The Simple Linear Regression Model

MATH 60604A: Statistical Modelling

This document provides supplementary information for the simple linear regression model (chapter 2 part 1).

The Model

The simple linear regression model is defined as follows

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

with $\epsilon_1, \ldots, \epsilon_n \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$. If the model is correctly specified, that is, if Y_i is indeed equal to $\beta_0 + \beta_1 X_i + \epsilon_i$, then $Y_1, \ldots, Y_n$ are independent and Normally distributed with mean and variance

$$\begin{aligned}
\mathrm{E}(Y_i|X_i) &= \mathrm{E}(\beta_0 + \beta_1 X_i + \epsilon_i|X_i) \\
&= \beta_0 + \beta_1 X_i + \mathrm{E}(\epsilon_i|X_i) \\
&= \beta_0 + \beta_1 X_i \\
\mathrm{var}(Y_i|X_i) &= \mathrm{var}(\beta_0 + \beta_1 X_i + \epsilon_i|X_i) \\
&= \mathrm{var}(\epsilon_i|X_i) \\
&= \sigma^2
\end{aligned}$$

That is, for a given value of X_i, Y_i is normally distributed with mean $\mathrm{E}(Y_i|X_i) = \beta_0 + \beta_1 X_i$ and constant variance σ^2.

To estimate the regression parameters, β_0 and β_1, we use the least squares criterion, that is, we find estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimize the sum of the squared errors

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2$$

Thus, $\hat{\beta}_0$ and $\hat{\beta}_1$ must satisfy

$$\frac{\partial}{\partial \beta_0} S(\beta_0, \beta_1) = -2 \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i) = 0$$
$$\frac{\partial}{\partial \beta_1} S(\beta_0, \beta_1) = -2 \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i) X_i = 0$$

which leads to estimators

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$
$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (X_i - \bar{X})^2} = \frac{S_{XY}}{S_{XX}}$$

Notice that both $\hat{\beta}_0$ and $\hat{\beta}_1$ are linear combinations of the Y_i:

$$\begin{aligned}
\hat{\beta}_1 = \frac{S_{XY}}{S_{XX}} &= \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{S_{XX}} \\
&= \frac{1}{S_{XX}} \left\{ \sum_{i=1}^{n} (X_i - \bar{X}) Y_i - \sum_{i=1}^{n} (X_i - \bar{X}) \bar{Y} \right\} \\
&= \frac{1}{S_{XX}} \left\{ \sum_{i=1}^{n} (X_i - \bar{X}) Y_i - \bar{Y} \sum_{i=1}^{n} (X_i - \bar{X}) \right\} \\
&= \frac{1}{S_{XX}} \sum_{i=1}^{n} (X_i - \bar{X}) Y_i \\
&= \sum_{i=1}^{n} C_i Y_i
\end{aligned}$$

where $C_i = \frac{X_i - \bar{X}}{S_{XX}}$, and

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} = \bar{Y} - \bar{X} \sum_{i=1}^{n} C_i Y_i$$

Since the Y_i are assumed to be normally distributed, then both $\hat{\beta}_0$ and $\hat{\beta}_1$ are also normally distributed as these estimators are both linear combinations of the Y_i.

We can find the mean and variance of $\hat{\beta}_0$ and $\hat{\beta}_1$ (treating X_i as fixed). For the mean,

$$\begin{aligned}
\mathrm{E}(\hat{\beta}_1) &= \mathrm{E}\left(\sum_{i=1}^n C_i Y_i\right) = \sum_{i=1}^n C_i \mathrm{E}(Y_i) \\
&= \sum_{i=1}^n C_i(\beta_0 + \beta_1 X_i) = \beta_0 \sum_{i=1}^n C_i + \beta_1 \sum_{i=1}^n C_i X_i \\
&= \beta_0 \times 0 + \beta_1 \times 1 \\
&= \beta_1 \\
\mathrm{E}(\hat{\beta}_0) &= \mathrm{E}\left(\bar{Y} - \bar{X}\sum_{i=1}^n C_i Y_i\right) = \mathrm{E}\left(\frac{1}{n}\sum_{i=1}^n Y_i - \bar{X}\sum_{i=1}^n C_i Y_i\right) \\
&= \frac{1}{n}\sum_{i=1}^n \mathrm{E}(Y_i) - \bar{X}\sum_{i=1}^n C_i \mathrm{E}(Y_i) = \frac{1}{n}\sum_{i=1}^n (\beta_0 + \beta_1 X_i) - \bar{X}\sum_{i=1}^n C_i(\beta_0 + \beta_1 X_i) \\
&= \beta_0 + \beta_1 \bar{X} - \bar{X}\beta_0 \sum_{i=1}^n C_i - \bar{X}\beta_1 \sum_{i=1}^n C_i X_i \\
&= \beta_0 + \beta_1 \bar{X} - \bar{X}\beta_0 \times 0 - \beta_1 \bar{X} \times 1 \\
&= \beta_0
\end{aligned}$$

