Notes for 751-752
Sections 17-18

Martin Lindquist*

November 8, 2017

*Thanks to Brian Caffo and Ingo Ruczinski for supplying their notes, upon which much of this document is based.

16 Effects of Departures from Assumptions

16.1 Under and overfitting

In linear models, we can characterize different forms of model misspecification. For this chapter consider the following:

Model 1: $\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \boldsymbol{\epsilon}$

Model 2: $\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\epsilon}$

where the $\boldsymbol{\epsilon}$ are assumed iid normals with variance σ^2. We further differentiate between the assumed model and the true model. If we assume Model 1 and Model 2 is true, we have underfit the model (i.e., omitted variables that were necessary). In contrast, if we assume Model 2 and Model 1 is true, we have overfit the model (i.e., included variables that were unnecessary).

16.1.1 Impact of underfitting

Let us begin by considering underfitting the model. That is we errantly act as if Model 1 is true, but in fact Model 2 is true. Such a situation would arise if there were unmeasured or unknown confounders. Then consider the bias of our estimate of $\boldsymbol{\beta}_1$.

$$\begin{aligned} E[\hat{\boldsymbol{\beta}}_1] &= E[(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{y}] \\ &= (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'(\mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2) \\ &= \boldsymbol{\beta}_1 + (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_2\boldsymbol{\beta}_2. \end{aligned}$$

Thus, $(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_2\boldsymbol{\beta}_2$ is the bias in estimating $\boldsymbol{\beta}_1$. Notice that there is no bias if $\mathbf{X}_1'\mathbf{X}_2 = \mathbf{0}$. Consider the case where both design matrices are centered. Then $\frac{1}{n-1}\mathbf{X}_1'\mathbf{X}_2$ is the empirical variance/covariance matrix between the columns of $\mathbf{X}_1$ and $\mathbf{X}_2$. Thus, if our omitted variables are uncorrelated with our included variables, then no bias exists. One way to try to force this in practice is to randomize the levels of the variables in $\mathbf{X}_1$. Then, the empirical correlation will be low with high probability. This is very commonly done when $\mathbf{X}_1$ contains only a single treatment indicator.

Example: 16.1 Suppose we fit

$$\mathbf{y} = \beta_0 + \beta_1\mathbf{x} + \boldsymbol{\varepsilon},$$

when the true model is

$$\mathbf{y} = \beta_0 + \beta_1\mathbf{x} + \beta_2\mathbf{x}^2 + \boldsymbol{\varepsilon}.$$

In this situation

$$(\mathbf{X}'\mathbf{X})^{-1} = \frac{1}{\sum(x_i - \bar{x})^2} \begin{pmatrix} \sum x_i^2/n & -\bar{x} \\ -\bar{x} & 1 \end{pmatrix}$$

and

$$\mathbf{X}'\mathbf{Z} = \begin{pmatrix} 1 & \dots & 1 \\ x_1 & \dots & x_n \end{pmatrix} \begin{pmatrix} x_1^2 \\ \vdots \\ x_n^2 \end{pmatrix} = \begin{pmatrix} \sum x_i^2 \\ \sum x_i^3 \end{pmatrix}.$$

Therefore the bias in $\hat{\boldsymbol{\beta}}$ is

$$(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}\beta_2 = \frac{\beta_2}{\sum(x_i - \bar{x})^2} \begin{pmatrix} (\sum x_i^2)^2/n - \bar{x}\sum x_i^3 \\ -\bar{x}\sum x_i^2 + \sum x_i^3 \end{pmatrix}.$$

Example: 16.2 Suppose we fit $y_{ij} = \mu_i + \varepsilon_{ij}$, when the true model is $y_{ij} = \mu_i + \eta z_{ij} + \varepsilon_{ij}$, with $i = 1, 2,\ j = 1, \dots, n_i$. In other words, we are comparing two groups, but ignore the covariate z. In matrix form the true model is $\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta} + \mathbf{X}_2\boldsymbol{\eta} + \boldsymbol{\varepsilon}$, or

$$\begin{pmatrix} y_{11} \\ \vdots \\ y_{1n_1} \\ y_{21} \\ \vdots \\ y_{2n_2} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 0 & 1 \\ \vdots & \vdots \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} + \begin{pmatrix} z_{11} \\ \dots \\ z_{1n_1} \\ z_{21} \\ \dots \\ z_{2n_2} \end{pmatrix} \eta + \begin{pmatrix} \varepsilon_{11} \\ \dots \\ \varepsilon_{1n_1} \\ \varepsilon_{21} \\ \dots \\ \varepsilon_{2n_2} \end{pmatrix}.$$

