Statistics 104

Midterm Examination 1 Solutions

October 27, 2004

1. (10 points) Indicate which of the following statements are true and briefly, for each of the others, show why they are false. You may simply correct the given statement as a way of showing why.

   a) (2 points) You wish to get a measure of the typical speed of vehicles on the interstate highway on which you are driving. So you adjust your speed until the number of vehicles passing you equals the number you are passing. Your current speed is the mean speed of vehicles currently on the interstate.

      False. The current speed is the median.

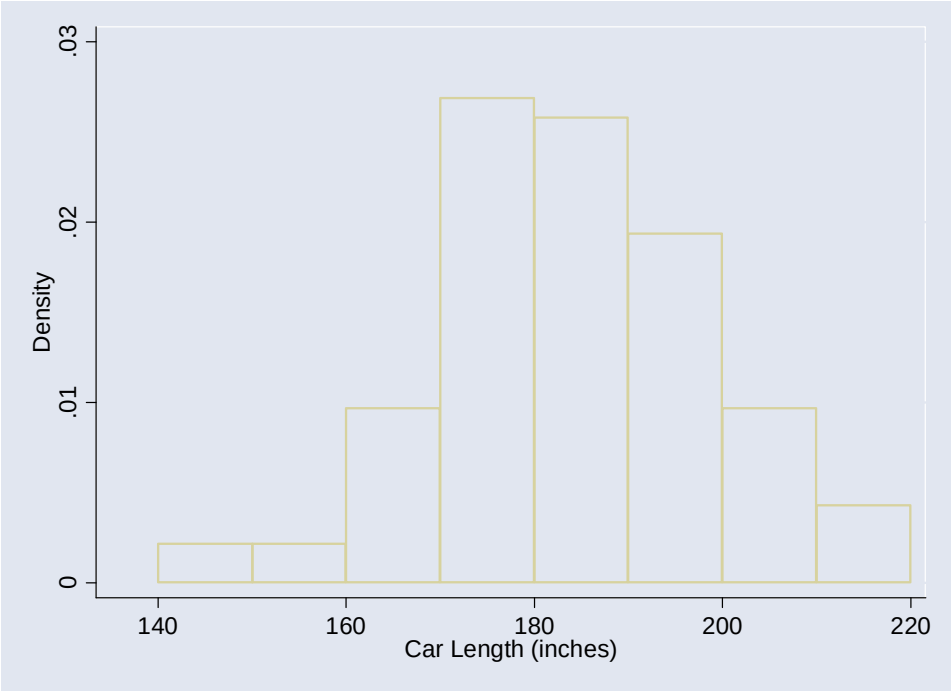
   b) (2 points) The Chapin Social Insight Test measures how accurately a subject appraises other people. Scores on the test are approximately normally distributed with mean 25 and standard deviation 5. The proportion of people who score between 15 and 35 on the test would be about 68%.

      False. There are two possible corrections

      i) The proportion of people who score between 15 and 35 on the test would be about **95**%.

      ii) The proportion of people who score between **20 and 30** on the test would be about **68**%.

c) (2 points) For the data displayed in the following histogram, the median car length is approximately 200 inches.



False. There are two possible corrections

i) the median car length is approximately **180** inches.

ii) the **3rd quartile** of car length is approximately 200 inches.

d) (2 points) In 1846, the Duke of Saxe-Cobourg and Gotha received a letter describing the chest measurements of 5738 recruits for the army. The measurements of chest circumference follow a distribution with a mean of 39.8 inches and a standard deviation of 2.0 inches. If the data was recoded so the measurements were in centimeters, the mean would be 15.67 cm and the standard deviation would be 0.79 cm (1 inch = 2.54 cm)

False. The mean should be 101.092 cm with a standard deviation of 5.08 cm.

e) (2 points) If an event *A* has a probability of 800%, it implies that the event must occur 8 times.

