Chapter 13
Study Guide

Regression Inference

INFERENCE FOR LINEAR REGRESSION AND CORRELATION

In Review Section 3, you learned how to describe and summarize bivariate data. You might wonder if the sample provides evidence that the two variables are associated in the population from which the sample was selected. In this section, you will learn inference techniques for bivariate numerical data.

OBJECTIVES

- Understand a simple regression model.
- Construct a confidence interval for the slope of the population regression line.
- Perform a test of significance for the slope of the population regression line.
- Interpret and communicate the results of the statistical analysis.

SIMPLE LINEAR REGRESSION MODEL

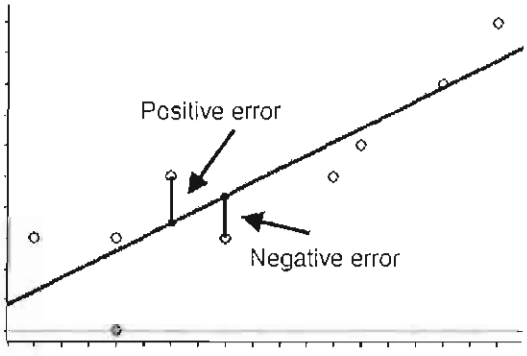
(*Introduction to Statistics & Data Analysis* 3rd ed. pages 690–700/4th ed. pages 742–752)

Recall from Review Section 3, that we use one variable x (the explanatory variable) to help explain or shed light on the variability of another variable y (the response variable).

The simple linear regression model is the linear model that is assumed to describe the relationship between x and y in the population data. It is written as

$$y = \alpha + \beta x + e$$

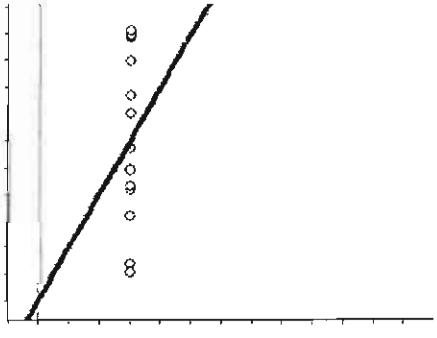
The relationship between two variables will not usually be perfectly linear, so the error e accounts for the random variability that exists. Think of the residuals as a sample from the population of errors. If a data point from the population lies above the regression line, e is positive and likewise, e is negative when a data point from the population lies below the regression line.



Our goal is to take a sample from a population and, using the data from our sample, be able to infer something about the population. In order to perform any inference procedures in the linear regression setting, we need to make four assumptions about the population and the associated errors.

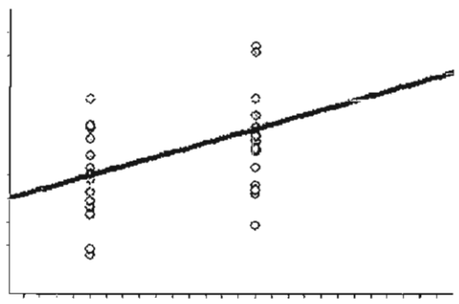
The first assumption is that the mean of all the errors at any particular x value is equal to zero for each and every value of x.

Below is a scatterplot to illustrate this idea; notice that the mean of the errors is zero (remember zero is the point on the line since we are talking about errors which are directed distances above and below the line). We are assuming that the population appears like this for each and every value of x.

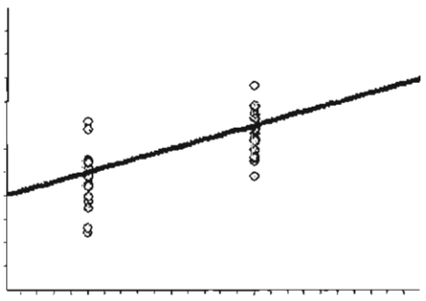


The second assumption we must make is that the spread of the distribution of e is the same for any particular value of x. This means that the standard deviation of e is the same for each and every value of x. The standard deviation of e is denoted σ_e.

Below is a scatterplot of what the population might look like for two specific x values; notice that the spread of the errors for each x is about the same. We are assuming that the population appears like this for each and every value of x.

