Some More Theory Underlying Simple Linear Regression

2018-09-11

One reason that OLS estimation is so useful is that, under a certain set of assumptions underlying the classical linear regression model, the estimators B_0 and B_1 have several desirable statistical properties. These properties include:

- The least squares estimators are *linear estimators*; they are linear functions of the observations. (This property helps us derive the sampling distributions for B_0 and B_1, which allows for statistical inference.)
- The least squares estimators are *unbiased estimators* of the population coefficients.
- The least squares estimators have sampling variances and a covariance.
- Of all the linear, unbiased estimators, the least squares estimators have the smallest sampling variance (most precise/efficient).
- Under the assumption of normality, the sampling distribution for the least squares estimators are also normally distributed; they are approximately normal under other conditions, especially with large sample sizes.
- Under the full set of assumptions, the least squares estimators are the maximum-likelihood estimators of the population coefficients.

The Classical Regression Model

We have previously defined the population regression model as

$$y_i = \beta_0 + \beta_1(x_i) + \epsilon_i.$$

where Y is assumed to be statistically and linearly related to X and ϵ. We also assume that the error term, ϵ, is a random variable.

Recall that the least squares estimators are:

$$B_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$

$$B_0 = \bar{y} - B_1(\bar{x})$$

Assumptions of the Classical Model

Let us now turn to assumptions governing the probability distribution of the errors, ϵ. Given that the model is correctly specified (**A.0**):

- **A.1:** ϵ_i has a mean of zero; $\mathbb{E}(\epsilon_i) = 0$ for all i.
- **A.2:** ϵ_i are homoscedastic; $\text{Var}(\epsilon_i)$ is constant (but unknown) for for all i

$$\begin{aligned}
\text{Var}(\epsilon_i) &= \mathbb{E}(\epsilon_i^2) - \big[\mathbb{E}(\epsilon_i)\big]^2 \\
&= \mathbb{E}(\epsilon_i^2) - 0 \\
&= \mathbb{E}(\epsilon_i^2) \\
&= \sigma^2_\epsilon
\end{aligned}$$

- **A.3:** Independence of errors; $\text{Cov}(\epsilon_i, \epsilon_j | x_i) = 0 \quad \text{for all } i \neq j$

For any value of X:

$$
\begin{aligned}
\text{Cov}(\epsilon_i, \epsilon_j) &= 0 \\
\mathbb{E}\Bigg[\bigg(\epsilon_i - \mathbb{E}[\epsilon_i]\bigg)\bigg(\epsilon_j - \mathbb{E}[\epsilon_j]\bigg) &= 0 \\
\mathbb{E}\Bigg[\bigg(\epsilon_i - 0\bigg)\bigg(\epsilon_j - 0\bigg) &= 0 \\
\mathbb{E}(\epsilon_i \epsilon_j) &= 0
\end{aligned}
$$

- **A.4:** X is a non-random variable (fixed) with finite, non-zero variance.

X fixed implies that the x_i values are controllable under repeated sampling from the same population. In most research in the social sciences, this is not the case. In those cases, we assume that X is measured without error and that X and ϵ are independent (uncorrelated).

- **A.5:** ϵ_i is normally distributed for all i.

If all five assumptions are met, we refer to this as the *strong classical regression model*. If only the first four assumptions are satisfied, we refer to it as the *weak classical regression model*.

Properties of Y

Since the X values, x_i, are fixed (**A.4**), the only source of variation in Y from repeated sampling is the variation in ϵ. Because of this, the probability distribution of Y is identical to the probability distribution for ϵ. Based on this, we can verify several properties of Y. (Note: For each of the following properties, we are assuming a fixed X so the expectation of Y is really the expectation of Y conditional on X.)

- **P.1:** $\mathbb{E}(Y) = \beta_0 + \beta_1(x_i)$

$$
\begin{aligned}
\mathbb{E}(Y) &= \mathbb{E}\Bigg(\beta_0 + \beta_1(x_i) + \epsilon_i\Bigg) \\
&= \mathbb{E}(\beta_0) + \mathbb{E}(\beta_1 x_i) + \mathbb{E}(\epsilon) \\
&= \beta_0 + \beta_1(x_i) + 0 \\
&= \beta_0 + \beta_1(x_i)
\end{aligned}
$$

