Name:_____ Date:_____

HW Chapter 15: Confidence Intervals for Least Square Regression Lines:

Surveys conducted using random digit telephone dialing increasingly have reliability issues because of the number of dialed households that refuse to participate. This *refusal rate*, that is, the percent of dialed households that refuse, may be linked to per capita income. Data were collected from a dozen counties in Florida and entered into a statistics package. Regression analysis was requested, with the following results.

Predictor	Coef	Stdev	t-ratio	p
Constant	0.3537	0.1686	2.10	0.062
Income	0.00000443	0.00001732	0.26	0.803

s = 0.1011 R-sq = 0.6% R-sq(adj) = 0.0%

1. What is the equation of the least-squares regression line?

2. What is the correlation? Interpret the correlation in the context of this problem.

3. County #10 had a per capita income of $8636 and a refusal rate of 0.191. What is the predicted refusal rate for a county with a per capita income of $8636, according to this least-squares model?

4. What is the value of the residual for the point (8636, 0.191)?

5. Would you be comfortable using this model to predict refusal rates for other counties in Florida? Explain your reasoning.

6. The output also reports the results of a test of significance. Summarize the results of this test.

There is some evidence that drinking moderate amounts of wine helps prevent heart attacks. The table gives data on yearly wine consumption (liters of alcohol from drinking wine, per person) and yearly deaths from heart disease (deaths per 100,000 people) in a dozen developed nations.

Country	Alcohol from wine	Heart disease deaths
Austria	3.9	167
Canada	2.4	191
Denmark	2.9	220
France	9.1	71
Ireland	0.7	300
Italy	7.9	107
Spain	6.5	86
Sweden	1.6	207
Switzerland	5.8	115
United Kingdom	1.3	285
United States	1.2	199
West Germany	2.7	172

1. Make a scatterplot that shows how wine consumption affects heart disease. Label the vertical axis from 50 to 300 with increments of 25

2. Use your calculator to obtain the LSRL equation and correlation.

3. Formulate null and alternative hypotheses about the slope of the true regression line.

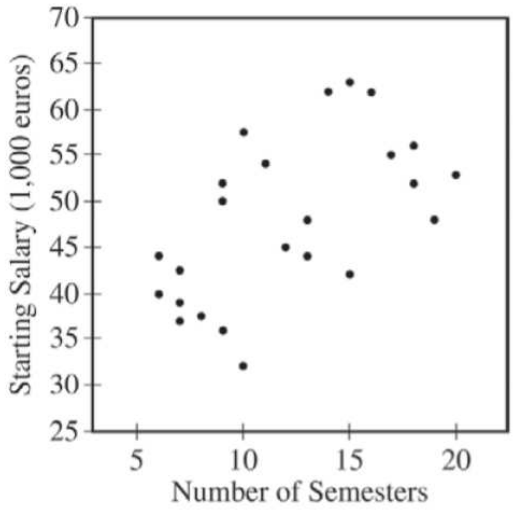
4. Report the sum of the 12 residuals and the sum of the squares of the residuals. What is the value of *s* (the standard error about the line)?

5. The model for regression inference has 3 parameters: α, β, and σ. Estimate these parameters from the data.

6. Computer output reports that the standard error of the slope is $SE_b = 3.511$. Use this to construct a 95% confidence interval for the slope β of the true regression line.

Q7 (AP 2016)

A newspaper in Germany reported that the more semesters needed to complete an academic program at the university, the greater the starting salary in the first year of a job. The report was based on a study that used a random sample of 24 people who had recently completed an academic program. Information was collected on the number of semesters each person in the sample needed to complete the program and the starting salary, in thousands of euros, for the first year of a job. The data are shown in the scatterplot below.



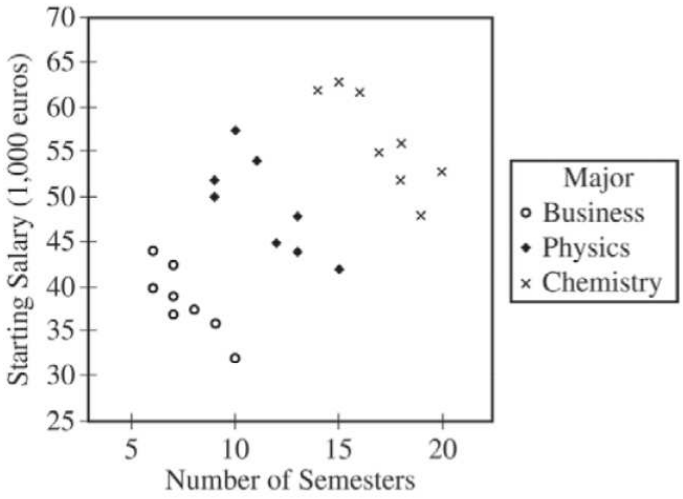
(a) Does the scatterplot support the newspaper report about number of semesters and starting salary? Justify your answer.

The table below shows computer output from a linear regression analysis on the data.

