Tutorial 11

YANG YANG

The Australian National University

Week 12, 2017

Overview

(b) Summary output from the initial model (*potatoes.lm*) is given at the top of page 2 of the *R* output, but details of the F statistic have been edited (replaced by question marks) and the analysis of variance (ANOVA) table is not shown. Fill in the details of the ANOVA table in the spaces shown below. [Hint: you could do this by working with basic formulae from the data, but it is a lot easier to work from other items given in the *R* output – if you are worried about making mistakes, then as well as writing your answers below, give some details of how you obtained these answers in your answer book, otherwise you will get no marks for any incorrect answers.] **(5 marks)**

Source	*Degrees of Freedom*	*Sum of Squares*	*Mean Square*	*F statistic*	*p-value*
Model (Regression)					
Residual (Error)					
Total					

```
> var(potatoes)
          Glucose     Weeks
Glucose 1734.4011 129.69231
Weeks    129.6923  38.76923
>
> summary(potatoes.lm)

Call:
lm(formula = Glucose ~ Weeks)

Residuals:
    Min      1Q  Median      3Q     Max
-48.357 -33.080  -7.357  28.241  67.536

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  110.560     20.950   5.277 0.000195 ***
Weeks          3.345      1.672   2.001 0.068562 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 37.54 on 12 degrees of freedom
Multiple R-squared:  0.2501,   Adjusted R-squared:  0.1877
F-statistic: ? on ? and ? DF,  p-value: ?
```

- $Total_{df} = n-1$, $Model_{df} = 1$, $Residual_{df} = n-2$
- SST $= Total_{df} \times Var(Y) = 13 \times 1734.4011 = 22547.2143$
- $R^2 = 1 - \frac{SSE}{SST}$ then $SSE = (1-0.2501) \times 22547.2143 = 16908.156$ and $SSR = 22547.2143 - 16908.156 = 5639.0583$
- $MSE = \frac{SSE}{Residual_{df}} = 1409.013$ and $MSR = \frac{SSR}{Model_{df}} = SSR = 5639.0583$
- $F = \frac{MSR}{MSE} = \frac{5639.0583}{1409.013} = 4.0021$;
  Alternatively, use $T^2 = F$ with T value given in the summary output $= 2.001$
- $p - value$ is found in the summary output $= 0.068562$

(d) There is also summary output for a second model (*potatoes.lm2*) on page 2 of the *R* output, which includes an additional term added to the initial model. If you are going to fit a model with this additional term, why should you still include the other terms from the initial model as well? Is this additional term a significant addition to the initial model? The ANOVA table is again not shown for this second model, but what would be the F statistic and degrees of freedom associated with this additional term? **(4 marks)**

```
>
> Weeks.sqd <- Weeks^2
> potatoes.lm2 <- lm(Glucose ~ Weeks + Weeks.sqd)
>
> summary(potatoes.lm2)

Call:
lm(formula = Glucose ~ Weeks + Weeks.sqd)

Residuals:
    Min      1Q  Median      3Q     Max
-15.619 -10.839  -7.357  13.446  21.167

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 198.1455    13.7219   14.44 1.70e-08 ***
Weeks       -19.3241     2.8971   -6.67 3.51e-05 ***
Weeks.sqd     1.0304     0.1282    8.04 6.23e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 14.95 on 11 degrees of freedom
Multiple R-squared:  0.891,	Adjusted R-squared:  0.8711
F-statistic: 44.94 on 2 and 11 DF,  p-value: 5.09e-06
```

- As a general rule, when fitting higher order terms (quadratic or interaction terms), we should always include all lower order terms to allow maximum flexibility in how the model fits the data.
- The quadratic term is fitted last in the model $T^2 = F = 8.04^2 = 64.6416$
- $p - value$ is still 6.23×10^{-6}

Question 3 continued

(b) In the context of model *church.lm2* on page 8 of the *R* output, are *Attendance* and *Employment* (grouped together) a significant addition to a model that already contains *Electoral_Roll*? Give full details of an appropriate hypothesis test. **(4 marks)**

```
> anova(church.lm2)
Analysis of Variance Table

Response: Annual_giving
               Df Sum Sq Mean Sq F value   Pr(>F)
Electoral_Roll  1 589.65  589.65 11.9140 0.003282 **
Attendance      1  64.60   64.60  1.3052 0.270067
Employment      1 189.88  189.88  3.8367 0.067809 .
Residuals      16 791.87   49.49
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
```

The underlying population model for church.lm2 *is:*

$$\text{Annual\_giving} = \beta_0 + \beta_1\text{Electoral\_Roll} + \beta_2\text{Attendance} + \beta_3\text{Employment} + \varepsilon$$

$$\varepsilon \textit{ iid } N\left(0, \sigma^2\right)$$

So we can do a nested model F- test for the addition of the two terms involving Attendance *and* Employment *to a model that already includes* Electoral_Roll:

$$H_o : \frac{\sigma^2_{Addition}}{\sigma^2_{Error}} = 1 \quad H_a : \frac{\sigma^2_{Addition}}{\sigma^2_{Error}} > 1 \quad \textit{or equivalently}$$

$$H_o : \beta_2 = \beta_3 = 0 \quad H_a : \textit{at least one of } \beta_2, \beta_3 \neq 0$$

Reject H_o *in favour of* H_a *if observed test statistic* (F):

$$F > F_{2,16}(0.95) = 3.634$$

$$F = \frac{(64.60 + 189.88)/(1+1)}{MS_{Residual}} = \frac{127.24}{49.49} = 2.57$$

So, we do not reject the null hypothesis and can therefore conclude that the additional terms are not a significant addition to the model (though the model church.lm2 *has obviously been affected by multicollinearity).*

SLR basic

- The errors are usually assumed to be independent, zero-mean, constant variance normal random variables.
- $\varepsilon_i \sim iid \ N(0, \sigma^2)$.
- interpretation of diagnostic plots
- Calculation of values in ANOVA table

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:

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

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.