- **P.2:** $\mathrm{Var}(Y) = \sigma_\epsilon^2$

$$
\begin{aligned}
\mathrm{Var}(Y) &= \mathbb{E}\Big(Y - \mathbb{E}(Y)\Big)^2 \\
&= \mathbb{E}\Big(\beta_0 + \beta_1(x_i) + \epsilon_i - \mathbb{E}\big(\beta_0 + \beta_1(x_i) + \epsilon_i\big)\Big)^2 \\
&= \mathbb{E}\Big(\beta_0 + \beta_1(x_i) + \epsilon_i - \mathbb{E}(\beta_0) - \mathbb{E}(\beta_1 x_i) - \mathbb{E}(\epsilon_i)\Big)^2 \\
&= \mathbb{E}\Big(\beta_0 + \beta_1(x_i) + \epsilon_i - \beta_0 - \beta_1 x_i - \mathbb{E}(\epsilon_i)\Big)^2 \\
&= \mathbb{E}\Big(\epsilon_i - \mathbb{E}(\epsilon_i)\Big)^2 \\
&= \mathrm{Var}(\epsilon_i) \\
&= \sigma_\epsilon^2
\end{aligned}
$$

- **P.3:** The observations, Y_i, are independent; $\mathrm{Cov}(Y_i, Y_j) = 0$ for $i \neq j$.

$$
\begin{aligned}
\mathrm{Cov}(Y_i, Y_j) &= \mathbb{E}\Big(Y_i - \mathbb{E}(Y_i)\Big)\Big(Y_j - \mathbb{E}(Y_j)\Big) \\
&= \mathbb{E}\Big(Y_i - (\beta_0 + \beta_1 x_i)\Big)\Big(Y_j - (\beta_0 + \beta_1 x_j)\Big) \\
&= \mathbb{E}(\epsilon_1)(\epsilon_j) \\
&= \mathrm{Cov}(\epsilon_i, \epsilon_j) \\
&= 0
\end{aligned}
$$

- **B.3:** The observations, Y_i, are normally distributed.

Since ϵ is normally distributed and $\beta_0 + \beta_1(x_i)$ is fixed, then Y_i is the sum of a random variable that is normally distributed and a constant $(\beta_0 + \beta_1(x_i) + \epsilon_i)$ which makes Y_i also normally distributed.

Properties of the OLS Estimators

By invoking some or all of these assumptions/properties, we can mathematically prove and derive certain other properties related to the OLS estimates.

Sampling Variance and Covariance of the Estimators

The sampling variances and covariance for the OLS estimators are:

$$\mathrm{Var}(B_1) = \frac{\sigma_e^2}{\sum(x_i - \bar{x})^2}$$

$$\mathrm{Var}(B_0) = \frac{\sigma_e^2 \sum x_i^2}{n\sum(x_i - \bar{x})^2}$$

$$\mathrm{Cov}(B_0, B_1) = \frac{\sigma_\epsilon^2 \bar{x}}{\sum(x_i - \bar{x})^2}$$

To derive these quantities, we will have to take advantage of the assumptions that the conditional variances are equal (**A.2**) and independence of errors/observations conditional on X (**A.3**).

Sampling Variance of the Estimators

We will start by deriving the sampling variance for B_1.

Since $B_1 = \sum w_i y_i$, then

$$\begin{aligned}\mathrm{Var}(B_1) &= \mathrm{Var}(\sum w_i y_i) \\ &= \sum \mathrm{Var}(w_i y_i) \\ &= \sum w_i^2 \mathrm{Var}(y_i)\end{aligned}$$

Now, recall that the variance of Y is conditional on X, then

$$\begin{aligned}\mathrm{Var}(y_i|x_i) &= \mathrm{Var}(\hat{y}_i + e_i \big| x_i) \\ &= \mathrm{Var}(\hat{y}_i \big| x_i) + \mathrm{Var}(e_i \big| x_i) + 2\mathrm{Cov}(\hat{y}_i, \epsilon_1 \big| x_i) \\ &= 0 + \mathrm{Var}(e_i \big| x_i) + 2(0) \\ &= \mathrm{Var}(e_i \big| x_i) = \sigma_\epsilon^2\end{aligned}$$

So,

$$
\begin{aligned}
\text{Var}(B_1) &= \sum w_i^2 \sigma_\epsilon^2 \\
&= \sigma_\epsilon^2 \sum w_i^2 \\
&= \sigma_\epsilon^2 \sum \left(\frac{x_i - \bar{x}}{\sum (x_i - \bar{x})^2} \right)^2 \\
&= \sigma_\epsilon^2 \frac{\sum (x_i - \bar{x})^2}{\sum (x_i - \bar{x})^2 \sum (x_i - \bar{x})^2} \\
&= \frac{\sigma_\epsilon^2}{\sum (x_i - \bar{x})^2}
\end{aligned}
$$

We can re-write this as

$$\text{Var}(B_1) = \frac{\sigma_\epsilon^2}{(n-1)s_x^2}$$

This helps us think about when the precision of B_1 will be high (low variance and SE):

- When the error variance, σ_ϵ^2, is small;
- When the sample size, n, is large; and
- When the variance in the predictor values, s_x^2, is large.

