Moodle Questions

Rebekah Robinson

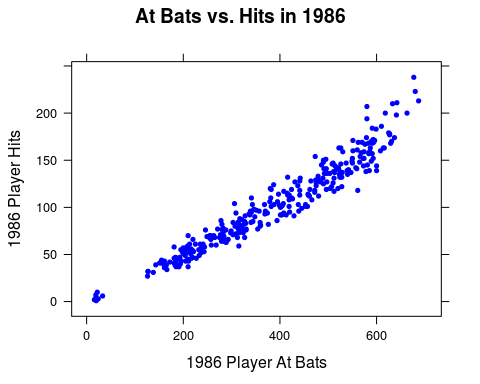
Tuesday, June 17, 2014

PART 1 # Baseball Scatterplot Code

Before starting this question, make sure to require the necessary package and load the appropriate dataset. Run the following code.

require(vcd)  
data(Baseball)  
View(Baseball)  
help(Baseball)

This data frame contains observations for 322 baseball players on 25 variables using data from 1986. **At bats**, atbat86, count the number of official plate appearances by a hitter. A **hit**, hits86, is credited to a hitter is they successfully make it to first base after they hit the ball into fair territory. The scatterplot is shown below.



Which of the following lines of code created this scatterplot?

xyplot(hits86~atbat86, data=Baseball)

xyplot(atbat86~hits86,data=Baseball)

xyplot(hits86~atbat86, groups=league86, data=Baseball, auto.key=TRUE)

xyplot(atbat86~hits86, groups=league86, data=Baseball, auto.key=TRUE)

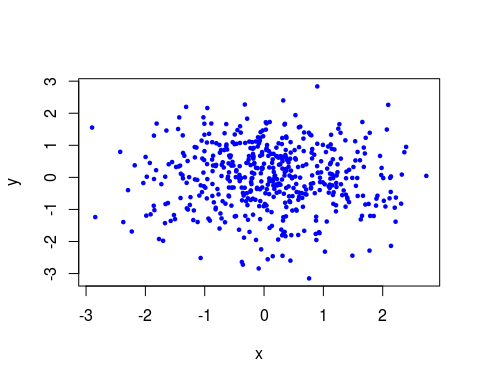
xyplot(hits86~atbat86|league86, data=Baseball)

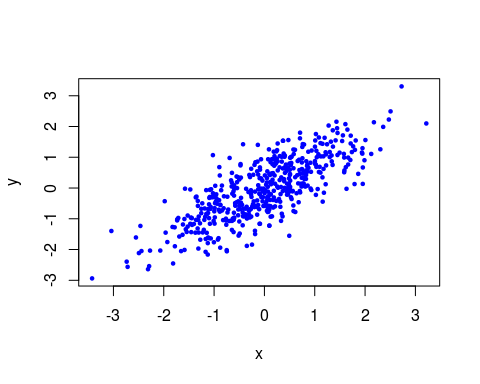
xyplot(atbat86~hits86|league86, data=Baseball)

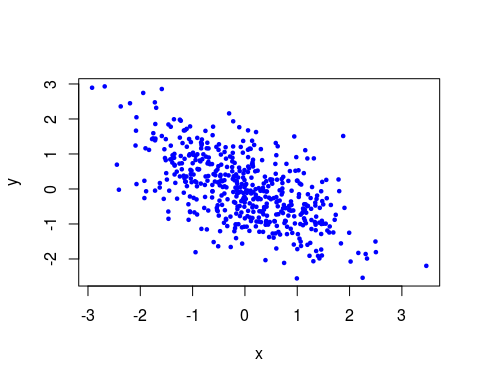
Just by looking at the scatterplot, which of the following is the *best* description of the association?

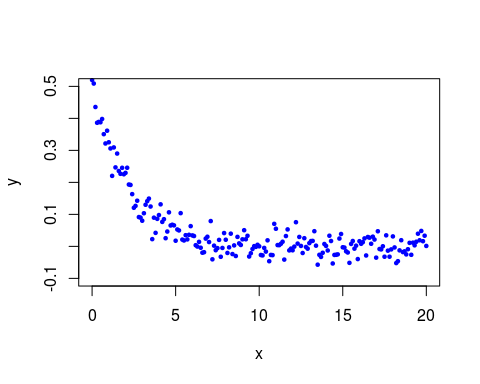
# Association Matching

Match each of the following scatterplots with the best description of the type of association between the variables.



