AP® STATISTICS
2010 SCORING GUIDELINES (Form B)

Question 6

Intent of Question

The primary goals of this investigative task were to assess students' ability to understand, apply and draw conclusions from a regression analysis beyond what they have previously studied. More specific goals were to assess students' ability to (1) interpret a slope coefficient and residual value; (2) interpret a confidence interval; (3) compare two regression models and draw appropriate conclusions.

Solution

Part (a):

The slope coefficient is 0.165. This means that for each additional square foot of size, the predicted price of the house increases by 0.165 thousand dollars, which is \$165. In other words, this model predicts that the average price of a house increases by \$165 for each additional square foot of a house's size.

Part (b):

The residual value of 49 for this house indicates that its actual price is 49 thousand dollars higher than the model would predict for a house of its size.

Part (c):

The average residual value for the eight houses with a swimming pool is:

$$\frac{(6 + 49 + (-18) + 42 + 1 + 50 + 9 + (-23) + 42)}{8} = \frac{149}{8} = 18.6 \text{ thousand dollars.}$$

The average residual value for the 17 houses with no swimming pool is:

$$\frac{(13 + 26 + (-45) + \ldots + (-58) + (-52) + 33)}{17} = \frac{-150}{17} = -8.8 \text{ thousand dollars.}$$

The residual averages suggest that the regression line tends to underestimate the price of homes with a swimming pool by about 18.6 thousand dollars and to overestimate the price of homes with no pool by about 8.8 thousand dollars. The difference between these two residual averages is $18.6 - (-8.8) = 27.4$ thousand dollars. This suggests that, for two houses of the same size, the house with a swimming pool would be estimated to cost \$27,400 more than the house with no swimming pool.

Part (d):

No, this confidence interval does *not* indicate a significant difference (at the 95 percent confidence level, equivalent to the 5 percent significance level) between the two slope coefficients because the interval includes the value zero.

AP® STATISTICS
2010 SCORING GUIDELINES (Form B)

Question 6 (continued)

Part (e):

If the two population regression lines do in fact have the same slope, the impact of a swimming pool is the (constant) vertical distance between the two lines. However, because the two fitted lines do not have the same slope, the distance between the two fitted lines depends on the size of the house. Using the available information, there are two acceptable approaches to estimating the impact of having a swimming pool.

Approach 1: Use the two fitted lines to predict the price of a house with and without a pool for a particular house size. For example, using the value of size = 2,250 square feet (which is near the middle of the distribution of house sizes), we find:

Predicted price for a 2,250 square-foot house with a swimming pool =
$-11.602 + 0.166 \times 2,250 = 361.898$ thousand dollars.

Predicted price for a 2,250 square-foot house with no swimming pool =
$-27.382 + 0.160 \times 2,250 = 332.618$ thousand dollars.

The difference in these predicted prices is $361.898 - 332.618 = 29.280$ thousand dollars, which is an estimate of the impact of a swimming pool on the predicted price of a 2,250 square-foot house. This is quite similar to the estimate based on residuals in part (c).

Approach 2: Because the slopes of the two sample regression lines were judged not to be significantly different, another acceptable approach would be to use the difference in the intercepts of the two fitted lines as an estimate of the vertical distance between the two population regression lines.

The difference in the intercepts of the two fitted lines is $-11.602 - (-27.382) = 15.780$ thousand dollars, which is an estimate of the impact of a swimming pool on the predicted price of a house, assuming this difference does not change with the size of the house. This is quite different from the estimate based on residuals in part (c).

Scoring

This question is scored in four sections. Section 1 consists of part (a); section 2 consists of part (b); section 3 consists of part (c); section 4 consists of parts (d) and (e). Each of the four sections is scored as essentially correct (E), partially correct (P) or incorrect (I).

Section 1 is scored as follows:

Essentially correct (E) if the response identifies the correct value for the slope coefficient and provides a correct interpretation in context.