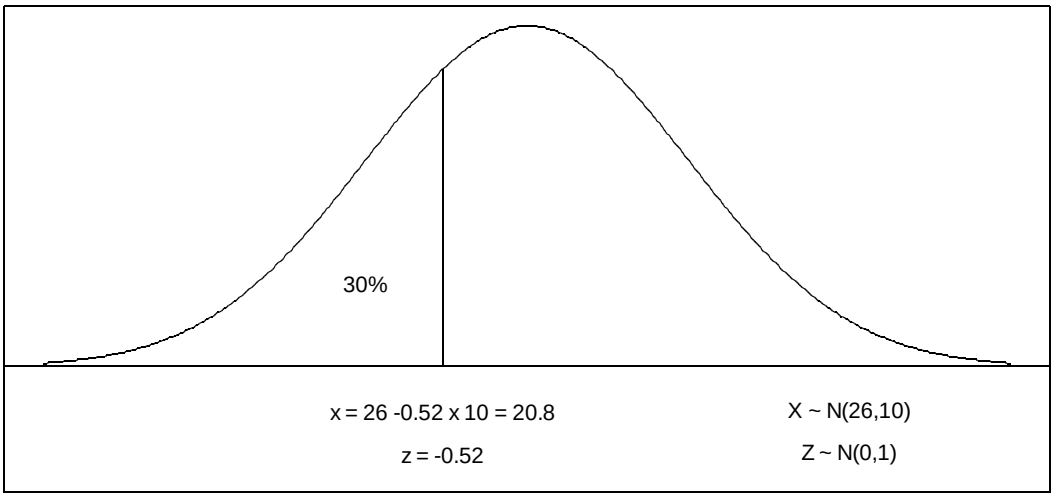
False. The maximum probability of any event is 100%. If an event must occur 8 times, then the probability of the base event must be 100%.

2. (12 points) Data collected as part of the 1980 Wisconsin Restaurant Survey, conducted by the University of Wisconsin Small Business Development Center is analyzed below. This survey was done primarily to "allow educators, researchers, and public policy makers to evaluate the status of Wisconsin's restaurant sector and to identify particular problems that it is encountering". Shown below are the summary statistics for the variable Wages (as a percentage of Sales) broken down by TypeFood (the type of restaurant).

Variable	TypeFood	N	Mean	StDev	Q1	Median	Q3
Wages	Fast Food	102	25.18	10.69	20.00	25.00	30.00
	Supper Club	67	24.84	11.85	20.00	25.00	30.00
	Other	65	25.75	10.37	20.00	27.00	33.00

a) (3 points) Assume that the wage data for the other restaurants follow a normal distribution with mean $\mu = 26$ and a standard deviation $\sigma = 10$. What Wage level do the lowest 30% of Other restaurants have?

P[Z = -0.52] = 0.3 $\Rightarrow z^* = -0.52 \Rightarrow x^* = 26 - 0.52 \times 10 = 20.8$

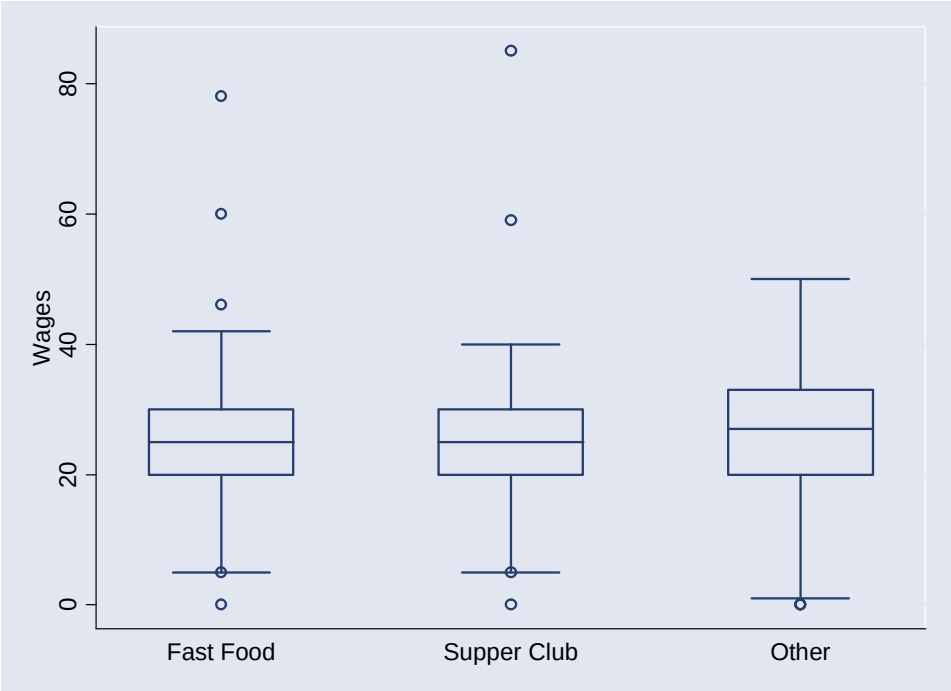


So the lowest 30% of Other restaurants have Wages less than 20.8.

b) (3 points) The largest Wage value for the Supper Club restaurants is 85. How much would be mean change for this group if this observation was dropped from the data set?

