WaterRunoff

Jillian Maw

2/9/2022

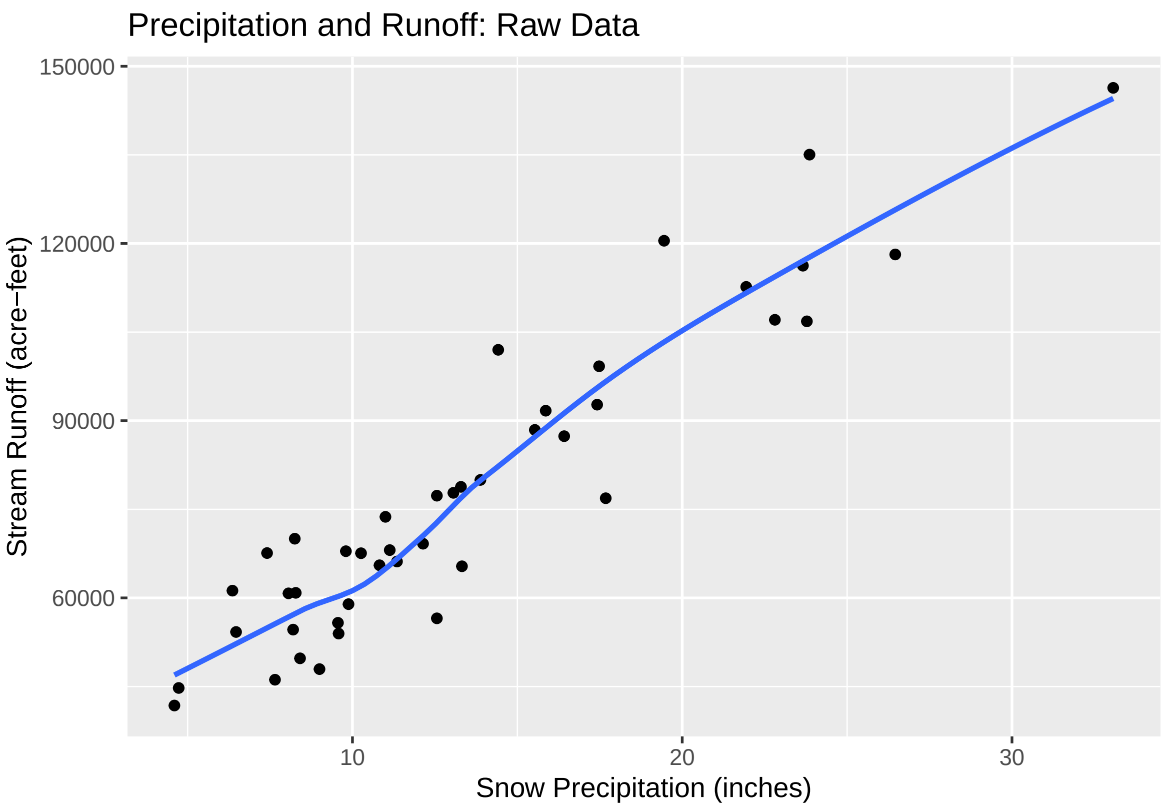
HOMEWORK ANALYSIS #3 - WATER AVAILABILITY

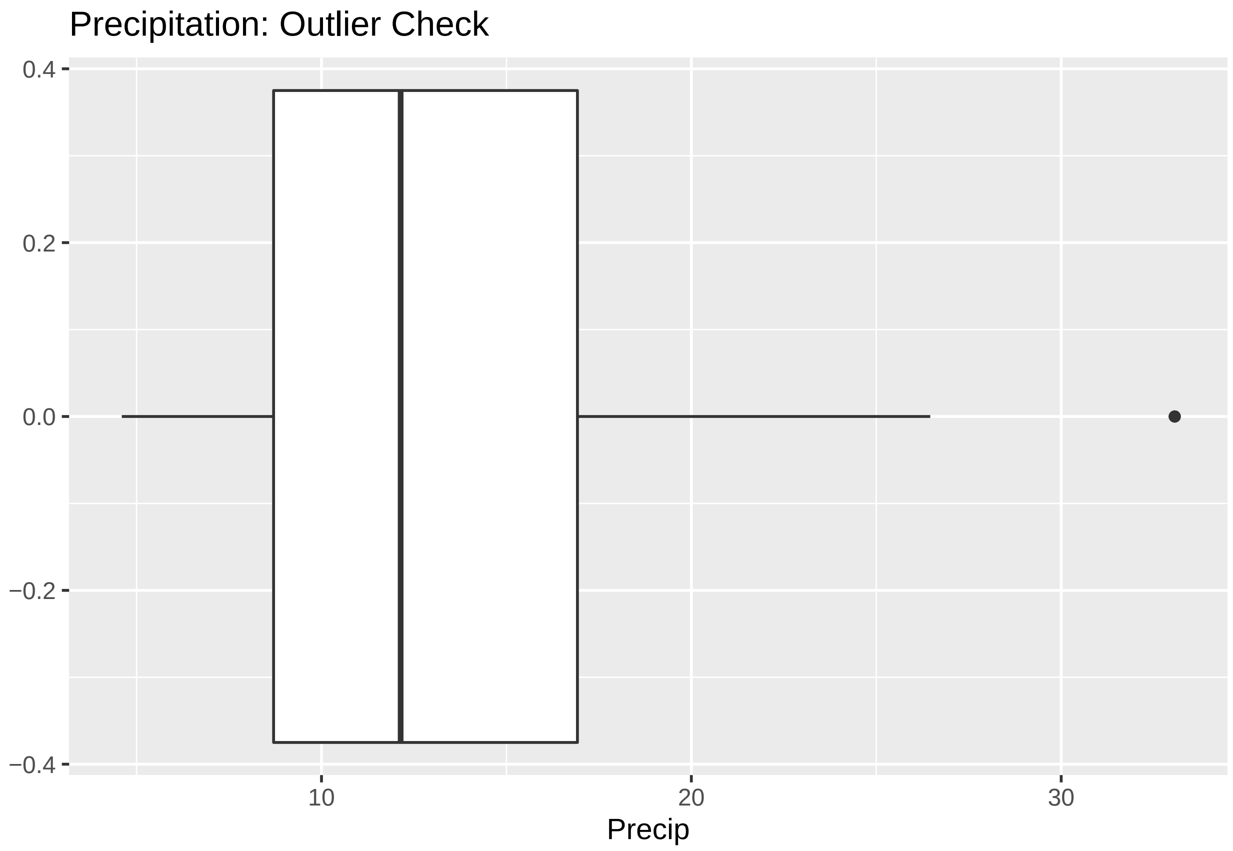
Climate change has left California particularly vulnerable to severe drought conditions. One factor affecting water availability in Southern California is stream runoff from snowfall (FYI: water in Utah is also heavily reliant on snowpack). If runoff could be predicted, engineers, planners, and policy makers could do their jobs more effectively because they would have an estimate as to how much water is entering the area.
The dataset `water.txt` compares the stream runoff in acre-feet of a river near Bishop, California (due east of San Jose) with snowfall, in inches, at a site in the Sierra Nevada mountains. For each of the following questions, assume that your audience are city water planners with moderate statistical training. Please attach your clearly commented code (R or Python) to the back of your answers as an appendix.

