

Regression in Practice: Variable Selection

Overview

1. Model selection, variable selection
2. Orthogonal case
3. General regression
4. Cross-validation
5. Origins of selection criteria
6. White/Sandwich estimator (computing)

Model Selection in Regression

Premise In practice, there is no such luxury as a “true regression model.”

You can make one up to experiment with different ideas in a simulation, but in reality you don’t know the data generating process. Hence, properties such as consistency of variable selection don’t make much sense from a practical point of view: give me more data, and I will find a different model.

PMSE Find the model that predicts the best when applied to new data.

Let Y and Y^* each denote n independent draws from some population associated with a collection of covariates, and let $M(y, x)$ denote a model that predicts based on input data y and covariates x . We want a model that minimizes the prediction mean squared error

$$PMSE = \mathbb{E}(Y^* - M(Y, X))^2$$

or a similar loss function (log loss, for instance). Notice that we assume that the *same* covariates describe both Y and Y^* . This is hence called the “in-sample” problem because there is no extrapolation.

Regression model Write the model as

$$Y = \mu + \epsilon, \quad E \epsilon = 0, \quad \text{Var } \epsilon = \sigma^2 I_n,$$

where each of Y, μ , and ϵ are $n \times 1$ vectors. Consequently, $\mathbb{E}Y = \mathbb{E}Y^* = \mu$.

In contrast to the more familiar regression model $Y = X\beta + \epsilon$, we do *not* assume the mean vector has the linear form $\mu = X\beta$ for some regression design X . The vector μ is the mean of the responses, whatever that may be.

Model selection How should one choose the set of explanatory variables $X = [X_1, X_2, \dots, X_k]$ which gives the “best” estimate of μ ? A popular approach is to pick the model which balances parsimony with goodness-of-fit is the prediction mean squared error, $PMSE$. For a collection of k explanatory variables X and estimator $\hat{\beta}$, this criterion is

$$PMSE(X) = \mathbb{E} \|Y^* - X\hat{\beta}\|^2/n, \tag{1}$$

where $Y^* = \mu + \epsilon^*$ and ϵ^* has mean zero, variance/covariance $\sigma^2 I_n$ and is uncorrelated with ϵ . Y^* is an independent “copy” of the observed data Y which has the same mean and variance.

Equivalent criterion Rather than look at predicting a new value, we could also look simply at the MSE of the estimator of μ since

$$\mathbb{E} \|Y^* - \hat{\mu}\|^2 = n\sigma^2 + \mathbb{E} \|\mu - \hat{\mu}\|^2$$

Selection in Models with Orthogonal Variables

“True model” \neq Best Suppose that $\mu = X\beta$ and $X'X = I_k$ (k elements in β , orthogonal explanatory variables). Extend X to an orthonormal matrix $Q = [X|X_0]$ where the $n - k$ columns of X_0 span X^\perp (i.e., X_0 are orthonormal and $X_0'X = 0$). Then

$$Q'Y = Q'X\beta + Q'\epsilon \quad \Rightarrow \quad Y = \mu + \epsilon$$

where $\mu = (\mu_1, \mu_2, \dots, \mu_k, 0, \dots, 0)$. The obvious estimator (MVUE) is $\hat{Y} = (y_1, \dots, y_k, 0, \dots, 0)$ for which

$$PMSE(k) = \mathbb{E} \|Y^* - \hat{Y}\|^2 = (n + k)\sigma^2$$

Notice the contribution to $PMSE$ from each term. Those for which $\mu_j = 0$ contribute σ^2 and those for which $\mu_j \neq 0$ contribute $2\sigma^2$. Hence if μ_j is small, we obtain smaller PMSE by estimating these by zero,

$$\mathbb{E}Y_j^2 = \mu_j^2 + \sigma^2 < 2\sigma^2 \text{ if } |\mu_j| < \sigma .$$

If we knew σ^2 , we could use a *testimator* such as

$$\hat{\mu}_j = \begin{cases} y_j & \text{if } |y_j| < \tau \\ 0 & \text{otherwise.} \end{cases}$$

τ in the definition of $\hat{\mu}$ is a threshold. Draw a graph to see how this works and see the similarity to Bayesian estimators.