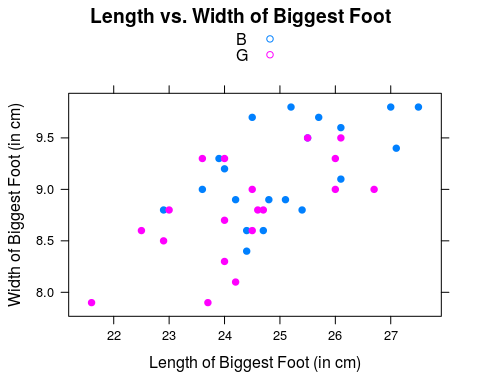






# KidsFeet Scatterplot Grouped by Sex

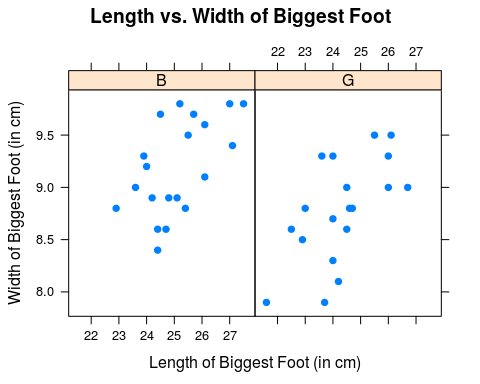
The *KidsFeet* dataframe contains data collected on 39 fourth grade students in Ann Arbor, MI, in October 1997. Two of the measurements taken on the children was the length, (*length*), and width, (*width*), of their longest foot. Based on the scatterplot shown below, there appears to be a positive linear association between the length and width of a fourth grader's biggest foot that holds even when the child's sex is accounted for. Which of the following lines of code created this scatterplot?



xyplot(width~length,groups=sex,data=KidsFeet, auto.key=TRUE)  
xyplot(length~width, groups=sex, data=KidsFeet, auto.key=TRUE)  
xyplot(width~length|sex,data=KidsFeet)  
xyplot(length~width|sex, data=KidsFeet)  
xyplot(width~length,groups=sex,data=KidsFeet)  
xyplot(length~width, groups=sex, data=KidsFeet)  
xyplot(width~length, data=KidsFeet)  
xyplot(length~width, data=KidsFeet)

# KidsFeet Parallel Scatterplots by Sex

The *KidsFeet* dataframe contains data collected on 39 fourth grade students in Ann Arbor, MI, in October 1997. Two of the measurements taken on the children was the length, (*length*), and width, (*width*), of their longest foot. Based on the scatterplot shown below, there appears to be a positive linear association between the length and width of a fourth grader's biggest foot that holds even when the child's sex is accounted for. Which of the following lines of code created this scatterplot?



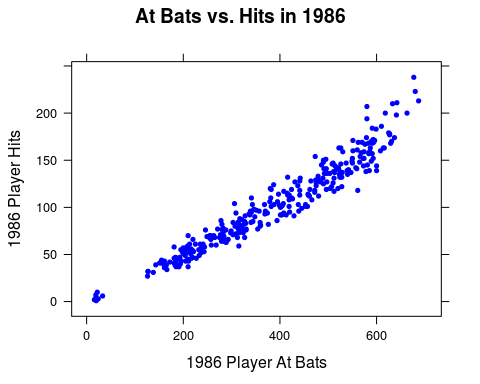
xyplot(width~length,groups=sex,data=KidsFeet, auto.key=TRUE)  
xyplot(length~width, groups=sex, data=KidsFeet, auto.key=TRUE)  
xyplot(width~length|sex,data=KidsFeet)  
xyplot(length~width|sex, data=KidsFeet)  
xyplot(width~length,groups=sex,data=KidsFeet)  
xyplot(length~width, groups=sex, data=KidsFeet)  
xyplot(width~length, data=KidsFeet)  
xyplot(length~width, data=KidsFeet)

PART 2 # Baseball Correlation

Before starting this question, make sure to require the necessary package and load the appropriate dataset. Run the following code.

require(vcd)  
data(Baseball)  
View(Baseball)  
help(Baseball)

This data frame contains observations for 322 baseball players on 25 variables using data from 1986. **At bats**, atbat86, count the number of official plate appearances by a hitter. A **hit**, hits86, is credited to a hitter is they successfully make it to first base after they hit the ball into fair territory. The scatterplot is shown below.