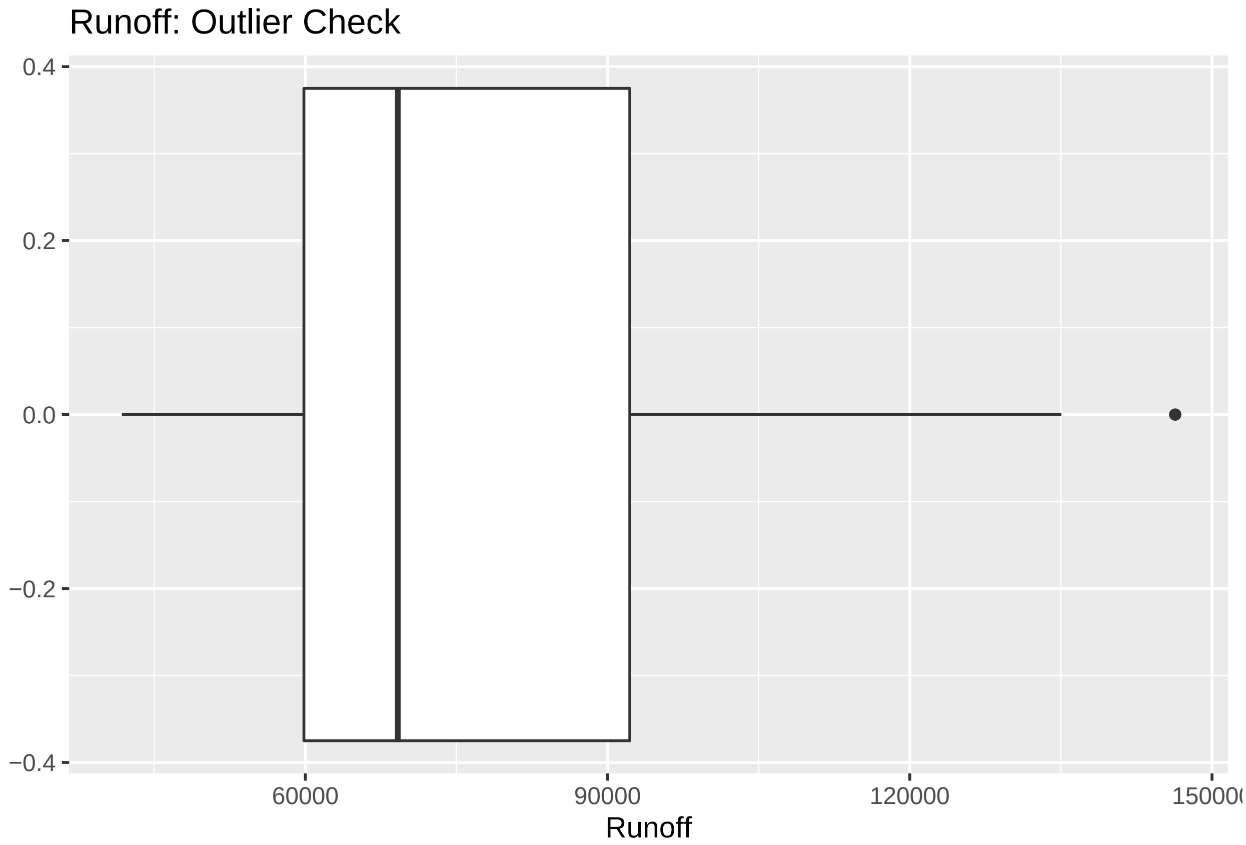
1. In your own words, summarize the overarching problem and any specific questions that need to be answered using the water data. Discuss how statistical modeling will be able to answer the posed questions.

We are determining whether the amount of precipitation of snowfall (measured in inches) in the Sierra Nevada mountains can predict the stream runoff (measured in acre-feet) of a river in Bishop, CA. If this prediction can be proven, it can assist engineers, planners, and policy makers by providing estimates for how much water can be used by cities in Southern California, an area prone to droughts. We need to be able to provide ranges of available water so the aforementioned groups of people can know the worst and best case scenarios for their towns and cities' water needs.

2. Using exploratory techniques (don't actually fit a model), explore the data to assess if a simple linear regression (SLR) model is suitable to analyze the water data. Justify your answer using any necessary graphics and relevant summary statistics that would suggest an SLR model would be successful at achieving the goals of the study.


Precipitation and Runoff: Raw Data
150000
120000
90000
60000
Stream Runoff (acre−feet)
10
20
30
Snow Precipitation (inches)


Precipitation: Outlier Check
0.4
0.2
0.0
−0.2
−0.4
10
20
30
Precip



It appears the data is roughly linear. While our correlation is high at 0.9384 out of 1, there may be two outliers, where a point has a snow precipitation of 33.07 inches and where a point has a stream runoff of 146,345 acre-feet, that are pulling the linear relationship in their directions stronger than the rest of the data would go without the outliers. The equal variance, independence, and normality assumptions seem to be fine at first glance, but we will check with statistical tests later in the report. We can also tell with the covariance of 152832.0828 that there is a positive relationship between snow precipitation and stream runoff.

3. Write out (in mathematical form with Greek letters) a justifiable SLR model that would help answer the questions in problem. Provide an interpretation of each mathematical term (β parameter) included in your model. Using the mathematical form, discuss how your model, after fitting it to the data, will be able to answer the questions in this problem.