We can perform a similar derivation to obtain the sampling variance of B_0 (not shown). The formula for the sampling variance for B_0, also offers us insight for when the precision of B_0 will be high (low variance and SE):

$$\text{Var}(B_0) = \frac{\sigma_e^2 \sum x_i^2}{n \sum (x_i - \bar{x})^2}$$

Because the denominator is similar to that of $\text{Var}(B_1)$, we will have high precision around B_0 when:

- The error variance, σ_ϵ^2, is small;
- The sample size, n, is large; and
- The variance in the predictor values, s_x^2, is large.

But, because of the added sum term in the numerator, when $\bar{x} \approx 0$, then that will essentially cancel out the sum terms in the numerator and denominator, which will lead to higher precision, so also when

- The X-values are centered near 0.

Covariance between the Estimators

The covariance between B_0 and B_1 is defined as:

$$\text{Cov}(B_0, B_1) = \frac{\sigma_\epsilon^2 \bar{x}}{\sum(x_i - \bar{x})^2}$$

We can derive this using the covariance and expectations rules.

$$\text{Cov}(B_0, B_1) = \mathbb{E}\Big[\big(B_0 - \mathbb{E}[B_0]\big)\big(B_1 - \mathbb{E}[B_1]\big)\Big]$$

$$\text{Cov}(B_0, B_1) = \mathbb{E}\Big[(B_0 - \beta_0)(B_1 - \beta_1)\Big]$$

Now we use the result that $B_0 - \beta_0 = -\bar{x}(B_1 - \beta_1)$. You can show this by using the result that $\bar{y} = B_0 + B_1(\bar{x})$, which implies $B_0 = \bar{y} - B_1(\bar{x})$. Substituting $-\bar{x}(B_1 - \beta_1)$ in for $B_0 - \beta_0$, we get:

$$\begin{aligned}
\text{Cov}(B_0, B_1) &= \mathbb{E}\Big[-\bar{x}(B_1 - \beta_1)^2\Big] \\
&= -\bar{x} \times \mathbb{E}\Big[(B_1 - \beta_1)^2\Big] \\
&= -\bar{x} \times \text{Var}(B_1) \\
&= -\bar{x} \times \frac{\sigma_\epsilon^2}{\sum(x_i - \bar{x})^2}
\end{aligned}$$

This formula provides insight into the sampling errors of the regression estimators. Since both σ_ϵ^2 and $\sum(x_i - \bar{x})^2$ are values greater than zero, the covariance between B_0 and B_1 depends on the sign of $\bar{x}$.

- If $\bar{x} > 0$, then $\text{Cov}(B_0, B_1) < 0$. This implies that the sampling errors $(B_0 - \beta_0)$ and $(B_1 - \beta_1)$ have opposite signs.
- If $\bar{x} < 0$, then $\text{Cov}(B_0, B_1) > 0$. This implies that the sampling errors $(B_0 - \beta_0)$ and $(B_1 - \beta_1)$ have the same signs.

Sampling Distributions of the Estimators

To derive the sampling distributions for the estimators, which are the basis for statistical inference, we need to also take advantage of the normality assumption (**A.5**). Recall that we used the t-distribution with $n-2$ degrees of freedom to test hypotheses about the slope and intercept. The general form of a hypotheses test for a regression coefficient, is

$$H_0 : \beta_j = b \quad \text{where b is the tested value}$$

To test this, we create a studentized test statistic using

$$\frac{B_j - b}{\text{SE}(B_j)}$$

This statistic follows a t-distribution with $n-2$ degrees of freedom.

Recall from introductory statistics if we could assume that the population was normally distributed, then the distribution of $T = \frac{\bar{y}-\mu}{\mathrm{SE}(\bar{y})}$ was t-distributed with $n-1$ degrees of freedom. Since $\bar{y}$ is a linear combination of the observations, the distribution of $\bar{y}$ is also normally distributed, and estimating the SE of $\bar{y}$ introduced additional error; making the distribution of T follow a t-distribution.