Partially correct (P) if the response identifies the correct value for the slope coefficient and provides a correct interpretation but not in context *OR* the response provides an incorrect value for the slope but provides a correct interpretation of this value in context *OR* the response identifies the correct value for the slope but the interpretation is incomplete because of one or more of the following errors:

- The interpretation does not mention "predicted" or "on average" or any other indication of a probabilistic rather than a deterministic relationship.

Question 6 (continued)

- The interpretation does not include the notion of each *additional* square foot of size by saying something like "for every square foot."
- The interpretation does not use units for the price variable, or it uses incorrect units for the price variable (e.g., dollars instead of thousands of dollars).

Incorrect (I) if there is no interpretation or if the interpretation does not warrant a score of P.

Note: It is possible to earn an E for section 1 without stating the actual numerical value of the slope, if a correct and well-communicated interpretation of the slope is given in context.

Section 2 is scored as follows:

Essentially correct (E) if the response provides a correct interpretation of the residual value, in context, including both direction and a comparison with the model's predicted or average value (e.g., actual price is higher than predicted).

Partially correct (P) if the response provides an interpretation of the residual value that fails to mention direction or that gives the incorrect direction *OR* if the response provides a correct interpretation of the residual value that includes direction, but that is not in context.

Incorrect (I) if there is no interpretation of the residual value *OR* the interpretation does not include direction and is not in context.

Section 3 is scored as follows:

Essentially correct (E) if the response correctly calculates averages of residual values both for houses with pools and houses without pools *AND* correctly reports the difference between those averages as the estimate of the impact of a swimming pool.

Partially correct (P) if the response either correctly calculates averages of residual values both for houses with pools and houses without pools but does not correctly report the difference between those averages as the estimate of the impact of a swimming pool *OR* incorrectly calculates one or both averages of residual values but does report the difference between those averages as the estimate of the impact of a swimming pool *OR* does not use all of the residual values but does use a reasonable set of residual values (such as houses of similar size) and correctly calculates both averages and correctly reports the difference between those averages as the estimate of the impact of a swimming pool.

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

Notes:

- If the student calculates some other measure of center for the two sets of residuals (e.g., medians) and reports the difference as the estimate of the impact of a swimming pool, this part can be scored, at best, partially correct (P).
- If the student estimates the values of the residuals from the residual plot rather than using the residuals provided in the table, the response can be scored as essentially correct (E), provided it is clear that this is what was done.

AP® STATISTICS
2010 SCORING GUIDELINES (Form B)

Question 6 (continued)

Section 4 is scored as follows:

Essentially correct (E) if the response includes all three of the following components:

1. Correctly notes that the confidence interval in part (d) includes zero and so the difference in the slopes is not statistically significant.
2. Calculates a reasonable estimate in part (e):
    - For approach 1, this includes choosing a house size within the range of the data and correctly computing the difference in predicted prices.
    - For approach 2, this includes appealing to the fact that the slopes were judged as not significantly different and computing the difference in intercepts.
3. Includes a comparison of the estimate in part (e) to the estimate in part (c).

Partially correct (P) if the response includes only one of (1) and (2) above.

Incorrect (I) if the response includes neither (1) nor (2) above.

Notes

- If the response uses approach 1, the difference between the two predicted values can range from 25.38 to 33.44, depending on the house size used.
- If the response uses approach 2, the constant vertical distance can be estimated from the graph showing the two regression lines rather than on the difference in intercepts, provided that the response makes it clear that this is what is being done.
- In the comparison with the estimate in part (c), an assessment of the size of the difference in estimates is not required. Statements that merely use phrases like "greater than," "about the same," etc. are acceptable for the comparison component of parts (d) and (e).
- If this section receives a score of partially correct only because the student neglects to compare the estimate in part (e) to the estimate in part (c), the response should be scored up if a decision on whether to score up or down is required.
- If the response subtracts the two fitted equations to obtain a general expression for the vertical distance between the two fitted lines as a function of house size, this should be considered an essentially correct approach for component 2 of section 4. The resulting expression is $15.580 + 0.006 \cdot (\text{size})$.
- If the student uses a house size outside the range of the data to compute the difference in predicted price, this can only be considered correct if the student appeals to the fact that the slopes of the sample regression lines are not significantly different.

