Chapter 9 – Regression Wisdom

1. Marriage age 2003.

a) The trend in age at first marriage for American women is very strong over the entire time period recorded on the graph, but the direction and form are different for different time periods. The trend appears to be somewhat linear, and consistent at around 22 years, up until about 1940, when the age seemed to drop dramatically, to under 21. From 1940 to about 1970, the trend appears non-linear and slightly positive. From 1975 to the present, the trend again appears linear and positive. The marriage age rose rapidly during this time period.

b) The association between age at first marriage for American women and year is strong over the entire time period recorded on the graph, but some time periods have stronger trends than others.

c) The correlation, or the measure of the degree of linear association is not high for this trend. The graph, as a whole, is non-linear. However, certain time periods, like 1975 to present, have a high correlation.

d) Overall, the linear model is not appropriate. The scatterplot is not "straight enough" to satisfy the condition. You could fit a linear model to the time period from 1975 to 1995, but this seems unnecessary. The ages for each year are reported, and, given the fluctuations in the past, extrapolation seems risky.

2. Smoking 2004.

a) The percent of men 18 – 24 who are smokers decreased dramatically between 1965 and 1990, but the trend has not been consistent since then.

b) The association between percent of men 18 – 24 who smoke and year is very strong from 1965 to 1990, but is erratic after 1990.

c) A linear model is not an appropriate model for the trend in the percent of males 18 – 24 who are smokers. The relationship is not straight.

3. Human Development Index.

a) Fitting a linear model to the association between HDI and GDPPC would be misleading, since the relationship is not straight.

b) If you fit a linear model to these data, the residuals plot will be curved downward.

c) Setting aside the single data point corresponding to Luxembourg will not improve the model. The relationship will still be curved.

4. HDI revisited.

a) Fitting a linear model to the association between the number of cell phones and HDI would be misleading, since the relationship is not straight.

b) The residuals plot will be curved downward.

5. Good model?

a) The student's reasoning is not correct. A scattered residuals plot, not high R^2, is the indicator of an appropriate model. Once the model is deemed appropriate, R^2 is used as a measure of the strength of the model.

b) The model may not allow the student to make accurate predictions. The data may be curved, in which case the linear model would not fit well.

6. Bad model?

a) The student's model may, in fact, be appropriate. Low R^2simply means that the model is not accurate. The model explains only 13% of the variability in the response variable. If the residuals plot shows no pattern, this model may be appropriate.

b) The predictions are not likely to be very accurate, but they may be the best that the student can get. R^2= 13% indicates a great deal of scatter around the regression line, but if the residuals plot is not patterned, there probably isn't a better model. The two variables that are being studied by the student have a weak association.

7. Movie Dramas.

a) The units for the slopes of these lines are millions of dollars per minutes of running time.

b) The slopes of the regression lines are the same. Dramas and movies from other genres have costs for longer movies that increase at the same rate.

c) The regression line for dramas has a lower y-intercept. Regardless of running time, dramas cost about 20 million dollars less than other genres of movies of the same running time.

8. Movie Ratings.

a) The slopes of the regression lines are approximately the same. The costs increase at about the same rate for all genres as the movies get longer.

b) Although the costs per minute are about the same, it costs about 20 million dollars less to make an R-rated movies than a movie of the other rating type with the same running time.

c) Omitting *King Kong* would make the slope for the PG-13 movies steeper. We would conclude that the cost per minute of PG-13 movies was greater than the cost per minute of movies with other rating.

9. Airline passengers.

a) According to the linear model, the use of the Oakland airport has been increasing by about 59,700 passengers per year, starting at about 282,000 passengers in 1990.

b) About 71% of the variability in the number of passengers can be accounted for by the model.

c) Errors in prediction based on the model have a standard deviation of 104,330 passengers.

d) No, the model would not be useful in predicting the number of passengers in 2010. This year would be an extrapolation too far from the years we have observed.

e) The negative residual is September of 2001. Air traffic was artificially low following the attacks on 9/11/2001.

10. Tracking hurricanes.

a) According to the linear model, tracking errors averaged about 292 nautical miles in 1970, and have increased an average of 5.23 nautical miles per year since then.

b) Residuals based on this model have a standard deviation of 42.87 nautical miles.

c) The linear model for the trend in predicting error is $\hat{Error} = 292.089 - 5.22924(Year - 1970)$.

$\hat{Error} = 292.089 - 5.22924(Year - 1970)$

$\hat{Error} = 292.089 - 5.22924(39)$

$\hat{Error} = 88.1$

The model predicts an error of only 88.1 nautical miles in 2009. However, the NHC has already achieved better than 125 nautical mile accuracy, so it seems safe to assume they can maintain the level of success even if the decline in prediction errors doesn't continue.

d) A tracking error of 90 nautical miles could be achieved is the trend fit by the regression model continues, but this is an extrapolation beyond the data.

e) We should be cautious in assuming that the improvements in prediction will continue at the same rate.