Predictor	Coef	SE Coef	T	P
Constant	34.018	4.455	7.64	0.000
Semesters	1.1594	0.3482	3.33	0.003
S = 7.37702	R-Sq = 33.5%		R-Sq(adj) = 30.5%	

(b) Identify the slope of the least-squares regression line, and interpret the slope in context.

An independent researcher received the data from the newspaper and conducted a new analysis by separating the data into three groups based on the major of each person. A revised scatterplot identifying the major of each person is shown below.



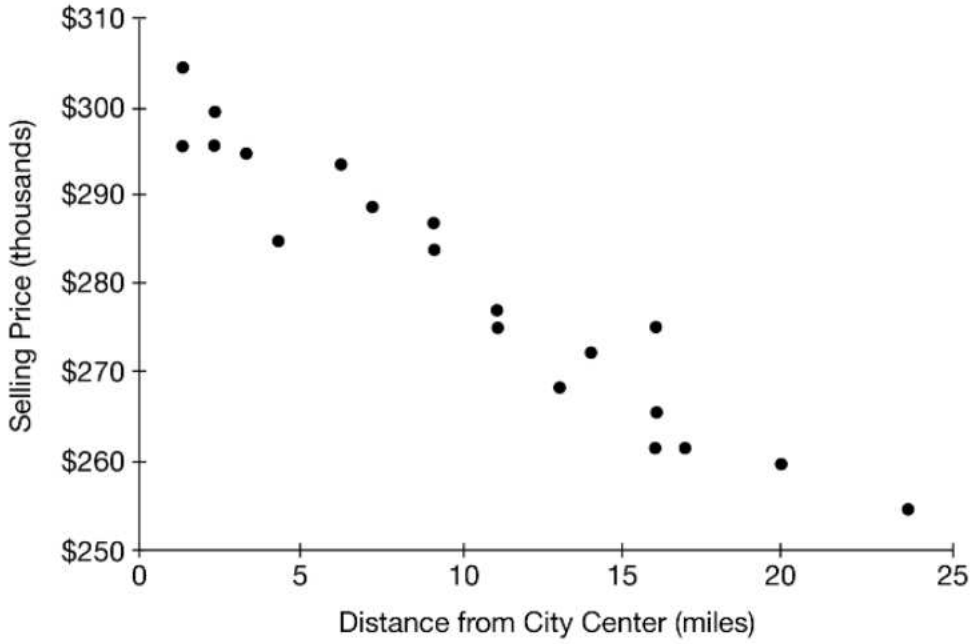
(c) Based on the people in the sample, describe the association between starting salary and number of semesters for the business majors.

(d) Based on the people in the sample, compare the median starting salaries for the three majors.

(e) Based on the analysis conducted by the independent researcher, how could the newspaper report be modified to give a better description of the relationship between the number of semesters and the starting salary for the people in the sample?

Q8 (AP 2019)

A real estate agent working in a large city believes that, for three-bedroom houses, the selling price of the house decreases by approximately $2,000 for every mile increase in the distance of the house from the city center. To investigate the belief, the agent obtained a random sample of 20 three-bedroom houses that sold in the last year. The selling price, in thousands of dollars, and the distance from the city center, in miles, for each of the 20 houses are shown in the scatterplot. The table shows computer output from a regression analysis of the data.



Predictor	Coef	SE Coef	T	P
Constant	301.7	1.80	167.17	0.000
Distance	−2.158	0.149	−14.45	0.000
	S = 4.4336		R-sq = 92.1%	

(b) Does the confidence interval contradict the agent's belief about the relationship between selling price and distance from the city center? Justify your answer.

SOLUTION: Q7 (2017)

Intent of Question

The primary goals of this question were to assess a student's ability to (1) use a scatterplot to comment on a report about the relationship between two variables and interpret the slope for the least-squares regression line summarizing this relationship; (2) describe the relationship between two variables in a scatterplot when a categorical variable is introduced and compare a characteristic of the distribution of a variable for different categories of individuals in a scatterplot; and (3) describe how the associations between two variables for each category of individuals in a scatterplot differ from the overall association in the same scatterplot.

Solution

Part (a):

The scatterplot supports the newspaper report about number of semesters needed to complete an academic program and starting salary because it shows a positive association between these two variables.

Part (b):

The slope is 1.1594. For each additional semester needed to complete an academic program, the predicted starting salary increases by €1,159.40.

Part (c):

For the business majors alone, there is a strong, negative, linear association between number of semesters and starting salary. Business majors who need a greater number of semesters to complete an academic program tend to have lower starting salaries.

Part (d):

Business majors have the lowest median starting salary at around €38,000, followed by physics majors at around €48,000, and then chemistry majors with the highest median starting salary at around €55,000.

Part (e):

The newspaper report should be modified to account for major. Overall, majors that take longer to complete tend to have higher starting salaries, with chemistry the highest, physics the next highest, and business the lowest. However, within a major, students who take a greater number of semesters tend to have lower starting salaries.

