Lab 5a: Regression

Stat 131A, Fall 2018

Learning Objectives:

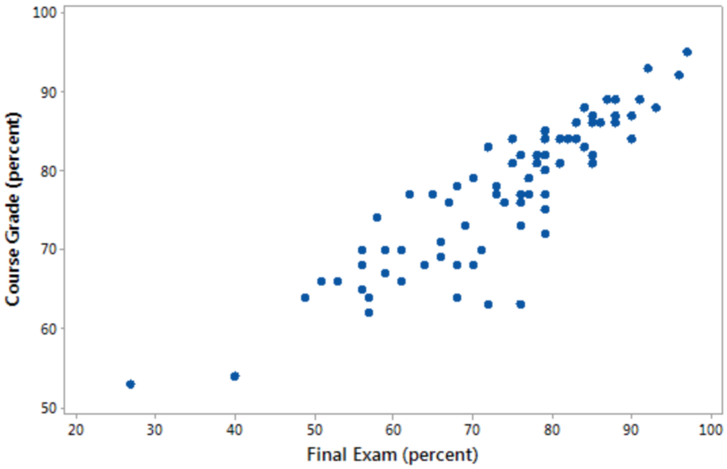
- Use a correlation coefficient to describe the direction and strength of a linear relationship.
- Distinguish between association and causation.
- Identify lurking variables that may explain an observed relationship.

General Instructions

- Write your solutions in an `Rmd` (R markdown) file.
- Name this file as `lab05a-first-last.Rmd`, where `first` and `last` are your first and last names (e.g. `lab05a-gaston-sanchez.Rmd`).
- Knit your `Rmd` file as an html document (default option).
- Submit your `Rmd` and `html` files to bCourses, in the corresponding lab assignment.

Problem 1

For the data set of this scatterplot, the least squares regression line is $Y = 31.72 + 0.62X$, where X represents the final exam score as a percent and Y represents the predicted course grade as a percent.



Notice that the regression line for this data set has slope 0.62. What is the most precise and accurate interpretation of the slope?

a. Students with higher final exam scores tend to have higher course grades.

b. For every 1% increase in a student's final exam score, we expect to see a 0.62% decrease in the course grade.

c. For every 1% increase in a student's final exam score, we expect to see a 0.62% increase in the course grade.

d. For every 1% increase in a student's final exam score, we expect to see a 62% increase in the course grade.

Problem 2

Using the equation of the regression line given above, what is the predicted course grade of a student who earns 85% on the final exam?

Problem 3

There is a fairly strong negative linear relationship between years of driving experience (X) and the dollar amount paid for car insurance each month (Y). The least-squares regression line is:

Predicted monthly car insurance premium $= 97 - 1.45 \times$ years of driving experience

Interpret the slope and intercept in the context of the data.

Problem 4

Suppose that we have the least squares regression line:

$$Y = 32 + 0.6X$$

and that the sum of the squares of the errors (SSE) for this line is 1,373. What can we conclude about the sum of the squares of the errors (SSE) for this data set and the line $Y = 32 + 1.6X$?

The sum of the squares of the errors (SSE) for the line $Y = 32 + 1.6X$ will be _______ 1,373:

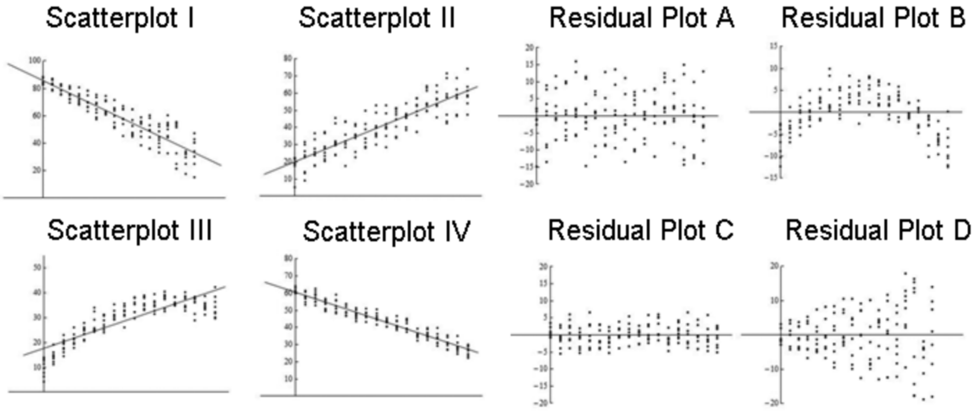
a. equal to
b. greater than
c. less than

Problem 5

An admissions officer is trying to choose between two methods for predicting first year scores. One method has an R.M.S. error of 7. Other things being equal, which should he choose? Why?