This comes from a theorem which says that (1) if Z is a standard normal variable and W is chi-squared distributed with ν degrees of freedom, and (2) Z and W are independent, then

$$T = \frac{Z}{\sqrt{W/\nu}}$$

will have a t-distribution with ν degrees of freedom.

In the case of the test of the mean, we can write $T = \frac{\bar{y}-\mu}{\mathrm{SE}(\bar{y})}$ in this form as:

$$T = \frac{\sqrt{n}(\bar{y}-\mu)/\sigma_y}{\sqrt{\left[(n-1)s_y^2/\sigma_y^2\right]/(n-1)}}$$

Thus, the distribution of T will be t-distributed with $n-1$ degrees of freedom.

For the regression estimators, under the assumption of normality (Assumption #5), the least squares estimators are also normally distributed. This is true since B_j is a linear combination of the observations, and we are assuming the observations to be normally distributed (linear shifts do not change the distribution). Thus, we have,

$$\begin{aligned} T &= \frac{B_j - b}{\mathrm{SE}(B_j)} \\ &= \frac{\frac{B_j-b}{\sigma_{B_j}}}{\frac{\mathrm{SE}(B_j)}{\sigma_{B_j}}} \end{aligned}$$

From this it is clear that the numerator of T is a standard normal variable. The denominator is

$$\begin{aligned} \frac{\mathrm{SE}(B_j)}{\sigma_{B_j}} &= \sqrt{\frac{\mathrm{Var}(B_j)}{\sigma_{B_j}^2}} \\ &= \sqrt{\frac{\frac{\mathrm{MSE}}{\sum(x_i-\bar{x})^2}}{\frac{\sigma_\epsilon^2}{\sum(x_i-\bar{x})^2}}} \\ &= \sqrt{\frac{\mathrm{MSE}}{\sigma_\epsilon^2}} \\ &= \sqrt{\frac{\frac{\mathrm{SSE}}{n-2}}{\sigma_\epsilon^2}} \\ &= \sqrt{\frac{\mathrm{SSE}}{\sigma_\epsilon^2(n-2)}} \end{aligned}$$

At this point we rely on a common theorem from regression theory which says that $\frac{\text{SSE}}{\sigma^2_\epsilon}$ is distributed as χ^2 with $n-2$ degrees of freedom and is independent of both B_0 and B_1. Relying on this,

$$T = \frac{B_j - b}{\text{SE}(B_j)} = \frac{z}{\sqrt{\frac{\chi^2(n-2)}{(n-2)}}}$$

Since z is a function of B_0 and B_1, then z and χ^2 are also independent, and it follows that $\frac{B_j - b}{\text{SE}(B_j)}$ is t-distributed with $n-2$ degrees of freedom.

Gauss–Markov Theorem

The Gauss–Markov Theorem is a powerful theorem that states that under the weak classical model (**A.1–A.4**), the least squares estimators have certain desirable properties. Both estimators are:

- Linear functions of the observations, Y_i;
- Unbiased estimators of the population coefficients;
- The most efficient (smallest sampling variance) unbiased linear estimators of the population coefficients.

Because of this theorem, we typically refer to the OLS coefficents as BLUE (Best Linear Unbiased Estimators).

Fox (2016) reminds us that the "best" in BLUE means that they have the smallest sampling variance of all the possible linear unbiased estimators. There may be a biased or non-linear estimator that produces a smaller sampling variance than the OLS estimator. It is also worth noting that if we also invoke the normality assumption (**A.5**), then the OLS estimators become "best" among all unbiased estimators (both linear and non-linear).

Proving this theorem is beyond the scope of the class, but an outline for this proof would entail:

- Show that B_0 and B_1 are linear functions of the observations; we can express each estimator as $\sum w_i y_i$ for some w_i.
- Show that B_0 and B_1 are unbiased; that $\mathbb{E}(B_0) = \beta_0$ and $\mathbb{E}(B_1) = \beta_1$
- Show that for any other unbiased linear estimator, say L_0 and L_1, that $\text{Var}(B_0) < \text{Var}(L_0)$ and $\text{Var}(B_1) < \text{Var}(L_1)$.

OLS Estimators are Maximum Likelihood Estimators

Given the assumptions of the strong classical model (**A.1–A.5**), we can show that the least squares estimators are also the maximum likelihood estimators. Recall that the likelihood is the probability of a set of parameters, computed as the joint density of the data, given a set of observations under a particular probability distribution.

