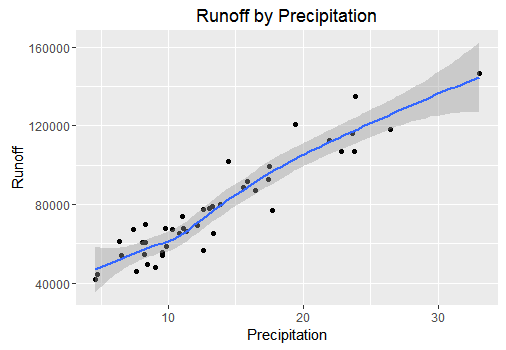
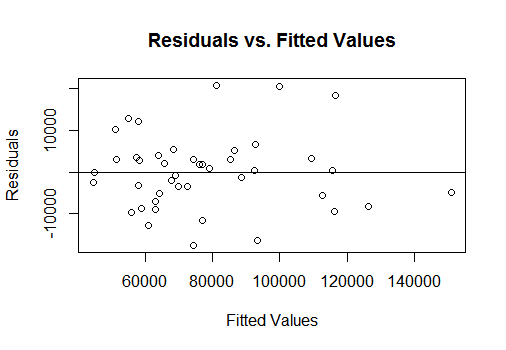
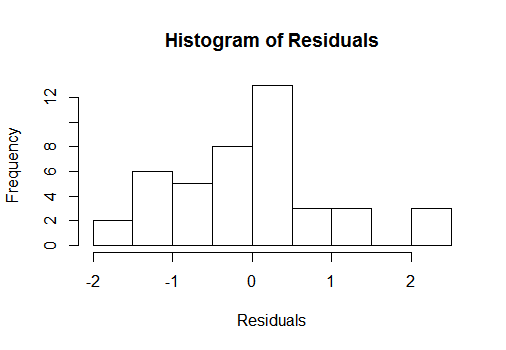
1. Severe drought conditions have plagued Southern California and made it difficult for engineers, planners, and policy makers to do their job because of their lack of knowledge of how much water is going to be entering the area. We have a dataset which compares stream runoff to the amount of precipitation. Our goal is to use this dataset to create a model that will allow us to predict runoff amounts using the precipitation amount in order to help planners, policy makers, etc. perform their jobs. Linear regression can help us with this task because we can use the regression line to make predictions.

2. We first assess whether a simple linear regression model is suitable to analyze this data. Looking at a scatterplot of the data, we see that there is a very strong, positive, linear relationship between precipitation and runoff. We also have a correlation of .938, which shows a strong, positive relationship.



Looking at our "Residuals vs. Fitted Values" plot, it doesn't look like there are any abnormal patterns in our residuals and they look to be equally distributed across the line. The "Histogram of Residuals" show that the residuals look normally distributed. There may be a bit of a right skew, but it looks all right. With these assumptions being met we can continue with a simple linear regression model.

3. We are actually going to use a centered model for this case. This just means that instead of using the original precipitation values, we will subtract the average precipitation value from them all. We do this because if you look at the original data you will notice that a precipitation of 0 will give you about 40,000 acre-feet for runoff, which doesn't make sense.

yi = β0 + β1x\* + εi where εi ~ N(0,σ2)

yi = value of the i­th observation of runoff

β0 =the average runoff value in a year when the precipitation amount is equal to the average precipitation amount

β1 = the average increase in runoff as the precipitation amount increases by 1

x\* = xi- = the precipitation value minus the average precipitation value, this is the value you will use to predict the ith runoff value

εi = the error associated with the ith observation

To use this model we are assuming linearity, independence, normality, equal variance. Once we fit the model, we will be able to plug precipitation values into the equation and predict runoff.