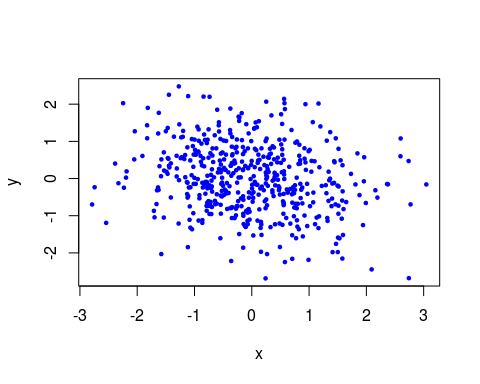


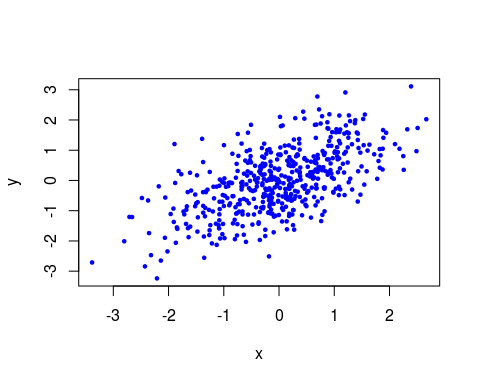
Just by looking at the scatterplot, which of the following is the most likely value for the correlation coefficient, ?

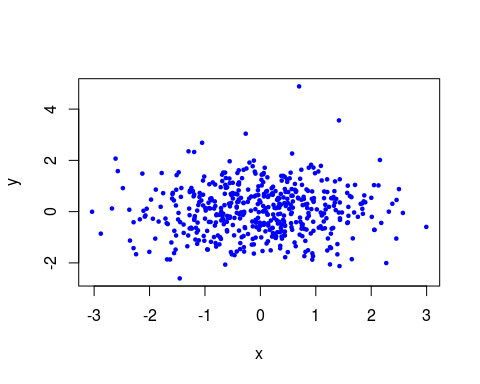
+1 0 -1 +0.25 +1.8 -0.009

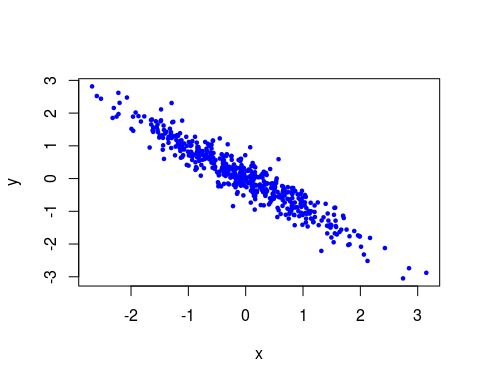
# Correlation Coefficient Matching

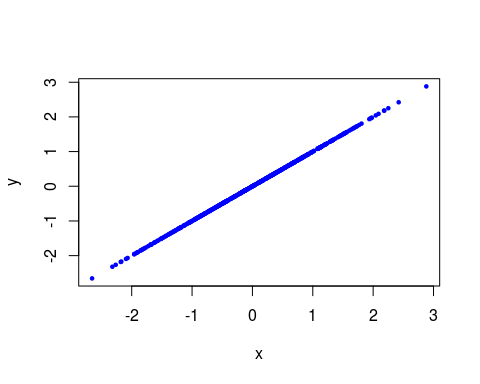
Match the following scatterplots with it's most likely value for the correlation coefficient, .