11. Unusual points.

a) **1)** The point has high leverage and a small residual.
2) The point is not influential. It has the *potential* to be influential, because its position far from the mean of the explanatory variable gives it high leverage. However, the point is not *exerting* much influence, because it reinforces the association.
3) If the point were removed, the correlation would become weaker. The point heavily reinforces the positive association. Removing it would weaken the association.
4) The slope would remain roughly the same, since the point is not influential.

b) **1)** The point has high leverage and probably has a small residual.
2) The point is influential. The point alone gives the scatterplot the appearance of an overall negative direction, when the points are actually fairly scattered.
3) If the point were removed, the correlation would become weaker. Without the point, there would be very little evidence of linear association.
4) The slope would increase, from a negative slope to a slope near 0. Without the point, the slope of the regression line would be nearly flat.

c) **1)** The point has moderate leverage and a large residual.
2) The point is somewhat influential. It is well away from the mean of the explanatory variable, and has enough leverage to change the slope of the regression line, but only slightly.
3) If the point were removed, the correlation would become stronger. Without the point, the positive association would be reinforced.
4) The slope would increase slightly, becoming steeper after the removal of the point. The regression line would follow the general cloud of points more closely.

d) **1)** The point has little leverage and a large residual.
2) The point is not influential. It is very close to the mean of the explanatory variable, and the regression line is anchored at the point $(\bar{x}, \bar{y})$, and would only pivot if it were possible to minimize the sum of the squared residuals. No amount of pivoting will reduce the residual for the stray point, so the slope would not change.
3) If the point were removed, the correlation would become slightly stronger, decreasing to become more negative. The point detracts from the overall pattern, and its removal would reinforce the association.
4) The slope would remain roughly the same. Since the point is not influential, its removal would not affect the slope.

12. More unusual points.

a) **1)** The point has high leverage and makes the large residual a bit smaller.
2) The point is influential. It is well away from the mean of the explanatory variable, and has enough leverage to change the slope of the regression line.
3) If the point were removed, the correlation would become stronger. Without the point, the positive association would be reinforced.
4) The slope would increase, becoming steeper after the removal of the point. The regression line would follow the general cloud of points more closely.

b) **1)** The point has high leverage and a small residual.
2) The point is influential. The point alone gives the scatterplot the appearance of an overall positive direction, when the points are actually fairly scattered.
3) If the point were removed, the correlation would become weaker. Without the point, there would be very little evidence of linear association.
4) The slope would decrease, from a positive slope to a slope near 0. Without the point, the slope of the regression line would be nearly flat.

c) **1)** The point has little leverage and a large residual.
2) The point is not influential. It is very close to the mean of the explanatory variable, and the regression line is anchored at the point $(\bar{x}, \bar{y})$, and would only pivot if it were possible to minimize the sum of the squared residuals. No amount of pivoting will reduce the residual for the stray point, so the slope would not change.
3) If the point were removed, the correlation would become slightly stronger. The point detracts from the overall pattern, and its removal would reinforce the association.
4) The slope would remain roughly the same. Since the point is not influential, its removal would not affect the slope.

d) **1)** The point has high leverage and a small residual.
2) The point is not influential. It has the *potential* to be influential, because its position far from the mean of the explanatory variable gives it high leverage. However, the point is not *exerting* much influence, because it reinforces the association.
3) If the point were removed, the correlation would become weaker. The point heavily reinforces the association. Removing it would weaken the association.
4) The slope would remain roughly the same, since the point is not influential.

13. The extra point.

1) Point e is very influential. Its addition will give the appearance of a strong, negative correlation like $r = -0.90$.

2) Point d is influential (but not as influential as point e). Its addition will give the appearance of a weaker, negative correlation like $r = -0.40$.

3) Point c is directly below the middle of the group of points. Its position is directly below the mean of the explanatory variable. It has no influence. Its addition will leave the correlation the same, $r = 0.00$.

4) Point b is almost in the center of the group of points, but not quite. Its addition will give the appearance of a very slight positive correlation like $r = 0.05$.

5) Point a is very influential. Its addition will give the appearance of a strong, positive correlation like $r = 0.75$.

14. The extra point revisited.

1) Point d is influential. Its addition will pull the slope of the regression line toward point d, resulting in the steepest negative slope, a slope of –0.45.

2) Point e is very influential, but since it is far away from the group of points, its addition will only pull the slope down slightly. The slope is –0.30.

