Stat 400, section 12.4 Inferences about $\mu_{Y \cdot x^*}$, Prediction of Y Values

notes by Tim Pilachowski

So far, we have calculated point estimates for parameters β_0 and β_1, and denoted our calculated statistics as $\hat{\beta}_0$ and $\hat{\beta}_1$. We then formulated an estimate of the true linear regression equation, $Y = \beta_0 + \beta_1 x$, and denoted it as $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x^*$. Once we have tested the utility of our model, and found it to be statistically significant, the y-values generated can be justifiably regarded either as a point estimate of $\mu_{Y \cdot x^*}$, the expected or true average value of Y when $x = x^*$, or as a prediction of the Y value that will result from a single observation made when $x = x^*$.

The point estimate or prediction by itself gives no information concerning how precisely $\mu_{Y \cdot x^*}$ has been estimated or Y has been predicted. This can be remedied by developing a confidence interval (CI) for $\mu_{Y \cdot x^*}$ and a prediction interval (PI) for a single Y value.

Before we obtain sample data, both $\hat{\beta}_0$ and $\hat{\beta}_1$ are subject to sampling variability – that is, they are both statistics whose values will vary from sample to sample.

For example, suppose (as we did in Lecture 12.1 Example A-2) that the true regression line that relates MASC score (x) to Math grade (Y) is $Y = 100 - 0.5x$.
The sample data we used gave us $\hat{Y} = 103.8958 - 0.52061x^*$.
A different sample might have given us $\hat{Y} = 99.8034 - 0.51223x^*$, or $\hat{Y} = 100.1123 - 0.48654x^*$.

That is, $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x^*$ itself varies in value from sample to sample, so it is a statistic. If the true intercept and slope of the population line are the values 100 and –0.5, respectively, and $x^* = 50$, then this statistic is trying to estimate the value $100 - 0.5(50) = 75$.
The estimate from the sample we used would be $103.8958 - 0.52061(50) = 77.8653$.
The estimates from the other two possible samples given as examples above would be $99.8034 - 0.51223(50) = 74.1919$ and $100.1123 - 0.48654(50) = 75.7853$ respectively.

Visually, we could see the differences between the (supposed) true regression line and the three (supposed) generated estimated regression lines by graphing them on the same grid.

The text provides such a visualization for its Example 10, which used simulations to generate 20 estimated lines of regression.
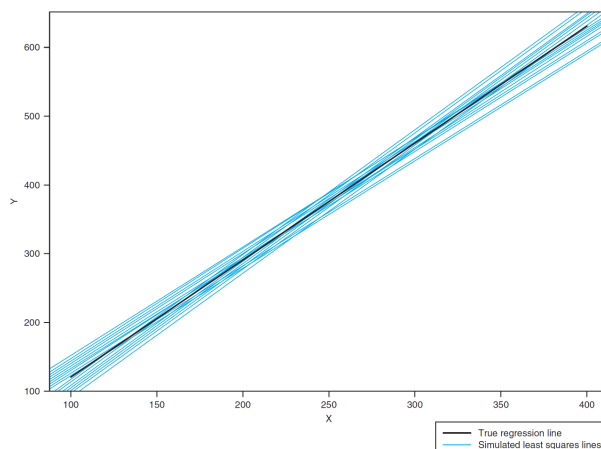
Note that there appears to be more variation in the value of $\hat{Y}$ on the far left and far right than there is "in the middle". That is, as we increase the distance of x^* from $\bar{x}$, we observe more variation in estimated/predicted values for random variable Y.

Methods for making inferences about β_1 were based on properties of the sampling distribution of the statistic $\hat{\beta}_1$. In the same way, inferences about the mean Y value $\beta_0 + \beta_1 x^*$ are based on properties of the sampling distribution of the statistic $\hat{\beta}_0 + \hat{\beta}_1 x^*$.

Substitution of the expressions for $\hat{\beta}_0$ and $\hat{\beta}_1$ into $\hat{\beta}_0 + \hat{\beta}_1 x^*$ followed by some algebraic manipulation leads to the representation of $\hat{\beta}_0 + \hat{\beta}_1 x^*$ as a linear function of the Y_i's.

$$\hat{\beta}_0 + \hat{\beta}_1 x^* = \sum_{i=1}^{n} \left[\frac{1}{n} + \frac{(x^* - \bar{x})(x_i - \bar{x})}{\sum (x_i - \bar{x})^2} \right] Y_i$$

Note that n, the x_i's and x^* are fixed values. Application of the change of variables rules of Lecture 3.6b (used again in chapter 5) to the linear function above gives the following properties.

