Chapter 4 – Simultaneous Inference and Other Topics

4.1 Joint Estimation of β_0 and β_1

We've obtained (1-α)100% confidence intervals for the slope and intercept parameters in Chapter 2. Now we'd like to construct a range of values (β_0, β_1) that we believe contains BOTH parameters with the same level of confidence. One way to do this is to construct each individual confidence interval at a higher level of confidence, namely: (1-(α/2))100% confidence intervals for β_0 and β_1 separately. The resulting ranges are called **Bonferroni Joint (Simultaneous) Confidence Intervals**.

Joint Confidence Level (1-α)100%	Individual Confidence Level (1-(α/2))100%
90%	95%
95%	97.5%
99%	99.5%

The resulting simultaneous confidence intervals, with a joint confidence level of at least (1-α)100% are:

$$b_0 \pm Bs\{b_0\} \qquad b_1 \pm Bs\{b_1\} \qquad B = t(1-(\alpha/4); n-2)$$

Bonferroni Joint (Simultaneous) Confidence intervals for a Family

The Bonferroni inequality can be extended to *g* simultaneous confidence intervals with family confidence coefficient 1-α by constructing each interval estimate with statement confidence coefficient 1-α/*g*.

4.2 Simultaneous Estimation of Mean Responses

Case 1: Simultaneous (1-α)100% Bounds for the Regression Line (Working-Hotelling's Approach)

$$\hat{Y}_h \pm Ws\{\hat{Y}_h\} \qquad W = \sqrt{2F(1-\alpha;2,n-2)}$$

Case 2: Simultaneous (1-α)100% Bounds at *g* Specific *X* Levels (Bonferroni's Approach)

$$\hat{Y}_h \pm Bs\{\hat{Y}_h\} \qquad B = t(1-(\alpha/2g); n-2)$$

4.3 Simultaneous Prediction Intervals for New Observations

Sometimes we wish to obtain simultaneous prediction intervals for *g* new outcomes.

Scheffe's Method:

$$\hat{Y}_h \pm Ss\{pred\} \qquad S = \sqrt{gF(1-\alpha; g, n-2)}$$

where $s\{pred\} = \sqrt{MSE(1 + \frac{1}{n} + \frac{(X_h - \overline{X})^2}{\sum (X_i - \overline{X})^2})}$ is the estimated standard error of the prediction.

Bonferroni's Method:

$$\hat{Y}_h \pm Bs\{pred\} \qquad B = t(1-\alpha/(2g); n-2)$$

Both *S* and *B* can be computed before observing the data, and the smaller of the two should be used.

4.4 Regression through the Origin

Sometimes it is desirable to have the mean response be 0 when the predictor variable is 0 (this is not the same as saying *Y* must be 0 when *X* is 0). Even though it can cause extra problems, it is an interesting special case of the simple regression model.

$$Y_i = \beta_1 X_i + \varepsilon_i \qquad \varepsilon_i \sim NID(0, \sigma^2)$$

We obtain the least squares estimate of β_1 (which also happens to be maximum likelihood) as follows:

$$Q = \sum \varepsilon_i^2 = \sum (Y_i - \beta_1 X_i)^2 \quad \Rightarrow \quad \frac{\partial Q}{\partial \beta_1} = 2\sum (Y_i - \beta_1 X_i)(-X_i)$$

$$\Rightarrow \quad -2\left[\sum X_i Y_i - b_1 \sum X_i^2\right] = 0 \quad \Rightarrow \quad \sum X_i Y_i = b_1 \sum X_i^2 \quad \Rightarrow \quad b_1 = \frac{\sum X_i Y_i}{\sum X_i^2}$$

The fitted values and residuals (which no longer necessarily sum to 0) are:

$$\hat{Y}_i = b_1 X_i \qquad e_i = Y_i - \hat{Y}_i$$

An unbiased estimate of the error variance σ^2 is:

$$s^2 = MSE = \frac{\sum (Y_i - \hat{Y}_i)^2}{n-1} = \frac{\sum e_i^2}{n-1}$$

Note that we have only estimated one parameter in this regression function.

Note that the following are linear functions of $Y_1, \ldots, Y_n$:

$$b_1 = \frac{\sum X_i Y_i}{\sum X_i^2} = \sum \frac{X_i}{\sum X_i^2} Y_i = \sum a_i Y_i \qquad a_i = \frac{X_i}{\sum X_i^2}$$

$$\Rightarrow \quad E\{b_1\} = E\left\{\sum a_i Y_i\right\} = \sum a_i E\{Y_i\} = \sum a_i \beta_1 X_i = \sum \frac{X_i}{\sum X_i^2} \beta_1 X_i = \beta_1 \frac{\sum X_i^2}{\sum X_i^2} = \beta_1$$

$$\Rightarrow \quad \sigma^2\{b_1\} = \sigma^2\{\sum a_i Y_i\} = \sum a_i^2 \sigma^2\{Y_i\} = \sum \left[\frac{X_i}{\sum X_i^2}\right]^2 \sigma^2 = \sigma^2 \frac{\sum X_i^2}{\left[\sum X_i^2\right]^2} = \frac{\sigma^2}{\sum X_i^2}$$

Thus, b_1 is an unbiased estimate of the slope parameter β_1, and its variance (and thus standard error) can be estimated as follows:

$$s^2\{b_1\} = \frac{s^2}{\sum X_i^2} = \frac{MSE}{\sum X_i^2} \qquad \Rightarrow \qquad s\{b_1\} = \sqrt{\frac{MSE}{\sum X_i^2}}$$

This can be used to construct confidence intervals for or conduct tests regarding β_1.

