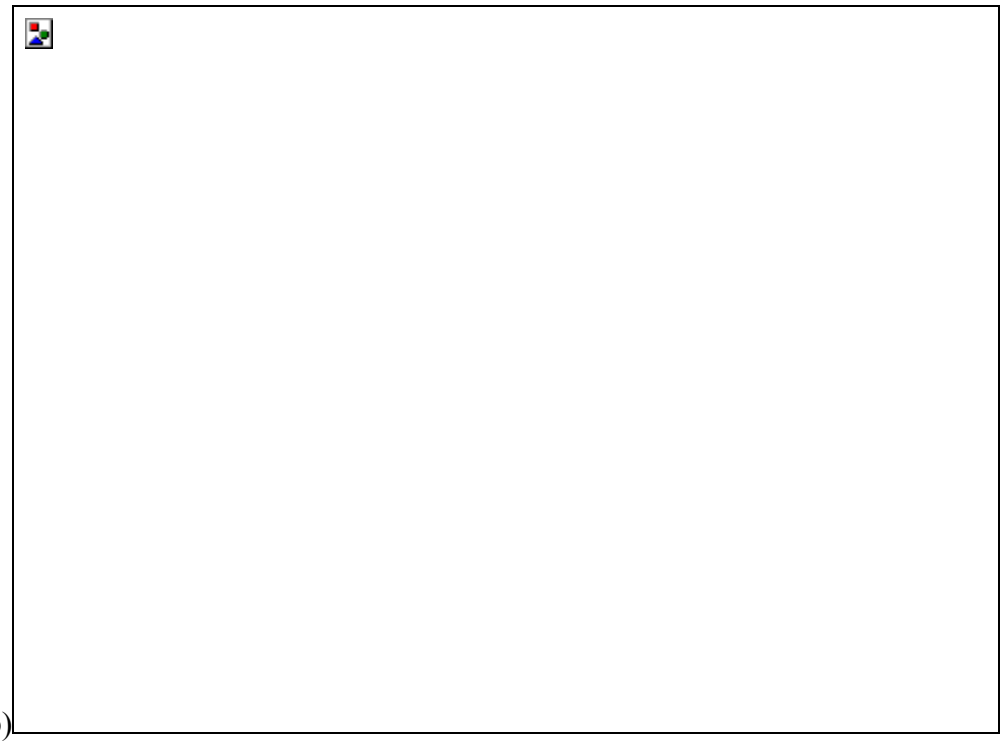
Workshop Statistics: Discovery with Data, Second Edition

Topic 11: Least Squares Regression II

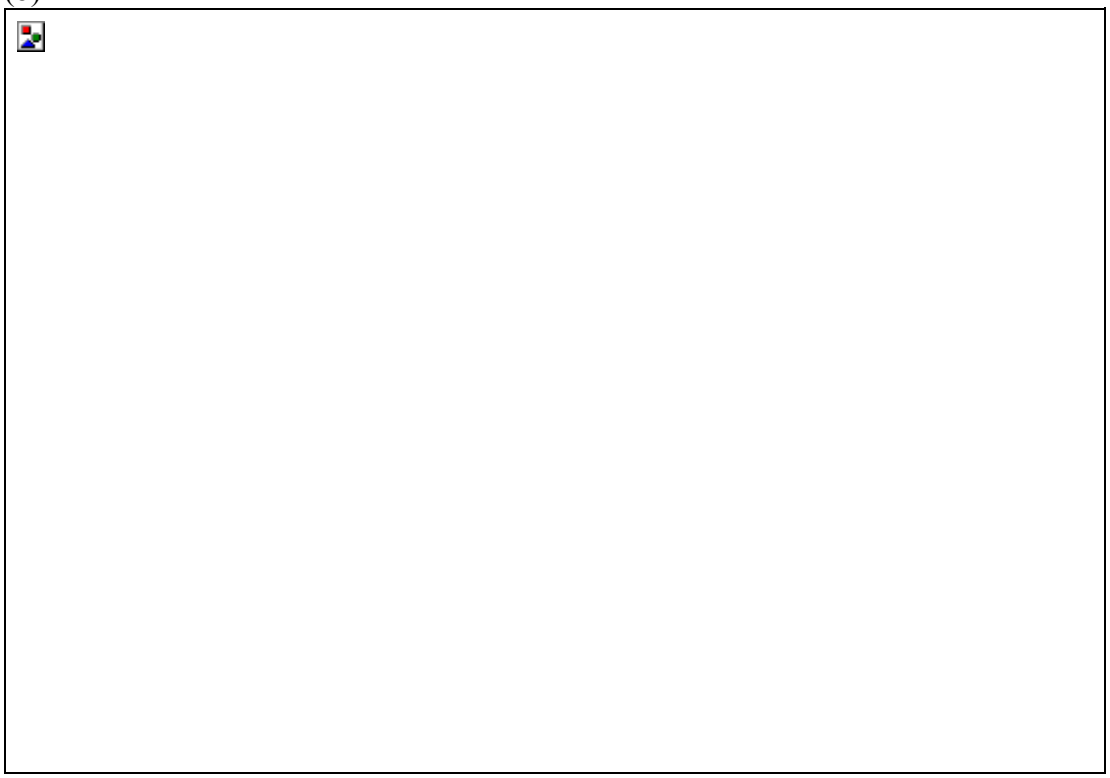
Activity 11-4: Planetary Measurements (*cont.*)

(a) 

A least squares line would not be a good fit for this relationship because it is not a linear relationship.



