Lesson 18

Simple Linear Regression

Learning Objectives

Upon completion of this lesson you should be able to do the following:

1. Recognize the equation that fits a line to bivariate data.
2. Calculate the least squares estimates of the slope and intercept for the straight line model.
3. Assess the quality of the linear model by considering the random error terms.
4. Estimate the variance of the random error terms.
5. Calculate the standard error of the slope estimate.
6. Generate an interval estimate for the slope parameter.
7. Perform a hypothesis test concerning the slope parameter.
8. Calculate and interpret the estimate of the linear correlation.
9. Calculate and interpret the estimate of the coefficient of determination.

Key Words

bivariate, model, dependent, independent, random error, y-intercept, slope, least squares, best fitting regression line, variance of the random error, inferences about the slope, coefficient of correlation, coefficient of determination

Concepts

This lesson presents a **bivariate** statistical method that analyzes and describes the relationship between two linearly related variables. The best-fitting simple linear equation will be estimated and used to describe the linear relationship between the two variables. The regression equation can be used to estimate the values of one variable based on values of the other variable.

Simple Linear Model

The linear model that fits a line to bivariate data has two parts. One part is just the equation of a line, y-intercept plus the slope times the value of the x-variable, and forms the deterministic portion of the model. The other part is the random error component in the model. The **model** is a simple linear equation that describes the relationship between the two variables.

Linear Regression Model with one x-term: $y = \beta_o + \beta_1 x + e$

where

y = ***dependent***, *response, or predicted variable.*

x = ***independent*** *or predictor variable.*

e = ***random error*** *component.*

β_o = ***y-intercept*** *of the true population regression line.*

β_1 = ***slope*** *of the true population regression line.*

The **slope** and the ***y*-intercept** for the line that best fits the bivariate data points are estimated from the sample data. The **least-squares** regression method generates the line that fits the data the best in terms of least squared vertical distance from the data values to the estimated regression line. The estimates for and are used in the following equation to calculate , the estimate of y, for a specific *x*-value.

Estimated Linear Regression Model: $\hat{y} = \hat{\beta}_o + \hat{\beta}_1 x$

where

$\hat{\beta}_o$ is the least-squares estimate of $\beta_o$and

$\hat{\beta}_1$ is the least-squares estimate of β_1.

The equations shown for the estimators of the slope and y-intercept generate the values that produce the **best fitting regression line** in terms of least squared vertical error from the data points to the line.

Fitting the Least Squares Regression Line

The following three calculation steps produce the least squares estimators for the slope and intercep

1. Calculate the sufficient sums.

$\sum x, \sum y, \sum x^2, \sum y^2$, and $\sum xy$

2. Calculate the three corrected sums of squares.

$$SS_x = \sum (x - \overline{x})^2 = \sum x^2 - \frac{\left(\sum x\right)^2}{n}.$$

$$SS_y = \sum (y - \overline{y})^2 = \sum y^2 - \frac{\left(\sum y\right)^2}{n}.$$

$$SS_{xy} = \sum (x - \overline{x})(y - \overline{y}) = \sum xy - \frac{\left(\sum x\right)\left(\sum y\right)}{n}.$$

3. The least squares estimators for the slope and intercept are:

$\hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}}$, $\hat{\beta}_o = \overline{y} - \hat{\beta}_1 \overline{x}$.

The estimates that result from the above equations for the slope and intercept provide

the best fitting regression line in terms of minimizing the squared vertical distance from the data to the estimated regression equation. The criterion for the best fitting line is that the sum of the squared differences between the observed y values and the estimated y values for each x-value is minimized. The slope and intercept equations above produce a line that minimizes the amount The sum of the squared vertical error from data to line is the least it can be for any possible line.

The assumptions for the model are based on the error term, . The errors, are assumed to be normally distributed random variables with mean equal to zero and common variance, σ_ε^2 The errors associated with any two observations are also assumed to be independent. The observed error or residual amount is the difference between the *y*-values observed in the data, denoted *y*, and the least squares estimate of y at some value of *x*, denoted $\hat{y}_{x=x_o}$. The random error term is ε.

The expected value of the random error term is zero, but each of the observed individual error terms differs from zero. The **variance of the random error** measures the dispersion of the error terms around the expected value of zero in terms of squared units. The symbol denotes the variance of the residual or error terms around the value 0. The square root of this variance, σ_ε, is the standard deviation of the residuals, which measures the dispersion of the error terms around the expected value of zero.

