Final Review

2. Simple Linear Regression

Estimating $\boldsymbol{\beta}$:
Our goal is to model the mean of a random variable Y using a quantitative variable X. If the relationship between them is linear, then we may model the mean for $i = 1, 2, \ldots, n$ as:

$$Y_i = E[Y|X = x_i] + e_i = \beta_0 + \beta_1 x_i + e_i$$

where e_i are random errors in Y_i such that $E[e_i] = 0$.

The values β_0 and β_1 are considered parameters which describe the population, or the "true" relationship between X and Y. In practice, these values are unknown, and estimated by data. The least squares estimate of the intercept is given by

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

and the slope by

$$\hat{\beta}_1 = \frac{SXY}{SXX}.$$

The first equation implies that the least squares line always goes through the point $(\bar{x}, \bar{y})$, and the second that slope has the same sign as correlation.

In matrix notation, we denote the design matrix X as:

$$\mathbf{X} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}$$

and the response vector $Y = [y_1\, y_2\, \ldots\, y_n]'$. Then the least squares estimate of the parameter vector $\boldsymbol{\beta} = [\beta_0\, \beta_1]'$ is $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$.

From Estimation to Inference:
We use the estimated model calculated from the data to make predictions about what the values of y will be. The predicted value of y is found by $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$. The difference between the actual value of y_i and its predicted value is called the residual $\hat{e}_i$; that is,

$$\hat{e}_i = y_i - \hat{y}_i.$$

Preliminary assumptions necessary in order to make inferences about the regression model using the methodology in this chapter are:

1. **L**inear relationship between x and y; that is, $Y_i = \beta_0 + \beta_1 x_i + e_i$, or $E[Y|X = x_i] = \beta_0 + \beta_1 x_i$.
2. **I**ndependence of the errors
3. **N**ormally distributed errrors with mean 0

4. **E**rrors have constant variance σ^2.

In matrix notation, the independence and constant variance assumptions imply that the covariance matrix of the errors is $\sigma^2\mathbf{I}$.

An unbiased estimate of the variance of the errors is given by:

$$s^2 = MSE = \frac{RSS}{n-2} = \frac{1}{n-2}\sum_{i=1}^{n} \hat{e}_i^2$$

To make inferences about a parameter, we usually find the distribution of the statistic used to estimate said parameter. In the case of the least squares estimates of the slope and the intercept, we note that they are linear combinations of the random variables y_i, which by the Central LImit Theorem are normally distributed. Normal distributions are fully defined by their mean and variance, so we can find the distribution of the vector $\hat{\boldsymbol{\beta}}$ by only determining its mean and variance:

$$E[\hat{\boldsymbol{\beta}}|\mathbf{X}] = \boldsymbol{\beta}$$
$$Var\left(\hat{\boldsymbol{\beta}}|X\right) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$$

Confidence and prediction intervals:

Confidence intervals estimate the location of the population mean for the y-variable at a selected value of the predictor variable x^*, while prediction intervals estimate the location of individuals at a selected value of the predictor variable x^*. A $100(1-\alpha)\%$ confidence interval for the mean value of y at $X = x^*$, the location of the population regression line at $X = x^*$, is given by:

$$\hat{y}^* \pm t_{\alpha/2,n-2}s\sqrt{\frac{1}{n} + \frac{(x^*-\bar{x})^2}{SXX}}$$

A $100(1-\alpha)\%$ prediction interval for the mean value of y at $X = x^*$, the location of the population regression line at $X = x^*$, is given by:

$$\hat{y}^* \pm t_{\alpha/2,n-2}s\sqrt{1 + \frac{1}{n} + \frac{(x^*-\bar{x})^2}{SXX}}$$

The difference between the two intervals is the 1, which measures the extra variability that an individual has.

ANOVA Table:

The ANOVA table breaks the variability in the y-variable down into two parts: one that the model explains, and one that it doesn't. The part that the model does explain is called the Sums of Squares Regression, SSReg or SSModel, the difference between the model and no model at all. The part that the model doesn't explain is the Residual Sums of Squares, RSS, or SSError. The total corrected sum of squares is denoted:

$$SST = SYY = \sum_{i}^{n}(y_i - \bar{y})^2$$

Regression sums of squares:

$$SSReg = \sum_{i}^{n} (\hat{y}_i - \bar{y})^2$$

Residual sums of squares:

$$RSS = \sum_{I}^{n} (y_i - \hat{y}_i)^2$$

The overall F-test statistic to test whether the overall model explains a significant amount of varibility in y is

$$F = \frac{SSReg/p}{RSS/(n-2)}$$

where p is the number of predictors in the model.

