

Linear Regression and Hypothesis Testing

AP Review

Terms

- ◆ Regression Equation (data)
 $\hat{Y} = a + bx$
- ◆ Regression Equation (population)
 $\hat{y} = \beta_0 + \beta_1x + \varepsilon$
- $a = y$ intercept: predicted value for y when $x = 0$
- $b =$ slope
- $\varepsilon =$ errors from population equation

Terms

- ◆ Residual = observed y – predicted y
- ◆ S_a = standard deviation of a
- ◆ S_b = standard deviation of b
- ◆ S = standard deviation of the residuals
- ◆ r = correlation coefficient
- ◆ r^2 = coefficient of determination

Terms

- ◆ df = degrees of freedom = $n - 2$
- ◆ t value = test statistic
- ◆ P-value = probability of t value given $H_0: t = 0$
- ◆ SST = Total sum of squares
(observed $y - \bar{y}$)²
- ◆ SSE = Sum of squares of the residuals (observed $y - \hat{y}$)²

Practice Multiple Choice

- ◆ MC questions tend to be
 - Definitions
 - Applications of formulas
 - Applications of definitions
 - Proper methodology

When regressing y on x ,
 y is called the:

- a. response variable.
- b. independent variable.
- c. predictor variable.
- d. explanatory variable.

When regressing y on x ,
 y is called the:

- a. response variable.
- b. independent variable.
- c. predictor variable.
- d. explanatory variable.

If the slope of the regression line is negative
and the coefficient of determination is .64,
then Pearson's correlation coefficient is:

- a. .64
- b. .8
- c. -.64
- d. -.8

If the slope of the regression line is negative and the coefficient of determination is .64, then Pearson's correlation coefficient is:

- a. .64
- b. .8
- c. -.64
- d. -.8

A value of $r = .40$ indicates that there is a:

- a. strong positive relationship between x and y .
- b. strong negative relationship between x and y .
- c. weak positive relationship between x and y .
- d. weak negative relationship between x and y .

A value of $r = .40$ indicates that there is a:

- a. strong positive relationship between x and y .
- b. strong negative relationship between x and y .
- c. weak positive relationship between x and y .
- d. weak negative relationship between x and y .

Which of the following is not a property of r ?

- a. r does not depend on the units of y or x .
- b. r is always between 0 and 1.
- c. r measures the strength of the linear relationship between x and y .
- d. r does not depend on which of the two variables is labeled x .

Which of the following is not a property of r ?

a. r does not depend on the units of y or x .

b. r is always between 0 and 1.

c. r measures the strength of the linear relationship between x and y .

d. r does not depend on which of the two variables is labeled x .

A good fitting regression line should have:

a. small r^2 and large s_e .

b. large r^2 and large s_e .

c. small r^2 and small s_e .

d. large r^2 and small s_e .

A good fitting regression line should have:

- a. small r^2 and large s_e .
- b. large r^2 and large s_e .
- c. small r^2 and small s_e .
- d. large r^2 and small s_e .**

A point is called an influential observation when:

- A. it has a large residual.
- B. it has a small residual.
- C. it plays a big role in determining the slope of the least squares line.
- D. it should be transformed using a power transformation.

A point is called an influential observation when:

A.it has a large residual.

B.it has a small residual.

C.it plays a big role in determining the slope of the least squares line.

D.it should be transformed using a power transformation.

Which of the following is not an assumption that is made about the random deviation e in a simple linear regression model?

A.The distribution of e is normal.

B.The standard deviation of e , depends upon the particular value of x .

C.The mean value of e is 0.

D.The random deviations e_1, e_2, \dots, e_n associated with different observations are independent of one another.

Which of the following is not an assumption that is made about the random deviation e in a simple linear regression model?

A. The distribution of e is normal.

B. The standard deviation of e , depends upon the particular value of x .

C. The mean value of e is 0.

D. The random deviations e_1, e_2, \dots, e_n associated with different observations are independent of one another.

The population standard deviation of y :

a. is the same for each value of x .

b. is different for each value of x .

c. is the same as σ (sub b)

d. has $n - 2$ degrees of freedom.

The population standard deviation of y :

- a. is the same for each value of x .
- b. is different for each value of x .
- c. is the same as σ (sub b)
- d. has $n - 2$ degrees of freedom.

Reading Computer Output

- ◆ Regression Problem on AP exam will require reading and interpreting computer output.
- ◆ Know where and how to interpret computer output.