3) Point c is directly below the middle of the group of points. Its position is directly below the mean of the explanatory variable. It has no influence. Its addition will leave the slope the same, 0.

4) Point b is almost in the center of the group of points, but not quite. It has very little influence, but what influence it has is positive. The slope will increase very slightly with its addition, to 0.05.

5) Point a is very influential. Its addition will pull the regression line up to its steepest positive slope, 0.85.

15. What's the cause?

1) High blood pressure may cause high body fat.
2) High body fat may cause high blood pressure.
3) Both high blood pressure and high body fat may be caused by a lurking variable, such as a genetic or lifestyle trait.

16. What's the effect?

1) Playing computer games may make kids more violent.
2) Violent kids may like to play computer games.
3) Playing computer games and violence may both be caused by a lurking variable such as the child's home life or a genetic predisposition to aggressiveness.

17. Reading.

a) The principal's description of a strong, positive trend is misleading. First of all, "trend" implies a change over time. These data were gathered during one year, at different grade levels. To observe a trend, one class's reading scores would have to be followed through several years. Second, the strong, positive relationship only indicates the yearly improvement that would be expected, as children get older. For example, the 4^{th} graders are reading at approximately a 4^{th} grade level, on average. This means that the school's students are progressing adequately in their reading, not extraordinarily. Finally, the use of average reading scores instead of individual scores increases the strength of the association.

b) The plot appears very straight. The correlation between grade and reading level is very high, probably between 0.9 and 1.0.

c) If the principal had made a scatterplot of all students' scores, the correlation would have likely been lower. Averaging reduced the scatter, since each grade level has only one point instead of many, which inflates the correlation.

d) If a student is reading at grade level, then that student's reading score should equal his or her grade level. The slope of that relationship is 1. That would be "acceptable", according to the measurement scale of reading level. Any slope greater than 1 would indicate above grade level reading scores, which would certainly be acceptable as well. A slope less than 1 would indicate below grade level average scores, which would be unacceptable.

18. Grades.

Perhaps the best way to start is to discuss the type of graph that would have been useful. The admissions officer should have made a scatterplot with a coordinate for each freshman, matching each individual's SAT score with his or her respective GPA. Then, if the cloud of points was straight enough, the officer could have attempted to fit a linear model, and assessed its appropriateness and strength.

As is, the graph of combined SAT score versus mean Freshman GPA indicates, very generally, that higher SAT achievement is associated with higher mean Freshman GPA, but that's about it.

The first concern is the SAT scores. They have been grouped into categories. We cannot perform any type of regression analysis, because this variable is not quantitative. We don't even know how many students are in each category. There may be one student with an SAT score in the 1500s, and 300 students in the 1200s. On this graph, these possibilities are given equal weight!

Even if the SAT scores were at all useful to us, the GPAs given are averages, which would make the association appear stronger than it actually is.

Finally, a connected line graph isn't a useful model. It doesn't simplify the situation at all, and may, in fact, give the false impression that we could interpolate between the data points.

19. Heating.

a) The model predicts a decrease in \$2.13 in heating cost for an increase in temperature of 1° Fahrenheit. Generally, warmer months are associated with lower heating costs.

b) When the temperature is 0° Fahrenheit, the model predicts a monthly heating cost of \$133.

c) When the temperature is around 32° Fahrenheit, the predictions are generally too high. The residuals are negative, indicating that the actual values are lower than the predicted values.

d)

$\hat{C} = 133 - 2.13(Temp)$

$\hat{C} = 133 - 2.13(10)$

$\hat{C} = \$111.70$

According to the model, the heating cost in a month with average daily temperature 10° Fahrenheit is expected to be \$111.70.

e) The residual for a 10° day is approximately –\$6, meaning that the actual cost was \$6 less than predicted, or \$111.70 – \$6 = \$105.70.

f) The model is not appropriate. The residuals plot shows a definite curved pattern. The association between monthly heating cost and average daily temperature is not linear.

g) A change of scale from Fahrenheit to Celsius would not affect the relationship. Associations between quantitative variables are the same, no matter what the units.

20. Speed.

a) The model predicts that as speed increases by 1 mile per hour, the fuel economy is expected to decrease by 0.1 miles per gallon.

b) For this model, the y-intercept is the predicted mileage at a speed of 0 miles per hour. It's not possible to get 32 miles per gallon if you aren't moving.

c) The residuals are negative for the higher gas mileages. This means that the model is predicting higher than the actual mileage.

d)

$\hat{mpg} = 32 - 0.1(mph)$

$\hat{mpg} = 32 - 0.1(50)$

$\hat{mpg} = 27$

When a car is driven at 50 miles per hour, the model predicts mileage of 27 miles per gallon.

e)

$\hat{mpg} = 32 - 0.1(mph)$

$\hat{mpg} = 32 - 0.1(45)$

$\hat{mpg} = 27.5$

When a car is driven at 45 miles per hour, the model predicts mileage of 27.5 miles per gallon. From the graph, the residual at 27.5 mpg is +1. The actual gas mileage is 27.5 + 1 = 28.5 mpg.

f) The association between fuel economy and speed is probably quite strong, but not linear.

g) The linear model is not the appropriate model for the association between fuel economy and speed. The residuals plot has a clear pattern. If the linear model were appropriate, we would expect scatter in the residuals plot.