(b)



The logarithmic transformation seems to produce a more linear relationship than the square root transformation.

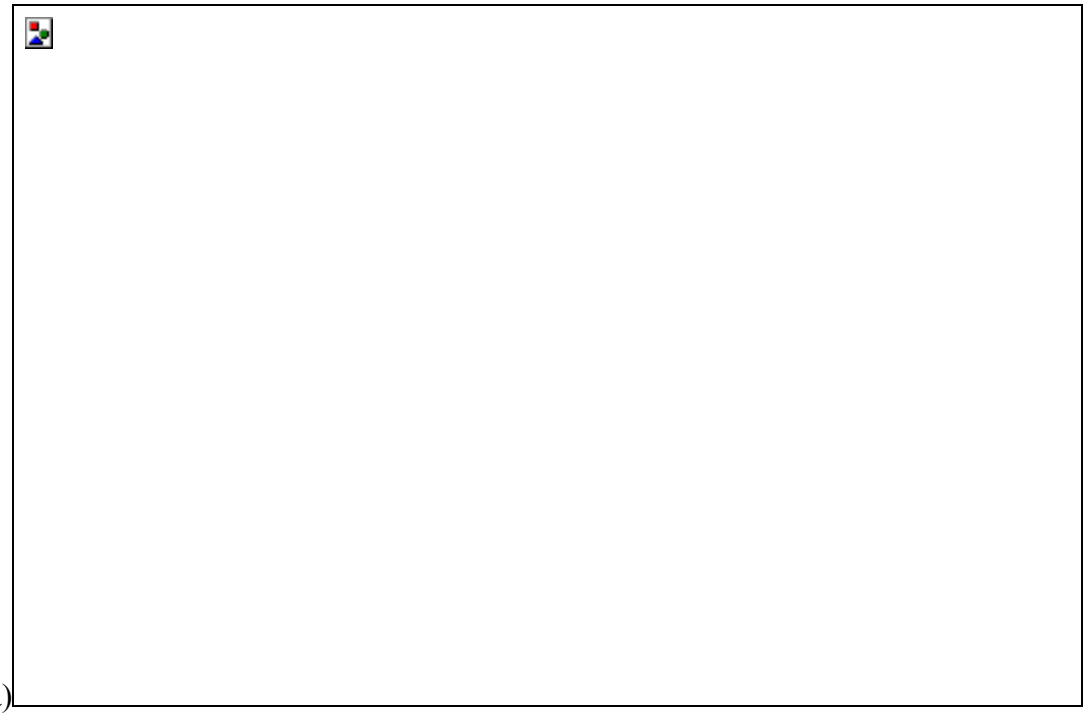
(c) log(distance) = 1.25 + .271 * position

(d) r^2 = .982

(e) 

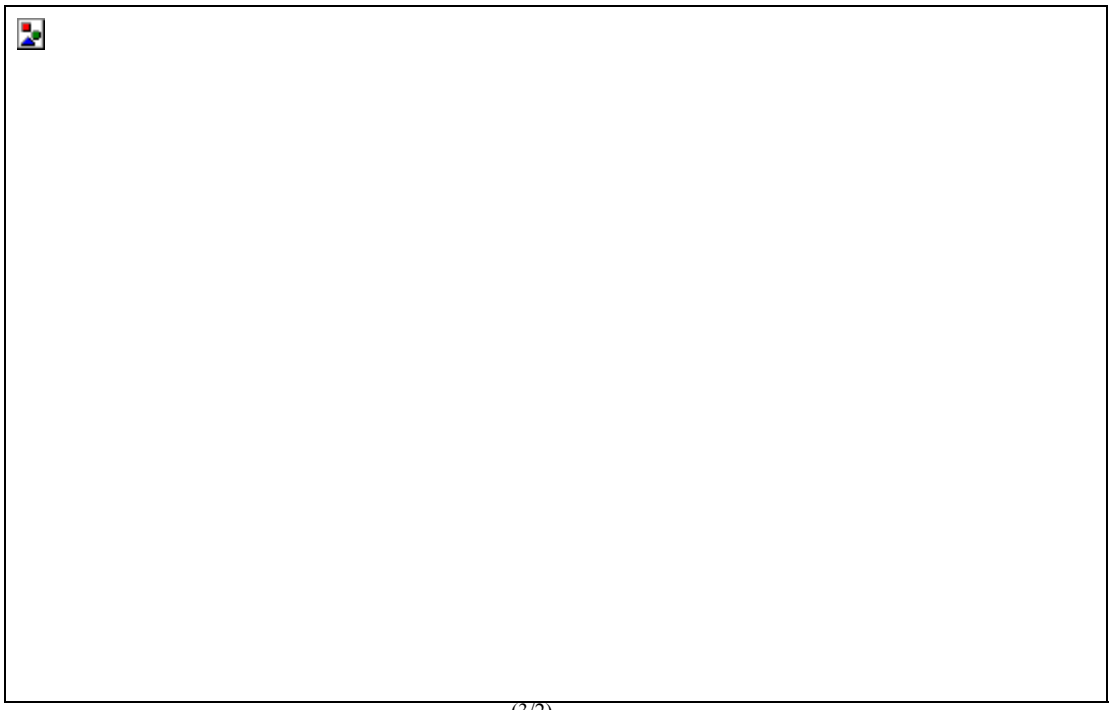
The residuals seem to be scattered randomly.