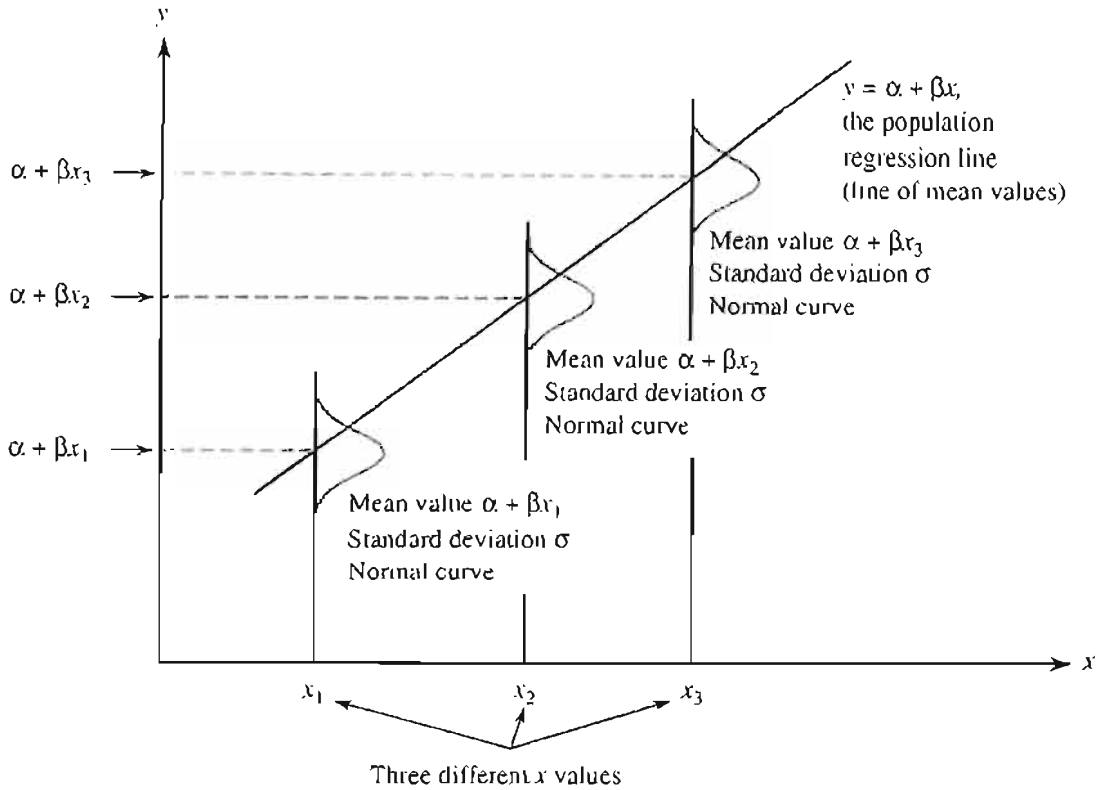


The third assumption we must make is that the distribution of e is normal for any particular value of x. Below is a scatterplot of what the population might look like for two specific x's; notice that the errors appear to be normally distributed about the regression line for each x. Assessing normality can be difficult. Remember, we are assuming that the population appears like this for each and every value of x.



The fourth and final assumption is that each error that corresponds to an observation is independent of each and every error that corresponds to another observation.

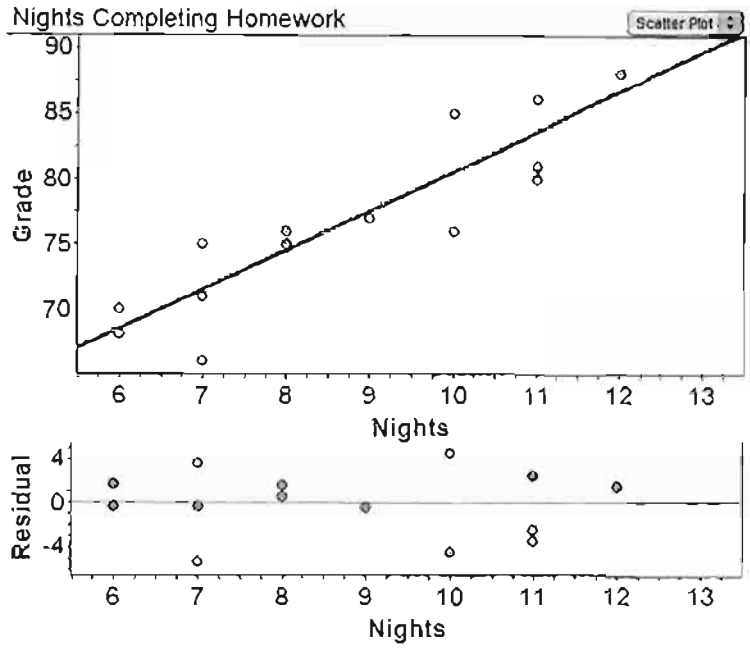
Now, let's put all four conditions together. The observations must be independent and for every value of x the errors must be normally distributed with a mean of zero and the same standard deviation. Below is a picture illustrating all of these ideas together.



EXAMPLE The following dataset gives the number of nights in a three-week time frame that students completed their homework and their quiz grade on the material taught during those three weeks. Is it reasonable to assume that the four assumptions of the linear regression model are true?

Number of Nights	11	12	8	12	7	6	11	7	10	6	10	11	7	9	8
Quiz Grade	81	88	76	88	66	70	80	75	85	68	76	86	71	77	75

First, we look at a scatterplot and residual plot for the data to confirm that a linear model is appropriate.



Based on the scatterplot and the residual plot, it appears that a linear model is appropriate. Now focus on the residual plot for assumptions 1–3.

