Model building and validation

The formulas are primarily derived from *Linear Regression Analysis, Fifth Edition* by Montgomery et al. and/or the materials from the STAT 6021 course taught by Dr. Woo at the University of Virginia. Except where specially cited, it is believed that the materials are well-known equations and concepts in the public domain. If you believe otherwise, please reach out to me through my Github account so that I can correct the material...

Specifically, see chapters 10 and 11 of the textbook.

The best model

To quote the book,

> Unfortunately, as we will see in this chapter, there is no unique definition of "best." Furthermore, there are several algorithms that can be used for variable selection, and these procedures frequently specify different subsets of the candidate regressors as best.

Translation: You will need to read the textbook or other texts to study the specific issues and considerations that should inform your judgements.

Model adequacy versus validity

- Model adequacy concerns with whether the model and data meet the theoretical requirements, e.g., linearity assumption, constant variance assumption, etc.
- Model validation is about checking whether the model works well in practice, e.g., provides good predictions, works on new data not in the original estimation data set, etc.

Purposes of regression

The choice of model will also depend on the intended purpose(s) of the model. The book notes a few (see p. 337). (JG: If I understand the book correctly, they are as follows).

- elucidating the data
- prediction
- estimating parameters of the population
- determining how to get a desired response

Comparing models for building

Summary

Comparison measure	Brief description
Coefficient of multiple determination (R_p^2)	You don't want the highest value necessarily, but you want a higher value. However, you may wish to consider using the adjusted version instead.
Adjusted R^2	Higher values are better. The adjustment helps by penalizing adding regressors that don't significantly improve the model.
Residual mean square	See below.
Mallow's C_p statistic	This estimates the total standardized mean square error. You want this low.
Aikike Information Criterion (AIC)	maximing entropy of the model
BIC	These are variations on AIC

Coefficient of multiple determination

The value is

$$R_p^2 = \frac{SS_R(p)}{SS_T} = 1 - \frac{SS_{Res}(p)}{SS_T}$$

where $SS_R(p)$ is the regression sum of squares for the p-many coefficients ($p-1$ included regressors and the intercept), and SS_{Res} is the residual sum of squares for the same.

To find a "satisfactory" (to quote the book) value of R_p^2, you can follow one of the following approaches:

- Look for a convenient bend where the curve slopes off as p increases. This is a judgement call.
- Use the Atkins criteria (below).

Let

$$R_0^2 = 1 - (1 - R_{K+1}^2)(1 + d_{a,n,K})$$

where

$$d_{a,n,K} = \frac{KF_{a,K,n-K-1}}{n-K-1}.$$

Then any value R_p^2 that exceeds R_0^2 is called an R^2-adequate (α) subset. Notice that the adequacy is based on a level of α. See p. 333 of the book for additional details and the reference to Atkins 1974.

Adjusted R^2

Let

$$R_{adj,p}^2 = 1 - \left(\frac{n-1}{n-p}\right)(1 - R_p^2)$$

be the adjusted version. You can look for when this stops increasing to see if models are ceasing to add sufficient additional value.

To quote the book:

> It can be shown (Edwards [1969]....) that if s regressors are added to the model, $R_{adj,p+s}^2$ will exceed $R_{adj,p}^2$ if and only if the partial F statistic for testing the significance of the s additional regressors exceeds 1.

Residual mean square

The residual mean square is

$$MS_{Res}(p) = \frac{SS_{Res}(p)}{n-p}$$

There are a few ways to pick the adequate model:

- The one with the minimum $MS_{Res}(p)$
- The one with an $MS_{Res}(p) \approx MS_{Res}^{\text{full}}$
- One near where changing adding/removing predictors causes an increase.

The book points out that where $MS_{Res}(p)$ is minimized, R^2_{adj} will be maximized. See p. 334 for the proof.

Mallow's C_p statistic

Consider the mean square error of a single fitted value:

$$\text{MSE}(\hat{y}_i) = E[\hat{y}_i - E(y_i)]^2.$$

Note that we use $E(y_i)$ here because $\hat{y}_i$ (in linear regresion) is attempting to estimate the mean y value for observation i, not a specific observed y_i.

We can build on this (see p. 334) and define the standardized total mean squared error as

$$\Gamma_p = \frac{1}{\sigma^2}\left\{\sum_{i=1}^{n}[E(y_i) - E(\hat{y}_i)]^2 + \sum_{i=1}^{n}\text{Var}(\hat{y}_i)\right\}.$$

Assuming near-zero bias, we can estimate Γ_p as

$$\hat{\Gamma}_p = C_p = \frac{SS_{Res}(p)}{\hat{\sigma}^2} - n + 2p$$

where C_p is Mallows C_p statistic.

It can be shown that

$$E[C_p|\text{Bias} = 0] = p.$$

We can estimate $\hat{\sigma}^2 = MS_{Res}^{full}$ using the residual mean squared from the full model. This can be problematic if the full model has many unnecessary predictors. The text discusses an alternative.

Akaike Information Criterion (AIC)

The AIC is a "peanlized log-likelihood measure" that is derived from "maximizing the expected *entropy* of the model". (JG: Compare this to the concepts used in building decision trees based on entropy and/or the Gini coefficient).