Activity 11-5: Planetary Measurements (*cont.*)

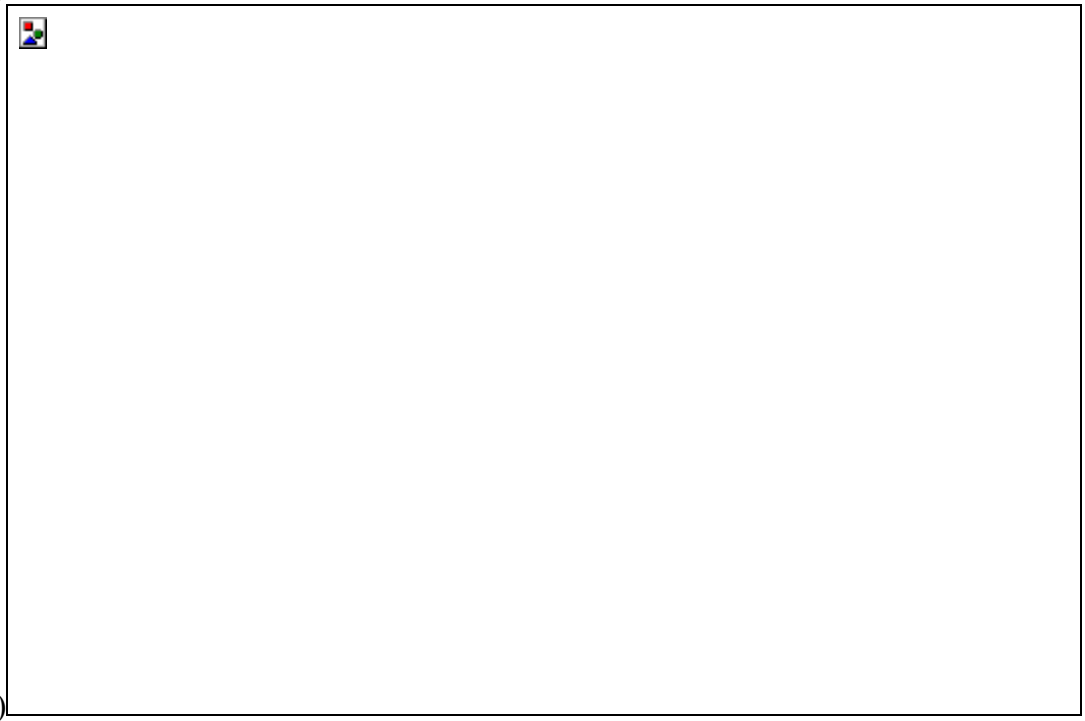
(a) 

There is a strong relationship between period of revolution and distance, though it is not linear.

(b) distance^(3/2)



(c) revolution = 11.2 + .409 * distance^(3/2)



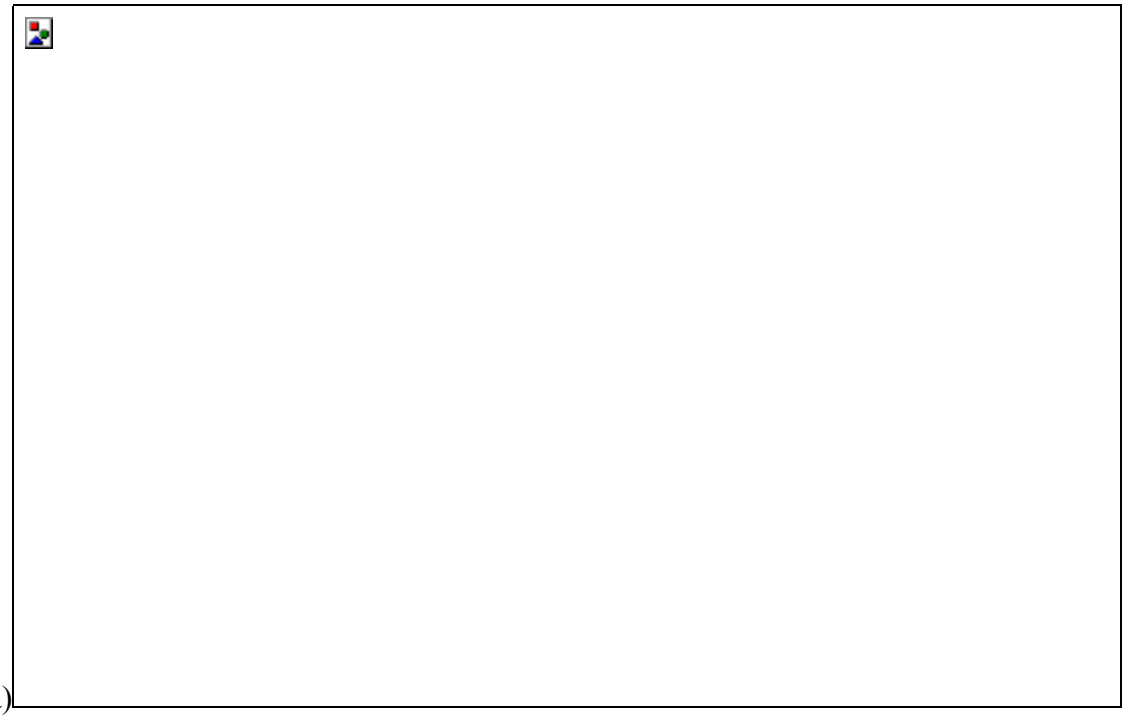
(d)

The residual plot appears to be randomly scattered, signifying that this transformation allows a linear fit.

