Programming Course

Lecture 10: Econometrics with R

Diego de Sousa Rodrigues
SciencesPo

31 mars 2022

Working with some other data

- `install.packages("UsingR")`
- `library(UsingR)`
- `data(galton)`

Working with some other data

- `install.packages("UsingR")`
- `library(UsingR)`
- `data(galton)`

- Let's look at the data first, used by Francis Galton in 1885
- Galton was a statistician who invented the term and concepts of regression and correlation, founded the journal Biometrika, and was the cousin of Charles Darwin
- The idea is to use parents'heights to predict childrens'heights

Working with some other data

- **Exercise** : Look at the marginal (parents disregarding children and children disregarding parents) distributions first plotting the histogram and then compare in the same graph childrens' heights and their parents' heights.

Working with some other data

- **Exercise** : How do you solve the following problem - Why do children of very tall parents tend to be tall, but a little shorter than their parents and why children of very short parents tend to be short, but a little taller than their parents ?

Regression to the mean

- Think of it this way, imagine if you simulated pairs of random normals
- The largest first ones would be the largest by chance, and the probability that there are smaller for the second simulation is high
- In other words $P(Y < x|X = x)$ gets bigger as x heads into the very large values.
- Similarly $P(Y > x|X = x)$ gets bigger as x heads to very small values

Fitting the best line

- $$\text{Child's Height} = \beta_0 + \text{Parent's Height}\ \beta_1$$

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

- $$\hat{\beta}_1 = \text{Cor}(Y, X)\frac{Sd(Y)}{Sd(X)} \quad \hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}$$

- **Exercise** : Calculate the values of $\hat{\beta}_0$ and $\hat{\beta}_1$ using the formulation above. How do the model above change if you decide to run a regression through the origin (i.e., if we force $\hat{\beta}_0 = 0$ we have $\hat{\beta}_1 = \frac{\sum_{i=1}^{n} Y_i X_i}{\sum_{i=1}^{n} X_i^2}$). See also the result you have by doing a regression by centering the data first.

Fitting the best line

- **Exercise** : Normalize the data $\left\{\frac{X_i - \overline{X}}{Sd(X)}, \frac{Y_i - \overline{Y}}{Sd(Y)}\right\}$ and calculate the values of the coefficients of the regression. What can you observe ?

Fitting the best line

- **Exercise** : Now plot the results and one line with the regression outcome, as well as a line with the outcome assuming the parent is the result, one assuming that $corr(X, Y) = 1$ and a point with the mean of x and mean of y.

Fitting the best line

- **Exercise** : Now plot the results and one line with the regression outcome, as well as a line with the outcome assuming the parent is the result, one assuming that $corr(X, Y) = 1$ and a point with the mean of x and mean of y.
- ```
  plot(galtonparent,galtonchild,pch=19,col="blue")
  abline(mean(y) - mean(x) * cor(y, x) * sd(y) /
  sd(x), sd(y) / sd(x) * cor(y, x), lwd = 3, col =
  "red")
  abline(mean(y) - mean(x) * sd(y) / sd(x) / cor(y,
  x), sd(y)/ cor(y, x) / sd(x), lwd = 3, col =
  "blue")
  abline(mean(y) - mean(x) * sd(y) / sd(x), sd(y) /
  sd(x), lwd = 2)
  points(mean(x), mean(y), cex = 2, pch = 19)
  ```
  ```

Interpreting the results

- **Exercise** :Now use the following data

