

## Chapter 10. Supplemental Text Material

### 10-1. The Covariance Matrix of the Regression Coefficients

In Section 10-3 of the textbook, we show that the least squares estimator of  $\beta$  in the linear regression model  $\mathbf{y} = \mathbf{X}\beta + \varepsilon$

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

is an unbiased estimator. We also give the result that the covariance matrix of  $\hat{\beta}$  is  $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$  (see Equation 10-18). This last result is relatively straightforward to show. Consider

$$V(\hat{\beta}) = V[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}]$$

The quantity  $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$  is just a matrix of constants and  $\mathbf{y}$  is a vector of random variables. Now remember that the variance of the product of a scalar constant and a scalar random variable is equal to the square of the constant times the variance of the random variable. The matrix equivalent of this is

$$\begin{aligned} V(\hat{\beta}) &= V[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}] \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'V(\mathbf{y})[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}]' \end{aligned}$$

Now the variance of  $\mathbf{y}$  is  $\sigma^2\mathbf{I}$ , where  $\mathbf{I}$  is an  $n \times n$  identity matrix. Therefore, this last equation becomes

$$\begin{aligned} V(\hat{\beta}) &= V[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}] \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'V(\mathbf{y})[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}]' \\ &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}]' \\ &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1} \end{aligned}$$

We have used the result from matrix algebra that the transpose of a product of matrices is just the produce of the transposes in reverse order, and since  $(\mathbf{X}'\mathbf{X})$  is symmetric its transpose is also symmetric.

### 10-2. Regression Models and Designed Experiments

In Examples 10-2 through 10-5 we illustrate several uses of regression methods in fitting models to data from designed experiments. Consider Example 10-2, which presents the regression model for main effects from a  $2^3$  factorial design with three center runs. Since the  $(\mathbf{X}'\mathbf{X})^{-1}$  matrix is symmetric because the design is orthogonal, all covariance terms between the regression coefficients are zero. Furthermore, the variance of the regression coefficients is

$$V(\hat{\beta}_0) = \sigma^2 / 12 = 0.0833\sigma^2$$

$$V(\hat{\beta}_i) = \sigma^2 / 8 = 0.125\sigma^2, i = 1, 2, 3$$

In Example 10-3, we reconsider this same problem but assume that one of the original 12 observations is missing. It turns out that the estimates of the regression coefficients does not change very much when the remaining 11 observations are used to fit the first-order model but the  $(\mathbf{X}'\mathbf{X})^{-1}$  matrix reveals that the missing observation has had a moderate effect on the variances and covariances of the model coefficients. The variances of the regression coefficients are now larger, and there are some moderately large covariances between the estimated model coefficients. Example 10-4, which investigated the impact of inaccurate design factor levels, exhibits similar results. Generally, as soon as we depart from an orthogonal design, either intentionally or by accident (as in these two examples), the variances of the regression coefficients will increase and potentially there could be rather large covariances between certain regression coefficients. In both of the examples in the textbook, the covariances are not terribly large and would not likely result in any problems in interpretation of the experimental results.

### 10-3. Adjusted $R^2$

In several places in the textbook, we have remarked that the adjusted  $R^2$  statistic is preferable to the ordinary  $R^2$ , because it is not a monotonically non-decreasing function of the number of variables in the model.

From Equation (10-27) note that

$$R_{Adj}^2 = 1 - \left[ \frac{SS_E / df_e}{SS_T / df_T} \right]$$

$$= 1 - \frac{MS_E}{SS_T / df_T}$$

Now the mean square in the denominator of the ratio is constant, but  $MS_E$  will change as variables are added or removed from the model. In general, the adjusted  $R^2$  will increase when a variable is added to a regression model only if the error mean square decreases. The error mean square will only decrease if the added variable decreases the residual sum of squares by an amount that will offset the loss of one degree of freedom for error. Thus the added variable must reduce the residual sum of squares by an amount that is at least equal to the residual mean square in the immediately previous model; otherwise, the new model will have an adjusted  $R^2$  value that is larger than the adjusted  $R^2$  statistic for the old model.

### 10-4. Stepwise and Other Variable-Selection Methods in Regression

In the textbook treatment of regression, we concentrated on fitting the full regression model. Actually, in moist applications of regression to data from designed experiments the experimenter will have a very good idea about the form of the model he or she wishes

to fit, either from an ANOVA or from examining a normal probability plot of effect estimates.

There are, however, other situations where regression is applied to **unplanned studies**, where the data may be observational data collected routinely on some process. The data may also be archival, obtained from some historian or library. These applications of regression frequently involve a moderately-large or large set of **candidate regressors**, and the objective of the analysts here is to fit a regression model to the “best subset” of these candidates. This can be a complex problem, as these unplanned data sets frequently have outliers, strong correlations between subsets of the variables, and other complicating features.

