Prediction of a new observation : note that this is a single observation, not the regression line.

First, the variance of a generic linear combination (from Chapter 1:1.27a & b)

$$T = aW + bX + cZ$$

$$E(T) = aE(W) + bE(X) + cE(Z)$$

$$Var(T) = a^2Var(W) + b^2Var(X) + c^2Var(Z) + 2(\text{Covariances})$$

$$Var(T) = a^2Var(W) + b^2Var(X) + c^2Var(Z)$$
$$abCov(W,X) + bcCov(X,Z) + acCov(X,Z)$$

If we are able to assume that the three terms are stochastically independent, then the covariances are equal to zero.

We have already seen a series of Linear Combinations

1) First we saw,

$$b_1 = \frac{\Sigma(X_i - \bar{X})Y_i}{\Sigma(X_i - \bar{X})^2} = \sum_{i=1}^{n} k_i Y_i$$

$$\text{so, } Var(b_1) = k_1^2Var(Y_1) + k_2^2Var(Y_2) + k_3^2Var(Y_3) + \ldots$$

since all Y_i at all values of X_i have the same variance (homogeneous), then

$$= \sum_{i=1}^{n} k_i^2 \, Var(Y_i)$$

and recall that $\sum_{i=1}^{n} k_i^2 = \frac{1}{\Sigma(X_i - \bar{X})^2}$

and that $Var(Y_i)$ is estimated by the MSE, then

$$Var(b_1) = \frac{MSE}{\Sigma(X_i - \bar{X})^2}$$

1) Show that b_1 is a linear combination of k_i $\left(= \frac{(X_i - \bar{X})}{\Sigma(X_i - X)^2} \right)$

a) $b_1 = \frac{\Sigma(X_i - \bar{X})(Y_i - \bar{Y})}{\Sigma(X_i - X)^2}$

where

b) $\Sigma(X_i - \bar{X})(Y_i - \bar{Y}) = \Sigma(X_i Y_i - \bar{X} Y_i - X_i \bar{Y} + \bar{X}\bar{Y})$

$= \Sigma(X_i - \bar{X})Y_i - \Sigma(X_i - \bar{X})\bar{Y}$

$= \Sigma(X_i - \bar{X})Y_i - \bar{Y}\Sigma(X_i - \bar{X})$

and since

$\Sigma(X_i - \bar{X}) = 0$

then

c) $\Sigma(X_i - \bar{X})(Y_i - \bar{Y}) = \Sigma(X_i - \bar{X})Y_i$

as a result,

d) $b_1 = \frac{\Sigma(X_i - \bar{X})(Y_i - \bar{Y})}{\Sigma(X_i - X)^2} = \frac{\Sigma(X_i - \bar{X})Y_i}{\Sigma(X_i - X)^2} = \frac{\Sigma(X_i - \bar{X})}{\Sigma(X_i - X)^2} Y_i = \sum_{i=1}^{n} k_i Y_i$

where

e) $\Sigma k_i = \sum_{i=1}^{n} \frac{(X_i - \bar{X})}{\Sigma(X_i - X)^2} = \frac{\Sigma(X_i - \bar{X})}{\Sigma(X_i - X)^2}$

note that $\Sigma k_i = 0$ since $\Sigma(X_i - \bar{X}) = 0$

now prove that $\sum_{i=1}^{n} k_i^2 = \frac{1}{\Sigma(X_i - \bar{X})^2}$

f) $\Sigma k_i^2 = \sum_{i=1}^{n} \left[\frac{(X_i - \bar{X})}{\Sigma(X_i - X)^2} \right]^2 = \sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{[\Sigma(X_i - \bar{X})^2]^2}$

$= \frac{1}{[\Sigma(X_i - \bar{X})^2]^2} \Sigma(X_i - \bar{X})^2 = \frac{1}{\Sigma(X_i - \bar{X})^2}$

2) Then we say that $\hat{Y}_i = b_0 + b_1 X_i$, also a linear combination.

$$\mathrm{Var}(\hat{Y}_i) = 1*\mathrm{Var}(b_0) + X_i*\mathrm{Var}(b_1) + 2*1*X_i*\mathrm{Cov}(b_0,b_1)$$

note that we do NOT assume that b_0 and b_1 are independent.

The covariance is included, not equal to zero.

Using previous definitions of $\mathrm{Var}(b_0)$ and $\mathrm{Var}(b_1)$, and the Gaussian multipliers from the $(X'X)^{-1}$ matrix for the covariance

$$\mathrm{Var}(\hat{Y}_i)=\sigma^2*1^2\left[\frac{1}{n}+\frac{\bar{X}^2}{\Sigma(X_i - \bar{X})^2}\right]+\sigma^2*X_i^2\left[\frac{1}{\Sigma(X_i - \bar{X})^2}\right]+2*1*X_i\sigma^2\left[\frac{-\Sigma X_i}{n\Sigma(X_i - \bar{X})^2}\right]$$

$$\mathrm{Var}(\hat{Y}_i) =\sigma^2\left[\frac{1}{n}+\frac{\bar{X}^2}{\Sigma(X_i - \bar{X})^2}+\frac{X_i^2}{\Sigma(X_i - \bar{X})^2}-\frac{2X_i\bar{X}}{\Sigma(X_i - \bar{X})^2}\right]$$

$$\mathrm{Var}(\hat{Y}_i) = \sigma^2\left[\frac{1}{n}+\frac{(X_i-\bar{X})^2}{\Sigma(X_i-\bar{X})^2}\right]$$

3) Now we want a confidence interval for a single (new) observation.

