1

The Logic of Ordinary Least Squares Estimation

This chapter begins the discussion of ordinary least squares (OLS) regression. OLS is the "workhorse" of empirical social science and is a critical tool in hypothesis testing and theory building. This chapter builds on the the discussion in Chapter 6, by showing how OLS regression is used to estimate relationships between and among variables.

1.1 Theoretical Models

Models, as discussed earlier, are an essential component in theory building. They simplify theoretical concepts, provide a precise way to evaluate relationships between variables, and serve as a vehicle for hypothesis testing. As discussed in Chapter 1, one of the central features of a theoretical model is the presumption of causality, and causality is based on three factors; time ordering (observational or theoretical), co-variation, and non-spuriousness. Of these three assumptions, co-variation is the one analyzed using OLS. The often repeated adage, "correlation is not causation" is key. Causation is driven by theory, but co-variation is the critical part of empirical hypothesis testing.

When describing relationships, it is important to distinguish between those that are `deterministic` versus `stochastic`. Deterministic relationships are "fully determined" such that, knowing the values of the independent variable, you can perfectly explain (or predict) the value of the dependent variable. Philosophers of Old (like Kant) imagined the universe to be like a massive and complex clock which, once wound up and set ticking, would permit perfect prediction of the future if you had all the information on the starting conditions. There is no "error" in the prediction. Stochastic relationships, on the other hand, include an irreducible random component, such that the independent variables permit only a partial prediction of the dependent variable. But that stochastic (or random) component of the variation in the dependent variable has a probability distribution that can be analyzed statistically.

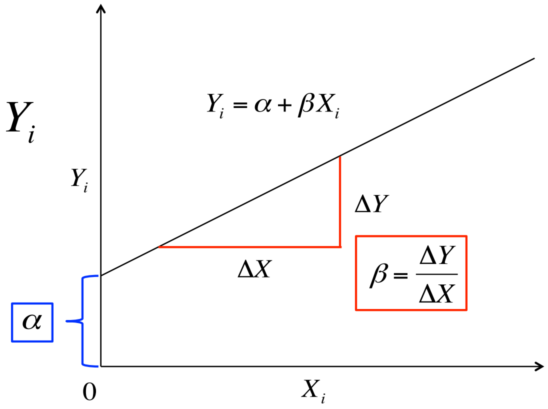

Y_i
Y_i
$Y_i = \alpha + \beta X_i$
ΔY
ΔX
$\beta = \frac{\Delta Y}{\Delta X}$
α
0
X_i

Figure 1.1: Deterministic Model

1.1.1 Deterministic Linear Model

The deterministic linear model serves as the basis for evaluating theoretical models. It is expressed as:

$$Y_i = \alpha + \beta X_i \tag{1.1}$$

A deterministic model is **systematic** and contains no error, therefore Y *is perfectly predicted by* X. This is illustrated in Figure 1.1. α and β are the model parameters, and are constant terms. β is the slope; the change in Y over the change in X. α is the intercept; the value of Y when X is zero.

Given that in social science we rarely work with deterministic models, nearly all models contain a stochastic, or random, component.

1.1.2 Stochastic Linear Model

The stochastic, or statistical, linear model contains a systematic component, $Y = \alpha + \beta$, and a stochastic component called the **error term**. The error term is the difference between the expected value of Y_i and the observed value of Y_i; $Y_i - \mu$. This model is expressed as:

$$Y_i = \alpha + \beta X_i + \epsilon_i \tag{1.2}$$

where ϵ_i is the error term. In the deterministic model, each value of Y fits along the regression line, however in a stochastic model the expected value of Y is conditioned by the values of X. This is illustrated in Figure 1.2.

Figure 1.2 shows the conditional population distributions of Y for several values of $X, p(Y|X)$. The conditional means of Y given X are denoted μ.

