Var}\left[\mathbf{S}^{-1}\epsilon|\mathbf{X}\right] &= \mathbf{S}^{-1}\mathrm{Var}\left[\epsilon|\mathbf{X}\right]\mathbf{S}^{-T} & (19)\\ &= \mathbf{S}^{-1}\mathbf{\Sigma}\mathbf{S}^{-T} & (20)\\ &= \mathbf{S}^{-1}\mathbf{S}\mathbf{S}^{T}\mathbf{S}^{-T} & (21)\\ &= \mathbf{I}. & (22) \end{aligned}$$

To sum up, if we know $\mathbf{\Sigma}$, we can estimate β by doing an ordinary least squares regression of $\mathbf{S}^{-1}\mathbf{Y}$ on $\mathbf{S}^{-1}\mathbf{X}$. The estimate is

$$\begin{aligned} \widehat{\beta} &= ((\mathbf{S}^{-1}\mathbf{X})^T\mathbf{S}^{-1}\mathbf{X})^{-1}(\mathbf{S}^{-1}\mathbf{X})^T\mathbf{S}^{-1}\mathbf{Y} & (23)\\ &= (\mathbf{X}^T\mathbf{S}^{-T}\mathbf{S}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{S}^{-T}\mathbf{S}^{-1}\mathbf{Y} & (24)\\ &= (\mathbf{X}^T\mathbf{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{\Sigma}^{-1}\mathbf{Y}. & (25) \end{aligned}$$

This looks just like our weighted least squares estimate, only with $\mathbf{\Sigma}^{-1}$ in place of $\mathbf{W}$.