AP® STATISTICS
2010 SCORING GUIDELINES (Form B)

Question 6 (continued)

Each essentially correct (E) section counts as 1 point. Each partially correct (P) section counts as ½ point.

4 **Complete Response**

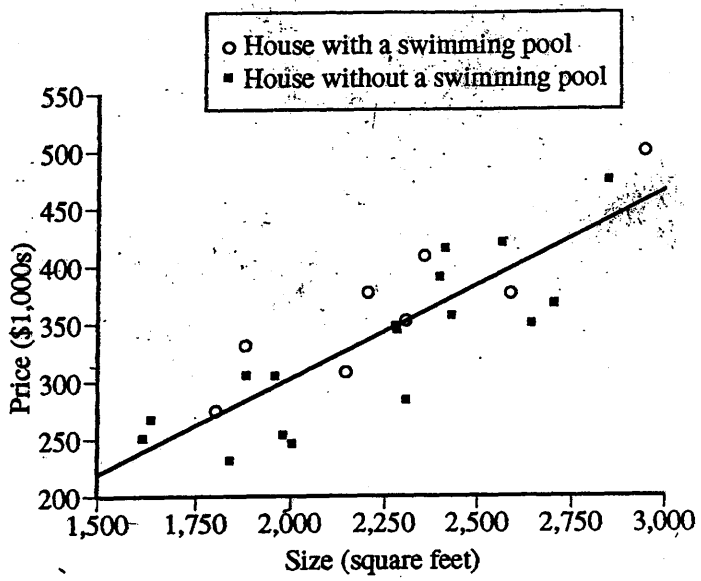
3 **Substantial Response**

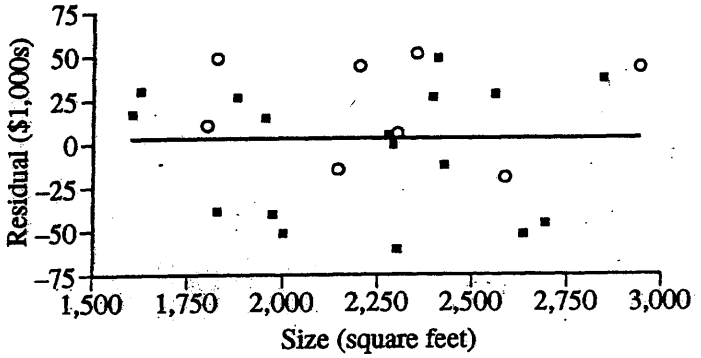
2 **Developing Response**

1 **Minimal Response**

If a response is between two scores (for example, 2½ points), use a holistic approach to determine whether to score up or down, depending on the overall strength of the response and communication. In deciding whether to score up or down, pay particular attention to the response to the investigative part of the question (section 4).

6A1





Linear Fit

Price = –28.144 + 0.165 Size

Summary of Fit

RSquare 0.722

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	–28.144	48.259	–0.58	0.5654
Size	0.165	0.0213	7.72	<.0001

(a) Interpret the slope of the least squares regression line in the context of the study.

Slope – with each additional square foot of size of house, the price of this house increases on average by 0.165 thousands of dollars

GO ON TO THE NEXT PAGE.

6A2

(b) The second house in the table has a residual of 49. Interpret this residual value in the context of the study.

$\hat{y} - \bar{y}$ - residual means that real value of price of the house differs (is bigger) than predicted By linear regression model value on 49 thousands of dollars

The real estate agent is interested in investigating the effect of having a swimming pool on the price of a house.

(c) Use the residuals from all 25 houses to estimate how much greater the price for a house with a swimming pool would be, on average, than the price for a house of the same size without a swimming pool.

average of residuals with pool ($\bar{x}_1$) = price of house. $\frac{149}{8} = 18,625$

average of residuals - // - without pool ($\bar{x}_2$) = $-\frac{150}{17} = -8,8235$

$(\bar{x}_1 - \bar{x}_2) = 18,625 - (-8,8255) = 27,4485$ thousands of dollars

GO ON TO THE NEXT PAGE.

