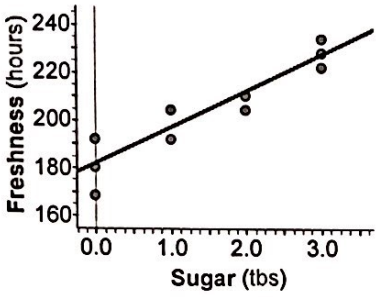
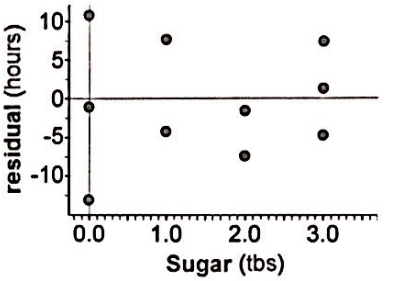
Key

3.2 Computer Output/ Regression Wisdom

Alternate Example: Fresh Flowers

For their second semester project, two AP Statistics students decided to investigate the effect of sugar on the life of cut flowers. They went to the local grocery store and randomly selected 12 carnations. All the carnations seemed equally healthy when they were selected. When they got home, the students prepared 12 identical vases with exactly the same amount of water in each vase. They put 1 tablespoon of sugar in three vases, 2 tablespoons of sugar in three vases, and 3 tablespoons of sugar in three vases. In the remaining 3 vases, they put no sugar. After the vases were prepared and placed in the same location, the students randomly assigned one flower to each vase and observed how many hours each flower continued to look fresh. A scatterplot, residual plot, and computer output from the regression are shown.





Predictor	Coef	SE Coef	T	P
Constant	181.200	3.635	49.84	0.000
Sugar	15.200	1.943	7.82	0.000

S = 7.52596 R-Sq = 86.0% R-Sq(adj) = 84.5%

(a) What is the equation of the least-squares regression line? Define any variables you use.

$\hat{Freshness} = 181.2 + 15.2\,(Sugar)$

(b) What is the correlation?

$r = \sqrt{.86} = 0.93$

(c) Is a linear model appropriate for this data? Explain.

Yes. The scatterplot shows a linear form. The residual plot shows no obvious pattern which means a linear model is appropriate. The correlation of 0.93 means there is a strong positive linear relationship btw sugar and freshness.

(d) Calculate and interpret the residual for the flower that got 3 tablespoons of sugar and lasted 222 hours.

$\hat{y} = 181.2 + 15.2(3)$

$\hat{y} = 226.8$

$Residual = y - \hat{y}$

$= 222 - 226.8$

$= -4.8$ hours

The diff btw the actual freshness and predicted freshness is 4.8 hrs. The model overpredicted by 4.8 hours how long the flower would stay fresh.

(e) What is r^2 and interpret its meaning?

$r^2 = .86$ 86% of the variation in freshness can be explained by the linear regression of freshness vs. sugar.

Chapter 3 Review "so far"

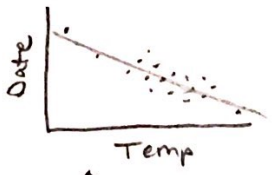
$s = 2.17$

The anticipation of the first blooms of spring flowers is one of the joys of April. One of the most beautiful is that of the Japanese cherry tree. Experience has taught us that, if the spring has been a warm one, the trees will bloom early, but if the spring has been cool, then the blossoms will appear later. Mr. Yamada is a gardener who has observed the date in April when the blossoms first appear for the last 24 years. His son, Hiro, went on the internet and found the average March temperature (in degrees Celsius) in his area for those years. For example, for the first year in the data table, the average March temperature was 4°C and the first bloom happened on the 14th day of April.

Temp	Date
4	14
5.4	8
3.2	11
2.6	19
4.2	14
4.7	14
4.9	14
4	21
4.9	9
3.8	14
4	13
5.1	11
4.3	13
1.5	28
3.7	17
3.8	19
4.5	10
4.1	17
6.1	3
6.2	3
5.1	11
5	6
4.6	9
4	11

1. Why should temperature be the explanatory variable?
2. Draw a scatterplot and discuss the noticeable features.
3. Calculate the least squares regression line and graph it on the scatterplot.
4. Interpret the slope in the context of the problem.
5. Interpret the y-intercept in the context of the problem. Is it a reasonable value in this context?
6. Calculate and interpret the value of the correlation coefficient.
7. If the temperature was measured in degrees Fahrenheit, how would this value change?
8. If r is high, can we conclude that warmer temperatures cause earlier blooms?
9. Calculate and interpret the residual for the first year in the data table.
10. Sketch the residual plot. What does it tell you?
11. Calculate and interpret the values of r^2 ~~and s~~ in the context of the problem.
12. If you were to use number of hours in April instead of number of days in April for the date of first bloom, how would the values of r^2 and s change?
13. Predict the date of first bloom for a year with an average March temperature of 3°C. How confident are you in your prediction?

1. Temp. should be the explanatory variable because the temp. influences when the trees bloom. The bloom date depends on the temp.

2. 

There is a strong, negative, linear relationship btw temp and date of 1st bloom. There could be one possible outlier at (1.5, 28).

3. $\hat{Date} = 33.12 - 4.69(Temp)$
4. As avg. temp increases by 1°C, predicted bloom date decreases by 4.69 days, on avg.
5. When avg. temp in March is zero, the date of the 1st bloom is 33rd day of April. Since there are only 30 days in April, this means the 1st bloom would happen in May. This seems reasonable.
6. $r = -.85$ There is a strong, negative linear relationship btw temp and bloom date
7. r would not change b/c it has no units
8. No.
9. $\hat{Date} = 33.12 - 4.69(4)$ $= 14.36$ Resid. $= y - \hat{y}$ $= 14 - 14.36$ $= -.36$

The diff btw actual bloom date and predicted bloom date is 0.36 days. The model overpredicted the bloom date by .36 days

10. 

The residual plot shows no obvious pattern which means a linear model is appropriate for the data

11. $r^2 = 0.72$ 72% of the variation in bloom date can be explained by avg. March temp.

$s = 2.17$ The predictions made using the model will be off by 2.17 days on average.

12. r^2 would not change b/c it has no units
s is expressed in the same units as the data

13. $\hat{Date} = 33.12 - 4.69(3) = 19.05$ days
This prediction seems reliable b/c 3°C is within the range of the data used to calculate the model.