Normal means What's the best estimator for μ if $Y_1^n \sim N(\mu, \sigma^2 I_n)$? The problem is so easy to solve as it looks. Need to pick the threshold τ , and for that we need

- An estimate of σ^2
- A means of handling multiplicity

For example, *hard thresholding* (Donoho-Johnstone, Foster-George) or the *Bonferroni rule* say to keep only those X's for which $Y_j^2 > 2\sigma^2 \log k$. To appreciate this approach, consider the distribution of $\hat{\beta}_j$. How big must Y_j^2 be in order to convince you that $\mu_j \neq 0$?

Selection in Regression

Bias/variance trade-off. Regression calculations using the least squares estimator lead to a simple expression for $PMSE$:

$$\begin{aligned} nPMSE(X) &= E\|Y^* - X\hat{\beta}\|^2 = E\|\epsilon^* + (\mu - X\hat{\beta})\|^2 \\ &= n\sigma^2 + E\|\mu - X(X'X)^{-1}X'(\mu + \epsilon)\|^2 \\ &= n\sigma^2 + E\|H\epsilon\|^2 + \|(I - H)\mu\|^2 \\ &= \underbrace{(n + k)\sigma^2}_{\text{variance}} + \underbrace{\mu'(I - H)\mu}_{\text{bias}^2} \\ &\approx \text{variance} + \text{bias}^2, \end{aligned}$$

The matrix $H = X(X'X)^{-1}X$ is the idempotent projection matrix (*a.k.a.*, “hat matrix) from regression (so that $H = H' = H^2$). Hence,

PMSE balances variance with bias. The larger the number of variables k , the larger the variance component and the smaller the bias.

Estimating PMSE via C_p . To use such a criterion, we need an estimate that may be computed without knowing μ . To choose the “best” explanatory variables, we need an *observable* metric. Taking the expected value of the observed residual sum-of-squares (*RSS*) gives

$$\begin{aligned} E(RSS) &= E\|Y - X\hat{\beta}\|^2 \\ &= E\|(I - H)(\mu + \epsilon)\|^2 \\ &= (n - k)\sigma^2 + \mu'(I - H)\mu, \end{aligned}$$

which matches the numerator of *PMSE* *except* that the sign of k is reversed. By adding $2k$ times $\hat{\sigma}^2$ (an estimator of the pure error variance) to the observed *RSS*,

$$pmse(X) = \frac{RSS(X) + 2k\hat{\sigma}^2}{n}.$$

Of course, this estimator dodges the issue of where to get such an estimate of σ^2 . The resulting criterion for model selection among regression models is known as Mallows’s C_p statistic.

Covariance criterion Efron characterizes these criteria as covariance penalties. Write the squared error as

$$\begin{aligned} (y_i - \hat{\mu}_i)^2 &= (y_i \pm \mu_i - \hat{\mu}_i)^2 \\ &= (y_i - \mu_i)^2 + (\hat{\mu}_i - \mu_i)^2 - 2(y_i - \mu_i)(\hat{\mu}_i - \mu_i) \end{aligned}$$

Re-group the terms as the following identity

$$(y_i - \mu_i)^2 + (\hat{\mu}_i - \mu_i)^2 = (y_i - \hat{\mu}_i)^2 + 2(y_i - \mu_i)(\hat{\mu}_i - \mu_i)$$

Now take expectations, noting that Y and Y^* have mean μ but that Y^* is independent of $\hat{\mu} = \mu(Y)$,

$$\underbrace{\mathbb{E}(y_i^* - \hat{\mu}_i)^2}_{PMSE} = \underbrace{\mathbb{E}(y_i - \hat{\mu}_i)^2}_{RSS} + 2 \text{Cov}(y_i, \hat{\mu}_i)$$

This characterization leads to many other criteria, including

- Stein’s unbiased estimate of risk,

- Ye's equivalent degrees of freedom,
- Tibshirani and Knight's covariance criterion.