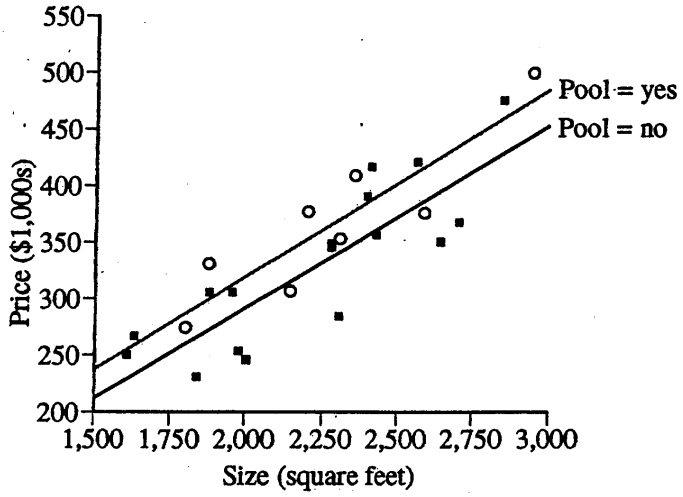
6A3

To further investigate the effect of having a swimming pool on the price of a house, the real estate agent creates two regression models, one for houses with a swimming pool and one for houses without a swimming pool. Regression output for these two models is shown below.

Linear Fit (Pool = yes)
Price = –11.602 + 0.166 size

Linear Fit (Pool = no)
Price = –27.382 + 0.160 size

o House with a swimming pool
▪ House without a swimming pool



(d) The conditions for inference have been checked and verified, and a 95 percent confidence interval for the true difference in the two slopes is $(-0.099, 0.110)$. Based on this interval, is there a significant difference in the two slopes? Explain your answer.

No, the difference between the two slopes is not significant, because the confidence interval for this difference includes zero.

GO ON TO THE NEXT PAGE.

6A4

(e) Use the regression model for houses with a swimming pool and the regression model for houses without a swimming pool to estimate how much greater the price for a house with a swimming pool would be than the price for a house of the same size without a swimming pool. How does this estimate compare with your result from part (c) ?

