Model Selection

Sean Hellingman ©

Regression for Applied Data Science (ADSC2020)

shellingman@tru.ca

Winter 2025

THOMPSON RIVERS UNIVERSITY

Topics

Introduction

- Assuming that we would like to **make inferences and our models pass the diagnostic checks** we can:
    1. Evaluate individual models performances
    2. Compare two or more models
- How much variability is accounted for?
- Information criteria based on the *likelihood* function.

Individual Models

Variable Significance

- Recall:
  - Hypothesis tests are conducted on each coefficient estimate ($H_0 : \beta_i = 0$).
  - *Simpson's Paradox:* the coefficient on an explanatory variable can depend on what other explanatory variables have been included in the model.
  - May need to include interaction terms or variable transformations.
- Generally, **we can omit variables** whose coefficients are insignificant in multiple estimated models.

Sum of Squares about the Mean

- Sum of squares about the mean (total variability from the grand mean):

- Partitioning the sum of squares:

$$ssTotal = ssR + ssE \tag{1}$$

- ssR: Sum of squares Regression (*explained* by regression)
- ssE: Sum of squares Error (*unexplained* by regression)

Analysis of Variance Table (ANOVA)

Source	df	Sum of Squares	Mean Squares	F-ratio
Regression	p	ssR	msR	$F = msR/msE$
Error	$n-p-1$	ssE	msE	
Total	$n-1$	$ssTotal$	$msTotal$	

- Where:
  - $msR = ssR/p$
  - $msE = ssE/(n-p-1)$
  - $msTotal = ssTotal/(n-1)$

(ANOVA) F-Test

- We use an F-test to test the overall model:

$$H_0 : \beta_1 = \beta_2 = ... = \beta_p = 0$$
$$H_1 : \text{At least one } \beta_j \neq 0.$$

ANOVA in R

- The `summary()` function gives us the results of the F-Test for our model.
- The `anova()` function gives us a breakdown of how much variability is accounted for by adding each variable to a smaller model.

Example 1

- Import the *Housing.csv* dataset into R and conduct the following tasks:
  1. Take a moment to familiarize yourself with the data.
  2. Estimate the following model: price $= 1 +$ area $+$ bathrooms
  3. Comment on the significance of the individual variables and the collective model (F-test).
  4. Comment on the variability accounted for by each variable using the `anova()` function.
     - **Caution:** The order that the variables are included matters if any correlation exists!

Coefficient of Determination (R^2)

- The **coefficient of determination** (R^2) is the proportion of the variation in the dependent variable that is explained by the independent variables.

$$R^2 = \frac{ssR}{ssTotal} \tag{2}$$

- *Percentage of variance explained by the regression model.*
- $0 \leq R^2 \leq 1$
- R^2 **always increases when more explanatory variables are added**
  - Even if they are junk!

Example 2

- Using the *Housing.csv* dataset and example code conduct the following tasks:
  1. Add the two simulated variables to your data frame.
  2. Estimate the same model from Example 1, but include X1 and X2 as explanatory variables.
  3. Are X1 or X2 significant in the model?
  4. Compare the R^2 values from the model in Example 1 and the model including X1 and X2.

Adjusted-R^2

- Using **the Adjusted**-R^2 the value only goes up when included explanatory variables account for more of the variability in the response.

$$\text{Adjusted-}R^2 = 1 - \frac{msE}{msTotal} = 1 - \frac{n-1}{n-(p+1)}(1-R^2)$$

Example 3

- Using the *Housing.csv* dataset and example code conduct the following tasks:
  1. Estimate the same model from Example 1, but include X1 and X2 as explanatory variables.
  2. Are X1 or X2 significant in the model?
  3. Compare the Adjusted-R^2 values from the model in Example 1 and the model including X1 and X2.
  4. What happens to the Adjusted-R^2 value when you add `hotwaterheating` to the model?

Comments on R^2

- Because of how the R^2 is calculated, it does not make sense to compare models with and without an intercept.

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$$

$$R_0^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i y_i^2}$$

- The higher the Adjusted-R^2, the better.
  - Bounded by 1.

- We will use the `anova()` function to compare models.

Multiple Models

Comparing Multiple Models

- To make proper inferences all models under consideration should pass the diagnostic checks.
- Only select more complex models when they are significantly better than a simpler model.
- Some methods of model selection:
  1. ANOVA (Not Adjusted-R^2)
  2. AIC
  3. BIC
  4. *Prediction Accuracy*

ANOVA

- **We can use an ANOVA table to test if the inclusion of more variables is significantly better at capturing variability in the response.**
- An F-test is used to make this comparison.
  - The null hypothesis is that the more complex model does not account for more of the variability.
- It can be useful to test the inclusion of blocks of explanatory variables.

`anova()` in R

- We can use the `anova(lm1,lm2)` function in R
- A small p-value (< 0.05) indicates that the complex model is significantly better at capturing the variability.
- A large p-value (> 0.05) indicates that there is very little difference and we should select the simpler model.

Example 4

- Use the `anova()` function to compare your model from Example 1 and the model from (4) in Example 3.
- Is the more complex model significantly better?

Akaike information criterion (AIC)

- The **Akaike information criterion (AIC)** estimates the relative amount of information lost by a given model: the less information a model loses, the higher the quality of that model.

- In other words, models with a lower AIC are said to be better based on this criterion.

$$AIC = 2k - 2\ln(\hat{L}).$$

  k is the number of estimated parameters.
  $\hat{L}$ is the maximized value of the likelihood function (estimation method).