Note that the above relies on $\sum_{i=1}^n C_i = 0$ *and* $\sum_{i=1}^n C_i X_i = 1$*, which can be showed as follows:*

$$\begin{aligned}
\sum_{i=1}^n C_i &= \sum_{i=1}^n \frac{(X_i - \bar{X})}{S_{XX}} = \frac{1}{S_{XX}}\left(\sum_{i=1}^n X_i - \sum_{i=1}^n \bar{X}\right) = \frac{1}{S_{XX}}\left(n\bar{X} - n\bar{X}\right) \\
&= 0 \\
\sum_{i=1}^n C_i X_i &= \sum_{i=1}^n \frac{(X_i - \bar{X})X_i}{S_{XX}} = \frac{1}{S_{XX}}\left(\sum_{i=1}^n (X_i - \bar{X})(X_i - \bar{X} + \bar{X})\right) \\
&= \frac{1}{S_{XX}}\left(\sum_{i=1}^n (X_i - \bar{X})^2 + \bar{X}\sum_{i=1}^n (X_i - \bar{X})\right) \\
&= \frac{1}{S_{XX}}\left(\sum_{i=1}^n (X_i - \bar{X})^2 + \bar{X} \times 0\right) = \frac{1}{S_{XX}}\left(\sum_{i=1}^n (X_i - \bar{X})^2\right) \\
&= 1
\end{aligned}$$

For the variance, we rely on the fact that the Y_i are, by assumption, independent (i.e., $\mathrm{cov}(Y_i, Y_j) =$

$0 \ \forall i \neq j$) and have constant variance σ^2 so that

$$
\begin{aligned}
\text{var}(\hat{\beta}_1) &= \text{var}\left(\sum_{i=1}^{n} C_i Y_i\right) = \sum_{i=1}^{n} C_i^2 \text{var}(Y_i) = \sum_{i=1}^{n} C_i^2 \sigma^2 \\
&= \sigma^2 \sum_{i=1}^{n} \left(\frac{(X_i - \bar{X})}{S_{XX}}\right)^2 = \frac{\sigma^2}{S_{XX}^2} \sum_{i=1}^{n} (X_i - \bar{X})^2 = \frac{\sigma^2}{S_{XX}^2} S_{XX} \\
&= \frac{\sigma^2}{S_{XX}} \\
\text{var}(\hat{\beta}_0) &= \text{var}(\bar{Y} - \hat{\beta}_1 \bar{X}) = \text{var}(\bar{Y}) + \bar{X}^2 \text{var}(\hat{\beta}_1) - 2\bar{X} \text{cov}(\bar{Y}, \hat{\beta}_1) \\
&= \frac{\sigma^2}{n} + \bar{X}^2 \frac{\sigma^2}{S_{XX}} - 0 \\
&= \sigma^2 \left(\frac{1}{n} + \frac{\bar{X}^2}{S_{XX}}\right)
\end{aligned}
$$

(Note that the above relies on showing $\text{cov}(\bar{Y}, \hat{\beta}_1) = 0$, *which can be shown to follow from the independence of the random error terms* ϵ_i*)*

Thus, we see that the assumption that $\epsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$ implies that (treating X_i as fixed)

$$
\begin{aligned}
&\Rightarrow Y_i \overset{iid}{\sim} \mathcal{N}(\beta_0 + \beta_1 X_i, \sigma^2) \quad \text{or, equivalently,} \quad \frac{Y_i - \beta_0 - \beta_1 X_i}{\sigma} \sim \mathcal{N}(0,1) \\
&\Rightarrow \hat{\beta}_1 \sim \mathcal{N}\left(\beta_1, \frac{\sigma^2}{S_{XX}}\right) \quad \text{or, equivalently,} \quad \frac{\hat{\beta}_1 - \beta_1}{\sigma/\sqrt{S_{XX}}} \sim \mathcal{N}(0,1) \\
&\Rightarrow \hat{\beta}_0 \sim \mathcal{N}\left(\beta_0, \sigma^2\left(\frac{1}{n} + \frac{\bar{X}^2}{S_{XX}}\right)\right) \quad \text{or, equivalently,} \quad \frac{\hat{\beta}_0 - \beta_0}{\sigma\sqrt{\frac{1}{n} + \frac{\bar{X}^2}{S_{XX}}}} \sim \mathcal{N}(0,1)
\end{aligned}
$$

This allows us to carry out hypothesis tests involving the regression parameters β_0, β_1.