Proposition:
Let $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x^*$ where x^* is some fixed value of x.
1. The mean value of $\hat{Y}$ is $E\left(\hat{Y}\right) = E\left(\hat{\beta}_0 + \hat{\beta}_1 x^*\right) = \mu_{\hat{\beta}_0 + \hat{\beta}_1 x^*} = \beta_0 + \beta_1 x^*$. That is, $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x^*$ is an unbiased estimator for $\mu_{Y \cdot x^*} = \beta_0 + \beta_1 x^*$.

2. The variance of $\hat{Y}$ is $V\left(\hat{Y}\right) = \left(\sigma_{\hat{Y}}\right)^2 = \sigma^2 \left[\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum (x_i)^2 - \dfrac{\left(\sum x_i\right)^2}{n}} \right] = \sigma^2 \left[\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}} \right]$ and $\sigma_{\hat{Y}} = \sqrt{V\left(\hat{Y}\right)}$.

The estimated standard deviation of $\hat{\beta}_0 + \hat{\beta}_1 x^*$, denoted by $s_{\hat{Y}}$ or $s_{\hat{\beta}_0 + \hat{\beta}_1 x^*}$, results from replacing σ with its estimate s: $s_{\hat{Y}} = s_{\hat{\beta}_0 + \hat{\beta}_1 x^*} = s \sqrt{ \frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}} }$.

3. $\hat{Y}$ has a normal distribution.

The variance of $\hat{\beta}_0 + \hat{\beta}_1 x^*$ is smallest when $x^* = \bar{x}$ and increases as x^* moves away from $\bar{x}$ in either direction. That is, the estimator of $\mu_{Y \cdot x^*} = \beta_0 + \beta_1 x^*$ is more precise when x^* is near the center of the x_i's than when it is far from the values at which observations have been made. This, in turn, implies that both the CI and PI are narrower for an x^* near $\bar{x}$ than for an x^* far from $\bar{x}$. Most statistical computer packages will provide both $\hat{\beta}_0 + \hat{\beta}_1 x^*$ and $s_{\hat{\beta}_0 + \hat{\beta}_1 x^*}$ for any specified x^* upon request.

Just as inferential procedures for β_1 were based on the t variable obtained by standardizing $\hat{\beta}_1$, a t variable obtained by standardizing $\hat{\beta}_0 + \hat{\beta}_1 x^*$ leads to a confidence interval and test procedures for our estimated linear regression equation.

Theorem: The variable $T = \dfrac{\hat{\beta}_0 + \hat{\beta}_1 x^* - (\beta_0 + \beta_1 x^*)}{S_{\hat{\beta}_0 + \hat{\beta}_1 x^*}} = \dfrac{\hat{Y} - (\beta_0 + \beta_1 x^*)}{S_{\hat{Y}}}$ has a t distribution with $n - 2$ degrees of freedom.

The rationales/proofs of this Theorem and the Proposition above both rely on the same observations about normality that we've already encountered and used in sections 12.1 through 12.3.

Appropriate algebraic manipulation of this T formula provides us a way to construct a confidence interval.

A $100(1 - \alpha)\%$ confidence interval for $\mu_{Y \cdot x^*} = \beta_0 + \beta_1 x^*$, the expected value of Y when $x = x^*$, is given by

$$\hat{\beta}_0 + \hat{\beta}_1 x^* \pm t_{\alpha/2, n-2}\left(s_{\hat{\beta}_0 + \hat{\beta}_1 x^*}\right) = \hat{y} \pm t_{\alpha/2, n-2}\left(s_{\hat{y}}\right).$$

This CI is centered at the point estimate for $\mu_{Y \cdot x^*}$ and extends out to each side by an amount that depends on the confidence level and on the extent of variability in the estimator on which the point estimate is based.

(Go to Example A-1 below.)