$$\mu_i \equiv E(Y_i) \equiv E(Y|X_i) = \alpha + \beta X_i \tag{1.3}$$

where

- $\alpha = E(Y) \equiv \mu$ when $X = 0$

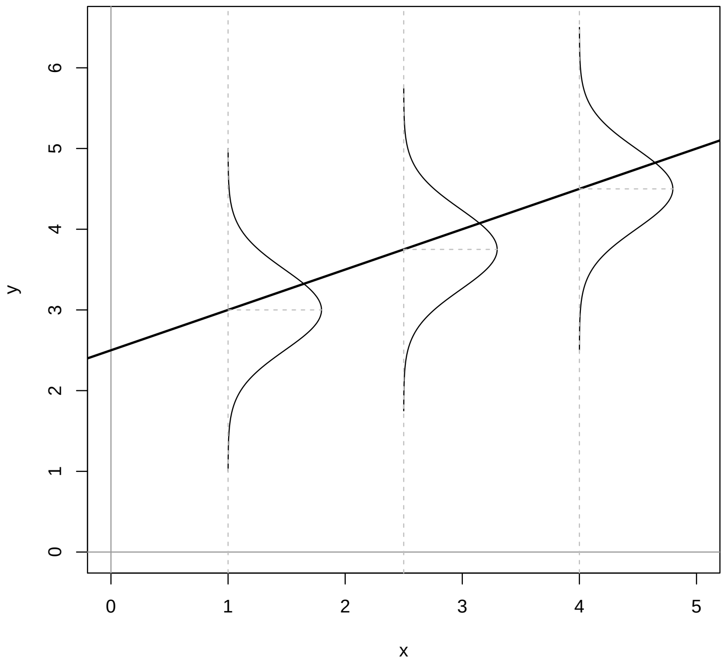


Figure 1.2: Stochastic Linear Model

- Each 1 unit increase in X increases $E(Y)$ by β

However, in the stochastic linear model variation in Y is caused by more than X, it is also caused by the error term ϵ. The error term is expressed as:

$$\begin{aligned}\epsilon_i &= Y_i - E(Y_i)\\ &= Y_i - (\alpha + \beta X_i)\\ &= Y_i - \alpha - \beta X_i\end{aligned}$$

Therefore;

$$\begin{aligned}Y_i &= E(Y_i) + \epsilon\\ &= \alpha + \beta X_i + \epsilon_i\end{aligned}$$

We make several important assumptions about the error term that are discussed in the next section.

Assumptions about the Error Term

There are three key assumptions about the error term; a) errors have identical distributions, b) errors are independent, c) errors are normally distributed.[1]

[1] Actually, we assume only that the **means** of the errors drawn from repeated samples of observations will be normally distributed – but we will deal with that wrinkle later on.

Error Assumptions

1. Errors have identical distributions

$$E(\epsilon_i^2) = \sigma_\epsilon^2$$

2. Errors are independent of X and other ϵ_i

$$E(\epsilon_i) \equiv E(\epsilon | x_i) = 0$$
$$\text{and}$$
$$E(\epsilon_i) \neq E(\epsilon_j) \text{ for } i \neq j$$

3. Errors are normally distributed

$$\epsilon_i \sim N(0, \sigma_\epsilon^2)$$

Taken together these assumption mean that the error term has a normal, independent, and identical distribution (normal i.i.d.). However, we don't know if, in any particular case, these assumptions are met. Therefore we must estimate a linear model.

1.2 Estimating Linear Models

With stochastic models we don't know if the error assumptions are met, nor do we know the values of α and β therefore we must estimate them. The stochastic model as shown in Equation 1.4 is estimated as:

$$Y_i = A + BX_i + E_i \tag{1.4}$$

where E_i is the **residual term**, or the estimated error term. Since no line can perfectly pass through all the data points, we introduce a residual, E, into the regression equation. Note that the predicted value of Y is denoted $\hat{Y}$; y-hat.

$$\begin{aligned} Y_i &= A + BX_i + E_i \\ &= \hat{Y}_i + E_i \\ E_i &= Y_i - \hat{Y}_i \\ &= Y_i - A - BX_i \end{aligned}$$

1.2.1 Residuals

Residuals measure prediction errors, of how far observation Y_i is from predicted $\hat{Y}_i$. This is shown in Figure 1.3.

The residual term contains the accumulation (sum) of errors that can result from measurement issues, modeling problems, and irreducible randomness. Ideally, the residual term

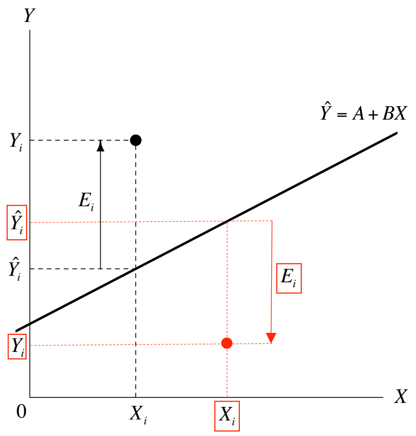


Figure 1.3: Residuals: Statistical Forensics

contains lots of small and independent influences that result in an overall random quality of the distribution of the errors. When that distribution is not random – that is, when the distribution of error has some systematic quality – the estimates of A and B may be biased. Thus, when we evaluate our models we will focus on the `shape` of the distribution of our errors.

The goal of regression analysis is to minimize the error associated with the model estimates. As noted, the residual term is the estimated error, or overall "miss" (e.g., $Y_i - \hat{Y}_i$). Specifically the goal is to minimize the sum of the squared errors, $\sum E^2$. Therefore, we need to find the values of A and B that minimize $\sum E^2$.