Then the bias in $(\hat{\mu}_1, \hat{\mu}_2)'$ is given by $(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_2\eta = \begin{pmatrix} \bar{z}_1 \\ \bar{z}_2 \end{pmatrix} \eta$. Hence, the group comparison given by $\hat{\mu}_1 - \hat{\mu}_2$ is unbiased if $\bar{z}_1 = \bar{z}_2$.

This example illustrates the effect of randomization. Suppose we randomly assign experimental units (for example patients) to the two groups. Then $\bar{z}_1 \approx \bar{z}_2$ for any covariate z, as long as groups are fairly large. Thus, randomization controls for bias due to unfitted covariates.

Our theoretical standard errors for the $\hat{\boldsymbol{\beta}}_1$ are still correct in that

$$\text{Var}(\hat{\boldsymbol{\beta}}_1) = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\sigma^2.$$

However, we still have to estimate σ^2.

We can also see the impact of underfitting on the bias of residual variance estimation.

$$\begin{aligned} E[(n-p_1)s^2] &= E[\mathbf{y}'(\mathbf{I}-\mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{y}] \\ &= (\mathbf{X}_1\boldsymbol{\beta}_1+\mathbf{X}_2\boldsymbol{\beta}_2)'\{\mathbf{I}-\mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\}(\mathbf{X}_1\boldsymbol{\beta}_1+\mathbf{X}_2\boldsymbol{\beta}_2) \\ &+ \text{trace}[\{\mathbf{I}-\mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\}\sigma^2] \\ &= \boldsymbol{\beta}_2'\mathbf{X}_2'\{\mathbf{I}-\mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\}\mathbf{X}_2\boldsymbol{\beta}_2+(n-p_1)\sigma^2 \end{aligned}$$

Therefore s^2 is biased upward. It makes sense that we would tend to overestimate the residual variance if we've attributed to the error structure variation that is actually structured and due to unmodeled systematic variation.

16.1.2 Impact of overfitting

Consider now fitting Model 2 when, in fact, Model 1 is true. There is no bias in our estimate of $\boldsymbol{\beta}_1$, since we have fit the correct model; it's just $\boldsymbol{\beta}_2 = \mathbf{0}$.

For $\text{var}(\hat{\boldsymbol{\beta}}_1)$ we have

$$\text{var}(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1} = \sigma^2\begin{pmatrix} \mathbf{X}_1'\mathbf{X}_1 & \mathbf{X}_1'\mathbf{X}_2 \\ \mathbf{X}_2'\mathbf{X}_1 & \mathbf{X}_2'\mathbf{X}_2 \end{pmatrix}^{-1} = \sigma^2\begin{pmatrix} (\mathbf{X}_1'\mathbf{X}_1)^{-1}+\mathbf{F}\mathbf{E}^{-1}\mathbf{F}' & -\mathbf{F}\mathbf{E}^{-1} \\ -\mathbf{E}^{-1}\mathbf{F}' & \mathbf{E}^{-1} \end{pmatrix},$$

where

$$\mathbf{F} = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_2,$$

and

$$\mathbf{E} = \mathbf{X}_2'\mathbf{X}_2 - \mathbf{X}_2'\mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_2 = \mathbf{X}_2'(\mathbf{I}-\mathbf{P}_{\mathcal{R}(\mathbf{X}_1)})\mathbf{X}_2.$$

Therefore,

$$\text{var}(\hat{\boldsymbol{\beta}}_1) = \sigma^2[(\mathbf{X}_1'\mathbf{X}_1)^{-1}+\mathbf{F}\mathbf{E}^{-1}\mathbf{F}'],$$

compared with $\sigma^2(\mathbf{X}_1'\mathbf{X}_1)^{-1}$ which would result from fitting the true model $E[\mathbf{y}] = \mathbf{X}_1\boldsymbol{\beta}_1$. In the above, $\mathbf{F}\mathbf{E}^{-1}\mathbf{F}'$ is positive definite unless $\mathbf{X}_1'\mathbf{X}_2 = \mathbf{0}$.