Summary. To recap, here are the important features:

- *PMSE* is one criterion for model selection. It measures how well (in terms of squared error) the model predicts a “new” set of observations, not the data that produce $\hat{\beta}$.
- *PMSE* is equivalent to selecting models to minimize the mean squared error of $X\hat{\beta}$ as estimator of μ .
- These selection methods use the residual sum-of-squares plus a so-called “penalty factor” which adds a term proportional to twice the number of estimated regression coefficients.

Cross Validation

Cross validation. Since the criterion is to select based on how well the fitted model predicts independent data, why not just do that? The problem is to decide how much data to use for estimating the model, and how much to “save” for validation.

Leave-one-out One method is to try to predict each observation, using the fit based on the other $n - 1$ values. The so-called cross-validation sum of squares is

$$CVSS = \sum_{i=1}^n (y_i - x_i' \hat{\beta}_{-i})^2 = \sum_i (y_i - \hat{y}_{-i})^2,$$

where the subscript “ $-i$ ” implies that the fit is obtained without using that observation. Using the result

$$\hat{\beta} - \hat{\beta}_{-i} = \frac{(X'X)^{-1} x_i e_i}{1 - h_i}, \quad e_i = y_i - x_i' \hat{\beta}, \quad h_i = H_{ii},$$

we have

$$\begin{aligned} y_i - \hat{y}_{-i} &= y_i - \hat{y}_i + \hat{y}_i - \hat{y}_{-i} \\ &= e_i + \frac{h_i e_i}{1 - h_i} \end{aligned}$$

$$= \frac{e_i}{1 - h_i}$$

implying that

$$CVSS = \sum_i \left(\frac{e_i}{1 - h_i} \right)^2 .$$

Generalized CV To make for faster calculation (especially in more complex models), an approximation known as “generalized” cross validation is used. It simply replaces h_i by the average leverage. In regression $\sum h_i = \text{tr}(H) = k$ and

$$\begin{aligned} GVSS &= \sum_i \left(\frac{e_i}{1 - \bar{h}} \right)^2 \\ &= \frac{RSS}{(1 - k/n)^2} = RSS(1 + k/n + \dots)^2 \\ &\approx RSS \left(1 + \frac{2k}{n} \right) \\ &\approx RSS + 2k\hat{\sigma}^2 , \end{aligned}$$

assuming that the fitted model is close to the true model so that RSS/n is in the right ballpark. The final approximation is the numerator in (??).

Further reading. Strategies for cross validation are studied in the article of P. Zhang (1993, *Annals of Stat*, 299-313) which suggests that you need to set aside more than used for estimation (“reverse” cross-validation). Others proposed leaving as few as 2 or 3 out at a time to improve the estimate of prediction error.

Criticism These methods are “optimistic” in the real world since new data may not come from the same population as the observed data (population drift). More narrowly, a technical criticism is that why should we want an unbiased estimator of the PMSE? The PMSE of a model that minimizes such a criterion is not unbiased.

Autoregressions and the FPE

PMSE in autoregressions. Consider choosing the “best” autoregression of the form

$$Y_t = \varphi_{p,1}Y_{t-1} + \cdots + \varphi_{p,p}Y_{t-p} + w_t = \varphi_p' Y_{p,t-1} + w_t,$$

where the vector notation uses $\varphi_p = (\varphi_{p,1}, \dots, \varphi_{p,p})'$, $Y_{p,t} = (y_t, y_{t-1}, \dots, y_{t-p+1})'$, and $\text{Var}(w_t) = \sigma^2$.

Simpler problem As usually framed, this identification problem is simpler than the choice of covariates in regression since it is carried out in a way that only considers the models hierarchically with all intervening lags. In regression, one must not only decide how many variables to add, but *which* variables to add. The hierarchical selection is from among k models, whereas the regression problem is among 2^k .

