Simple Regression Model (Part III)

Pedram Jahangiry

Jon M. Huntsman School of Business
Utah State University

Standard assumptions for the linear regression model

Assumption SLR.1 **Linear in Parameters**

In the population model, the dependent variable, y, is related to the independent variable, x, and the error (or disturbance), u, as

$$y = \beta_0 + \beta_1 x + u, \tag{2.47}$$

where β_0 and β_1 are the population intercept and slope parameters, respectively.

Assumption SLR.2 **Random Sampling**

We have a random sample of size n, $\{(x_i, y_i): i = 1, 2, \ldots, n\}$, following the population model in equation (2.47).

Standard assumptions for the linear regression model (cont'd)

Assumption SLR.3 **Sample Variation in the Explanatory Variable**

The sample outcomes on x, namely, $\{x_i, i = 1, \ldots, n\}$, are not all the same value.

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 > 0$$

The values of the explanatory variables are not all the same (otherwise it would be impossible to study how different values of the explanatory variable lead to different values of the dependent variable)

Assumption SLR.4 **Zero Conditional Mean**

The error u has an expected value of zero given any value of the explanatory variable. In other words,

$$\mathrm{E}(u|x) = 0.$$

$$E(u_i|x_i) = 0$$

The value of the explanatory variable must contain no information about the mean of the unobserved factors

Unbiasedness of OLS

THEOREM 2.1

UNBIASEDNESS OF OLS:

Using Assumptions SLR.1 through SLR.4,

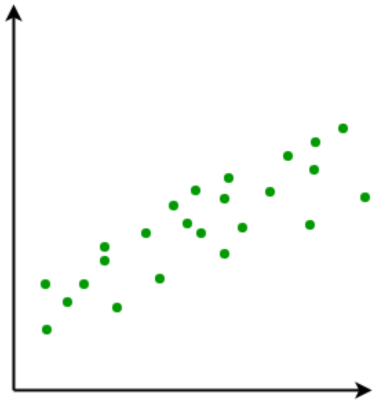
$$E(\hat{\beta}_0) = \beta_0 \text{ and } E(\hat{\beta}_1) = \beta_1,$$

for any values of β_0 and β_1. In other words, $\hat{\beta}_0$ is unbiased for β_0, and $\hat{\beta}_1$ is unbiased for β_1.

Interpretation of unbiasedness

- The estimated coefficients may be smaller or larger, depending on the sample that is the result of a random draw
- However, on average, they will be equal to the values that characterize the true relationship between y and x in the population
- "On average" means if sampling was repeated, i.e. if drawing the random sample and doing the estimation was repeated many times
- In a given sample, estimates may differ considerably from true values. We can never know for sure whether this is the case

Homoskedasticity (Equal variances)



Homoscedasticity

Assumption SLR.5 **Homoskedasticity**

The error u has the same variance given any value of the explanatory variable. In other words,

$$\text{Var}(u|x) = \sigma^2.$$

The value of the explanatory variable must contain no information about the variability of the unobserved factors

THEOREM 2.2

SAMPLING VARIANCES OF THE OLS ESTIMATORS

Under Assumptions SLR.1 through SLR.5,

$$\mathrm{Var}(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \sigma^2/\mathrm{SST}_x,$$

and

$$\mathrm{Var}(\hat{\beta}_0) = \frac{\sigma^2 n^{-1}\sum_{i=1}^{n} x_i^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2},$$

The sampling variability of the estimated regression coefficients depends on 3 things:

1. Variability of the unobserved factors (σ^2)
2. Variation in the explanatory variable $var(X)$ or SST_X
3. Number of observations n

Estimating the error variance

Remember!