You may need to do t-test manually using given outputs.

ANOVA Table

Source	D.F.	Sum of Squares	Mean Square	F	P-value
Regression	1	$\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$	$\frac{\sum_{i=1}^{n}(\hat{Y}_i-\bar{Y})^2}{1}$	$\frac{MSREG}{MSE}$	$P(T^2 \geq \frac{MSREG}{MSE})$ $T^2 \sim F(1, n-2)$
Error	$n-2$	$\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$	$\frac{\sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2}{n-2}$		
Total	$n-1$	$\sum_{i=1}^{n}(Y_i - \bar{Y})^2$			

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

You can include these in your cheat sheet.

Confidence Intervals and Prediction Intervals

A 100(1-α)% confidence interval for a given value of x, x_0:

$$\hat{y} \pm t_{\alpha/2,n-2} \times s_\epsilon \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}^2}}$$

A 100(1-α)% prediction interval for a given value of x, x_0:

$$\hat{y} \pm t_{\alpha/2,n-2} \times s_\epsilon \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}^2}}$$

Alternatively, current Assignment 2 Q1 (g)

Rarely calculate CI or PI using these formulas.

Hypothesis test of the $\rho_{x,y}$

Step One: $H_0 : \rho_{x,y} = 0$ v.s. $H_A : \rho_{x,y} \neq 0$

Step Two: Test Statistic $= \frac{r-0}{se(r)} = \frac{r\sqrt{n-2}}{1-r^2}$

Step Three: Refers to the t distribution table with $n-2$ degrees of freedom and find the critical values.

Step Four: Compare the calculated test statistics with the critical values and make a decision.

Step Five: Conclusion

Similar structure and formula with a t-test

Assumptions for MLR models

- We assume uncorrelated (independence) and homoscedastic (constant variance) errors.
- We generally assume that the ϵ_i's are normally distributed with zero mean and constant variance.
- We assume that the underlying true relationship between the response and the predictors is a linear one. $\rightarrow$ "**linear in the parameters**"
- Be careful when the given model has transformation in the response. Betas no longer correspond to the original scale of Y.

Standardised residuals

- Internally studentised residuals:
$r_i = \frac{e_i}{s_\epsilon\sqrt{1-h_{ii}}} = \frac{e_i}{\sqrt{MSE(1-h_{ii})}}$

- Externally studentised residuals:
$t_i = \frac{e_i}{s_{-i}\sqrt{1-h_{ii}}} = \frac{e_{i,-i}}{s_{-i}/\sqrt{1-h_{ii}}}$

- Interpreting internally/externally studentised residual plots.

Influence Statistics

- $DFFITS_i$: removal of the i^{th} data point affects the associated fitted value for this point $\longrightarrow |DFFITS_i| > 2\sqrt{p/n}$
- $DEBETAS_i$: each data point's influence on the estimated parameters $\longrightarrow |DEBETAS_i| > 2/\sqrt{n}$
- $COVRATIO_i$: the i^{th} data point influence overall performance of the model $\longrightarrow COVRATIO_i > 1 + 3p/n$ or $COVRATIO_i < 1 - 3p/n$
- Interpretation in relative terms or regards to cut-off values

F-test and T-test for MLR

- F-tests are sequential tests.
- T-tests are marginal tests $\longrightarrow$ p-values are the same even if we change the order of predictors.
- Interpret estimated coefficients.
- Nested F-test

Hypothesis test for outliers

Externally studentised residuals:
$t_i = \frac{e_i}{s_{-i}\sqrt{1-h_{ii}}} = \frac{e_{i,-i}}{s_{-i}/\sqrt{1-h_{ii}}}$

- t_i follows a student's t distribution with $n-p-1$ degrees of freedom under assumption that the i^{th} data point does not suffer from a location shift
- $H_0 : \Delta_i = 0$ vs $H_A : \Delta_i \neq 0$
- qt(0.975, df=error.df-1)

Get yourself familiar with some definitions

Normally you will not required to write down definitions. But sometimes you will be asked to explain how it works.

- sequential variable selection or `step()` function (Page 40 Lecture Notes)
- Mallow's C_p, $PRESS_p$ and $R^2_{adjusted}$ together with plots (Q2 (d) and Q4 (f) of Tutorial 5; Page36-40 Lecture Notes)

Things may go into your cheat sheet

- Complex formulas which are not easy to remember
- Standard hypothesis test
- Framework of interpretations to: main residual plot, Q-Q plot, Cook's distance plot, leverage plot
- "Don't forget to do ..."

Make your own study plan for the final exam!


Thank you!