Assumption 1: The mean of e at any given x value is equal to 0. This assumption is for each and every value of x. This assumption cannot be checked; the least squares regression line will always have residuals that have a mean of zero. We must assume that the mean of the errors is zero for the population. This assumption is reasonable when the pattern in the scatterplot is linear and there is no obvious pattern in the residual plot.

Assumption 2: σ_e is the same for each and every value of x. The scatter about the regression line does not appear to be much larger for some x values than for others, so this assumption seems reasonable.

Assumption 3: The distribution of e is normal for any particular value of x. For this assumption to be reasonable, points should tend to cluster near the line, with fewer points as you move away from the line. This appears to be the case here, so this assumption seems reasonable.

Assumption 4: The observations are independent. If no students worked together on the homework assignments or on the quiz, it seems reasonable to assume that students' homework and quiz grades are independent.

Since all four assumptions are reasonable, it makes sense to use a linear regression model.

Notice that all of the assumptions are about the population where the population is the collection of all students' number of nights of homework preparation and their corresponding quiz grade. Usually, we only have information from a sample taken from the desired

population. Using a least squares regression line, we can estimate the population parameters from sample data.

The estimated regression line for the population is

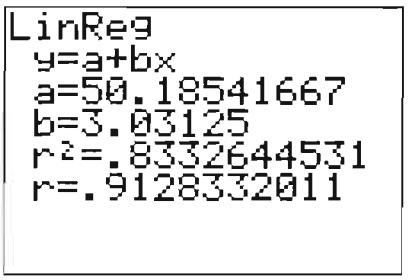
$$\hat{y} = a + bx$$

Where a is a point estimate for α and b is a point estimate for β.

Let x^* be a specific value of x. Then, $a + bx^*$ has two different meanings:

1. It is an estimate of the mean y for all observations when
2. It is the predicted y when

EXAMPLE Refer to the previous example about number of nights a student completed their homework and their quiz grade. Estimate α and β for the population regression model.



The least squares regression line for our sample data is

$$\hat{y} = 50.185 + 3.031x$$

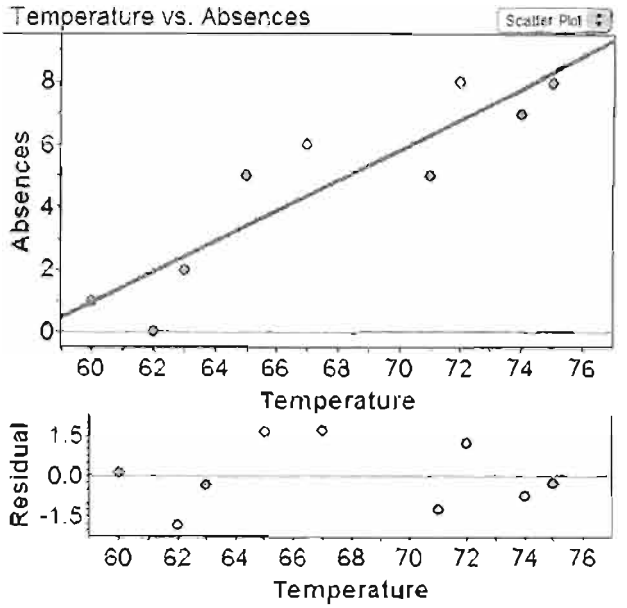
where x is the number of nights that a student completed his or her homework and $\hat{y}$ is the student's predicted quiz grade.

Therefore, our estimate for α is 50.185 points and our estimate for β is 3.031 points per night of homework completed.

SAMPLE PROBLEM 1 Data on outside temperature at 9 a.m. and number of students absent from school is given for 9 randomly selected days during the school year. Is a linear regression model reasonable? If so, estimate the equation of the population regression line. Find $\hat{y}$ when and interpret this value in context.

Outside Temperature (°F)	60	62	63	65	67	71	72	74	75
Number of Students Absent	1	0	2	5	6	5	8	7	8

SOLUTION TO PROBLEM 1 First, look at a scatterplot and a residual plot for the data.



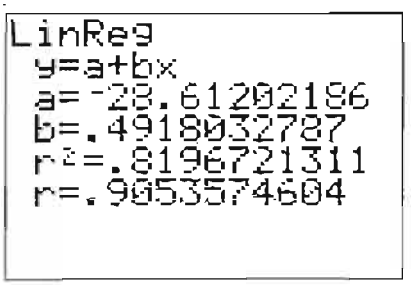
A linear model seems appropriate for this data. Now, we must check the four assumptions.

Assumption 1: for each and every value of x. This assumption is reasonable when the pattern in the scatterplot is linear and there is no obvious pattern in the residual plot, which appears to be the case here.

Assumption 2: is the same for each and every value of x. The scatter about the regression line does not appear to be much larger for some x values than for others, so this assumption seems reasonable.

Assumption 3: The distribution of e is normal for any particular value of x. For this assumption to be reasonable, points should tend to cluster near the line, with fewer points as you move away from the line. This appears to be the case here, so this assumption seems reasonable.