1. Error terms $u_i = y_i - \beta_0 - \beta_1 x_i$ are not observable, so we need to come up with an estimate for that!
2. Residuals $\hat{u}_i = y_i - \hat{y}_i$ are observable.

$$Var(u_i|x_i) = \sigma^2 = Var(u_i)$$

Homoskedasticity implies that the variance of u does not depend on x, i.e. equal to the unconditional variance

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(\hat{u}_i - \bar{\hat{u}}_i)^2 = \frac{1}{n}\sum_{i=1}^{n}\hat{u}_i^2$$

One could estimate the variance of the errors by calculating the variance of the residuals in the sample; unfortunately this estimate would be biased

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}\hat{u}_i^2$$

An **unbiased estimate** of the error variance can be obtained by substracting the **number of estimated regression coefficients** from the number of observations

THEOREM 2.3

UNBIASED ESTIMATION OF σ^2

Under Assumptions SLR.1 through SLR.5,

$$E(\hat{\sigma}^2) = \sigma^2.$$

Calculation of standard errors

The estimated standard deviations of the regression coefficients are called "standard errors."
They measure how precisely the regression coefficients are estimated.

$$se(\hat{\beta}_1) = \sqrt{\widehat{Var}(\hat{\beta}_1)} = \sqrt{\hat{\sigma}^2 / SST_x}$$

$$se(\hat{\beta}_0) = \sqrt{\widehat{Var}(\hat{\beta}_0)} = \sqrt{\hat{\sigma}^2 n^{-1} \sum_{i=1}^{n} x_i^2 / SST_x}$$

Plug in $\hat{\sigma}^2$ for the unknown σ^2

Standard Error of the Regression (SER) : It is an estimate of the standard deviation in the unobservables affecting y

$$\hat{\sigma} = \sqrt{\hat{\sigma}^2}$$

Most regression packages (including R) report the value of SER along with the R^2 , $\widehat{\beta_0}$, $\widehat{\beta_1}$, $se(\widehat{\beta_0})$, $se(\widehat{\beta_1})$, t-stats, p-values, and ….

In R, the SER is names as "Residual Standard error"

Put it all together!

THE GAUSS-MARKOV ASSUMPTIONS FOR SIMPLE REGRESSION

For convenience, we summarize the **Gauss-Markov assumptions** that we used in this chapter. It is important to remember that only SLR.1 through SLR.4 are needed to show $\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased. We added the homoskedasticity assumption, SLR.5, to obtain the usual OLS variance formulas (2.57) and (2.58).

Assumption SLR.1 (Linear in Parameters)

In the population model, the dependent variable, y, is related to the independent variable, x, and the error (or disturbance), u, as

$$y = \beta_0 + \beta_1 x + u,$$

where β_0 and β_1 are the population intercept and slope parameters, respectively.

Assumption SLR.2 (Random Sampling)

We have a random sample of size n, $\{(x_i, y_i): i = 1, 2, \ldots, n\}$, following the population model in Assumption SLR.1.

Assumption SLR.3 (Sample Variation in the Explanatory Variable)

The sample outcomes on x, namely, $\{x_i, i = 1, \ldots, n\}$, are not all the same value.

Assumption SLR.4 (Zero Conditional Mean)

The error u has an expected value of zero given any value of the explanatory variable. In other words,

$$\mathrm{E}(u|x) = 0.$$

Assumption SLR.5 (Homoskedasticity)

The error u has the same variance given any value of the explanatory variable. In other words,

$$\mathrm{Var}(u|x) = \sigma^2.$$

Regression function in R is *lm(* $y \sim x$ *,* $data$*)*
Use *summary()* on *lm()* to see the regression results in R.

```
> summary(lm(   salary  ~   roe   ,   ceosal1 ))

Call:
lm(formula = salary ~ roe, data = ceosal1)

Residuals:
    Min      1Q  Median      3Q     Max 
-1160.2  -526.0  -254.0   138.8 13499.9 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   963.19     213.24   4.517 1.05e-05 ***
roe            18.50      11.12   1.663   0.0978 .  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1367 on 207 degrees of freedom
Multiple R-squared:  0.01319,	Adjusted R-squared:  0.008421 
F-statistic: 2.767 on 1 and 207 DF,  p-value: 0.09777
```

Rmarkdown version:

```r
reg <- lm ( salary ~ roe , ceosal1    )
stargazer(reg, type="text")
```

```
===============================================
                        Dependent variable:
                    ---------------------------
                              salary
-----------------------------------------------
roe                          18.501*
                            (11.123)

Constant                    963.191***
                            (213.240)

-----------------------------------------------
Observations                   209
R2                            0.013
Adjusted R2                   0.008
Residual Std. Error    1,366.555 (df = 207)
F Statistic            2.767* (df = 1; 207)
===============================================
Note:               *p<0.1; **p<0.05; ***p<0.01
```