21. Interest rates.

a) $r = \sqrt{R^2} = \sqrt{0.774} = 0.88$. The correlation between rate and year is +0.88, since the scatterplot shows a positive association.

b) According to the model, interest rates during this period increased at about 0.25% per year, starting from an interest rate of about 0.64%in 1950.

c) The linear regression equation predicting interest rate from year is

$\hat{Rate} = 0.640282 + 0.247637(Year - 1950)$

$\hat{Rate} = 0.640282 + 0.247637(50)$

$\hat{Rate} = 13.022$

According to the model, the interest rate is predicted to be about 13% in the year 2000.

d) This prediction is not likely to have been a good one. Extrapolating 20 years beyond the final year in the data would be risky, and unlikely to be accurate.

22. Ages of couples 2003.

a) The correlation between age difference and year is $r = \sqrt{R^2} = \sqrt{0.751} \approx -0.8666$. The negative value is used since the scatterplot shows that the association is negative, strong, and linear.

b) The linear regression model that predicts age difference from year is: $(Men - \hat{W}omen) = 35.0617 - 0.016565(Year)$. This model predicts that each passing year is associated with a decrease of approximately 0.017 years in the difference between male and female marriage age. A more meaningful comparison might be to say that the model predicts a decrease of approximately 0.17 years in the age difference for every 10 years that pass.

c)

$(Men - \hat{W}omen) = 35.0617 - 0.016565(Year)$

$(Men - \hat{W}omen) = 35.0617 - 0.016565(2015)$

$(Men - \hat{W}omen) \approx 1.68$

According to the model, the age difference between men and women at first marriage is expected to be approximately 1.68 years. (This figure is very sensitive to the number of decimal places used in the model.)

d) The latest data point is before the year 2003. Extrapolating for 2015 is risky because it depends on the assumption that the trend in age at first marriage will continue in the same manner.

23. Interest rates revisited.

a) The values of R^2 are approximately the same, so the models fit comparably well, but they have very different slopes.

b) The model that predicts the interest rate on 3-month Treasury bills from the number of years since 1950 is $\hat{Rate} = 21.0688 - 0.356578(Year - 1950)$. This model predicts the interest rate to be 3.24%, a rate much lower than the prediction from the other model.

c) We can trust the newer prediction, since it is in the middle of the data used to generate the model. Additionally, the model accounts for 74.5% of the variability in interest rate.

d) Since 2020 is at least 15 years after the last year included in the newer model. It would be extremely risky to use this, or any, model to make a prediction that far into the future.

24. Ages of couples, again.

a) The data from the late 1800s to 1950 are high leverage points. Since they generally follow the same linear trend as the 1975 – 1998 data, those data points increase the correlation and the R^2 value.

b) The residuals plot shows a slight bend, so the linear model is probably appropriate, but we should proceed with caution.

c) For every 10 years that pass, the model predicts a decrease of approximately 0.30 years in average age difference at first marriage.

d) The *y*-intercept is the prediction of the model in year 0, over 2000 years ago. An extrapolation that far into the past is not meaningful. The earliest year for which we have data is 1975.

25. Gestation.

a) The association would be stronger if humans were removed. The point on the scatterplot representing human gestation and life expectancy is an outlier from the overall pattern and detracts from the association. Humans also represent an influential point. Removing the humans would cause the slope of the linear regression model to increase, following the pattern of the non-human animals much more closely.

b) The study could be restricted to non-human animals. This appears justifiable, since one could point to a number of environmental factors that could influence human life expectancy and gestation period, making them incomparable to those of animals.

c) The correlation is moderately strong. The model explains 72.2% of the variability in gestation period of non-human animals.

d) For every year increase in life expectancy, the model predicts an increase of approximately 15.5 days in gestation period.

e)

$$\hat{Gest} = -39.5172 + 15.4980(LifEx)$$
$$\hat{Gest} = -39.5172 + 15.4980(20)$$
$$\hat{Gest} \approx 270.4428$$

According to the linear model, monkeys with a life expectancy of 20 years are expected to have gestation periods of about 270.5 days. Care should be taken when assessing the accuracy of this prediction. First of all, the residuals plot has not been examined, so the appropriateness of the model is questionable. Second, it is unknown whether or not monkeys were included in the original 17 non-human species studied. Since monkeys and humans are both primates, the monkeys may depart from the overall pattern as well.

