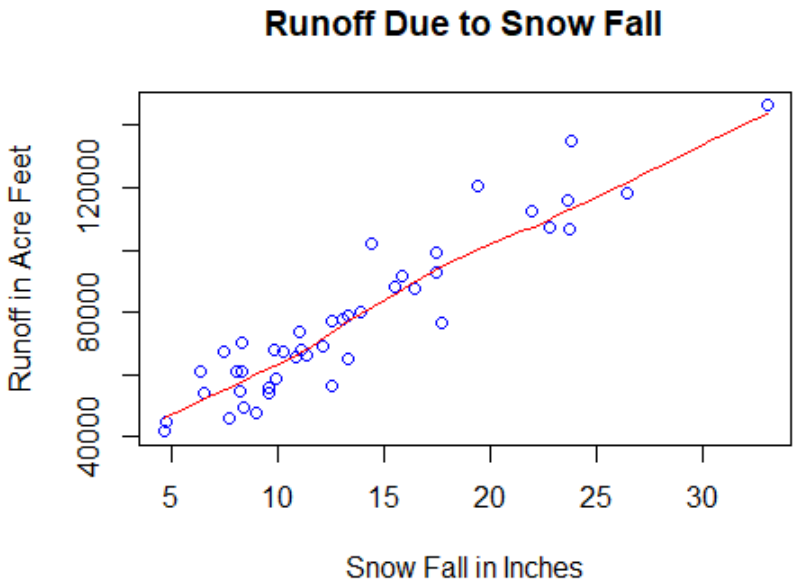
Homework 3: Stream Runoff in California

Kevin Toney

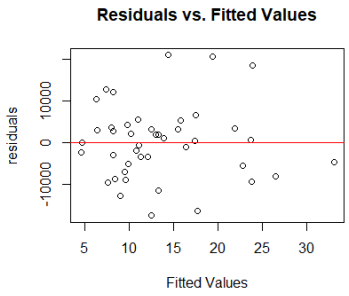
October 7, 2017

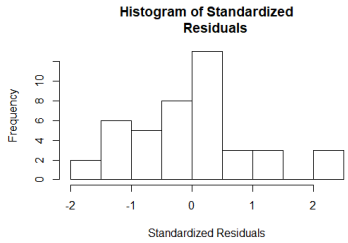
Question 1. Because of California's severe droughts in recent years, the purpose of the analysis is predict stream runoff from snowfall. Also, we are interested in determining the relationship between two measurements. If runoff can be predicted effectively, engineers, planners and policy makers can help more people because of their improved decision making. Statistical modeling, like linear regression, will enable us to input snow fall values into a linear equation and get the predicted amount of water runoff we can expect in Bishop, California, in the future.

Question 2. Consider the scatter plot of the data below.



According to the plot, the stream runoff data and the snow fall data seems to have a positive linear relationship. Therefore, as snow fall increases, we expect the stream runoff in acre feet to increase as well. Also, a correlation coefficient of 0.94, which is close a perfect score of positive 1, confirmed the relationship of runoff and snow fall is linear. There may be two outliers in the data, specifically when snow fall equals approximately 18 and 23 inches. Please keep these values in mind





Since simple linear regression's purpose is to create a mathematical equation that models the runoff data and the snow fall data, simple linear regression will achieve the goals of our analysis to predict. For another reason we can use simple linear regression, consider the two plots below.

According to the two above plots, we may use simple linear regression because the relationship is linear, the residuals (variation from the mean) are normally distributed. Additionally, stream run off measurements are independent of each other. Finally, the residuals vs. fitted values plot showed the variance of the stream runoff data points, from the average, is equal throughout the data. Therefore, Simple linear regression will appropriately predict the values of runoff and see the relationship it has with snow fall.

Question 3. In order to perform simple linear regression, we will fit the data to a regression model without transforming or centering the data. For your information, centering the snow fall measurements around the mean isn't appropriate for our analysis. The simple linear regression model, in mathematical form, is:

$y_i \sim N(\beta_0 + \beta_1 * x_i, \sigma^2)$ i = One to the number of recorded data values (43).

y_i = the predicted runoff in acre feet for the ith observation in the recorded data.

