MONASH University

MONASH BUSINESS SCHOOL

Statistical Thinking (ETC2420/ETC5242)

Multiple regression

Week 10

Learning Goals for Week 10

- Apply multiple regression models
- Diagnose issues related to multicollinearity
- Apply model performance measures
- Formulate a general strategy for building a regression model

Recommended reading for Week 10:

- Chapter 6 in ISRS

Re-cap Week 9

- We reviewed estimation of simple linear regression
- Touched upon testing and interval estimation
- Learnt about diagnosing potential problems using:
  - residuals
  - Leverage and distance

Residuals

- The estimated model is the "explained" part of y, and the residuals are the "unexplained" part
- If we have done a good job of explaining y using x, the residuals should be:
  - random (*i.e.* no pattern)
  - Normally distributed
- If we find a pattern etc, this suggests that we need to investigate alternatives
- We can transform the data, add new variables (including dummy variables), etc
- Different specifications

(Potential) influential observations

- We used leverage and Cook's distance.
- These identified **observations** that could have a large impact on our regression.
- They are informal methods, using a "rule of thumb" threshold
- Potentially influential observations need to be investigated
- We can drop them and see what happens to the regression
- Then we decide what to do
  - We could remove them (but this may change our analysis)
  - We could add new variables (e.g. a dummy variable)

We need to always keep in mind our research purpose

- Now lets look at some other measures

Leave One Out Cross Validation (LOOCV)

- **LOOCV is a method for validating a model**
- **Leverage** is related to **LOOCV** for regression models

$$LOOCV = \frac{1}{n}\sum_{i=1}^{n} e_{[i]}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_{[i]})^2 = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{e_i}{1-h_{ii}}\right)^2$$

Here

- $e_{[i]} = y_i - \hat{y}_{[i]}$ is the i^{th} **case-deleted residual**
- $\hat{y}_{[i]}$ is the **predicted** value for the i^{th} observation
  - using model estimated with the i^{th} case deleted
- e_i is the **OLS residual** based on all of the data, and
- h_{ii} is the i^{th} **leverage** value from the OLS fit
- $\Rightarrow$ This means we can calculate **LOOCV** without fitting all n models!
  - (rather than fitting the n different regressions that leave out just one observation)

What is LOOCV doing?

- It is the average squared **case-deleted residual**
- That is, we remove and observation so that our new sample size is **n-1**
- We then estimate our regression over the **n-1** sample
- Then we use the estimated coefficients to predict the omitted dependent variable $\hat{y}_{[i]}$
- The case deleted residual is the difference between this prediction and the observed (omitted) y_i

What is LOOCV doing?

- So it is using many training sets of size *n-1* to predict many single test sets of observed dependent variables
- We can use this to compare predictive fit of different models
- Since it is related to leverage, we can just use the residuals and leverage statistics and not estimate all the models
  - NOTE: other methods may require the estimation of all the models
- Let's look at the tutorial exercise in R. . .

Multiple regression

How to decide which regressors to include in a model?

- Look at the signs and sizes of estimated coefficients - do they make sense?
- Which regressors are significant?
  - need to be careful in case of multicollinearity
- Does the model fit well?
  - we will move past R^2 today
- Before checking fit, let's consider the potential for **multicollinearity**
- Multicollinearity occurs when **regressors are highly correlated** with each other

Multicollinearity

We assume that each “x” variable provides unique information about y.

- if 2 regressors are closely related:
  - we can’t disentangle their influence
  - they may have low p-values but are important in explaining y
- Many ways to deal with it
  - we will say one is redundant and remove it
- Sometimes we don’t care (eg forecasting)

Variance inflation factor (VIF)

- The VIF can help identify regressors that are closely related
- The VIF is defined as:

$$\frac{1}{1 - R_j^2},$$

  where R_j^2 is computed by regressing variable j on all other variables
- VIF is a measure the **degree of collinearity between the explanatory variables**
- **Values greater than 10** are considered to be high.
  - VIF > 10 implies $R_j^2 > 0.9$

Why is it called Variance Inflation Factor?

- When x_k is correlated with x_j, for $j \neq k$, then estimate $s(b_k)$ will tend to be large

Why would multicollinearity inflate variance of estimates?

- Uncertainty in the **unique** value of β_k

- **A VIF is a measure concerning a regressor**
- **Note**: this is unlike leverage and Cook's D which are concerned with particular observations

Variance of b_1

$$Var(b_1) = \frac{\sigma^2}{SSE(1 - R_1^2)}$$

,

where R_1^2 is computed by regressing variable 1 on all other explanatory variables.

So if R_1^2 is close to one

- then the other variables explain a lot of x_1
- so the denominator of $Var(b_1)$ is small
- which means that $Var(b_1)$ gets bigger
- or is inflated

Model fit

- Assuming that multicollinearity is not a problem
- We would like our model to “fit” or explain y well.
- We already know about R^2, but we cannot use this to compare models (in general)
- There are many ways to assess model fit

But which model?