Therefore, the variance assuming Model 2 will always be greater than the variance assuming Model 1. Note that at no point did we utilize which model was actually true. Thus we arrive at an essential point, adding more regressors into a linear model necessarily increases the standard error of the ones already included. This is called "variation inflation". The estimated variances need not go up, since σ^2 will go down as we include variables. However, the central point is that one concern with including unnecessary regressors is inflating a component of the standard error needlessly.

If we fit Model 2, but Model 1 is correct, then our variance estimate is unbiased. We've fit the correct model, we just allowed the possibility that $\boldsymbol{\beta}_2$ was non-zero when it is exactly zero. Therefore s^2 is unbiased for σ^2. However, recall too that

$$\frac{(n-p_1-p_2)s_2^2}{\sigma^2} \sim \chi^2_{n-p_1-p_2},$$

where the subscript 2 on s_2^2 is used to denote the fitting where Model 2 was assumed. Similarly,

$$\frac{(n-p_1)s_1^2}{\sigma^2} \sim \chi^2_{n-p_1},$$

where s_1^2 is the variance assuming Model 1 is true. Using the fact that the variance of a Chi squared is twice the degrees of freedom, we get that

$$\frac{\text{Var}(s_2^2)}{\text{Var}(s_1^2)} = \frac{(n-p_1)^2}{(n-p_1-p_2)^2}.$$

Thus, despite both estimates being unbiased, the variance of the estimated variance under Model 2 is higher.

16.1.3 Summary

The lesson in this subsection is that overfitting does not introduce bias into regression coefficient estimates, but it does inflate their variances. In comparison to underfitting, we have the following:

	Effect of Underfitting	Effect of Overfitting
$\hat{\boldsymbol{\beta}}$	biased	unbiased
$\hat{\mathbf{y}}$	biased	unbiased
s^2	biased upward	unbiased
$\text{var}(\hat{\boldsymbol{\beta}})$	still $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$	$>$ than necessary

16.2 Effects of a Mis-Specified Covariance Matrix

Assume that we have specified $E[\mathbf{y}] = \mathbf{X}\boldsymbol{\beta}$ correctly, but suppose that $\text{var}(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{V}$, when we assume that $\text{var}(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}$.

In the full rank case the parameter estimates are still unbiased, but

$$\text{var}(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}.$$

Also, in most cases s^2 is biased, since

$$E[s^2] = \frac{\sigma^2}{n-p}\text{tr}[\mathbf{V}(\mathbf{I}-\mathbf{H})].$$

Example: 16.3 The effect of non-constant variance in the two-sample t-test:

Assume the model $y_{ij} = \mu_i + \varepsilon_{ij}$, $\text{var}(\varepsilon_{ij}) = \sigma_i^2$, $i = 1, 2$, $j = 1, \ldots, n_i$. The usual t-statistic for forming a confidence interval for $\mu_1 - \mu_2$ is

$$T = \frac{\bar{y}_1 - \bar{y}_2 - (\mu_1 - \mu_2)}{S(n_1^{-1} + n_2^{-1})^{1/2}},$$

where

$$s^2 = \frac{1}{n-2} \sum_i \sum_j (y_{ij} - \bar{y}_i)^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n-2}.$$

Here, $n = n_1 + n_2$, and s_i^2 is the sample variance in the ith group. Now, if $\sigma_1^2 = \sigma_2^2$ and ε_{ij} is normally distributed, then $T \sim t_{n-2} \approx N(0, 1)$ for large n. However, assume that $\sigma_1^2 \neq \sigma_2^2$. Then, heuristically, $s^2 \approx \frac{1}{n}(n_1\sigma_1^2 + n_2\sigma_2^2)$ for large n, and T is approximately normally distributed with mean 0 and

$$\text{var}(T) \approx \frac{\text{var}(\bar{y}_1 - \bar{y}_2)}{\frac{1}{n}(n_1\sigma_1^2 + n_2\sigma_2^2)(n_1^{-1} + n_2^{-1})} = \frac{n_1^{-1}\sigma_1^2 + n_2^{-1}\sigma_2^2}{\frac{1}{n}(n_1\sigma_1^2 + n_2\sigma_2^2)(n_1^{-1} + n_2^{-1})} = \frac{\frac{\sigma_1^2}{\sigma_2^2} + \frac{n_1}{n_2}}{\frac{n_1}{n_2}\frac{\sigma_1^2}{\sigma_2^2} + 1}.$$