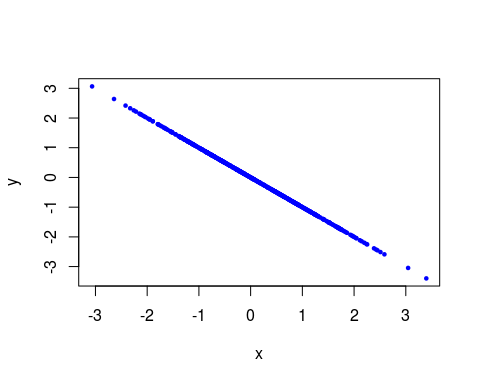












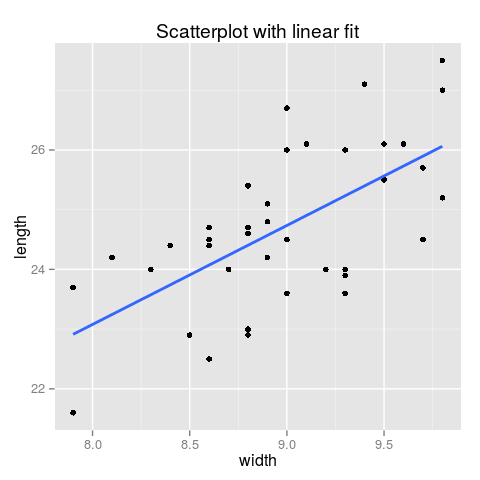
# KidsFeet Regression Analysis Code

The *KidsFeet* dataframe contains data collected on 39 fourth grade students in Ann Arbor, MI, in October 1997. Two of the measurements taken on the children were the length in centimeters, (*length*), and width in centimeters, (*width*), of their longest foot. This data could be used to answer the following

*Research Question:* How is the width of a fourth-grade student's foot related to the length?

Which of the following lines of code resulted in the following plot and regression analysis? **Hint: More than one choice may be correct.**

##   
## Linear Regression  
##   
## Correlation coefficient r = 0.6411   
##   
## Equation of Regression Line:  
##   
## length = 9.8172 + 1.6576 \* width   
##   
## Residual Standard Error: s = 1.0248   
## R^2 (unadjusted): R^2 = 0.411



lmGC(width~length,data=KidsFeet,graph=TRUE)  
lmGC(width~length,data=KidsFeet)  
lmGC(length~width,data=KidsFeet,graph=TRUE)  
lmGC(length~width,data=KidsFeet)  
  
KidsFeetMod<-lmGC(length~width,data=KidsFeet,graph=TRUE)  
KidsFeetMod  
  
KidsFeetMod<-lmGC(width~length,data=KidsFeet,graph=TRUE)  
KidsFeetMod  
  
KidsFeetMod<-lmGC(length~width,data=KidsFeet)  
KidsFeetMod  
  
KidsFeetMod<-lmGC(width~length,data=KidsFeet)  
KidsFeetMod

# KidsFeet Regression Line Slope Interpretation

The *KidsFeet* dataframe contains data collected on 39 fourth grade students in Ann Arbor, MI, in October 1997. Two of the measurements taken on the children were the length in centimeters, (*length*), and width in centimeters, (*width*), of their longest foot. This data could be used to answer the following

*Research Question*: How is the width of a fourth-grade student's foot related to the length?

Which of the following sentences interprets the **slope** of the regression line in terms of how the length of a student's foot changes as the width increases?

##   
## Linear Regression  
##   
## Correlation coefficient r = 0.6411   
##   
## Equation of Regression Line:  
##   
## length = 9.8172 + 1.6576 \* width   
##   
## Residual Standard Error: s = 1.0248   
## R^2 (unadjusted): R^2 = 0.411

For every 1 centimeter increase in the width of a fourth-grader's foot, the predicted length increases by 1.6576 centimeters.

For every 1 inch increase in the width of a fourth-grader's foot, the predicted length increases by 1.6576 inches.

For every 1 inch increase in the length of a fourth-grader's foot, the predicted width increases by 1.6576 inches.

For every 1 centimeter increase in the width of a fourth-grader's foot, the predicted length decreases by 1.6576 centimeters.

For every 1 inch increase in the width of a fourth-grader's foot, the predicted length increases by 9.8172 inches.

# KidsFeet Regression Line Intercept

The *KidsFeet* dataframe contains data collected on 39 fourth grade students in Ann Arbor, MI, in October 1997. Two of the measurements taken on the children were the length in centimeters, (*length*), and width in centimeters, (*width*), of their longest foot. This data could be used to answer the following

*Research Question:* How is the width of a fourth-grade student's foot related to the length?

Which of the following is the -intercept of the regression line?

