Regression Analysis in R

Kaloyan Ganev

2022/2023

Contents

Introduction

Introduction

- Note: This lecture is not intended to replace an entire econometrics course!
- In statistical analysis, very often the goal is to try to explain the behaviour of one variable through the values of other variables
- The former is often called the *dependent* variable, while the latter are called *predictors*
- Establishing the nature of such relationships allows to potentially control the dependent variable by altering the values of predictors
- Also, this allows forecasting the future values of the dependent variable

What Is Regression Analysis?

- A method for studying functional relationships among variables
- Can be applied to different kinds of data: cross-sectional, time-series, or panel
- Relates to a *quantitative* dependent variable, predictors (regressors) can be both quantitative and qualitative
- Important: Regression assumes causality but it cannot prove causality alone!
- *Ceteris paribus* and regression analysis. . .
- In this lecture, we focus on cross-section-data regression analysis only

Simple Linear Regression

Simple Linear Regression

- Used to study the relationship between two variables
- Also called *bivariate regression*
- Formalization:

$$y = \alpha + \beta x + \varepsilon \quad (*)$$

- Here, y is the dependent variable, x is the regressor (predictor, covariate,. . .), and ε is the stochastic disturbance (error) term
- α is the *intercept parameter*, while β is the *slope parameter*
- It is linear as it defines a linear relationship
- Note: Linearity pertains to parameters not to variables!
- Examples

$$y = \alpha + \beta \ln x + \varepsilon \quad \text{(Linear)}$$

$$y = \alpha + \frac{1}{\beta} x + \varepsilon \quad \text{(Non-linear)}$$

Simple Linear Regression (2)

- Assume we work with $(*)$
- Using the *ceteris paribus* condition (which implies that there is no change in ε, too),

$$\Delta y = \beta \Delta x$$

- Usually it is assumed that

$$\mathsf{E}(\varepsilon) = 0 \quad (**)$$

- This is already a statement about the distribution of ε
- Assume also that

$$\mathsf{E}(\varepsilon | x) = \mathsf{E}(\varepsilon) \quad (***)$$

  i.e. x and ε are independent (therefore uncorrelated)[1]

[1] Sometimes the term *orthogonal* is also used.

Simple Linear Regression (3)

- Combine $(**)$ with $(***)$ to get

$$\mathsf{E}(\varepsilon|x) = 0$$

- Take conditional expectations in $(*)$ using the latter; this yields

$$\mathsf{E}(y|x) = \alpha + \beta x$$

  i.e. the conditional mean of y given x is $\alpha + \beta x$
- This relationship gives the *population regression function (PRF)*
- $\mathsf{E}(y|x)$ is also called the *systematic part* of y, while ε is the *unsystematic (unexplained)* one

OLS Estimation of Regression Parameters

- Assume a sample of size n is taken from the population (n values for x and n values for y)
- Since we know which x's and y's match each other, we can write the sample as

$$\{(x_i, y_i), \, i = 1, 2, \ldots, n\}$$

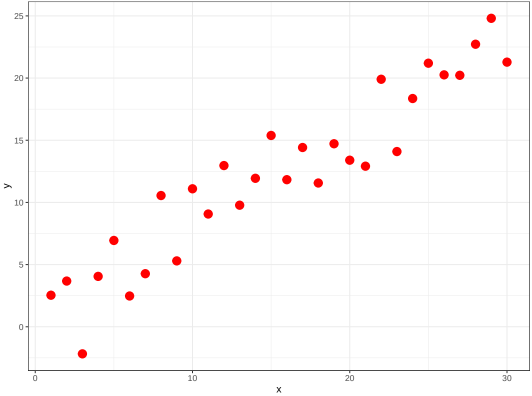
- Also, assuming that the data is generated according to $(*)$, we can write

$$y_i = \alpha + \beta x_i + \varepsilon_i, \quad i = 1, 2, \ldots, n$$

- Recall that, on the one hand, the relationship between x and y is linear but on the other it is non-exact (due to the presence of random disturbances)