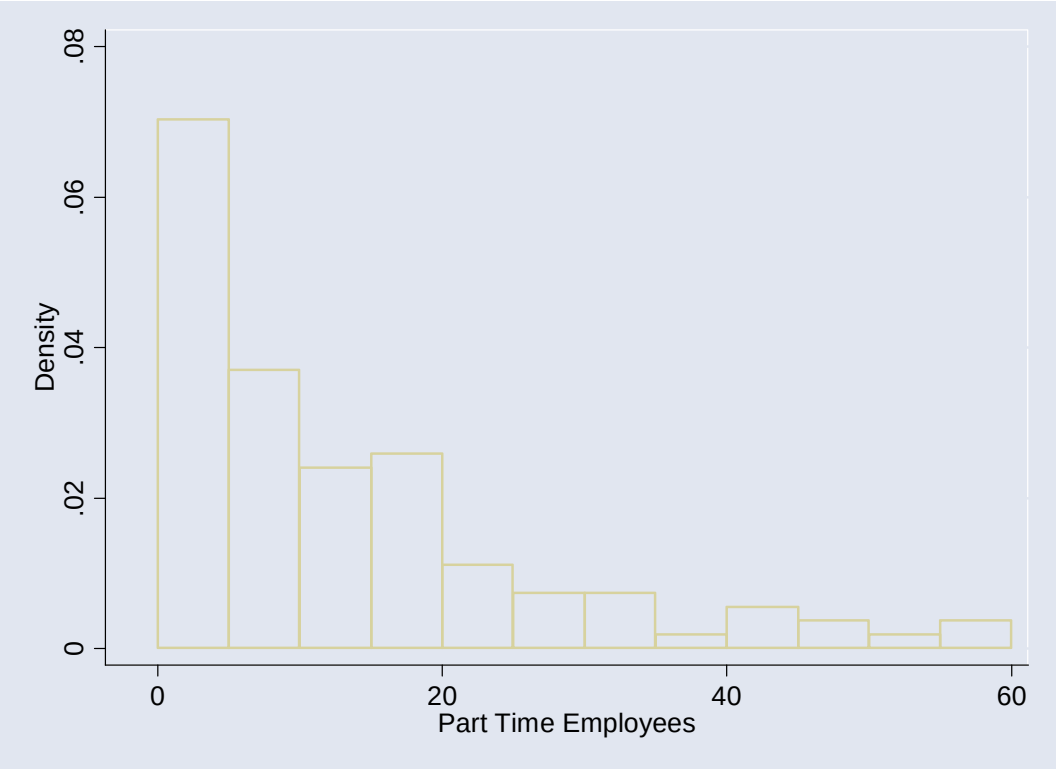
Total of all 67 Supper Club Restaurants is 67 x 24.84 = 1664.28. Remove the largest value from this leaves 1664.28 – 85 = 1579.28. These remaining 66 restaurants have an average of 23.93 (= 1579.28 ÷ 66).

c) (3 points) Describe briefly what the relationship between wages (as a percentage of sales) and restaurant type. Below is a side by side box plot of the data.



There is very little difference between the three types of restaurants in the Wages they pay. The medians and the quartiles are very close together. One noticeable difference, though not particularly important, is the existence of outliers (based in the 1.5IQR rule) for the fast food and supper club restaurants, but not for the others.

d) (3 points) One variable that might influence the amount spent in Wages is the number of part time employees. A histogram of the number of part time employees for the Fast Food restaurants is displayed below. Describe the major features of the distribution displayed in the plot.



This distribution is skewed strongly to the right. It appears that median value is around 10 (the true median is 8). This distribution has a very long right tail.

3. (15 points) A prolific novel writer has an assistant, who, among other things, proofreads drafts of the manuscripts for errors. The writer also proofreads the drafts. The table below gives the probability distribution of an error being detected, for the writer and the assistant.