The estimate for the variance of the residuals, $\hat{\sigma}_\varepsilon^2$, is:

$$\hat{\sigma}_\varepsilon^2 = S_\varepsilon^2 = \frac{\sum(y-\hat{y})^2}{n-2} = \frac{SS_y - \hat{\beta}_1 SS_{xy}}{n-2}.$$

Inferences on the Slope

The equations for the estimates of the slope and intercept for the best fitting regression line were shown in the prior section. The spread of the data about the line was estimated with the variance of the residuals. The estimate of the intercept indicates what predicted value for y would occur when the x variable has the value zero. The estimate of the slope indicates how much y is expected to change for a one-unit increase in the x variable. The estimated slope is a gauge of the effect of the independent variable, X, on the dependent variable, Y. What is implied about the relationship between the variables if the slope of the regression equation was not significantly different from zero? What implications about the relationship between *x* and *y* could be drawn if the regression line was a flat horizontal line?

Consider the slope as a population parameter. The point estimator for the slope is the least squares estimator discussed on the prior page. The standard error of the estimator of the slope needs to be considered. The standard error of the slope estimate is directly affected by the spread of the data about the regression line. That is the spread estimated by the variance of the residuals, so that value appears in the equation for the standard error of the slope. **Inferences about the slope** includes estimating the slope with a confidence interval and testing values of the slope in a hypothesis test.

Parameter is β_1

Point Estimator for β_1 is $\hat{\beta}_1$

Standard Error of $\hat{\beta}_1$ is $\sigma_{\hat{\beta}_1} = \sqrt{\frac{\sigma_\varepsilon^2}{SS_x}}$

Estimated Standard Error of $\hat{\beta}_1$ is $\hat{\sigma}_{\hat{\beta}_1} = S_{\hat{\beta}_1} = \sqrt{\frac{S_\varepsilon^2}{SS_x}}$

$(1-\alpha)100\%$ Confidence Interval to Estimate β_1 is $\hat{\beta}_1 \pm t_{\frac{\alpha}{2}(n-2)} \cdot S_{\hat{\beta}_1}$.

Test Statistic To Test $H_o : \beta_1 = \beta_{1_o}$ is $t = \frac{\hat{\beta}_1 - \beta_{1_o}}{S_{\hat{\beta}_1}} \sim t(n-2)$ if H_o is true.

Linear Correlation

How well does a line describe the relationship between *x* and *y*? If the bivariate data lie close to a line the linear relationship between the two variables is strong. In that circumstance a regression line is a good estimator for *y* based on values of the *x* variable. The linear correlation between *x* and *y* can be estimated to address whether a line is a good model to describe the relationship between the variables.

The Pearson product moment **coefficient of correlation**, *r*, estimates the linear correlation between *x* and *y*. The calculation of Pearson's *r* is short once you have the corrected sums of squares, SS_x, SS_y, SS_{xy}.

$$r = \frac{SS_{xy}}{\sqrt{SS_x \cdot SS_y}}$$

The value of *r*, the estimated linear correlation, provides information about the strength and the direction of the linear relationship between *x* and *y*. The possible range of values for *r* is between –1 and +1, including those values. The magnitude of *r* gives information about the strength of the linear relationship between *x* and *y*. As the value of *r* gets closer to zero, a weaker linear relationship is indicated. When *r* starts to approach –1 or +1, a stronger linear relationship is indicated. The sign of *r*, whether the value is positive or negative, denotes information about the direction of the linear relationship. If *r* is negative, then *x* and *y* are inversely or negatively related; and as the values of one variable increase, the values of the other variable decrease. If *r* is positive, then the variables are directly or positively related; and as the values of one variable increase, the values of the other variable also increase.

The **coefficient of determination** is the square of the coefficient of correlation, *r*. The value tells in a percentage how much of the fluctuation or variance in the *y* variable is determined or explained by the linear equation involving *x*. The value can be calculated as shown in the equation below, or for simple linear regression when a line is fitted to the data it can be calculated by simply squaring the value calculated for *r*.

$$r^2 = \left(\frac{SS_{xy}}{\sqrt{SS_x \cdot SS_y}} \right)^2 = 1 - \frac{SSE}{SSTotal} = 1 - \frac{\sum (y - \hat{y})^2}{\sum (y - \bar{y})^2}.$$

Once it has been established that x and y are linearly related and the best-fitting linear regression equation is estimated the equation can be used to estimate the mean of Y or to **predict** an individual y value. The prediction of an individual y value is often referred to as a **forecast.**

The estimated regression equation that appears:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

is used for two basic purposes:

1. Estimating the mean or average value of y at some specific x-value.
2. Predicting or forecasting an individual y-value at some specific x-value.