Focusing on β_1 (although similar results can be shown for β_0 as well), suppose we're interested in testing

$$H_0 : \beta_1 = 0 \quad \text{vs.} \quad H_1 : \beta_1 \neq 0$$

Then, under H_0, that is, if β_1 is truly equal to 0, we have that

$$\frac{\hat{\beta}_1 - 0}{\sigma/\sqrt{S_{XX}}} \sim \mathcal{N}(0,1)$$

We could use the above as our test statistic, however, the difficulty is that we usually do not know the true value of the variance parameter σ^2. An unbiased (and intuitive) estimator for σ^2 is given by:

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \frac{1}{n-2}\sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2$$

Notice the similarity between $\hat{\sigma}^2$ here and the classical sample variance $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$. The idea is that we can use the estimated regression model, $\hat{\beta}_0 + \hat{\beta}_1 X_i$, to estimate the mean $\mathrm{E}(Y_i|X_i)$, and then to have an unbiased estimator we divide by $n-2$ since we used two parameters (β_0 and β_1) to model the mean.

We can derive the distribution of this estimator and show that

$$\left(\frac{n-2}{\sigma^2}\right)\hat{\sigma}^2 \sim \chi^2_{n-2}$$

that is, the variance estimator $\hat{\sigma}^2$ scaled by a factor $\frac{n-2}{\sigma^2}$ follows a χ^2 distribution with $n-2$ degrees of freedom. It can also be shown that $\hat{\sigma}^2$ is independent of $\hat{\beta}_1$. Thus, we can show that under $H_0 : \beta_1 = 0$, the test statistic

$$t = \frac{\hat{\beta}_1 - 0}{\hat{\sigma}/\sqrt{S_{XX}}} = \frac{\hat{\beta}_1}{\hat{se}(\hat{\beta}_1)} \sim t_{n-2}$$

that is, the test statistic follows a Student t distribution with $n-2$ degrees of freedom.

Note: the Student t distribution can be derived by working with the Normal distribution and the χ^2 distribution. Suppose Z follows a standard normal distribution, $Z \sim \mathcal{N}(0, 1)$, and V follows a χ^2 distribution with ν degrees of freedom, $V \sim \chi^2_\nu$, and supposed Z and V are independent. Then the ratio $\frac{Z}{\sqrt{V/\nu}}$ follows a Student t distribution with ν degrees of freedom.

We can also compute confidence intervals from this:

$$\hat{\beta}_1 \pm t_{n-2,\alpha/2}\hat{se}(\hat{\beta}_1)$$

Note that similar results extend to the multiple linear regression model where there are several explanatory variables $X_1, \ldots, X_n$.

Model Assumptions

The simple linear regression model $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$ relies on the assumption that $\epsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$, that is

1. the error terms $\epsilon_1, \ldots, \epsilon_n$ are **independent** random variables

    or, equivalently, $Y_1, \ldots, Y_n$ are independent random variables

2. the error terms have **mean zero** $\mathrm{E}(\epsilon_i) = 0$

    or, equivalently, $\mathrm{E}(Y_i|X_i) = \beta_0 + \beta_1 X_i$

3. the error terms have **constant variance** $\mathrm{var}(\epsilon_i) = \sigma^2$ (**homoscedasticity**)

    or, equivalently, $\mathrm{var}(Y_i|X_i) = \sigma^2$

4. the error terms ϵ_i follow a **normal** distribution

    or, equivalently, Y_i follow a normal distribution

Each assumption is key in establishing certain results for the linear regression model. It is important to verify these model assumptions as deviations from these assumptions can cause the analysis to no longer be valid and ultimately lead to incorrect conclusions.

1. Independent errors

Independence plays an important role in determining the variance of the estimators $\hat{\beta}_0$ and $\hat{\beta}_1$. If observations are dependent, then the variance $se(\hat{\beta}_j)$ and the corresponding estimator $\hat{se}(\hat{\beta}_j)$, for $j = 0, 1$, are incorrect and thus conclusions from hypothesis tests and C.I.s could be incorrect. (Note that generally, violations of the independence assumption will lead to an underestimation of the variance, although this is not necessarily always the case.)

How to fix this: consider a linear mixed model which allows for correlated responses (or a time series model is the data is of that form).

2. Mean specification

The model assumes that $\mathrm{E}(\epsilon_i) = 0$ and thus that $\mathrm{E}(Y_i|X_i) = \beta_0 + \beta_1 X_i$. This assumption implies that the mean model is correctly specified, i.e. the effect of X_i on Y_i is indeed linear and, moreover, all important explanatory variables that affect the mean of Y_i have been included in the model. If this assumption is not met, it could lead to biased estimators.