##   
## Linear Regression  
##   
## Correlation coefficient r = 0.6411   
##   
## Equation of Regression Line:  
##   
## length = 9.8172 + 1.6576 \* width   
##   
## Residual Standard Error: s = 1.0248   
## R^2 (unadjusted): R^2 = 0.411

9.8172 0.6411 1.6576 1.0248 0.411

# KidsFeet Predictions

The *KidsFeet* dataframe contains data collected on 39 fourth grade students in Ann Arbor, MI, in October 1997. Two of the measurements taken on the children were the length in centimeters, (*length*), and width in centimeters, (*width*), of their longest foot. This data could be used to answer the following

*Research Question:* How is the width of a fourth-grade student's foot related to the length?

Which of the following is the -intercept of the regression line?

##   
## Linear Regression  
##   
## Correlation coefficient r = 0.6411   
##   
## Equation of Regression Line:  
##   
## length = 9.8172 + 1.6576 \* width   
##   
## Residual Standard Error: s = 1.0248   
## R^2 (unadjusted): R^2 = 0.411

Use the following regression analysis to answer the following question:

What is the predicted foot length (in centimeters) for a fourth grader with a foot that is 8.8 centimeters wide? Answers are all rounded to the nearest tenth.

##   
## Linear Regression  
##   
## Correlation coefficient r = 0.6411   
##   
## Equation of Regression Line:  
##   
## length = 9.8172 + 1.6576 \* width   
##   
## Residual Standard Error: s = 1.0248   
## R^2 (unadjusted): R^2 = 0.411

24.4 24.8 -4.8 9.8 24.2

# KidsFeet Residuals

The *KidsFeet* dataframe contains data collected on 39 fourth grade students in Ann Arbor, MI, in October 1997. Two of the measurements taken on the children were the length in centimeters, (*length*), and width in centimeters, (*width*), of their longest foot. This data could be used to answer the following

*Research Question:* How is the width of a fourth-grade student's foot related to the length?

Allison is a fourth-grade student. Suppose the width of a Allison's dominant foot is 9.25 centimeters wide. Run the following lines of code below to see what the regression line predicts the length of her foot to be, in centimeters.

KidsFeetMod<-lmGC(length~width,data=KidsFeet)  
predict(KidsFeetMod,9.25)

Suppose that Allison's foot is really 26.5 centimeters long. Which of the following numbers is the **residual** difference between the actual length and the predicted length of Allison's foot?

1.3497 -1.3497 17 -17 15.6503 -15.6503

# Explanatory and Response Matching

For the following research question, which variable should be the explanatory variable and which should be the response variable?

*Research Question:* For MAT 111, how are midterm exam scores related to final exam scores?

# KidsFeet Squared Correlation

The *KidsFeet* dataframe contains data collected on 39 fourth grade students in Ann Arbor, MI, in October 1997. Two of the measurements taken on the children were the length in centimeters, (*length*), and width in centimeters, (*width*), of their longest foot. This data could be used to answer the following

*Research Question:* How is the width of a fourth-grade student's foot related to the length?

##   
## Linear Regression  
##   
## Correlation coefficient r = 0.6411   
##   
## Equation of Regression Line:  
##   
## length = 9.8172 + 1.6576 \* width   
##   
## Residual Standard Error: s = 1.0248   
## R^2 (unadjusted): R^2 = 0.411

The statistic for these data is 0.411. Which of the following sentences is the correct interpretation of this statistic?

41.1% of the children in the study have wider than average feet.

41.1% of the variation in the length of children's feet is explained by the linear relationship between the width and the length of a child's foot.

41.1% of the variation in the width of children's feet is explained by the linear relationship between the width and the length of a child's foot.

For every additional centimeter wide a student's foot is, the length of their foot increases by 0.411.

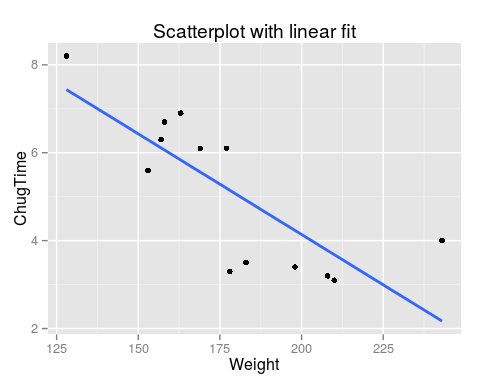
If the width of a child's dominant foot is 0 centimeters, the predicted length would be 0.411 centimeters.