In other words, $\text{var}(T) \approx 1$ and therefore $T \approx N(0, 1)$ for large n if either $\sigma_1^2 = \sigma_2^2$ (i.e. the equal variance assumption holds), or if $n_1 = n_2$ (i.e. the sample sizes are equal, regardless of equality of variances).

Example: 16.4 (cont.) Recall the 95% confidence interval for $\mu_1 - \mu_2$:

$$\text{CI} = [\ \bar{y}_1 - \bar{y}_2 - t_{n-2}^{.025} s(n_1^{-1} + n_2^{-1})^{1/2}\ ,\ \bar{y}_1 - \bar{y}_2 + t_{n-2}^{.025} s(n_1^{-1} + n_2^{-1})^{1/2}\ ].$$

The error rate of this confidence interval is

$$P(\mu_1 - \mu_2 \notin \text{CI}) = P(|T| > t_{n-2}^{.025}) \approx P(|N(0, v)| > t_{n-2}^{.025}),$$

where $v = (\frac{\sigma_1^2}{\sigma_2^2} + \frac{n_1}{n_2})/(\frac{n_1}{n_2}\frac{\sigma_1^2}{\sigma_2^2} + 1)$.

Some values of the error rate based on the above normal approximation are given in the table below. The error rate does not deviate too far from the nominal value of 0.05 unless both the sample sizes and the variances differ substantially between groups.

$\downarrow$	σ_1^2/σ_2^2						
n_1/n_2	$\frac{1}{8}$	$\frac{1}{4}$	$\frac{1}{2}$	1	2	4	8
$\frac{1}{2}$	0.011	0.016	0.028	0.050	0.080	0.110	0.133
1	0.050	0.050	0.050	0.050	0.050	0.050	0.050
2	0.133	0.110	0.080	0.050	0.028	0.016	0.011
4	0.237	0.179	0.110	0.050	0.016	0.004	0.001
8	0.331	0.237	0.133	0.050	0.011	0.001	0.000

16.3 Effects of Non-normality

Finally, let us suppose we have correctly specified the model:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \;\; E[\boldsymbol{\varepsilon}] = \mathbf{0}, \;\; \mathrm{var}(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I},$$

but suppose that $\boldsymbol{\varepsilon}$ is not necessarily multivariate normal.

We have seen previously that $\hat{\boldsymbol{\beta}}$ is unbiased, and $\mathrm{var}(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$, without requiring any distributional assumptions. Thus, normality is not required to fit a linear model. However, normality of the coefficient estimates $\hat{\boldsymbol{\beta}}$ is needed to compute confidence intervals and perform tests. As $\hat{\boldsymbol{\beta}}$ is a weighted sum of $\mathbf{y}$, the Central Limit Theorem guarantees that it will be normally distributed if the sample size is large enough. Thus, tests and confidence intervals can be based on the associated t-statistic in these settings. However, in many settings, bootstrap procedures may be more appropriate.

17 Model selection

We often have data on a large number of explanatory variables and wish to build a model using some subset of them. The problem becomes choosing between different competing linear models. If the model is too small we 'underfit' the data. This leads to poor predictions and high bias, but low variance. If the model is too big we 'overfit' the data. This leads to poor predictions and high variance, but low bias. When the model is 'just right', we balance bias and variance to get good predictions.

One approach towards model selection is to consider all possible subsets of the pool of explanatory variables and find the 'best' model according to some predetermined criteria. Another approach is to use a search algorithm to find the 'best' model. The latter is usually more efficient when the number of potential variables is large. In both cases different criteria may be used to select what constitutes the best model. Popular choices include Adjusted R^2, Mallow's C_p, AIC and BIC. These criteria assign scores to each model and allow us to choose the model with the best score.