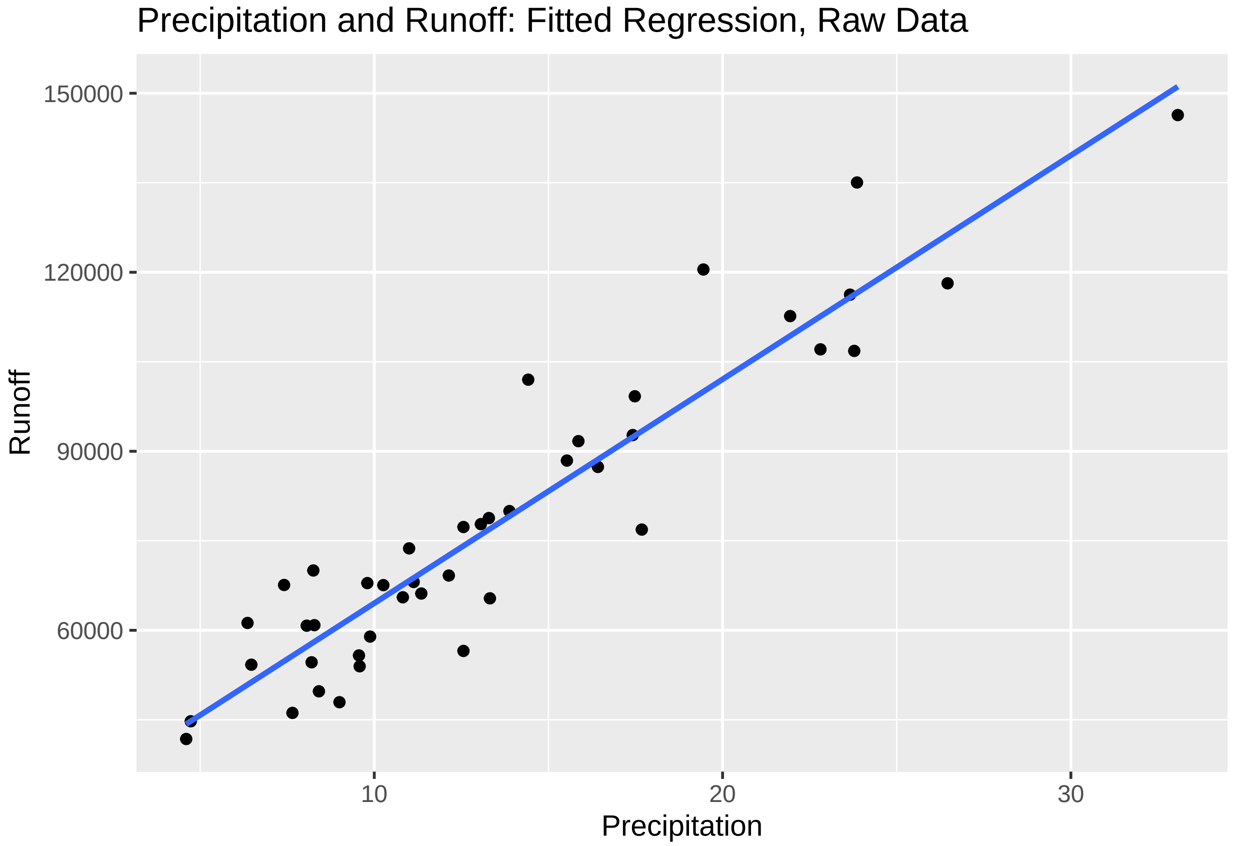
$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \text{where } \epsilon_i = N(0,\ \sigma^2)$$

In this model, we assume we are able to meet the assumptions of linearity, independence, equal variance, and normality to create a simple linear regression. Above, y_i represents the response variable stream runoff (measured in acre-feet) of a river near Bishop, California, at a given measurement observation i. There are a total of 43 measurements used to create this simple linear regression model. The x_i represents the explanatory variable snow precipitation, in inches, at a site in the Sierra Nevada mountains, at a given measurement observation i, which we are using to explain the response variable of stream runoff (using statistical modeling). β_0 represents the intercept coefficient, which says when x_i, the snow precipitation measurement, is 0, the average y_i, the stream runoff measurement, is the intercept coefficient. The slope coefficient β_1 states that as the explanatory variable x_i increases, the average response variable y_i increases by the slope coefficient. In this model's case, it means that as the snow precipitation (in inches) increases by

1, the average stream runoff (in acre-feet) also increase by β_1. The ϵ_i represents the residual errors, or the difference from the true average of snow precipitation. The N is short for Normal distribution, meaning the simple linear regression model's residuals follow a Normal distribution's shape and behaviors, standardized at a mean of 0 and a standard deviation of σ^2. The symbol σ^2 represents the variance of the data around the regression line fitted to the data by this model. Another way to think of the variance is that it is the square of the standard deviation. The standard deviation shows that for any x_i, 99.7% of the response variables will be within 3 standard deviations of the regression line made by $\beta_0 + \beta_1 x_i$, the intercept coefficient plus the product of the slope intercept and explanatory variable. Using this model on the collected data will enable us to predict runoff by entering the amount of snow precipitation into the equation, allowing engineers, planners, and policy makers to have an estimate for available water.

4. Fit your model in #3 to the water data and summarize the results by displaying the fitted model in equation form (do NOT just provide a screen shot of the R or Python output). Interpret each of the fitted parameters in the context of the problem. Provide a plot of the data with a fitted regression line.

Fitting our data to the aforementioned model, we get $\hat{y}_i = 27,014.5874 + 3,752.4856\ x_i)$. In this model, $\hat{y}_i$ represents the estimated average stream runoff (in acre-feet) for measurement observation i. The x_i represents the snow precipitation (in inches) for measurement observation i. The intercept coefficient has been replaced by 27014.5874 acre-feet, which says when the snow precipitation is 0 inches, the average stream runoff is 27014.5874 acre-feet. The slope coefficient has been replaced by 3752.4856 inches, which says as the snow precipitation increases by 1 inch, the average stream runoff increases by 3752.4856 acre-feet.