- Therefore $2k$ is a penalization term for adding more parameters to the model.

Bayesian information criterion (BIC)

- The **Bayesian information criterion (BIC)** also estimates the relative amount of information lost by a given model: the less information a model loses, the higher the quality of that model.

- In other words, models with a lower BIC are said to be better based on this criterion.

$$BIC = k\ln(n) - 2\ln(\hat{L}).$$

  k is the number of estimated parameters and n is the number of observations.
  $\hat{L}$ is the maximized value of the likelihood function (estimation method).

- Therefore $k\ln(n)$ is a **larger** penalization term for adding more parameters to the model.

AIC and BIC in R

- AIC: `AIC(lm1,lm2)`
- BIC: `BIC(lm1,lm2)`
- **Remember, models with the smallest values are considered better.**
- These are two different methods and they may result in different preferences when we are comparing models.

Example 5

- Use the `AIC()` and `BIC()` functions to compare your model from Example 1 and the model from (4) in Example 3.
- What do these result imply?

Stepwise Selection

- We can let R select a model for us based on one of the criteria.
- We can list all the variables under consideration and R will search for the *best* model.
- Algorithm directions:
  1. Forwards
     - Intercept model $\rightarrow$ add one variable at a time.
  2. Backwards
     - Full model $\rightarrow$ remove one variable at a time.
  3. Both (Exhaustive)
     - Intercept model $\rightarrow$ add and remove variables.

Forward Selection

- Start with an empty model (intercept only) then add terms until the *best* model is found (based on AIC):
  - `intercept.model <- lm(response ~ 1, data = data)`
  - `full.model <- lm(response ~ ., data = data)`
    - *Does not need to be all variables, can be a set under consideration.*
  - `forward <- step(intercept.model, direction='forward', scope=formula(full.model), trace=0)`
    - `trace = 1` *Shows each step*
  - `forward$anova` *Shows the results*
  - `forward$coefficients` *Shows the estimates*
  - `k = log(nrow(data))` *BIC*

Example 6

- Use the forward selection algorithm to obtain the *best* model from the *Housing.csv* dataset.
- Use the BIC next.
- What are your thoughts on these models?

Backward Selection

- Start with a full model (all variables under consideration) then remove terms until the *best* model is found (based on AIC):
  - `intercept.model <- lm(response ~ 1, data = data)`
  - `full.model <- lm(response ~ ., data = data)`
    - *Does not need to be all variables, can be a set under consideration.*
  - `backward <- step(full.model, direction='backward', scope=formula(full.model), trace=0)`
    - `trace = 1` *Shows each step*
  - `backward$anova` *Shows the results*
  - `backward$coefficients` *Shows the estimates*
  - `k = log(nrow(data))` *BIC*

Example 7

- Use the backward selection algorithm to obtain the *best* model from the *Housing.csv* dataset.
- Use the BIC next.
- What are your thoughts on these models?

Both (Exhaustive) Selection

- Start with an empty model (intercept only) then add and remove terms (all combinations) until the *best* model is found (based on AIC):
  - `intercept.model <- lm(response ~ 1, data = data)`
  - `full.model <- lm(response ~ ., data = data)`
    - *Does not need to be all variables, can be a set under consideration.*
  - `both <- step(intercept.model, direction='both', scope=formula(full.model), trace=0)`
    - `trace = 1` *Shows each step*
  - `both$anova` *Shows the results*
  - `both$coefficients` *Shows the estimates*
  - `k = log(nrow(data))` *BIC*

Example 7

- Finally, use the exhaustive selection algorithm to obtain the *best* model from the *Housing.csv* dataset.
- Use the BIC next.
- What are your thoughts on these models?

Comments on Stepwise Selection

- **These algorithms are not a substitute for common sense.**
- To include all possible pairwise interactions:

  $(\texttt{variable}_1\ \texttt{+...+}\ \texttt{variable}_p)^2$

- *You can also try polynomial terms in your models.*
- **It may be better to try some shrinkage algorithms if you have many explanatory variables**

Repeated Observations

- If we have repeated observation in a group, we may treat continuous variables as categorical.
    - This allows for more flexibility in the model.
- Such variables *may* be included as factors in a linear regression model.
- Resulting in more coefficients to estimate.

Example 8

- Examine the scatterplots provided in the code to see where repeated observation may allow for more flexible models.
- Estimate new linear regression models using repeated observation as factors.
- Do these models appear to be better?
- Compare the factor model with the continuous model using the `anova()` function.

Exercise 1

- Using the *Wages.csv* dataset:
    - Estimate multiple linear regression models for the dependent variable *Salary*.
    - Compare your models using ANOVA, AIC, and BIC.
        - What is the best combination of variables you can come up with?

Exercise 2

- Following your model selection in Exercise 1, use the selection algorithms we covered to select the *best* linear regression model.
    - Do not be afraid to test interactions and polynomial terms.

- How different are all of the models you uncovered?

References & Resources

1. Kaplan, Daniel T. (2017). *Statistical Modelling: A Fresh Approach. (Second Edition).* Retrieved from https://dtkaplan.github.io/SM2-bookdown/

2. Fox, J. (2015). *Applied regression analysis and generalized linear models (Third Edition).* Sage Publications.

- anova()
- R^2
- `AIC()`
- `step()`
- `Shrinkage Methods`