26. Swim the lake 2006.

a) Only 1.3% of the variability in lake swim times is accounted for by the linear model.

b) The slope of the regression, 5.14, means that the model predicts that lake swim times are increasing by about 5.14 minutes per year. This means that lake swimmers are generally getting slower. However, this model has very weak predicting power, and an outlier, so we shouldn't put too much faith in our prediction.

c) Removing this outlier is probably a good idea, since it doesn't belong with the other data points, but its removal probably wouldn't change the regression much. The fact that the point has a large residual indicates that it didn't have much leverge. If it had leverage, it would have dominated the regression, and had a small residual. It would be nice to have a scatterplot to look at, in addition to the residuals plot. There could be other outliers that don't show up in the residuals plot.

27. Elephants and hippos.

a) Hippos are more of a departure from the pattern. Removing that point would make the association appear to be stronger.

b) The slope of the regression line would increase, pivoting away from the hippos point.

c) Anytime data points are removed, there must be a justifiable reason for doing so, and saying, "I removed the point because the correlation was higher without it" is not a justifiable reason.

d) Elephants are an influential point. With the elephants included, the slope of the linear model is 15.4980 days gestation per year of life expectancy. When they are removed, the slope is 11.6 days per year. The decrease is significant.

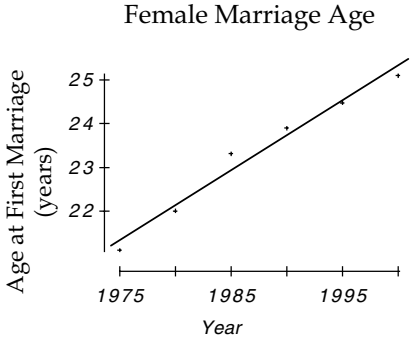
28. Another swim 2006.

a) The smaller value of s_e means that errors in prediction are smaller for this model than the original model.

b) The regression accounts for only 4.1% of the variation in lake swim times, but it appears that Lake Ontario swimmers are getting slower, at a rate of about 6.2 seconds per year.

29. Marriage age 2003 revisited.

a) Modeling decisions may vary, but the important idea is using a subset of the data that allows us to make an accurate prediction for the year in which we are interested. We might model a subset to predict the marriage age in 2010, and model another subset to predict the marriage age in 1911.



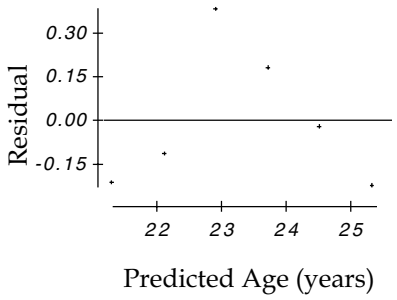
In order to predict the average marriage age of American women in 2010, use the data points from the most recent trend only. The data points from 1975 – 2000 look straight enough to apply the linear regression model.

Regression output from a computer program is given below, as well as a residuals plot.

Dependent variable is: **Age**
No Selector
R squared = 97.5% R squared (adjusted) = 96.9%
s = 0.2684 with 6 - 2 = 4 degrees of freedom

Source	Sum of Squares	df	Mean Square	F-ratio
Regression	*11.2801*	*1*	*11.2801*	*157*
Residual	*0.288190*	*4*	*0.072048*	

Variable	Coefficient	s.e. of Coeff	t-ratio	prob
Constant	*-295.819*	*25.51*	*-11.6*	*0.0003*
Year	*0.160571*	*0.0128*	*12.5*	*0.0002*



The linear model used to predict average female marriage age from year is: $\hat{Age} = -295.819 + 0.161(Year)$. The residuals plot shows a curved pattern (although the number of points used is small, so this may just be random variation), and the value of R^2 is high. 97.5% of the variability in average female age at first marriage is accounted for by variability in the year. The model predicts that each year that passes is associated with an increase of 0.161 years in the average female age at first marriage.

$$\hat{Age} = -295.819 + 0.161(Year)$$
$$\hat{Age} = -295.819 + 0.161(2010)$$
$$\hat{Age} \approx 27.79$$

According to the model, the average age at first marriage for women in 2010 will be 27.79 years old. Care should be taken with this prediction, however. It represents an extrapolation of 10 years beyond the highest year, and the residuals plot shows a pattern.

