Homework 3

Simple Linear Regression Model Inference

AUTHOR

Zeb Sorenson

Data and Description

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 Runoff Water data set compares the **stream runoff (column 2)** (in acre-feet) of a river near Bishop, California (due east of San Jose) with **snowfall (column 1)** (in inches) at a site in the Sierra Nevada mountains. The data set contains 43 years' worth of measurements. Download the water.txt file from Canvas, and put it in the same folder as this R Markdown file.

0. Replace the text "< PUT YOUR NAME HERE >" (above next to "author:") with your full name.

1. Read in the data set, and call the dataframe "water". Print a summary of the data and make sure the data makes sense.

```
water <- read.csv("~/Desktop/Stat 330/water.txt", sep="")

head(water)
```

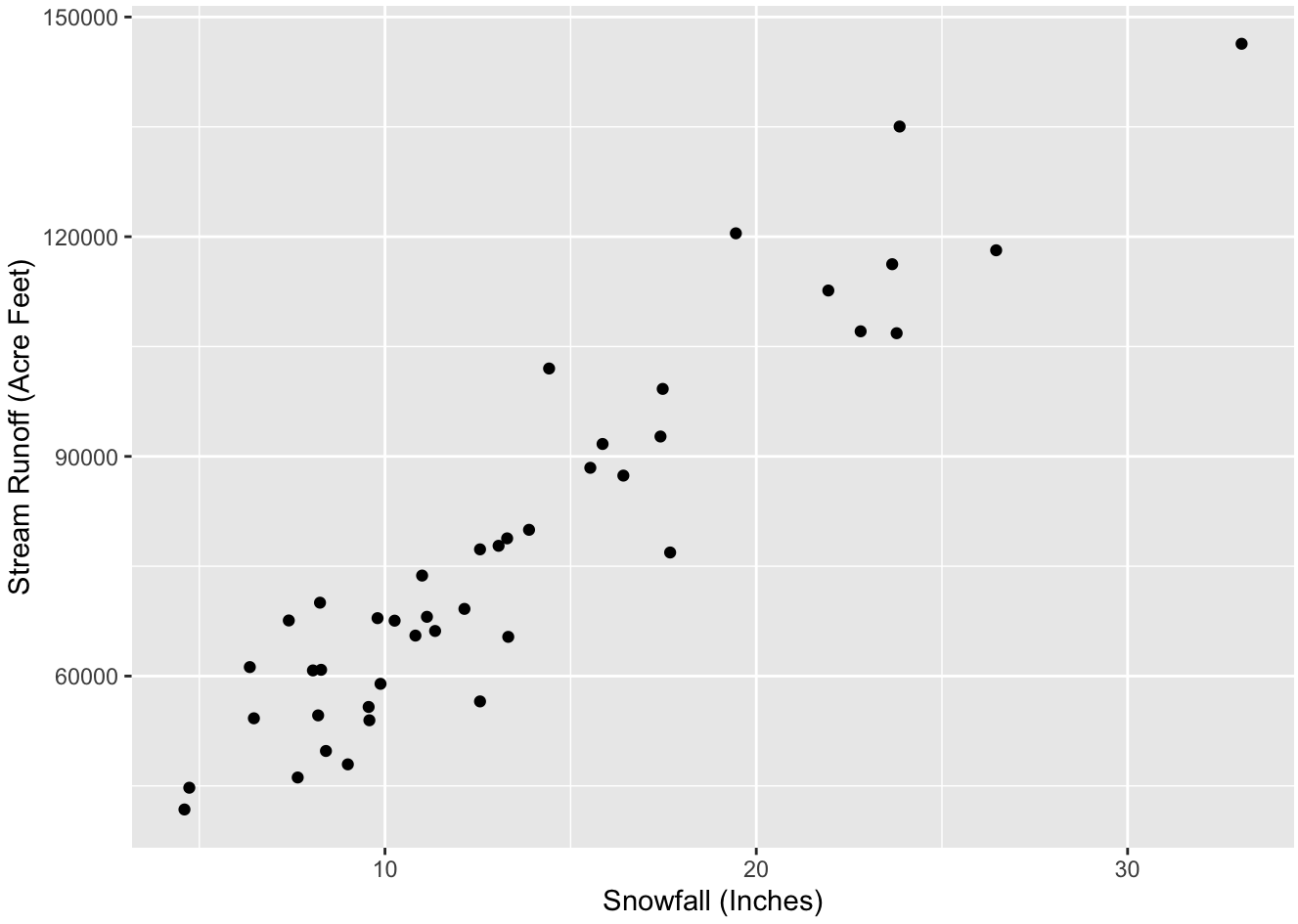
```
  Precip Runoff
1   6.47  54235
2  10.26  67567
3  11.35  66161
4  11.13  68094
5  22.81 107080
6   7.41  67594
```

```
summary(water)
```

```
     Precip           Runoff      
 Min.   : 4.600   Min.   : 41785  
 1st Qu.: 8.705   1st Qu.: 59857  
 Median :12.140   Median : 69177  
 Mean   :13.522   Mean   : 77756  
 3rd Qu.:16.920   3rd Qu.: 92206  
 Max.   :33.070   Max.   :146345  
```