Assumption 4: The observations are independent. Because the days were randomly selected, it is reasonable to assume that the observations are independent.



The least squares regression line for our sample data is

$$\hat{y} = -28.612 + 0.492x$$

where x is the outside temperature and $\hat{y}$ is the predicted number of students absent from class. This equation is the estimate of the population regression line.

When $x = 68$,

$$\hat{y} = -28.612 + 0.492(68) = -28.612 + 33.443 = 4.831$$

There are two ways to correctly interpret this value:

1. The average number of students absent on days when it is 68°F outside is 4.831.
2. Our model predicts that 4.831 students will be absent on an individual day where it is 68°F.

This example highlights what happens on most problems when we check the four assumptions—we did not have sufficient evidence to doubt the assumptions, so we assumed they were true and proceeded.

INFERENCES ABOUT THE SLOPE OF THE REGRESSION LINE

(Introduction to Statistics & Data Analysis 3rd ed. pages 702–710/4th ed. pages 755–762)

Remember from Review Section 3 that the slope of the least squares regression line is the average change in y for a one-unit change in x. We now are concerned with β, the slope of the population regression line. It is interpreted the same way—only we are talking about the population, not a sample.

When the four assumptions for a linear regression model are met, the following is true for the slope of the sample regression line, b:

1. The mean value of b for all possible random samples is β. This means that b is an unbiased estimator of β.
2. The standard deviation of b for all possible random samples is .
3. The sample statistic b has a normal distribution.

Since the sample statistic b has a normal distribution, the statistic will have a standard normal distribution.

Just as before, this presents a problem. We rarely, if ever, know the value of σ. We use s_e to estimate σ and use a t interval (test) instead of a z interval (test). The distribution of $t = \dfrac{b - \beta}{\dfrac{s_e}{\sqrt{S_{xx}}}}$ has a t distribution with

$df = n - 2$.

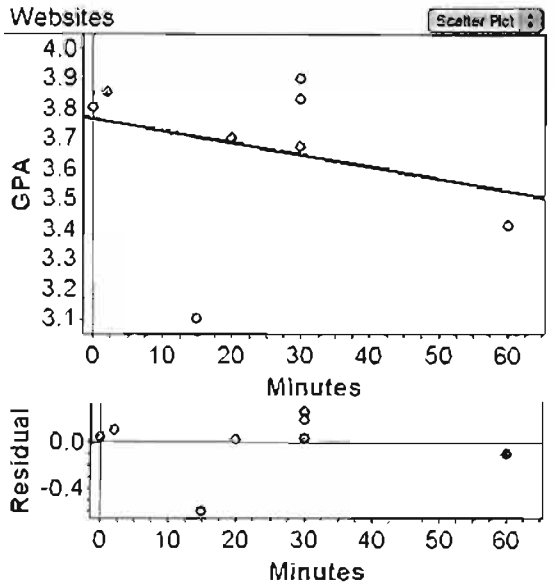
EXAMPLE A teacher is interested in determining if there is a relationship between GPA and the number of minutes each day a student spends on social networking websites. She selects a random sample of nine students. Use the data given below to construct an interval estimate the slope of the population regression line relating GPA and the number of minutes spent on social networking websites.

Minutes Spent	20	60	15	30	0	30	0	2	30
Grade Point Average	3.70	3.41	3.10	3.90	3.80	3.67	3.80	3.85	3.83

We are interested in knowing about β (the true mean change in GPA per additional minutes spent on social networking websites). The question asks for an estimate—this signals to us that we should find a confidence interval.

NAME OF INTERVAL: *t* interval for the slope of the population regression line

CONDITIONS: First, look at a scatterplot and a residual plot of the sample data (shown below). A linear model seems appropriate for this data.



Assumption 1: for each and every value of *x*. This assumption is reasonable when the pattern in the scatterplot is linear and there is no obvious pattern in the residual plot, which appears to be the case here.

Assumption 2: is the same for each and every value of *x*. The scatter about the regression line does not appear to be much larger for some *x* values than for others, so this assumption seems reasonable.

Assumption 3: The distribution of *e* is normal for any particular value of *x*. For this assumption to be reasonable, points should tend to cluster near the line, with fewer points as you move away from the line. This appears to be the case here, so this assumption seems reasonable.

Assumption 4: Because the sample was a random sample, the observations are independent.

CALCULATIONS:

Model of Websites — Simple Regression

Response attribute (numeric): GPA
Predictor attribute (numeric): Minutes

Sample count:	9
Equation:	GPA = -0.00397999 Minutes + 3.7560
r:	-0.299352
r-squared:	0.089612
Slope:	-0.00397999 +/- 0.0113378
SE Slope:	0.00479474

For a confidence level of 95% and df = 9 − 2 = 7, the appropriate t critical value is 2.365. From the given output, $b = 0.00398$ and $s_b = 0.00479$. The interval is then

estimate ± t^* (SE_b)
−0.00398 ± (2.365)(0.00479)
(−0.0153, 0.0074)

CONCLUSION We are 95% confident that the true mean change in GPA for each additional minute spent on social networking sites each day is between −0.0153 and 0.0074.

