AREC422 Notes

Youpei Yan

February 5, 2019

1 R & Rstudio

Install R, then Rstudio. Rstudio is a user-friendly interface to clearly view graphs, data tables, code, and output at the same time.

Open a file in Rstudio, which is an "R script". It's an R file ended with ".r". We'll submit R files for the assignments.

Four parts of the Rstudio: R script, Console, Environment & Code history, and packages & files.

- R script:
  - A final r.file that you can use to document your final, corrected code
  - You can write comments using "#" to make the code clearer for someone to read
  - To view the result of this r.file, select the lines in it and click "Run"
- Console :
  - It's the "field experiment" part, a scratching paper for you.
  - Try your code here and view the results.
  - If it's correct, copy it to the r.file; if it's wrong, try again with no consequences.
  - You can copy the results of the lines (maybe a regression results with coefficients) and copy it as a comment into the r script.
- Environment: you can view your data (if it's a data file) or values (can be strings, numbers, lists, etc.) here. It keeps the record of things you have.
- Packages & files:
  - Sometimes, the built-in functions are not enough, we need to download some user-written functions instead of programming a complicated thing by ourselves.
  - Plot a picture or view a function's help can also be found here.

Try some code here:
x <- 2+2
x <- 2+3
Note that the value x is overlapped.

We can setup a working directory (see the instruction in HW1).

We can try some summary statistics:

price <- c(1,2,3,4,5,6)
sum(price)
mean<-mean(price)
var(price)
length(price)

Note that mean(price) = sum(price)/length(price). We can also try to put the variance formula starting with:
deviation <-price - mean

Note that deviation contains 6 numbers here from the Environment.

Add ^2 after deviation squares each of the numbers in it.

We can try some regression:

demand <- c(10,9,7,6,3,1) # Define the variable demand
reg1 <- lm(demand ~ price) # lm() defines a linear model
summary(reg1) # We view the result of this regression

Look at the Environment part, we just defined a new list called reg1.

2 Simple Linear Regression: Definition & Estimators

Define a Simple Linear Regression:

- $y = \beta_0 + \beta_1 x + u$: two-variable linear regression model
- y: Dependent var., predicted var.
- x: Independent var., explanatory var., control var., predictor var., regressor
- u: error term, disturbance (unobserved)
- β_1: slope parameter
  e.g.: 1 unit increase in price, will decrease the demand by 1.83 units.
  one-unit change in x has the *same* effect on y
- β_0: intercept parameter, constant term

Note that the defined relationship between x and y is not always true, that's why we observe residuals in R results. (Also note that residuals are not the same as errors.)

The model is made up, remember? Most of the time, it is not the true relationship:
e.g.: $yield \sim rain\_fall$: we omitted the effects of fertilizers, temperatures, number of hours people

working in the field, etc.
e.g.: $wage \sim degree$: we omitted sector, discrimination of gender or race, intelligence, etc.
More importantly, these may not be linear functions.

So, we need assumptions: Assumptions make the model legit; Violating assumptions means that the model is wrong and the coefficients are not trustworthy.

Assumption 1: $E(u) = 0$. (Data points are around the fitted line).
Assumption 2: u and x are uncorrelated, or formally: $E(u|x) = 0$. It is called the "zero conditional mean" assumption.

For instance, if we have a true model:
$\widehat{pollution} = 10 + 5firm\_numbers + car\_numbers + 1.5polluted\_rivers$
but we only estimate this model: $pollution \sim firm\_numbers$, then $car\_numbers$ and $polluted\_rivers$ will be in the error terms.

We know that more firms could mean higher car densities, we end up with $E(u|x) \neq 0$, and we may get a different estimation:
$\widehat{pollution} = 6 + 7firm\_numbers$

In this case, we have different impact of firm numbers (wrong estimation). We over-estimate the impact of firms because some effect of car numbers are absorbed by the variable $firm\_numbers$.

Assume that both of the assumptions are valid for a model (relationship), how to calculate the slope & the intercept with a given dataset?

We start with our two assumptions:

$$\begin{cases} E(u) = 0 \\ E(xu) = 0 \end{cases}$$

Then:

$$\begin{cases} E(y - \beta_0 - \beta_1 x) = 0 \\ E[x(y - \beta_0 - \beta_1 x)] = 0 \end{cases}$$

After that:

$$\begin{cases} \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \\ \frac{1}{n}\sum_{i=1}^{n}[x_i(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)] = 0 \end{cases}$$

We have two equations with 2 unknowns: $\hat{\beta}_0$ and $\hat{\beta}_1$. We add hats on them because even the models are correct, they may not be the true β_0 and β_1, due to the sample collection.

Fitted value for y: $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$
Note that $\hat{y}_i$ is the values on the fitted line, it's not the same as the original y.

Residual: $\hat{u}_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$
Sum of squared residuals (SSR)

$$\sum_{i=1}^{n} \hat{u}_i^2 = \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

We need it as small as possible to estimate the line (which means the estimation of $\hat{\beta}_0$ and $\hat{\beta}_1$.

$$min_{\hat{\beta}_0,\hat{\beta}_1} \sum_{i=1}^{n} \hat{u}_i^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

also returns two equations (using the first order conditions by taking the derivatives). These two equations are the same as the equations we used above.

Anyway, we can derive $\hat{\beta}_0$ and $\hat{\beta}_1$ using either of the methods, and get the results:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{sample\ covariance}{sample\ variance\ of\ x_i}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

They are called the ordinary least square estimates (OLS estimates).