FPE criterion Akaike (1969) proposed to select the order p by choosing the model that minimizes a form of prediction mean squared error. This criterion is defined

$$FPE(p) = E(y_{t+1}^* - \hat{\varphi}_p' Y_{p,t}^*)^2,$$

where $\hat{\varphi}$ denotes the vector of parameter estimates computed from the observations y_1, y_2, \dots, y_n which are *independent* of y_1^*, y_2^*, \dots . As in the motivation for C_p , one bases the estimation on one series and uses the resulting estimates to predict an independent series which has the same statistical properties as the original data.

Properties. If indeed the process is $\text{AR}(p)$ (or lower order), then

$$\begin{aligned} FPE(p) &= \sigma^2 + E(\varphi_p - \hat{\varphi}_p)(\varphi_p - \hat{\varphi}_p)' E(Y_{t,p}^* Y_{t,p}^{*'}) \\ &= \sigma^2 + \text{tr}(\text{Var}(\hat{\varphi}_p)\Gamma_p), \end{aligned} \tag{2}$$

where $\Gamma_p = [\gamma(i-j)]_{(i,j=1,\dots,p)}$ is the $p \times p$ covariance matrix of p adjacent observations. Since

$$n \text{Var}(\hat{\varphi}_p) \approx \sigma^2 \Gamma_p^{-1},$$

when the true model is an autoregression of order p , the FPE simplifies in this case to

$$FPE(p) = \sigma^2(1 + p/n),$$

Again, this expression *assumes* that the order of the fitted model meets or exceeds the order of the true model.

A usable criterion. To obtain a computable model-selection criterion, we need an estimator for σ^2 . Again, as in regression, consider the residual sum-of-squares from the fitted model. In an autoregression of order p ,

$$E(RSS) = E \sum_{t=1}^n \hat{w}_t^2 \approx (n-p)\sigma^2,$$

assuming one is using a complete data procedure (like a Kalman filter covered later) that does not “lose” the initial p observations. Hence an approximately unbiased estimator for σ^2 is $RSS/(n-p)$. The criterion is then

$$fpe(p) = \left(\frac{RSS}{n-p} \right) (1 + p/n) = \hat{\sigma}^2 \frac{n+p}{n-p} \approx \hat{\sigma}^2 \left(1 + \frac{2p}{n} \right)$$

if we let $\hat{\sigma}^2 = \sum \hat{w}_t^2/n$ denote the biased estimate of variance often associated with maximum likelihood. This last form looks quite similar to C_p , but this similarity is somewhat artificial since we kept bias in the estimator $\hat{\sigma}^2$ to get the factor $2p$ (twice the number of fitted coefficients) to appear.

AIC You can see the origins for the AIC criterion (Definitions 2.1, 2.2) in these expressions: the idea is to penalize the estimator by adding a multiple of the number of estimated parameters to the RSS. In a Gaussian model, the RSS is proportional to the negative log-likelihood. AIC makes this connection more formal.

White/Sandwich Estimator

Variance of OLS slope The variance of the usual OLS estimator of β in the linear regression model is

$$Y = X\beta + \epsilon \quad \hat{\beta} = (X'X)^{-1}X'Y$$

for which

$$\text{Var } \hat{\beta} = (X'X)^{-1}X' \text{Var}\epsilon X(X'X)^{-1}$$

In the case of independent errors, $\text{Var}\epsilon = \sigma^2 I_n$ and the variance of the slope estimator reduces to $\text{Var}\hat{\beta} = \sigma^2(X'X)^{-1}$.

Sandwich estimator To avoid the assumption of independent errors, estimate $\text{Var}\epsilon$ using the residuals, as in

$$\text{var } \hat{\beta} = (X'X)^{-1}X'VX(X'X)^{-1}$$

where the $n \times n$ matrix V is a “direct” estimate of the error variance using the residuals, as in

$$V = \text{diag}(e_i^2)$$

This estimator provides robustness in the case of heteroscedasticity. A similar block diagonal form provides a robust estimate of variance in the case of autocorrelation.