  ```
  data(diamond)
  diamond
  head(diamond)
  ```

  Do the following :

  1. Plot the fitted regression line and data
  2. Estimate a linear regression model and interpret the coefficients
  3. Do a regression to the mean and interpret the coefficients of the slope and the intercept
  4. Reescale the value of x multiplying it by 10 and interpret the coefficients
  5. Predict the price of the diamond using the following vector
     ```
     newx <- c(0.16, 0.27, 0.34)
     ```

Interpreting the results

- **Exercise** : Obtain the residuals manually using the `predict` function. After this check by comparing this with R's built-in `resid` function. Finally, do a residual plot of the data of diamond with the carat data in the x-axis. What can you observe ?

Estimating the variance

- Suppose you have the model
  $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$ where $\epsilon_i \sim N\left(0, \sigma^2\right)$
- The ML estimate of σ^2 is $\frac{1}{n}\sum_{\mathrm{i=1}}^{\mathrm{n}} \mathrm{e}_{\mathrm{i}}^2$, the average squared residual. Most people use

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

  so that we have $E\left[\hat{\sigma}^2\right] = \sigma^2$
- **Exercise** : Estimate the variance of the previous model using the `data(diamond)`. To calculate the residuals estimate first the following : $\hat{\beta}_1 = \mathrm{Cor}(\mathrm{Y}, \mathrm{X})\frac{\mathrm{Sd(Y)}}{\mathrm{Sd(X)}}, \quad \hat{\beta}_0 = \overline{\mathrm{Y}} - \hat{\beta}_1\overline{\mathrm{X}}$, and finally the residuals.

Results

- $$\sigma^2_{\hat{\beta}_1} = \text{Var}\left(\hat{\beta}_1\right) = \sigma^2 / \sum_{i=1}^{n} (X_i - X)^2$$
  $$\sigma^2_{\hat{\beta}_0^2} = \text{Var}\left(\hat{\beta}_0\right) = \left(\frac{1}{n} + \frac{\overline{X}^2}{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}\right)\sigma^2$$
- **Exercise** : Calculate the variance and the standard errors for the estimators of the model using `data(diamond)`.

Results

By knowing the t-statistic is given by :

$$\frac{\hat{\beta}_{\mathrm{j}} - \beta_{\mathrm{j}}}{\hat{\sigma}_{\hat{\beta}_{\mathrm{j}}}}$$

and that it follows a t distribution with $n - 2$ degrees of freedom and a normal distribution for large n we can calculate the p-values using the following formulation :

```
pBeta0 <- 2 * pt(abs(tBeta0), df = n - 2, lower.tail =
FALSE)
```

- **Exercise** : Calculate the t-values and the p-values for both estimates using `data(diamond)`.

Results

Organizing your results :

- `coefTable <- rbind(c(beta0, seBeta0, tBeta0, pBeta0), c(beta1, seBeta1, tBeta1, pBeta1))`
- `colnames(coefTable) <- c("Estimate", "Std. Error", "t value", "P(>|t|)")`
- `rownames(coefTable) <- c("(Intercept)", "x")`
- `coefTable`

Exercise : Check whether your results are the same as the ones in :

```
fit <- lm(y ~ x);
summary(fit)coefficients.
```

Prediction of outcomes

Exercise : Can someone explain what the code below is doing ?

- ```
  library(ggplot2)
  newx = data.frame(x = seq(min(x), max(x), length =
  100))
  p1 = data.frame(predict(fit, newdata=
  newx,interval = ("confidence")))
  p2 = data.frame(predict(fit, newdata =
  newx,interval = ("prediction")))
  ```
  ```

Multivariate Regression

- `require(datasets); data(swiss); ?swiss`
- `install.packages("GGally")`
- `library(datasets); data(swiss); require(stats); require(graphics)`
- `pairs(swiss, panel = panel.smooth, main = "Swiss data", col = 3 + (swiss$Catholic > 50))`

Exercise : What is the graph generated about ? Do a regression of fertility against the other variables and interpret the coefficient of agriculture. After run a regression of fertility against agriculture only and explain the coefficient.

Multivariate regression

- Why does the signal reverse ?
- **Exercise** : Now add the following variable `z <- swiss$Agriculture + swiss$Education` and then run `lm(Fertility ~ . + z, data = swiss)` . What does it happen ?

Multivariate regression

- `require(datasets);data(InsectSprays); require(stats); require(ggplot2)`
- `summary(lm(count ~ spray, data = InsectSprays))$coef`

Exercise : Solve the problem by hard coding the dummy variable using as the reference group the spray A. How do you interpret the results ?

Multivariate regression

- `require(datasets);data(InsectSprays); require(stats); require(ggplot2)`
- `summary(lm(count ~ spray, data = InsectSprays))$coef`

Exercise : Solve the problem by hard coding the dummy variable using as the reference group the spray A. How do you interpret the results ?

1. What happens if you include all the 6 variables ?

2. And what about when you omit the intercept ?

3. Do a reordering of level using the function `relevel` we had used previously and using as a reference group the group C.

Model selection

Exercise : Run the following code

- `fit1<-lm(Fertility Agriculture,data=swiss)`
- `fit3 <- update(fit1, Fertility   Agriculture + Examination + Education, data=swiss)`
- `fit5 <- update(fit1, Fertility   Agriculture + Examination + Education + Catholic + Infant.Mortality, data=swiss)`
- `anova(fit1, fit3, fit5)`

Which model do you select based on this result ?