17.1 Model Selection Criteria

Before we start, recall that we can partition the data into sums of squares:

$$\text{SST}_p = \text{SSE}_p + \text{SSR}_p$$

where

$$\begin{aligned} \text{SST}_p &= ||\mathbf{y} - \bar{y}\mathbf{J}_n||^2 = \mathbf{y}'(\mathbf{I} - \mathbf{H_J})\mathbf{y} \\ \text{SSE}_p &= ||\mathbf{y} - \hat{\mathbf{y}}||^2 = \mathbf{y}'(\mathbf{I} - \mathbf{H_X})\mathbf{y} \\ \text{SSR}_p &= ||\hat{\mathbf{y}} - \mathbf{J}_n\bar{y}||^2 = \mathbf{y}'(\mathbf{H_X} - \mathbf{H_J})\mathbf{y} \end{aligned}$$

Here the subscript p refers to the fact that the model includes p parameters.

We can use these values to compute the coefficient of multiple determination, which is given by

$$R_p^2 = \frac{\text{SSR}_p}{\text{SST}_p} = 1 - \frac{\text{SSE}_p}{\text{SST}_p}.$$

This represents the proportion of the total variability explained by our model. It is guaranteed to take values between 0 and 1. High values imply that the explanatory variables are useful in explaining the response and low values imply that the explanatory variables are not useful. However, since R_p^2 increases with the size of the model, it is not a good criterion for variable selection. It would always choose to include all variables. Alternative approaches that penalize for unnecessary model complexity are needed.

17.1.1 Adjusted R_p^2

One such approach is the adjusted coefficient of multiple determination, which uses the mean squares instead of the sums of square, i.e.

$$R_{a,p}^2 = 1 - \frac{\text{MSE}_p}{\text{MST}_p} = 1 - \left(\frac{n-1}{n-p}\right)\frac{\text{SSE}_p}{\text{SST}_p}.$$

Since the term includes the number of model parameters, p, it penalizes for model complexity.

17.1.2 Mallow's C_p

Mallow's C_p is a criteria to assess fits when models with different numbers of parameters are being compared. It can be expressed as:

$$C_p = \frac{SSE_p}{s^2} - n + 2p.$$

Here the MSE of the full model is used to estimate s^2. If the model includes all important variables, then

$$E(SSE_p) = (n-p)\sigma^2.$$

If s^2 provides a good estimate of σ^2 then

$$E(C_p) \approx \frac{(n-p)\sigma^2}{\sigma^2} - n + 2p = p.$$

Values close to the corresponding p indicate a good model. The smaller the value, the better the model fits.

17.1.3 Information Criteria

Many methods are based on combining a term based on the log-likelihood with one based on penalizing model complexity. This provides a means to balance model fit with model complexity when assessing the best fitting models, providing a way to penalize unnecessarily complicated models.

To illustrate, let $f(\mathbf{y}|\mathbf{X}, \boldsymbol{\beta})$ be the density of the response $\mathbf{y}$. For a sample of n observations $(\mathbf{x}_i, y_i)$, $i = 1, \ldots n$, the log-likelihood is given by

$$\log(L) = \sum_{i=1}^{n} log(f(y_i|\mathbf{x}_i, \boldsymbol{\beta}).$$

Let $\ell(\hat{\beta}_p)$ be the log-likelihood evaluated at the MLE under the model with p parameters. Standard information criteria are of the form:

$$-2\ell(\hat{\boldsymbol{\beta}}) - \phi(n, p).$$

Here the first term represents model fit, and the second a penalized model complexity. In the linear model setting $\ell(\hat{\boldsymbol{\beta}}) \propto -n \log(SEE_p/n)$.

Akaike's Information Criterion (AIC) tries to balance the conflicting demands of model accuracy and parsimony. For the linear model with Guassianity assumption it can be expressed as:

$$AIC_p = n \log(SSE/n) + 2p.$$

here low values indicate a better model.

Several modifications of AIC have been suggested. For example, the Bayesian Information Criterion (BIC) is defined as:

$$BIC_p = n \log(SSE/n) + \log(n)p.$$

Again, low values indicate a better model.