R^2, the coefficient of determination of the regression line, is defined as the proportion of the total sample variability in the Y variable explained by the regression model:

$$R^2 = \frac{SSReg}{SST} = 1 - \frac{RSS}{SST}$$

This quantity is called R^2 because it is the correlation between Y and X. It is arguably one of the most commonly used statistics.

3. Diagnostics and Transformations for Simple Linear Regression

Estimates and confidence intervals must be based on a valid model. If the model is not valid, estimates and confidence do not have their advertised meaning. Transformations change the relationship between the predictor and the response, the goal being to find a linear relationship. Transformations may also change the variability of the residuals from non-constant to constant.

One way of checking whether a valid model has been fit is to plot residuals versus x and look for patterns. If no pattern is found then this indicates that the model provides an adequate summary of the data. If a pattern is found then the shape of the pattern provides information on the function of x that is missing from the model.

Many authors recommend that an effective plot to diagnose non-constant error variance is a plot of

$$|\text{Standardized Residuals}|^{0.5} \text{ against x}$$

The square root is used to reduces skewness in the absolute values. When non-constant variance exists, it is often possible to transform one or both of the regression variables to produce a model in which the error variance is constant.

Transformations:
Count data are often modeled using the Poisson distribution. If the response variable Y follows a Poisson distribution with mean λ, the variance is equal to the mean. In this case, the appropriate tranformation for Y is the square root.

Justification: Consider the following Taylor series expansion:

$$f(Y) = f(E[Y]) + f'(E[Y])(Y - E[Y]) + \ldots$$

Then the variance of $f(Y)$ can be found by:

$$Var(f(Y)) \approx [f'(E[Y])]^2 \, \text{Var}(Y)$$

If the transformation is $f(Y) = \sqrt{Y}$ and $\text{Var}(Y) = \lambda = E[Y]$ then:

$$\text{Var}\left(\sqrt{Y}\right) \approx \left[0.5\,(E[Y])^{-0.5}\right]^2 \text{Var}(Y) = \left[0.5\lambda^{-0.5}\right]^2 \lambda = \text{constant}$$

Logarithms can be used to estimate percentage effects. For the regression model

$$\log(Y) = \beta_0 + \beta_1 \log(x) + e$$

the slope β_1 approximately equal to the ratio of the percentage changes in Y and x. For small β_1, then, the slope can be interpreted as the percentage change in Y for a one percent change in x.

Because the relationship between normally distributed random variables is linear, we may wish to find a transformation that makes the predictor and response variables normally distributed. The Box-Cox procedure aims to find a power transformation that transforms variables to a normal distribution. The procedure is based on maximum likelihood. The result of the Box-Cox transformation method may be a transformed variable that is not very clsoe to normally distributed.

Another method for transforming the response and/or predictor variables to overcome problems due to nonlinearity is the inverse response plot. If the true regression model between Y and X is given by

$$Y = g(\beta_0 + \beta_1 x + e)$$

where g is a function generally unknown, the model can be turned into a simple linear regression model by transforming Y by g^{-1} since

$$g^{-1}(Y) = \beta_0 + \beta_1 x + e.$$

If x has an elliptically contoured distribution then g-1 can be estimated from the scatter plot of Y (on the horizontal axis) and$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ (on the vertical axis). Such a plot is called an inverse response plot.

Leverage Points:
Data points which exercise considerable influence on the fitted model are called leverage points. A leverage point is a point whose x-value is distant from the other x-values. A point is a bad leverage point if its Y-value does not follow the pattern set by the other data points. In other words, a bad leverage point is a leverage point which is also an outlier while a good leverage point is a leverage point which is NOT also an outlier.

The predicted value for Y is a weighted average of the observed values of Y:

$$\hat{y}_i = h_{ii} y_i + \sum j \neq i h_{ij} y_j$$

where

$$h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_j (x_j - \bar{x})^2}$$

There is a problem when this weighted average has too heavy a weight on the observed value that matches the current desired prediction. We are then relying too heavily on our observed data, and a different data set may result in a drastically different inference.