N = the normal distribution. The normal distribution, as you may remember makes a bell curve. In the normal distribution, a majority of the possible values are near the mean of the data. A smaller portion of the data can be found in the tails.

β_0 = The runoff, on average, we can expect if the snow fall equals 0.

β_1 = The average amount the runoff in acre-feet increases if the snow fall increases by one inch.

x_i = each ith snow fall measurement.

σ^2 = the average difference the runoff values are from their average.

In order to use simple linear regression, we assumed the relationship between stream runoff and snow fall is linear. Also, we assumed the residuals (variation of each data point from the mean) follow a normal distribution. Then, we assumed the variances of the runoff values from the average line are equal throughout range the of the data. Finally, we assumed the stream runoff values are not affected by each other.

According to the mathematical form of our linear regression model, $y_i \sim N(\beta_0+\beta_1*x_i, \sigma^2)$, we will be able to plug in a value for snow fall(x_i), and receive a predicted value for stream runoff (y_i). Also, we can interpret β_1 to understand the relationship stream runoff has with snow fall.

Question 4. The linear regression model, according to our data, is the following: $y_i \sim N(27,014.6 + 3,752.5 * x_i, \sigma^2 = 8,922)$

y_i = the predicted runoff in acre feet for the ith observation in the recorded data.

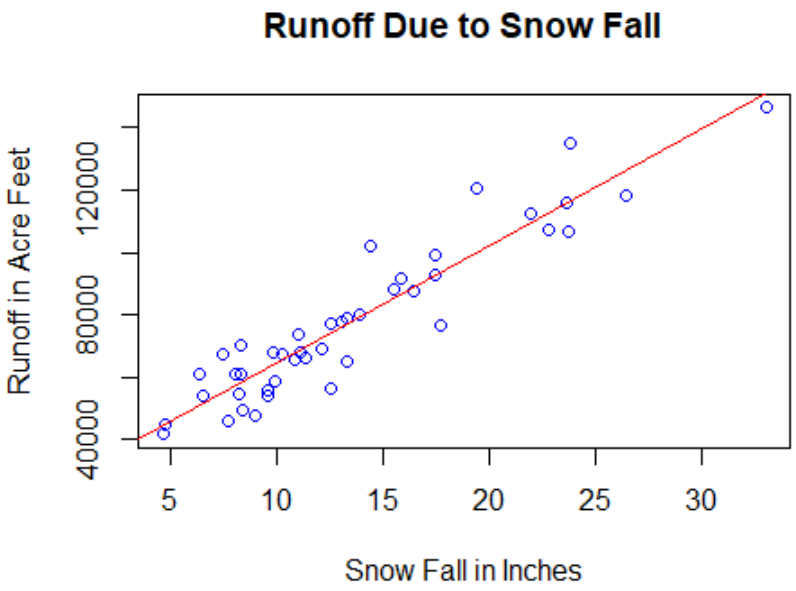
27,014.6 = $\hat{\beta}_0$ = the average stream runoff is predicted to be 27,014.6 acre feet if the snow fall is 0 inches.

3,752.5 = $\hat{\beta}_1$ = on average, the stream runoff is predicted to increase by 3,752.5 acre feet if the snow fall rises by 1 inch.

x_i = each ith snow fall measurement.

$\hat{\sigma}^2$ = 8,922 = the predicted average distance the stream run off varies away from the mean line is 8,922 acre feet.

The data, plotted with the simple linear regression line is below.