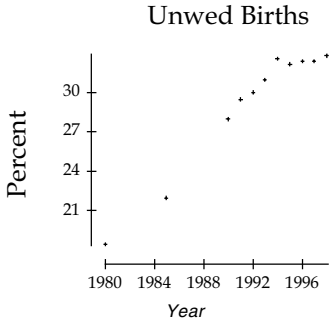
b) This prediction is for a year that is 10 years higher than the highest year for which we have an average female marriage age. Don't place too much faith in this extrapolation.

c) An extrapolation of more than 50 years into the future would be absurd. There is no reason to believe the trend would continue. In fact, given the situation, it is very unlikely that the pattern would continue in this fashion. The model given in part a) predicts that the average marriage age will be 34.27 years. Realistically, that seems quite high.

30. Unwed births.

The analysis that follows is one of several good models that may be used to predict the percentage of unwed births. The important feature to recognize is that these data consist of two distinct trends. Your modeling decisions may vary slightly from these, but that is fine, as long as those decisions are justified.

A scatterplot (at the right) of year vs. percent of unmarried births shows two distinct trends. From 1980 to 1994, there is a



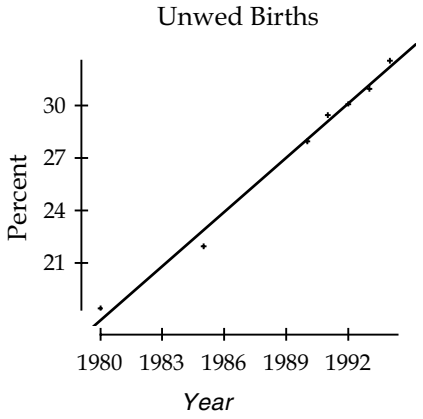
strong, positive linear association between year and percent of unmarried births. For the years 1995 to 1998, there is also a strong, positive, linear association, but the percent of unmarried births increases much more slowly from year to year, almost to the point of being flat. Two linear models will fit the relationship well.

Model I (1980—1994)

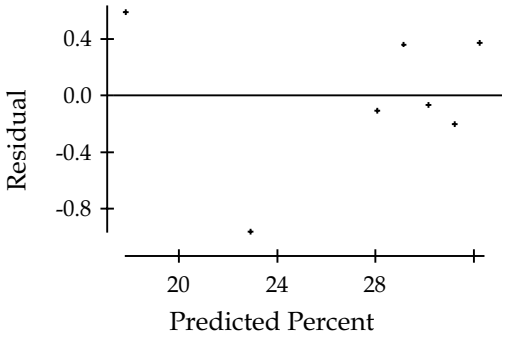
Dependent variable is: **Births**
No Selector
R squared = 99.0% R squared (adjusted) = 98.8%
s = 0.5645 with 7 - 2 = 5 degrees of freedom

Source	Sum of Squares	df	Mean Square	F-ratio
Regression	*160.621*	*1*	*160.621*	*504*
Residual	*1.59314*	*5*	*0.318628*	

Variable	Coefficient	s.e. of Coeff	t-ratio	prob
Constant	*-2021.41*	*91.25*	*-22.2*	*≤ 0.0001*
Year	*1.02991*	*0.0459*	*22.5*	*≤ 0.0001*



$\hat{\%} = -2021.41 + 1.0299(year)$ is a good model for the years 1980 – 1994. A scatterplot of the relationship, with regression line, is shown above and to the right. $R^2 = 99\%$, so the model explains 99% of the variability in percent of unmarried births. The residuals plot (at the right) is scattered, indicating an appropriate model.

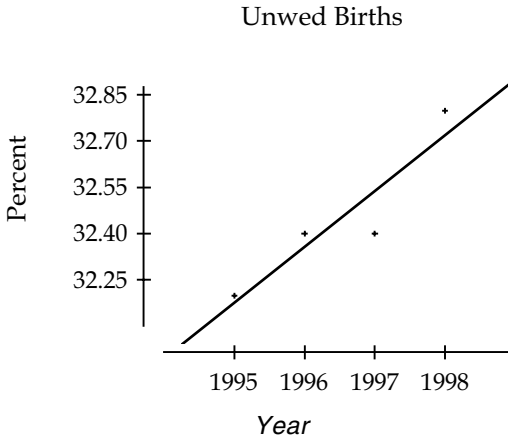


Model II (1995—1998)

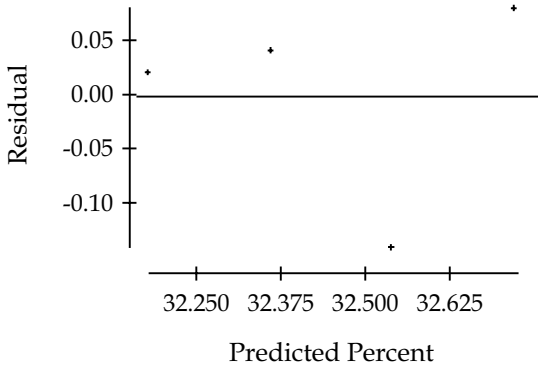
Dependent variable is: **Births**
No Selector
R squared = 85.3% R squared (adjusted) = 77.9%
s = 0.1183 with 4 - 2 = 2 degrees of freedom