Activity 11-6: Car Data (*cont.*)

(a)-(b) Answers will vary from student to student.
(c) 1/4 mile = 20.5 - .000987 * weight; Corvette's residual value: -3.147835
(d) 1/4 mile = 20.5 - .000958 * weight; The slope hasn't changed much, and the intercept hasn't changed at all.
(e) Based on our answer to (d), the Corvette is not a very influential observation. Its removal does not significantly change the least squares regression line.

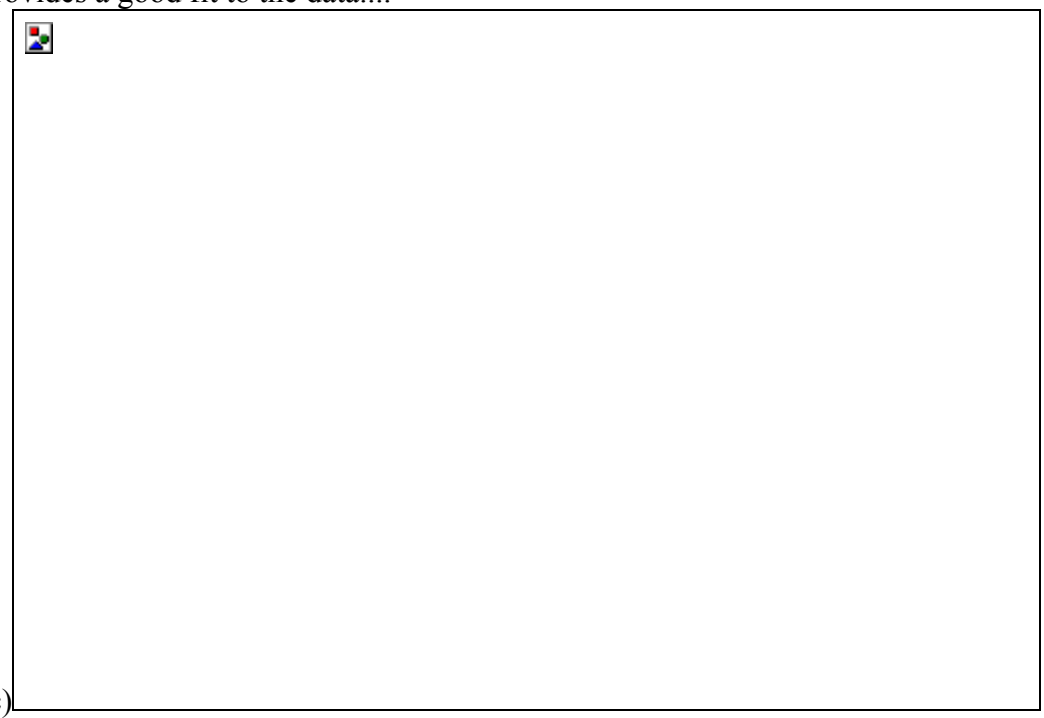
Activity 11-7: Broadway Shows (*cont.*)

(a) 

receipts = -$84,835 + $65.40 * attendance
(b) $36,937
(c) -$100,243
(d) Answers will vary from student to student.
(e) without *The Lion King*: receipts = -$74,081 + $63.40 * attendance
without *Footloose*: receipts = -$81,969 + $65.70 * attendance
The Lion King is more influential.
(d) When we add $100,000 to the gross receipts for *The Lion King*, the regression line becomes: receipts = -$104,343 + $69.20 * attendance. When we subtract $100,000 from the gross receipts for *The Lion King*, the regression line becomes: receipts = -$65,327 + $61.70 * attendance. When we add $100,000 to the gross receipts for *Footloose*, the regression line becomes: -$82,118 + $65.70 * attendance. When we subtract $100,000 from the gross receipts for *Footloose*, the regression line becomes: -$87,552 + $65.20 * attendance. *The Lion King* is definitely more influential than *Footloose*.

Activity 11-8: Solitaire (*cont.*)