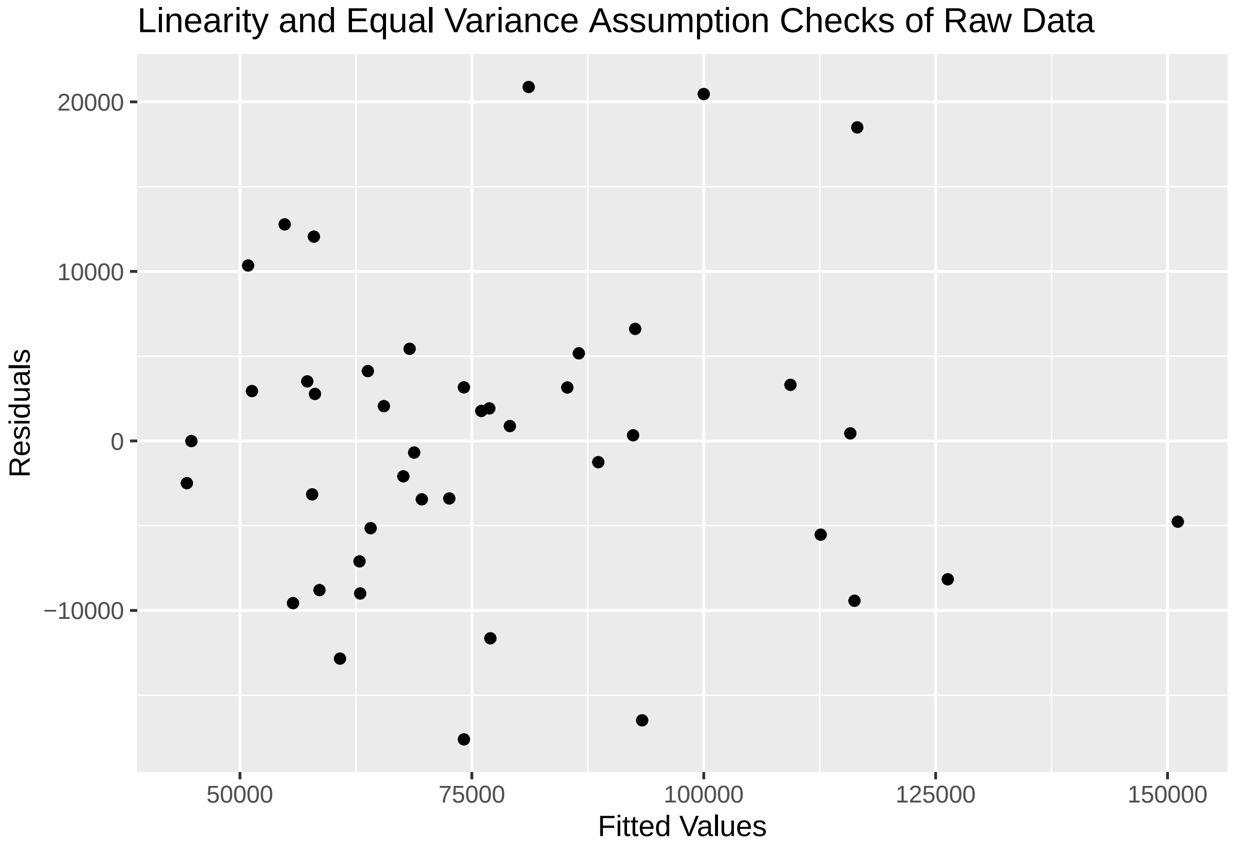


5. List then justify your model assumptions using appropriate graphics or summary statistics.

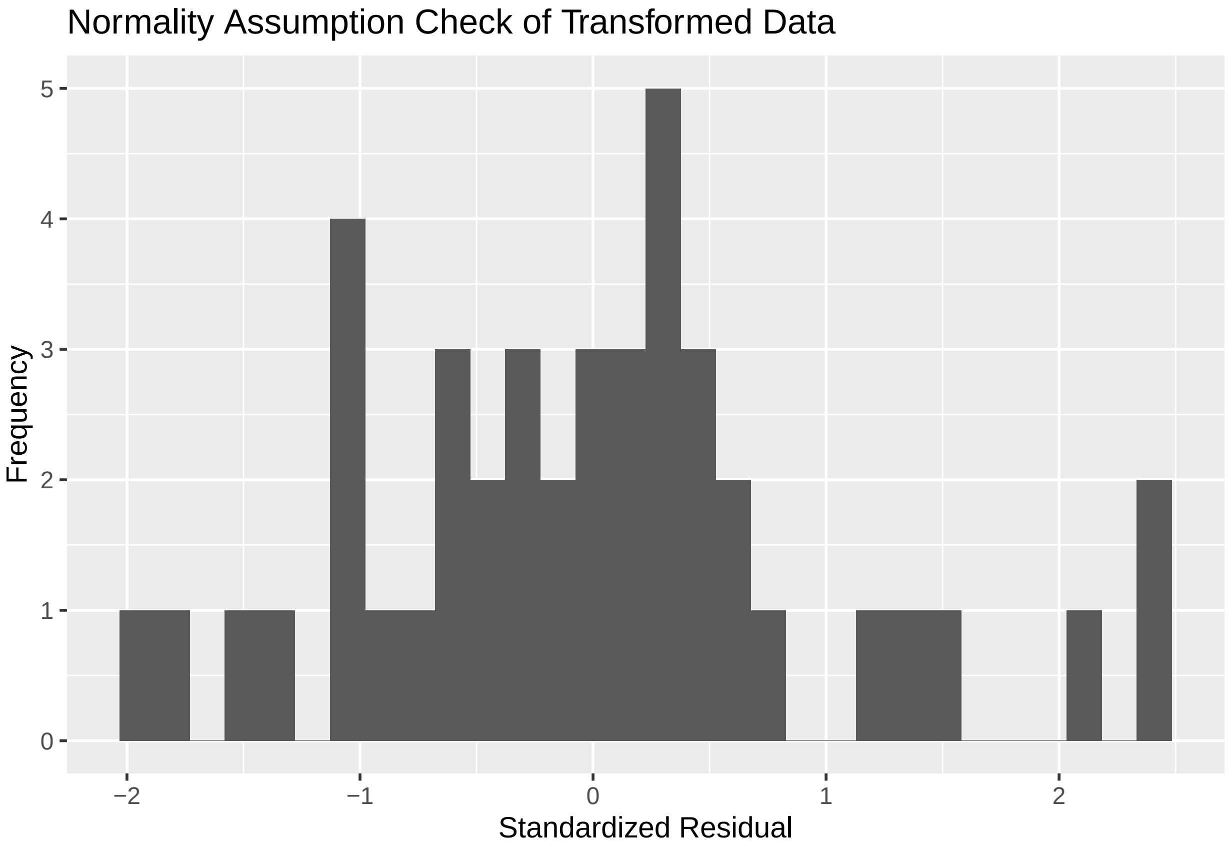
The model assumptions we made were that the model was linear, independent, followed a Normal distribution, and has equal variance. We will prove this with graphics and summary statistics below. We will also review the outliers that could throw off the Normal distribution assumption, and prove it ultimately will not.

First, we will prove the linear, independent, and equal variance assumptions by checking a plot comparing fitted values (predicted value for a point in the dataset) to the residuals (the difference between an observed and a predicted value) in a scatterplot. If there are no patterns, it is independent. If there is a constant variance, it is linear and has equal variance. As can be seen below, the linearity and independence assumptions seem to be proven correct.

In addition, for the independence assumption, it seems highly unlikely that one year's snow precipitation will affect the snow precipitation of years following, especially with climate change making the measured area prone to droughts.



Next, we will prove the data follows a Normal distribution by plotting the standardized residuals (residuals transformed to show their difference from the data's mean, if the mean was 0) on a histogram. Below, we see the histogram shows a generally Normal distribution with the exception of two possible outliers like we highlighted above, but we can prove the Normal distribution is good enough with an additional statistical test, which we will explain below the histogram.



Below, we show the results for the One-sample Kolmogorov-Smirnov test, also called a KS-test, and the Jarque-Bera test for normality, also called a JB-test, which conducts hypothesis tests on whether or not a data set follows a Normal distribution or not. For the KS-test, the null hypothesis is that the data comes from a Normal distribution, while the alternative hypothesis is that the data does *not* come from a Normal distribution. For the JB-test, the null hypothesis is that the data's distribution is not skewed, whereas the alternative hypothesis is that the data's distribution is skewed. We set the p-value to be 0.05, to prove significance. Because the One-sample Kolmogorov-Smirnov test fails to reject the null hypothesis and the studentized Breusch-Pagan test fails to reject the null hypothesis, as seen in the test results below, we accept that we have the Normal assumption met for our data.