A popular rule is to classify a point x_i as a point of high leverage in a simple linear regression model if $h_{ii} > 4/n$. Strategies for dealing with bad leverage points are: 1.) Remove invalid data points 2.) Fit a different regression model.

The i^{th} standardized residual r_i is given by

$$r_i = \frac{\hat{e}_i}{s\sqrt{1 - h_{ii}}}$$

where $s = \sqrt{\frac{1}{n-2} \sum_j \hat{e}_j^2}$ is the estimate of σ obtained from the model. The standardized residual measures the extent to which the ith case is outlying.

When points of high leverage exist, instead of looking at residual plots, it is generally more informative to look at plots of standardized residuals since plots the residuals will have non-constant variance even if the errors have constant variance. The other advantage of standardized residuals is that they immediately tell us how many estimated standard deviations any point is away from the fitted regression model. We follow the common practice of labelling points as outliers in small to moderate size data sets if the standardized residual for the point falls outside the interval from -2 to 2. In very large data sets, we shall change this rule to -4 to 4. Identification and examination of any outliers is a key part of regression analysis.

Cook's distance is a function of both the standardized residual and the leverage. A popular cutoff for Cook's distance is $4/(n-2)$, but in practice it is more important to find points with a Cook's distance that is very different from other points and examine them.

The assumption of normal errors is needed in small samples for the validity of t-distribution based hypothesis tests and confidence intervals and for all sample sizes for prediction intervals. This assumption is generally checked by looking at the distribution of the residuals or standardized residuals. A common way to assess normality of the errors is to look at what is commonly referred to as a normal probability plot or a normal qq-plot of the standardized residuals.

4. Weighted Least Squares

An alternative way of coping with non-constant error variance is to use weighted least squares.

Consider the simple linear regression model

$$Y = \beta_0 + \beta_1 x + e$$

where the e_i have mean 0 but variance σ^2/w_i. When w_i is very large then the variance of e_i is close to 0. In this situation, the estimates of the regression parameters β_0 and β_1 should be such that the fitted line at x_i should be very close to y_i. On the other hand, when w_i is very small then the variance of e_i is very large. In this situation, the estimates of the regression parameters β_0 and β_1 should take little account of the values (x_i, y_i). In the extreme situation that w_i is 0 then the variance of ei is equal to infinity and the ith case (x_i, y_i) should be ignored in fitting the line: this is equivalent to deleting the i^{th} case (x_i, y_i) from the data and fitting the line based on the other $n-1$ cases.

Thus, we need to take account of the weights wi when estimating the regression parameters β_0 and β_1. This is achieved by considering the following weighted version of the residual sum of squares

$$WRSS = \sum_i w_i (y_i - \hat{y}_{Wi})^2 = \sum_i w_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2.$$

If the simple linear regression model

$$Y_i = \beta_0 + \beta_1 x_i + e_i$$

where the e_i have mean 0 but variance σ^2/w_i. If we multiply the previous equation by $\sqrt{w_i}$, we get:

$$\sqrt{w_i} Y_i = \sqrt{w_i}\beta_0 + \sqrt{w_i}\beta_1 x_i + \sqrt{w_i} e_i$$

where the $\sqrt{w_i} e_i$ have mean 0 but variance $\left(\sqrt{w_i}\right)^2 \sigma^2/w_i = \sigma^2$. Thus it is possible to calculate the weighted least squares fit of the simple linear regression model by calculating the least squares fit to the weighted model. The weighted model is a multiple linear regression model with two predictors and no intercept.

To see this, define transformations:

$$\begin{aligned} Y_{newi} &= \sqrt{w_i} Y_i \\ x_{1newi} &= \sqrt{w_i} \\ x_{2newi} &= \sqrt{w_i} x_i \\ e_{newi} &= \sqrt{w_i} e_i \end{aligned}$$

And then we can rewrite the weighted model as:

$$Y_{newi} = +\beta_0 x_{1newi} + \beta_1 x_{2newi} + e_{newi}.$$

For this transformed model, the variance of the i^{th} residual does not depend on the weight w_i.

Weighted least squares are most commonly used in the special case when Y_i is the sum, average, or median of n_i observations, so that $\text{Var}(Y_i) \propto n_i$ or $\frac{1}{n_i}$. In these cases we would take $w_i = \frac{1}{n_i}$ or n_i.