EXAMPLE Airlines are interested in knowing if there is a relationship between number of times people travel in a year and the weight of their luggage. A random sample of adult Americans who took at least one trip in the last year was used to produce the computer output below. Suppose it is reasonable to believe that the assumptions for inference are met. Does the sample provide evidence of a relationship between number of trips and luggage weight? Use the Minitab printout below.

Regression Analysis: WeightOfLuggage versus NumberOfTimes

The regression equation is
WeightOfLuggage = 35.3 - 0.157 NumberOfTimes

Equation of the estimated regression line

Predictor	Coef	SE Coef	T	P
Constant	35.2813	0.5244	67.28	0.000
Registrations	-0.15698	0.03879	-4.05	0.002

p-value
Test statistic
Estimated SEs

S = 0.768322 — s_e
R-Sq = 62.1%
R-Sq(adj) = 58.3%

Analysis of Variance — n - 2

Source	DF	SS	MS	F	P
Regression	1	9.6660	9.6660	16.37	0.002
Residual Error	10	5.9032	0.5903		
Total	11	15.5692			

We are interested in knowing about β (the true mean change in baggage weight per additional trip). The question asked for a decision in the form of a yes or no answer—is there enough evidence to show that there is a relationship between number of times a person travels in a year and the weight of their luggage? To answer this question, we will perform a hypothesis test about the slope of the regression line.

NAME OF TEST Linear regression t test (two tailed)

HYPOTHESES

$H_0 : \beta = 0$

$H_a : \beta \neq 0$

β = the true mean change in baggage weight per additional trip

$\alpha = 0.05$

CONDITIONS The problem states that the conditions for inference are met.

CALCULATIONS From the given computer output, t = -4.05, P-value = 0.002, and df = 10.

CONCLUSION If it were true that , we would get this result or one more extreme than this less than 0.2% of the time. Since this P-value is less than α = 0.05, we reject the null hypothesis. We have sufficient evidence to conclude that there is a linear relationship between the number of times people travel and the weight of their luggage.

AP Tip

Be sure to know how to read a computer output. Know where you can find the statistics to answer the questions you might be asked. Some of the information on these outputs is extraneous and you will not need this information nor are you expected to know what it means. Be sure to practice finding the statistics you will need to know!

SAMPLE PROBLEM 2 Each year, most students take a test that measures their reading achievement. The score reports this in terms of the grade level at which the child is reading. The following scores are for a random sample of six children in a particular U.S. public school. Suppose it is reasonable to believe that the assumptions for inference are met. Is there convincing evidence that mean grade level score increases by more than 1 with each additional year of school?

Grade Level	1	3	4	7	9	12
Grade Level Equivalence	0.83	3.2	4.1	6.7	9.1	12.2

SOLUTION TO PROBLEM 2

Name of Test: t test (one tailed) for slope of regression line

Hypotheses:

$H_0 : \beta = 1$

$H_a : \beta > 1$

β = the true mean change in grade level score per additional year of school

$\alpha = 0.05$

Conditions: The problem states that the conditions for inference are met.

Calculations: Computer regression output is shown here, or a graphing calculator can be used.

```
The regression equation is
Grade Level Reading Score = - 0.068 + 1.01 Grade in School

Predictor          Coef  SE Coef      T      P
Constant        -0.0676   0.1714  -0.39  0.713
Grade in School  1.01488  0.02424  41.86  0.000


S = 0.222195   R-Sq = 99.8%   R-Sq(adj) = 99.7%
```

Remember that the hypothesized value of the slope is 1. From the computer output, we can compute the value of the test statistic and then find the associated *P*-value (the values of the test statistic and the *P*-value aren't the ones shown in the computer output; the default computer output is for a null hypothesis of $\beta = 0$.

$$t = \frac{b-1}{s_b} = \frac{.015-1}{0.024} = 0.625$$

$$df = 4$$

$$P - \text{value} = 0.283$$

CONCLUSION If it were true that we would see a result like this or one more extreme than this about 28.3% of the time. Since this p-value is larger than $\alpha = 0.05$, we fail to reject the null hypothesis. We do not have evidence that the mean increase in grade level score for each additional year of school is greater than 1.

INTERPRETATION OF RESULTS OF HYPOTHESIS TESTING

(Introduction to Statistics & Data Analysis 3rd ed. pages 737–740/4th ed. pages 788–789)

Linear regression is often a useful way to summarize bivariate data. Researchers use linear regression hypothesis tests to make inferences about the way the two variables are related. You should ask yourself the following questions when you evaluate research that involves relating two variables.

1. Which variable is the response variable? Is it quantitative?
2. If the research uses a sample to make inferences about the population, is it reasonable to assume the conditions were met?
3. Does the model appear to be useful? Are the results of the hypothesis test given? What is the p-value of the test?
4. Has the linear model been used in an appropriate way? Has the research avoided extrapolation?
5. If a correlation coefficient is given, is it in the context of a test of significance? Are the results interpreted accurately?