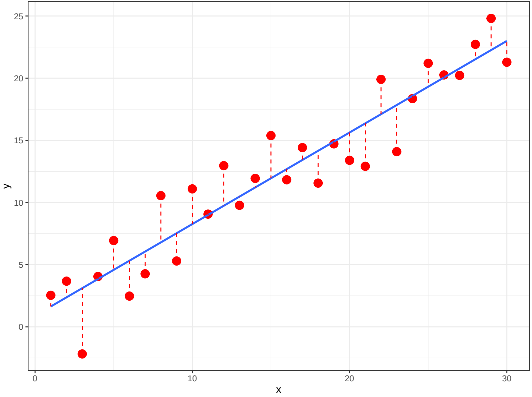
OLS Estimation of Regression Parameters (2)

- Assume that a scatterplot of the data looks as in the following figure



OLS Estimation of Regression Parameters (3)

- The goal is to find the values of α and β that correspond to the straight line passing as close as possible to all data points
- This implies that the sum of deviations from individual points to the line should be the least possible



OLS Estimation of Regression Parameters (4)

- Denote those optimal values of parameter estimates by $\widehat{\alpha}$ and $\widehat{\beta}$
- Then, for each pair of observations, the following should hold:

$$y_i = \widehat{\alpha} + \widehat{\beta}x_i + \widehat{\varepsilon}_i$$

- Obviously, the fitted values of y will be

$$\widehat{y}_i = \widehat{\alpha} + \widehat{\beta}x_i$$

i.e.

$$y_i - \widehat{y}_i = \widehat{\varepsilon}_i$$

OLS Estimation of Regression Parameters (5)

- Sum both sides of the latter over all i

$$\sum_{i=1}^{n}(y_i - \widehat{y}_i) = \sum_{i=1}^{n}\widehat{\varepsilon}_i$$

- In fact, this is what we are supposed to minimize
- However, this turns out to be inappropriate as
  1. There are positive and negative deviations from the regression line
  2. They would be receiving equal weights in summation despite the observations being at different distances from the line
  3. This would lead to a very small sum (in the population, this sum is exactly 0)

OLS Estimation of Regression Parameters (6)

- Therefore, it is a better idea to minimize the sum of squares

$$\min_{\widehat{\alpha},\widehat{\beta}} \sum_{i=1}^{n} \widehat{\varepsilon}_i^2 = \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \widehat{\alpha} - \widehat{\beta} x_i)^2$$

- This is where “least squares” come from
- After minimization is carried out, the estimator of β is

$$\widehat{\beta} = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n} (x_i - \overline{x})^2} \left(= \frac{\mathrm{Cov}(x,y)}{\mathrm{Var}(x)} \right)$$

OLS Estimation of Regression Parameters (7)

- Using

$$y_i = \widehat{\alpha} + \widehat{\beta}x_i + \widehat{\varepsilon}_i$$

we can write

$$\widehat{\varepsilon}_i = y_i - \widehat{\alpha} - \widehat{\beta}x_i$$

- Sum over all i and divide throughout by n

$$\frac{1}{n}\sum_{i=1}^{n}\widehat{\varepsilon}_i = \frac{1}{n}\sum_{i=1}^{n}y_i - \widehat{\alpha} - \widehat{\beta}\frac{1}{n}\sum_{i=1}^{n}x_i$$

or

$$0 = \overline{y} - \widehat{\alpha} - \widehat{\beta}\overline{x}$$

- Therefore:

$$\widehat{\alpha} = \overline{y} - \widehat{\beta}\overline{x}$$

A Practical Example in R

- To estimate a simple linear regression model in R, the `lm` command is used
- We will used some data from Gujarati (2004), p. 81 – data on food expenditures and total expenditures of 55 households in India
- We will try to build a linear model of the relationship between food expenditures and total expenditures
- The data are contained in `foodexp.csv`

A Practical Example

- Load the data into an R data frame

```
expend <- read.csv("foodexp.csv")
```

- Make a scatterplot:

```
fig1 <- ggplot(expend, aes(x = totalexp, y = foodexp)) +
  geom_point(col = "red", size = 4, alpha = 0.5) +
  theme_bw()

fig1
```