# Regression Analysis Matching

A researcher gathered data on 13 college-aged males chugging a 12-ounce can of a certain beverage. The researcher recorded the weight (**Weight**), in pounds, and the chugtime (**ChugTime**), in seconds, of each subject.

The regression analysis, along with a plot, is shown below.

##   
## Linear Regression  
##   
## Correlation coefficient r = -0.7879   
##   
## Equation of Regression Line:  
##   
## ChugTime = 13.2975 + -0.0458 \* Weight   
##   
## Residual Standard Error: s = 1.1234   
## R^2 (unadjusted): R^2 = 0.6208



Match each of the following numbers with their correct description.

-0.7879 The numerical description of the strong negative relationship between a college male's weight and the amount of time it takes for him to chug a 12 ounce beverage.

13.3 A college-aged male weighing 0 pounds is predicted to take 13.2975 seconds to chug a 12 ounce beverage.

-0.0458 For every one pound increase in the weight of a college male, the predicted chug time of a 12 ounce beverage decreases by 0.0458 seconds.

1.123 A measurement of the variation in the data from the regression line that is affected by a change in scale of the response variable.

0.6208 The proportion of variation in the chug times that is explained by the weights of the college males.

PART 3

# KidsFeet Error 1

The *KidsFeet* dataframe contains data collected on 39 fourth grade students in Ann Arbor, MI, in October 1997. Two of the measurements taken on the children were the length in centimeters, (*length*), and width in centimeters, (*width*), of their longest foot. This data could be used to answer the following

*Research Question:* How is the width of a fourth-grade student's foot related to the length?

##   
## Linear Regression  
##   
## Correlation coefficient r = 0.6411   
##   
## Equation of Regression Line:  
##   
## length = 9.8172 + 1.6576 \* width   
##   
## Residual Standard Error: s = 1.0248   
## R^2 (unadjusted): R^2 = 0.411

True or False? Sam is a senior in high school whose dominant foot measures 10.8 centimeters wide. The regression line shown above could be used to make a prediction about how long Sam's foot is.

# KidsFeet Error 2

True or False? For school children, foot size is correlated with reading skills. Thus, learning new words makes children's feet grow bigger.

# KidsFeet Error Matching

The *KidsFeet* dataframe contains data collected on 39 fourth grade students in Ann Arbor, MI, in October 1997. Two of the measurements taken on the children were the length in centimeters, (*length*), and width in centimeters, (*width*), of their longest foot. This data could be used to answer the following

*Research Question:* How is the width of a fourth-grade student's foot related to the length?

##   
## Linear Regression  
##   
## Correlation coefficient r = 0.6411   
##   
## Equation of Regression Line:  
##   
## length = 9.8172 + 1.6576 \* width   
##   
## Residual Standard Error: s = 1.0248   
## R^2 (unadjusted): R^2 = 0.411

Match the following statements with the *type of error* they commit.

Extrapolation: Sam is a senior in high school whose dominant foot measures 10.8 centimeters wide. The regression line shown above could be used to make a prediction about how long Sam's foot is.

Association does not imply causation: For school children, foot size is correlated with reading skills. Thus, learning new words makes children's feet grow bigger.

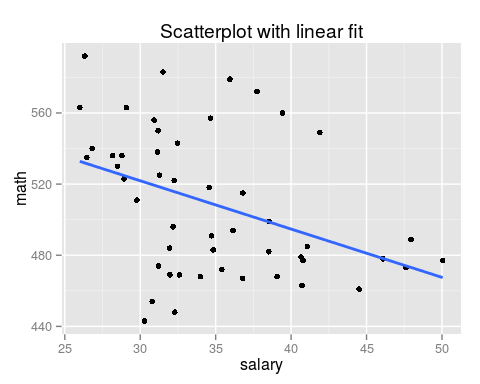
# SAT Misinterpretations

Data was gathered on the SAT (a college entrance exam) for each of the fifty states in the U.S. A research was interested in answering the following

*Research Question:* How is a state's average annual teacher salary, in thousands of dollars, related to that state's average student performance on the math portion of the SAT?

Study the following regression analysis.