For example, suppose that the relation between Y_i and X_i is in fact quadratic such that $\mathrm{E}(Y_i|X_i) = \beta_0 + \beta_1 X_i + \beta_2 X_i^2$. If we use a simple linear regression model of the form $\mathrm{E}(Y_i|X_i) = \alpha_0 + \alpha_1 X_i$, it is clear that the model is not appropriate and the parameter α_1 does not adequately measure the effect of X_i on Y_i since the relationship is not linear.

Consider another example: suppose that the true model is given by $E(Y_i|X_i, Z_i) = \beta_0+\beta_1 X_i+\beta_2 Z_i$ so that the mean of Y_i is influenced by both the explanatory variable X_i and Z_i. Further suppose that $E(Z_i|X_i) = a + bX_i$ so that X_i and Z_i are themselves related. Then, if only X_i is included as an explanatory variable in the model, we have that

$$\begin{aligned}
\mathrm{E}(Y_i|X_i) &= E(\beta_0 + \beta_1 X_i + \beta_2 Z_i \mid X_i) \\
&= \beta_0 + \beta_1 X_i + \beta_2 \mathrm{E}(Z_i|X_i) \\
&= \beta_0 + \beta_1 X_i + \beta_2 (a + bX_i) \\
&= (\beta_0 + a\beta_2) + (\beta_1 + b\beta_2)X_i \\
&= \beta_0^* + \beta_1^* X_i
\end{aligned}$$

That is, the misspecified model that only includes X_i (and omits Z_i) is actually estimating β_1^* (which reflects the effect of both X_i and Z_i) rather than β_1 (which reflects only the effect of X_i). Thus the estimator will be biased as the model is really estimating β_1^* rather than β_1. The exception to this is if $b = 0$, that is, if X_i and Z_i are unrelated. In this case, the misspecified model will still yield an unbiased estimator for β_1. Note that Z here is a *confounder* - we will revisit this concept. This is problematic when dealing with observational data.

Note that in general, if the mean model is misspecified such that there are too many variables in the model (thus some variables are "useless"), the estimators of the regression parameters are usually unbiased, but the variances of the estimators will be larger as we are estimating extra unnecessary parameters.

How to fix this: reformulate the mean specification to properly specify the mean of Y, i.e., include all relevant explanatory variables as well as any necessary transformations (e.g. a quadratic term X^2, etc.).

3. Constant Variance (Homoscedasticity)

The constant variance assumption is used when deriving the variance of the estimators. If this assumption is not met, the estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ will still be unbiased, however, the estimated standard deviations $\hat{se}(\hat{\beta}_0)$ and $\hat{se}(\hat{\beta}_1)$ will no longer be valid. As a result, C.I.'s and hypothesis tests will be not be valid and may lead to incorrect conclusions.

How to fix this: there are different approaches that can be taken. Either consider a different model (perhaps within the GLM framework), or consider a *variance stabilizing transformation* of the response variable.

Often, when the assumption of constant variance is violated, it is because the true underlying distribution is not normal but rather has a form such that the variance $\mathrm{var}(Y|X)$ is somehow related to $\mathrm{E}(Y|X)$ (i.e. is a function of X). For example, for a Poisson distribution, it can be shown that the mean and variance are equal. In that case, one can chose to use a different model or use a transformation. Variance-stabilizing transformations consider a *transformation* of the response variable such that the variance no longer depends on the mean. That is, rather than carrying out linear regression on Y, a transformation Y^* is chosen such that $\mathrm{var}(Y^*|X)$ does not depend on $\mathrm{E}(Y^*|X)$ and the linear regression is carried out on Y^*. The form of the transformation depends on the type of data and can be chosen empirically (i.e. based on the observed data). For example, for Poisson observations where $\mathrm{E}(Y) = \mathrm{var}(Y)$ an appropriate variance-stabilizing transformation is $Y^* = \sqrt{Y}$.

4. Normality

The assumption of normality allows us to establish the distribution of the test statistic $t = \frac{\hat{\beta}_j}{\hat{se}(\hat{\beta}_j)}$, $j = 0, 1$ as being that of Student t. This assumption, however, is mostly important when the sample size is small. For large sample sizes, we can rely on the central limit theorem to construct C.I.'s and carry out hypothesis tests.

How to fix this: if the sample size is large, asymptotic theory ensures that the results are still (asymptotically) valid. If the sample size is small, a different model should be used that is appropriate for non-normal observations.

References

[1] Lumley, T., Diehr, P., Emerson, S. & Chen, L. The Importance of the Normality Assumption in Large Public Health Data Sets. *Annual Review of Public Health* 23:1, 151-169 , 2002.

[2] Montgomery D. C., Peck, E. A., & Vining, G. G., *Introduction to Linear Regression Analysis.* Wiley, New York, 2012.

[3] Wakefield, J., *Bayesian and Frequentist Regression methods.* Springer, New York, 2013.