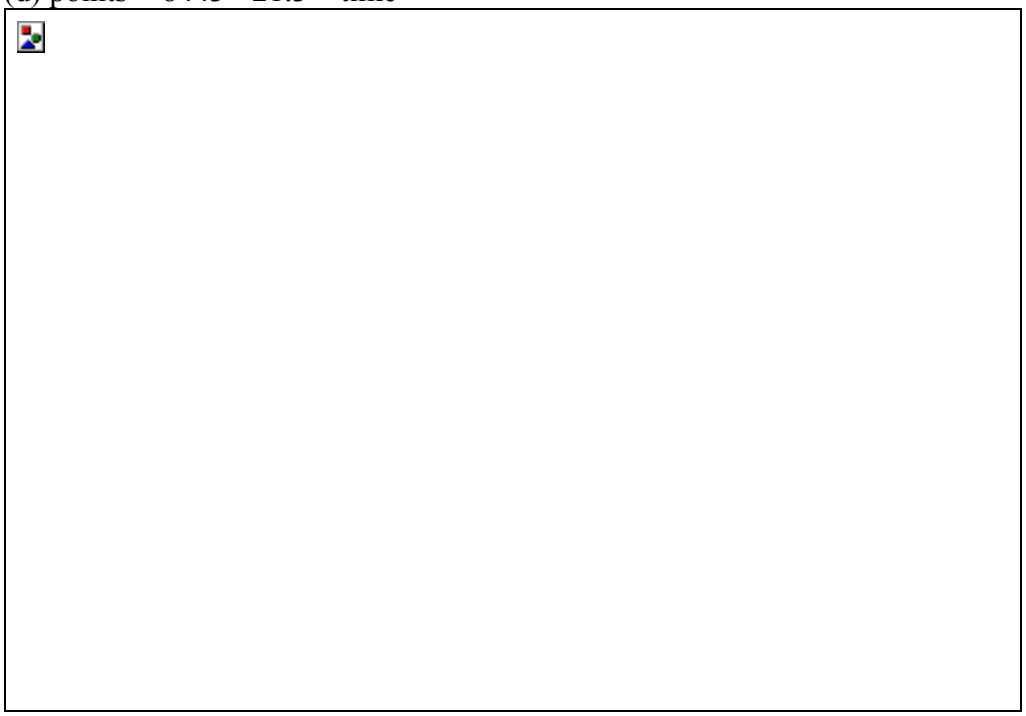
(a) $r = -.979$
(b) $r^2 = 95.9\%$; This is very close to 100%. This would suggest that a linear regression provides a good fit to the data....



(c)

This relationship does not appear to be linear, but rather curved.

(d) points = 8443 - 21.3 * time



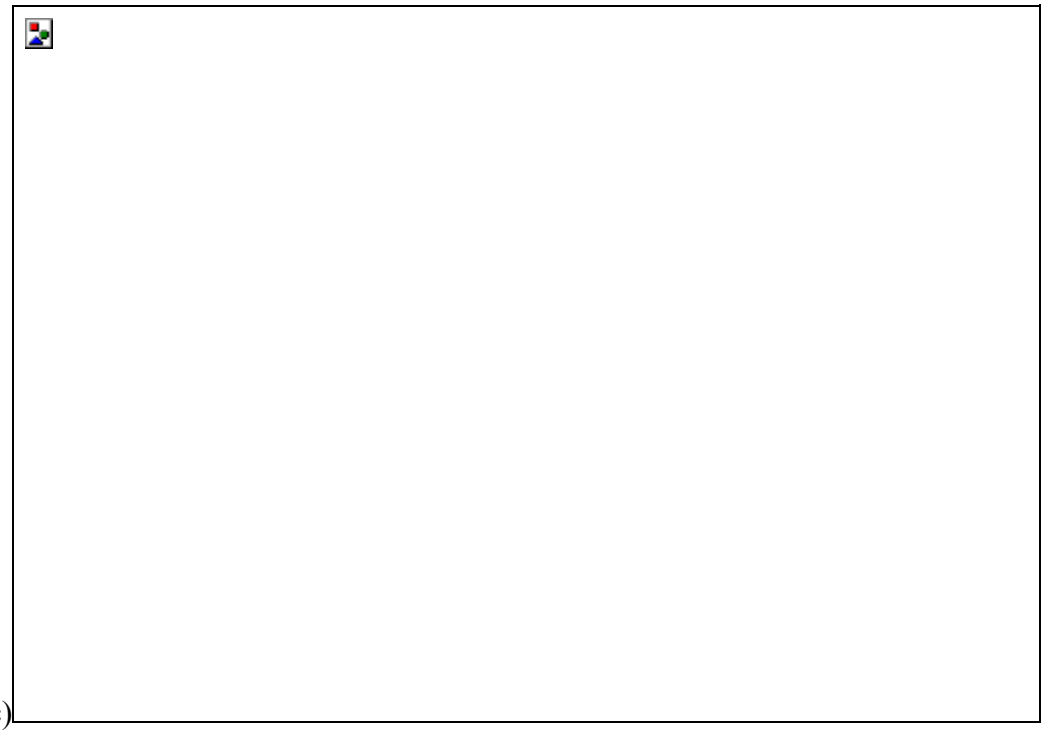
The least squares regression line does not appear to summarize well the relationship between points and time.

(e) 4609 points; about right (maybe a slight overestimate since residuals tend to be

negative between 150 and 200 seconds)
(f) 2 minutes: 5887 points; underestimate (residuals tend to be positive below 150 seconds)
4 minutes: 3331 points; underestimate (residuals tend to be positive above 220 seconds)

Activity 11-9: College Football Players (*cont.*)