We can also find the solutions for the parameters in matrix form by:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}(\mathbf{X}'\mathbf{W}\mathbf{Y})$$

However, many situations exist in which the variance is not constant and in which it is not straightforward to determine the correct model for the variance. In these situations, the use of weighted least squares is problematic.

5. Multiple Linear Regression

Multiple linear regression includes the special case of polynomial regression, in which the predictors are a single predictor x along with its polynomial powers (x^2, x^3, etc.). In polynomial regression, we can display results like residual plots on a single two-dimensional graph.

Multiple regressions with three or more predictor variables, on the other hand, can only be displayed on multiple two and three-dimensional graphs. The important idea behind inference for multiple regression is that we control for the levels of other variables when interpreting each individual slope.

The first test to be conducted is the Analysis of Variance to test $H_0 : \beta_1 = \beta_2 = \ldots \beta_p = 0$ against the alternative that at least one of the $\beta_i \neq 0$. After the overall F-test is found to be significant, we may conduct t-tests for each of the individual variables. The individual t-tests test whether that variable explains a significant amount of variability in Y after all the other variables are already added to the model.

Model Reduction:

To test whether a set of variables may be removed from the model, a partial F-test can be conducted. The F-test for testing $H_0 : \beta_1 = \beta_2 = \ldots \beta_k = 0$ for $k < p$ against the alternative that at least one of the β_i, $i = 1, \ldots, k$ is not equal to 0 is given by:

$$F = \frac{(RSS(reduced) - RSS(full))/(dferr_{reduced} - dferr_{full})}{RSS(full)/dferr_{full}}$$

Parallel, Same Intercept, and Unrelated Regression Lines:
In the case when we have one predictor variable x and one dummy variable d, we may consider a parallel regression line model:

$$Y = \beta_0 + \beta_1 x + \beta_2 d + e$$

A model with the same intercepts but different slopes:

$$Y = \beta_0 + \beta_1 x + \beta_2 x d + e$$

Or a model with different intercepts and different slopes:

$$Y = \beta_0 + \beta_1 x + \beta_2 d + \beta_3 x de$$

For the parallel regression lines model, the coefficient for the dummy variable measures the additive effect due to being in the group denoted by a 1 in the dummy variable.

6. Diagnostics and Transformations for Multiple Linear Regression

The multiple linear regression model is $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$ where $\text{Var}(\mathbf{e}) = \sigma^2\mathbf{I}$ and $\mathbf{I}$ is the $(n \times n)$ identity matrix. The predicted values are given by $\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$ where $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$. Then the residuals are given by:

$$\hat{\mathbf{e}} = \mathbf{Y} - \hat{\mathbf{Y}} = \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}$$

Leverage and Outliers:
Recall that data points that have considerable influence on the location of the fitted model are called leverage points. Leverage measures the extent to which the fitted model is attracte to the given data point. We are therefore interested in the relationship between $\hat{\mathbf{Y}}$ and $\mathbf{Y}$. The $(n \times n)$ matrix $\mathbf{H}$ is commonly called the hat matrix since pre-multiplying $\mathbf{Y}$ by $\mathbf{H}$ changes $\mathbf{Y}$ into $\hat{\mathbf{Y}}$:

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \mathbf{H}\mathbf{Y}$$

where

$$\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$

A popular rule for leverage points in multiple linear regression identifies points as points of high leverage if the i^{th} diagonal of $\mathbf{H}$, h_{ii}, is greater than $2(p+1)/n$.

When the correct model has been fit, a plot of standardized residuals r_i against any predictor will have a random scatter of points around the horizontal axis and constant variability as we look along the horizontal axis. An implication of these features is that any pattern is indicative of an incorrect model.

Multivariate issues:
Added variable plots enable us to visually assess the effect of each predictor, having adjusted for the effects of the other predictors. Assuming a valid model, the added variable plot produces points randomly scattered around a straight line through the origin with slope equal to the slope in the model. The added variable plot also enables the user to identify any data points which have too much influence on the estimate of that slope.

Variance inflation factors are found by the term $1/(1-R_j^2)$, where R_j^2 denotes the value of R^2 obtained from the regression of x_j on all the other predictors. The variance inflation factor tells us the amount that the variance of the j^{th} predictor is inflated by due to correlation with some linear combination of the other predictors. This can be seen in the formula:

$$\text{Var}(\hat{\beta}_j) = \frac{1}{1-R_j^2} \times \frac{\sigma^2}{(n-1)S_{x_j}^2}, j = 1, \ldots, p$$

A common cutoff for the variance inflation factor is 5. If the VIF exceeds 5 we will conclude that the associated regression coefficient is poorly estimated due to multicollinearity among the predictors.