The difference between AIC and BIC lies in the severity of the penalty. The penalty is larger for BIC when $n > 8$. Hence, BIC tends to favor more parsimonious models compared to AIC which has a tendency to overfit (i.e., include too many explanatory variables).

17.1.4 PRESS

The prediction sum of squares (PRESS) criterion measures how well the fitted values for a subset model can predict the observed response. The PRESS statistic is defined as:

$$PRESS_p = \sum_{i=1}^{n} (y_i - \hat{y}_{(i),i})^2$$

Thus, the PRESS statistic is the sum of squared deleted residuals.

The model with the smallest PRESS statistic is considered 'best'. Leaving one item out at a time is known as leave-one-out cross-validation. This allows us to predict the performance of the model on holdout data.

17.2 Methods for Model Selection

17.2.1 Exhaustive Search

One approach is to consider all possible subsets of the pool of explanatory variables and find the 'best' model according to of the criteria outlined above. However, if have 15 predictors there are 2^{15} different models (even before considering interactions, transformations, etc.).

17.2.2 Step-wise Methods

When the number of explanatory variables is large it is not feasible to fit all possible models. Instead, it is more efficient to use a search algorithm to find the best model. A number of such algorithms exist, including forward selection, backward elimination and stepwise regression.

Let, us begin by setting up the problem. Assume we are choosing from a set of P possible explanatory variables v_k, $k = 1, ..P$. In each algorithm our goal is to find the subset of v_k that best balances model fit and parsimony. We discuss each algorithm in detail.

Forward Selection

1. Fit the P simple linear regression models:

$$y_i = \beta_0 + \beta_1 v_{ki} + \epsilon_i \;\; k = 1, \dots P.$$

2. Set $x_1 = v_k$, where v_k is the variable that has the most significant regression coefficient (i.e. the smallest p-value) If no variable is significant (e.g., none of the p-values are smaller than a preset significance level α) the algorithm stops.

3. Lock in the variable x_1, and repeat the procedure with models that include two explanatory variables:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 v_{ki} + \epsilon_i, \;\; k = 1, \dots P, \;\; v_k \neq x_1.$$

   Set $x_2 = v_k$, where v_k is the variable that has the most significant regression coefficient. If no variable is significant stop the algorithm.

4. Continue the procedure until no remaining v_k generate a p-value that is smaller than the preset significance level α.

The criteria for choosing whether to include a new variable can vary. As an alternative to using p-values one can instead use a criteria such as the AIC. In each step choose the variable whose inclusion lowers the AIC the most. If no variables lower the AIC than stop the algorithm.

Backward Elimination

1. Start by fitting a model that includes all possible variables:

$$y_i = \beta_0 + \beta_1 v_{1i} + \cdots + \beta_P v_{Pi} + \epsilon_i.$$

2. Find the variable v_k which has the least significant regression coefficient (i.e. the largest p-value). If its p-value is smaller than some preset significance level, stop the algorithm, otherwise drop the variable.

3. Next, fit the largest model excluding the dropped v_k. Find the variable which has the least significant regression coefficient. If its p-value is smaller than some preset significance level, stop the algorithm, otherwise drop the variable.

4. Continue until the algorithm stops.

Alternatively, AIC or BIC can be used as a criteria for determining whether to drop variables. Start with a full model. In each step choose the variable whose exclusion lowers the AIC the most. If the exclusion of any variable does not lower the AIC than stop the algorithm.

Stepwise Regression

1. Start in the same manner as in forward selection and add the most significant variable from a series of P simple linear regressions.

2. Once a new variable has been included in the model, check other variables already included in the model for their partial significance. Remove the least significant explanatory variable whose p-value is greater than the preset significance level.

3. Continue until no variables can be added and none removed, according to the specified criteria.

Again, note that AIC can be used instead of p-values.

17.3 Shrinkage Methods

The subset selection procedure is a discrete process, as individual variables are either in or out. This method can have high variance in that a different dataset from the same source can result in a totally different model. Shrinkage methods allow a variable to be partly included in the model. That is, the variable is included but with a shrunken co-efficient. Here popular methods include ridge regression, the Lasso, or the Elastic Net.