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  std.resids
## D = 0.11131, p-value = 0.621
## alternative hypothesis: two-sided
```

```
##
##  Jarque-Bera test for normality
##
## data:  std.resids
## JB = 1.3587, p-value = 0.362
```

We then have a statistical test for checking the Equal Variance assumption for our model. Checking the scatterplot of fitted values versus residuals above, it appears that we have a model with linearity and mostly equal variance, can proceed with the Breusch-Pagan test, or BP-test, to confirm equal variance. The BP-test conducts hypothesis tests on whether or not a data set has homoskedasticity, or equal variance. The null hypothesis assumes that the data has homoskedasticity, while the alternative hypothesis assumes that the data has heteroskedasticity. We set the p-value to be 0.05, to prove significance. The code results below shows the Breusch-Pagan test produces a p-value that fails to reject the null hypothesis, so we accept that we have the Equal Variance assumption met for our data.

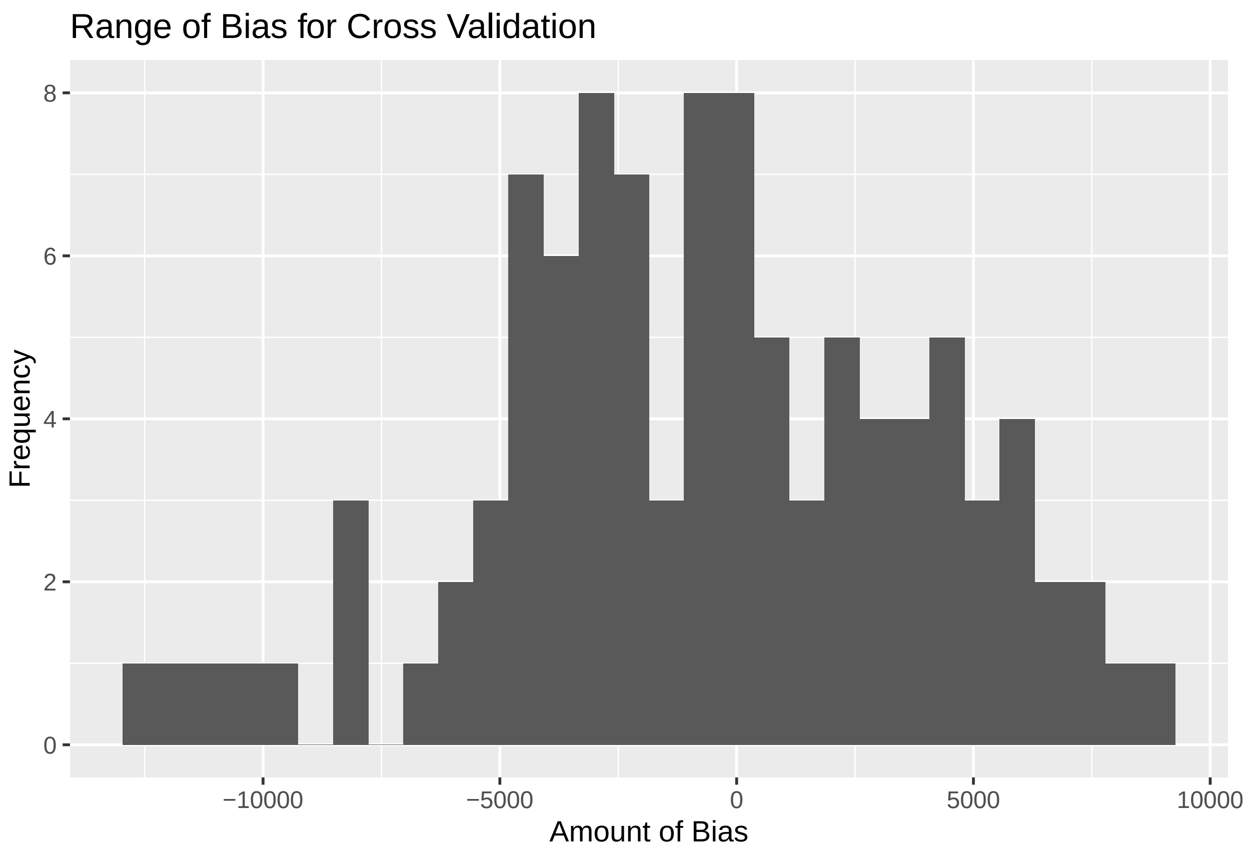
```
##
##  studentized Breusch-Pagan test
##
## data:  water.slm
## BP = 1.0277, df = 1, p-value = 0.3107
```

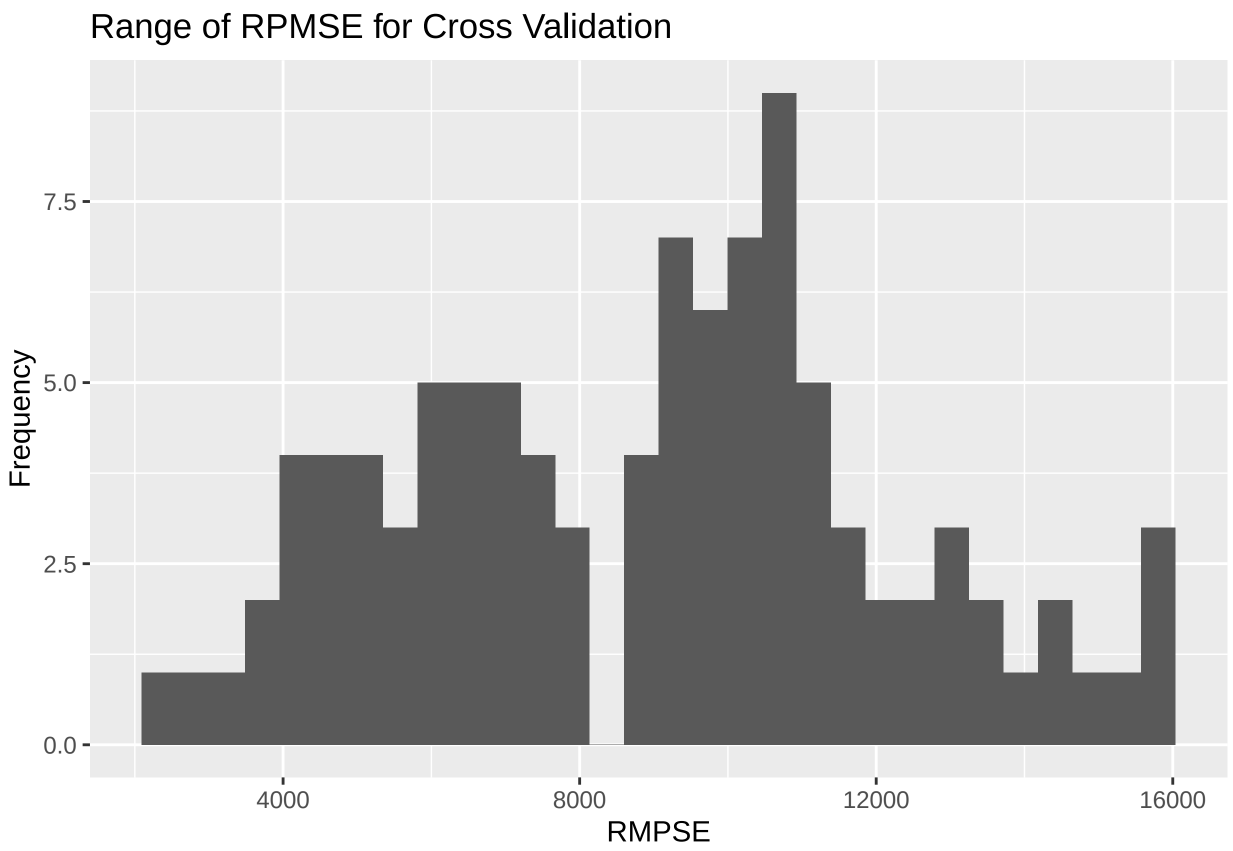
Finally, we will check for outliers that could affect the Normal assumption of the data set. We use Cook's formula to check how much the data set's simple linear regression is affected by each individual point. We will flag those points for future consideration, but due to the passing of the earlier KS-test and JB-test, we will leave the points alone for this report.

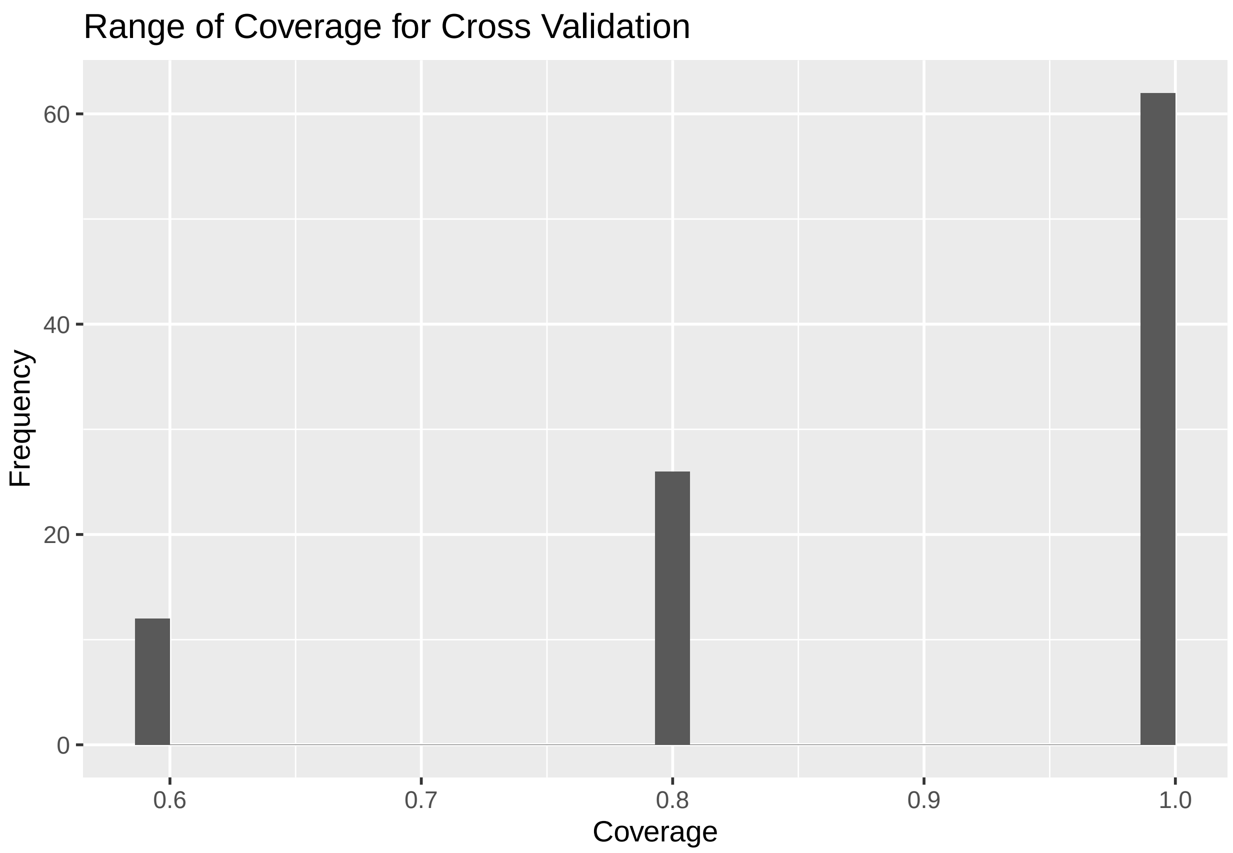
We flagged two observations as possibly affecting the data, the observation where Precipitation is 19.45 inches and Runoff is 120,463 acre-feet and the observation where Precipitation is 23.86 inches and Runoff is 135,043 acre-feet, instead of the originally assumed point where Precipitation is 33.07 or where Runoff is 146,345 acre-feet.

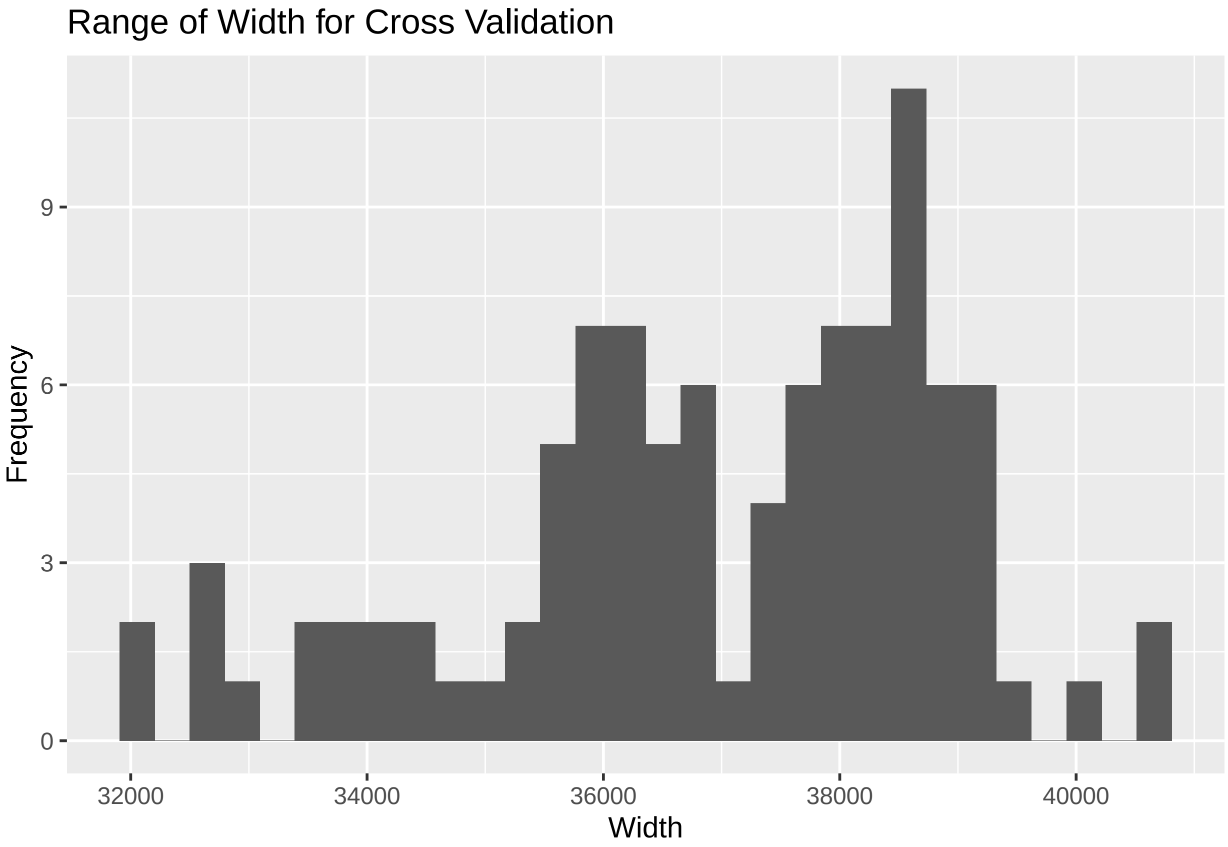
6. Assess the fit *and* predictive capability of your model. Discuss on the level of your target audience (e.g. interpret your model R^2). Draw a conclusion about how "good" your predictions are relative to the range of the response variable. Be sure to include any necessary statistics that show your prediction intervals are working as you expect.