Consider an equation which estimates sales based on advertising cost measured in 100's of dollars. Assume the best fitting least squares regression equation is

$$\hat{y} = -0.1 + 0.7(x),$$

which is equivalent to

$$\hat{sales} = -0.1 + 0.7\,(advertising\ cost).$$

Two questions can be answered with this equation. Assume we are concerned with advertising cost of \$400, which is $x = 4$. One question that can be addressed is "What is the *average* or mean sales when advertising cost is \$400?" Another question is, "What is the *predicted* or forecast sales when advertising cost is \$400?"

The answer to both of these questions is

$$\hat{sales} = -0.1 + 0.7(4) = 2.7.$$

The estimate of average sales is \$2700 when advertising cost is \$400. The predicted value of sales is \$2700 the next time that advertising cost is \$400.

The same value, \$2700, was used to estimate an average sales and to predict a future sales amount, but the standard errors in these two cases are different. The amount $\sigma_{\hat{y}}$ is the standard error when the estimated equation is used to estimate the mean of y. The amount $\sigma_{(y-\hat{y})}$ is the

standard error when $\hat{y}$ is used to predict a future y-value. The equations for these two standard errors are

$$s_{\hat{y}} = s\sqrt{\frac{1}{n} + \frac{(x_p - \bar{x})^2}{SS_{xx}}}$$

$$s_{(y-\hat{y})} = s\sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{SS_{xx}}}.$$

Notice the similarity of the equations, but also take note of the difference—the extra one in the standard error for prediction. The error associated with a forecast or predicted value has the error related to the estimate of the mean and also the additional error of the data from the regression model measured by the s quantity. The extra "1" in the lower equation increases the standard error for prediction by the amount of the estimate of the variance of the residuals.

A confidence interval for the mean value of y at a given x and a prediction interval for a forecast or individual value of y at a given x can be constructed. A prediction interval is constructed in the same manner as the confidence interval with the least squares estimate of y based on the regression equation as the center point of the interval. The prediction interval is always wider than the confidence interval if the same confidence level and same x-value are used since the standard error of prediction is larger than the standard error for estimation.

The basic notion of using an equation to describe the relationship between two or more variables is basic to statistics. The concept is involved in many of the processes examined later in this course. The next lesson discusses linear models that have additional x-terms that fit geometric shapes other than lines to bivariate and multivariate data.

DETAILS FOR INFERENCES ON THE SLOPE OF A REGRESSION LINE

HYPOTHESIS TEST

1. **State the set of hypotheses.**

   $H_o : \beta_1 = \beta_{1_o}$

   $H_a : \beta_1 \neq \beta_{1_o}$ or $H_a : \beta_1 < \beta_{1_o}$ or $H_a : \beta_1 > \beta_{1_o}$

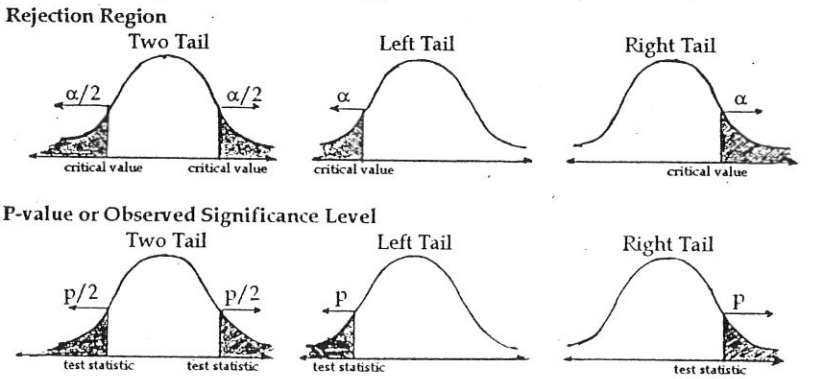
2. **Take a random sample of bivariate data. Calculate the test statistic.**

$$t = \frac{\hat{\beta}_1 - \beta_{1_o}}{s_{\hat{\beta}_1}} = \frac{\hat{\beta}_1 - \beta_{1_o}}{\frac{s_e}{\sqrt{ss_x}}} = \frac{\hat{\beta}_1 - \beta_{1_o}}{\frac{\sqrt{\frac{SS_{yy} - \hat{\beta}_1 SS_{xy}}{n-2}}}{\sqrt{ss_x}}}.$$

3. **Identify the distribution of the test statistic.**

   The variable $t = \dfrac{\hat{\beta}_1 - \beta_{1_o}}{S_{\hat{\beta}_1}} \sim t(n-2)$ if H_o is true.