One measure of the success of knee surgery is postsurgical range of motion for the knee joint. This range of motion was recorded for 12 patients who had knee surgery following a knee dislocation. The age of each patient was recorded also. The following computer output is from this data. (POD 5.11)

Predictor	Coef	StDev	T	P		
Constant	107.58	11.12	9.67	0.000	a	Sa
Age	0.8710	0.4146	2.10	0.062	b	Sb

S = 10.42 R-Sq = 30.6% R-Sq(adj)=23.7%

Se r^2

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	479.2	479.2	4.41	0.062
Residual Error	10	1085.7	108.6	df	SSE
Total	11	1564.9	SST		

Note: P-value assumes $H_a: \beta \neq 0$ or $p \neq 0$

From the computer output, find

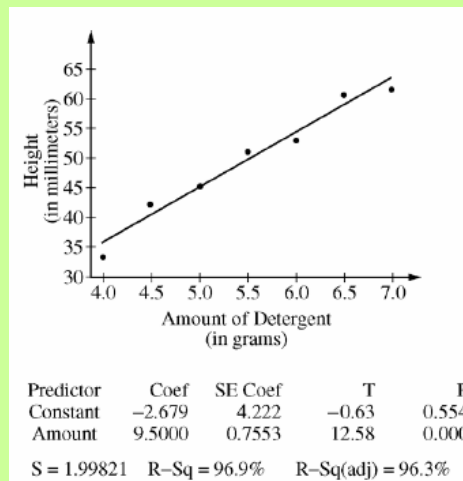
- a. Equation of least squares regression line
- b. Correlation coefficient & interpret
- c. the slope
- d. S
- e. If there is a useful linear relationship
- f. 95% confidence interval for the true slope
- g. a point estimate for range of motion for a 27 year old.

LSRL 2006

A manufacturer of dish detergent believes the height of soapsuds in the dishpan depends on the amount of detergent used. A study of the suds' heights for a new dish detergent was conducted. Seven pans of water were prepared.

The temperature of the water was the same for each pan. An amount of dish detergent was assigned at random to each pan, and that amount of detergent was added to the pan. Then the water in the dishpan was agitated for a set amount of time, and the height of the resulting suds was measured.

A plot of the data and the computer output from fitting a least squares regression line to the data are shown below.



You have 12 minutes to answer the following questions about these soapsuds.

- a. Write the equation of the fitted regression line. Define any variables used in this equation.
- b. Note that $s = 1.99821$ in the computer output. Interpret this value in the context of this study.
- c. Identify and interpret the standard error of the slope.

Part a

$$\hat{Y} = -2.679 + 9.5x$$

Where

X = amount of detergent in grams

\hat{Y} = predicted height of soapsuds

Part b

- ◆ $S=1.99821$ is the standard deviation of the residuals. It is a typical amount of variation between the observed height and the predicted height of the soapsuds.
- ◆ It also is a measure of the variation in height of the soapsuds for any given amount of detergent.

Part c

- ◆ The standard error of the slope is 0.7553. This estimates the variability in the sampling distribution of the estimated slope from one experiment to another.

Part d

Is it reasonable to use this equation to predict the height of the soapsuds if 3.25g of detergent is used?

No, because the domain of data is from 4 as a minimum to 7 as a maximum. 3.25 is not in this interval

Part e

Find the average height of soapsuds if 5.3g of detergent is used.

$$\hat{Y} = -2.679 + 9.5 (5.3) = 47.671$$

I would predict soapsuds of approximately 47.671mm for 5.3g of detergent.

Part f

Find a 95% confidence interval for the population slope.

$$9.5 \pm 2.571 \cdot 0.7553 = (7.558, 11.44)$$

I am 95% confident that the slope of the true regression line lies in the interval from 7.558 to 11.44.

Part g

The manufacturer believes that the true slope of the regression line is 9.0. Run a hypothesis test with $H_0: \beta = 9$ against $H_a: \beta > 9$

$$t = (9.5 - 9) / .7553 = .6620$$

$$P(t > .6220, df = 5) = .2806$$

The p-value is larger than any reasonable α , so we fail to reject H_0 . There is not enough evidence to say that the slope of the true regression lines is larger than 9.

What should you write?

- ◆ If you find a value, whether \hat{y} , b , interval or hypothesis test, interpret it in context of the problem.
- ◆ Indicate that values found by the regression line is approximate
- ◆ Indicate that y value found by the regression line for a given value of x is approximate, or is an average value of y for that x .

What should you write?

- ◆ Do not write more than is needed to answer the question. Essays are expected to be a few sentences at most.
- ◆ You do not need to “fill up” all the space allowed on the exam.