	Writer Detects	Writer Misses
Assistant Detects	0.70	0.15
Assistant Misses	0.10	0.05

a) (2 points) What is P[writer misses an error]? What is P[assistant misses an error]?

P[writer misses an error] = 0.15 + 0.05 = 0.2

P[assistant misses an error] = 0.10 + 0.05 = 0.15

b) (2 points) What is P[writer misses an error **and** the assistant misses the same error]?

0.05 (take straight from table)

c) (3 points) What is P[writer misses an error **given** that the assistant misses the error] (i.e. P[writer misses | assistant misses])?

P[writer misses | assistant misses] = 0.05 ÷ 0.15 = 0.33

d) (2 points) Are the two events, "writer misses" and "assistant misses" independent? Please justify briefly.

No since P[writer misses | assistant misses] ? P[writer misses an error]. Another justification would be that P[writer misses an error] P[assistant misses an error] ? P[writer misses an error **and** the assistant misses the same error].

e) (3 points) Let X be the random variable representing the number of times each error is detected by the writer and the assistant ($X = 0$ if both miss the error, $X = 1$ if one detects the error and the other misses it, and $X = 2$ if both detect the error). Give the probability distribution for X based on the above probability table

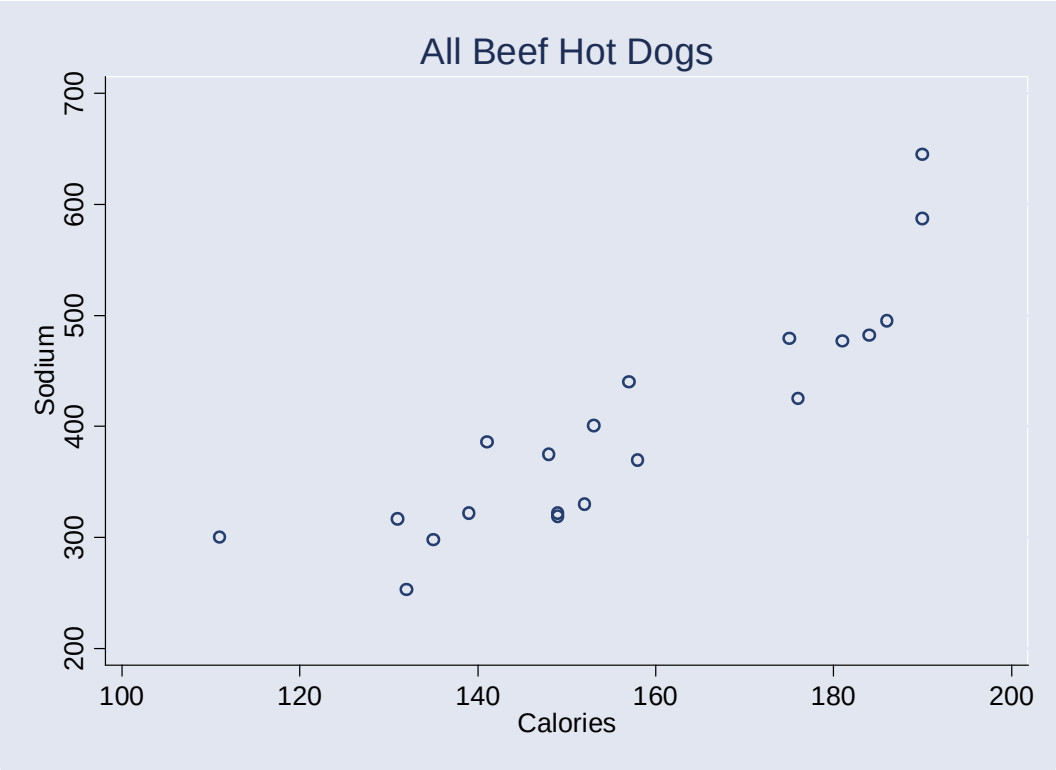
X	0	1	2
p_i	0.05	0.25	0.70

f) (3 points) Find the standard deviation σ of X. You may assume that the mean μ of X is 1.65.