A well known weakness of regression modeling based on observational data is that the observed association between two variables may be because both are related to a third variable that has been omitted from the regression model. This phenomenon is commonly referred to as spurious correlation. The third variable is commonly called either an omitted variable or a confounding covariate.

Transformations:
Similarly to simple linear regression, the Box-Cox procedure may be used to transform the predictor variables to multivariate normality, and then transform the response variable to normality, given the predictors. The Box-Cox method searches only for power transformations of the variables.

Marginal model plots allow us to assess whether the mean function has been fit correctly, as with simple linear regression.

Transforming the response variable using the log transformation without transforming the predictor x_k results in the interpretation that for every 1 unit change in the predictor, the model predicts a $100 \times \beta_k\%$ change in Y, as long as another predictor has been transformed using the log transformation. We use percent/percent interpretations when both x_k and y have been transformed using a log transformations.

7. Variable Selection

We assume that the full model is a valid regression model and consider methods for choosing hte best model from a class of multiple regression models using variable selection methods. In general, the more predictor variables included in a valid model the lower the bias of the predictions but the higher the variance. Including too many predictors is called over-fitting while the opposite is under-fitting. The two key aspects of variable selection methods are:

1. Evaluating each potential subset of p predictor variables
2. Deciding on the collection of potential subsets

In general, methods in Chapter 7 result in models with p-values that are much too small, as the model has been fit to the data set at hand. The naive use of inference procedures that do not take the model selection step into account can be highly misleading.

We consider two big ideas in this chapter:

1. Determining which subsets will be evaluated: backward, forward, stepwise, or all possible subsets
2. Using different criteria for choosing a best model: $R^2_{adj}, AICC, BIC, C(p)$.

If possible, we prefer to consider all possible subsets. When that is computationally too slow or impossible, we decide to use a selection criterion. Both the forward and stepwise methods begin by considering models with only one predictor variable, and so may be a good choice for situations when many variables are under consideration.

A popular strategy with respect to using different criteria to choose a best model is to compute all the different criteria for choosing a best model and compare the models selected by each of them. With a defined number of terms in the regression model, all the criteria agree that the best choice is the set of predictors with the smallest value of the residual sum of squares. Only when comparing models of different numbers of predictors do the selection methods give different results.

Assessing the predictive ability of regression models
Given that the model selection process changes the properties of the standard inferential procedures, a standard approach to assessing the predictive ability of different regression models is

to evaluate their performance on a new data set (i.e., one not used in the development of the models). In practice, this is often achieved by randomly splitting the data into

1. A training data set
2. A test data set

The training data set is used to develop a number of regression models, while the test data set is used to evaluate the performance of these models. Splitting the data into a training set and a test set Snee (1977, p. 421) demonstrated the advantages of splitting the data into a training set and a test set such that "the two sets cover approximately the same region and have the same statistical properties". Random splits, especially in small samples, do not always have these desirable properties.

8. Logistic Regression

Binomial Data:
This chapter considers the situation in which the response variable is based on "yes" / "no" responses, such as whether or not a patient has a heart attack. We begin by considering the case of a single predictor variable x. In this case

$$(Y_i|x_i) \sim Bin(m_i, \theta(x_i)), i = 1, \ldots, n$$

The sample proportion of "successes" at each value of the predictor variable is given by $y_i m_i$. Notice that

$$E[y_i/m_i|x_i] = \theta(x_i) \text{ and } \mathrm{Var}(y_i/mi|x_i) = \theta(x_i)(1-\theta(x_i))/m_i.$$

The sample proportion of successes in the data, y_i/m_i, is used as the response variable because:

1. y_i/m_i is an unbiased estimate of $\theta(x_i)$.
2. y_i/m_i varies between 0 and 1.