Generalized Linear Models

- Frequently we care about outcomes that have two values : Alive/dead, win/loss, success/failure
- This case is called **Binary**, **Bernoulli** or **0/1 outcomes**
- A collection of exchangeable binary outcomes for the same covariate data are called **Binomial outcomes**
- We will use the Ravens Baltimore win and loss data

Linear Regression

$$\mathrm{RW_i} = \mathrm{b_0} + \mathrm{b_1 RS_i} + \mathrm{e_i},$$

- $\mathrm{RW_i}$ - 1 if Ravens win and 0 otherwise
- $\mathrm{RS_i}$ - number of points Ravens scored
- b_0 - probability of Ravens win if they score 0 points
- b_1 - increase in probability of Ravens win for each additional point
- e_i - residual variation

Linear Regression

Exercise : Run a linear model in R using the model above and explain the main problems with the results.

Odds

- **Binary outcome 0/1**

$$\mathrm{RW_i}$$

- **Probability (0,1)**

$$\mathsf{Pr}\left(\mathrm{RW_i} \mid \mathrm{RS_i}, \mathrm{b_0},\ \mathrm{b_1}\right)$$

- **Odds (0, ∞)**

$$\frac{\mathsf{Pr}\left(\mathrm{RW_i} \mid \mathrm{RS_i}, \mathrm{b_0},\ \mathrm{b_1}\right)}{1-\mathsf{Pr}\left(\mathrm{RW_i} \mid \mathrm{RS_i}, \mathrm{b_0},\ \mathrm{b_1}\right)}$$

- **Log odd** $(-\infty, \infty)$

$$\log\left(\frac{\mathsf{Pr}\left(\mathrm{RW_i} \mid \mathrm{RS_i}, \mathrm{b_0},\ \mathrm{b_1}\right)}{1-\mathsf{Pr}\left(\mathrm{RW_i} \mid \mathrm{RS_i}, \mathrm{b_0},\ \mathrm{b_1}\right)}\right)$$

Linear versus Logistic Regression

- **Linear**

$$\mathrm{RW_i} = \mathrm{b_0} + \mathrm{b_1 RS_i} + \mathrm{e_i}$$
$$\mathrm{E}\left[\mathrm{RW_i} \mid \mathrm{RS_i}, \mathrm{b_0},\ \mathrm{b_1}\right] = \mathrm{b_0} + \mathrm{b_1 RS_i}$$

- **Logistic**

$$\Pr\left(\mathrm{RW_i} \mid \mathrm{RSS_i}, \mathrm{b_0},\ \mathrm{b_1}\right) = \frac{\exp(\mathrm{b_0}+\mathrm{b_1 RS_i})}{1+\exp(\mathrm{b_0}+\mathrm{b_1 RS_i})}$$
$$\log\left(\frac{\Pr(\mathrm{RW_i}|\mathrm{RS_i},\mathrm{b_0},\ \mathrm{b_1})}{1-\Pr(\mathrm{RW_i}|\mathrm{RS_i},\mathrm{b_0},\ \mathrm{b_1})}\right) = \mathrm{b_0} + \mathrm{b_1 RS_i}$$

Interpreting Logistic Regression

$$\log\left(\frac{\mathsf{Pr}\left(\mathrm{RW_i} \mid \mathrm{RS_i}, \mathrm{b_0},\ \mathrm{b_1}\right)}{1 - \mathsf{Pr}\left(\mathrm{RW_i} \mid \mathrm{RS_i}, \mathrm{b_0},\ \mathrm{b_1}\right)}\right) = \mathrm{b_0} + \mathrm{b_1}\mathrm{RS_i}$$

- b_0 - Log odds of Ravens win if they score zero points
- b_1 - Log odds ratio of win probability for each point scored (compared to zero points)
- $exp(b_1)$ - Odds ratio of win probability for each point scored (compared to zero points)

Interpreting Logistic Regression

- Imagine that you are playing a game where you flip a coin with success probability p
- If it comes up heads, you win X. If it comes up tails, you lose Y
- What should we set X and Y for the game to be fair?