Problem 6

Below are four scatterplots with regression lines shown and four corresponding residual plots. Match the scatterplot to its residual plot.



a. Scatterplot I matches

b. Scatterplot II matches

c. Scatterplot III matches

d. Scatterplot IV matches

Problem 7

Use the scatter diagrams and residual plots of the previous question. Match each description with the appropriate residual plot.

a. The residuals show a curved pattern.

b. The pattern in the residual plot suggests that predictions based on the linear regression line will result in greater error as we move from left to right through the range of the explanatory variable.

c. Residuals are randomly scattered with no distinctive pattern.

d. Based on the residual plots, for which two scatterplots does the regression line fail to capture important aspects of the relationship between the variables?

Problem 8

A law school finds the following relationship between LSAT scores and first-year scores:

- average LSAT score = 165, $SD = 5$
- average first-year score = 65, $SD = 10$
- $r = 0.6$

a) The admissions officer uses the regression line to predict first-year scores from LSAT scores. The r.m.s. error of the line is __________. Options:

i. 5

ii. 10

iii. $\sqrt{1 - 0.6^2} \times 5$

iv. $\sqrt{1 - 0.6^2} \times 10$

b) One of the students is chosen at random; you have to guess his first-year score, without being told his LSAT score. How would you do this?

c) Your r.m.s. error would be _________.

i. 5

ii. 10

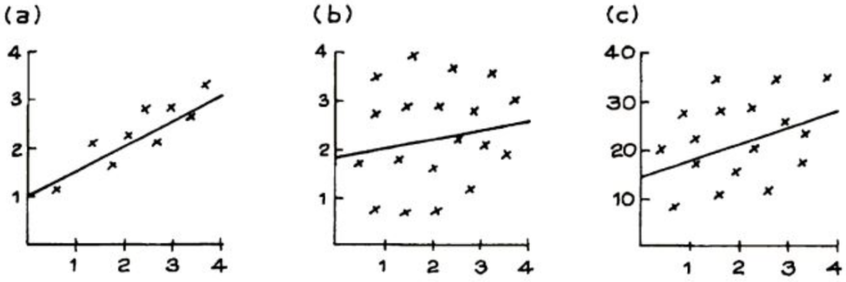
iii. $\sqrt{1 - 0.6^2} \times 5$

iv. $\sqrt{1 - 0.6^2} \times 10$

d) Repeat parts (b) and and (c) if you are allowed to use his LSAT score.

Problem 9

Below are three scatter diagrams. The regression line has been drawn across each one, by eye. In each case, guess whether the r.m.s. error is 0.2, or 1, or 5.



Problem 10

Suppose that we want to examine the relationship between high school GPA and college GPA. We collect data from students at a local college. The linear regression predicted college GPA = 1.07 + 0.62 * high school GPA.

One student has a high school GPA of 3.00 and a college GPA of 3.15. What is the residual for this student?

a. 0.22

b. -0.22

c. 0.15

d. -0.15

Problem 11

Researchers conduct a study of obesity in children. They measure body mass index (BMI), which is a measure of weight relative to height. High BMI is an indication of obesity.

Data from a study published in the Journal of the American Dietetic Association shows a fairly strong positive linear association between mother's BMI and daughter's BMI ($r = 0.506$). This means that obese mothers tend to have obese daughters.

Which of the following are valid interpretations of r^2?

a. The regression line will correctly predict the daughter's BMI 25% of the time.