- A linear relationship seems to be present
- Run the regression model:

```
mod1 <- lm(foodexp ~ totalexp, data = expend)
```

- The estimation output is stored in a list object

A Practical Example (2)

- Before we look at the regression output, add the regression line to the plot:

```
fig2 <- fig1 +
  geom_smooth(method='lm', formula= y~x, se = F)

fig2
```

- The list object that is created contains a lot of information
- This information can be extracted with specific commands

A Practical Example (3)

- For example, to get regression output you can issue:

```
print(mod1)
```

or

```
summary(mod1)
```

- This is not the only possible info to extract
- Take a look at the output of

```
names(mod1)
```

(More info on those e.g. here: https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/lm)

A Practical Example (4)

- ANOVA of the model:

```
anova(mod1)
```

- Confidence intervals for parameter estimates:

```
confint(mod1, level = 0.95)
```

- Fitted values:

```
fitted(mod1)
```

- Residuals:

```
resid(mod1)
```

A Practical Example (4)

- Prediction:

```
predict(mod1)
```

- Note, however, that this will only reproduce fitter values
- If you want to predict specific values of the dependent variable given specific values of the independent variable:

```
new.df <- data.frame(totalexp=c(1:50))
predict(mod1,newdata=new.df)
```

A Second Example

- We will use data from Woodridge (2012) on CEO salaries and sales
- Load the data

```
load("ceosal1.RData")
```

- To estimate a linear regression between the log of CEO salary and firm sales:

```
mod_ceosal <- lm(log(salary) ~ log(sales), data = data)
```

- Note that the variables are in logs
- Therefore, the slope coefficient is interpreted as elasticity
- This model is a *constant-elasticity* one

A Second Example (2)

- Why is the slope coefficient interpreted as elasticity?
- Start by assuming the general linear relationship between the natural logs of two variables x and y
$$\ln y = \beta \ln x$$
- Transform both sides by exponentiation:
$$e^{\ln y} = e^{\beta \ln x} \Leftrightarrow y = e^{\beta \ln x}$$
- Recall that $(e^x)' = e^x$
- Differentiate both sides with respect to x (use the chain rule)
$$\frac{dy}{dx} = \underbrace{e^{\beta \ln x}}_{=y} \beta \frac{1}{x} = \beta \frac{y}{x} \Rightarrow \beta = \frac{dy}{y} / \frac{dx}{x}$$
- The latter is exactly the definition of elasticity

A Second Example (3)

- Look at

```
summary(mod_ceosal)
```

- How do we interpret the parameter estimates? The intercept? The slope?
- We will leave the remaining part of the regression output for later
- We first need to pay attention at the conceptual level

Goodness of Fit

- From the FOC of the least squares problem follows directly that

$$\sum_{i=1}^{n} \widehat{\varepsilon}_i = 0$$

- Also, those FOCs imply that the sample covariance of x and ε is 0[2]
- The pair of means $(\overline{x}, \overline{y})$ always lies on the regression line (follows from $0 = \overline{y} - \widehat{\alpha} - \widehat{\beta}\overline{x} \Leftrightarrow \overline{y} = \widehat{\alpha} + \widehat{\beta}\overline{x}$)
- Now recall that

$$y_i = \widehat{y}_i + \widehat{\varepsilon}_i \quad (\spadesuit)$$

- We will use this to derive a measure of goodness of fit

[2] In particular, follows from $\sum x_i \widehat{\varepsilon}_i = 0$

Goodness of Fit (2)

- Subtract $\overline{y}$ from both sides of (♠)

$$(y_i - \overline{y}) = (\widehat{y}_i - \overline{y}) + \widehat{\varepsilon}_i$$

- Square both sides and sum over all i

$$\sum_{i=1}^{n}(y_i - \overline{y})^2 = \sum_{i=1}^{n}(\widehat{y}_i - \overline{y})^2 + 2\sum_{i=1}^{n}(\widehat{y}_i - \overline{y})\widehat{\varepsilon}_i + \sum_{i=1}^{n}\widehat{\varepsilon}_i^2$$