2. Create (and print) a scatterplot of the data with variables on the appropriate axes. Make you plot look professional (e.g., axis labels are descriptive). You should save your plot as an object to be used throughout the rest of the assignment.

```
water_scatter_plot <- ggplot(data = water, aes(Precip, Runoff)) +
  geom_point()+
  xlab("Snowfall (Inches)") +
  ylab("Stream Runoff (Acre Feet)")

print(water_scatter_plot)
```



3. Calculate (and report) the correlation coefficient. Use that and the scatterplot to briefly describe the relationship between Stream Runoff and Snowfall.

```
snow_runnoff_cor<- cor(water$Precip, water$Runoff)
```

```
print(snow_runnoff_cor)
```

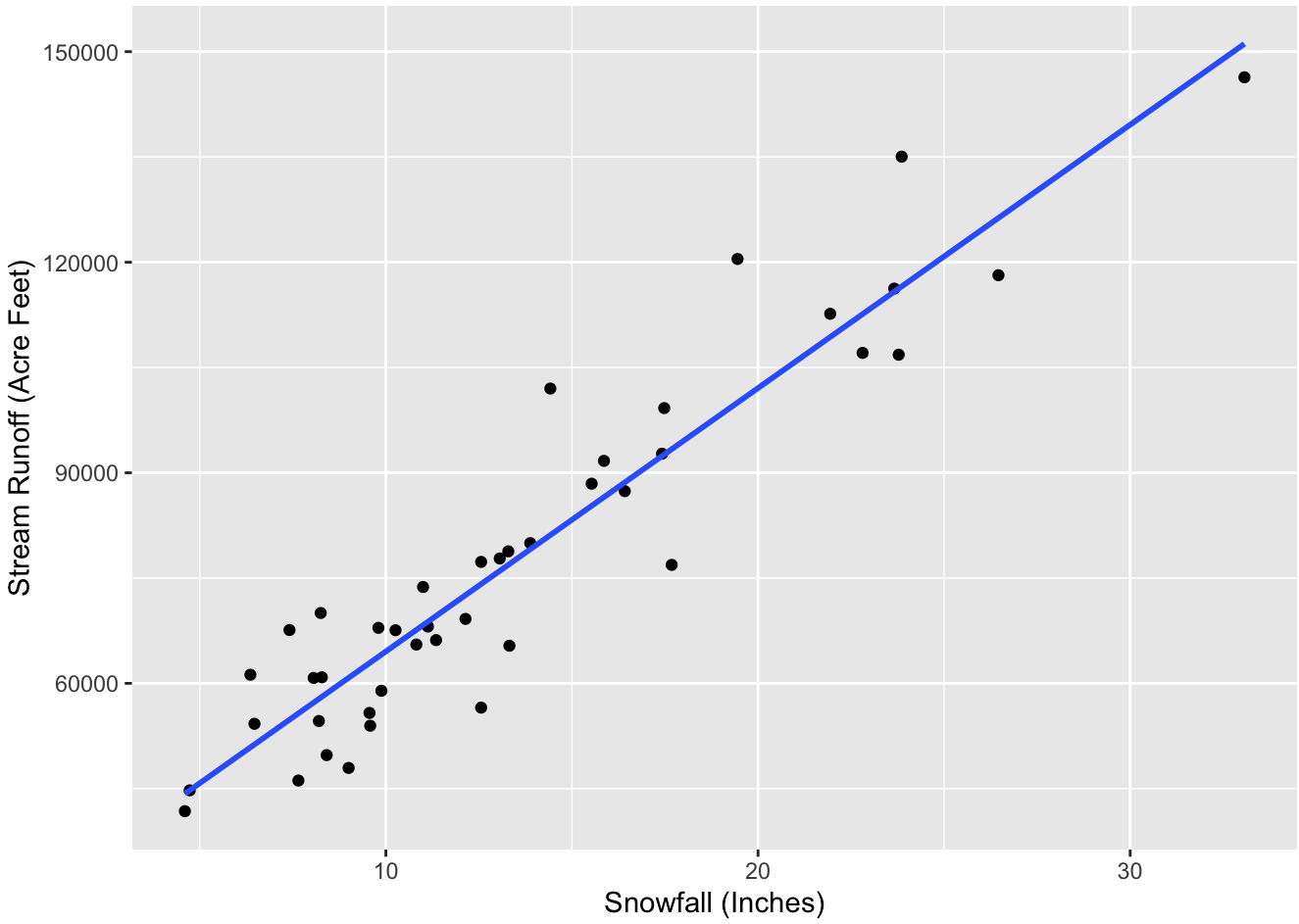
```
[1] 0.938436
```

Looking at both our scatter plot and out correlation coefficient of 0.938436, we see that Runoff and Snowfall have a positive, strong, linear relationship with each other. When one increases, the other does as well in a linear fashion.