Keep in mind the following limitations of linear regression inference:

- Small samples can show a weak linear pattern due to chance not due to a relationship in the population parameters.
- As with all inference, linear regression inference is not appropriate if the sample is not a random sample.
- As with all inference, linear regression inference procedures are not reasonable if the necessary conditions are not met.

Inference for Linear Regression and Correlation: Student Objectives for the AP Exam

- You should be able to find a point estimate β.
- You should be able to construct a confidence interval to estimate the slope of the population regression line.
- You should be able to perform a hypothesis test for the slope of a population regression line.

Multiple-Choice Questions

Questions 1–7 refer to the following information:

Data on x = number of powerboat registrations in Florida and y = number of manatee deaths in Florida for 31 randomly selected years was used to produce the following computer output. You can assume that all conditions needed for inference are met.

Regression Analysis: Manatees versus Registrations

Predictor	Coef	SE Coef	T	P
Constant	-40.641	5.840	-6.96	0.000
Registrations	0.125006	0.007867	15.89	0.000

S = 7.87985 R-Sq = 89.7% R-Sq(adj) = 89.3%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	15677	15677	252.48	0.000
Residual Error	29	1801	62		
Total	30	17477			

1. The estimate for the population regression line is
   (A) predicted manatees = –40.641 + 5.84(registrations)
   (B) predicted manatees = 0.125 + 0.007867(registrations)
   (C) predicted manatees = 5.840 + 0.007867(registrations)
   (D) predicted manatees = –40.641 + 0.125006(registrations)
   (E) predicted manatees = 7.87985 + 0.125006(registrations)

2. The standard error of the slope is shown to be 0.0079. Interpret this value in context.
   (A) The standard deviation of the number of powerboat registrations is 0.0079.
   (B) The standard deviation of the number of manatees killed is 0.0079.
   (C) If many samples were taken and the slope of each least squares regression line was recorded, the estimated standard deviation of these slopes is 0.0079.
   (D) If many samples were taken and the slopes of each least squares regression line were recorded, the difference in the estimated change in number of manatees killed per powerboat registration and the true change in number of manatees killed per powerboat registration is on average 0.0079.
   (E) The distance between the true change in number of manatees killed per powerboat registration and the sample change in number of manatees killed per powerboat registration is 0.0079.

3. Which of the following is a confidence interval for the mean change in number of manatee deaths associated with an increase of 1 powerboat registration?
   (A) $-40.641 \pm t^*(5.84)$
   (B) $0.125 \pm t^*(0.0079)$
   (C) $0.125 \pm t^*(15.89)$
   (D) $0.125 \pm t^*(0.0079/)$
   (E) $0.125 \pm t^*(0.0079/)$

4. The value of the t test statistic for testing is
   (A) 7.87985
   (B) -6.96
   (C) 15.89
   (D) 252.48
   (E) 89.3

5. What is the value of degrees of freedom that would be used in determining the P-value is a hypothesis test of $H_0 : \beta = 0$?
   (A) 1
   (B) 28
   (C) 29
   (D) 30
   (E) 32

6. In a test of $H_0 : \beta = 0$ versus $H_a : \beta \neq 0$ with a significance level of 0.05, the decision would be to
   (A) accept H_0
   (B) fail to reject H_a
   (C) fail to reject H_0
   (D) reject H_0
   (E) reject H_a

7. The estimated standard deviation of the residuals is
   (A) 5.84
   (B) 7.87985
   (C) 0.0079
   (D) 1801
   (E) 89.7

Questions 8–15 refer to the following information:

Each person in a random sample of adults was asked to estimate the average number of minutes of television watched per day and their yearly salary. The following data resulted.

Minutes of TV Watched	36	57	60	62	70	76	101
Annual Salary (in thousands)	89	45	30	50	55	67	30

This data was used to produce the following regression output.

Regression Analysis: Salary versus TV

```
The regression equation is
Salary = 94.4 - 0.638 TV


Predictor      Coef  SE Coef      T      P
Constant      94.40    25.60   3.69  0.014
TV          -0.6382   0.3736  -1.71  0.148


S = 18.2023   R-Sq = 36.9%   R-Sq(adj) = 24.2%
```

The researcher wishes to determine if these data provide evidence of a linear relationship between number of minutes spent watching television each day and annual salary. You can assume that all conditions required for inference are met.