The equation for that observation is

$$Y_i = b_0 + b_1 X_i + \epsilon_i$$

or

$$Y_i = \hat{Y}_i + \epsilon_i$$

We assumed independence once before (each Y_i independent of others). We are now going to assume independence again. We assume that the residuals are independent of the model (ie. assume that ϵ_i are independent of $\hat{Y}_i$).

So the variance of single observations will be

$$\mathrm{Var}(Y_i) = \mathrm{Var}(\hat{Y}_i) + \mathrm{Var}(\epsilon_i) + \left[2*\mathrm{Cov}(\hat{Y}_i,\epsilon_i) = 0\right]$$

We know from previous work that

$$\mathrm{Var}(\hat{Y}_i) = \sigma^2 \left[\frac{1}{n} + \frac{(X_i - \bar{X})^2}{\Sigma(X_i - X)^2}\right]$$

$$\mathrm{Var}(\epsilon_i) = \sigma^2$$

therefore

$$\mathrm{Var}(Y_i) = \sigma^2 \left[\frac{1}{n} + \frac{(X_i - \bar{X})^2}{\Sigma(X_i - X)^2}\right] + \sigma^2$$

or

$$\mathrm{Var}(Y_i) = \sigma^2 \left[1 + \frac{1}{n} + \frac{(X_i - \bar{X})^2}{\Sigma(X_i - X)^2}\right]$$

where the estimator of σ^2 is MSE

Note that both your textbook and I have been using σ^2 for both $\mathrm{Var}(Y_i)$ and for $\mathrm{Var}(\epsilon_i)$. Each is "the variance", but they are variances of different things. A better notation perhaps is $\mathrm{Var}(\epsilon_i) = \sigma^2_\epsilon$

There is another confidence interval of potential interest between

$$\mathrm{Var}(\hat{Y}_i) = \sigma^2 \left[\frac{1}{n} + \frac{(X_i - \bar{X})^2}{\Sigma(X_i - X)^2}\right] \text{, the regression line}$$

and

$$\mathrm{Var}(Y_i) = \sigma^2 \left[1 + \frac{1}{n} + \frac{(X_i - \bar{X})^2}{\Sigma(X_i - X)^2}\right] \text{, a new observation}$$

This is the confidence interval for the mean of a new sample taken at some particular value of X_i, where m is the size of the new sample. This cannot be as narrow as the confidence interval for the regression, but should be narrower than the confidence interval for a single sample. This CI is given by,

$$\mathrm{Var}(Y_i) = \sigma^2 \left[\frac{1}{m} + \frac{1}{n} + \frac{(X_i - \bar{X})^2}{\Sigma(X_i - X)^2}\right] \text{, } \bar{X} \text{ for a new sample}$$

Example : From *vial breakage regressed on number of airline transfers* example
Place a confidence interval on the breakage for 3 transfers for a single new observation.

$$s^2_{Y_i} = \text{MSE}\left[1 + \frac{1}{n} + \frac{(X_i - \bar{X})^2}{\Sigma(X_i - \bar{X})^2}\right]$$

$$= 2.2\left[1 + \frac{1}{10} + \frac{(3-1)^2}{20 - \frac{10^2}{10}}\right] = 2.2\left[1 + \frac{1}{10} + \frac{4}{10}\right] = 2.2*1.5 = 3.3$$

we previously calculated the variance of the regression line at $s^2_{\hat{Y}_i} = 1.1$. Note that the variance of a single point is $s^2_{\hat{Y}_i} + s^2 = 1.1 + 2.2 = 3.3$

$$s_{Y_i} = \sqrt{3.3} = 1.816$$

since $t_{\frac{\alpha}{2}, 8\,df} = 2.306$, then

$$P(\hat{Y}_{X=3} - t_{1-\frac{\alpha}{2}, n-2}\, s_{Y_i} \leq E(\hat{Y}) \leq \hat{Y}_{X=3} + t_{1-\frac{\alpha}{2}, n-2}\, s_{Y_i}) = 1\text{-}\alpha$$

$$P(22.2 - 2.306*1.816 \leq E(\hat{Y}) \leq 22.2 + 2.306*1.816) = 1\text{-}\alpha$$

$$P(18.011 \leq E(\hat{Y}) \leq 26.389) = 0.95$$

SAS will calculate confidence intervals for either the regression line (option CLM) or for individual points (option CLI). But not for a new sample.

Check this against the SAS output

Suppose were were to ship 4 cases through 3 transfers. What is the confidence interval for the mean breakage of 4 cases?

$$s^2_{\bar{Y}_i} = \text{MSE}\left(\frac{1}{m} + \frac{1}{n} + \frac{(X_i - \bar{X})^2}{\Sigma(X_i - X)^2}\right) = 2.2\left(\frac{1}{4} + \frac{1}{10} + \frac{(3-1)^2}{20 - \frac{10^2}{10}}\right)$$

$$= 2.2\left(\frac{1}{4} + \frac{1}{10} + \frac{4}{10}\right) = 2.2*0.75 = 1.65$$

$$s_{\bar{Y}_i} = \sqrt{1.65} = 1.2845$$

since $t_{\frac{\alpha}{2}, 8\,df} = 2.306$, then

$$P(\bar{Y}_{X=3} - t_{1-\frac{\alpha}{2},n-2}\, s_{\bar{Y}_i} \le E(\bar{Y}) \le \bar{Y}_{X=3} + t_{1-\frac{\alpha}{2},n-2}\, s_{\bar{Y}_i}) = 1\text{-}\alpha$$

$$P(22.2 - 2.306*1.2845 \le E(\bar{Y}) \le 22.2 + 2.306*1.2845) = 1\text{-}\alpha$$

$$P(19.238 \le E(\bar{Y}) \le 25.162) = 0.95$$

The MEAN of the 4 cases falls in this range.

The CI for the regression line is narrower

The CI for individual points is higher