Tests of hypotheses about $\hat{\beta}_0 + \hat{\beta}_1 x$ are based on the test statistic T obtained by replacing $\beta_0 + \beta_1 x^*$ in the numerator of the T formula by the null value μ_0. The test would be upper-, lower-, or two-tailed according to the inequality in H_a.

Rather than calculate an interval estimate for $\mu_{Y \cdot x^*}$, an investigator may wish to obtain an interval of plausible values for the value of Y associated with some future observation when the independent variable has value x^*.

A CI refers to a parameter, or population characteristic, whose value is fixed but unknown to us. In contrast, a future value of Y is not a parameter but instead a random variable; for this reason we refer to an interval of plausible values for a future Y as a **prediction interval** rather than a confidence interval.

The error of estimation (i.e. for a CI) is $\beta_0 + \beta_1 x^* - \left(\hat{\beta}_0 + \hat{\beta}_1 x\right)$, a difference between a fixed (but unknown) quantity and a random variable.
The error of prediction (i.e. for a PI) is $Y - \left(\hat{\beta}_0 + \hat{\beta}_1 x\right)$, a difference between two random variables.

There is thus more uncertainty in prediction than in estimation, so a prediction interval will be wider than a confidence interval.

Because the future value Y is independent of the observed Y_i's,

$$\text{Variance of prediction error} = V\left(Y - \left(\hat{\beta}_0 - \hat{\beta}_1 x^*\right)\right)$$
$$= V(Y) + V\left(\hat{\beta}_0 - \hat{\beta}_1 x^*\right)$$
$$= \sigma^2 + \sigma^2\left[\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}\right]$$
$$\sigma^2\left[1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}\right].$$

Furthermore, because $E(Y) = \beta_0 + \beta_1 x^*$ and $E\left(\hat{\beta}_0 + \hat{\beta}_1 x^*\right) = \beta_0 + \beta_1 x^*$, the expected value of the prediction error is $E\left(Y - \left(\hat{\beta}_0 + \hat{\beta}_1 x^*\right)\right) = 0$. It can then be shown that the standardized variable $T = \dfrac{Y - \left(\beta_0 + \beta_1 x^*\right)}{S\sqrt{1 + \dfrac{1}{n} + \dfrac{(x^* - \bar{x})^2}{S_{xx}}}}$

has a t distribution with $n - 2$ degrees of freedom.

Appropriate algebraic manipulation of this T formula provides us a way to construct a prediction interval.

A $100(1 - \alpha)\%$ prediction interval for a future Y observation to be made when $x = x^*$, is given by

$$\hat{\beta}_0 + \hat{\beta}_1 x^* \pm t_{\alpha/2, n-2}\left(s\sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}}\right) = \hat{y} \pm t_{\alpha/2, n-2}\sqrt{s^2 + \left(s_{\hat{y}}\right)^2}.$$

The interpretation of the prediction level $100(1 - \alpha)\%$ is analogous to that of previous prediction levels. If the formula is used repeatedly, in the long run the resulting intervals will actually contain the observed y values $100(1 - \alpha)\%$ of the time.

Example A: A paper in *Measurement and Evaluation in Counseling and Development* (Oct 90, pp. 121–127) discussed a survey instrument called the *Mathematics Anxiety Scale for Children* (MASC). Suppose the MASC was administered to ten fifth graders with the following results:

MASC Score	67	37	70	40	35	65	40	35	30	40
Math grade (%)	75	85	60	90	80	75	70	90	95	80

From Lectures 12.1 and 12.2, we have

$$\bar{x} = 45.9,\ S_{xy} \approx -1075,\ S_{xx} \approx 2064.9,\ S_{yy} \approx 1000,\ \hat{\beta}_1 \approx -0.52061,\ \hat{\beta}_0 \approx 103.8958,\ \hat{\beta}_1 \approx -0.52061$$

$$\text{SSE} \approx 440.3482,\ \text{SST} \approx 1000,\ \text{SSR} \approx 559.6518,\ s^2 \approx 55.04353,\ r^2 \approx 0.559652.$$

1. Construct 95% confidence intervals for the mean Math grade for all students who have a MASC score of a) 50, and b) 35. Compare and contrast the two intervals.

2. Construct 95% prediction intervals for the mean Math grade for all students who have a MASC score of a) 50, and b) 35. Compare and contrast the two intervals with each other, and with the CIs constructed above.