Source	Sum of Squares	df	Mean Square	F-ratio
Regression	*0.162000*	*1*	*0.162000*	*11.6*
Residual	*0.028000*	*2*	*0.014000*	

Variable	Coefficient	s.e. of Coeff	t-ratio	prob
Constant	*-326.920*	*105.6*	*-3.09*	*0.0905*
Year	*0.180000*	*0.0529*	*3.40*	*0.0766*

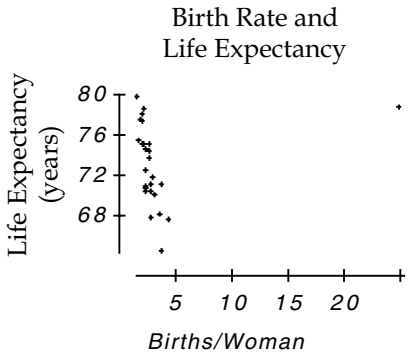


$\hat{\%} = -326.92 + 0.18(year)$is a good model for the years 1995 – 1998. Although not as accurate as the first model, R^2= 85.3%, which means that the model accounts for 85.3% of the variability in percent of unmarried births. The residuals plot is scattered, indicating an appropriate model. Great care should be taken in using this model for predictions, since it was developed from only four data points. The slope of the regression line indicates that for each year that passes, the model predicts an increase of only 0.18% unmarried births. The rate may have actually leveled out.



31. Life expectancy 2004.

a) The scatterplot of births per woman and life expectancy is at the right. The association is strong, linear, and negative. Countries with higher life expectancies tend to have a lower number of births per woman. There is one outlier, Costa Rica, with an unreasonable 24.9 births per woman, and a life expectancy of 78.7 years.



b) Costa Rica has a birth rate of 24.9 births per woman, which common sense would indicate is impossible. There is ample justification to leave that point out of all further calculations. Probably, a decimal point was incorrectly shifted from 2.49 births per woman, but there is no way to be sure. Omit this strange data point!

c)

Dependent variable is: **Life Expectancy**
No Selector
R squared = 63.3% R squared (adjusted) = 61.7%
s = 2.387 with 25 - 2 = 23 degrees of freedom

Source	Sum of Squares	df	Mean Square	F-ratio
Regression	*226.037*	*1*	*226.037*	*39.7*
Residual	*131.023*	*23*	*5.69666*	

Variable	Coefficient	s.e. of Coeff	t-ratio	prob
Constant	*84.4971*	*1.902*	*44.4*	*≤ 0.0001*
Births/Wo...	*-4.43993*	*0.7048*	*-6.30*	*≤ 0.0001*

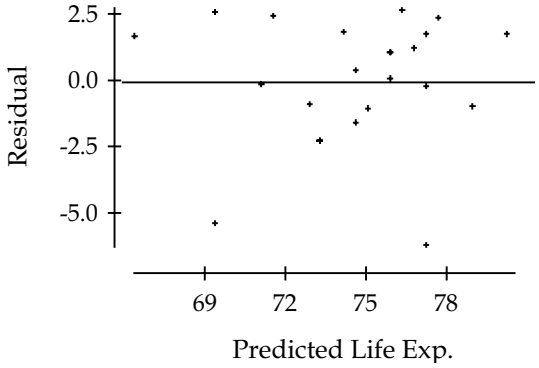
Without Costa Rica, R^2= 63.3%, so $r = \sqrt{R^2} = \sqrt{0.633} = -0.796$.

63.3% of the variability in life expectancy is explained by variability in the number of births per year.

(If you assumed that 24.9 births per woman was a data entry error meant to be 2.49 births per woman, the revised value of R^2 is 59.5%, making $r = \sqrt{R^2} = \sqrt{0.595} = -0.771$.)

d) The linear model that predicts a country's life expectancy from the number of births per woman (without Costa Rica) is $\hat{LifeExp} = 86.8137 - 4.35623(Births)$. (If the number of births per woman in Costa Rica is changed to 2.49, the linear model that predicts a country's life expectancy is $\hat{LifeExp} = 83.8394 - 4.49366(Births)$

e) The linear model, without Costa Rica, is an appropriate model. The residuals plot shows no pattern. (If the number of births per woman is changed to 2.49, the linear model is also appropriate. The revised residuals plot shows no pattern.)



f) According to the model, each additional birth per woman is expected to correspond with a decrease in life expectancy of approximately 4.36 years (or 4.49 years, using the revised model). When the number of births per woman is 0, the model predicts that the life expectancy will be 86.8 years (or 84.8 years, using the revised model). This figure is an extrapolation below the range of the data, and doesn't hold any meaning.

