AP® CollegeBoard

AP Statistics
Sample Student Responses
and Scoring Commentary

Inside:

Free Response Question 1

☑ **Scoring Guideline**

☑ **Student Samples**

☑ **Scoring Commentary**

AP® STATISTICS
2018 SCORING GUIDELINES

Question 1

Intent of Question

The primary goals of this question were to assess a student's ability to (1) identify various values in regression computer output; (2) interpret the intercept of a regression line in context; (3) interpret the coefficient of determination (r^2) in context; and (4) identify an outlier from a scatterplot.

Solution

Part (a):

The estimate of the intercept is 72.95. It is estimated that the average time to finish checkout if there are no other customers in line is 72.95 seconds.

Part (b):

The coefficient of determination is $r^2 = 73.33\%$. This value indicates that 73.33% of the variability in the times it takes customers to finish checkout, including time waiting in line, can be explained by knowing how many customers are in line in front of the selected customer.

Part (c):

The outlier is the point with $x = 3$ and y close to 0. This point is considered an outlier because the combination of x and y values differs from the pattern of the rest of the data. Specifically, the value of y (time to finish checkout) is much lower than would be expected when there are $x = 3$ customers in line in front of the selected customer, given the remaining data.

Scoring

Parts (a), (b), and (c) are scored as essentially correct (E), partially correct (P), or incorrect (I).

Part (a) is scored as follows:

Essentially correct (E) if the response satisfies the following three components:
1. Correctly identifies 72.95 as the intercept.
2. Communicates the concept of a y-intercept in a context that includes both time and zero customers.
3. Indicates that the value of the intercept is a prediction by using language such as "predicted," "estimated," or "average" value of y.

Partially correct (P) if the response includes only two of the three components.

Incorrect (I) if the response includes at most one of the three components.

Notes:
- Regression equations (such as $\hat{y} = 72.95 + 174.40x$) cannot be used to satisfy identification of the intercept in component 1, unless the intercept is explicitly labeled as such.
- A regression equation cannot be used to satisfy component 3.
- Incorrect regression equations are treated as extraneous and do not affect the scoring of any component.
- A response that interprets 72.95 as a slope does not satisfy components 1 or 2.

Part (b) is scored as follows:

Essentially correct (E) if the response satisfies the following three components:
1. Correctly identifies 73.33% as the coefficient of determination.
2. Provides a correct (possibly generic) interpretation of r^2.
3. Interpretation includes context.

Partially correct (P) if the response satisfies only two of the three components;
> OR

if the response satisfies the three components, but reverses the roles of number of customers in line and time to finish checkout in the interpretation.

Incorrect (I) if the response satisfies at most one of the three components.

Notes:
- In component 2 the correct interpretation of the coefficient of determination can take any of several equivalent forms, such as:
  - The percent variability in y that is attributed to the linear relationship between y and x or between x and y.
  - The proportion of the total variability in the dependent variable y that is explained by the independent variable x.
  - The proportion of variation in y that is accounted for by the linear model.
  - The proportionate reduction of total variation of the y values that is associated with the use of the independent variable x.
  - The proportionate reduction in the sum of the squares of vertical deviations obtained by using the least-squares line instead of the naïve prediction of $\bar{y}$.
- In component 2 common *incorrect* interpretations of the coefficient of determination include:
  - The percent variability in the *predicted* y values that is explained by the linear relationship between y and x.
  - The percent variability in the *data* that is explained by the linear relationship between y and x.
  - The percent variability that is explained by the linear relationship between y and x.
  - The percent variability in y that is *on average* explained by the linear relationship between y and x.
- For component 3 context must include mention of time or customers.

Part (c) is scored as follows:

Essentially correct (E) if the response satisfies the following two components:
1. Correctly identifies the outlier.
2. Describes an unusual feature of the identified scatter plot point, relative to the remaining data points, that is sufficient to identify it as the outlier. Examples include:
   - The combination of x and y values is unusual compared to the other points.
   - The value of y is much lower than would be expected (or predicted), given the remaining data.
   - The residual for the point is unusually large relative to the other residuals.

Partially correct (P) if the response satisfies component 1 but does not satisfy component 2.