The joint density of the errors is:

$$\prod_{i=1}^{n} f(\epsilon_i;\ 0, \sigma^2_\epsilon) = \left(2\pi\sigma^2_\epsilon\right)^{-\,n/2} \times e^{-\frac{1}{2\sigma^2_\epsilon}\sum \epsilon_i^2}$$

Using properties **P.3** (independence of Ys) and **P.4** (normality of Ys), and that ϵ_i is a linear function of y_i, we can write the likelihood of the parameters given the observations and the normal probability distribution of Y as,

$$\mathcal{L}\bigg(\beta_0, \beta_1, \sigma^2_\epsilon \mid Y, n\bigg) = \left(2\pi\sigma^2_\epsilon\right)^{-\,n/2} \times e^{-\frac{1}{2\sigma^2_\epsilon}\sum \left(Y_i - \beta_0 - \beta_1 x_i\right)^2}$$

Or, in log-likelihood form,

$$\log \mathcal{L}\bigg(\beta_0, \beta_1, \sigma_\epsilon^2 \mid Y, n\bigg) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma_\epsilon^2) - \frac{1}{2\sigma_\epsilon^2}\sum\big(Y_i - \beta_0 - \beta_1 x_i\big)^2$$

If we differentiate this expression with respect to each of the parameters, β_0, β_1, and σ_ϵ^2, we get

$$\begin{aligned}
\frac{\partial \log \mathcal{L}}{\partial \beta_0} &= \frac{1}{\sigma_\epsilon^2}\sum\big(Y_i - \beta_0 - \beta_1 x_i\big) \\
\frac{\partial \log \mathcal{L}}{\partial \beta_1} &= \frac{1}{\sigma_\epsilon^2}\sum x_i\big(Y_i - \beta_0 - \beta_1 x_i\big) \\
\frac{\partial \log \mathcal{L}}{\partial \sigma_\epsilon^2} &= -\frac{n}{2\sigma_\epsilon^2} + \frac{1}{2\sigma_\epsilon^4}\sum\big(Y_i - \beta_0 - \beta_1 x_i\big)^2
\end{aligned}$$

After setting each of these equal to zero and solving, we find that

$$\sum y_i = nB_0 + B_1 \sum(x_i)$$

$$\sum x_i y_i = B_0 \sum x_i + B_1 \sum x_i^2$$

Hence the maximum likelihood estimators for the two coefficients are equivalent to the OLS estimators of these parameters. We also find that,

$$\begin{aligned}
\hat{\sigma}_\epsilon^2 &= \frac{1}{n}\sum\big(Y_i - B_0 - B_1 x_i\big)^2 \\
&= \frac{1}{n}\sum\big(e_i\big)^2
\end{aligned}$$

Thus the maximum likelihood estimate for the error variance is not the same as the OLS estimate for error variance (the OLS version divides by $n - 2$).

The maximum likelihood estimators for the coefficients (which are the same as the OLS estimators) possess several asymptotic (large sample) properties; most importantly normality. Because of this, with a large sample size, the sampling distribution for the estimates will be approximately t-distributed with $n - 2$ degrees of freedom.

Implications for Applied Researchers

If the assumptions underlying the strong classical regression model (**A.1**–**A.5**) are all valid, then the OLS estimators B_0 and B_1 are good estimators of β_0 and β_1. They are unbiased and efficient and have accurate sampling variances and covariance.

Of course, any of the assumptions may be challenged either on a priori substantive grounds, or post hoc, via empirical examination of the residuals. If one (or more) of the assumptions are violated, then some of the properties may be compromised. If this is the case, one can often transform the data in some way or use an alternative estimation technique.

Violation of the normality assumption causes the least number of problems. Under non-normality, B_0 and B_1 are still BLUE, and s_e^2 is still an unbiased estimator of σ_ϵ^2. However, the under non-normality, the use of the F- and t-distributions for inference is questionable especially if the sample size is small. If the sample size is large, the sampling distributions of B_0 and B_1 are approximately normal and subsequently the use of the F- and t-distributions for inference is justified.

If Assumptions **A.1–A.4** are violated, then B_0 and B_1 are no longer BLUE. Moreover, the formulas used to compute the sampling variances will also be incorrect (which also affects inference).