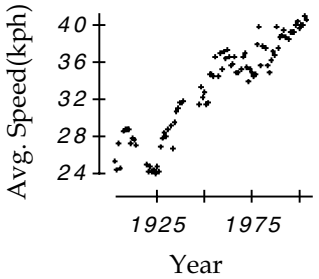
g) The government leaders should not suggest that women have fewer children in order to raise the life expectancy. Although there is evidence of an association between the birth rate and life expectancy, this does not mean that one causes the other. There may be lurking variables involved, such as economic conditions, social factors, or level of health care.

32. Tour de France 2007.

a) The association between average speed and year is positive, moderate, but not quite linear. Generally, average speed of the winner has been increasing over time. There are several periods where the relationship is curved, but since 1950, the relationship has been much more linear. There are no races between 1915 and 1918 or between 1940 and 1947, presumably because of the two World Wars in Europe at the times.

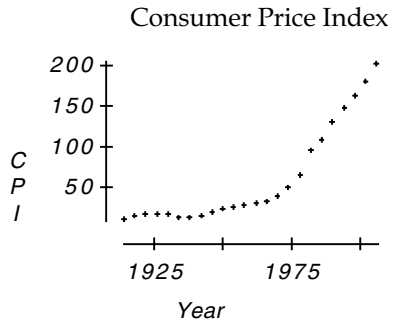
b) $\hat{Avgspeed} = -275.227 + 0.158(Year)$

c) The conditions for regression are not met. Although the variables are quantitiative, and there are no outliers, the relationship is not straight enough in the early part of the 20th century to fit a regression line.

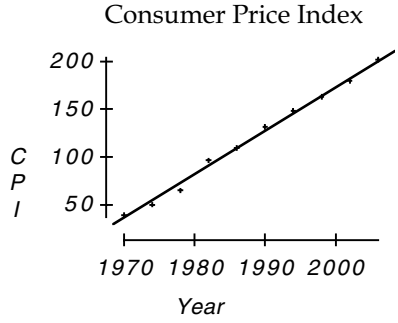


33. Inflation 2006.

a) The trend in Consumer Price Index is strong, non-linear, and positive. Generally, CPI has increased over the years, but the rate of increase has become much greater since approximately 1970. Other characteristics include fluctuations in CPI in the years prior to 1950.

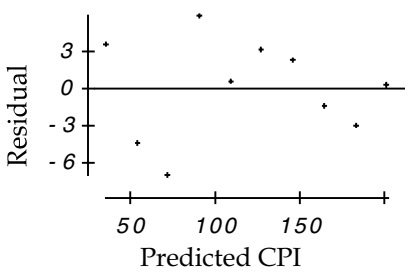


b) In order to effectively predict the CPI in 2016, use only the most recent trend. The trend since 1970 is straight enough to apply the linear model. Prior to 1970, the trend is radically different from that of recent years, and is of no use in predicting CPI for 2016.



The linear model that predicts CPI from year is $\hat{CPI} = -9052.42 + 4.61(year)$. $R^2 = 99.5\%$, meaning that the model predicts 99.5% of the variability in CPI. The residuals plot shows no pattern, so the linear model is appropriate. According to the model, the CPI is expected to increase by \$4.61 each year, for 1970—2006.

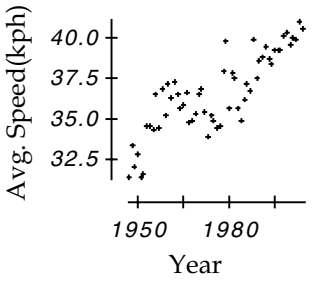
$$\hat{CPI} = -9052.42 + 4.61(year)$$
$$\hat{CPI} = -9052.42 + 4.61(2016)$$
$$\hat{CPI} = 241.34$$



As with any model, care should be taken when extrapolating. If the pattern continues, the model predicts that the CPI in 2006 will be approximately \$241.34.

34. Second stage 2007.

a) There is still some curving in the beginning of the period, but the relationship is straighter. The new linear model is $\hat{Avgspeed} = -202.852 + 0.121(Year)$.



b) According to the linear model, the average winning speed increases by about 0.121 kph per year.

c) Hinault's 1979 time has a residual of 2.92 kph. He raced much faster than the model would predict. But Armstrong's 2001 time was actually below the model's prediction and had a residual of –0.327 kph. Hinault's performance was more remarkable for its era.