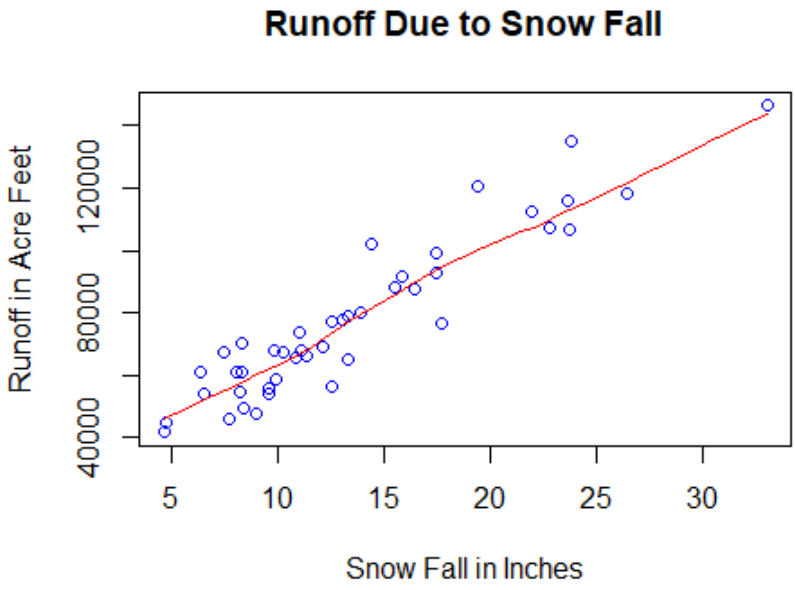


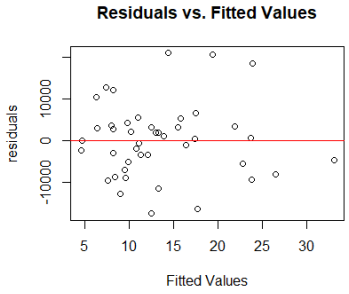
This model fits the data with the least amount of prediction error, and it is indicated by the red line.

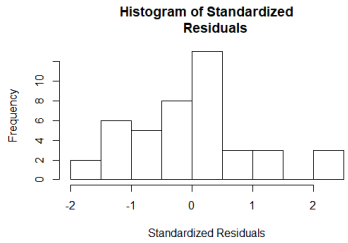
Question 5. As stated earlier, the assumptions we need to make in order to perform simple linear regression are fulfilled. The scatter plot of the runoff values compared to the snow fall show a linear relationship, as shown below. Moreover, the correlation score of 0.94 further supports the conclusion there is a linear relationship.

Secondly, the histogram of standardized residuals, i.e. the differences of the stream runoff centered around 0, shows the residuals follow a normal distribution. To validate this conclusion further, I performed a test called the KS test. The test returned a p-value of 0.621. The p-value is not small enough to prove the residuals don't come from a normal distribution. According to this test and the plot, the residuals follow a normal distribution.

According to the residuals vs. fitted values plot, the residuals do not follow any noticeable pattern. Therefore, stream runoff values must be independent of each other.







The residuals vs. fitted values plot may show variances along the regression line that are unequal to each other. Nonetheless, the variances aren't unequal enough. Additionally, a test, called the BP Test, gave us a p-value above 0.05, which isn't low enough to support the idea that the variances are unequal throughout the average line. Our assumptions for simple linear regression are justified.

In linear regression, we determined how reliable the linear regression model is to predict the stream runoff by interpreting a score called R^2. According to our simple linear regression model, the R^2 value is 0.88. Since this score is close to 1, which is a perfect score, our model is reliable. In fact, 88 percent of the stream runoff's variation in our data is explained by the linear model.

After running predictions many many times, we determined we are under-predicting the stream runoff values. On average, the predictions were 8,833.5 acre feet away from the mean stream runoff values. When we compared these differences to the range of the water runoff, this predictive differences wasn't significant to lessen our confidence in our prediction model.

Then, we used these predictions to create prediction intervals. We are 95 percent confident these intervals contain the future stream runoff values. On average, 88 percent of the

prediction intervals contained the actual water runoff values. On average, our prediction intervals had a range of 37,201.25 acre feet.