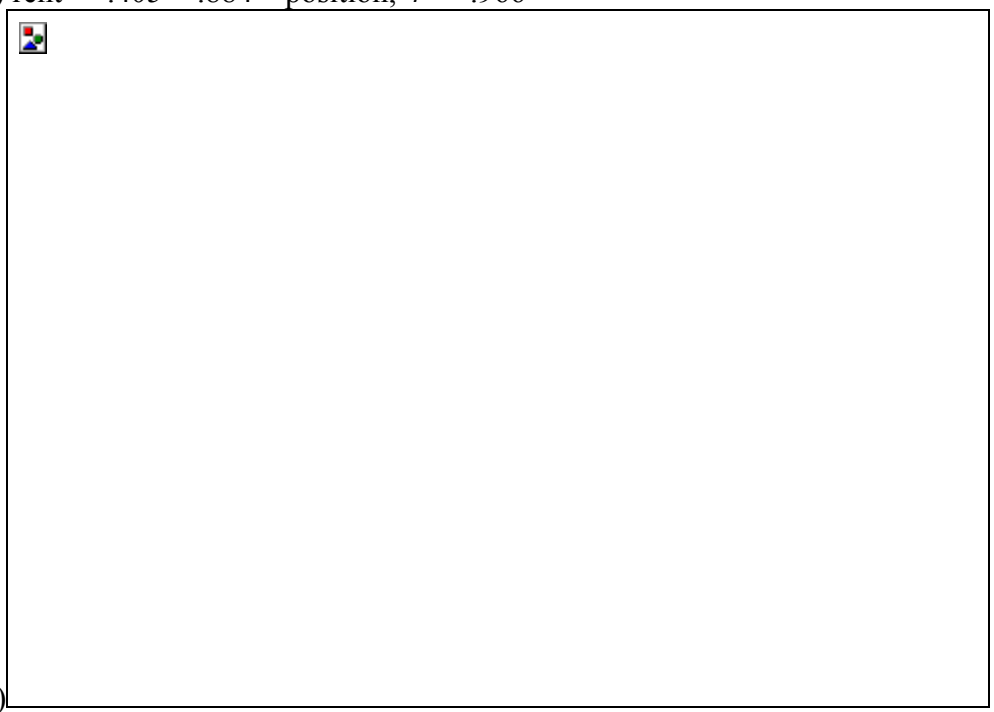
(a) 
(b) Answers will vary from student to student.

(c) 

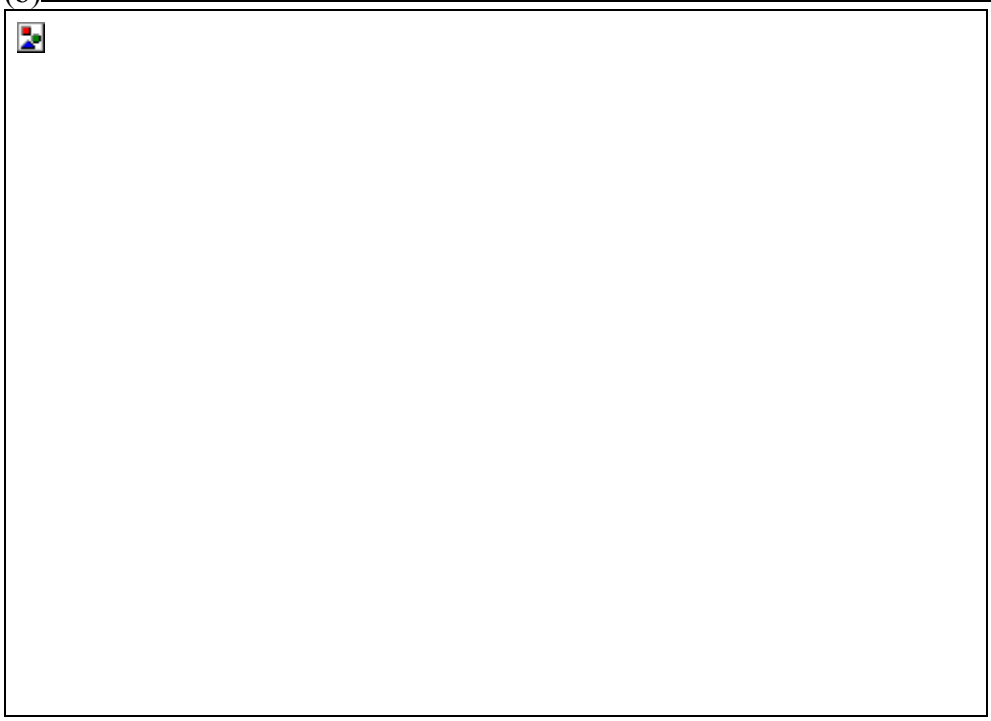
The residuals plot does seem to correspond to the above scatterplot. The residuals are centered at zero as the regression line went through the midal of the data. Players with jersey numbers between 50 and 80 tend to have weights above the regression line (positive residuals) and players with jersey numbers above 80 tend to have weights below the regression line (negative residuals).

Activity 11-10: Monopoly Prices (*cont.*)

