STAT 500

Selection Methods for
Multiple Linear Regression Models

Multiple Regression: Model Selection

- Importance of Model Selection
  - Including too few variables in the model leads to inaccurate estimates of coefficients and response means
  - Including too many variables leads to unnecessary excess variability in estimates of the coefficients and mean response

Model Selection Theory

- Consider two models
  - model A ("fit") $\mathbf{Y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$
  - model B ("true") $\mathbf{Y} = X\boldsymbol{\beta} + Z\boldsymbol{\gamma} + \boldsymbol{\epsilon}$
- Fitting model A (we have omitted the variables in Z) leads to a biased estimate of the regression coefficients:

$$\mathbf{b} = (X^T X)^{-1} X^T \mathbf{Y}$$

$$\begin{aligned}
E(\mathbf{b}) &= E((X^T X)^{-1} X^T \mathbf{Y}) \\
&= (X^T X)^{-1} X^T E(\mathbf{Y}) \\
&= (X^T X)^{-1} X^T (X\boldsymbol{\beta} + Z\boldsymbol{\gamma}) \\
&= (X^T X)^{-1} X^T X\boldsymbol{\beta} + (X^T X)^{-1} X^T Z\boldsymbol{\gamma} \\
&= \boldsymbol{\beta} + (X^T X)^{-1} X^T Z\boldsymbol{\gamma}
\end{aligned}$$

Model Selection Theory

- Estimates of the mean responses based on model A may be biased:

$$\hat{\mathbf{Y}} = X\mathbf{b} = X(X^TX)^{-1}X^T\mathbf{Y} = P_X\mathbf{Y}$$

where $P_X = X(X^TX)^{-1}X^T$ projects vectors into the space spanned by the columns of X. Then,

$$\begin{aligned} E(\hat{\mathbf{Y}}) &= E(P_X\mathbf{Y}) \\ &= P_X E(\mathbf{Y}) \\ &= P_X(X\boldsymbol{\beta} + Z\boldsymbol{\gamma}) \\ &= X\boldsymbol{\beta} + P_X Z\boldsymbol{\gamma} \end{aligned}$$

because $P_X X = X(X^TX)^{-1}X^TX = XI_{n\times n} = X$.

Model Selection Theory

- Bias in $\hat{\mathbf{Y}}$ based on model A:

$$
\begin{aligned}
\boldsymbol{\delta} &= E(\hat{\mathbf{Y}}) - E(\mathbf{Y}) \\
&= (X\boldsymbol{\beta} + P_X Z\boldsymbol{\gamma}) - (X\boldsymbol{\beta} + Z\boldsymbol{\gamma}) \\
&= (P_X - I)Z\boldsymbol{\gamma}
\end{aligned}
$$

- No bias if $\boldsymbol{\gamma} = \mathbf{0}$ or each column in Z is in the column space of X, e.g. the correct model has $E(\mathbf{Y}) = X\boldsymbol{\beta}$

Model Selection Theory

- Another result: Estimate of the error variance based on model A may be biased.

$$E(MS_{error}) = \sigma^2 + \frac{\boldsymbol{\gamma}^T Z^T (I - P_X) Z \boldsymbol{\gamma}}{n-k-1} = \sigma^2 + \frac{1}{n-k-1} \Sigma_i \text{Bias}(\widehat{Y}_i)^2$$

- $p = k + 1$ is the number of parameters in the MLR model.
- Omitting useless terms ($\gamma = 0$): $E(MS_{error}) = \sigma^2$.
- Omitting needed terms ($\gamma \neq 0$): $E(MS_{error}) > \sigma^2$.

Model Selection Theory

- The total variance of $\hat{\mathbf{Y}}$:

$$
\begin{aligned}
\sum_i \text{Var}(\hat{Y}_i) &= \text{trace}(\text{Var}(\hat{\mathbf{Y}})) \\
&= \text{trace}(\text{Var}(P_X \mathbf{Y})) \\
&= \text{trace}(P_X(\sigma^2 I)P_X^T) \\
&= \sigma^2 \text{trace}(P_X) \\
&= \sigma^2(k+1)
\end{aligned}
$$

Because P_X is symmetric and idempodent, i.e.,

$$P_X P_X^T = P_X P_X = P_X,$$

it has $k+1$ eigenvalues equal to one, and the rest are zero.