Let take for estimation size = 2000

$\hat{y}_{\text{pool yes}}(2000) = 320,398$

$\hat{y}_{\text{pool non}}(2000) = 292,618$

$\hat{y}_{\text{pool yes}} - \hat{y}_{\text{pool non}} = 27,78$ \$1000

How much the price would be greater

estimate in part e > estimate in part c

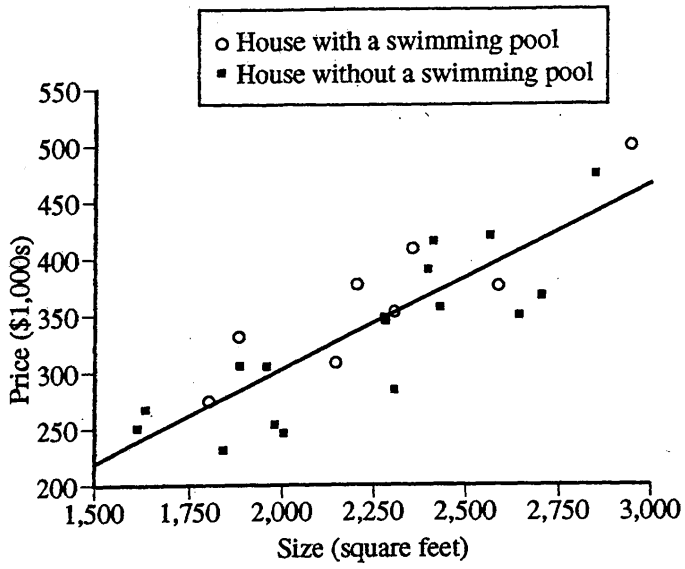
STOP

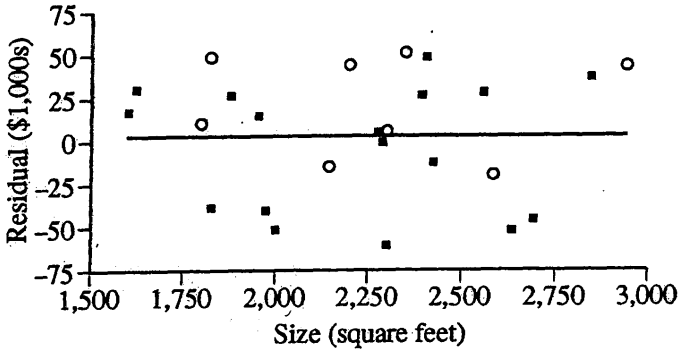
END OF EXAM

THE FOLLOWING INSTRUCTIONS APPLY TO THE COVERS OF THE SECTION II BOOKLET.

- **MAKE SURE YOU HAVE COMPLETED THE IDENTIFICATION INFORMATION AS REQUESTED ON THE FRONT AND BACK COVERS OF THE SECTION II BOOKLET.**
- **CHECK TO SEE THAT YOUR AP NUMBER LABEL APPEARS IN THE BOX(ES) ON THE COVER(S).**
- **MAKE SURE YOU HAVE USED THE SAME SET OF AP NUMBER LABELS ON ALL AP EXAMS YOU HAVE TAKEN THIS YEAR.**

6B1





Linear Fit

Price = –28.144 + 0.165 Size

Summary of Fit

RSquare 0.722

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	–28.144	48.259	–0.58	0.5654
Size	0.165	0.0213	7.72	<.0001

(a) Interpret the slope of the least squares regression line in the context of the study.

the slope of the least squares regression line is 0.165, which indicates that, on average, the price of houses in that part of city is expected to increase by 165 dalars for each additional square feet of the house.

GO ON TO THE NEXT PAGE.

(b) The second house in the table has a residual of 49. Interpret this residual value in the context of the study.

The residual = actual − predicted, so the residual of 49 means that the actual price of the second house is $49,000 more expensive than the predicted price calculated by the linear model. (actual price = $-28.144+0.165(1875)+49 = 330$ thousand dollars)

The real estate agent is interested in investigating the effect of having a swimming pool on the price of a house.

(c) Use the residuals from all 25 houses to estimate how much greater the price for a house with a swimming pool would be, on average, than the price for a house of the same size without a swimming pool.

because $\bar{x}_{\text{residuals for houses with pool}} = \frac{6+49-18+42+11+50-23+42}{8} = 18.625$

$\bar{x}_{\text{residuals for houses without pool}}$

$= \frac{13+26-45+22+10-46-57+1-2-69+23+44-19+26-58-52+33}{17}$

$= -8.82$

So the price for a house with a swimming pool would be, on average, $18.625-(-8.82) = 27.445$ thousand dollars than the price for a house of the same size without a swimming pool.

GO ON TO THE NEXT PAGE.

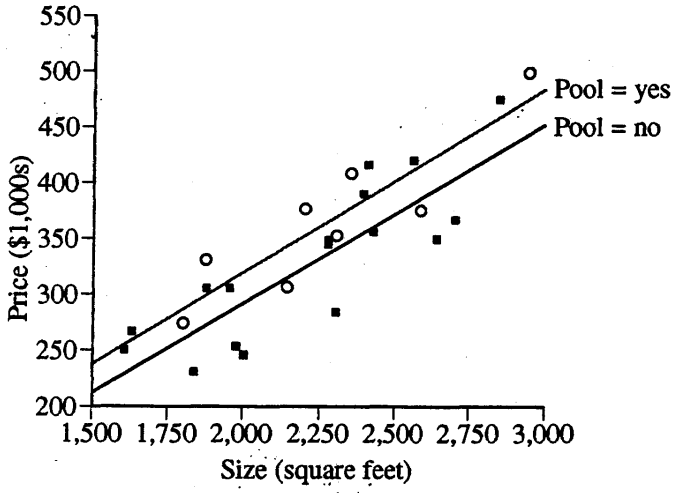
6B3

To further investigate the effect of having a swimming pool on the price of a house, the real estate agent creates two regression models, one for houses with a swimming pool and one for houses without a swimming pool. Regression output for these two models is shown below.