(a) rent = -.405 + .884 * position; r^2 = .966



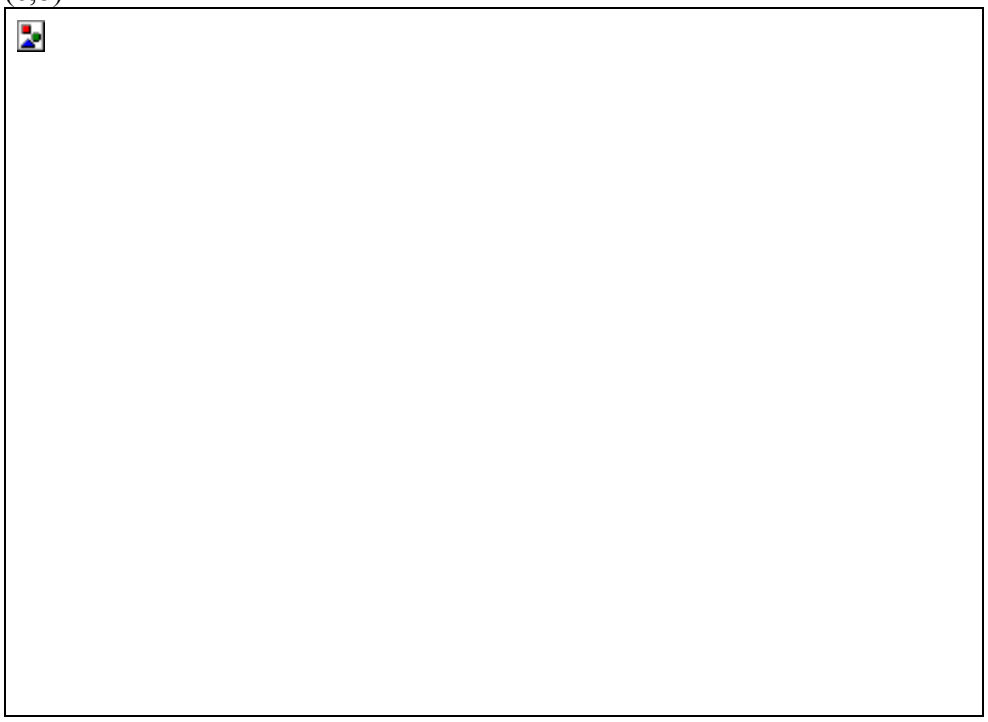
(b)



Althought the least squares regression line is such that the data points fall about it fairly well, the residual plot has a definite pattern with only a couple of outliers, meaning that a linear fit is not the best option in this case.

(c,a) rent = -4.7 + .106 * price; r^2 = .988

(c,b)

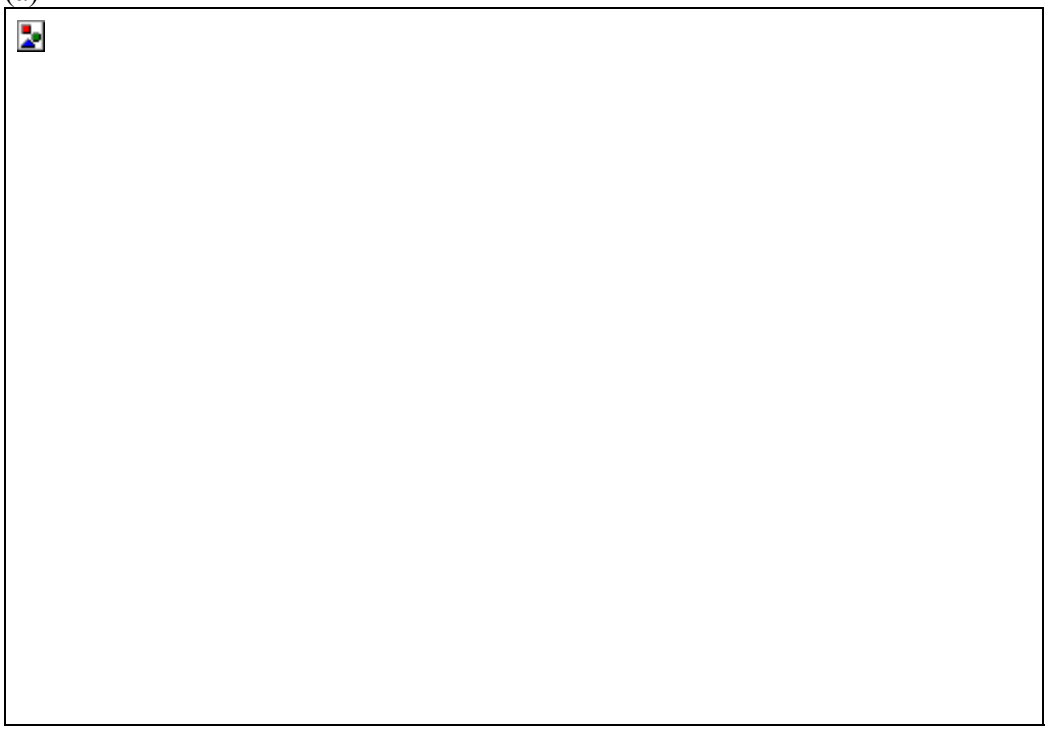


Althought the least squares regression line is such that the data points fall about it fairly well, the residual plot has a definite pattern with only a couple of outliers, meaning that a linear fit is not the best option in this case.

(d) The scatterplots of the residuals vs. the explanatory variables show definite patterns. This causes us to doubt the adequacy of the linear model in both these situations.

Activity 11-11: Box Office Blockbusters (*cont.*)

(a)

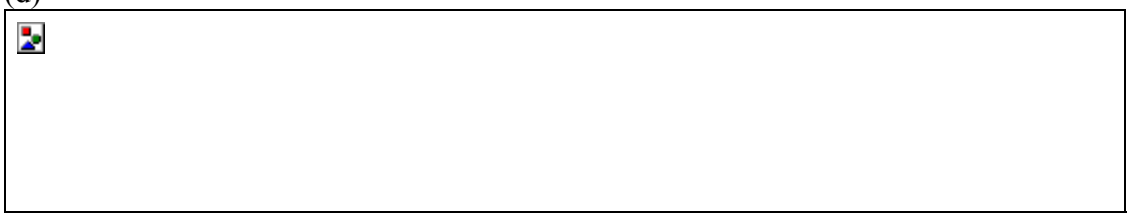


r = .882; There appears to be a fairly strong positive association.

