MAT 2377 3X (Spring 2011)
Simple Linear Regression (Inference)
Sections 11.4.1, 11.5, 11.6, 11.8

§11.4: t-test concerning β_0 or β_1.

The **Simple Linear Regression Model** with **normal random errors** is

$$Y = \beta_0 + \beta_1 x + \epsilon,$$

where β_0 and β_1 are unknown constants, x is a value taken by the predictor X and ϵ is **random error**.

We will assume that ϵ is a **normal** random variable with mean 0 and variance σ^2. That is,

$$E(\epsilon) = 0 \qquad \text{and} \qquad V(\epsilon) = \sigma^2$$

Consequences:

- Y is follows $N(\beta_0 + \beta_1 x, \sigma^2)$ distribution.
- The estimators $\widehat{\beta}_0$ and $\widehat{\beta}_0$ are linear combinations of $Y_1, \ldots, Y_n$, that is linear combinations of independent normals, thus they are both normal random variables. That is,

$$\widehat{\beta}_0 \sim N(\beta_0, \sigma^2_{\widehat{\beta}_0}) \quad \text{and} \quad \widehat{\beta}_0 \sim N(\beta_0, \sigma^2_{\widehat{\beta}_0}).$$

- The standardized estimators are standard normals, i.e.

$$\frac{\widehat{\beta}_0 - \beta_0}{\sigma_{\widehat{\beta}_0}} \sim N(0,1) \quad \text{and} \quad \frac{\widehat{\beta}_1 - \beta_1}{\sigma_{\widehat{\beta}_1}} \sim N(0,1).$$

- As we replace the standard errors by the estimated standard errors we get t random variables with $\nu = n - 2$ degrees of freedom, that is

$$\frac{\widehat{\beta}_0 - \beta_0}{\widehat{\sigma}_{\widehat{\beta}_0}} \sim t(n-2) \quad \text{and} \quad \frac{\widehat{\beta}_1 - \beta_1}{\widehat{\sigma}_{\widehat{\beta}_1}} \sim t(n-2),$$

where

$$\widehat{\sigma}_{\widehat{\beta}_1} = \sqrt{\frac{\hat{\sigma}^2}{S_{xx}}} \quad \text{and} \quad \widehat{\sigma}_{\widehat{\beta}_0} = \sqrt{\hat{\sigma}^2 \left[\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}\right]}.$$

t-tests concerning β_0: Suppose that we have a hypothesis test with the following null hypothesis $H_0 : \beta_0 = \beta_{0,0}$ where $\beta_{0,0}$ is some real number. We will use the following test statistic

$$T_0 = \frac{\widehat{\beta}_0 - \beta_{0,0}}{\widehat{\sigma}_{\widehat{\beta}_0}} = \frac{\widehat{\beta}_0 - \beta_{0,0}}{\sqrt{\widehat{\sigma}^2 \left[\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}\right]}}.$$

where T_0 follows a t distribution with $\nu = n - 2$ degrees of freedom when H_0 is true.

t-tests concerning β_1: Suppose that we have a hypothesis test with the following null hypothesis $H_0 : \beta_1 = \beta_{1,0}$ where $\beta_{1,0}$ is some real number. We will use the following test statistic

$$T_0 = \frac{\widehat{\beta}_1 - \beta_{1,0}}{\widehat{\sigma}_{\widehat{\beta}_1}} = \frac{\widehat{\beta}_1 - \beta_{1,0}}{\sqrt{\widehat{\sigma}^2 / S_{xx}}}.$$

where T_0 follows a t distribution with $\nu = n - 2$ degrees of freedom when H_0 is true.

Test For the Significance of the Regression: The following test allows us to test how significance of the predictor X in predicting the response Y.

$$H_0 : \beta_1 = 0 \quad \text{against} \quad H_1 : \beta_1 \neq 0.$$

Interpretation:

- Failure to reject H_0 means that there is no linear relationship between Y and X.
- Rejecting H_0 means the linear relationship between Y and X is significant.

Example 1: Consider the data from Examples 1, 2 and 3 from Notes 14. Recall that the point estimates for β_0 and β_1 are respectively $\widehat{\beta}_0 = 0.069242$ and $\widehat{\beta}_1 = 0.003829$. Furthermore, the estimated standard errors are $\widehat{\sigma}_{\widehat{\beta}_0} = 0.100974$ and $\widehat{\sigma}_{\widehat{\beta}_1} = 0.000438$.

(a) Test for the significance of the regression with $\alpha = 5\%$.

(b) Do the data support the claim that $\beta_0 > 0.1$ at a level of significance of 5%?

§11.5: Interval Estimation

We known that

$$\frac{\widehat{\beta}_0 - \beta_0}{\widehat{\sigma}_{\widehat{\beta}_0}} \sim t(n-2) \quad \text{and} \quad \frac{\widehat{\beta}_1 - \beta_1}{\widehat{\sigma}_{\widehat{\beta}_1}} \sim t(n-2),$$

where

$$\widehat{\sigma}_{\widehat{\beta_1}} = \sqrt{\frac{\hat{\sigma}^2}{S_{xx}}} \quad \text{and} \quad \widehat{\sigma}_{\widehat{\beta_0}} = \sqrt{\hat{\sigma}^2 \left[\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}\right]}.$$

Hence a $(1-\alpha) \times 100\%$ confidence interval for β_0 is

$$\widehat{\beta}_0 \pm t_{\alpha/2,n-2}\, \widehat{\sigma}_{\widehat{\beta_0}} = \widehat{\beta}_0 \pm t_{\alpha/2,n-2} \sqrt{\hat{\sigma}^2 \left[\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}\right]},$$

and a $(1-\alpha) \times 100\%$ confidence interval for β_1 is

$$\widehat{\beta}_1 \pm t_{\alpha/2,n-2}\, \widehat{\sigma}_{\widehat{\beta_1}} = \widehat{\beta}_1 \pm t_{\alpha/2,n-2} \sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}.$$

Example 2: Consider the data from Examples 1, 2 and 3 from Notes 14. Recall that the point estimates for β_0 and β_1 are respectively $\widehat{\beta}_0 = 0.069242$ and $\widehat{\beta}_1 = 0.003829$. Furthermore, the estimated standard errors are $\widehat{\sigma}_{\widehat{\beta}_0} = 0.100974$ and $\widehat{\sigma}_{\widehat{\beta}_1} = 0.000438$.

(a) Construct a 95% confidence interval for β_0.

(b) Give a 95% confidence interval for β_1.

Estimating the mean response

Given a specified value of the predictor X, say x_0, we would like to estimate the mean response, that is

$$\mu_{Y|x_0} = \beta_0 + \beta_1 \, x_0.$$

We can use the value on the estimated regression line as a point estimate, i.e.

$$\widehat{\mu}_{Y|x_0} = \widehat{\beta}_0 + \widehat{\beta}_1 \, x_0.$$

Properties of the estimated mean response:

- Its expectation is

$$E[\widehat{\mu}_{Y|x_0}] = \beta_0 + \beta_1 \, x_0 = \mu_{Y|x_0}.$$

  Hence it is unbiased for estimating $\mu_{Y|x_0}$.

- Its variance is

$$V[\widehat{\mu}_{Y|x_0}] = \sigma^2 \, \left[\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}} \right]$$

- Standardization with the estimated standard error:

$$\frac{\widehat{\mu}_{Y|x_0} - \mu_{Y|x_0}}{\sqrt{\widehat{\sigma}^2 \, \left[\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}} \right]}} \sim t(n-2).$$

Interval Estimation:

A $100\,(1-\alpha)\%$ confidence interval for the mean response at a value $x = x_0$, say $\mu_{Y|x}$, is

$$\widehat{\mu}_{Y|x_0} \pm t_{\alpha/2, n-2} \sqrt{\widehat{\sigma}^2 \, \left[\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}} \right]}.$$

§Section 11.6: Prediction of new observations

Goal: To predict a new or future response Y_0 corresponding to a specified level x_0 of the predictor.

Prediction: We can use the following

$$\widehat{Y}_0 = \widehat{\beta}_0 + \widehat{\beta}_1 \, x_0$$

as a point estimator of the new or future value of the response Y_0.

Error in Prediction: We will define the error in prediction as

$$\mathbf{e} = Y_0 - \widehat{Y}_0.$$

The expectation of the error in prediction is

$$E[\mathbf{e}] = E[Y_0] - E[\widehat{Y}_0] = (\beta_0 + \beta_1 \, x_0) - (\beta_0 + \beta_1 \, x_0) = 0$$

and the variance of the error in prediction is

$$V[\mathbf{e}] = V[Y_0] + V[\widehat{Y}_0] = \sigma^2 + \sigma^2 \left[\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{XX}} \right] = \sigma^2 \left[1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{XX}} \right],$$

since we are assuming that new or future value Y_0 is independent of the current observations $Y_1, \ldots, Y_n$.

If we use $\widehat{\sigma}^2$ to estimate σ^2, it can be shown that, then

$$\frac{Y_0 - \widehat{Y}_0}{\sqrt{\widehat{\sigma}^2 \, \left[1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{XX}} \right]}} \sim t(n-2).$$

A $100\,(1-\alpha)\%$ prediction interval for new or future value response Y_0 are the value x_0 is given by

$$\widehat{y}_0 \pm t_{\alpha/2, n-2} \sqrt{\widehat{\sigma}^2 \, \left[1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{XX}} \right]}$$

Example 3: Consider the data from Examples 1, 2 and 3 from Notes 14. Recall that the estimated regression line is

$$\widehat{y} = 0.069242 + 0.003829\, x,$$

the point estimate for σ^2 is $\widehat{\sigma}^2 = 0.025298$. Furthermore, $n = 10$, $S_{xx} = 132000$ and $\overline{x} = 200$.

(a) Give a 95% confidence interval for the mean evaporation coefficient at a velocity of $x_0 = 140$.

(b) Give a 95% prediction interval for a new or future evaporation coefficient at a velocity of $x_0 = 140$.

§**Section 11-8: Correlation Analysis**

Scenario: We will assume that both X and Y are random variables. We would like to measure the linear association between the two random variables.

Definition: The covariance between X and Y is

$$\sigma_{XY} = E[(X - \mu_X)(Y - \mu_Y)].$$

Properties:

1. If X and Y are independent, then $\sigma_{XY} = 0$.
2. If there is a (statistical) linear association between X and Y, then the sign of the covariance should be the same as the sign of the slope of the line.

Definition: The correlation coefficient between X and Y is

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \, \sigma_Y}.$$

Properties:

1. $-1 \leq \rho_{XY} \leq 1$
2. It is only equal to 1 or -1, if the points fall exactly on a line with probability 1.
3. If X and Y are independent, then $\rho_{XY} = 0$.
4. If there is a (statistical) linear association between X and Y, then the sign of the correlation should be the same as the sign of the slope of the line.

These properties motivate the use of the correlation as a measure of the strength of the linear association between X and Y. In practice, the joint distribution of X and Y is unknown so we must estimate ρ_{XY}.

Consider the following random sample $(X_1, Y_1), \ldots, (X_n, Y_n)$, we define the **sample correlation coefficient** as

$$R = \frac{\sum_{i=1}^{n} (X_i - \overline{X})\,(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n} (X_i - \overline{X})^2 \cdot \sum_{i=1}^{n} (Y_i - \overline{Y})^2}} = \frac{S_{XY}}{\sqrt{S_{XX}\, S_{YY}}}.$$

Remark: Recall that the slope of the estimated regression line is

$$\widehat{\beta}_1 = \frac{S_{XY}}{S_{XX}}.$$

Thus,

$$R = \frac{S_{XY}}{\sqrt{S_{XX}\, S_{YY}}} = \frac{S_{XY}}{S_{XX}} \sqrt{\frac{S_{XX}}{S_{YY}}} = \widehat{\beta}_1 \sqrt{\frac{S_{XX}}{S_{YY}}}.$$

Hence R and $\widehat{\beta}_1$ are closely related.

Testing $\rho = 0$: Suppose that we would like to test

$$H_0 : \rho = 0 \quad \text{against} \quad H_1 : \rho \neq 0.$$

We will use the following test statistic

$$T_0 = \frac{R\sqrt{n-2}}{\sqrt{1-R^2}} = \frac{\widehat{\beta}_1 - 0}{\widehat{\sigma}_{\widehat{\beta}_1}} \sim t(n-2).$$

Example 4: Consider the data from Examples 1, 2 and 3 from Notes 14. Recall that

$$S_{xx} = 132000, \quad S_{xy} = 505.4, \quad S_{yy} = 2.13745.$$

(a) Compute the sample correlation coefficient between X and Y.

(b) Test $H_0 : \rho_{XY} = 0$ against $H_1 : \rho_{XY} \neq 0$ at $\alpha = 5\%$.