Question 6. In order to test that there is no relationship between snowfall and runoff, we determined if the null hypothesis (H_o) is proven false or not.
H_o = There is no relationship between stream runoff and snow fall.

H_a = There is a relationship between the two factors.

Through the linear regression model, the p-value, or the probability of getting runoff measurements as extreme or more extreme than our data, of this test is less than 0.05. Therefore, the null hypothesis was proven to be false. There is a relationship between stream runoff and snow fall. The relationship is this: on average, the stream runoff is predicted to increase by 3,752.5 acre feet if the snow fall rises by 1 inch.

Question 7. We are 95 percent confident if the snow fall equals 0 inches, the average water runoff would be between 20,513.98 acre feet and 33,515.20 acre feet.

Also, we are 95 percent confident if the snow fall increases by 1 inch, the average water runoff would increase by a value between 3,316.81 acre feet and 4,188.16 acre feet.

Question 8. In the winter of 2013-2014, the area by the stream in Bishop California only received 4.5 inches of snowfall. If this amount of snowfall happened again, we predict the stream runoff would be between 25,254.2 acre feet and 62,547.3 acre feet. Please take this prediction with a grain of salt. The minimum snow fall in our data has a value of 4.6 inches. Therefore, the snow fall of 4.5 inches is outside of the data's range. We do not know if the relationship of the two factors outside of our data is still linear. Therefore, we should not base any decisions off of this prediction.

