STAT 151A Lecture 27

Henry Liev

29 October 2025

Remark 0.1 (Variable Selection)

Given a set of covariates
$\vec{x}_1, \ldots, \vec{x}_p$, which of these belong in our design matrix
Given a set of different models based on some data, $\vec{x}_1, \ldots, \vec{x}_p$ and $\vec{y}$, which is best?
What might benefits be (relative to keeping all the variables)?

(1) Overfitting $\to$ Bias/Variance tradeoff $\to$ trying to approximate the noise because it has too much flexibility

(2) Interpretability $\to$ more coefficients and harder to interpret each coefficient

(3) Some variables may just not matter (or highly collinear with things in the model) $\to$ if model needs to be deployed very fast or simply, or if you need to collect new data

Variable selection is typically quite important in prediction settings
Often not a good idea in causal inference settings
If you leave out a variable, critics may wonder how different your $\hat{\beta}$ would have been
After variable selection, the usual p-values and confidence intervals cannot be trusted

Remark 0.2 (How do we select variables for a model?)

Incremental F-test:
Fit two models, one bigger and one smaller
Compare them with an F-test
If it does not reject, keep the smaller model
Does not work well at large scale to select good predictive models
F-tests require normality, don't really need it for prediction
Testing two at a time is very slow
Multiple testing: errors are inevitable when we make many tests
Tests aren't designed for good prediction $\to$ sometimes a coefficient is significant but still small and doesn't change $\hat{y}$ much and sometimes the interval for $\hat{\beta}_j$ covers zero but it still large and can give better predictions

Remark 0.3 (Criterion-Based Model Selection)
Model 1, Model 2, . . . , Model K
Criterion F: models $\to \mathbb{R}$
f(model j) $\to$ Score j $\to$ choose the model with the best score
2 questions:

(1) What criterion will be good?

(2) How to organizee search across models (if K is really really big)

Criteria

- Bad Criterion: $R^2 = \frac{RegSS}{TotSS}$, will ignore overfitting, always goes up with more variables
- Better choice: $R^2_{adj} = 1 - \frac{n-1}{n-(p+1)} \cdot \frac{ErrSS}{TotSS}$
- Mallow's C_p
- Cross-Validation errors
- AIC
- BIC

Remark 0.4 (Mallow's C_p)
Idea: Variance/Bias decomposition
We want to learn about some parameter μ_{true}
We have a random variable V with $\mathbb{E}(V) = \mu_V$
What happens if we use V to estimate μ_{true}
Evaluate MSE $\mathbb{E}[V - \mu_{true}]^2 = \mathbb{E}[V - \mu_V + \mu_V - \mu_{true}] = \mathbb{E}[V - \mu_v]^2 + 2\mathbb{E}[(V - \mu_V)(\mu_V - \mu_{true})] + \mathbb{E}[\mu_V - \mu_{true}]^2 = \mathbb{E}[V - \mu_V]^2 + [\mu_v - \mu_{true}]^2 = \mathrm{Var}(V) + \mathrm{Bias}^2(V)$
Connection to linear models:
Suppose we have a design matrix $\mathbf{X} \in \mathbb{R}^{n\times(p+1)}$ suppose also we have a reduced model with $p_m + 1 < p + 1$ columns
If the full model is "correct", $\mathbb{E}[\hat{\beta}_{full}] = \beta$ unbiased, how does MSE of $\hat{\beta}$ change when we use $\hat{\beta}_m$ instead?
Get some bias, but may lose some variance, overall MSE could go up or down, Mallow's C_p: estimate MSE of $\hat{\beta}$
More specifically:
Estiamtes $\frac{\mathbb{E}(\hat{y}_m - \mathbf{X}\beta)^2}{\sigma^2}$ as $(p_m + 1) + (p - p_m)(F_m - 1)$ where F_m is the F-statistic for testing model against a full model
$2(p_m + 1) - n + \frac{ErrSS}{\hat{\sigma}^2}$

Remark 0.5 (Cross-Validation Errors)
One problem with Mallow's C_p based on estimating $\mathbb{E}(\mathbf{X}\hat{\beta}_m - \mathbf{X}\beta) = \mathbb{E}(\hat{y} - \mathbb{E}(y))$
In real life, we want to compare $\hat{y}$ to y
CV error gets this more directly
$CV_{error} = \frac{1}{n}\sum_{i=1}^n (y_i - \hat{y}_{-i})^2$, $\mathbf{X}\hat{\beta}_{-i}$, we approximate without i, leave one out CV error