##   
## Linear Regression  
##   
## Correlation coefficient r = -0.4013   
##   
## Equation of Regression Line:  
##   
## math = 603.3648 + -2.7157 \* salary   
##   
## Residual Standard Error: s = 37.2068   
## R^2 (unadjusted): R^2 = 0.1611



Which of the following statements is an **incorrect** interpretation of the analysis? There may be more than one correct answer.

The correlation between average annual teacher salary and average SAT math subscore is -0.4013. (Correct)

Average annual teacher salary and average SAT math subscore are negatively correlated. (Correct)

States that pay lower teacher salaries tend to produce students to score higher on the math portion of the SAT. (Correct)

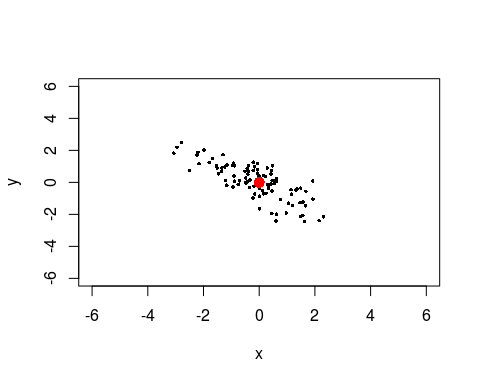
Paying teachers lower salaries causes students to perform better on the math portion of the SAT. (Incorrect)

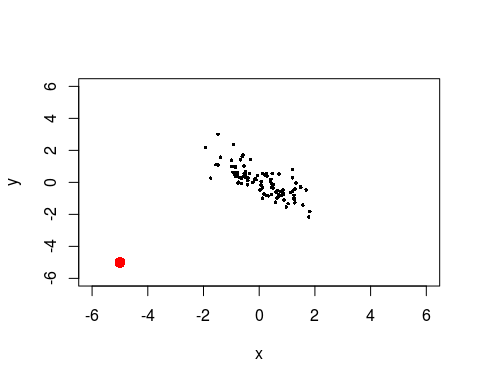
If a state that currently pays an average teacher salary of $40,000 per year would raise teacher pay by $10,000 next year, the average math subscore on the SAT would go down by 27 points. (Incorrect)

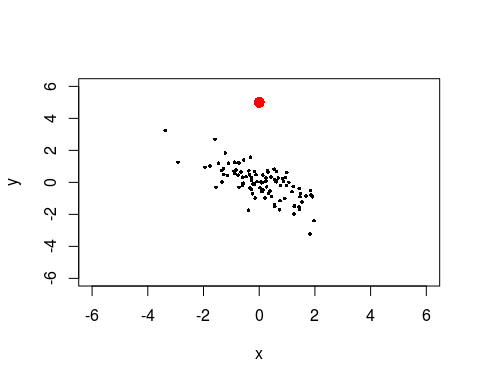
This analysis provides evidence that paying teachers $100,000 per year would be detrimental to student's math ability. An average teacher salary of $100,000 would result in an average math subscore of 332 on the SAT. (Incorrect)

# Influential Observation

Which of the following placement of the red point in the scatterplots below would have the most influence over the regression line?





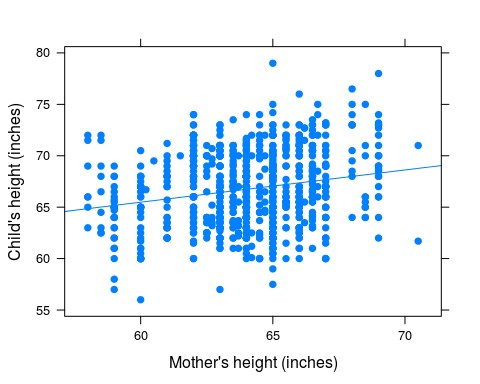


# Prediction

We'll look at the Galton data frame from the mosaicData package:

data(Galton)  
View(Galton)  
help(Galton)

Here's a scatterplot of the heights of the mothers and the heights of their children:



We would like to use the data to predict the height of a child from the height of her mother. As described in class and in the Course Notes, use the predict() function to predict the height of a child (the time of Galtons' study) if the child's mother was known to be 64 inches tall. Record your answer (without units) as a decimal number, correct to at least two decimal places.

66.73

motherMod <- lmGC(height~mother,data=Galton)  
predict(motherMod,x=64)

## Predict height is about 66.73,  
## give or take 3.513 or so for chance variation.

# Give or Take

This question refers to the Galton data.