Assumption A.1 Not Valid

Assumption **A.1** was that ϵ_i has a mean of zero; $\mathbb{E}(\epsilon_i) = 0$ for all i. If we let $\epsilon_i = h + \tilde{\epsilon}$ where h is a non-zero constant and $\tilde{\epsilon}$ is a random error term that obeys the properties of the weak classical regression model then $\mathbb{E}(\epsilon_i) \neq 0$, since

$$\begin{aligned} \mathbb{E}(\epsilon_i) &= \mathbb{E}(h + \tilde{\epsilon}) \\ &= \mathbb{E}(h) + \mathbb{E}(\tilde{\epsilon}) \\ &= h \ \ (\neq 0) \end{aligned}$$

The regression model becomes,

$$\begin{aligned} y_i &= \beta_0 + \beta_1(x_i) + \epsilon \\ y_i &= \beta_0 + \beta_1(x_i) + h + \tilde{\epsilon} \\ y_i &= \bigg(\beta_0 + h\bigg) + \beta_1(x_i) + \tilde{\epsilon} \end{aligned}$$

Using B_0 to estimate β_0 results in a biased estimate since

$$\begin{aligned} \mathbb{E}\bigg(\beta_0 + h\bigg) &= \mathbb{E}(\beta_0) + \mathbb{E}(h) \\ &= \beta_0 + h \ \ (\neq \beta_0) \end{aligned}$$

Thus the estimator B_0 is biased when this assumption is volated. However, the estimator B_1 is still BLUE. Unfortunately, we can never test this assumption in practice since $\sum e_i = 0$ always holds when using OLS.

Heteroscedasticity

Assumption **A.2** was that ϵ_i are homoscedastic; $\text{Var}(\epsilon_i)$ is constant (but unknown) for for all i. If the variance is not constant, then we have,

$$\text{Var}(\epsilon_i) = \sigma_i^2$$

Here the error variance varies from one observation to another (note the i subscript). We can show (but do not) as part of the Gauss–Markov Theorem that

$$B_1 = \beta_1 + \sum w_i \epsilon_i$$

$$B_0 = \beta_0 + \Big(\beta_1 - B_1\Big)\bar{x} + \bar{\epsilon}$$

and subsequently that $\mathbb{E}(B_0) = \beta_0$ and $\mathbb{E}(B_1) = \beta_1$ (take the expectations of both sides). Since in both equations, heteroscedasticity would affect only the last term, and would not change the expectations of those terms, even with non-constant variance, the estimators are still unbiased.

If we assume a heteroskedastic variance,

$$y_i = \beta_0 + \beta_1(x_i) + \epsilon \qquad \text{where } \mathrm{Var}(\epsilon_i) = \sigma_i^2$$

We can transform these heteroskedastic variances into constant variances by weighting them by $^1/_{\sigma_i}$. Doing so we obtain,

$$\frac{y_i}{\sigma_i} = \beta_0\left(\frac{1}{\sigma_i}\right) + \beta_1\left(\frac{x_i}{\sigma_i}\right) + \frac{\epsilon}{\sigma_i} \qquad \text{where } \sigma_i \neq 0$$

We can express this as

$$y_i^* = \beta_0 V_i^* + \beta_1 x_i^* + \epsilon_i^*$$

where $y_i^* = \frac{y_i}{\sigma_i}$, $V_i^* = \sigma_i^{-1}$, and $\epsilon_i^* = \frac{\epsilon}{\sigma_i}$. If we hold the values for x_i^* (and subsequently V_i^*) fixed for all i, then

$$\mathbb{E}(\epsilon_i^*) = \frac{\mathbb{E}(\epsilon_i)}{\sigma_i} = 0$$

$$\mathrm{Var}(\epsilon_i^*) = \frac{\mathrm{Var}(\epsilon_i)}{\sigma_i^2} = 1$$

$$\mathrm{Cov}(\epsilon_i^*, \epsilon_j^*) = \frac{\mathbb{E}(\epsilon_i, \epsilon_j)}{\sigma_i \sigma_j} = 0$$

Thus, *under the weighted transformations*, all of the assumptions of the weak classical regression model still hold and we can apply the Gauss–Markov Theorem using OLS on the transformed model. The transformation leads to the derivation of *weighted least squares* estimators which are BLUE.

Non-Independence

Assumption **A.3** is independence of errors; $\mathrm{Cov}(\epsilon_i, \epsilon_j | x_i) = 0 \quad$ for all $i \neq j$. Under this violation, the estimators are still unbiased, but no longer have the smallest sampling variances. Moreover, the variance estimates we get using OLS estimation tend to be too small when the independnece assumption is violated; leading to p-values that are too small (higher chance of making a type I error).

References

Fox, J. (2016). *Applied regression analysis & generalized linear models* (3rd ed.). Thousand Oaks, CA: Sage.