A R Code:

```
rm(list=ls())

#Because of California's severe droughts in recent years,
#the purpose of the analysis is predict stream runoff from snowfall. Also,
#we are interested in determining the relationship
#between runoff and snow fall.
#If runoff can be predicted effectively, engineers, planners
#and policy makers can help more people with their decision making.

#don't take this section.
#The response variable is stream runoff in acre-feet.
#The river we are studying is near Bishop, California.
#The explanatory variable is snowfall, in inches.

#statistical modeling, like linear regression, will help us to input snow
fall values
```

```
#and get the predicted amount of
#run off we can expect in Bishop, California at a particular time.


#############
#Problem 2
#############
water.dat <- read.table("Fall 2017/STAT 330/Homework/Homework3/water.txt",
header = T, sep="")
plot(water.dat$Precip, water.dat$Runoff)
scatter.smooth(water.dat$Precip, water.dat$Runoff, ylab="Runoff in Acre Feet"
               xlab="Snow Fall in Inches", main="Runoff Due to Snow Fall",
               col="blue", lpars = list(col="red"))
#lpars makes the line red. lpars can also change the line thickness
and style.

#the data seems to have a positive linear relationship.
#There seems to be two outliers, specifically when snow fall equals
#approximately 18 and 23 inches.

cov(water.dat$Precip, water.dat$Runoff)
cor(water.dat$Precip, water.dat$Runoff)
#The scatter plot of the data is shown below.
#According to the scatter plot, and a correlation coefficient of 0.94,
#the relationship of Runoff and Snow Fall is linear. Also, the
#relationsip is positive. Therefore, if snow fall increase,
#we expect the runoff to increase as well.

#Since simple linear regression's purpose is to create
#a mathematical  equation that models the runoff data and the snow fall
#data, simple linear regression will achieve the goals of our analysis.

residuals <- linear_model$residuals
plot(water.dat$Precip, residuals, xlab = "Fitted Values", ylab="residuals",
     main = "Residuals vs. Fitted Values")
abline(0,0, col="red")

library(MASS)
library(lmtest)
st.res <- stdres(linear_model)
#test for residual normality.
hist(st.res, xlab = "Standardized Residuals", main="Histogram of
Standardized Residuals")

#Also, we may use simple linear regression because the relationship is linear
```

```
#the residuals (variation from the mean) are normally distributed.
Stream run off
#measurements are independent of each other. Finally, the residuals vs.
fitted values plot show
#the variance of the stream runoff is equal throughout the data.
#We will be able to predict the values of runoff and see the relationship
#it has with snow fall because of simple linear regression.


####################
#Problem 3
####################
# Try Centering the X's to make
#beta0 more interpretable.
linear_model <- lm(Runoff ~ Precip, data = water.dat)
summary(linear_model)

#see if centering the data is best for this analysis.
#The simple linear regression model, in mathematical form, is
#$y_i \sim N(\beta_0 + \beta_1*x_i, \sigma^2)$

#i = One to the number of recorded data values (43).

#$y_i$ = the predicted runoff in acre feet for the ith
observation in the recorded data.

#N = the normal distribution. In the normal distribution, a majority
#of the possible values are near the mean of the data.
#A smaller portion of the data can be found in the tails.
The normal distribution,
#as you may remember makes a bell curve.

#$\beta_0$ = The runoff, on average, we can expect if the
#snow fall equals 0.

#$\beta_1$ = The average amount the runoff in acres feet increases
#if the snow fall increases by one inch.

#$x_i$ = each ith snow fall measurement.

#$\sigma^2$ = the average amount the runoff values are from the average.

water.dat$Pre.cent <- water.dat$Precip - mean(water.dat$Precip)
plot(water.dat$Pre.cent, water.dat$Runoff)
#looking at the centered data, a snow fall of -10 inches doesn't
```

```
make common sense.
#I don't think centering the snow fall agrees witih common sense.

#In order to use simple linear regression, we assumed the
#relationship between stream runoff and snow fall is linear.
#Also, we assumed the residuals (variation of each data
point from the mean)
#follow a normal distribution.
#Then, we assumed the variances of the runoff values from
the average line are equal
#throughout range the of the data.
#Finally, we assumed the stream runoff values are not
affected by each other.

#According the mathematical form of our linear regression model,
#$y_i \sim N(\beta_0 + \beta_1*x_i, \sigma^2)$, we will be able
#to plug in a value for snow fall($x_i$), and
#receive a predicted value for stream runoff ($y_i). Also,
#we can interpret $\beta_1$ to understand the relationship
#stream runoff has with snow fall.

############
#Problem 4
############
summary(linear_model)

#The linear regression model, according to our data, is the following:
#$y_i \sim N(27,014.6 + 3,752.5*x_i, \sigma^2 = 8,922)$

#$y_i$ = the predicted runoff in acre feet for the ith observation in
the recorded data.

#27,014.6 = $\hat\beta_0$ = the average stream runoff is predicted to be
#27,014.6 acre feet if the snow fall is 0 inches.

#3,752.5 = $\hat\beta_1$ = on average, the stream runoff is predicted to
#increase by 3,75w.5 acre feet if the snow fall rises by 1 inch.