- It is easy to show that the middle sum in the RHS equals 0, so

$$\sum_{i=1}^{n}(y_i - \overline{y})^2 = \sum_{i=1}^{n}(\widehat{y}_i - \overline{y})^2 + \sum_{i=1}^{n}\widehat{\varepsilon}_i^2$$

Goodness of Fit (3)

- Note that in the latter, the LHS sum measures the variation of the actual y values around their mean
- By analogy the first sum in the RHS measures the variation of explained (fitted) values arount the mean
- Based on the above, define

$$SST = \sum_{i=1}^{n} (y_i - \overline{y})^2; \quad SSE = \sum_{i=1}^{n} (\widehat{y}_i - \overline{y})^2; \quad SSR = \sum_{i=1}^{n} \widehat{\varepsilon}_i^2$$

- Correspondingly, those are the *total sum of squares*, the *explained sum of squares*, and the *residual sum of squares*

Goodness of Fit (4)

- We can write this simply as

$$SST = SSE + SSR,$$

or, the total variation in y equals the explained variation, plus the unexplained (residual) one

- Divide both sides by SST

$$1 = \frac{SSE}{SST} + \frac{SSR}{SST}$$

- The ratio $\frac{SSE}{SST} = 1 - \frac{SSR}{SST}$ measures the share of total variation explained by the fitted model
- It is called *coefficient of determination* and is denoted by R^2

Goodness of Fit (5)

- Since all sums of squares are non-negative, $0 \leq R^2 \leq 1$
- Take a look at the two regressions that we already estimated
- They both have relatively low values of R^2
- This is not uncommon for cross-sectional data
- It can be shown that R^2 is the square of the sample correlation coefficient between y_i and $\widehat{y}_i$
- While informative, it should not be relied too much upon to judge a model

Properties of OLS Estimators

Before we discuss them, let's state explicitly a set of assumptions underlying the so-called classical linear regression model (CLRM)

1. The regression model is linear in parameters
2. There is a random sample of data on (x, y) of size n
3. Not all values of x are one and the same
4. $\mathrm{E}(\varepsilon|x) = 0$ (exogeneity of the regressor)
5. Homoskedasticity: $\mathrm{Var}(\varepsilon|x) = \sigma^2$
6. No autocorrelation:[3] $\mathrm{E}(\varepsilon_i \varepsilon_j) = 0$, $i \neq j$
7. Normality: $\varepsilon \sim N(0, \sigma^2)$

Note: The last assumption is related only to inference, not to OLS computation.

[3] Follows from 2., but anyway stated explicitly.

Properties of OLS Estimators (2)

Theorem 1

Given the CLRM assumptions, the OLS estimators of α and β are unbiased, i.e.

$$\mathrm{E}(\widehat{\alpha}) = \alpha, \quad \mathrm{E}(\widehat{\beta}) = \beta$$

- Interpretation of unbiasedness: the OLS estimators do not *systematically* overestimate or underestimate the true values of the parameters
- The proof is left to the curious
- The violation of assumptions 1-4 leads to biasedness[4]

[4] The normality and the homoskedasticity assumptions play no role in this respect.

Properties of OLS Estimators (3)

- We will quickly demonstrate the validity of this theorem
- Recall that the OLS estimator of β is

$$\widehat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

- Consider the numerator:

$$\begin{aligned}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) &= \sum_{i=1}^{n}(x_i - \overline{x})y_i - \sum_{i=1}^{n}(x_i - \overline{x})\overline{y} = \\ &= \sum_{i=1}^{n}(x_i - \overline{x})y_i - \overline{y}\sum_{i=1}^{n}(x_i - \overline{x}) = \sum_{i=1}^{n}(x_i - \overline{x})y_i\end{aligned}$$

Properties of OLS Estimators (4)

- With this,

$$\widehat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})y_i}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

