MONASH University

MONASH BUSINESS SCHOOL

Statistical Thinking (ETC2420/ETC5242)

Associate Professor Catherine Forbes

Week 10: Multiple regression

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

Assigned reading for Week 10:

- Chapter 6 in ISRS

Notes

- This material builds on videos from Week 9 (pp25-39)
- Only one more (short) set of videos and slides (Week 11) with new content

Update on Labs and Homeworks

Labs

- Lab 9 submission quiz now covers ONLY Part A of Lab 9 (due Weds week 10)
- We will continue on with Lab 9 in the week 10 Labs
- Lab 10 submission quiz will cover Part B of Lab 9 (due Weds week 11)
- Lab 11 is the final lab (due Weds Week 12)

Homeworks

- Both assignments will be released together
- Both are group assignments
- HW2 submission via Moodle quiz
  - EVERY group member must submit separate quiz
  - Everyone in group will get the average group mark
  - Focus on terminology and numerical results
- HW3 submission via assignment upload
  - RMarkdown and PDF (from HTML)
  - Focus on explanations

How to decide which regressors to include in a model?

- First plot regressors against each other in a **scatterplot matrix**
- Useful to include the response variable too (if there is room!)
- Use the **ggscatmat()** function from the **GGally** R package for this
- Before checking fit, let's consider the potential for **multicollinearity**
- Multicollinearity occurs when **regressors are highly correlated** with each other

Variance inflation factor (VIF)

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

+ 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.

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

Note that unlike leverage and Cook’s D

- which are concerned with particular observations

A VIF is a measure concerning a regressor

But which model?

- With p regressors, including the intercept term
- 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
+ Here we just used the **fitall()** function
+ Can do bootstrap and many other things!

May also be helpful to use
+ **purrr** R package to vectorize operations
+ **stringr** R package to work with labels (strings) more easily

Model performance measures: Adjusted R^2

Cannot use R^2 or **maximised log-Likelihood**

- These will generally increase with more regressors
- **Not helpful** for choosing the regressors!

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** $negAIC = -2k + 2\ell(\hat{\theta})$
- $\Rightarrow$ negAIC is a **penalised maximum likelihood** method

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

- Or **equivalently, maximise**
  $negAIC = -2(p+1) + 2\ell(\ (b_0, b_1, \ldots, b_{p-1}),\ \hat{\sigma}^2)$
- $\Rightarrow$ 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

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

- Or **equivalently, maximise** $negBIC = -(p+1) \ln n + 2\ell(\ (b_0, b_1, \dots, b_{p-1}),\ \hat{\sigma}^2)$
- $\Rightarrow$ 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

Note: We **maximise** *negAIC* and *negBIC* when comparing along side $adjR^2$

- We maximise **negAIC**, **negBIC** and $adjR^2$
- Hopefully all will agree on which is the **best model**!
- If not all methods agree, use 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

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

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