The percent of variability in stream runoff explained by snow precipitation is 88.07. We conducted a cross validation procedure on 5 randomly selected observations of our test data, we calculated an average bias of -653.6322848. This means our predictions are, on average, lower than the true average stream runoff. We also calculated an average Root Predictive Mean Square Error of 8905.9954884, which means our predictions are off, on average, 8905.9954884 acre-feet. Considering the size of most rivers, this seems a reasonable amount of error. To see how far our predictions ranged, we calculated the width to be 36954.1462977 acre-feet, on average. In addition, the coverage, or the percentage of prediction intervals that contain the true average stream runoff, to be 0.9.


Range of Bias for Cross Validation
Frequency
8
6
4
2
0
−10000
−5000
0
5000
10000
Amount of Bias


Range of RPMSE for Cross Validation
Frequency
0.0
2.5
5.0
7.5
4000
8000
12000
16000
RMPSE


Range of Coverage for Cross Validation
Frequency
60
40
20
0
0.6
0.7
0.8
0.9
1.0
Coverage



7. Carry out a test that there is no relationship between snowfall and runoff (i.e., write out the hypotheses, report an appropriate p-value, and conclude in context).

With simple linear regression models, we want to perform a hypothesis test on β_1 (explained above in step 3) to test whether or not there is a linear relationship between our explanatory and response variables. To see if there is a relationship between snow precipitation and stream runoff, we create two opposing hypotheses to test. The null hypothesis is that β_1 is 0, and therefore proves there is not a linear relationship. The alternative hypothesis is that $beta_1$ is *not* 0, and therefore proves there *is* a linear relationship. We will set the p-value (the probability of obtaining hypothesis test results at least as extreme as the results actually observed, under the assumption that the null hypothesis is correct) at 0.05, a widely accepted academic standard. If the p-value is less than 0.05, we can conclude that snow precipitation has a statistically significant, linear effect on stream runoff. The null and alternate hypothesis are mathematically:

$$H_0: \ \beta_1 = 0 H_a: \ \beta_1 \neq 0$$

Below, we see the results of the test:

```
## [1] "Hypothesis Tests:"
```

```
##
## Call:
## lm(formula = Runoff ~ Precip, data = water)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -17603.8  -5338.0    332.1   3410.6  20875.6
##
## Coefficients:
##             Estimate Std. Error t value             Pr(>|t|)
## (Intercept)  27014.6     3218.9   8.393       0.000000000193 ***
## Precip        3752.5      215.7  17.394 < 0.0000000000000002 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 8922 on 41 degrees of freedom
## Multiple R-squared:  0.8807, Adjusted R-squared:  0.8778
## F-statistic: 302.6 on 1 and 41 DF,  p-value: < 0.00000000000000022
```

Assuming the null hypothesis is true, the probability of observing a slope of 0.01, or more extreme, is essentially 0. Therefore, we conclude there is a statistically significant, linear effect that snow precipitation has on stream runoff.

8. Construct 95% confidence intervals for the slope and intercept parameters and interpret these intervals in the context of the problem.

We are 95% confident that if we sampled repeatedly from our model, the true value of $beta_0$ would be between 20513.9780517 and 33515.1966499. In context, that means if snow precipitation was 0 inches, the average amount of stream runoff in acre-feet would be between 20513.9780517 acre-feet and 33515.1966499 acre-feet. We are 95% confident that if we sampled repeatedly from our model, the true value of $beta_1$ would be between 3316.8091054 and 4188.1620664. In context, this means that for every 1 inch increase of snow precipitation, the average stream runoff increases by between 3316.8091054 acre-feet and 4188.1620664 acre-feet.

9. In a recent winter, the site only received 4.5 inches of snowfall. What do you predict will be the associated runoff? Provide a 95% predictive interval and interpret the interval in the context of the problem. Do you have any hesitations performing this prediction (hint: you should)? Describe these hesitations and their potential impact on your prediction.

Predicting the associated Runoff of an observation that had a Precipitation of 4.5 inches is outside the scope of this data and it’s regression model. We should not extrapolate the associated Runoff of a value outside the limits of the regression model because we do not know how much additional variability we may need to account for to give an accurate prediction. Nevertheless, if such warnings are ignored, we are 95% confident that if the Precipitation was 4.5 inches, the associated Runoff would be between 25254.2022134 acre-feet and 62547.3427612 acre-feet, on average.

Appendix of Code