Model Selection Theory

- Adding a predictor to model A (adding a column to X that is not a linear combination of the columns already in X)
  - decreases bias (or may leave it the same) of $\hat{\mathbf{Y}}$
  - increases the total variance of the estimates of the response means $\hat{\mathbf{Y}}$, because the column rank of X, which is also the rank of the new P_X, increases by 1.
- If we fit the "true" model, model B, then
  - bias $= 0$
  - variance $= \sum_i \text{Var } \hat{Y}_i$
    $= \sigma^2(k + 1 + \text{dim}(Z))$

Model Selection Criteria

- How many explanatory variables? Which ones?
- Criteria for identifying the "best" model
  - R^2
  - adj R^2
  - C_p
  - AIC
  - BIC

Model Selection Criteria – R^2

$$R^2 = \frac{SS_{\text{model}}}{SS_{\text{Total}}}$$

- Larger values indicate better model
- Maximizing R^2 is equivalent to minimizing SS_{error}
- R^2 never decreases when adding an explanatory variable to model
- Most useful for comparing two models with the same number of explanatory variables

Model Selection Criteria – adj R^2

$$\text{adj } R^2 = 1 - \frac{MS_{\text{error}}}{SS_{\text{Total}}/(n-1)}$$

- Larger values indicate better model
- Maximizing adj R^2 equivalent to minimizing $MS_{\text{error}} = \hat{\sigma}^2$
- Does not necessarily increase when adding an explanatory variable to model
- Most useful in comparing models with different numbers of explanatory variables

Model Selection Criteria – C_p

$$C_p = \frac{SS\text{error}}{\hat{\sigma}^2} - (n - 2(k+1))$$

- SS_{error} from fitted model
- $\hat{\sigma}^2$ is MS_{error} for model containing all explanatory variables
- $p = k + 1$ is the number of coefficients in the fitted model

Model Selection Criteria – C_p

- The rationale behind C_p statistic is to minimize $E[\Sigma_i(\hat{Y}_i - E(Y_i))^2]$, the mean squared error of the predictions, $MSEP$= bias^2 + variance.

$$
\begin{aligned}
MSEP &= E[\Sigma_i(\hat{Y}_i - E(\hat{Y}_i) + E(\hat{Y}_i) - E(Y_i))^2] \\
&= E[\Sigma_i(\hat{Y}_i - E(\hat{Y}_i))^2] + E[\Sigma_i(E(\hat{Y}_i) - E(Y_i))^2] \\
&= \Sigma_i \text{Var}(\hat{Y}_i) + \Sigma_i \text{Bias}(\hat{Y}_i)^2 \\
&= \sigma^2(k+1) + E(SS_{error}) - \sigma^2(n-k-1) \\
&= E(SS_{error}) - \sigma^2(n - 2(k+1))
\end{aligned}
$$

- The second to last line uses the previously obtained relationship of bias and $E(MS_{error})$.

Model Selection Criteria – C_p

$$C_p = \frac{SS\text{error}}{\hat{\sigma}^2} - (n - 2(k+1))$$

- Full name – Mallow's C_p
- Good models have C_p around $p = k + 1$
  – Why?
- $C_p < p$ is no problem (sampling error)

Model Selection Criteria – C_p

- Large C_p indicates poor model

- Let m denote the size of biggest possible model with $m-1$ explanatory variables and m regression coefficients.

- For the model containing all explanatory variables, $C_p = m$

- Limited to MLR models

Model Selection: Mallow's C_p

- C_p is related to the F-test that the submodel with only p explanatory variables is acceptable

$$C_p = (m - p)(F - 1) + p$$

If $F < 2$ then $C_p < m$ and the data do not provide enough evidence on bias to reject the submodel

- C_p focuses on prediction

- You can think of using C_p to minimize

$$SS_{error} + \text{[penalty for p]}$$

so do AIC and BIC

Model Selection Criteria – AIC

$$\text{AIC} = n \log(SS_{\text{error}}/n) + 2(k+1)$$

- Full name – Akaike Information Criterion
- Smaller values indicate better models
- Favors models with a slightly larger number of explanatory variables, i.e., may include a few non-significant explanatory variables
- Not limited to MLR models

Model Selection Criteria – BIC

$$\text{BIC} = n \log(SS_{\text{error}}/n) + (k+1)\log(n)$$

- Full name – Bayesian Information Criterion
- Smaller values indicate better models
- Leads to smaller models than AIC (larger penalty for explanatory variables)
- Not limited to MLR models

Model Selection Summary

- Many different approaches
- Measures that focus on fit
  - R^2 (fit using SS_{error}): bad
  - adjusted R^2 (fit using MS_{error})

Model Selection Summary

- Measures that combine fit and complexity
  - general idea: fit + penalty for model complexity
  - Mallows C_p: least penalty
  - AIC: larger penalty
  - BIC: largest penalty (usually)

Model Selection Summary

- Often C_p, AIC and BIC lead to same model
  - When they differ, smaller penalty will tend to select more variables
  - C$_p$ selects most variables
  - BIC selects fewest

Model Comparison Example – Grandfather Clocks

Model	R^2	adj R^2	C_p	AIC	BIC
Numbid	15.62%	12.81%	484.299	379.953	382.884
Age	53.24%	51.68%	255.914	361.065	363.997
Age & Numbid	89.23%	88.49%	39.361	316.065	320.462
Age & Numbid*	95.39%	94.89%	4	290.938	296.801

Multiple Regression: Model Selection

- Significance of explanatory variables depends on presence of other explanatory variables in model.

- Cannot make independent decisions about significance of explanatory variables.

- How do we decide which explanatory variables to be included in the final model?

