Chapter 9: Model Building

• With confirmatory observational studies, the goal is to determine whether (or how) the response is related to one or more particular (pre-specified) explanatory variables.

• **Exploratory** observational studies are done when we have little previous knowledge of exactly which explanatory variables are related to the response.

• We may have a large list of **potentially** useful predictor variables for our model.

• Variable selection procedures can help us “screen out” unimportant predictors and build a useful model.

**First steps:** Often involve plots.

•

•

•

• Once a reasonable set of potential predictors is identified, formal model selection is begun.

• If the set of predictors is large (more than 20 or so), we may use stepwise procedures to reduce the number of variables under consideration.

  **Forward Stepwise Regression**

• A procedure for adding (or deleting) one variable **at a time** to a model.
• Suppose we have $K$ potential predictors. Steps:

**Note:** We should choose $\alpha$-to-enter to be somewhat smaller than $\alpha$-to-remove. Book example:

• “Forward Selection,” “Backward Elimination,” and “Backward Stepwise Regression” are similar procedures – see page 368 for details about these.
• Once we reduce the set of potential predictors to a reasonable number, we can examine all possible models and choose the “best” model(s) based on some criterion.

Possible criteria:
(1) Choose the model with the largest adjusted $R^2$:

Note: This is equivalent to choosing the model with the smallest MSE.

• Note that if irrelevant variables are added to the model, $p$ increases and so

• Thus $R^2_a$ penalizes a model that is

(2) Choose the model with the smallest Akaike Information Criterion (AIC): With the normal-error model,

• The first two terms represent $-2 \ln L$ (where $L =$ maximized likelihood function) for the normal model.
• Like $R^2_a$, using AIC as a criterion favors models with small SSE, but penalizes models with too many variables (large $p$).
Choose model with the smallest Schwarz Bayesian Criterion (SBC), also known as the Bayesian Information Criterion (BIC).

BIC is similar to AIC, but for $n \geq 8$, the BIC “penalty term” is more severe.

Choose model using Mallows’ $C_p$:

Measures the bias in the regression model, relative to the “full” model having all the candidate predictors.

If the model is unbiased, meaning

then

Goals: (i) Choose candidate model for which $C_p$ is relatively small. (ii) Choose candidate model for which $C_p \approx p$ (= the number of parameters in that candidate model.)

Criteria (1)-(4) may yield different “best” models. Our goal is to find a model that balances (i) A good fit to the data (ii) Low bias (iii) Parsimony (less complexity)

All else being equal, a simpler model is often easier to interpret and work with.
Model Validation

• It is often desired to check our chosen model’s predictive ability with “independent” data.
• This could be done through:

(1) Collecting new data (typically impractical)
(2) Data splitting (cross-validation)

• We measure the predictive ability with the mean-squared prediction error:

• MSPR should be “close” to MSE from the training-set model.

Note: Data splitting is most useful with large data sets.
Note: The training set should be at least as big as the validation set.
(3) \textit{n-fold Cross-Validation}

\begin{itemize}
    \item Can be used for smaller data sets.
    \item For each observation \( i = 1, \ldots, n \), we delete the \( i \)-th observation. Fit the model with the other \( n-1 \) observations, and use fitted model to predict the \( i \)-th response. Let \( \hat{Y}_{i(i)} \) be this predicted value.
    \item Do this for all \( n \) observations, and add the squared prediction errors: Prediction Sum of Squares (PRESS) is:
\end{itemize}

\begin{itemize}
    \item If PRESS is only slightly larger than model SSE, then our model has good predictive ability.
\end{itemize}

Example (Surgical Unit Final Model):

\textbf{Diagnostic Measures}

To check for the proper functional form for a predictor variable, we could use:

\textbf{Plots of residuals against each individual predictor:}

\begin{itemize}
    \item A clear curved pattern may suggest the predictor should enter the model in a curvilinear manner.
\end{itemize}
Added-variable (Partial Regression) Plots:
• For any predictor $X_j$:

What to Look For:
• Flat Pattern with near zero slope:

• Linear Pattern with nonzero slope:

• Curved Pattern with nonzero slope:

Example (Life insurance data):

Example (Bodyfat data):
Outliers and Influential Observations

• Outliers are individual observations that are in some way separated from the bulk of the data set.
• In regression, we may have:
  (1) Outliers in $Y$ value
  (2) Outliers in $X$ value(s)
  (3) Outliers in both $Y$ and $X$ value(s)

SLR example:

• Which point will have the most influence on the regression line?

• Outliers are often easily seen with a scatterplot in SLR.

• In multiple regression, we rely on complex diagnostics.
Detecting Outliers in Y: Studentized Residuals

• The residuals, are measured in the same units as the response.

• To obtain a unit-free residual, we divide by the standard error of $e_i$:

This is called the **internally studentized residual** for the $i$-th observation.

**Rule of Thumb**: An observation with $|r_i| > 2.5$ may be considered an outlier (in $Y$).

**Note**: An **externally studentized residual** involves the MSE calculated with the $i$-th observation deleted.

• Here, a formal t-test allows us to declare an observation an outlier if its externally studentized residual
Detecting Outliers in $X$

- The diagonal elements $h_{ii}$ of the hat matrix (also called the leverage values) measure how far each observation is from the center of the $X$ space.

Note:

- If a leverage value $h_{ii}$ is large, this means the $i$-th observation may potentially have a large influence on the fitted regression equation (but it is not always the case).

Note:

Recall:

**Rule of Thumb**: The $i$-th observation is a high-leverage point if its
Detecting Influential Observations

- An observation is influential if its exclusion (or inclusion) from the analysis causes major changes in the fit of the regression function.

Picture:

- We focus on two main measures of influence.
- Both measure (for each $i = 1, \ldots, n$) the difference between the fitted line with observation $i$ included and the fitted line with observation $i$ deleted.

DFFITS:

Cook’s Distance:

Rules of Thumb: The $i$-th observation may be influential if
Note: DFBETAS is another measure that reveals the influence of an observation on the estimation of each regression coefficient.

Example 1 (Bodyfat data, 3 predictors):

Example 2 (Surgical unit data, 4 predictors):

- Handling outliers and influential points is quite subjective.
- Analyst should closely examine observation(s) in question before excluding them from the analysis.
- If they are truly representative of the relevant population, better to leave them in the data set.
- Advanced methods (e.g., ridge regression) can reduce influence of unusual observations without deleting them.

- A drawback of the single-deletion detection methods studied here: What if a pair of points is influential?
- These methods may not detect the points.

Picture: