NATIONAL MATH + SCIENCE INITIATIVE

AP Statistics

Linear Regression

Presenter Notes

2016-2017 EDITION

We have intentionally included more material than can be covered in most Student Study Sessions to account for groups that are able to answer the questions at a faster rate. Use your own judgment, based on the group of students, to determine the order and selection of questions to work in the session. Be sure to include a variety of types of questions (multiple choice and free response) in the time allotted.

Student Study Session - Presenter Notes – Descriptive Statistics

Thank you for agreeing to present at one of NMSI's Saturday Study Sessions. We are grateful you are sharing your time and expertise with our students. Saturday mornings can be a "tough sell" for students, so we encourage you to incorporate strategies and techniques to encourage student movement and engagement. Suggestions for different presentation options are included in this document. If you have any questions about the content or about presentation strategies, please contact Mathematics Director Charla Holzbog at cholzbog@nms.org or AP Statistics Content Specialist Penny Smeltzer at psmeltzer@nms.org.

The material provided contains many released AP multiple choice and free response questions as well as some AP-like questions that we have created. The goal for the session is to let the students experience a variety of both types of questions to gain insight on how the topic will be presented on the AP exam. It is also beneficial for the students to hear a voice other than their teacher in order to help clarify their understanding of the concepts.

Suggestions for presenting:
The vast majority of the study sessions are on Saturday and students and teachers are coming to be WOWed! We want activities to engage the students as well as prepare them for the AP Exam. The following presenter notes include pacing suggestions (you only have 50 minutes!), solutions, and recommended engagement strategies.

Suggestions on how to prepare:

- The notes/summaries on the last page(s) are for reference. We want the students' time during the session to be focused on the questions as much as possible and not taking or reading the notes. As the questions are presented during the session, you may wish to refer the students back to those pages as needed. It is not our intent for the sessions to begin with a lecture over these pages.
- As you prepare, work through the questions in the packet noting the level of difficulty and topic or skill required for the questions.
- Design a plan for what questions you would like to cover with the group depending on their level of expertise. Some groups will be ready for the tougher questions while other groups will need more guidance and practice on the easier ones. Create an easy, medium, and hard listing of the questions prior to the session. This will allow you to adjust on the fly as you get to know the groups. In most instances, there will **not** be enough time to cover all the questions in the packet. Use your judgement on the amount of questions to cover based on the students' interactions. Remember to include both multiple choice and free response type questions. Discussions on test taking strategies and scoring of the free response questions are always great to include during the day.
- The concepts should have been previously taught; however, be prepared to "teach" the topic if you find out the students have not covered the concept prior in class. In sessions where multiple schools come together, you might have a mixture of students with and without prior knowledge on the topic. You will have to use your best judgement in this situation.
- Consider working through some free response questions before the multiple choice questions, or flipping back and forth between the two types of questions. Sometimes, if free response questions are saved for the last part of the session, it is possible students only get practice with one or two of them and most students need additional practice with free response questions.

The Plan

You only have 50-55 minutes, so watch your timing and give students the maximum amount of practice they can handle in the short amount of time.

In a Nutshell

1. 20-25 minutes: PowerPoint on Swine problem
2. 15-20 minutes: Multiple choice questions
3. 15-20 minutes: Addition Free Response work

Detailed Plan

1. As students walk in or at the starting time, have them load the swine data into their calculators. Begin using the PowerPoint to assist you in student discussions over the questions. It is important to note that the PowerPoint is NOT to be used as a click and note taking aide. Instead, have students discuss with a partner the appropriate response before clicking to an answer. There are a few additional slides in the PowerPoint that are not on the student sheet to illicit more thinking about the data.

2. Have students work in groups of 3 or 4 to complete the multiple choice questions. Encourage discussions among the student groups. Then go through the answers using the PowerPoint. It might be a good idea to have students stand up with their answer choice (to stretch a bit and let you know how they did). For instance, for question 1, stand if you chose A. Stand if you chose B, etc. You can monitor the confidence shown as students stand.

Note: Question 3 causes confusion with many students. They tend to miss the word "increase". They may benefit from you showing them that the correct answer may be found by simply multiplying the value 5 by the slope OR by finding the predicted value of 2 temperatures using x values that differ by 5 and then subtracting the results. (Substitute 20 and 25 in for x, find the predicted y values, and subtract.)

3. Choose another Free Response question for students to try on their own.

Answers - Multiple Choice

1. A (AP Course Description Q1)
   Point A has the largest vertical distance from the LSRL.

2. C (AP Style Question)
   Point C has the most leverage of the points labeled and improves the strength of the correlation coefficient since it fits the linear trend well and is an outlier in the *x* direction.

3. B (1997 Q28) $3.3(5) = 16.5$

4. A (AP Style Question)
   $\hat{y} = 2.3 + 0.37(4) = 3.78$; therefore the residual is $\text{resid} = 7 - 3.78 = 3.22$

5. E (AP Style Question)
   Since correlation does not depend on units, the correlation stays the same. $r = 0.75$

6. C (AP Style Question)
   $\hat{y} = 16.6 + 0.65(20) = 29.6$; since the residual is -4.6 pounds, $-4.6 = y - 29.6$ which yields $y = 25$ pounds

7. E (AP Style Question)
   Since the response variable is the number of degrees of motion, the slope of the LSRL represents the predicted change in degrees of motion for each unit change in the years the player has played professional baseball.

8. C (AP Style Question)
   $r^2 = .908$ which means that 90.8% of the variation in home price is explained by the linear relationship with the size of the home.

9. D (NMSI question)

10. E (AP Style Question)
    Since the residual plot for Regression II shows random scatter, Regression II provides an appropriate linear fit. Since there is a linear relationship between $\ln x$ and $\ln y$, there is a non-linear relationship between *x* and *y*.

Answers Additional Free Response
2005B Q3 (extended)
Solution

Part (a):
Yes, the linear model is appropriate for these data. The scatterplot shows a strong, positive, linear association between the number of railcars and fuel consumption, and the residual plot shows a reasonably random scatter of points above and below zero.

Part (b):
According to the regression output, fuel consumption will increase by 2.209 units for each additional railcar. Since the fuel consumption cost is \$42 per unit, the average cost of fuel per mile will increase by approximately $(\$42)(2.209) = \$92.78$ for each railcar that is added to the train.

Part (c):
The regression output indicates that $r^2 = 97.1\%$ or 0.971. Thus, 97.1% of the variation in the fuel consumption values is explained by using the linear regression model with number of railcars as the explanatory variable.

Part (d):
No, the data set does not contain any information about fuel consumption for any trains with less than 20 cars. Using the regression model to predict the fuel consumption for a train with 5 railcars, known as extrapolation, is not reasonable.

Scoring
Each part is scored as essentially correct (E), partially correct (P), or incorrect (I).

Part (a) is essentially correct (E) if the model is deemed appropriate AND the explanation clearly indicates:
- There is a linear pattern in the scatterplot; OR
- There is no pattern in the residual plot.

Part (a) is partially correct (P) if the:
- Model is deemed appropriate AND the student refers to the scatterplot or residual plot but fails to state the relevant characteristic of the plot; OR
- Student refers to the relevant characteristic of the scatterplot or residual plot without deeming model appropriate.

Part (a) is incorrect (I) if the student:
- States that the model is appropriate without an explanation; OR
- States that the model is inappropriate; OR

Makes a decision based only on numeric values from the computer output.

Part (b) is essentially correct (E) if the point estimate for the slope (2.209 or 2.21) and the fuel consumption cost per unit ($42) are used to calculate the correct point estimate ($\$92.778 \approx \$92.78$ or $92.82).

Part (b) is partially correct (P) if only the point estimate for the slope (2.209 or 2.21) is stated with a supporting calculation or interpretation.

Part (c) is essentially correct (E) if the student states:

- 97.1% of the variation in fuel consumption is explained by the linear regression model; OR
- 97.1% of the variation in fuel consumption is explained by the number of railcars.

Part (c) is partially correct (P) if the student makes one of the above statements using $\text{R-Sq(adj)} = 96.8\%$.

Part (d) is essentially correct (E) if the student states that this is unreasonable due to extrapolation.

Part (d) is partially correct (P) if the student states this is:

- Unreasonable but provides a weak explanation; OR
- Reasonable even though it is considered a slight extrapolation.

Note: Any answer appearing without supporting work is scored as incorrect (I).

Each essentially correct (E) response counts as 1 point, each partially correct (P) response counts as $\frac{1}{2}$ point.

4 Complete Response
3 Substantial Response
2 Developing Response
1 Minimal Response

Note: If a response is in between two scores (for example, 2.5 points), use a holistic approach to determine whether to score up or down depending on the strength of the response and communication.

Solution to additional parts

(e) $r = 0.985$. This value of r, indicates a strong, positive, linear relationship between the number of railcars and the fuel consumption in units per mile.

(f) The predicted fuel consumption for the train with 40 railcars is

$\hat{y} = 8.2689 + 2.209(40)$

$= 96.6289 \frac{\text{units}}{\text{mile}}$

$\text{resid} = 99 - 96.6289 = 2.3711 \frac{\text{units}}{\text{mile}}$. This means that the regression equation under predicted the fuel consumption by $2.3711 \frac{\text{units}}{\text{mile}}$ for the train with 40 railcars.

(g) The predicted fuel consumption for a train with 33 cars is

$\hat{y} = 8.2689 + 2.209(33)$

$= 81.1659 \frac{\text{units}}{\text{mile}}$

So the total cost would be $\text{cost} = 81.1659 \frac{\text{units}}{\text{mile}} (775 \text{ miles}) \left(\frac{\$42}{\text{unit}} \right) = \$2,641,950$

(h) Since this point does not fit the trend of the data, this point would cause the correlation coefficient to decrease in strength.

(i)Since this point is close to $\bar{x}$, the slope will likely not change very much.

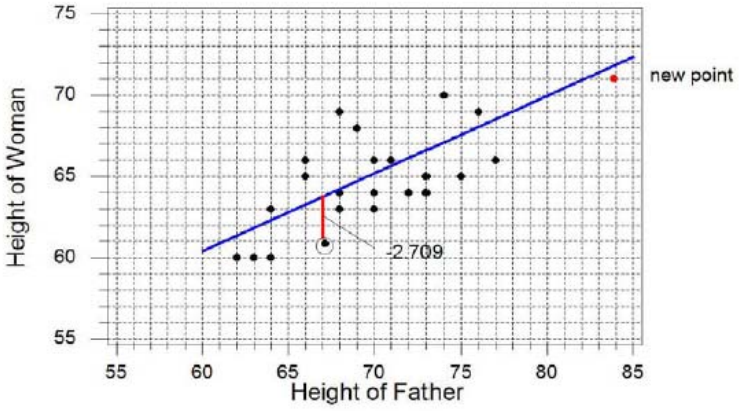
2007 B Q4

Intent of Question

The goals of this question are to assess a student's ability to: (1) plot a least squares regress line; (2) examine a residual; and (3) discuss the effect of an additional observation on an estimated correlation coefficient and on the least squares estimate of the slope of a line.

Solution

Parts (a) and (b):



When $x = 67$, $\hat{y} = 35.1 + 0.427(67) = 63.709$ and the residual $= y - \hat{y} = 61 - 63.709 = -2.709$.

Part (c):

See the new point indicated in the plot above. The slope would remain about the same since the new point is consistent with the linear pattern in the original plot (i.e., close to the line).

The correlation coefficient would increase. We know that $b = r\frac{s_y}{s_x}$. The added point will increase s_x more than it will increase s_y so $\frac{s_y}{s_x}$ will be less than 1. If the slope is to stay the same, *r* must increase.

OR

This point fits the pattern well and has an *x* value that is far from $\bar{x}$.
(This is what most students wrote.)

Scoring

This problem is scored in 4 sections. Section 1 consists of the graphical parts of (a) and (b) together. Section 2 consists of the numerical parts of (b). Section 3 consists of the first part of (c). Section 4 consists of the second part of (c).

Each section is scored as either essentially correct (E), partially correct (P), or incorrect (I).

Section 1 (graphical parts of a and b) is essentially correct (E) if:
1. the regression line is drawn correctly on the scatterplot;
2. the point $(67, 61)$ is circled and the vertical segment corresponding to the residual is drawn on the scatterplot.

Section 1 is partially correct (P) if the response includes one of the above two elements.

Section 2 (numerical part of b) is essentially correct (E) if the residual is correctly computed as -2.709;
OR
the response states that the residual was approximated using the graph, a reasonable value for the residual is given, and the sign of the residual is correct.

Section 2 is partially correct (P) if the magnitude of the residual is correct but the sign is wrong.

Section 3 (first part of (c)) is essentially correct (E) if it:
1. states the slope will remain about the same (or change slightly);
2. provides an explanation based on the new point fitting the pattern of the original plot.

Section 3 is partially correct (P) if it states that the slope will be about the same, but the explanation is missing or incorrect.

NOTE: If the line is drawn incorrectly in part (a), and the answer to this part is consistent with the line drawn, section 3 is essentially correct (E).

Section 4 (second part of (c)) is essentially correct (E) if it:

1. states that the value of the correlation coefficient will increase;
2. provides an explanation based on the relative changes in s_x and s_y

   OR

   based on the fact that the new point fits the pattern AND is far out in the *x* direction,

   OR

   because the linear pattern is stronger.

Section 4 is partially correct (P) if it states that the value of the correlation coefficient will increase, but the explanation is missing or incorrect.

NOTE: If the response just says that the correlation coefficient will increase because the point is close to the line, section 4 is partially correct.

4 Complete Response

All four sections essentially correct

3 Substantial Response

Three sections essentially correct and no sections partially correct

OR

Two sections essentially correct and two sections partially correct

2 Developing Response

Two sections essentially correct and no sections partially correct

OR

One section essentially correct and two sections partially correct

OR

Four parts partially correct

1 Minimal Response

One section essentially correct and no sections partially correct

OR

No sections essentially correct and two sections partially correct

If a response is between two scores (for example, $2\frac{1}{2}$ points), use a holistic approach to determine whether to score up or down depending on the strength of the response and communication.