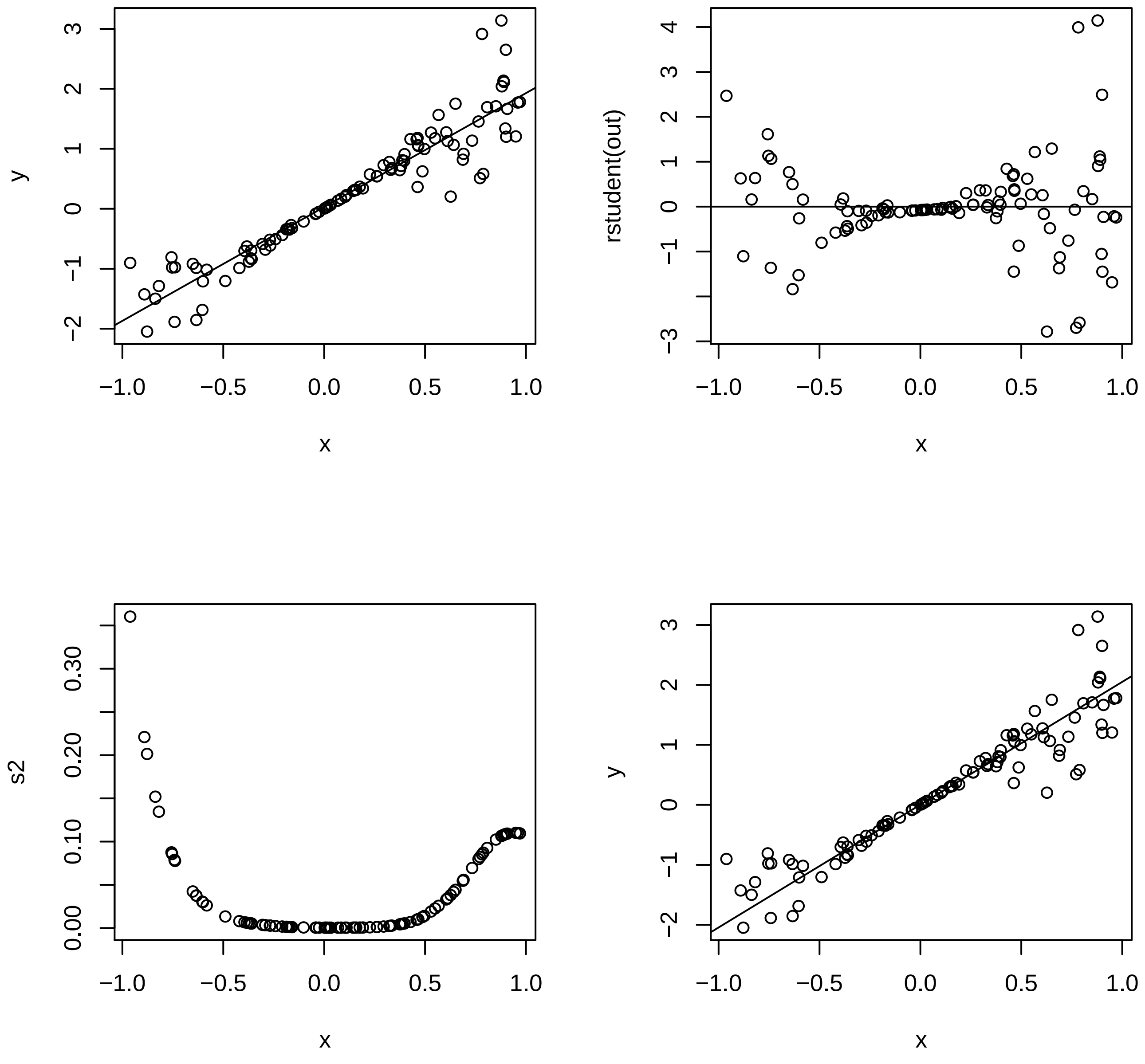

FIGURE 1: *Top left: data with fitted line. Top right: residuals. Bottom left: Plot of estimates variances. Bottom right: Data with weigthed least squares line.*

Approach 2. This resemblance is no mere coincidence. We can write the WLS problem as that of minimizing $(\mathbf{Y} - \mathbf{X}\beta)^T\mathbf{W}(\mathbf{Y} - \mathbf{X}\beta)$, for a diagonal matrix $\mathbf{W}$. Suppose we try instead to minimize

$$(\mathbf{Y} - \mathbf{X}\beta)^T\mathbf{W}(\mathbf{Y} - \mathbf{X}\beta)$$

for a *non-diagonal*, but still symmetric and positive-definite, matrix $\mathbf{W}$. This is called a **generalized least squares** (GLS) problem. Every single step we went through before is still valid, because none of it rested on $\mathbf{W}$ being diagonal, so

$$\widehat{\beta}_{GLS} = (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\mathbf{Y}. \tag{26}$$

We have just seen is that if we set $\mathbf{W} = \mathbf{\Sigma}^{-1}$, we also get this solution when we transform the variables so as to de-correlate the noise, and then do ordinary least squares. This should at least make it *plausible* that this is a good way to estimate β in the face of correlated noise.

To go beyond plausibility, refer back to §3. At no point in our reasoning did we actually rely on $\mathrm{Var}\left[\epsilon|\mathbf{X}\right]$ being diagonal. It follows that if we set $\mathbf{W} = \mathrm{Var}\left[\epsilon|\mathbf{X}\right]^{-1}$, we get the linear, unbiased estimator of minimum variance. If we believe that the noise is Gaussian, then this is also the maximum likelihood estimator.

Where Do the Covariances Come From? In general we cannot estimate Σ. So we can only use this approach when Σ comes from external information. For example, the data may have some time series structure or spatial structure which suggests the form of Σ.

7 WLS versus Model Errors

When you find that your residuals from an initial model have non-constant variance or are correlated with each other, there are (at least) two possible explanations. One is that the fluctuations around the regression line really are heteroskedastic and/or correlated. In that case, you should try to model that variance and those correlations, and use WLS or GLS. The other explanation is that something is wrong with your model. If there's an important predictor variable which is just missing from your model, for example, then its contribution to the response will be part of your residuals. If that omitted variable is larger in some parts of the data than in others, or if the omitted variable has correlations, then that will make your residuals change in magnitude and be correlated. More subtly, having the wrong functional form for a variable you do include can produce those effects as well.