STA 580 — Spring 2009 — Dr. Charnigo

Lecture 12

Simple linear regression

First motivation. In the last part of Lecture 11, I described how to construct point and interval estimates for the population correlation between two continuous variables X and Y. That methodology is useful if there is not an obvious choice for the response (dependent) variable and the explanatory (independent) variable. But what can we do if we regard Y as the response variable and X as the explanatory variable, especially if we are interested in predicting Y from X?

Second motivation. In Lecture 10, I described the one-way analysis of variance. That methodology is useful if we have a continuous variable Y whose mean may depend on the level of some factor. If X is a categorical variable representing the factor and has numerical designations for the factor levels, then the one-way analysis of variance is asking whether $m(x) := E[Y|X = x]$ depends on x.

For instance, in "Obstetrics" on page 620, Y is birthweight and X is the mother's smoking status. We can adopt the convention that $X = 1$ for mothers who never smoked, $X = 2$ for mothers who gave up smoking during pregnancy, $X = 3$ for mothers who smoked lightly during pregnancy, and $X = 4$ for mothers who smoked heavily during pregnancy. Then $m(1) = E[Y|X = 1]$ is the mean birthweight of infants born to mothers who never smoked, $m(2) = E[Y|X = 2]$ is the mean birthweight of infants born to mothers who gave up smoking during pregnancy, $m(3) = E[Y|X = 3]$ is the mean birthweight of infants born to mothers who smoked lightly during pregnancy, and $m(4) = E[Y|X = 4]$ is the mean birthweight of infants born to mothers who smoked heavily during pregnancy. Asking whether $m(1) = m(2) = m(3) = m(4)$ is equivalent to asking whether $m(x)$ depends on x.

Now suppose that X is a continuous variable instead of a categorical variable. Can we determine whether and how $m(x)$ changes with x?

Fixed versus random explanatory variable. When employing simple linear regression, the methodology to be described in this lecture, we will have measurements $x_1, \ldots, x_n$ and $y_1, \ldots, y_n$ for our n subjects (or other units of analysis). Before we choose the sample, there exist random conceptualizations of $y_1, \ldots, y_n$ that we denote $Y_1, \ldots, Y_n$. We say that the explanatory variable is random if there also exist random conceptualizations of $x_1, \ldots, x_n$ that we denote $X_1, \ldots, X_n$. We say that the explanatory variable is fixed if there do not exist such random conceptualizations.

In "Hypertension" on page 548, the explanatory variable is fixed because the investigators predetermined the ages at which the 90^{th} percentile of systolic blood pressure would be recorded. Specifically, the investigators decided *a priori* that x_1 should equal 1, x_2 should equal 2, and so forth. So $x_1, \ldots, x_n$ do not arise from a random process.

In "Environmental Health" on page 547, the explanatory variable is random because the investigators did not predetermine the volumes of traffic at which carbon monoxide concentrations would be recorded. Rather, the investigators observed the volumes of traffic and carbon monoxide concentrations simultaneously.

The following material on simple linear regression is theoretically predicated on the assumption that the explanatory variable should be fixed. In practice, however, I have never seen anybody decline to apply this material on account of having a random explanatory variable.

The statistical model. Our statistical model for "simple linear regression" is (Equation 11.2)

$$Y_i = \alpha + \beta x_i + \epsilon_i,$$

where the ϵ_i are independent normal random variables with mean 0 and (unknown) variance σ^2. Hence, the simple linear regression model entails the assumption that (Equation 11.1)

$$m(x) = \alpha + \beta x.$$

Why make such an assumption? First, we have reduced the problem of estimating $m(x)$ at all possible values of x to the vastly simpler problem of estimating two parameters. Second, any smooth function can be locally approximated by a linear function.[1]

Interpreting the regression coefficients. We refer to α and β as regression coefficients. More specifically, we call α the intercept (coefficient) and β the slope (coefficient). The intercept is the mean response when $X = 0$ (if having $X = 0$ is possible) and, in a graphical display, represents the position at which the "regression line" $m(x) = \alpha + \beta x$ crosses the vertical axis. The slope is the change in the mean response associated with a one-unit increase in X and, in a graphical display, represents rise over run for the regression

[1] Let x_0 be a typical value for X and consider the Taylor expansion

$$m(x) = m(x_0) + (x - x_0)\, m'(x_0) + \frac{(x - x_0)^2}{2}\, m''(x_0) + r(x),$$

where $r(x)$ is an appropriate remainder term. If $|m''(x_0)|$ is small and x is close to x_0, then

$$m(x) \approx \alpha + \beta x \quad \text{with} \quad \alpha := m(x_0) - x_0\, m'(x_0) \quad \text{and} \quad \beta := m'(x_0)$$

constitutes an excellent approximation.

line. A positive slope implies a tendency for large values of Y to accompany large values of X (Figure 11.3-a), a negative slope implies a tendency for small values of Y to accompany large values of X (Figure 11.3-b), and a zero slope implies neither tendency (Figure 11.3-c).

The principle of least squares. We can estimate α and β using the principle of least squares. Let r_0 and r_1 denote guesses for α and β. We can assess how good these guesses are by considering the sum of squares $\sum_{i=1}^{n}(y_i - r_0 - r_1 x_i)^2$, which is the sum of squared vertical distances from the data points to the line $r_0 + r_1 x$. If $r_0 + r_1 x$ is close to the regression line $\alpha + \beta x$, then the sum of squares will be small; if $r_0 + r_1 x$ is not close to the regression line, then the sum of squares will be large. The principle of least squares says that the best guesses for α and β are the values of r_0 and r_1 for which the sum of squares is minimized (Figure 11.4). Thus, letting a and b denote the values of r_0 and r_1 for which the sum of squares is minimized, we adopt a and b as our estimates of α and β.

To clarify the preceding, let us revisit "Hypertension" on page 548. If we take $r_0 = 114.94$ and $r_1 = 0$, then (one can show that) the sum of squares equals 1586.94. If we take $r_0 = 97.382$ and $r_1 = 1.951$, then the sum of squares equals 33.96. Moreover, there is no way to make the sum of squares smaller than 33.96. Hence, $a = 97.382$ and $b = 1.951$ are the least squares estimates of α and β.

Formulas for the least squares estimates. Let L_{xx}, L_{yy}, and L_{xy} be as defined in Lecture 11. The least squares estimates of α and β are (Equation 11.3)

$$b = L_{xy}/L_{xx} \quad \text{and} \quad a = \bar{y} - b\bar{x} = \left(\sum_{i=1}^{n} y_i - b \sum_{i=1}^{n} x_i \right) / n.$$

Inferences about regression coefficients

Sums of squares. The quantity $\hat{y} := a + bx$ is both an estimate of $m(x)$ and a prediction for a value of the response given that the explanatory variable has value x. We refer to $\hat{y}$ as a "fitted value" or a "predicted value", with some preference for fitted value in the former context and for predicted value in the latter context. If $x = x_i$, then the corresponding $\hat{y}$ is often denoted $\hat{y}_i$.

As with the one-way analysis of variance in Lecture 10, we may decompose a gross measure of variability into parts accounted for by X and not accounted for by X (Equation 11.5, Figure 11.5):

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

or

$$\text{Tot SS} = \text{Reg SS} + \text{Res SS}.$$

Above, Reg SS quantifies the improvement in prediction attained by considering X, rather than ignoring X and making the same prediction for everyone, while Res SS quantifies how far off the improved predictions are from the actual values of the response.

Computational formulas for desk calculators. We have (Equation 11.6) Tot SS $= L_{yy}$, Reg SS $= L_{xy}^2/L_{xx}$, and Res SS $=$ Tot SS $-$ Reg SS.

ANOVA test for model significance. Let Reg MS $:=$ Reg SS/1, and let Res MS $:=$ Res SS$/(n-2)$. Reject $H_0 : \beta = 0$ in favor of $H_1 : \beta \neq 0$ if $f :=$ Reg MS/Res MS $> f_{1,n-2,1-\alpha}$ (Equation 11.7). Rejection of H_0 implies that X is useful in predicting Y.

Example (Sums of squares, ANOVA test for model significance). Refer to "Hypertension" on page 548 and {SLRExample.pdf}. Page 3 of {SLRExample.pdf} shows a plot of the data points with the estimated regression line $97.382 + 1.951x$ superimposed. The estimated regression line provides an excellent fit to the data points.

The "Analysis of Variance" box on page 1 of {SLRExample.pdf} shows that Reg SS $= 1552.98$, Res SS $= 33.96$, and Tot SS $= 1586.94$. Moreover, Reg MS $= 1552.98$ and Res MS $= 2.264$, so that $f = 685.93$. The corresponding p-value is less than 0.0001, so we reject $H_0 : \beta = 0$ and conclude that X is useful in predicting Y.

In the next box on page 1, Root MSE is the square root of Res MS (an estimate for σ), Dependent Mean is simply $\bar{y}$, and Coeff Var is Root MSE divided by Dependent Mean expressed as a percentage.

Test for the slope coefficient. To test $H_0 : \beta = \beta_0$ against $H_1 : \beta \neq \beta_0$, compute

$$t := \frac{b - \beta_0}{\sqrt{\text{Res MS}/L_{xx}}}$$

and reject H_0 if $|t| > t_{n-2,1-\alpha/2}$. If $\beta_0 = 0$ (Equation 11.8), this procedure is equivalent to the ANOVA test for model significance since, in that case, $t^2 = f$ and $t^2_{n-2,1-\alpha/2} = f_{1,n-2,1-\alpha}$.

Example (test for the slope coefficient). In the "Parameter Estimates" box on page 1 of {SLRExample.pdf}, we see that $a = 97.382$ and $b = 1.951$. Moreover, $\sqrt{\text{Res MS}/L_{xx}}$ equals 0.0745. Hence, with $\beta_0 = 0$,

$$t = \frac{1.951}{0.0745} = 26.19,$$

for which the corresponding p-value is less than 0.0001. As before, we reject $H_0 : \beta = 0$.

Test for the intercept coefficient. To test $H_0 : \alpha = \alpha_0$ against $H_1 : \alpha \neq \alpha_0$, compute

$$t := \frac{a - \alpha_0}{\sqrt{\text{Res MS}(1/n + \bar{x}^2/L_{xx})}}$$

and reject H_0 if $|t| > t_{n-2,1-\alpha/2}$.[2]

Confidence intervals for regression coefficients. The $100(1-\alpha)\%$ confidence intervals for the intercept α and slope β are (Equation 11.10)

$$a \pm t_{n-2,1-\alpha/2}\sqrt{\text{Res MS}(1/n + \bar{x}^2/L_{xx})}$$

and

$$b \pm t_{n-2,1-\alpha/2}\sqrt{\text{Res MS}/L_{xx}}.$$

Prediction and estimation of the mean response

Prediction. Suppose that we want to predict a future value of the response given that the explanatory variable will have value x. If we had to provide a

[2] Using α in two different ways (intercept coefficient, significance level) is poor but conventional notation. Fortunately, there seems to be little potential for confusion.

single number, we would quote $\hat{y} = a + bx$. However, just as we can supply a confidence interval along with a point estimate, we can supply a prediction interval along with a predicted value. The $100(1-\alpha)\%$ prediction interval is (Equation 11.11)

$$\hat{y} \pm t_{n-2,1-\alpha/2}\sqrt{\text{Res MS}(1 + 1/n + [x-\bar{x}]^2/L_{xx})}.$$

Example (prediction). Suppose that we plan to take systolic blood pressure measurements on another group of 16 year-old boys. Let Y denote the 90^{th} percentile in systolic blood pressure for this group of 16 year-old boys. A single-number prediction is

$$\hat{y} = 97.382 + 1.951 \cdot 16 = 128.6,$$

as reported in the "Predicted Value" column on page 2 of {SLRExample.pdf} in row 16. The limits of the 95% prediction interval may be found in the "95% CL Predict" columns as 125.1 and 132.1.

Estimation of the mean response. Suppose that we want an interval estimate of $m(x)$ for a specific x. The $100(1-\alpha)\%$ confidence interval is constructed around the point estimate $\hat{y}$ and has the form (Equation 11.12)

$$\hat{y} \pm t_{n-2,1-\alpha/2}\sqrt{\text{Res MS}(1/n + [x-\bar{x}]^2/L_{xx})}.$$

This is identical to the formula for the prediction interval except that the "1" under the square root is absent.

Example (estimation of the mean response). In the "95% CL Mean" columns on page 2 of {SLRExample.pdf}, we see that the limits of the 95% confidence interval for $m(16)$ are 127.2 and 130.0. Hence, the average of the 90^{th} percentile in systolic blood pressure over all possible groups of 16 year-old boys is believed to lie between 127.2 and 130.0, even though the prediction interval for a specific group of 16 year-old boys is 125.1 to 132.1. The intuitive explanation for the wider prediction interval is that the next group of 16 year-old boys to be observed may or may not be typical.

Goodness of fit

Introduction. After fitting a simple linear regression model, we may want to assess "goodness of fit". This actually entails two questions:

• How much of the variability in the response is accounted for by the simple linear regression model (i.e., by the explanatory variable)?

• Is the simple linear regression model appropriate? That is, are the underlying assumptions reasonable?

Explaining variability in the response. The fraction of variability accounted for by the simple linear regression model is (Definition 11.14)

$$R^2 := \text{Reg SS/Tot SS} = 1 - \text{Res SS/Tot SS}.$$

A few comments are in order.

• One can show that $R^2 = r^2$, the square of the sample (Pearson) correlation between X and Y.

• We have $0 \leq R^2 \leq 1$, with $R^2 = 1$ implying that the data fall exactly on the estimated regression line.

• There is no universal standard for what constitutes a "good" R^2. In some physical science problems, one may have R^2 above 0.90. In some behavioral science problems, one may have R^2 less than 0.30.

Example (explaining variability in the response). Referring to page 1 of {SLRExample.pdf}, we see that $R^2 = 0.9786$. Hence, almost 98% of the variability in the 90^{th} percentile of systolic blood pressure is accounted for by the simple linear regression model (i.e., accounted for by age).

Checking model assumptions. In testing $H_0 : \mu_1 = \mu_2$ against $H_1 : \mu_1 \neq \mu_2$ in Chapter 8, we could use box plots to determine whether the underlying assumptions were reasonable. We could also construct box plots to determine whether the assumptions underlying the one-way ANOVA and two-way ANOVA in Chapter 12 were tenable. Unfortunately, informative box plots cannot generally be constructed when we have a continuous explanatory variable.[3] Hence, we must proceed differently to check the modeling assumptions in Chapter 11.

The main idea is to define and analyze quantities called residuals. A residual is the difference between the actual value of the response for someone in the sample and the predicted value for that person: $y_i - \hat{y}_i = y_i - (a + bx_i)$. Graphically, a residual is the (appropriately signed) vertical distance from a data point to the estimated regression line. With some regret, I must leave the details on checking linear regression modeling assumptions to your next statistics course (CPH 630/STA 681, CPH 930, or STA 671).

[3]With a very large sample size, one can discretize X and then create box plots.