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

Notes:
- In the absence of any point being circled on the graph, component 1 can still be satisfied by explicitly referring to the coordinates of the outlier. Valid coordinates for outlier identification must specify an x value of 3 and a y value that is strictly between 0 and 250.
- A response that does not make a comparison to the remaining data points, such as stating the outlier has a large residual or is nowhere near the regression line, does not satisfy component 2.
- A response that makes a comparison to the remaining data points based upon an unusual feature that is *insufficient* for outlier identification, such as stating the point is the only point with that particular y value, does not satisfy component 2.
- In the absence of explicit numerical calculation, a response that appeals to the influence that the outlier has on the regression coefficient estimates or on the sample correlation coefficient does not satisfy component 2.

Question 1 (continued)

4   Complete Response

Three parts essentially correct

3   Substantial Response

Two parts essentially correct and one part partially correct

2   Developing Response

Two parts essentially correct and no parts partially correct

OR

One part essentially correct and one or two parts partially correct

OR

Three parts partially correct

1   Minimal Response

One part essentially correct

OR

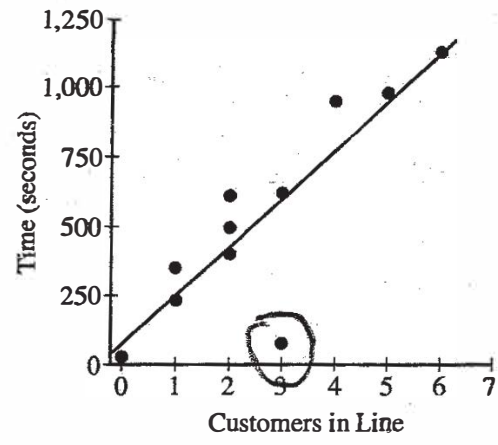No parts essentially correct and two parts partially correct

STATISTICS
SECTION II
Part A
Questions 1-5
Spend about 1 hour and 5 minutes on this part of the exam.
Percent of Section II score—75

Directions: Show all your work. Indicate clearly the methods you use, because you will be scored on the correctness of your methods as well as on the accuracy and completeness of your results and explanations.

1. The manager of a grocery store selected a random sample of 11 customers to investigate the relationship between the number of customers in a checkout line and the time to finish checkout. As soon as the selected customer entered the end of a checkout line, data were collected on the number of customers in line who were in front of the selected customer and the time, in seconds, until the selected customer was finished with the checkout. The data are shown in the following scatterplot along with the corresponding least-squares regression line and computer output.

Predictor	Coef	SE Coef	T	P
Constant	72.95	110.36	0.66	0.525
Customers in line	174.40	35.06	4.97	0.001

S = 200.01	R-Sq = 73.33%	R-Sq (adj) = 70.37%

GO ON TO THE NEXT PAGE.

(a) Identify and interpret in context the estimate of the intercept for the least-squares regression line.

The y intercept is 72.95 seconds. If there are no people in front of the customer, we would expect them to be finished with checkout in 72.95 seconds.

(b) Identify and interpret in context the coefficient of determination, r^2.

r^2 is 73.33%. 73.33% of the variation in checkout time is accounted for by the linear relationship between customers in line and checkout time.

(c) One of the data points was determined to be an outlier. Circle the point on the scatterplot and explain why the point is considered an outlier.

That point is an outlier because it does not follow the pattern the rest of the data points follow and is very far from the rest of the data.

GO ON TO THE NEXT PAGE.
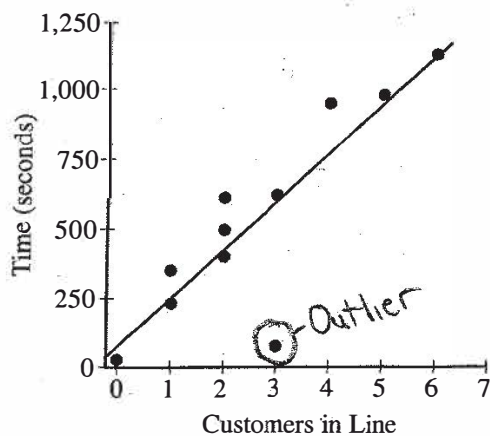
STATISTICS

SECTION II

Part A

Questions 1-5

Spend about 1 hour and 5 minutes on this part of the exam.