$\sigma^2 = 0.05(0 - 1.65)^2 + 0.25(1 - 1.65)^2 + 0.70(2 - 1.65)^2 = 0.3275$

$\sigma = 0.5723$

4. (18 points) Consumer Reports magazine's laboratory examined calorie and sodium content (in milligrams) for a number of major brands of hot dogs. There are three different types of hot dogs included in the dataset: All Beef, Meat (mainly pork and beef), and Poultry. We will examine the All Beef hot dogs only. Below is a scatterplot of the data and edited regression output from Stata for predicting sodium level by calories




```
. regress sodium calories if hotdog == "Beef"

  Source |       SS       df       MS              Number of obs =      20
---------+------------------------------           F(  1,    18) =   66.48
   Model |  156884.515     1  156884.515           Prob > F      =  0.0000
Residual |  42480.0349    18  2360.00194           R-squared     =  XXXXXX
---------+------------------------------           Adj R-squared =  XXXXXX
   Total |   199364.55    19  10492.8711           Root MSE      =   48.58

------------------------------------------------------------------------------
  sodium |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
---------+--------------------------------------------------------------------
calories |   4.013269   .4922259     8.15   0.000     2.979141    5.047397
   _cons |  -228.3313   77.96608    -2.93   0.009    -392.1319   -64.53064
```

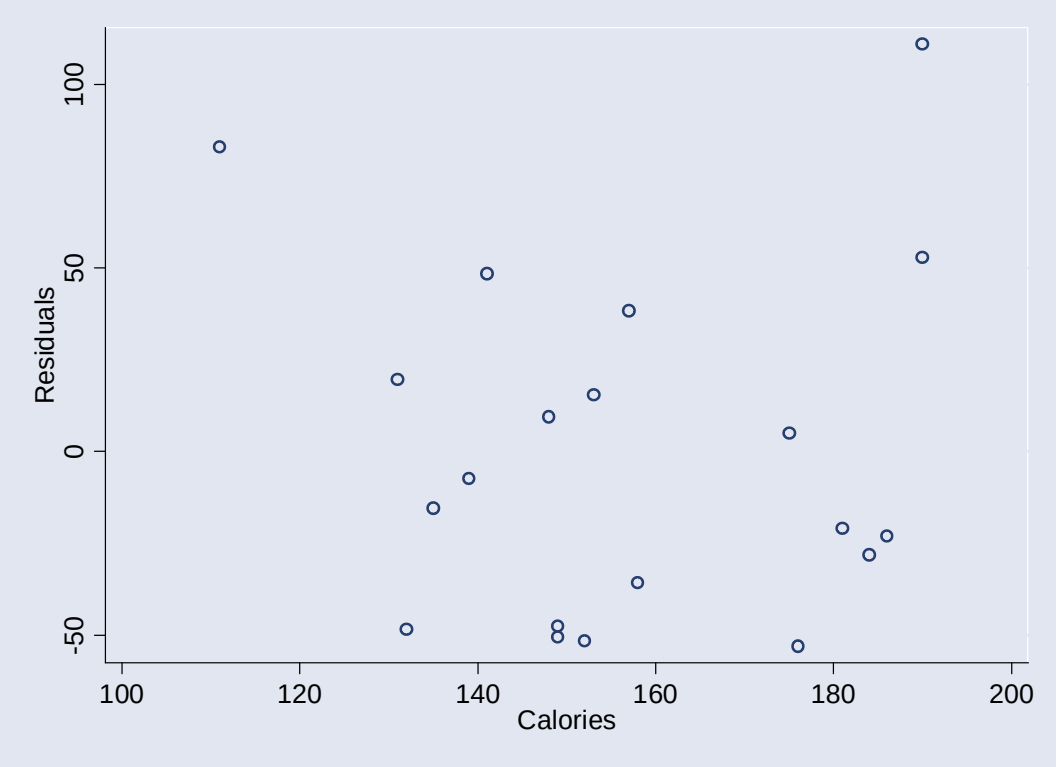
a) (3 points) Suppose that a new brand of all beef hot dog is entering the market. If it contains 111 calories, what you predict its sodium level is?

Fit = -228.33 + 4.013 x 111 = 217.113

b) (3 points) There is already one hot dog on the market that has 111 calories. Its sodium level is 300 mg. What is the residual for this observation?

Residual = 300 – 217.113 = 82.887

c) (3 points) The following is the residual plot for this regression. Is there any evidence that a straight line description in not reasonable for this dataset?



The linearity assumption seems somewhat reasonable, though there are 2 observation, one where calories = 111 (smallest x) and one where calories ≈ 190 (largest x) which have large residuals. This can either suggest some curvature or these observations might be outliers.

d) (3 points) For the observation that has calories = 111, is there any evidence that this observation is an outlier.