4. **Generate the rejection region or the p-value to make the decision of whether to reject the null hypothesis or not.**



5. **Form a conclusion in words.** If the null hypothesis is rejected then the data do support the alternative hypothesis; if the null hypothesis is not rejected then the data do not support the alternative hypothesis.

(1-α) 100% CONFIDENCE INTERVAL

$$\hat{\beta}_1 \pm t_{\alpha/2(n-2)} \cdot s_{\hat{\beta}_1} \Rightarrow \hat{\beta}_1 \pm t_{\alpha/2(n-2)} \cdot \frac{s_e}{\sqrt{ss_x}} \Rightarrow \text{(lower bound of CI, upper bound of CI)}$$

EXAMPLE

SIMPLE LINEAR REGRESSION

Before the example, let's look at the basic calculations and equations involved.

a. Calculate: $\Sigma x = x_1 + x_2 + \ldots + x_n$

$$\Sigma y = y_1 + y_2 + \ldots + y_n$$

$$\Sigma x^2 = x_1^2 + x_2^2 + \ldots + x_n^2$$

$$\Sigma y^2 = y_1^2 + y_2^2 + \ldots + y_n^2$$

$$\Sigma xy = x_1y_1 + x_2y_2 + \ldots + x_ny_n$$

b. Calculate: $SS_{xx} = \Sigma x^2 - \dfrac{(\Sigma x)^2}{n}$

$$SS_{yy} = \Sigma y^2 - \frac{(\Sigma y)^2}{n}$$

$$SS_{xy} = \Sigma xy - \frac{(\Sigma x)(\Sigma y)}{n}$$

c. Then the basic calculating equations are

$$\hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}}$$

$$\hat{\beta}_0 = \frac{\Sigma y}{n} - \left(\hat{\beta}_1\right)\frac{\Sigma x}{n}$$

$$r = \frac{SS_{xy}}{\sqrt{SS_{xx}SS_{yy}}}$$

$$\hat{\sigma} = s = \sqrt{\frac{SS_{yy} - \hat{\beta}_1\left(SS_{xy}\right)}{n-2}} = \sqrt{\frac{\Sigma y^2 - \hat{\beta}_0\left(\Sigma y\right) - \hat{\beta}_1\left(\Sigma xy\right)}{n-2}}$$

$$\hat{\sigma}_{\hat{\beta}_1} = S_{\hat{\beta}_1} = \sqrt{\frac{S^2}{SS_{xx}}}$$

$$\hat{\sigma}_{\hat{y}} = s_{\hat{y}} = s\sqrt{\frac{1}{n} + \frac{(x_p - \bar{x})^2}{SS_{xx}}}$$

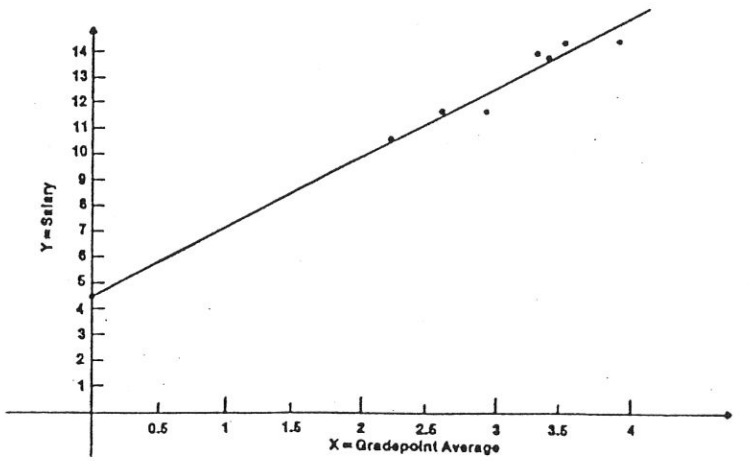
$$\hat{\sigma}_{(y-\hat{y})} = s_{(y-\hat{y})} = s\sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{SS_{xx}}}$$

Now, let's look at the example.