#$x_i$ = each ith snow fall measurement.

#8,922 = $\hat\sigma^2$ = the predicted average distance the stream run
#off varies away from the mean line.

#plot the fitted regression line on the data.
plot(water.dat$Precip, water.dat$Runoff, ylab="Runoff in Acre Feet",
```

```
     xlab="Snow Fall in Inches", main="Runoff Due to Snow Fall",
     col="blue")
abline(reg = linear_model, col="red")

##########
#Problem 5
##########


#The scatter plot of the runoff values compared to the snow fall
#show a linear relationship.
library(MASS)
library(lmtest)
st.res <- stdres(linear_model)

#test for residual normality.
hist(st.res)
ks.test(st.res, "pnorm")
#the p-value is not small enough to prove the residuals don't come from a
#normal distribution. According to this test and the plot, the
#residuals follow a normal distribution.

#test for equal variance and independence
residuals <- linear_model$residuals
plot(water.dat$Precip, residuals, xlab = "Fitted Values", ylab="residuals",
     main = "Residuals vs. Fitted Values")
abline(0,0, col="red")
#the residuals do not follow any noticeable pattern. The
#response values must be independent of each other.
bptest(linear_model)
#The residuals vs. fitted values plot may show variances which are unequal.
#Nonetheless, the variances aren't unequal enough to say the
variances aren't equal.
#Additionally, a test, called the BP Test, gave us a p-value
above 0.05, which isn't
#low enough to
support the idea that the variances are unequal throughout the average line.

#Therefore, the assumptions we reported in question three are all met correct

#In linear regression, we determine how reliable the
linear regression model is to predict
#the stream runoff by interpreting a score called $R^2$.
#According to our simple linear regresssion model, the $R^2$ is 0.88.
#Since this score is close to 1, which is a perfect score,
```

```
our model is relaible.
#In fact, 88 percent of the stream runoff in our data is
explained by the linear model.

#Find predictive accuracy through cross validation studies.
n.cv <- 250
bias <- rep(NA, n.cv)
rpmse <- rep(NA, n.cv)
pred.int.width <- rep(NA, n.cv)
coverage <- rep(NA, n.cv)
#create two NULL vectors

for(i in 1:n.cv) {
  ## Step 1: split data into test and training sets
  adv.test <- sample(1:nrow(water.dat), 5)
  test.data <- water.dat[adv.test,]
  train.data <- water.dat[-adv.test,]

  ## Step 2: Fit model to training data
  my.model <- lm(Runoff ~ Precip, data = train.data)
  #if I am using a predict.lm statement, I need to fit the model
  #using the column names of the data frame, not variables
  #I created before.
  ## Step 3: predict for test data
  test.preds <- predict.lm(my.model, newdata = test.data)
  ## Step 4: calculate the bias and RPMSE
  bias[i] <- mean((test.preds - test.data$Runoff))
  rpmse[i] <- sqrt(mean((test.preds - test.data$Runoff)^2))

  pred.int <- predict.lm(my.model, newdata= test.data,
  interval = "prediction", level=0.95)
  #do my prediction intervals contain the five data points from the test data
  coverage[i] <- pred.int[,2] < test.data$Runoff &
  test.data$Runoff < pred.int[,3]
  pred.int.width[i] <- mean(pred.int[,3] - pred.int[,2])
  #get the average of the differences between the upper and the lower
  #that number equals the prediction interval width.
}

#Are we overpredicting or underpredicting values compared to the mean line?
mean(bias)
#We are underpredicting the values. The mean predictive bias is negative.

mean(rpmse)
#On average, the predictions are 8,833.5 acre feet away from the mean stream
```

```
#When we compare these differences to the range of the stream runoff values,
#this predictive differences aren't signifcant to lessen our confidence
in our prediction model.

#On average, 88 percent of the prediction intervals contain
#the actual stream runoff values.
mean(coverage)

#On average, our prediction intervals have a range of 37,201.25 acre feet.
mean(pred.int.width)



##########
#Problem 6
##########

#In order to test that there is no relationship between snowfall and runoff,
#we will see if the null hypothesis ($H_o$) is proven false or not.
#$H_o$ = There is no relationship between stream run off and snow fall.
#$H_a$ = There is a relationship between the two factors.
summary(linear_model)
#The p-value of this test is less than 0.05. Therefore, the null hypthoesis
#is false. There is a relationship.


###########
#Problem 7
###########
confint(linear_model, level = 0.95)
#We are 95 percent confident if the snow fall equals 0 inches,
#the average water runoff would be between 20,513.98 acre feet and 33,515.20

#Also, we are 95 percent confident if the snow fall increases by 1 inch,
#the average water runoff would increase by a value between 3,316.81 acre
feet and 4,188.16 acre feet.



###########
#Problem 8
###########

#In the winter of 2013-2014, the area by the stream in Bishop California only
predict.lm(linear_model, newdata = data.frame(Precip=4.5),
```

```
interval = "prediction")
#We predict the stream runoff would be between 25,254.2 acre feet
and 62,547.3 acre feet.
#Please take this prediction with a grain of salt. The minimum snow fall
data point
#has a value of 4.6 inches. Therefore, the snow fall of 4.5
inches is outside of the data's range.
#We do not know if the relationship of the two factors outside of our data
#is still linear.
```