STAT 308 – Chapter 16

Background Information

We have discussed multiple types of regression methods, including simple and multiple linear regression with numeric and categorical variables, polynomial regression, and regression on transformed response variables. We will now answer the question: given a set of possible variables with or without interactions and polynomial terms, what subset of these variables should we choose to provide the "best'' fitting model.

Maximum Model

Maximum Model: The model with all of the explanatory variables, potentially including polynomial and interaction terms, that we would want to include in our regression model

How to determine if we should include a polynomial term in the maximum model?

- Observe the scatterplots of each explanatory variable on the x-axis compared to the response on the y-axis
- If the plot looks non-linear, add a quadratic, cubic, or higher-order polynomial term to your model, which ever you think looks appropriate

Example

Reconsider the `Boston` dataset in the `MASS` package where we wish to fit a model predicting the median home value of Boston homes in the suburbs. There are 12 potential numeric explanatory variables we can use in our model. Determine if we should include any polynomial terms in our maximum model.

How to determine if we should include an interaction term in the maximum model?

- Observe 3-d scatterplots of each pair of explanatory variables compared to the response on the y-axis
- If the plot looks like there is an apparent pattern between the two variables, add an interaction term

Example

Determine if we should include any interaction terms in our maximum model for median home value.

Type I Error: Including a predictor in our linear model that is truly non-signficant

Type II Error: Excluding a predictor in our linear model that is truly signficant

If adding a polynomial or interaction term appears borderline (i.e. not sure if we should add the term or not), **add the term into the maximum model**. This decreases the chances of making Type II Error. Excluding these significant terms will **introduce bias** into our model.

However, the possibility of **overfitting** is possible in our maximum model. Overfitting occurs either when

- Terms that are not statistically significant are included in our model
- Terms that are related to other statistically significant terms in our model are also included (e.g. including both first floor and basement square footage in a regression model for home sales prices)

Let's see the regression results for our maximum model for median home values in Boston suburbs.

Criteria for determining "best'' regression model

The most obvious criteria is to choose the model that explains that largest amount of variation in the response (i.e. the largest $r^2 = \frac{\text{SSM}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}}$). However, as we have explained previously, r^2 will always increase when we add additional terms to our regression model. So what should we use?

For the following criteria, define a reduced model with p predictors as

$$Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \epsilon$$

with SSE(p) as the sum of squared errors for the p-predictors model. Denote $SST = \sum (y_i - \bar{y})^2$ as the total sums of squares for the response Y, and k be the number of predictors in the maximum model.

Criteria choices:

- Adjusted r^2: $r^2_{adj,p} = 1 - \frac{\text{SSE}(p)/(n-p)}{\text{SST}/(n-1)} = 1 - (1-r^2)\frac{n-1}{n-p}$
  - Adjusted the regular r^2 for the number of parameters in the model
- Akaike Information Criterion (AIC): $AIC = n\log(\text{SSE}(p)/n) + 2p$
  - Compares the "log-likelihood" (based on the sum of square errors) and adds a penalty for the number of parameters in the model
- F-statistic: $F_p = \frac{\text{SSE}(p) - \text{SSE}(k)/(k-p)}{\text{SSE}(k)/(n-(k+1))} = \frac{\text{SSE}(p) - \text{SSE}(k)/(k-p)}{\text{MSE}(p)}$
  - Can compute a p-value for F_p using an F-distribution with $k-p$ and $n-(k+1)$ degrees of freedom used to
- Mallow's C_p: $C_p = \frac{\text{SSE}(p)}{\text{MSE}(k)} - (n - 2(p+1)) = (k-p)F_p + (2p - k + 1)$
  - This value helps to determine the number of variables to be put in the model, since this value is close to $p+1$ if $\text{MSE}(p) \approx \text{MSE}(k)$

We will typically want to choose the model that either **maximizes** the adjusted r^2 or **minimizes** the AIC or Mallow's C_p. Most often, we choose Mallow's C_p since that method simplifies the decision about how many predictors to include in the model

Selecting "best'' regression model

If we want to compare all possible models, this means we have to compare all possible subsets of the maximum model. It turns out there are 2^k possible combinations of models. For a large k, this is computationally infeasible. We need to have a different method of choosing the "best'' regression model.

Backward Selection

1. Start with a model that contains all possible predictors (i.e. $Y \sim \mathcal{N}(\beta_0, \sigma^2)$)
2. For each of the possible variable to be removed from the model

- Calculate Mallow's C_p when adding each variable individually to the model
- Add the variable to the model that has the lowest Mallow's C_p

3. Repeat (2) for each variable not already in the model
4. If the Mallow's C_p does not decrease any further, do not add any additional variables to the model, and return the final model.

Example

Determine the "best'' model for the `Boston` dataset from our maximum model using backward selection.

Forward Selection

1. Start with a model that contains no predictors (i.e. $Y \sim \mathcal{N}(\beta_0, \sigma^2)$)
2. For each of the possible variables to be removed from the model

- Calculate selection criterion when removing each variable individually from the model
- Remove the variable from the model that has the biggest impact on the model (i.e. largest increase in adjusted r^2, largest decrease in AIC/Mallow's C_p)

3. Repeat (2) for each variable not already in the model
4. If the criterion does not change any further, do not remove any additional variables from the model, and return the final model

Stepwise Selection

Stepwise selection starts the same as backward selection, except that it incorporates "re-examination". In other words, variables that were previously removed from the model using backward selection are now allowed to be reentered into the model if adding the variable back in minimizes or maximizes the specific criterion.

Example

Determine the "best'' model for the `Boston` dataset from our maximum model using stepwise selection.