Statistics 203: Introduction to Regression and Analysis of Variance

Model Selection: General Techniques

Jonathan Taylor

Today

- Outlier detection / simultaneous inference.
- Goals of model selection.
- Criteria to compare models.
- (Some) model selection.

Crude outlier detection test

- If the studentized residuals are large: observation may be an outlier.
- Problem: if n is large, if we "threshold" at $t_{1-\alpha/2,n-p-1}$ we will get many outliers by chance even if model is correct.
- Solution: Bonferroni correction, threshold at $t_{1-\alpha/2n,n-p-1}$.

Bonferroni correction

- If we are doing many t (or other) tests, say $m > 1$ we can control overall false positive rate at α by testing each one at level α/m.
- Proof:

$$
\begin{aligned}
P\left(\text{at least one false positive}\right) \\
&= P\left(\cup_{i=1}^m |T_i| \geq t_{1-\alpha/2m, n-p-1}\right) \\
&\leq \sum_{i=1}^m P\left(|T_i| \geq t_{1-\alpha/2m, n-p-1}\right) \\
&= \sum_{i=1}^m \frac{\alpha}{m} = \alpha.
\end{aligned}
$$

- Known as "simultaneous inference": controlling overall false positive rate at α while performing many tests.

Simultaneous inference for β

- Other common situations in which simultaneous inference occurs is "simultaneous inference" for β.
- Using the facts that

$$\widehat{\beta} \sim N\left(\beta, \sigma^2 (X^t X)^{-1}\right)$$

$$\widehat{\sigma}^2 \sim \sigma^2 \cdot \frac{\chi^2_{n-p}}{n-p}$$

along with $\widehat{\beta} \perp \widehat{\sigma}^2$ leads to

$$\frac{(\beta - \widehat{\beta})^t (X^t X)(\widehat{\beta} - \beta)/p}{\widehat{\sigma}^2} \sim \frac{\chi^2_p / p}{\chi^2_{n-p}/(n-p)} \sim F_{p,n-p}$$

- $(1-\alpha) \cdot 100\%$ simultaneous confidence region:

$$\left\{ \beta : (\beta - \widehat{\beta})^t (X^t X)(\widehat{\beta} - \beta) \leq p \widehat{\sigma}^2 F_{p,n-p,1-\alpha} \right\}$$

Model selection: goals

- When we have many predictors (with many possible interactions), it can be difficult to find a good model.
- Which main effects do we include?
- Which interactions do we include?
- Model selection tries to “simplify” this task.

Model selection: general

- This is an "unsolved" problem in statistics: there are no magic procedures to get you the "best model."
- In some sense, model selection is "data mining."
- Data miners / machine learners often work with very many predictors.

Model selection: strategies

- To "implement" this, we need:
  - a criterion or benchmark to compare two models.
  - a search strategy.
- With a limited number of predictors, it is possible to search all possible models.

Possible criteria

- R^2: not a good criterion. Always increase with model size –> "optimum" is to take the biggest model.
- Adjusted R^2: better. It "penalized" bigger models.
- Mallow's C_p.
- Akaike's Information Criterion (AIC), Schwarz's BIC.

Mallow's C_p

- $$C_p(\mathcal{M}) = \frac{SSE(\mathcal{M})}{\widehat{\sigma}^2} - n + 2 \cdot p(\mathcal{M}).$$
- $\widehat{\sigma}^2 = SSE(F)/df_F$ is the "best" estimate of σ^2 we have (use the fullest model)
- $SSE(\mathcal{M}) = \|Y - \widehat{Y}_{\mathcal{M}}\|^2$ is the SSE of the model $\mathcal{M}$
- $p(\mathcal{M})$ is the number of predictors in $\mathcal{M}$, or the degrees of freedom used up by the model.
- Based on an estimate of

$$\begin{aligned}
&\frac{1}{\sigma^2}\sum_{i=1}^n \mathbb{E}\left((Y_i - \mathbb{E}(Y_i))^2\right) \\
&= \frac{1}{\sigma^2}\sum_{i=1}^n \mathbb{E}\left((Y_i - \widehat{Y}_i)^2\right) + \text{Var}(\widehat{Y}_i)
\end{aligned}$$

AIC & BIC

- Mallow's C_p is (almost) a special case of Akaike Information Criterion (AIC)

$$AIC(\mathcal{M}) = -2 \log L(\mathcal{M}) + 2 \cdot p(\mathcal{M}).$$

- $L(\mathcal{M})$ is the likelihood function of the parameters in model $\mathcal{M}$ evaluated at the MLE (Maximum Likelihood Estimators).
- Schwarz's Bayesian Information Criterion (BIC)

$$BIC(\mathcal{M}) = -2 \log L(\mathcal{M}) + p(\mathcal{M}) \cdot \log n$$

Maximum likelihood estimation

- If the model is correct then the log-likelihood of (β, σ) is

$$\log L(\beta, \sigma | X, Y) = -\frac{n}{2}\left(\log(2\pi) + \log \sigma^2\right) - \frac{1}{2\sigma^2}\|Y - X\beta\|^2$$

where Y is the vector of observed responses.

- MLE for β in this case is the same as least squares estimate because first term does not depend on β
- MLE for σ^2:

$$\frac{\partial}{\partial \sigma^2} \log L(\beta, \sigma)\bigg|_{\widehat{\beta}, \widehat{\sigma}^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\|Y - X\widehat{\beta}\|^2 = 0$$

- Solving for σ^2:

$$\widehat{\sigma}^2_{MLE} = \frac{1}{n}\|Y - X\widehat{\beta}\|^2 = \frac{1}{n} SSE(\mathcal{M})$$

Note that the MLE is biased.

AIC for a linear model

- Using $\widehat{\beta}_{MLE} = \widehat{\beta}$

$$\widehat{\sigma}^2_{MLE} = \frac{1}{n} SSE(\mathcal{M})$$

we see that the AIC of a multiple linear regression model is

$$AIC(\mathcal{M}) = n\left(\log(2\pi) + \log(SSE(\mathcal{M})) - \log(n)\right) + 2(n + p(\mathcal{M}) + 1)$$

- If σ^2 is known, then

$$AIC(\mathcal{M}) = n\left(\log(2\pi) + \log(\sigma^2)\right) + \frac{SSE(\mathcal{M})}{\sigma^2} + 2p(\mathcal{M})$$

which is almost $C_p(\mathcal{M}) + K_n$.

Search strategies

- "Best subset": search all possible models and take the one with highest R_a^2 or lowest C_p.
- Stepwise (forward, backward or both): useful when the number of predictors is large. Choose an initial model and be "greedy".
- "Greedy" means always take the biggest jump (up or down) in your selected criterion.

Implementations in R

- "Best subset": use the function `leaps`. Works only for multiple linear regression models.
- Stepwise: use the function `step`. Works for any model with Akaike Information Criterion (AIC). In multiple linear regression, AIC is (almost) a linear function of C_p.
- Here is an example.

Caveats

- Many other "criteria" have been proposed.
- Some work well for some types of data, others for different data.
- These criteria are not "direct measures" of predictive power.
- Later – we will see cross-validation which is an *estimate* of predictive power.