Linear Fit (Pool = yes)
Price = −11.602 + 0.166 size

Linear Fit (Pool = no)
Price = −27.382 + 0.160 size

o House with a swimming pool
▪ House without a swimming pool



(d) The conditions for inference have been checked and verified, and a 95 percent confidence interval for the true difference in the two slopes is (−0.099, 0.110). Based on this interval, is there a significant difference in the two slopes? Explain your answer.

~~the critical~~ because the interval is (−0.099, 0.110), the ~~mean difference~~ predicted

$\mu_d = .1045$

~~For a test for differ~~

For a two-sample t-test for differences, we check the

$H_0: \mu_1 - \mu_2 = 0$

$H_a: \mu_1 - \mu_2 \neq 0$

since

GO ON TO THE NEXT PAGE.

6B4

(e) Use the regression model for houses with a swimming pool and the regression model for houses without a swimming pool to estimate how much greater the price for a house with a swimming pool would be than the price for a house of the same size without a swimming pool. How does this estimate compare with your result from part (c) ?

Price = -11.602 + 0.166 size ①

Price = -27.382 + 0.160 size ②.

① - ② ⇒ : Price difference = 15.78 + 0.006 size.

the result is similar to the result from part (c), for example, if the size of the house is 1875, the Price difference is equal to 15.78 + 0.006(1875) = 27.03, very close to 27.445 resulted from part (c).

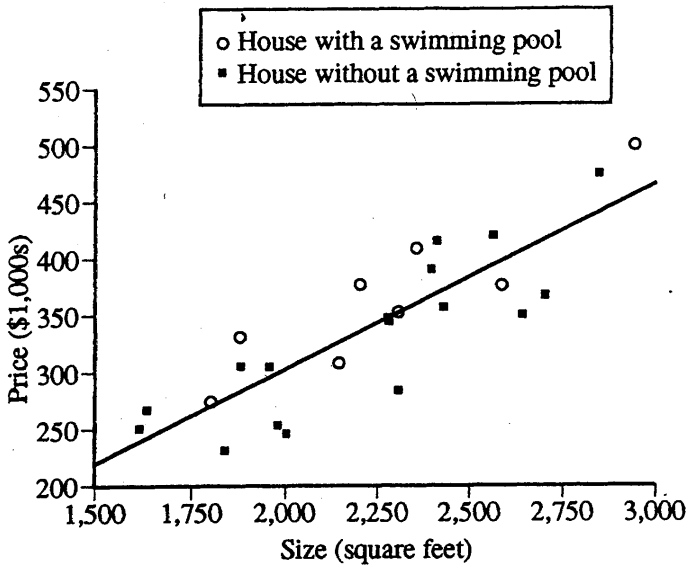
STOP

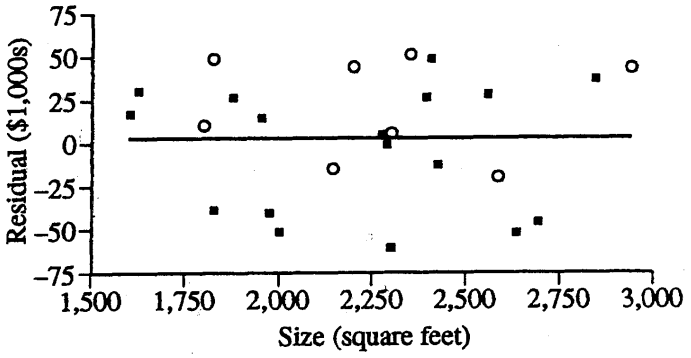
END OF EXAM

THE FOLLOWING INSTRUCTIONS APPLY TO THE COVERS OF THE SECTION II BOOKLET.

