PREDICTION FROM LEAST SQUARES REGRESSION

CONSIDER 2 VARIABLE REGRESSION

$$Y_i = \beta_1 + \beta_2 X_i + u_i$$

THE OLS ESTIMATORS OF THE COEFFICIENTS ARE

$$\hat{\beta}_2 = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2}$$

$$\hat{\beta}_1 = \bar{Y} - \hat{\beta}_2 \bar{X}$$

THE VARIANCE OF $\hat{\beta}_2 = \dfrac{\sigma^2}{\sum (X_i - \bar{X})^2}$

VARIANCE OF $\hat{\beta}_1 = \left[\dfrac{1}{n} + \dfrac{\bar{X}^2}{\sum (X_i - \bar{X})^2}\right]\sigma^2$

$$\mathrm{Cov}(\hat{\beta}_1, \hat{\beta}_2) = -\frac{\bar{X}}{\sum (X_i - \bar{X})^2}.$$

[HINT: RECALL $\mathrm{Var}(\hat{\beta}) = \sigma^2 (X'X)^{-1}$ & IS $X'X$ IN THIS 2-VAR REGRESSION CASE

$$X'X = \begin{bmatrix} n & \sum X \\ \sum X & \sum X^2 \end{bmatrix}$$]

NOW SUPPOSE THAT WE WANT TO USE THE OLS PREDICTION IS VALUE OF Y FOR $X = X_0$. THE

$$\hat{Y}_0 = \hat{\beta}_1 + \hat{\beta}_2 X_0$$

THE ACTUAL VALUE Y_0 WILL RELATE TO X_0 AS

$$Y_0 = \beta_1 + \beta_2 X_0 + u_0$$

THUS THE PREDICTION ERROR WILL BE

$$e_0 = Y_0 - \hat{Y}_0 = -(\hat{\beta}_1 - \beta_1) - (\hat{\beta}_2 - \beta_2)X_0 + u_0$$

$$E(e_0) = -E[\hat{\beta}_1 - \beta_1] - X_0 E[\hat{\beta}_2 - \beta_2] + E(u_0)$$

$$= 0$$

THUS $\hat{Y}_0$ IS AN UNBIASED PREDICTOR OF Y_0

CONSIDER THE VARIANCE OF THE PREDICTION ERROR

$$\operatorname{Var}(e_0) = E\left[(\hat{\beta}_1 - \beta_1)^2\right] + X_0^2 E\left[(\hat{\beta}_2 - \beta_2)^2\right] + 2X_0 E\left[(\hat{\beta}_1 - \beta_1)(\hat{\beta}_2 - \beta_2)\right] + E\left[u_0^2\right]$$

(NOTE: u_0 IS INDEPENDENT PAST U_i'S)

$$= \operatorname{var}(\hat{\beta}_1) + X_0^2 \operatorname{var}(\hat{\beta}_2) + 2X_0 \operatorname{cov}(\hat{\beta}_1, \hat{\beta}_2) + \sigma_u^2$$

$$= \sigma_u^2 \left[\frac{1}{n} + \frac{\bar{X}^2}{\sum(X_i - \bar{X})^2}\right] + \sigma^2 \frac{X_0^2}{\sum(X_i - \bar{X})^2} - \frac{2X_0\bar{X}}{\sum(X_i - \bar{X})^2}\sigma_u^2 + \sigma_u^2$$

$$= \sigma_u^2 \left[1 + \frac{1}{n} + \frac{\bar{X}^2 + X_0^2 - 2X_0\bar{X}}{\sum(X_i - \bar{X})^2}\right]$$

$$= \sigma_u^2 \left[1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum(X_i - \bar{X})^2}\right]$$

OF COURSE, WE DO NOT KNOW σ_u^2 & MUST USE s_u^2 FOR CONFIDENCE INTERVALS & SIGNIFICANCE TESTS.

$$\frac{e_0 - E(e_0)}{se(e_0)} \sim t_{n-2}$$

THUS $$\frac{Y_0 - \hat{Y}_0}{se(e_0)} = t \sim t_{n-2}$$

NOTE THAT $\Pr\left[-t^*_{\alpha/2} \le t \le t^*_{\alpha/2}\right] = 1 - \alpha$

HENCE $$\Pr\left[-t^*_{\alpha/2} \le \frac{Y_0 - \hat{Y}_0}{se(e_0)} \le t^*_{\alpha/2}\right] = 1 - \alpha$$

HENCE $\Pr\left[-se(e_0)\, t^*_{\alpha/2} \le Y_0 - \hat{Y}_0 \le se(e_0)\, t^*_{\alpha/2}\right] = 1-\alpha$

OR $\Pr\left[\hat{Y}_0 - se(e_0)\, t^*_{\alpha/2} \le Y_0 \le \hat{Y}_0 + se(e_0) \cdot t^*_{\alpha/2}\right] = 1-\alpha$

THIS DEFINES A CONFIDENCE INTERVAL FOR THE PREDICTION.

WE NOW CONSIDER THE K-VARIABLE REGRESSION

$$y = X\beta + u$$

$$\hat{\beta} = (X'X)^{-1}X'y \qquad var(\hat{\beta}) = \sigma^2 (X'X)^{-1}$$

NOW CONSIDER THE PREDICTION OF y FOR $x = x_0$:

$$\hat{y}_0 = x_0 \hat{\beta}$$

AGAIN, $y_0 = x_0\beta + u_0$

$$\hat{y}_0 - y_0 = x_0(\hat{\beta} - \beta) + u_0$$

$$E(\hat{y}_0 - y_0) = x_0 E(\hat{\beta} - \beta) + E(u_0) = 0$$

$$Var(\hat{y}_0 - y_0) = x_0 E\left[(\hat{\beta} - \beta)(\hat{\beta} - \beta)'\right] x_0' + E(u_0^2)$$

$$= \sigma^2 \left[1 + x_0 (X'X)^{-1} x_0'\right]$$

AS BEFORE, WE USE $\hat{\sigma}^2$ & DEFINE THE STATISTIC

$$\frac{\hat{y}_0 - y_0}{se[(\hat{y}_0) - y_0]} \sim t_{n-k}$$

& GET THE RELEVANT CONFIDENCE INTERVAL