8. The appropriate hypotheses are
   (A) $H_0 : \beta = 0;\ H_a : \beta \neq 0$
   (B) $H_0 : b = 0;\ H_a : b \neq 0$
   (C) $H_0 : \beta \neq 0;\ H_a : \beta = 0$
   (D) $H_0 : \beta = 0;\ H_a : \beta < 0$
   (E) $H_0 : \beta = 0;\ H_a : \beta > 0$

9. What is the value of the correlation coefficient for these data?
   (A) –0.369
   (B) –0.607
   (C) 0.148
   (D) 0.369
   (E) 0.607

10. What proportion of variability in salary can be explained by the linear relationship between TV watched and salary?
    (A) 0
    (B) 0.148
    (C) 0.369
    (D) 0.374
    (E) 18.202

11. The p-value for the test of $H_0 : \beta = 0$ versus $H_0 : \beta \neq 0$ is 0.148. A correct interpretation of this value in context is
    (A) The probability that we committed a Type I error is 0.149.
    (B) The probability that we committed a Type I or Type II error is 0.149.
    (C) If we were to take many samples of people and ask their TV viewing habits and their salary, 14.9% of the tests would yield different results.
    (D) If minutes watching TV and salary are not linearly related, we would get samples like this one or more extreme 14.9% of the time just by chance.
    (E) If minutes watching TV and salary are associated, we would detect the relationship only 14.9% of the time.

12. Using a significance level of 0.05, the correct conclusion for the test of $H_0 : \beta = 0$ versus $H_0 : \beta \neq 0$ in context is
    (A) Since the p-value is greater than the significance level, we accept the null hypothesis that minutes watching TV and salary are linearly related.
    (B) Since the p-value is greater than the significance level, we reject the null hypothesis that minutes watching TV and salary are not linearly related.
    (C) Since the p-value is greater than the significance level, we fail to reject the null hypothesis that minutes watching TV and salary are not linearly related.
    (D) Since the p-value is greater than the significance level, we fail to reject the null hypothesis that minutes watching TV and salary are linearly related.
    (E) Since the p-value is greater than the significance level, we have evidence that minutes watching TV and salary are not linearly related.

13. Which of the following is a correct description of a Type II error in context?
    (A) We conclude that minutes watching TV and salary are not linearly related when, in fact, they are linearly related.
    (B) We conclude that minutes watching TV and salary are linearly related when, in fact, they are not linearly related.
    (C) We conclude that minutes watching TV and salary are not linearly related when, in fact, they are not linearly related.
    (D) We conclude that minutes watching TV and salary are linearly related when, in fact, they are linearly related.
    (E) A Type II error happens 2% of the time when we run similar tests on similar samples.

14. Which of the following is not a necessary condition for inference about the slope of the population regression line?
    (A) The observations must be independent of one another.
    (B) The residuals for every value of x must have a mean of zero.
    (C) The residuals must be normally distributed at each x value.
    (D) The residuals have the same standard deviation for each value of x.
    (E) The sample size must be large.

15. Which of the following is true of all significance tests for the slope of a population regression line?
    (A) The alternative hypothesis is that there is a linear relationship between x and y.
    (B) The degrees of freedom for this test is one less than the sample size.
    (C) We always assume the four necessary assumptions are true no matter what the residual plot displays.
    (D) If the slope of the sample regression line is small, then we know we will have a small p-value.
    (E) The null hypothesis assumes no linear relationship between x and y.

FREE-RESPONSE PROBLEMS

1. Each person in a random sample of 42 students at a large university was asked how much they paid per month for housing and how far from campus they lived (in miles). The resulting data was used to produce the following regression output. You can assume that any conditions needed for inference are met.

```
The regression equation is
Housing cost = 452 - 9.25 Distance

Predictor      Coef  SE Coef       T       P
Constant      452.1    178.1    2.54   0.039
Distance   -9.2472   0.2145  -43.11   0.000

S = 221.098   R-Sq = 99.6%   R-Sq(adj) = 99.6%
```

    (a) Estimate the slope of the population regression line using a 90% confidence interval.
    (b) Interpret this interval in context.
    (c) A recent news report stated that students pay a premium to live near campus. Evaluate this statement based on the interval from part (a).

2. Each person in a random sample of adult males with children was asked his height and the height of his first full grown child, resulting in the following data.

Height of Father (inches)	68	68	73	72	74	71	73	73	71	70
Height of Child (inches)	61	65	65	73	61	64	65	70	66	63

    Is there convincing evidence of a linear relationship between the height of fathers and the height of their child?

Answers

Multiple-Choice Questions

1. **D.** (*Introduction to Statistics & Data Analysis* 3rd ed. page 215/4th ed. pages 748–749).

2. **C.** (*Introduction to Statistics & Data Analysis* 3rd ed. page 698/4th ed. page 750).

3. **B.** (*Introduction to Statistics & Data Analysis* 3rd ed. pages 704–706/4th ed. pages 757–758).

4. **C.** (*Introduction to Statistics & Data Analysis* 3rd ed. page 710/4th ed. pages 760–762).

5. **C.** (*Introduction to Statistics & Data Analysis* 3rd ed. page 705/4th ed. pages 757-758).

6. **D.** (*Introduction to Statistics & Data Analysis* 3rd ed. page 710/4th ed. pages 760-762).

7. **B.** (*Introduction to Statistics & Data Analysis* 3rd ed. page 700/4th ed. page 752).