- **MAKE SURE YOU HAVE COMPLETED THE IDENTIFICATION INFORMATION AS REQUESTED ON THE FRONT AND BACK COVERS OF THE SECTION II BOOKLET.**
- **CHECK TO SEE THAT YOUR AP NUMBER LABEL APPEARS IN THE BOX(ES) ON THE COVER(S).**
- **MAKE SURE YOU HAVE USED THE SAME SET OF AP NUMBER LABELS ON ALL AP EXAMS YOU HAVE TAKEN THIS YEAR.**

6C1





Linear Fit

Price = –28.144 + 0.165 Size

Summary of Fit

RSquare 0.722

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	–28.144	48.259	–0.58	0.5654
Size	0.165	0.0213	7.72	<.0001

(a) Interpret the slope of the least squares regression line in the context of the study.

As the size of the house increases by 0.165 square feet, the price of the house increases by 28.144 thousand dollars.

GO ON TO THE NEXT PAGE.

6 of 12

(b) The second house in the table has a residual of 49. Interpret this residual value in the context of the study.

The second house with a price of $330,000 and 1,875 sq. feet (and a pool) deviates from the least squared Regression Line by ~~49,000~~ a y value of 49, or $49,100.

The real estate agent is interested in investigating the effect of having a swimming pool on the price of a house.

(c) Use the residuals from all 25 houses to estimate how much greater the price for a house with a swimming pool would be, on average, than the price for a house of the same size without a swimming pool.

$\frac{149}{8} = 18.625$ ← [avg. of residual for houses w/ a pool

$\frac{-150}{17} = -8.8235$ ← avg. of residuals for houses w/ no pools

$18.625 + -8.8235 = 9.8015$

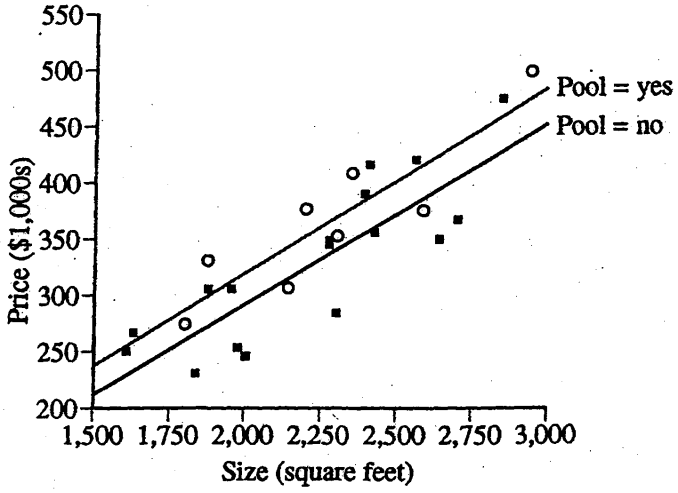
On average, a house with a pool ~~cost~~ is $9801.50 greater in price than a house w/ out a pool, according to this sample.

To further investigate the effect of having a swimming pool on the price of a house, the real estate agent creates two regression models, one for houses with a swimming pool and one for houses without a swimming pool. Regression output for these two models is shown below.

Linear Fit (Pool = yes)
Price = –11.602 + 0.166 size

Linear Fit (Pool = no)
Price = –27.382 + 0.160 size

o House with a swimming pool
▪ House without a swimming pool



(d) The conditions for inference have been checked and verified, and a 95 percent confidence interval for the true difference in the two slopes is (–0.099, 0.110). Based on this interval, is there a significant difference in the two slopes? Explain your answer.

Based on the interval (-0.099, 0.110) for the true difference of the two slopes, there is not a significant difference in the two slopes because the confidence interval includes 0, meaning it is possible that there is NO difference between the slopes, and if there is a difference, it is extremly small (almost negligable).

GO ON TO THE NEXT PAGE.

6 C4

(e) Use the regression model for houses with a swimming pool and the regression model for houses without a swimming pool to estimate how much greater the price for a house with a swimming pool would be than the price for a house of the same size without a swimming pool. How does this estimate compare with your result from part (c) ?