Percent of Section II score—75

Directions: Show all your work. Indicate clearly the methods you use, because you will be scored on the correctness of your methods as well as on the accuracy and completeness of your results and explanations.

1. The manager of a grocery store selected a random sample of 11 customers to investigate the relationship between the number of customers in a checkout line and the time to finish checkout. As soon as the selected customer entered the end of a checkout line, data were collected on the number of customers in line who were in front of the selected customer and the time, in seconds, until the selected customer was finished with the checkout. The data are shown in the following scatterplot along with the corresponding least-squares regression line and computer output.

Predictor	Coef	SE Coef	T	P
Constant	72.95	110.36	0.66	0.525
Customers in line	174.40	35.06	4.97	0.001

S = 200.01   R-Sq = 73.33%   R-Sq (adj) = 70.37%

GO ON TO THE NEXT PAGE.

(a) Identify and interpret in context the estimate of the intercept for the least-squares regression line.

The estimate of the intercept, which is 72.95 seconds, means that if there are no customers in the line, the predicted time to finish checkout is 72.95 seconds.

(b) Identify and interpret in context the coefficient of determination, r^2.

The r^2 value of 73.33% means that about 73.33% of the variation of time to finish checkout, y, can be explained by the least-squares regression line of customers in line, x, and time to finish checkout, y.

(c) One of the data points was determined to be an outlier. Circle the point on the scatterplot and explain why the point is considered an outlier.

This point is considered an outlier because its value is very far from the predicted value of the least-squares regression line. The point's value is about 100 while, when there are 3 customers in line, the LSRL predicts a value of about 600.

GO ON TO THE NEXT PAGE.
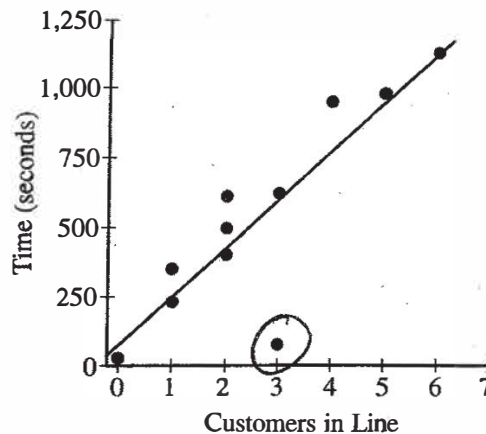
STATISTICS
SECTION II
Part A
Questions 1-5
Spend about 1 hour and 5 minutes on this part of the exam.
Percent of Section II score—75

Directions: Show all your work. Indicate clearly the methods you use, because you will be scored on the correctness of your methods as well as on the accuracy and completeness of your results and explanations.

1. The manager of a grocery store selected a random sample of 11 customers to investigate the relationship between the number of customers in a checkout line and the time to finish checkout. As soon as the selected customer entered the end of a checkout line, data were collected on the number of customers in line who were in front of the selected customer and the time, in seconds, until the selected customer was finished with the checkout. The data are shown in the following scatterplot along with the corresponding least-squares regression line and computer output.

Predictor	Coef	SE Coef	T	P
Constant	72.95	110.36	0.66	0.525
Customers in line	174.40	35.06	4.97	0.001
S = 200.01		R-Sq = 73.33%	R-Sq (adj) = 70.37%	

GO ON TO THE NEXT PAGE.

(a) Identify and interpret in context the estimate of the intercept for the least-squares regression line.

The estimate given for the intercept is 72.95 seconds. This means that with 0 customers in line, one would expect the time to finish checkout to be approximately 72.95 seconds.

(b) Identify and interpret in context the coefficient of determination, r^2.

r^2 is given by the output as 70.37%, or 0.7037. This indicates that approximately 70.37% of the variance of results from the expected times (regression line) is accounted for by the least-squares regression line.

(c) One of the data points was determined to be an outlier. Circle the point on the scatterplot and explain why the point is considered an outlier.