(b) revenue2 = $3,150,142 +$.566 * revenue1 [or 3.15 + $.566 revenue 2 in millions of dollars]

(c) *Star Wars: The Phantom Menace* has the largest (in absolute value) residual at $11,566,714 [11.547]. This movie had an incredible first weekend, partially due to the fact that it opened on a Wednesday, so it had two extra days to earn money for first weekend revenues. There was a large drop in second weekend revenues for this movie.

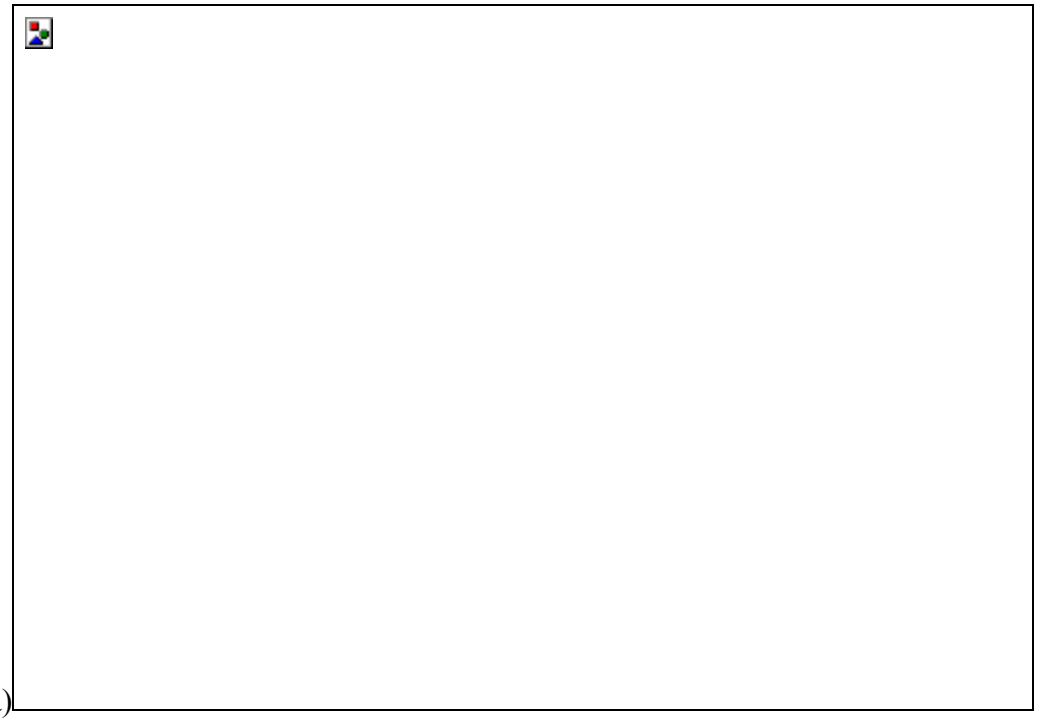
(d)



The mean is 0. The median is about -.48851 (values will vary with rounding differences). There are nine negative residuals, a proportion of .6. The regression line seems to be overestimating the second week box office revenue for most movies.

(e) *Star Wars: The Phantom Menace* is most likely to be an influential observation. revenue2 = $6,689,792 + $.43 * revenue1 [or 6.69 + .43 revenue1]; The equation has more than doubled its y-intercept, and the slope has decresed by .136.

Activity 11-12: College Enrollments

(a) 

r = .935

There is a strong positive association between enrollement and number of faculty.

(b) faculty = 2.5 + .0698 * enrollment

(c) For about every 100 additional students, 7 faculty members are added on average.

(d) Answers will vary from student to student, but they might point to college 9 as influential and college 28 as having a high residual.

Activity 11-13: Gestation and Longevity (*cont.*)

Answers will vary from student to student based on the longevity guess. However, they should not that there is no other data near the longevity of a human so it is hard to rely on this regression equation to predict the genstation for humans.

Activity 11-14: Gestation and Longevity (*cont.*)

(a) gestation = 21.7 + 13.1 * longevity

(b) 283.7 days

(c) about 26 years (26.206)

(d) longevity = 6.64 + .0335 * gestation

(e) about 19 years (18.867)

(f) The regression equation for predicting longevity from gestation is not algebraically equivalent to the regression equation for prediction gestation from longevity. If it was, the second equation would have been -1.66 + .076(gestation). Minimizing the residuals in one direction can produce a line much different than the one that minimizes the residuals in the other direction.. Therefore, the equations can look very different based on which

variable is the explanatory, and which is the response.

Activity 11-15: Turnpike Tolls (*cont.*)

(a) There appear to be definite patterns in this scatterplot.
(b) Despite the high correlation coefficient, the patterns in this residual scatterplot give us reason to suspect that the relationship between toll and mileage is not completely linear. Refering back to the original scatterplot of toll vs. mileage in Activity 10-12, one can see some definite curves about the regression line.