Suppose we are interested in predicting or estimating starting salaries from college grade-point averages (GPAs). We sampled seven people asking each of them their GPA and their starting salary. The bivariate data, the (x, y) data points, appear as follows:

x(GPA)	y(Salary)	x^2	y^2	xy
2.58	11.5	6.6564	132.25	29.67
3.27	13.8	10.6929	190.44	45.126
3.85	14.5	14.8225	210.85	55.825
3.50	14.2	12.25	201.64	49.7
3.33	13.5	11.0889	182.25	44.955
2.89	11.6	8.3521	134.56	33.524
2.23	10.6	4.9729	112.36	23.638
$\sum x = 21.65$	$\sum y = 89.7$	$\sum x^2 = 68.8357$	$\sum y^2 = 1163.75$	$\sum xy = 282.438$

1. Graph the (x, y) data, that is, draw the scatter diagram.



2. Calculate the appropriate corrected sums of squares.

$$SS_{xx} = \Sigma x^2 - \frac{(\Sigma x)^2}{n} = 68.8357 - \frac{1}{7}(21.65)^2 = 1.875342856$$

$$SS_{yy} = \Sigma y^2 - \frac{(\Sigma y)^2}{n} = 1163.75 - \frac{1}{7}(89.7)^2 = 14.3085714$$

$$SS_{xy} = \Sigma xy - \frac{(\Sigma x)(\Sigma y)}{n} = 282.438 - \frac{1}{7}(21.65)(89.7) = 5.00871$$

Remember to keep all the possible digits past the decimal. Do not round these numbers.

3. Estimate the linear correlation between GPA and salary.

$$r = \frac{SS_{xy}}{\sqrt{SS_{xx} \cdot SS_{yy}}} = \frac{5.00871428}{\sqrt{(1.875342856)(14.3085714)}} = 0.9669 \approx 0.97$$

An r value of 0.97 indicates the variables have a strong positive linear correlation. We could do a good job of estimating salary with a linear equation based on grade-point average. The estimated correlation coefficient is r.

4. Give the coefficient of determination. What does this value represent?

$$r^2 = (0.9669)^2 = 0.935.$$

An r^2 value of 0.935 indicates that about 93.5% of the variance of starting salary can be explained by the linear equation based on grade-point average.

5. Calculate the least-squares regression equation. State it.

First, estimate the slope, β_1, with $\hat{\beta}_1$.

$$\hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}} = \frac{5.00871428}{1.875342856} = 2.670825905$$

Estimate the y-intercept, β_0, with $\hat{\beta}_0$:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1(\bar{x}) = \frac{\Sigma y}{n} - (2.670825905)\left(\frac{\Sigma x}{n}\right) = 4.553802732.$$

The estimated regression equation is (graph onto plot):

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1(x)$$

so

$$\hat{y} = 4.553802732 + 2.670825905(x).$$

Say we wanted to estimate the average starting salary for college graduates who have a *GPA* = 3.0, the estimate is

$$\hat{y}_{x=3.0} = 4.553802732 + 2.670825905(3.0) = 12.57.$$

We expect someone to make about \$12,570 as starting salary if their *GPA* = 3.0.

6. Calculate the standard deviation of the residuals, s:

$$s = \sqrt{\frac{\Sigma(y-\hat{y})^2}{n-2}} = \sqrt{\frac{\Sigma y^2 - a(\Sigma y) - b(\Sigma xy)}{n-2}} = 0.431547807$$

Note: $s^2 = (0.431547807)^2 = 0.18623351$.

7. Calculate the standard errors of $\hat{\beta}_1$, $\hat{y}$, and for forecasting.

$$s_{\hat{\beta}_1} = \sqrt{\frac{s^2}{SS_{xx}}} = \sqrt{\frac{0.18623351}{1.875342856}} = 0.315129148$$

$$s_{\hat{y}\cdot x=3.0} = s\sqrt{\frac{1}{n} + \frac{(x_p - \bar{x})^2}{SS_{xx}}} = 0.431547807\sqrt{\frac{1}{7} + \frac{\left(3.0 - \frac{21.65}{7}\right)^2}{1.875342856}}$$
$$= 0.165713736$$

$$s_{(y-\hat{y})x=3.2} = s\sqrt{\frac{1}{n} + \frac{(x_p - \bar{x})^2}{SS_{xx}} + 1} = 0.431547807\sqrt{1 + \frac{1}{7} + \frac{\left(3.2 - \frac{21.65}{7}\right)^2}{1.875342856}}$$
$$= 0.46257788$$

where x_p is the x-value of concern: $x_p = 3.0$ for $s_{\hat{y}}$ since that is the x-value used in number 11. $x_p = 3.2$ for $s_{(y-\hat{y})}$ since that is the x-value in number 10.

