Introduction to Regression with R

February 5, 2019

The Regression Model

This is a linear regression model

$$Y = \beta_0 + \beta_1 X + U \tag{1}$$

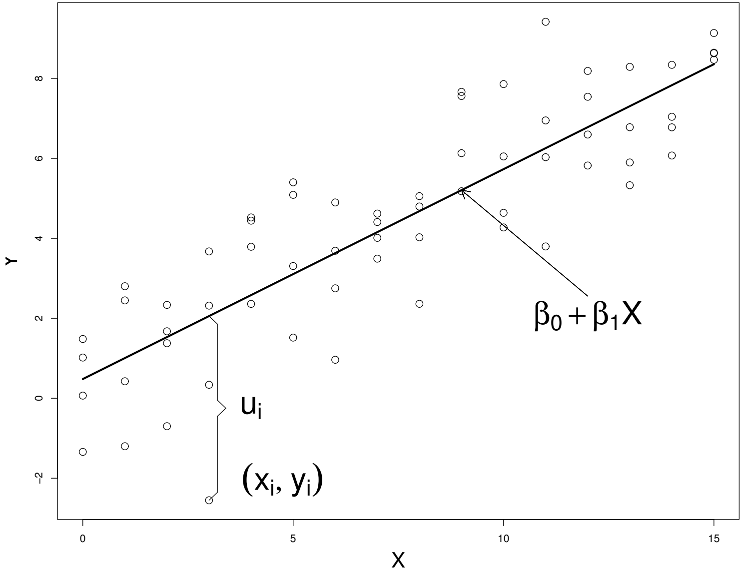
where

Y : Dependent variable (also called respond, regressand and explained variable).

X : Independent variable (also called covariate, regressor and explanatory variable).

U : Error term, $E(U) = 0$. The error term includes all factors that affects Y except for X.

The Regression model



The regression line as a conditional average

Let us take the conditional mean of Y w.r.t. X

$$E(Y|X) = E(\beta_0 + \beta_1 X + U|X) = \beta_0 + \beta_1 X, \tag{2}$$

$\beta_0 + \beta_1 X$ is the regression line.

- For the regression line to actually be a **conditional average**, we have to assume that $E(U|X) = E(U) = 0$.
- If this assumption is wrong, the regression line is not a conditional average but merely a **linear projection**
  - A linear approximation to the conditional average

Least squares (LS)

First, we collect a (random) sample of data $(X_i, Y_i),\ i = 1, \ldots, n$

- Carl Friedrich Gauss realized, in the late 18^{th} century, that we need a measure between the observation points and the line.
- His choice was

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

and the LS estimators are based on the minimization problem:

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

The LS estimators

$$\hat{\beta}_1 = \frac{\sum(Y_i - \bar{Y})(X_i - \bar{X})}{\sum(X_i - \bar{X})^2},$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X},$$

the fitted (predicted) values ($\hat{E}(Y_i|X_i)$)

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i,$$

and the residuals

$$\hat{U}_i = Y_i - \hat{Y}_i.$$

Exercise

Use the data *Prestige* in the package *car*. Consider the following model:

$$\text{prestige} = \beta_0 + \beta_1 \text{income} + U$$

1. We compute $\hat{\beta}_1$ using the formula from the previous slide and then you compute:
2. $\hat{\beta}_0$
3. and $\hat{Y}_i$
4. and $\hat{U}_i$

Distributions of $\hat{\beta}_0$ and $\hat{\beta}_1$

- The $\hat{\beta}_0$ and $\hat{\beta}_1$ are estimators and random variables.
- They will vary between different samples and have distributions with probabilities of each possible outcome
- In small samples we have to assume $U \sim N(0, \sigma^2)$, then

$$T = \frac{\hat{\beta}_j - \beta_j}{S_{\hat{\beta}_j}} \sim t_{n-2}$$

- However if n is large we don't need the normality assumption, then

$$T = \frac{\hat{\beta}_j - \beta_j}{S_{\hat{\beta}_j}} \overset{a.}{\sim} N(0, 1)$$

- We will come back to the details in coming lectures

T-testing of hypotheses about the parameters in a regression model (P-values)

1. $H_0: \ \beta_j = \beta_{j0}$
2. $$H_a: \begin{cases} \beta_j > \beta_{j0} & \text{(upper-tail alternative)} \\ \beta_j < \beta_{j0} & \text{(lower-tail alternative)} \\ \beta_j \neq \beta_{j0} & \text{(two-tailed alternative)} \end{cases}$$
3. Test statistic: $T = T_0 = \frac{\hat{\beta}_j - \beta_{0j}}{S_{\hat{\beta}_j}}$ under H_0
4. $$\text{P-value} = \begin{cases} P(T > t_0) & \text{(upper-tail alternative)} \\ P(T < t_0) & \text{(lower-tail alternative)} \\ 2P(T > |t_0|) & \text{(two-tailed alternative)} \end{cases}$$

T-tests given by default by Statistical packages

1. $H_0: \ \beta_j = 0$
2. $H_a: \ \beta_j \neq 0$
3. Test statistic: $T = T_0 = \frac{\hat{\beta}_j}{S_{\hat{\beta}_j}}$ under H_0
4. $\text{P-value} = 2P\left(T > |t_0|\right)$

i.e. for the coefficient on X the H_0 is: X does not have an effect on Y.

