MONASH University

MONASH BUSINESS SCHOOL

Statistical Thinking (ETC2420/ETC5242)

Re-sampling with regression models

Week 11

Learning Goals for Week 11

- Review hypothesis testing and Confidence intervals
- Apply randomisation techniques from earlier in the semester to regression coefficients.

Review Multiple Linear Regression

- Recall that linear regression provides us with an estimate of the average of y for a given value of x (or *conditional* upon x)
- MLR gives an **estimate** of the impact upon the average of y conditional upon x, after controlling for the other variables
- Our aim is for our regression to "explain" the variability in the dependent variable
- We know that there will be some unexplained (random) part of the dependent variable that we can't explain

Model evaluation tools

- With this in mind, we looked at some ways to asses our model

Model evaluation tools

- Coefficient significance
- Sensible coefficients (sign and size)
- Assess observations:
  - Influential observations: Leverage (X only); Cook's D (Leverage and residual)
  - LOOCV (case-deleted residuals - uses leverage and residuals as well)
    - measure of fit by observation
- Check residuals for patterns/Normality
- Multicollinearity (VIF) - is related to hypothesis testing
- We then discussed some remedies

Model selection

- Used fit statistics to select "best" or preferred model
  - Use *adj* R^2, *negAIC* and *negBIC*
  - **NEVER** use $log - likelihood$ or R^2
- Remember that fit is just one assessment component
- Once we have our preferred model, we still need to use the other assessment tools to decide if it is adequate for our research purpose
- If there are issues, discuss potential remedies
- **MUST** remember our research purpose when deciding on penalties
- So W9, 10 and 11 lectures and tutorials all go together!

CLT-based tests and confidence intervals

- Use the **lm()** function in R for estimated coefficients and their (estimated) standard errors

Due to the availability of an appropriate CLT result

- Can undertake **hypothesis test** for individual regression coefficient β_k
- Can construct **confidence interval** for individual regression coefficient β_k
- for for any $k = 0, ..., p-1$.

CLT-based hypothesis tests

$$H_0 : \beta_k = 0 \;\; vs \;\; H_1 : \beta_k \neq 0$$

Under H_0, $\frac{b_k}{s(b_k)}$ has (approximately) a t_{n-p} distribution

CLT-based confidence intervals

A $(1-\alpha) \times 100\%$ Confidence interval for β_k is given by:

$$b_k \pm t_{\alpha/2, n-p} SE(b_k)$$

Quick Hypothesis test review

- A test of significance is asking if the variable x_k helps to predict y, after controlling for the other variables in the regression
- The population coefficient of x_k (β_k) quantifies the predictive effect of x_k
- We are asking if our sample supports the null hypothesis or not
- So we are testing if the sample evidence, reflected by our estimate b_k, is likely to have come from the distribution implied by H_0

Quick Hypothesis test review

- CLT uses the t-statistic, which incorporates the sample variation in b_k using the $s.e.(b_k)$
- The permutation test simulates the sampling distribution assuming that H_0 is true
- The way we sample incorporates the variation in the data
- Both methods test the same thing, but in different ways

Permutation tests for regression

We have used a **permutation test** previously to formally decide if two groups have the same proportion

- The idea was to **break** the connection between group and promotion outcome
- To **force null hypothesis** (H_0 : no difference between groups) **to hold**
- And generate an approximate **sampling distribution of the test statistic** $p_2 - p_1$

For a **regression**, we test $H_0 : \beta_k = 0$ vs $H_1 : \beta_k \neq 0$

- For any $k = 1, 2, ..., p - 1$ (note no testing for β_0)
- we **need to break any existing association between regressor x_k and y in our sample**
- We do this via permutations (shuffling) the values of x_k over different observations

Permutation-based hypothesis tests for regression

Procedure for coefficient β_k ($k > 1$) based on R permuted samples

Want to test, **for some** $k = 1, 2, ..., p-1$ **(but not for** $k = 0$**)**,

$$H_0 : \beta_k = 0 \;\; vs \;\; H_1 : \beta_k \neq 0$$

1. Create an $(R \times 1)$ **vector** to store all b_k regression coefficients from each permutation sample
2. Repeat for each permutation replication - sampling **WITHOUT** replacement
   - Permute column of tibble containing regressor x_k **only** - keep all other rows of the data frame in order
   - **Fit the regression model** to the permuted data frame
   - **Save** b_k in the i^{th} entry of the storage vector
3. Plot a histogram of the permutation-generated b_k values
   - Draw a vertical red-lines corresponding to the data-based b_k and $-b_k$ value
   - Compute percentage of permutation-generated $abs(b_k)$ values exceeding data-based b_k value
   - We can do one-sided tests

Simulated data example

- Let’s use the simulated data from the tutorial to do a permutation test.

Randomised confidence interval

The $(1-\alpha) \times 100\%$ confidence interval for b_k states

We are $(1-\alpha) \times 100\%$ confident that the **TRUE** β_k lies somewhere within the interval

So we are $(1-\alpha) \times 100\%$ confident that if x_k increased by 1 unit, y will change on average by β_k units, after controlling for the other regressors.

1. Notice that we do not say estimated, as this is a statement about β_k, not b_k
2. This is a general statement - if the units and the other regressors are known, they should be included in the interpretation
3. We must say *after controlling...* (unless it is a simple regression)
4. It is not a probability statement

Bootstrap-based confidence intervals for regression

Bootstrap-based CI for a regression coefficient

1. Create an $(R \times p)$ matrix to store all regression coefficients from each bootstrap sample
   - R rows, one for each for bootstrap sample
   - p columns for number of regression coefficients in model
2. Repeat for each bootstrap replication
   - Sample **rows** of the data frame **with replacement** (use `slice`)
   - **Fit the regression model** for each bootstrap sample
   - **Save all regression coefficients** in a row of the storage matrix
3. Compute bootstrap-based confidence interval for β_k
   - Select the $\alpha/2\%$ and $(1-\alpha/2)\%$ quantiles of the column $(k+1)$ corresponding to β_k
   - (These are the end points of the $(1-\alpha) \times 100\%$ bootstrap-CI for β_k)

Can use for each b_k for $k = 0, 1, 2, ..., p-1$

Simulated data example

- Let’s use the simulated data from the tutorial to do a bootstrap.

Next week

- Revision
  - outline of exam/formula sheet
  - General advice
  - General admin
- Answer your questions (so come prepared)

Tutorials

- Revision quiz questions
  - Use Kahoot!, so have some fun
  - You will get the questions and answers for revision