8. Test the hypothesis that the true slope is zero, $\beta_1 = 0$.

$$H_0 : \beta_1 = 0 \text{ vs. } H_a : \beta_1 \neq 0$$

$$\textit{Test Statistic} = t_{calc.} = \frac{\hat{\beta}_1 - 0}{s_{\hat{\beta}_1}} = \frac{2.670825905}{0.315129148} = 8.4753$$

Such a large test statistic value indicates a very small OSL, so there is very strong evidence against the null hypothesis, $\beta_1 = 0$. Reject $H_0 : \beta_1 = 0$ and conclude that the slope is not 0, $\beta_1 \neq 0$.

9. Since we've concluded $\beta_1 \neq 0$, let's estimate β_1 with a 95% confidence interval. The basic form of the 95% confidence interval to estimate the slope is

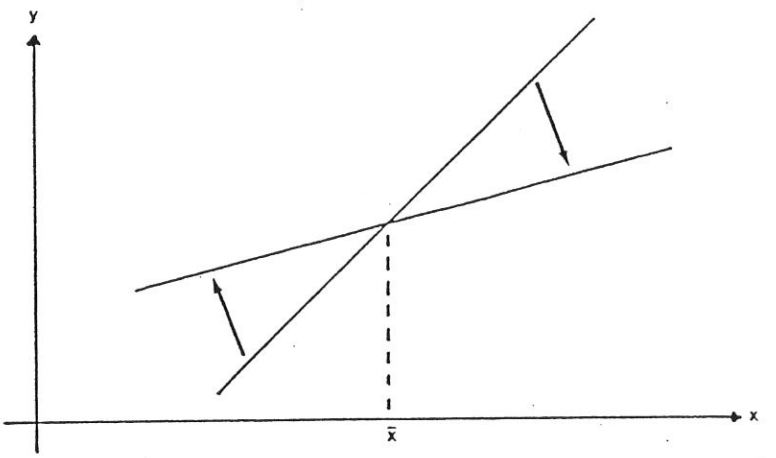
$$\hat{\beta}_1 \pm \left[t_{(n-2)\frac{.05}{2}} \cdot s_{\hat{\beta}_1} \right]$$

$2.670825905 \pm [2.571 \cdot 0.315129148]$

$2.670825905 \pm [0.810197040]$

$\rightarrow (1.861,\ 3.481)$.

It is estimated that the true slope, β_1, is between 1.86 and 3.48 units. This confidence interval on the slope describes a fan-shaped region with the most narrow section at the $\bar{x}$-value.



Fan-Shaped Region Defined by Confidence Interval on the Slope.

10. Estimate with an interval an individual's starting salary if the person has a 3.2 GPA. This is a prediction interval on a forecast value. The basic form of the internal associated with 95%:

$$\hat{y}_{x=3.2} \pm \left[t_{(n-2)\frac{.05}{2}} \cdot s_{(y-\hat{y})\cdot x_p=3.2} \right]$$
$$13.1012 \pm [2.571 \cdot 0.46257788]$$
$$13.1012 \pm [1.18928772]$$
$$\rightarrow (11.9119,\ 14.2905).$$

We are *fairly* sure that an individual who has a 3.2 GPA could find a job with starting salary between \$11,920 and \$14,290.

11. Give an interval estimate for the average starting salary for people who make a 3.0 GPA. This is a confidence interval for $\mu_{y\cdot x=3.0}$; the form is (95% confidence level):

$$\hat{y}_{x=3.0} \pm \left[t_{(n-2)\frac{.05}{2}} \cdot s_{\hat{y}\cdot x=3.0} \right]$$
$$12.567 \pm [2.571 \cdot 0.165713736]$$
$$12.567 \pm [0.42605]$$
$$\rightarrow (12.14,\ 12.99).$$

We are fairly sure that the average starting salary for people who have a *GPA* = 3.0 is between \$12,140 and \$12,990. Recognize that in the above we used $\hat{y}_{x=3.0}$ to estimate $\mu_{y\cdot x=3.0}$.

12. Conclude that starting salary and GPA have a very strong positive linear relationship with a correlation coefficient of 0.97. The least-squares regression equation that best describes the linear relationship between starting salary and GPA is $\hat{y} = 4.55 + 2.67(x)$, where y is starting salary and x is GPA.