8. **A.** (*Introduction to Statistics & Data Analysis* 3rd ed. page 707/4th ed. page 759).

9. **B.** (*Introduction to Statistics & Data Analysis* 3rd ed. page 707/4th ed. page 744).

10. **C.** This is the critical value associated with 1% of the area in each tail in a *t* distribution with 5 degrees of freedom (*Introduction to Statistics & Data Analysis* 3rd ed. page 709/4th ed. pages 760–761).

11. **D.** (*Introduction to Statistics & Data Analysis* 3rd ed. page 709/4th ed. pages 760–761).

12. **C.** (*Introduction to Statistics & Data Analysis* 3rd ed. page 709/4th ed. pages 760–761).

13. **A.** (*Introduction to Statistics & Data Analysis* 3rd ed. page 531/4th ed. page 583).

14. E. The residuals must be normally distributed for each and every *x* (*Introduction to Statistics & Data Analysis* 3rd ed. page 707/4th ed. page 744).

15. E. (*Introduction to Statistics & Data Analysis* 3rd ed. pages 707–710/4th ed. pages 758–762).

FREE-RESPONSE QUESTIONS

1. (a) The question states that the conditions for inference are met, so it is OK to use a t confidence interval to estimate the slope. With $n = 42$, df = 40 and the t critical value for a 90% confidence interval is 1.68. The confidence interval is then

$-9.2472 \pm (1.68)(0.2145)$

-9.2472 ± 0.3604

$(-9.6076, -8.8868)$

(b) On average, the mean housing cost decreases by somewhere between \$8.89 and \$9.61 with each additional mile from campus.

(c) Housing costs do tend to decrease as you move farther away from campus. A student who chose to live 5 miles from campus might expect to pay about \$45 less per month than someone who lives adjacent to the campus, and a student who lives 10 miles from campus would expect to pay about \$90 less per month.

(*Introduction to Statistics & Data Analysis* 3rd ed. pages 703–706/4th ed. pages 756–758).

2. **Name of Test:** t test for slope of regression line

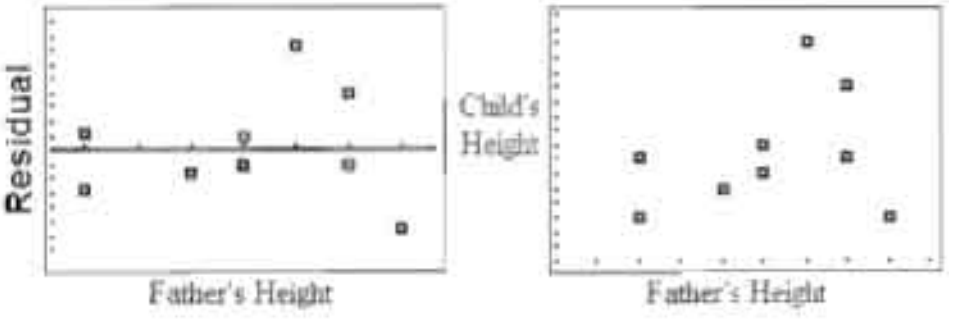
Hypotheses:

$H_0 : \beta = 0$

$H_a : \beta \neq 0$

Conditions:

First, look at a scatterplot and a residual plot of the sample data (shown below). A linear model seems appropriate for this data.



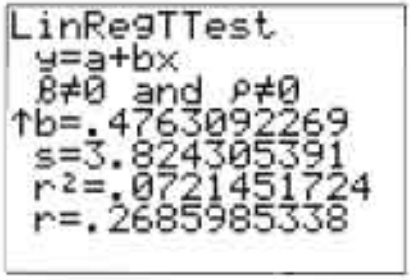
Assumption 1: for each and every value of x. This assumption is reasonable when the pattern in the scatterplot is linear and there is no obvious pattern in the residual plot, which appears to be the case here.

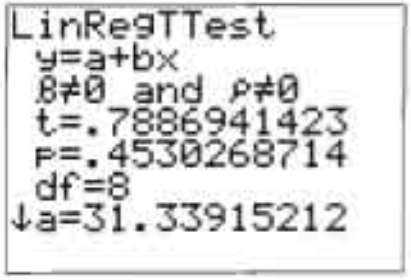
Assumption 2: is the same for each and every value of x. The scatter about the regression line does not appear to be much larger for some x values than for others, so this assumption seems reasonable.

Assumption 3: The distribution of e is normal for any particular value of x. For this assumption to be reasonable, points should tend to cluster near the line, with fewer points as you move away from the line. This appears to be the case here, so this assumption seems reasonable.

Assumption 4: Because the sample was a random sample, the observations are independent.

Calculations:





$T = 0.79$ with 8 degrees of freedom, P-value = 0.453

Conclusion:

Since the P-value of 0.453 is greater than any reasonable alpha level, we fail to reject the null hypothesis. There is not convincing evidence that the father's height and the child's height are linearly related

(*Introduction to Statistics & Data Analysis* 3rd ed. pages 707–710/4th ed. pages 758–762).