Interpreting Logistic Regression

- Imagine that you are playing a game where you flip a coin with success probability p
- If it comes up heads, you win X. If it comes up tails, you lose Y
- What should we set X and Y for the game to be fair ?
- $$\mathrm{E}[\text{ earnings }] = \mathrm{Xp} - \mathrm{Y}(1-\mathrm{p}) = 0$$
  $$\frac{\mathrm{Y}}{\mathrm{X}} = \frac{\mathrm{p}}{1-\mathrm{p}}$$

Interpreting Logistic Regression

- Imagine that you are playing a game where you flip a coin with success probability p
- If it comes up heads, you win X. If it comes up tails, you lose Y
- What should we set X and Y for the game to be fair ?
- $$\mathrm{E}[\text{ earnings }] = \mathrm{Xp} - \mathrm{Y}(1-\mathrm{p}) = 0$$
  $$\frac{\mathrm{Y}}{\mathrm{X}} = \frac{\mathrm{p}}{1-\mathrm{p}}$$
- The odds can be said as "**How much should you be willing to pay for a p probability of winning a dollar ?"**
  - If $p > 0.5$ you have to pay more if you lose than you get if you win
  - If $p < 0.5$ you have to pay less if you lose than you get if you win

Visualizing fitting logistic regression curves

Exercise : Create a vector x from $(-10, 10)$ with $length = 1000$, create a variable called `beta0=0` and the vector `beta1s = seq(.25, 1.5, by = .1)`. After this create the vector `y = 1 / (1 + exp( -1 * ( beta0 + beta1 * x ) ))` and plot in the same graph all the possible results of (x,y) for each one of the `beta1s` with the x-axis varying from (-10,10) and the y axis from (0,1).

Visualizing fitting logistic regression curves

Exercise : Create a vector x from $(-10, 10)$ with $length = 1000$, create a variable called `beta1=1` and the vector `beta0s = seq(-2, 2, by = .5)`. After this create the vector `y = 1 / (1 + exp( -1 * ( beta0 + beta1 * x ) ))` and plot in the same graph all the possible results of (x,y) for each one of the `beta0s` with the x-axis varying from (-10,10) and the y axis from (0,1).

Simulating data and seeing the fitted value

Exercise : Create a vector `x = seq(-10, 10, length = 1000)`, set `beta0 = 0; beta1 = 1` and create a vector `p = 1 / (1 + exp(-1 * (beta0 + beta1 * x)))`. Plot the results.

Simulating data and seeing the fitted value

Exercise : Create a vector `x = seq(-10, 10, length = 1000)`, set `beta0 = 0; beta1 = 1` and create a vector `p = 1 / (1 + exp(-1 * (beta0 + beta1 * x)))`. Plot the results.
1. Now do the following simulation `y = rbinom(prob = p, size = 1, n = length(p))` . Plot the results of this simulation.
2. Finally use the glm to run a regression of y from x using the following code `fit = glm(y ~ x, family = binomial)` . Plot in the same graph the points of your simulation and the fitted values of the regression above. What can you observe ?

Coming back to the data

Exercise : Run a glm simulation of the following model using the binomial family of the model below :

$$\mathrm{RW_i = b_0 + b_1 RS_i + e_i}$$

Coming back to the data

Exercise : Run a glm simulation of the following model using the binomial family of the model below :

$$\mathrm{RW_i} = \mathrm{b_0} + \mathrm{b_1 RS_i} + \mathrm{e_i}$$

- Plot the fit of the model
- To interpret the coefficients we need to take the exponential of them. Do this procedure
- What is the interpretation of the coefficient in this case ?
- Run the following code `anova(logRegRavens,test="Chisq")` . What can you conclude ?

Interpreting odds ratio

- They are not probabilities
- Odds ratio of $1 =$ no difference in odds
- Log odds ratio of $0 =$ no difference in odds
- Odds ratio < 0.5 or > 2 commonly seen as having a "Moderate effect"

Using Poisson distribution

Exercise : Use the data `grogger` and do a linear regression of average duration of sentence against duration of unemployment. What is the problem with this specification ?

Using Matrix Algebra

Exercise : In this exercise our goal is to estimate the coefficients using matrix algebra. We know that the following is true for a model $y = X\beta + \epsilon$

- Estimates : $\hat{\beta} = (X^T X)^{-1} X^T y$
- Fitted values : $\hat{y} = X\hat{\beta}$
- Residuals : $\hat{\epsilon} = y - \hat{y}$
- Residual sum of squares : $RSS = \hat{\epsilon}^T \hat{\epsilon}$

Now use the `data(mtcars)`. The goal is to estimate a regression in which $y = mpg$ and $X = (1, hp, wt)$. Create those variables and estimate the coefficients of this regression. Compute also the fitted values and the residuals. After finish compare your results to the lm function in R.