The mean response at X_h for this model is: $E\{Y_h\} = \beta_1 X_h$ and its estimate is $\hat{Y}_h = b_1 X_h$, with mean and variance:

$$E\{\hat{Y}_h\} = E\{b_1 X_h\} = X_h E\{b_1\} = X_h \beta_1$$

$$\sigma^2\{\hat{Y}_h\} = \sigma^2\{b_1 X_h\} = X_h^2 \sigma^2\{b_1\} = \sigma^2 \frac{X_h^2}{\sum X_i^2} \quad \Rightarrow \quad s^2\{\hat{Y}_h\} = MSE \frac{X_h^2}{\sum X_i^2}$$

This can be used to obtain a confidence interval for the mean response when $X=X_h$.

The estimated prediction error for a new observation at $X=X_h$ is:

$$s^2\{pred\} = s^2\{Y_{h(new)} - \hat{Y}_h\} = s^2\{Y_{h(new)}\} + s^2\{\hat{Y}_h\} = s^2 + \frac{s^2 X_h^2}{\sum X_i^2} = MSE\left[1 + \frac{X_h^2}{\sum X_i^2}\right]$$

This can be used to obtain a prediction interval for a new observation at this level of X.

Comments Regarding Regression through the Origin:

- R: lm(Y~X-1)
- SAS: Proc reg; model Y=X /NOINT;
- You should test whether the true intercept is 0 when *X*=0 before proceeding.
- If *X*=0 is not an important value of *X* in practice, there is no reason to put this constraint into the model.
- R^2 is no longer constrained to be bigger than 0, the error sum of squares from the regression can exceed the total corrected sum of squares. The coefficient of determination loses its interpretation of being the proportion of variation in *Y* that is "explained" by *X*.

4.5 Effects of Measurement Errors

Measurement Errors in *Y*

This causes no problems as the measurement error in *Y* becomes part of the random error term, which represents effects of many unobservable quantities. This is the case as long as the random errors are independent, unbiased, and not correlated with the level of *X*.

Measurement Errors in *X*

Problems do arise when the measurement of the predictor variable is measured with error. This is particularly the case when the observed (reported) X_i* level is the true level X_i plus a random error term. In this case the random error terms are not independent of the reported levels of the predictor variable, causing the estimated regression coefficients to be biased and not consistent. See textbook for a mathematical development. Certain methods have been developed for particular forms of measurement error. See *Measurement Error Models* by W.A. Fuller for a theoretical treatment of the problem or *Applied Regression Analysis* by J.O. Rawlings, S.G. Pantula, and D.A. Dickey for a brief description.

4.6 Inverse Predictions

Sometimes after we fit (or calibrate) a regression model, we can observe *Y* values and wish to predict the *X* levels that generated the outcomes. Let $Y_{h(new)}$ represent a new value of *Y* we have just observed, or a desired level of *Y* we wish to observe. In neither case, was this observation part of the sample. We wish to predict the *X* level that led to our observation, or the *X* level that will lead to our desired level. Consider the estimated regression function:

$$\hat{Y} = b_0 + b_1 X$$

Now we observe a new outcome $Y_{\mathrm{h(new)}}$ and wish to predict the X value corresponding to it, we can use an estimator that solves the previous equation for X. The estimator and its (approximate) estimated standard error are:

$$\hat{X}_{h(new)} = \frac{Y_{h(new)} - b_0}{b_1} \qquad s\{predX\} = \sqrt{\frac{MSE}{b_1^2}\left[1 + \frac{1}{n} + \frac{(\hat{X}_{h(new)} - \overline{X})^2}{\sum (X_i - \overline{X})^2}\right]}$$

Then, an approximate (1-α)100% Prediction Interval for $X_{\mathrm{h(new)}}$ is:

$$\hat{X}_{h(new)} \pm t(1-\alpha/2; n-2) s\{predX\}$$

4.7 Choosing *X* Levels

Issues arising involving choices of X levels and sample sizes include:

- The “range” of X values of interest to experimenter
- The goal of research: inference concerning the slope, predicting future outcomes, understanding the shape of the relationship (linear, curved,…)
- The cost of collecting measurements

Note that all of our estimated standard errors depend on the number of observations and the spacing of X levels. The more spread out, the smaller the standard errors, generally. However, if we wish to truly understand the shape of the response curve, we must space the observations throughout the set of X values. See quote by D.R. Cox on page 170 of textbook.