- Denote

$$k_i = \frac{(x_i - \overline{x})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

- Thus,

$$\widehat{\beta} = \sum_{i=1}^{n} k_i y_i$$

- From the latter it is obvious that $\widehat{\beta}$ is a linear estimator

Properties of OLS Estimators (5)

- This result can also be written as

$$\begin{aligned}
\widehat{\beta} &= \sum_{i=1}^{n} k_i(\alpha + \beta x_i + \varepsilon_i) = \\
&= \alpha \sum_{i=1}^{n} k_i + \beta \sum_{i=1}^{n} k_i x_i + \sum_{i=1}^{n} k_i \varepsilon_i
\end{aligned}$$

- But $\sum_{i=1}^{n} k_i = 0$ and $\sum_{i=1}^{n} k_i x_i = 1$ so

$$\widehat{\beta} = \beta + \sum_{i=1}^{n} k_i \varepsilon_i \quad (\heartsuit)$$

- Take expectations to see $\widehat{\beta}$ is unbiased (the result with respect to $\widehat{\alpha}$ is shown by analogy)

Variances of OLS Estimators

- Using $(\heartsuit)$ and the homoskedasticity assumption,

$$\begin{aligned}\mathsf{Var}(\widehat{\beta}) &= \mathsf{E}(\widehat{\beta} - \mathsf{E}(\widehat{\beta}))^2 = \mathsf{E}(\widehat{\beta} - \beta)^2 = \\ &= \mathsf{E}\left(\sum_{i=1}^{n} k_i \varepsilon_i\right)^2 = \ldots = \sigma^2 \sum_{i=1}^{n} k_i^2\end{aligned}$$

- Using the definition of k_i,

$$\sum_{i=1}^{n} k_i^2 = \frac{1}{\sum_{i=1}^{n}(x_i - \overline{x})^2},$$

so

$$\mathsf{Var}(\widehat{\beta}) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

Variances of OLS Estimators (2)

- By analogy, the variance of $\widehat{\alpha}$ equals

$$\mathrm{Var}(\widehat{\alpha}) = \frac{\sigma^2 \sum_{i=1}^{n} x_i^2}{n \sum_{i=1}^{n} (x_i - \overline{x})^2}$$

- We can also calculate the covariance of $\widehat{\alpha}$ and $\widehat{\beta}$
- Combining $\widehat{\alpha} = \overline{y} - \widehat{\beta}\overline{x}$ and $\mathrm{E}(\widehat{\alpha}) = \overline{y} - \beta\overline{x}$ leads to

$$\widehat{\alpha} - \mathrm{E}(\widehat{\alpha}) = \widehat{\alpha} - \alpha = -\overline{x}(\widehat{\beta} - \beta)$$

Variances of OLS Estimators (3)

- Using the latter,

$$\begin{aligned}\mathrm{Cov}(\widehat{\alpha}, \widehat{\beta}) &= \mathrm{E}(\widehat{\alpha} - \mathrm{E}(\widehat{\alpha}))(\widehat{\beta} - \mathrm{E}(\widehat{\beta})) = \mathrm{E}(\widehat{\alpha} - \alpha)(\widehat{\beta} - \beta) = \\ &= \mathrm{E}[-\overline{x}(\widehat{\beta} - \beta)(\widehat{\beta} - \beta)] = -\overline{x}\,\mathrm{E}(\widehat{\beta} - \beta)^2 = -\overline{x}\,\mathrm{Var}(\widehat{\beta})\end{aligned}$$

- In the formulas for the variances of $\widehat{\alpha}$ and $\widehat{\beta}$ however stands the theoretical variance of ε equal to σ^2
- A sample estimate is needed to replace it
- Following a similar line of reasoning, it turns out that

$$\widehat{\sigma}^2 = \frac{\sum_{i=1}^{n} \widehat{\varepsilon}_i^2}{n-2}$$

Back to the Last Example

- We first calculate the residual variance as it is needed in the other formulas, too

```
sigmahat_sq <- sum((mod_ceosal$residuals)^2)/(length(data$
    salary) - 2)
```