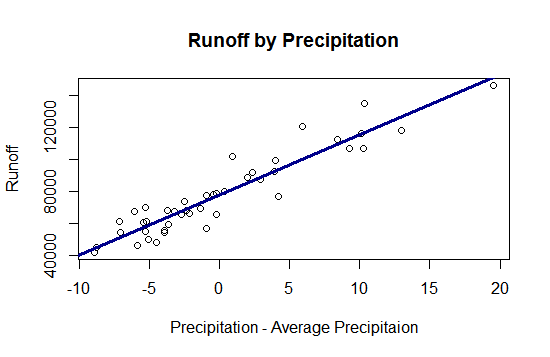
4.

i  = 77,756 + 3,752.5x\*

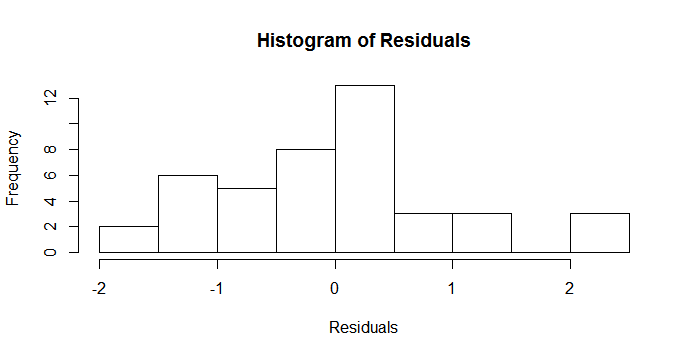
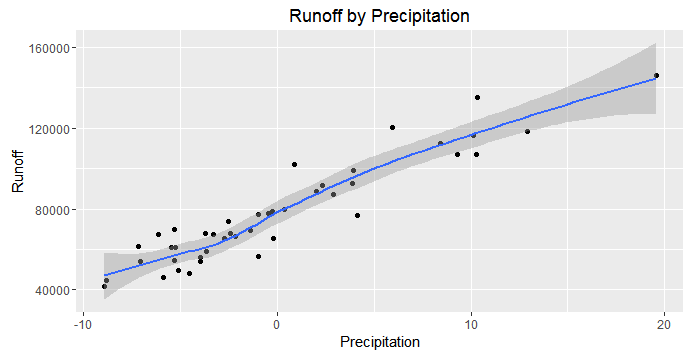
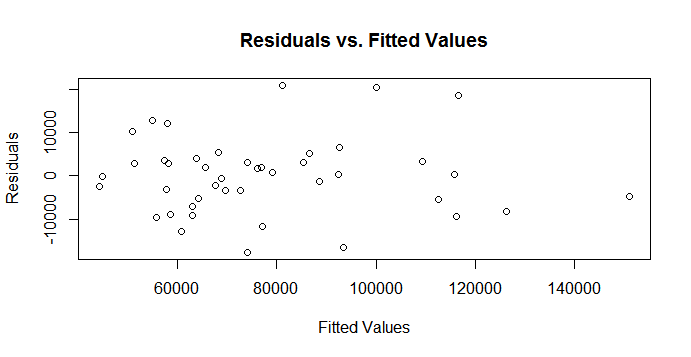
where x\* = xi -

β0 = If the precipitation is equal to the average precipitation amount then we would expect an average runoff of 77,756 acre-feet.

β1 = If we increase the difference between the precipitation amount and the average precipitation amount by 1 inch, we expect an average increase in runoff of 3,752.5 acre-feet.



To the right is a plot with a fitted regression line lying on top of the data.

5.

Our model assumes linearity, equal variance, normality, and independence. The scatterplot of the data show linearity. The residuals versus fitted values have equal variance across the line. The histogram of residuals shows that the residuals are normally distributed. We might see a slight patter in the residuals vs. fitted values plot, and it is hard to say that the Runoff amount at one point in time does not affect the Runoff amount at another point in time. However, we will assume independence and will use this model to make predictions for runoff.

This model has an R2 measure of .8807. This means that 88% of the variation in Runoff can be explained by the precipitation amount. We ran a cross-validation to measure our predictive accuracy. We found a predictive bias of 154. This means that we are generally a little high on our predictions. However, comparing this to the scale for Runoff, that is very small. Our predictions are about 8,580 acre-feet off on average as well. We have a coverage of 91%, which means that 91% of the prediction intervals we create do indeed contain the true Runoff value. The average width of our prediction intervals is 37, 070 acre-feet.