Multiple Regression Selection Techniques

- Different methods for searching among models
  - All possible subsets of a given group of explanatory variables
  - Stepwise model selection
    - Backward elimination
    - Forward selection
    - Stepwise (Mixed) selection

Model Selection – All Possible Subsets

- Set of k explanatory variables
- Fit all $2^k - 1$ possible models
- Compare models using some criterion (like adjR^2, C_p, AIC or BIC)
- Works up to about $k = 20$ (i.e., takes a reasonable amount of time to process $2^k - 1$ possible models)
- Review the best models of each size - $1, 2, \ldots, k$

Model Selection – Stepwise Methods

- Enter or delete one variable at a time from model according to algorithm
- Less time to compute than all possible subsets
- Possible algorithms
  - Forward selection
  - Backward elimination (selection)
  - Stepwise selection

Model Selection – Forward Selection

1. Start with only intercept in model
2. Fit all one variable models, select the explanatory variable with the largest correlation with the response as long as effect test for variable is statistically significant (p-value $< \alpha_{entry}$)
3. Add to the model the next explanatory variable that reduces the SS_{error} the most as long as effect test for variable is statistically significant (p-value $< \alpha_{\text{entry}}$)
4. Repeat step 3 until no significant variables can be added to the model

Model Selection – Backward Elimination

1. Begin with the largest possible model (all k explanatory variables)

2. Do an effects test for each explanatory variable and compute the p-value.

3. Delete the variable with the least significant effect test (largest p-value) as long as p-value $\geq \alpha_{stay}$

4. Fit the model again. Repeat step 3. Stop when there is no explanatory variable with an effect test with p-value $\geq \alpha_{stay}$

Model Selection – Stepwise Selection

1. Start with only intercept in model
2. Fit all one variable models, select the explanatory variable with the largest correlation with the response as long as effect test for variable is statistically significant (p-value $< \alpha_{entry}$)
3. Add to the model the next explanatory variable that reduces the SS_{error} the most as long as effect test for variable is statistically significant (p-value $< \alpha_{\text{entry}}$)
4. Examine each variable in the current model to make sure effect test for variable is still significant (p-value $< \alpha_{stay}$). If not, delete variable from model.
5. Repeat steps 3-4 until there are no changes
6. Note: need $\alpha_{entry} \leq \alpha_{stay}$ to avoid never ending loops

Difficulties with Model Selection

- High Correlation between Some Explanatory Variables (called multicollinearity)

- Example: Suppose x_j and x_l have a high correlation (near -1 or 1) and are both in model.
  - Significance test for either β_j or β_l: does x_j or x_l significantly add to the model that includes all other explanatory variables?

  - Once one of the variables is in the model, the other is not likely to significantly add to model due to their close association.

Assessing Impact of High Correlation

- Pairwise correlation matrix for explanatory variables
  - $r > |0.7|$
- Models with and without highly correlated explanatory variables.
  - Large change in estimated coefficients, standard errors, and p-values.
- Models with a significant F-test statistic and many or all non-significant t-test statistics.

Variance Inflation Factor (VIF)

- Measures the degree to which the standard error of an estimated coefficient $\widehat{\beta}_j$ is inflated by the correlations with the other explanatory variables.

$$\text{VIF}_j = \frac{1}{1 - R_j^2}$$

where R_j^2 is the R^2 values from the MLR with response variable x_j on the remaining explanatory variables.

Variance Inflation Factor (VIF)

- Explanatory variables with $\text{VIF}_j > 4$ should be investigated further.

- Explanatory variables with $\text{VIF}_j > 10$ indicate severe multi-collinearity.

Solutions to Multicollinearity

- Fit model as is; don't assess significance of individual explanatory variables.

- Select only certain explanatory variables for model to remove highly correlated variables.

- Rely more on theoretical or contextual basis (rather than statistical) for inclusion of variables in model.

Misuses of Model Selection

Observational studies:

- Including an explanatory variable in the model does not imply a causal relationship. Wrong to claim that:
  - Included $\Rightarrow$ variable *causes* change in Y
  - Omitted $\Rightarrow$ variable has *no effect* on Y
  - Omitted $\Rightarrow$ variable is *unimportant*
- DO NOT focus only on estimated coefficients for the selected model
  - Depends on which other variables are included in the model
  - Could be many other reasonable models

Misuses of Model Selection

- Overemphasis on choice of variables in model
  - e.g. repeat a study on a different population
  - Find predictors A, B, D, H in pop. 1
    and predictors A, F, L, M in pop. 2
  - Is A more important?
  - Do the two populations respond differently?
- Extrapolation
  - Model provides good predictions across region of X values included in the study
  - May not be valid outside that region

Model Selection: after selection?

- Still need to examine model assumptions – Model Diagnostics
- Need to examine case diagnostics
- Have we overfit to this particular data set?
  - Rule of thumb: sample size $n >$ 6-10 $\times m$.
  - If sample size is a lot fewer, e.g. 15 candidate variables, $n = 40$ obs, fitted model predicts current data well, new data poorly
  - Model validation