- From it, we also find the residual standard error

```
resid_se <- sqrt(sigmahat_sq)
```

- This allows to make a comparison with regression output (as everything that follows)

Back to the Last Example (2)

- The variances and standard errors of $\widehat{\alpha}$ and $\widehat{\beta}$ are found using

```
var_ahat <- sigmahat_sq * sum((log(data$sales)^2)) /
(sum((log(data$sales) - mean(log(data$sales)))^2) * length(data
    $salary))

se_ahat <- sqrt(var_ahat)

var_bhat <- sigmahat_sq / sum((log(data$sales) - mean(log(data$
    sales)))^2)

se_bhat <- sqrt(var_bhat)
```

- The covariance of the two (alongside with the built-in function):

```
cov_ahat_bhat <- -mean(log(data$sales)) * var_bhat

vcov(mod_ceosal)
```

Back to the Last Example (3)

- Given the above, the t-statistics can be computed:

```
t_ahat <- coef(mod_ceosal)[1]/se_ahat
t_bhat <- coef(mod_ceosal)[2]/se_bhat
```

... and the corresponding p-values

```
df_t <- length(data$salary) - 2
pval_t_ahat <- 2 * (1 - pt(t_ahat, df_t))
pval_t_bhat <- 2 * (1 - pt(t_bhat, df_t))
```

Back to the Last Example (4)

- Finally, we can compute R^2

```
TSS <- sum((log(data$salary) - mean(log(data$salary)))^2)
ESS <- TSS - RSS
R_sq <- ESS/TSS
```

- The adjusted R^2 tries to impose a penalty for increasing the number of regressors relative to the number of observations:

$$\overline{R}^2 = 1 - (1 - R^2)\frac{n-1}{n-p-1},$$

where p is the number of regressors

- In R,

```
nobs <- length(data$salary)
ncoef <- length(coef(mod_ceosal)) - 1
R_sq_adj <- 1 - (1 - R_sq) * (nobs - 1) / (nobs - ncoef - 1)
```

Multiple linear regression

Multiple linear regression

- We will use a dataset on student grades
- We will try to model them by means of the number of books read and the number of lectures attended
- The data is contained in `data1_1.sav`, an SPSS file
- To read this file into R, we need the `foregin` package:

```
library(foreign)
```

Multiple linear regression (2)

- The data is read with the following command into a data frame:

```
gradedata <- read.spss("data1_1.sav", to.data.frame = T)
```

- Change column names to lowercase:

```
names(gradedata) <- tolower(names(gradedata))
```

- Attach the data frame so you can use variable names directly:

```
attach(gradedata)
```

Multiple linear regression (3)

- Run the following regression model:

```
mod2 <- lm(grade ~ books + attend)
```

- View the model output summary:

```
summary(mod2)
```

- Mind the interpretation of the regression coefficients!
- Note: To run the model without an intercept term:

```
mod2 <- lm(grade ~ 0 + books + attend)
```

or

```
mod2 <- lm(grade ~ -1 + books + attend)
```

Polynomial regression

Polynomial regression

- Use the `polynomial.csv` file:

```
poly <- read.csv("poly.csv")
names(poly) <- tolower(names(poly))
attach(poly)
```

- Run the following models:

```
p1 <- lm(y ~ x)
p2 <- lm(y ~ x + I(x^2))
p3 <- lm(y ~ x + I(x^2) + I(x^3))
p4 <- lm(y ~ x + I(x^2) + I(x^3) + I(x^4))
```

- Note: these are called *nested models*
- The function `I()` is interpreted as “as is”

Polynomial regression (2)

- To display the values of x and y on a graph:

```
plot(x,y)
```

- If you want to compare the fit and the actual values:

```
lines(x, fitted(p1), lwd=2, col="red"))
```

References

References

- Gujarati, D., and D. Porter (2008): *Basic Econometrics*, McGraw-Hill Irwin, 5th edn.
- Wooldridge, J. (2012): *Introductory Econometrics*, Cengage Learning, 5th edn.