For the height of a child whose mother was known to be 64 inches tall, what is the prediction standard error? Record your answer (without units) as a decimal number, correct to at least two decimal places.

3.513

# A 95%-Prediction Interval

This question refers to the Galton data.

Find a 95%-prediction interval the height of a child whose mother was known to be 64 inches tall. Record the upper bound of your answer (without units) as a demcimal number, correct to at least two decimal places.

motherMod <- lmGC(height~mother,data=Galton)  
predict(motherMod,x=64,level=0.95)

## Predict height is about 66.73,  
## give or take 3.513 or so for chance variation.  
##   
## 95%-prediction interval:  
## lower.bound upper.bound   
## 59.839140 73.629370

# A Cubic Fit

The following R command will import a new data frame called fc:

fc <- read.csv("/mat111/Additional\_Datasets/fuelConsumption.csv")

Here is the data:

fc

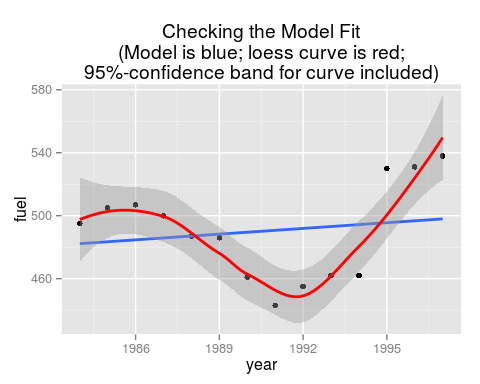
## year fuel  
## 1 1984 495  
## 2 1985 505  
## 3 1986 507  
## 4 1987 500  
## 5 1988 487  
## 6 1989 486  
## 7 1990 461  
## 8 1991 443  
## 9 1992 455  
## 10 1993 462  
## 11 1994 462  
## 12 1995 530  
## 13 1996 531  
## 14 1997 538

The variable **fule** gives the mean number of gallons used per car in the U.S., for each of the given years. we would like to use the data to predcit the mean number fo gallons used per car in the year 1998, the first year not in the data.

We make a check a model with a linear fit:

lmGC(fuel~year,data=fc,check=TRUE)

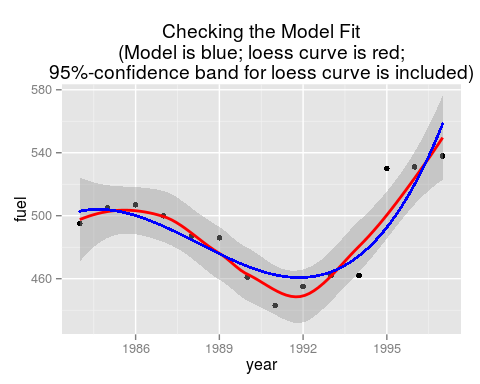
##   
## Linear Regression  
##   
## Correlation coefficient r = 0.1658   
##   
## Equation of Regression Line:  
##   
## fuel = -1915.956 + 1.2088 \* year   
##   
## Residual Standard Error: s = 31.304   
## R^2 (unadjusted): R^2 = 0.0275



Then we make aand check a model with a third-degree (cubic) fit:

polyfitGC(fuel~year,data=fc,degree=3,check=TRUE)

## Polynomial Regression, Degree = 3  
##   
## Residual Standard Error: s = 16.7889   
## R^2 (unadjusted): R^2 = 0.7669



Choose the code below that gives the most reliable prediction. along with an explanation of why that particular prediction is better.

Option A:

fc1 <- lmGC(fuel~year,data=fc) predict(fc1,x=1998)

This is better because the regression line is used to predict y values from knonw x values.

Option B:

fc1 <- lmGC(year~fuel,data=fc) predict(fc1,x=1998)

This is better because the regression line is used to predict y values from knonw x values.

Option C:

fc3 <- polyfitGCGC(fuel~year,data=fc,degree=3) predict(fc3,x=1998)

This is better because the regression line strays outside the 95%-confidence band around the loess curve, whereas the cubic fit stays within the band.

Option D:

fc3 <- polyfitGCGC(year~fuel,data=fc,degree=3) predict(fc3,x=1998)

This is better because the regression line strays outside the 95%-confidence band around the loess curve, whereas the cubic fit stays within the band.