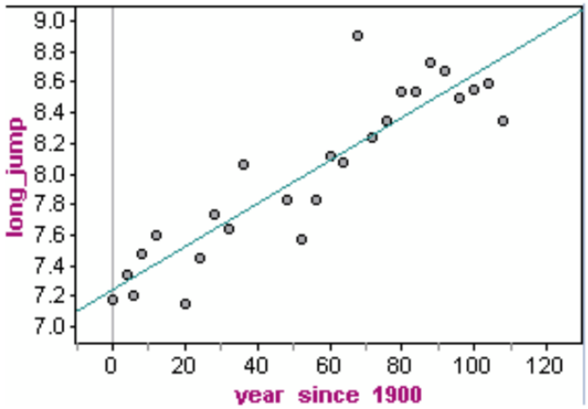
b. 25% of the daughter BMI measurements are explained by the mother BMI measurements.

c. The regression line based on mother BMI measurements accounts for 25% of the total variation in observed daughter BMI measurements.

d. 75% of the total variation in observed daughter BMI measurements is due to individual variation in the daughters that cannot be explained by the linear relationship with mother BMI.

Problem 12

The scatterplot below shows Olympic gold medal performances in the long jump from 1900 to 1988. The long jump is measured in meters.



a. The Olympics were not held in 1940 because of World War II. If the Olympics had happened in 1940, how could you estimate the gold medal winning distance in the long jump for that year?

b. For the regression line predicted long jump = 7.24 + 0.014 (year since 1900), what does the 7.24 tell us?

   i) 7.24 meters is the predicted value for the long jump in 1900.

   ii) 7.24 meters is the actual value for the long jump in 1900.

   iii) 7.24 is the minimum value for the long jump from 1900 to 1988.

   iv) 7.24 meters is the predicted increase in the winning long jump distance for each additional year after 1900.

Problem 13

For a linear regression model with square of the correlation r2 and standard error se, which of the following best describes the size of r2 and se that you would most want for your model?

a. r^2 is small and RMSE is small

b. r^2 is big and RMSE is big

c. r^2 is small and RMSE is big

d. r^2 is big and RMSE is small

Problem 14

Which of the following statements is true of a least-squares regression line? Check all that apply.

a. The least-squares regression line is chosen so that the sum of the squares of the residuals is as small as possible.

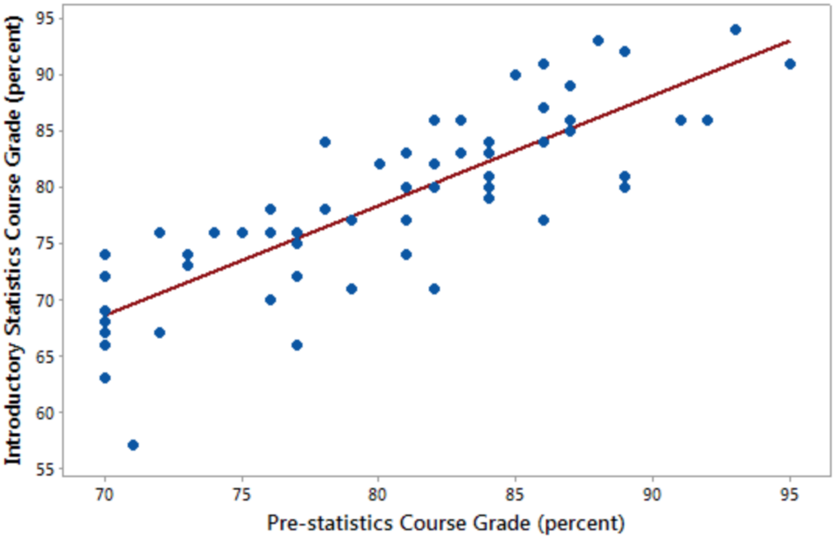
b. The least-squares regression line is the only line with the smallest sum of the squares of the errors.

c. If all the points are on a line, then the sum of the squares of errors is zero.

d. The sum of the squares of the residuals is always equal to r^2.

Problem 15

We recorded the pre-statistics course grade (in percent) and introductory statistics course grade (in percent) for 60 community college students. Then we generated the following scatterplot of the data.



For this linear regression model, $r^2 = 0.70$. What does this mean?

a. Our linear regression model explains 70% of the total variation in the introductory statistics course grade.

b. There will be about 70% of the data along the regression line.

c. The pre-statistics course grade explains 70% of the introductory statistics course grade.

d. Our linear regression model explains 70% of the total variation in the pre-statistics course grade.