Notice that the variance of the response y_i/m_i depends on $\theta(x_i)$ and as such it is not a constant. In addition, this variance is also therefore unknown. In general, the shape of the underlying function $\theta(x)$ is not a straight line. Instead it appears S-shaped, with very low values of the x-variable resulting in zero probability of "success" and very high values of the x-variable resulting in a probability of "success" equal to one. A popular choice for the S-shaped function is the logistic function:

$$\theta(x) = \frac{exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)} = \frac{1}{1 + \exp(-[\beta_0 + \beta_1 x])}$$

Solving this last equation for $\beta_0 + \beta_1 x$ gives

$$\log\left(\frac{\theta(x)}{1-\theta(x)}\right) = \beta_0 + \beta_1 x$$

Thus, if the chosen function is correct, a plot of the logit, $\log\left(\frac{\theta(x)}{1-\theta(x)}\right)$ against x will produce a straight line.

Odds:
The quantity $\frac{\theta(x)}{1-\theta(x)}$ is known as odds. The odds in favor of success are defined as the ratio of the probability that success will occur to the probability that success will not occur. The odds in a logistic regression are in the form of odds in favor of a success (in contrast to horse races, which model the odds of failure).

The parameters for logistic regression are found by the method of maximizing the log likelihood function, rather than least squares. This has to be done using an iterative method such as Newton-Raphson or iteratively reweighted least squares.

The standard approach to testing $H_0 : \beta_1 = 0$ and finding confidence intervals for the parameters is to use Wald z-statistics rather than the t-statistics from linear regression methods.

In logistic regression the concept of the residual sum of squares is replaced by a concept known as the deviance. In the case of logistic regression the deviance is defined to be

$$G^2 = 2\sum_i \left[y_i \log\left(\frac{y_i}{\hat{y}_i}\right) + (m_i - y_i)\log\left(\frac{m_i - y_i}{m_i - \hat{y}_i}\right)\right]$$

where $\hat{y}_i = m_i\hat{\theta}(x_i)$. The degrees of freedom associated with the deviance is given by $n -$ (number of β's estimated) The deviance associated with a given logistic model (M) is based on comparing the maximized log-likelihood under (M) with the maximized log-likelihood under (S), the so-called saturated model that has a parameter for each observation. In fact, the deviance is given by twice the difference between these maximized loglikelihoods. The saturated model (S) estimates $\theta(x_i)$ by the observed proportion of successes in the data at x_i; that is, by y_i/m_i.

When each m_i, the number of trials at x_i, is large enough, the deviance can be used as a goodness of fit test for the logistic regression model. If we wish to test the null hypothesis that the logistic regression model is appropriate against the alternative that it is inappropriate so that a saturated model is needed, we can use the deviance G^2. Under the null hypothesis and when each m_i is large enough, the deviance G^2 is approximately distributed as χ^2_{n-p-1}, where $n =$ the number of binomial samples and $p =$ the number of predictors in the model.

We can also use the difference in deviance values to compare nested models. The difference in the deviances is to be compared to a χ^2 distribution with degrees of freedom equal to the difference in degrees of freedom between the two models. Wald tests and tests based on the difference in deviances can result in different p-values.

An alternative measure of the goodness of fit of a logistic regression model is the Pearson χ^2 statistic. The deviance and deviance residuals are preferred since their distribution is closer to that of least squares residuals, however.

Deviance residuals are analogous to Pearson residuals, and are defined by

$$r_{Deviance,i} = \text{sign}\left(y_i/m_i - \hat{\theta}(x_i)\right) g_i$$

where $G^2 = \sum_i g_i^2$. Furthermore, standardized deviance residuals (essentially z-scores - these should have the same variance) are defined to be:

$$sr_{Deviance,i} = \frac{r_{Deviance,i}}{\sqrt{1 - h_{ii}}}$$

Binary Data:
A very important special case of logistic regression occurs when all the m_i equal 1. Such data are called binary data. In this situation the goodness-of-fit measures χ^2 and G^2 are problematic and plots of residuals can be difficult to interpret.

The deviance does not provide an assessment of the goodness of fit of the model when all the m_i are equal to 1. Furthermore, the distribution of the deviance is not χ^2, even approximately, although the difference between deviances is approximately χ^2. Residual plots are thus problematic when the data are binary.

Transformations:
When the predictor variable X is normally distributed with a different variance for the two values of Y, the log odds ratio is a quadratic function of x. When the variances are equal, the log odds ratio is a linear function of x. When we have p predictors, if the covariance matrix of the predictors differs across the two groups than the log odds ratio is a function of $x_i, {}_i^2$, and$x_i x_j$. The product term $x_i x_j$ is needed as a predictor if the covariances of x_i and x_j differ across the two values of y, that is, if the regression of x_i on x_j (or vice versa) has a different slope for the two values of y.