There are several techniques that have been developed for selecting the best subset regression model. Generally, these methods are either **stepwise-type** variable selection methods or **all possible regressions**. Stepwise-type methods build a regression model by either adding or removing a variable to the basic model at each step. The forward selection version of the procedure begins with a model containing none of the candidate variables and sequentially inserts variables into the model one-at-a-time until a final equation is produced. In backward elimination, the procedure begins with all variables in the equation, and then variables are removed one-at-a-time to produce a final equation. Stepwise regression usually consists of a combination of forward and backward stepping. There are many variations of the basic procedures.

In all possible regressions with  $K$  candidate variables, the analyst examines all  $2^K$  possible regression equations to identify the ones with potential to be a useful model. Obviously, as  $K$  becomes even moderately large, the number of possible regression models quickly becomes formidably large. Efficient algorithms have been developed that implicitly rather than explicitly examine all of these equations. For more discussion of variable selection methods, see textbooks on regression such as Montgomery and Peck (1992) or Myers (1990).

### 10-5. The Variance of the Predicted Response

In section 10-5.2 we present Equation (10-40) for the variance of the predicted response at a point of interest  $\mathbf{x}'_0 = [x_{01}, x_{02}, \dots, x_{0k}]$ . The variance is

$$V[\hat{y}(\mathbf{x}_0)] = \sigma^2 \mathbf{x}'_0 (\mathbf{X}'\mathbf{X})^{-1} \mathbf{x}_0$$

where the predicted response at the point  $\mathbf{x}_0$  is found from Equation (10-39):

$$\hat{y}(\mathbf{x}_0) = \mathbf{x}'_0 \hat{\beta}$$

It is easy to derive the variance expression:

$$\begin{aligned} V[\hat{y}(\mathbf{x}_0)] &= V(\mathbf{x}'_0 \hat{\beta}) \\ &= \mathbf{x}'_0 V(\hat{\beta}) \mathbf{x}_0 \\ &= \sigma^2 \mathbf{x}'_0 (\mathbf{X}'\mathbf{X})^{-1} \mathbf{x}_0 \end{aligned}$$

Design-Expert calculates and displays the confidence interval on the mean of the response at the point  $\mathbf{x}_0$  using Equation (10-41) from the textbook. This is displayed on the point prediction option on the optimization menu. The program also uses Equation (10-40) in the contour plots of prediction standard error.

### 10-6. Variance of Prediction Error

Section 10-6 of the textbook gives an equation for a prediction interval on a future observation at the point  $\mathbf{x}'_0 = [x_{01}, x_{02}, \dots, x_{0k}]$ . This prediction interval makes use of the variance of prediction error. Now the point prediction of the future observation  $y_0$  at  $\mathbf{x}_0$  is

$$\hat{y}(\mathbf{x}_0) = \mathbf{x}'_0 \hat{\beta}$$

and the prediction error is

$$e_p = y_0 - \hat{y}(\mathbf{x}_0)$$

The variance of the prediction error is

$$\begin{aligned} V(e_p) &= V[y_0 - \hat{y}(\mathbf{x}_0)] \\ &= V(y_0) + V[\hat{y}(\mathbf{x}_0)] \end{aligned}$$

because the future observation is independent of the point prediction. Therefore,

$$\begin{aligned} V(e_p) &= V(y_0) + V[\hat{y}(\mathbf{x}_0)] \\ &= \sigma^2 + \sigma^2 \mathbf{x}'_0 (\mathbf{X}'\mathbf{X})^{-1} \mathbf{x}_0 \\ &= \sigma^2 [1 + \mathbf{x}_0 (\mathbf{X}'\mathbf{X})^{-1} \mathbf{x}_0] \end{aligned}$$

The square root of this quantity, with an estimate of  $\sigma^2 = \hat{\sigma}^2 = MS_E$ , appears in Equation (10-42) defining the prediction interval.

### 10-7. Leverage in a Regression Model

In Section 10-7.2 we give a formal definition of the leverage associated with each observation in a design (or more generally a regression data set). Essentially, the leverage for the  $i$ th observation is just the  $i$ th diagonal element of the “hat” matrix

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

or

$$h_{ii} = \mathbf{x}'_i (\mathbf{X}'\mathbf{X})^{-1} \mathbf{x}_i$$

where it is understood that  $\mathbf{x}'_i$  is the  $i$ th row of the  $\mathbf{X}$  matrix.

There are two ways to interpret the leverage values. First, the leverage  $h_{ii}$  is a measure of **distance** reflecting how far each design point is from the center of the design space. For example, in a  $2^k$  factorial all of the cube corners are the same distance  $\sqrt{k}$  from the

design center in coded units. Therefore, if all points are replicated  $n$  times, they will all have identical leverage.

Leverage can also be thought of as the maximum potential influence each design point exerts on the model. In a near-saturated design many or all design points will have the maximum leverage. The maximum leverage that any point can have is  $h_{ii} = 1$ . However, if points are replicated  $n$  times, the maximum leverage is  $1/n$ . High leverage situations are not desirable, because if leverage is unity that point fits the model exactly. Clearly, then, the design and the associated model would be vulnerable to outliers or other unusual observations at that design point. The leverage at a design point can always be reduced by replication of that point.