Tutorial 1

YANG YANG

The Australian National University

Week 3, 2017

Overview

1. Review of last week's lectures
2. Question One
3. Question Two

Simple linear regression models

- $Y = \beta_0 + \beta_1 X + \varepsilon$.
- Use the least square method to find the regression coefficients $\hat{\beta}_0$ and $\hat{\beta}_1$.
- Minimising $\sum_{i=1}^{N} \varepsilon_i^2 = \sum_{i=1}^{N}(Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{N}(Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i))^2$ gives
  $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$ and
  $\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2}$.

Assumptions

General assumptions:

- The Sample is representative of the population of interest;
- The explanatory variable(s), the X variables are measured without error $\longrightarrow$ all the error is in the Y direction;
- A model of the proposed form is appropriate.

SLR assumptions:

- The errors are usually assumed to be independent, zero-mean, constant variance normal random variables.
- $\varepsilon_i \sim iid\ N(0, \sigma^2)$.

SLR model ANOVA table

Source	degrees of freedom	Sum of Squares	Mean Squares	F value
Regression	1	SSR	MSR=SSR/1	MSR/MSE
Error	n-2	SSE	MSE=SSE/(n-2)	
Total	n-1	SST=SSR+SSE		

- $SSR = \sum(\hat{Y}_i - \bar{Y})^2$.
- $SSE = \sum(Y_i - \hat{Y}_i)^2$.
- $SST = \sum(Y_i - \bar{Y})^2$.

Type I and Type II errors

	Decision	
	Accept H_0	Reject H_0
H_0 (true)	Correct decision	Type I error (α error)
H_0 (false)	Type II error (β error)	Correct decision

- $P(\text{Type I error}) = \alpha$ (significance level).
- $1 - \alpha$ is called the confidence.
- There is a strong relationship between Type I and Type II errors. For a given sample size, we can't reduce both errors at the same time. $\longrightarrow$ Increase the sample size.

Hypothesis Test (t-test) on β_1

Step One: Clearly state hypotheses:

$H_0: \ \beta_1 = 0$
$H_0: \ \beta_1 > 0$

Step Two: Calculate test statistic:

$t = \frac{\hat{\beta}_1 - E[\beta_1 | H_0]}{se(\hat{\beta}_1)}$ where $se(\hat{\beta}_1) = \frac{\hat{\sigma}}{\sqrt{S_{xx}}}$.

Step Three: Make a decision according to the decision rule:

Find the critical value and compare it with calculated test statistic. Alternatively, compare p-value to the given significance level.

Assessing the results

Plots:

- Residuals versus fitted values.

(i) constant variance;

(ii) patterns.

- Normal Q-Q plot

Summary measure:

(i) $R^2 = \frac{SS_{Regression}}{SS_{Total}}$;

(ii) Diagnostic statistics, e.g. Cook's distance.

Q1 (a) to (c)

(a) Simple linear regression with summary output; hypothesis test for β_1
- Build a simple linear regression model in R.
- Check summary output of the regression model.
- Do a formal hypothesis test of β_1.

(b) Plotting SLR; making predictions
- Make a scatter and then superimpose the SLR line.
- Use vector multiplication to do predictions.

(c) Calculate the centroid of the data.

Part (a) to (c)

Download the data file "**Lubricant.csv**" from wattle and put it into the directory folder of R/RStudio

(a) Simple linear regression with summary output; hypothesis test for β_1.
- **lm(response variable∼independent variable)**
- **summary(**lm(Res.∼Ind.)**)**
- **sqrt(MSE/S_{xx})**

(b) Plotting SLR; making predictions
- Use **$** to extract SLR coefficients.
- Use **abline(SLR coefficients)** to impose a straight line on the scatter plot.
- **c(1,x-value)%*%(SLR intercept, slope)**

(c) Calculate the centroid of the data.
- **mean()**

Part (d) to (f)

(d) **anova(SLR model)**

(e) **plot(fitted(),residuals())**

(f) Use logical arguments to split the original dataset into four subsets. Make a scatter plot then impose SLR lines. **legend()** is used to add a legend to our plot. Details see "help(legend)".

Question 2

Do a hypothesis test manually with the null "$H_0 : \beta_1 = 120$" with the help of the following R output.

```
> summary(LACE.lm)

Call:
lm(formula = Score ~ Day)

Residuals:
     Min       1Q   Median       3Q      Max
-24.9604 -12.9918  -0.4289  10.2145  26.9767

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   -5.564      9.108  -0.611    0.555
Day          173.587      1.403 123.759   <2e-16 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 16.77 on 10 degrees of freedom
Multiple R-squared:  0.9993, Adjusted R-squared:  0.9993
F-statistic: 1.532e+04 on 1 and 10 DF,  p-value: < 2.2e-16
```

Hypotesis test of $\hat{\beta}_1$

- $t = \frac{\hat{\beta}_1 - E[\beta_1|H_0]}{se(\hat{\beta}_1)}$ where $se(\hat{\beta}_1) = \frac{\hat{\sigma}}{\sqrt{S_{xx}}}$.
- $\hat{\beta}_1 = 173.587$, $E[\beta_1|H_0] = 120$, $se(\hat{\beta}_1) = 1.403$.
- $t = \frac{173.587-120}{1.403} = 38.19458$.
- Compare test statistic to $t_{10,0.95} = 1.8125$.
  $38.19458 >> 1.8125$. Reject the null hypothesis and conclude that β_1 is significantly larger than 120.

This question would become more difficult if the standard error of regression $se(\hat{\beta}_1)$ of the summary output is not given. We need to follow the solution's path and calculate $se(\hat{\beta}_1)$ first.