If one rejects H_0 it is common to say that X *has a significant effect on* Y.

Example using the lm() function

Now we will use the lm() function in R for computing the estimates in the previous exercise. Check the help file for lm() (?lm()).

Use the data *Prestige* in the package *car*. Consider the following model:

$$\text{prestige} = \beta_0 + \beta_1 \text{income} + U$$

Does *income* have a significant effect on *prestige*?

Why LS?

1. Practical
   - Most researchers know how LS works and how to interpret the results.
   - All computing softwares have preprogrammed functions for LS estimation (Excel, Minitab, Matlab, R, gretl ...).
2. Theoretically attractive
   - LS is consistent, asymptotic normal and also BLUE ("Best Linear Unbiased Estimator") under a special assumption (more about this latter).

Why not LS?

- A major concern is that we assume a linear functional form.
  - Why? Because it is simple, mathematically
- Because more general models requires more data to give significant results.
  - There is often a choice between a general specification and no significant results and a linear model and significant results

R^2

- How well does the linear relationship explains the data? One measure of fit is the so called R^2 value.

$$R^2 = \frac{\sum(\hat{Y}_i - \bar{Y})^2}{\sum(Y_i - \bar{Y})^2} = \frac{ESS}{TSS} = 1 - \frac{\sum(Y_i - \hat{Y}_i)^2}{\sum(Y_i - \bar{Y})^2} = 1 - \frac{RSS}{TSS}.$$

where

ESS: Explained (variation) sum of squares

RSS: Residual (variation) sum of squares

TSS: Total (variation) sum of squares, and

$$TSS = ESS + RSS$$

R^2 is a consistent estimator of the following measure:

$$1 - \frac{\sigma_U^2}{\sigma_Y^2}$$

R^2

- R^2 is interpreted as the relative fraction of 'variation' of Y explained by X
- The range of R^2 is limited to

  $0 \leq R^2 \leq 1$
- If X do not explain any variation in $Y \Leftrightarrow \hat{\beta}_1 = 0 \Rightarrow$

$$\sum(\hat{Y}_i - \bar{Y})^2 = \sum(\hat{\beta}_0 + \hat{\beta}_1 X_i - \bar{Y})^2 = \sum(\hat{\beta}_0 - \bar{Y})^2 =$$

$$\underbrace{=}_{[\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}]} \sum(\bar{Y} - \bar{Y})^2 = 0 \Rightarrow \underline{\underline{R^2 = 0}}$$

- If X explains all variation in $Y \Rightarrow \hat{Y}_i = Y_i \Rightarrow \underline{\underline{R^2 = 1}}$

The Standard Error of the Regression

- The Standard Error of the Regression is defined as

$$S_U = \sqrt{S_U^2} = \sqrt{\frac{\sum \hat{U}_i^2}{n-2}} = \sqrt{\frac{\sum (Y_i - \hat{Y}_i)^2}{n-2}}$$

This is an estimator of the standard deviation of U, and thus measures the spread around the $E(Y|X) - line$.

Exercise

Compute R^2 for the Prestige regression in the previous example. Compare your r^2 to the ones reported by lm().

Remarks

- A relatively small R^2 value (e.g 0.1) and a large S_U do not imply that X is not important; only that there is other factors accumulated in U that also explain variation in Y
- X could still be the single variable that explains most variation in Y, but of course the probability for this decreases with R^2.
- How large R^2 is supposed to be is arbitrary and differs between different research disciplines and applications.

Regression when the explanatory variables is binary

Binary variables, also called Dummy or indicator variables, e.g.

$$D = \begin{cases} 1 & \text{if} \quad \text{Female} \\ 0 & \text{if} \quad \text{Male} \end{cases} \quad \text{or} \quad D = \begin{cases} 1 & \text{if} \quad \text{Medicine} \\ 0 & \text{if} \quad \text{Placebo} \end{cases}$$

$$Y = \beta_0 + \beta_1 D + U$$

$$\begin{aligned} E(Y|D=0) &= \beta_0 \\ E(Y|D=1) &= \beta_0 + \beta_1 \\ E(Y|D=1) - E(Y|D=0) &= \beta_1 \end{aligned}$$

- Thus in the simple regression framework, with one binary regressor, the interpretation of β_1 is the difference in average of Y given group belonging, $D = 1$ or $D = 0$.

Next

- Next time we start with the second exercise set. You will mainly work with the lm() function in R. Estimate regressions and interpret results based on today's lecture.
- Next lecture we will examine some matrix algebra related to least-squares regression (when the number of X-variables are more than one).