Note that for a fixed set of data {A,B}, each possible choice of values for A and B corresponds to a specific residual sum of squares, $\sum E^2$. This can be expressed by the following functional form:

$$S(A, B) = \sum_{i=1}^{n} E_i^2 = \sum (Y_i - \hat{Y}_i)^2 = \sum (Y_i - A - BX_i)^2 \tag{1.5}$$

Minimizing this function requires specifying estimators for A and B such that $S(A, B) = \sum E^2$ is at the lowest possible value. Finding this minimum value requires the use of calculus, which will be discussed in the next chapter. Before that we walk through a quick example of simple regression.

1.3 An Example of Simple Regression

The following example uses a measure of peoples' political ideology to predict their perceptions of the risks posed by global climate change. OLS regression can be done using the `lm` function in `R`. For this example, we are using the `tbur` data set.

What's in E?

Measurement Error

- Imperfect operationalizations
- Imperfect measure application

Modeling Error

- Modeling error/mis-specification
- Missing model explanation
- Incorrect assumptions about associations
- Incorrect assumptions about distributions

Stochastic ”noise”

- Unpredictable variability in the dependent variable

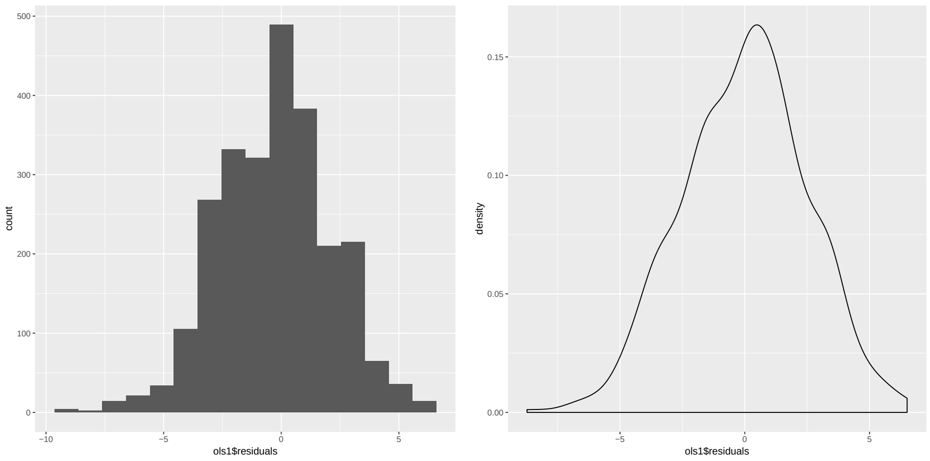
```
ols1 <- lm(ds$glbcc_risk~ds$ideol)
summary(ols1)

## 
## Call:
## lm(formula = ds$glbcc_risk ~ ds$ideol)
## 
## Residuals:
##    Min     1Q Median     3Q    Max 
## -8.726 -1.633  0.274  1.459  6.506 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 10.81866    0.14189   76.25   <2e-16 ***
## ds$ideol    -1.04635    0.02856  -36.63   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
## 
## Residual standard error: 2.479 on 2511 degrees of freedom
##   (34 observations deleted due to missingness)
## Multiple R-squared:  0.3483,Adjusted R-squared:  0.348 
## F-statistic:  1342 on 1 and 2511 DF,  p-value: < 2.2e-16
```

The output in R provides a quite a lot of information about the relationship between the measures of ideology and perceived risks of climate change. It provides an overview of the distribution of the residuals; the estimated coefficients for A and B; the results of hypothesis tests; and overall measures of model "fit" – all of which we will discuss in detail in later chapters. But, for now, note that the estimated B for ideology is negative, which indicates that as the value for ideology *increases*—in our data this means more conservative—the perceived risk of climate change *decreases*. Specifically, for each one unit increase in the ideology scale, perceived climate change risk decreases by -1.0463463.

We can also examine the distribution of the residuals, using a histogram and a density curve. This is shown in Figure 1.4. Note that we will discuss residual diagnostics in detail in future chapters.

```
data.frame(ols1$residuals) %>%
  ggplot(aes(ols1$residuals)) +
  geom_histogram(bins = 16)
data.frame(ols1$residuals) %>%
  ggplot(aes(ols1$residuals)) +
  geom_density(adjust = 1.5)
dev.off()
```



(a) Histogram

(b) Density

Figure 1.4: Residuals of Simple Regression Example

For purposes of this Chapter, be sure that you can run the basic bivariate OLS regression model in *R*. If you can – congratulations! If not, try again. And again. And again...