- Consider a regression with p regressors, including the intercept term
- We want to identify the "best" model from all possible models
- How many possible models?
  - assume we always keep an intercept $\Rightarrow$ 2^{p-1} models
- May exclude certain regressors due to VIFs being too large
  - Still may have a large number of possible models

Fit all possible models (Ensemble)

Ultimately we want to fit and compare all possible models

- i.e. we consider an **ensemble** of models

Use **meifly** R package

For

- Exploratory model analysis
- Fit and graphical explore ensembles of linear models
- We will just use the **fitall()** function
- Can do bootstrap and many other things!

Model performance measures: Adjusted R^2

We cannot use R^2 or **maximised log-Likelihood**

- These will generally increase with more regressors
- **Not helpful** for choosing the regressors!
- So we ignore them

What about using **adjusted-**R^2?

$$adjR^2 = 1 - \left(\frac{n-1}{n-p}\right)\frac{SSE}{SSTo}$$

where p is the number of regressors (including the intercept)

Why??

- Because R^2 will always go up (or stay the same) if you add a new regressor
- We penalise for increasing the number of regressors

Model performance measures: AIC

In a **general** model setting, other penalised measures include

- **Akaike information criterion (AIC)** for model containing parameter θ
  - Where θ is comprised of k components

$$AIC = 2k - 2\ell(\hat{\theta})$$

Choose model where AIC **is minimised** (comparing all possible competing models)

- Or **equivalently, maximise** $negAIC = -2k + 2\ell(\hat{\theta})$
  - We can plot fit measures, so we use negAIC so that the graphs look similar

AIC for linear regression models

For linear regression models, $\theta = (b_0, b_1, \ldots, b_{p-1}, sigma^2)$

$$AIC = 2(p+1) - 2\ell((b_0, b_1, \ldots, b_{p-1}), \hat{\sigma}^2)$$

- The **log-likelihood function** for a linear model is

$$2\ell(\ (b_0, b_1, \ldots, b_{p-1}),\ \hat{\sigma}^2) = c - n \ln \hat{\sigma}^2 - \frac{1}{\hat{\sigma}^2} \sum_{i=1}^{n} (y_i - \hat{y}_i(b_0, b_1, \ldots, b_{p-1}))^2$$

Choose regressors for model where *AIC* is minimised

- $\Rightarrow$ negAIC is a **penalised maximum likelihood** method

Model performance measures: BIC

- The **Bayesian information criterion (BIC)** for k components in parameter θ (in the general model setting)

$$BIC = k \ln(n) - 2\ell(\hat{\theta})$$

And choose model where BIC is minimised

For linear regression

$$BIC = (p+1)\ln(n) - 2\ell(\ (b_0, b_1, \ldots, b_{p-1}),\ \hat{\sigma}^2)$$

- Or **equivalently, maximise** $negBIC = -(p+1)\ln(n) + 2\ell(\ (b_0, b_1, \ldots, b_{p-1}),\ \hat{\sigma}^2)$
- $\Rightarrow$ negBIC is a **penalised maximum likelihood** method

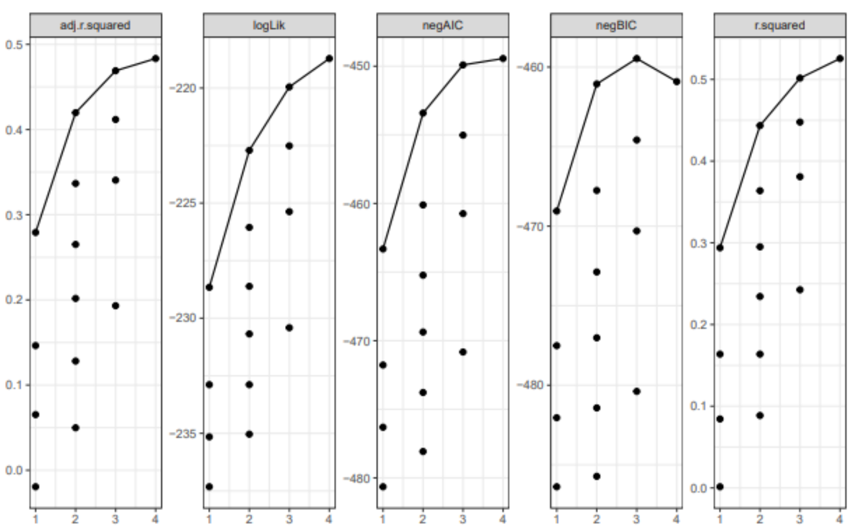
Ensemble output

From **meifly** R package we can

- **Extract the model fit statistics**, adjusted-R^2, AIC, BIC, for each model
- **Display each model fit statistic against the number of regressors** in the model

- We maximise **negAIC**, **negBIC** and $adjR^2$
- Hopefully all will agree on which is the **best model**!
- If not all methods agree, use other criteria to help assess how different is the best model from the next best model
- Can then consider residuals and other diagnostics on a small set of **good** model choices
  - (*i.e.* the stuff from last week!)

Tutorial Example



General strategy

There are many ways to devise a strategy for choosing regressors

Here we consider **automated** methods, but they may miss important aspects

- May need transformations
- May have influential observations
- May still have some multicollinearity

Don't forget the purpose!

Always need to consider the **purpose** intended for the model

- Forecasting
- Finding potential associations between regressors and response
- Understanding of causal factors
- etc. . .

Resource:

- Regression Diagnostics: Identifying Influential Data and Sources of Collinearity