If the densities $f(x|Y = j), j = 0, 1$ are skewed the log odds ratio can depend on both x and $\log(x)$. It is often easiest to assess the need for both by including them both in the model so that their relative contributions can be assessed directly. If the skewed predictor can be transformed to have a normal distribution conditional on Y, then just the transformed version of X should be included in the logistic regression model.

If the conditional distribution of X is Poisson with means λ_1 and λ_0 for the two values of Y, the log odds ratio is a linear function of x. When X is a dummy variable, it can be shown that the log odds ratio is also a linear function of x.

9. Serially Correlated Errors

In many situations data are collected over time. It is common for such data sets to exhibit serial correlation, that is, results from the current time period are correlated with results from earlier time periods. Thus, these data sets violate the assumption that the errors are independent, an important assumption necessary for the validity of least squares based regression methods.

It is common statistical practice to look at values of the correlation between Y and the various values of lagged Y for different periods. Such values are called autocorrelations. The autocorrelation of lag l is the correlation between Y and values of Y lagged by l periods, i.e. between Y_t and Y_{t-l}. In formula notation:

$$\text{Autocorrleation}(l) = \frac{\sum_{i=l+1}^{n}(y_t - \bar{y})(y_{t-l} - \bar{y})}{\sum_{t=n}^{n}(y_t - \bar{y})^2}$$

Autocorrelations are declared to be statistically significantly different from zero if they are less than $-2/\sqrt{n}$ or greater than $2/\sqrt{n}$ (i.e. if they are more than two standard errors away from zero).

We consider the simplest situation, when Y_t can be predicted from a single predictor X_t and the errors follow an autoregressive process of order 1, AR(1), that is:

$$Y_t = \beta_0 + \beta_1 X_t + e_t, \text{ where } e_t = \rho\, e_{t-1} + \nu_t, \text{ and } \nu_t \sim N(0, \sigma_\nu^2)$$

Thus the correlation between the errors that are at lag l apart from each other is $\text{Corr}(e_t, e_{t-l}) = \rho^l, l = 1, 2, \ldots$. When $|\rho| < 1$, these correlations get smaller as l increases.

The least squares estimate of $\hat{\beta}_1$ is unbiased for β_1; however, its variance is the usual variance of the least squares estimate multiplied by a function of ρ. Thus using least squares ignoring autocorrelation will results in consistent estimates of the slope but incorrect estimates of the variance of the slope estimate, invalidating the results of confidence intervals and hypothesis tests. We use generalized least squares estimation instead of the usual methods.

Instead of assuming the errors are independent, we assume that $\mathbf{e} \sim N(\mathbf{0}, \Sigma)$, where the covariance matrix does not have 0's in the off-diagonal elements. Instead, it is given by:

$$\Sigma = \frac{\sigma_\nu^2}{1-\rho^2}\begin{bmatrix} 1 & \rho & \cdots & \rho^{n-1} \\ \rho & 1 & \cdots & \rho^{n-2} \\ \vdots & \vdots & \ddots & \vdots \\ \rho^{n-1} & \rho^{n-2} & \cdots & 1 \end{bmatrix}$$

The generalized least squares (GLS) estimator of β, then, is:

$$\hat{\beta}_{GLS} = (\mathbf{X}'\Sigma^{-1}\mathbf{X})^{-1}(\mathbf{X}'\Sigma^{-1}\mathbf{Y}).$$

Transformations: To use least squares diagnostics, we can transform the AR(1) model into a model with uncorrelated errors. The Cholesky decomposition of the covariance matrix Σ into $\Sigma = SS'$ gives a transformation with iid errors. Let:

$$\begin{aligned} \mathbf{Y}^* &= S^{-1}\mathbf{Y} \\ \mathbf{X}^* &= S^{-1}\mathbf{X} \\ \mathbf{e}^* &= S^{-1}\mathbf{e} \end{aligned}$$

Then $\mathbf{Y}^* = \mathbf{X}^*\beta + \mathbf{e}^*$ provides a linear model with iid errors.

Ignoring autocorrelations can produce misleading model diagnostics. It is instead recommended that one use least squares diagnostics based on $\mathbf{Y}^*$ and $\mathbf{X}^*$ for model checking.