Scoring

This question is scored in three sections. Section 1 consists of parts (a) and (b), section 2 consists of parts (c) and (d), and section 3 consists of part (e). Sections 1, 2, and 3 are scored as essentially correct (E), partially correct (P), or incorrect (I).

Section 1 is scored as follows:

Essentially correct (E) if the response includes the following five components:

1. In part (a) the response addresses the positive association.
2. In part (a) the response uses the positive association to justify that the scatterplot supports the newspaper report.
3. In part (b) the response correctly identifies the numerical value of the slope from the computer output.
4. In part (b) the response interprets the slope as the change in starting salary for each additional semester, in context.
5. In part (b) the interpretation of slope includes nondeterministic language (e.g., "predicted starting salary" or equivalent) when interpreting the slope.

Partially correct (P) if the response includes three or four of the five components.

Incorrect (I) if the response does not meet the criteria for E or P.

Notes:

- In part (a) the response can use phrases such as "positive association (correlation, relationship)," "increasing relationship," or describe a positive association (e.g., "the starting salaries are higher when there is a greater number of semesters") to satisfy component 1. However, describing the relationship between only two points does not satisfy this component.
- In part (a) comments about linearity and strength should be ignored, even if the response implies these are required (e.g., "Yes, because the relationship is strong, positive, and linear.").
- In part (a) responses that answer "no," "maybe," "kind of," "somewhat," or equivalent do not satisfy component 2.
- In part (a) a response that says "no, because the three clusters of points each have a negative association" does not satisfy component 1 or component 2.
- In part (a) no context is required, but variable names are required in part (b) to satisfy component 4.
- In part (b) a response that states the equation $\hat{y} = 34.018 + 1.1594x$ does not satisfy component 3 unless the slope is specifically identified or used in the interpretation.
- In part (b) a response that incorrectly identifies the numerical value of the slope can still satisfy components 4 and 5 using the incorrect value.
- In part (b) the 1-unit increase in number of semesters must be stated or implied to satisfy component 4 (e.g., for each semester, for every semester). A response that states or implies an unspecified number of semesters does not satisfy this component (for example, as semesters increase).
- In part (b) examples of nondeterministic language include "predicted starting salary," "expected starting salary," "estimated starting salary," "typical starting salary," "average starting salary," "starting salary, on average," "our model predicts," and so on. However, "about," "approximately," and "according to the model" do not satisfy component 5.
- In part (b) no units are required for the change in predicted starting salary (which means it is OK to say 1.1594 euros, 1159.40, 1.1594, 1,159.40 dollars, and so on).
- In all parts it is acceptable if a response refers to salary rather than starting salary.

Section 2 is scored as follows:

Essentially correct (E) if the response includes the following five components:

1. In part (c) the response states that the association is negative.
2. In part (c) the response states that the association is strong *OR* linear *OR* both.
3. In part (c) the response refers to both variables (semesters, salary) in context.
4. In part (d) the response correctly compares the three majors.
5. In part (d) the response provides reasonable values for the median salaries or refers to "median starting salaries" when describing a characteristic of the graph.

Partially correct (P) if the response includes three or four of the five components.

Incorrect (I) if the response does not meet the criteria for E or P.

Notes:

- In part (c) the response can use phrases such as "negative association (correlation, relationship)," "decreasing relationship," "inversely related," or describe a negative association (for example, "the starting salaries are smaller when there is a greater number of semesters") to satisfy component 1. However, describing the relationship between only two points does not satisfy this component.
- In part (c) responses that describe the relationship incorrectly as weak or nonlinear do not satisfy component 2, even if the other characteristic is described correctly.
- Only drawing a line on the scatterplot does not satisfy component 2.
- In part (d) a response that says "chemistry has the highest median starting salary and business has the lowest median starting salary" (or the equivalent) implies that physics is in the middle, and satisfies component 4.
- In part (d) if no values are provided for the medians, the response must use the phrase "median starting salary" at least once to satisfy component 5.
- In all parts it is acceptable if a response refers to salary rather than starting salary.

Section 3 is scored as follows:

Essentially correct (E) if the response states that there is a negative association for each of the majors *AND* the response notes the overall positive association.

Partially correct (P) if the response states that there is a negative association for each of the majors BUT does not note the overall positive association;
OR
if the response states that there is a negative association for one or two specific majors (for example, for business majors) *AND* the response notes the overall positive association.

Incorrect (I) if the response does not meet the criteria for E or P.

Notes:

- Additional ways to note the overall positive association include stating that the original report was correct, stating that majors that take longer to complete tend to have higher starting salaries, or the equivalent.

- A response that does not explicitly name the majors or refer to the three majors but makes a general statement such as "However, for an academic major, there is a negative association" satisfies the requirement to state the negative association for each of the majors.
- A response that states the original report was wrong, or incorrect, or should be retracted, etc., cannot satisfy the requirement to note the overall positive association. However, a response that says the original report might be misleading (or the equivalent) but still notes the overall positive association satisfies the requirement to note the overall positive association.
- In all parts it is acceptable if a response refers to salary rather than starting salary.