It is defined (for least squares) as

$$AIC = n\ln(\frac{SS_{Res}}{n}) + 2p.$$

In the more general scenario, where L is the likelihood function of a particular model that isn't necessarily least squares, it is

$$AIC = -2\ln(L) + 2p.$$

Bayesian information criteria (BICs)

There are two popular kinds of BICs.

The Schwartz BIC

It is defined as

$$\text{BIC}_{Sch} = -2\ln(L) + p\ln(n),$$

which for least squares is

$$\text{BIC}_{Sch} = n\ln\left(\frac{SS_{Res}}{n}\right) + p\ln(n).$$

The Sawa BIC

We don't define it here.

PRESS statistics

Let

$$\text{PRESS}_p = \sum_{i=1}^{n}[y_i - \hat{y}_{(i)}]^2 = \sum_{i=1}^{n}\left(\frac{e_i}{1-h_{ii}}\right)^2,$$

where $e_{(i)} = y_i - \hat{y}_{(i)}$ and $y_{(i)}$ is the value of the i predicted response if you fit a model without that data point. PRESS stands for "prediction error sum of squares" and is due to Allen. See p. 151 of the textbook for these details and more.

You could potentially look for models where this is small. However, it's not really as simple as that.

Validating a selected model

Types of performance to validate

The following are worth evaluating:

- Effective prediction of values ranges inside and outside values used during build
- Effective prediction of new scenarios outside original data set specification
- Stability of coefficients across different kinds of fittings
- Reasonable signs (positive versus negative) and magnitudes of coefficients with known meanings
- Avoidance of multicollinearity

It's best to test as many as possible because you don't always know how a model will be used.

Techniques

The textbook notes three validation techniques:

- Evaluating the model using analytical methods, background knowledge, and theory.
- Evaluating the prediction on new data
- Subsetting in which you train it on a subset of your data and validate it on the remaining data.

(JG: Note that the book doesn't prefer the term "cross-validation", but I do.)

Using data splitting

You can divide your data up and fit the model with one subset and validate it on the other. The book calls this "data splitting" and calls the two sets of data the "estimation data" (used to fit the model) and the "prediction data" (used to evaluate the effectiveness on new data).

Using the PRESS statistic

The book recommends using the PRESS statistic on the prediction set to calculate a "prediction" R^2:

$$R^2_{Prediction} = 1 - \frac{PRESS}{SS_T}.$$

An absolute value close to 1 is good.

Choosing how to form subsets

Some methods include

- Breaking the data based on time sequence (earlier versus later)
- Breaking the data up arbitrarily (although you may not necessarily pick the best division possible)
- Using the DUPLEX algorithm (see below)

The DUPLEX algorithm consists of transforming the data so that predictors are orthonormalized and that the data forms a rough sphere and then picking off the pair of points furthest apart to be split into each subset. The process of finding pairs is repeated until the subsets are fully formed.

Evaluating subsets

However chosen, the subsets can be checked by looking at a value that (effectively) measures the relative volumes of the subsets, in terms of how data is spread out:

$$\left(\frac{|\mathbf{X}'_E\mathbf{X}_E|}{|\mathbf{X}'_P\mathbf{X}_P|}\right)^{1/p},$$

where $|\mathbf{K}|$ is the determinant and the matrices are the points in the estimation and prediction subsets.

Caveats

- The book says that Snee recommends at least $2p + 25$ observations.
- The book suggests eliminating near neighbors in the full data set.
- The book notes that you can potentially refit using the full data set and compare it to the reduced subset fit. You can potentially decrease standard errors this way.
- The book discusses double-cross validation.

Data from planned experiments

Designed experiments let you exercise some congtrol over the data collected. Because of this, you can attempt to pick predictors in a way that avoid multicollinearity, attempt to collect data in case you need higher order terms, etc.

Errors of point estimators

The material in this section is sourced from *Mathematical Statistics with Applications, ed. 6* by Wackerly et al. See the book for additional information.

The bias of an estimate is

$$B(\hat{\theta}) = E(\hat{\theta}) - \theta$$

The mean squared error for an estimate is

$$\text{MSE}(\hat{\theta}) = E((\hat{\theta} - \theta)^2).$$

It can be shown that

$$\text{MSE}(\hat{\theta}) = \text{Var}(\hat{\theta}) + (B(\hat{\theta}))^2.$$

(Caveat by JG: The text MSE is also used to represent the related concept of "mean square for error", discussed in the notebooks for single and multiple linear regression. Mean squared errors of point estimators is concerned with the theoretical values (assuming a known population), while mean square for error is normally calculated from samples.)

Ancillary Material

The alias matrix

Let the full model be

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

Then in reducing the model to p many predictors by removing r many predictors, the $\mathbf{X}$ can be partitioned into two matrices, $\mathbf{X}_p$ and $\mathbf{X}_r$ such that the full model can be re-written as

$$\mathbf{y} = \mathbf{X}_p\boldsymbol{\beta}_p + \mathbf{X}_r\boldsymbol{\beta}_r + \boldsymbol{\varepsilon}.$$

Then

$$\mathbf{A} = (\mathbf{X}_p'\mathbf{X}_p)^{-1}\mathbf{X}_p'\mathbf{X}_r$$

is the alias matrix. The textbook describes its use, which we shall not repeat here.