```r
knitr::opts_chunk$set(echo = FALSE, include = FALSE)
library(tinytex)
library(ggplot2)
library(MASS)
library(normtest)
library(lmtest)
library(SciViews)
options(scipen = 999)
water = read.table("~/R programming/STAT_330/water.txt", sep = ' ', header = TRUE)
```

```r
#x=Precip is explanatory variable; y=Runoff is response variable
water.scatter <- ggplot(data = water, mapping=aes(x=Precip, y=Runoff)) +
  geom_point() +
  geom_smooth(se=FALSE) +
  xlab('Snow Precipitation (inches)') +
  ylab('Stream Runoff (acre-feet)') +
  ggtitle('Precipitation and Runoff: Raw Data')
suppressMessages(print(water.scatter))
#Calculate the correlation and covariance between Speed and Distance
water.cov <- round(cov(water$Precip, water$Runoff), digits = 4)
water.cor <- round(cor(water$Precip, water$Runoff), digits = 4)
#Provide a summary of the main features of the data
water.Precip.box <- ggplot(data = water, mapping=aes(Precip)) +
  geom_boxplot() +
  ggtitle("Precipitation: Outlier Check")
water.Runoff.box <- ggplot(data = water, mapping=aes(Runoff)) +
  geom_boxplot() +
  ggtitle("Runoff: Outlier Check")
suppressMessages(print(water.Precip.box))
suppressMessages(print(water.Runoff.box))
#Fit a simple linear model to the water data
water.slm <- lm(formula=Runoff~Precip, data=water)
#Identify the estimates beta sub zero, beta sub one, and sigma squared.
water.beta.0 <- round(as.numeric(coef(water.slm)["(Intercept)"]), digits = 4)
water.beta.1 <- round(as.numeric(coef(water.slm)["Precip"]), digits = 4)
water.var <- round(sigma(water.slm)^2, digits = 4)
water.r2 <- round(summary(water.slm)$r.squared, digits = 4)
#prettyNum(x, big.mark = ",", scientific = FALSE)
#Add your estimated regression line to the scatterplot you created above.
water.est.reg <- ggplot(water, aes(x=Precip,y=Runoff)) +
  geom_point() +
  geom_smooth(method="lm",se=FALSE) +
  xlab('Precipitation') +
  ylab('Runoff') +
  ggtitle('Precipitation and Runoff: Fitted Regression, Raw Data')
suppressMessages(print(water.est.reg))
# Draw a fitted values vs. residuals plot to check the L and E assumption.
fit.vs.resids.1 <- ggplot(water, aes(x=water.slm$fitted.values, y=water.slm$residuals)) +
  geom_point() +
  xlab('Fitted Values') +
  ylab('Residuals') +
  ggtitle('Linearity and Equal Variance Assumption Checks of Raw Data')
suppressMessages(print(fit.vs.resids.1))
# Draw a histogram (or density plot) of standardized residuals to check the N assumption.
standardized.residuals <- stdres(water.slm)
water.freq <- ggplot() +
  geom_histogram(mapping=aes(x=standardized.residuals)) +
  xlab('Standardized Residual') +
  ylab('Frequency') +
  ggtitle('Normality Assumption Check of Transformed Data')
suppressMessages(print(water.freq))
# Conduct a KS and JB test for normality.
std.resids <- stdres(water.slm)
```

```r
ks.test(std.resids, "pnorm")
jb.norm.test(std.resids)
# Conduct a BP test for equal variance.
bptest(water.slm)
# Identify any outlying observations using Cook’s distance.
cooks <- cooks.distance(water.slm)
outlier.where <- 4/length(water$Precip)
outliers <- water[which(cooks>outlier.where),]
#outliers <- round(cooks[which(cooks>outlier.where)], 4) to get cook's distance of outliers
#cross validation
#set seed for reproducibility
set.seed(2)
n.cv <- 100 #Number of CV studies we’ll run
bias <- rep(NA, n.cv) #n.cv empty biases (one for each CV)
RPMSE <- rep(NA, n.cv) #n.cv empty RPMSE (one for each CV)
coverage <- rep(NA, n.cv) #n.cv empty coverage (one for each CV)
width <- rep(NA, n.cv) #n.cv empty width (one for each CV)
n.test <- 5 #How big my test set is
for(i in 1:n.cv){
  # Choose which obs. to put in test set
  test.obs <- sample(1:nrow(water), n.test)

  # Split data into test and training sets
  test.set <- water[test.obs,]
  train.set <- water[-test.obs,]

  # Using training data to fit a (possibly transformed) model
  train.lm <- lm(Runoff~Precip,data=train.set)

  # Predict test set
  test.preds <- predict.lm(train.lm, newdata=test.set, interval="prediction")

  # Calculate bias
  bias[i] <- mean(test.preds[,1]- test.set$Runoff)

  # Calculate RPMSE
  RPMSE[i] <- sqrt(mean((test.preds[,1] - test.set$Runoff)^2))

  #coverage
  coverage[i] <- mean((test.preds[,2] < test.set$Runoff) &
                      (test.preds[,3] > test.set$Runoff))

  #width
  width[i] <- mean(test.preds[,3] - test.preds[,2])
}

water.stddev <- round(sigma(water.slm), digits = 4)
CV.bias <- ggplot() +
  geom_histogram(mapping=aes(x=bias)) +
  xlab('Amount of Bias') +
  ylab('Frequency') +
  ggtitle('Range of Bias for Cross Validation')
suppressMessages(print(CV.bias))
```

```r
CV.RPMSE <- ggplot() +
  geom_histogram(mapping=aes(x=RPMSE)) +
  xlab('RMPSE') +
  ylab('Frequency') +
  ggtitle('Range of RPMSE for Cross Validation')
suppressMessages(print(CV.RPMSE))

CV.coverage <- ggplot() +
  geom_histogram(mapping=aes(x=coverage)) +
  xlab('Coverage') +
  ylab('Frequency') +
  ggtitle('Range of Coverage for Cross Validation')
suppressMessages(print(CV.coverage))

CV.width <- ggplot() +
  geom_histogram(mapping=aes(x=width)) +
  xlab('Width') +
  ylab('Frequency') +
  ggtitle('Range of Width for Cross Validation')
suppressMessages(print(CV.width))
print("Hypothesis Tests:")
summary(water.slm)
#Do I need to center the data first? Only if I center everything
# center.data <- water
# center.data$Precip <- center.data$Precip - mean(center.data$Precip)
# center.slm <- lm(Runoff~Precip,data=center.data)
water.CI <- confint(water.slm, level = 0.95)
#confint(center.slm, level = 0.95)
water.predict <- data.frame(Precip = 4.5)
prediction <- predict.lm(water.slm, newdata=water.predict,
                         interval="prediction",
                         level=0.95)
#End of homework's code
```