By the 2SE rule, this observation is not an outlier as 82.9 < 2 x 48.58 = 97.16. However in this plot is appears that the observation and the larger observation at 190 (residual > 100) are extreme as the rest of the observations have residuals in the range of -50 to 50. As the standard deviation isn't resistant, these two observations are greatly inflating the

standard deviation of the residuals and making it harder to declare observations outliers by the 2SE rule.

e) (3 points) Is the correlation between calories and sodium level closest to -0.9, -0.5, 0, 0.5, or 0.9?

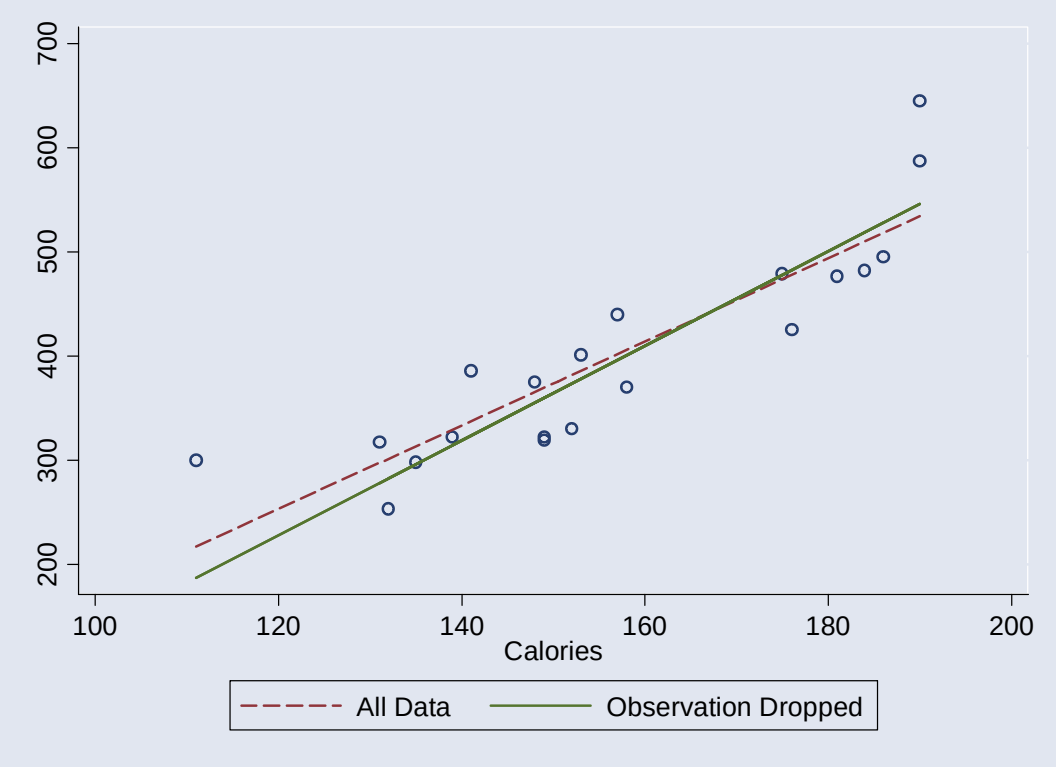
As there is an increasing trend in the data, the first 3 possibilities are not reasonable. As the relationship is strong, only a correlation of 0.9 makes sense.

The true value of 0.88 could be determined from the Stata output as

$$R^2 = \frac{SSModel}{SSTotal} = \frac{156884.52}{199364.55} = 0.787$$

The square root of this gives the correlation. (This was not expected as a possible approach for answering thise question.)

f) (3 points) One of the observation in the data set has a fairly large influence on the estimated regression line as can be seen in the following plot. The regression line using all the data is the dashed line and the regression line based on the data with this influential point omitted is the solid line. Which observation is the omitted observation given that is the one that has the largest influence on the fitted regression line. (Give approximate *x* and *y* co-ordinates.)



The observation at (111, 300) is the influential point. When it is included in the analysis, the regression line is pulled towards this point (dashed line). However when it is omitted from the analysis, the regression line has a much steeper slope.

The second most influential point in the data set is the observation at (190, 650). However it can't be the dropped observation as the fitted line with the point included in the analysis should be closer to this point that the fitted line from the analysis omitting this point.