The point at approximately (3, 100) is considered an outlier because it is significantly further from the regression line, or expected time for its number of customers, than any other point. This negatively affects the accuracy of the regression line.

GO ON TO THE NEXT PAGE.

AP® STATISTICS
2018 SCORING COMMENTARY

Question 1

Overview

The primary goals of this question were to assess a student's ability to (1) identify various values in regression computer output; (2) interpret the intercept of a regression line in context; (3) interpret the coefficient of determination in context; and (4) identify an outlier from a scatterplot.

Sample: 1A
Score: 4

In part (a) the response correctly recognizes the value of the intercept, satisfying component 1. The response then communicates the concept of an intercept using context that incorporates both time and zero customers; this satisfies component 2. Because the interpretation of the intercept indicates that the value is a prediction, as indicated by "we would expect them to be finished with checkout in 72.95 seconds," component 3 is satisfied. This response includes all three components; therefore, part (a) was scored as essentially correct. In part (b) the response correctly recognizes the value of r^2, satisfying component 1. The response provides a correct interpretation of r^2; this satisfies component 2. Because the interpretation is made using context, component 3 is satisfied. The response includes all three components; therefore, part (b) was scored as essentially correct. In part (c) the outlier is circled on the scatterplot, satisfying component 1. The response gives valid reasoning why the circled point is the outlier, relative to the remaining data points, by stating the circled point "does not follow the pattern the rest of the data points follow," and that the circled point is "very far from the rest of the data." Component 2 is satisfied. The response includes both components; therefore, part (c) was scored as essentially correct. Because three parts were scored as essentially correct, the response earned a score of 4.

Sample: 1B
Score: 3

In part (a) the response correctly recognizes the value of the intercept, satisfying component 1. The response also communicates the concept of an intercept using context that incorporates both time and zero customers; this satisfies component 2. Because the interpretation of the intercept indicates that the value is a prediction, as indicated by "the predicted time to finish checkout," component 3 is satisfied. The response includes all three components; therefore, part (a) was scored as essentially correct. In part (b) the response correctly recognizes the value of r^2, satisfying component 1. The response provides a correct interpretation of r^2; this satisfies component 2. Because the interpretation is made using context, component 3 is satisfied. The response includes all three components; therefore, part (b) was scored as essentially correct. In part (c) the outlier is circled on the scatterplot, satisfying component 1. The response does not, however, give sufficient reasoning to explain why the circled point is the outlier. The statement, "its value is very far from the predicted value of the least-squares regression line," does not make a comparison to the variation in the remaining data points, nor does it reference the distance between the y-coordinate of the circled point and the value of the prediction at $x = 3$. Therefore, in the absence of any comparison against the distances of all other points to the regression line, the response does not satisfy component 2. The response includes component 1 but not component 2; therefore, part (c) was scored as partially correct. Because two parts were scored as essentially correct, and one part was scored as partially correct, the response earned a score of 3.

Sample: 1C
Score: 2

In part (a) the response correctly recognizes the value of the intercept, satisfying component 1. The response then communicates the concept of an intercept using context that incorporates both time and zero customers; this satisfies component 2. The interpretation of the intercept indicates that the value is a prediction in two different ways; the first is indicated by "one would expect the time to finish checkout to be," and the other is indicated by "approximately." Either of these indications satisfies component 3. The response includes all three components; therefore, part (a) was scored as essentially correct. In part (b) the response gives an incorrect value of r^2, so component 1 is not satisfied. The response incorrectly interprets r^2 in terms of the percent of the variance in "results from the expected times," instead of percent variance in the observed times. Therefore the response does not satisfy component 2. Because the interpretation is made using context, component 3 is satisfied. The response includes only one of the three components, consequently, part (b) was scored as incorrect. In part (c) the outlier is circled on the scatterplot, satisfying component 1. The response gives valid reasoning why the circled point is the outlier by stating the circled point is "significantly farther from the regression line … than any other point." It is this portion of the response that compares the circled point to the remaining data points and, therefore, the response satisfies component 2. Because the response includes both components, part (c) was scored as essentially correct. Because two parts were scored as essentially correct, and one part was scored as incorrect, the response earned a score of 2.