House w/ a pool @ 2,000 sq feet

price = -11.60 + 0.166(2000)

[price = 320,400]

House w/out a pool @ 2,000 sq feet

price = -27.382 + 0.160(2000)

[price = 292,618]

320,400 − 292,618 = $27,782 ← Difference btwn same size houses, with and without pools

This estimate is much larger than my estimate of the price difference between the two houses from part c) which was ≈ $1800. (approx. 3 times as large!)

STOP

END OF EXAM

THE FOLLOWING INSTRUCTIONS APPLY TO THE COVERS OF THE SECTION II BOOKLET.

- **MAKE SURE YOU HAVE COMPLETED THE IDENTIFICATION INFORMATION AS REQUESTED ON THE FRONT AND BACK COVERS OF THE SECTION II BOOKLET.**
- **CHECK TO SEE THAT YOUR AP NUMBER LABEL APPEARS IN THE BOX(ES) ON THE COVER(S).**
- **MAKE SURE YOU HAVE USED THE SAME SET OF AP NUMBER LABELS ON ALL AP EXAMS YOU HAVE TAKEN THIS YEAR.**

AP® STATISTICS
2010 SCORING COMMENTARY (Form B)

Question 6

Sample: 6A
Score: 4

Part (a) of this response includes a correct interpretation of the slope, in context, so section 1, consisting of part (a), was scored as essentially correct. Section 2, consisting of part (b), was also scored as essentially correct because the residual of 49 is correctly interpreted in context. In part (c) residual averages are computed separately for houses with pools and for houses without pools, and the difference in the residual averages is correctly calculated; thus section 3, consisting of part (c), was scored as essentially correct. In part (d) the response correctly states that there is no significant difference in the slopes and provides appropriate justification based on the given confidence interval. In part (e) a house size of 2,000 square feet, which is within the range of house sizes in the sample, is chosen, and the difference in price for a house of this size with a pool and a house of this size without a pool is computed. This estimate is then compared with the estimate in part (c). Section 4, consisting of parts (d) and (e), therefore includes all three components needed to receive a score of essentially correct. The entire answer, based on all four sections, was judged a complete response and earned a score of 4.

Sample: 6B
Score: 3

Correct interpretations, in context, of the slope and the provided residual are given in parts (a) and (b), so section 1, consisting of part (a), and section 2, consisting of part (b), were each scored as essentially correct. In part (c) residual averages are computed separately for houses with pools and for houses without pools, and the difference in the residual averages is correctly calculated and reported as the estimate of the impact of a pool. Section 3, consisting of part (c), was scored as essentially correct. In part (d) the response does not use the given confidence interval to reach a decision and begins to set up an incorrect hypothesis test. In part (e) the two given regression equations are subtracted to obtain an expression that describes the difference in price for houses with and without a pool as a function of size. This expression is used to correctly estimate the price difference for a house of 1,875 square feet, and the resulting estimate is correctly compared with the estimate in part (c). Thus component (1) of section 4, consisting of part (d), is not correct, but components (2) and (3) of this section, comprising part (e), are correct, so section 4 was scored as partially correct. The entire answer, based on all four sections, could have been scored up to a 4 or down to a 3. Based on the incorrect response in part (d), this response was judged to be substantial rather than complete and so earned a score of 3.

Sample: 6C
Score: 2

In part (a) the slope is not interpreted correctly, so section 1, consisting of part (a), was scored as incorrect. In part (b) the residual is interpreted as the amount by which the actual price differs from the predicted price, but the direction of the difference is not specified; thus section 2, consisting of part (b), received a score of partially correct. In part (c) the two residual averages are added rather than subtracted to produce the estimate, so section 3, consisting of part (c), was scored as partially correct. In part (d) the response correctly states that there is no significant difference in the slopes and provides appropriate justification based on the given confidence interval. In part (e) a house size of 2,000 square feet, which is within the range of house sizes in the sample, is chosen, and the difference in price for a house of this size with a pool and a house of this size without a pool is computed. This estimate is then compared with the estimate in part (c). Section 4, consisting of parts (d) and (e), therefore includes all three required components and received a score of essentially correct. The entire answer, based on all four sections, was judged a developing response and earned a score of 2.