6. Testing the null hypothesis that there is no relationship between snowfall and runoff, we found a p-value of .00000000000000022. We can effectively reject the null hypothesis and accept the alternative hypothesis that there is indeed a relationship between snowfall and runoff.

7. We are 95% confident that on an average year of precipitation, the average runoff amount will be between 75,008 and 80,503 acre-feet. We are also 95% confident that as the precipitation increases by 1 inch, the average increase in Runoff will be between 3,316 and 4,188 acre-feet.

8. With the site receiving 4.5 inches of snowfall, we predict that the associated runoff will be 43,900 acre-feet, with 95% confidence that the associated runoff will lie between 25,254 and 62,547 acre-feet. However, we are very wary about making this prediction because the snowfall amount of 4.5 inches is outside the bounds of our original data.

# Appendix

setwd("~/3Fall2016/stat330/Homework")

# Read in Data #

water <- read.table("water.txt",header=TRUE)

head(water)

# Libraries #

library(lmtest)

library(MASS)

# Scatterplot and Correlation #

require(ggplot2)

water.plot <- ggplot(water,aes(x=Precip,y=Runoff))

water.plot + geom\_point() + geom\_smooth() + theme\_grey() +

ggtitle("Runoff by Precipitation") +

xlab("Precipitation") + ylab("Runoff")

with(water,{cor(Precip,Runoff)})

# Centered Model and Regular Model#

water$cent <- water$Precip - mean(water$Precip)

lm.water <- with(water,{lm(Runoff~Precip)})

lm.cent <- with(water,{lm(Runoff~cent)})

summ <- summary(lm.cent)

summary(lm.water)

# Assumptions #

plot(lm.water$fitted.values,lm.water$residuals,

main="Residuals vs. Fitted Values",

xlab="Fitted Values",ylab="Residuals")

abline(h=0)

hist(stdres(lm.water),main="Histogram of Residuals",xlab="Residuals")

# Plot with Regression Line #

plot(x=water$cent,water$Runoff,

main="Runoff by Precipitation",

xlab="Precipitation - Average Precipitation",

ylab="Runoff")

abline(reg=lm.cent,col="blue",lwd=3)

# Centered Assumptions #

require(ggplot2)

cent.plot <- ggplot(water,aes(x=cent,y=Runoff))

cent.plot + geom\_point() + geom\_smooth() + theme\_grey() +

ggtitle("Runoff by Precipitation") +

xlab("Precipitation") + ylab("Runoff")

with(water,{cor(cent,Runoff)})

plot(lm.cent$fitted.values,lm.cent$residuals,

main="Residuals vs. Fitted Values",

xlab="Fitted Values",ylab="Residuals")

hist(stdres(lm.cent),main="Histogram of Residuals",xlab="Residuals")

# Cross-Validation #

pred.width <- numeric(0)

coverage <- numeric(0)

bias <- numeric(0)

rpmse <- numeric(0)

for (i in 1:1000){

n <- 4

sampl <- sample(1:length(water$cent),n)

train <- water[-sampl,]

test <- water[sampl,]

train.lm <- with(train,{lm(Runoff~cent)})

pred.cent <- data.frame(cent=test$cent,Runoff=test$Runoff)

pred.run <- predict.lm(train.lm,pred.cent)

pred.cent$predrun <- pred.run

bias[i] <- mean(pred.cent$predrun - pred.cent$Runoff)

rpmse[i] <- sqrt(mean((pred.cent$predrun - pred.cent$Runoff)^2))

pred.int <- predict.lm(train.lm,test,

interval="prediction",level=.95)

covers <- mean(pred.int[,2] < test$Runoff &

test$Runoff < pred.int[,3])

coverage[i] <- covers

int.width <- mean(pred.int[,3] - pred.int[,2])

pred.width[i] <- int.width

}

mean(bias)

mean(rpmse)

mean(coverage)

mean(pred.width)

# Parameter Confidence Intervals #

conf.parms <- confint(lm.cent)

conf.parms

# A few Predictions #

preds <- data.frame(cent=c(4.5-mean(water$Precip)))

preds$pred <- predict.lm(lm.cent,newdata=preds,

interval='prediction',level=.95)

preds