4. Add the OLS regression line to the scatterplot you created in 2. Show the plot.

```
water_scatter_plot + geom_smooth(method = "lm", se = FALSE)
```

```
`geom_smooth()` using formula = 'y ~ x'
```



5. Fit a simple linear regression model to the data (no transformations), and save the residuals and fitted values to the `water` dataframe. Print a summary of the linear model.

```
water_lm <- lm(Runoff ~ Precip, data = water)

water$residuals <- water_lm$residuals

water$fitted <- water_lm$fitted.values
```

```
head(water)
```

```
  Precip Runoff  residuals    fitted
1   6.47  54235  2941.8309  51293.17
2  10.26  67567  2051.9105  65515.09
3  11.35  66161 -3444.2988  69605.30
4  11.13  68094  -685.7519  68779.75
5  22.81 107080 -5528.7836 112608.78
6   7.41  67594 12773.4945  54820.51
```

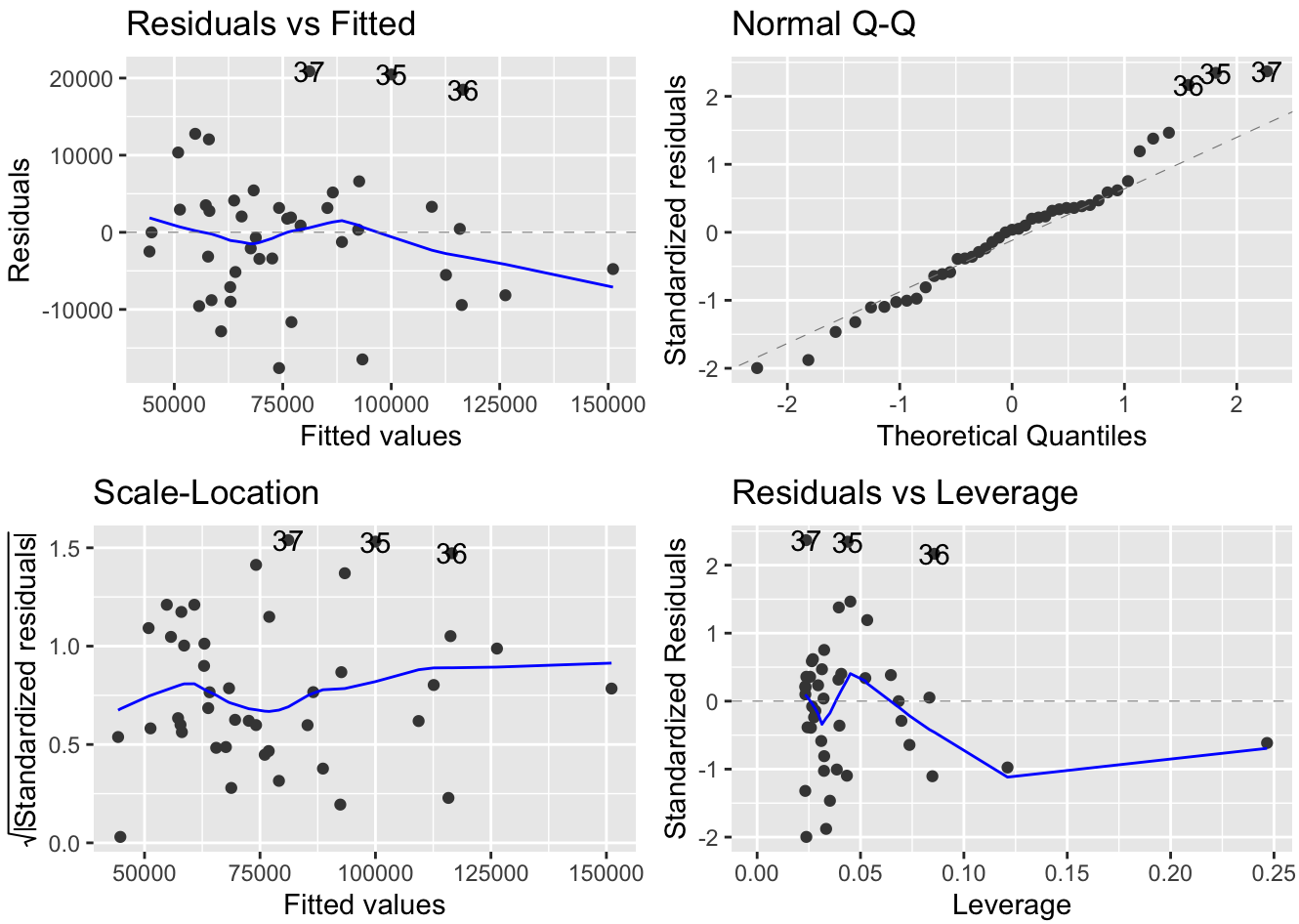
```
summary(water_lm)
```

```
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 1.93e-10 ***
Precip        3752.5      215.7  17.394  < 2e-16 ***
---
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: < 2.2e-16
```

Questions 6 to 10 involve using diagnostics to determine if the linear regression assumptions are met and if there are influential observations. For each assumption, (1) perform appropriate diagnostics to determine if the assumption is violated, and (2) explain whether or not you think the assumption is violated and why you think that.

6. (L) X vs Y is linear

```
autoplot(water_lm)
```



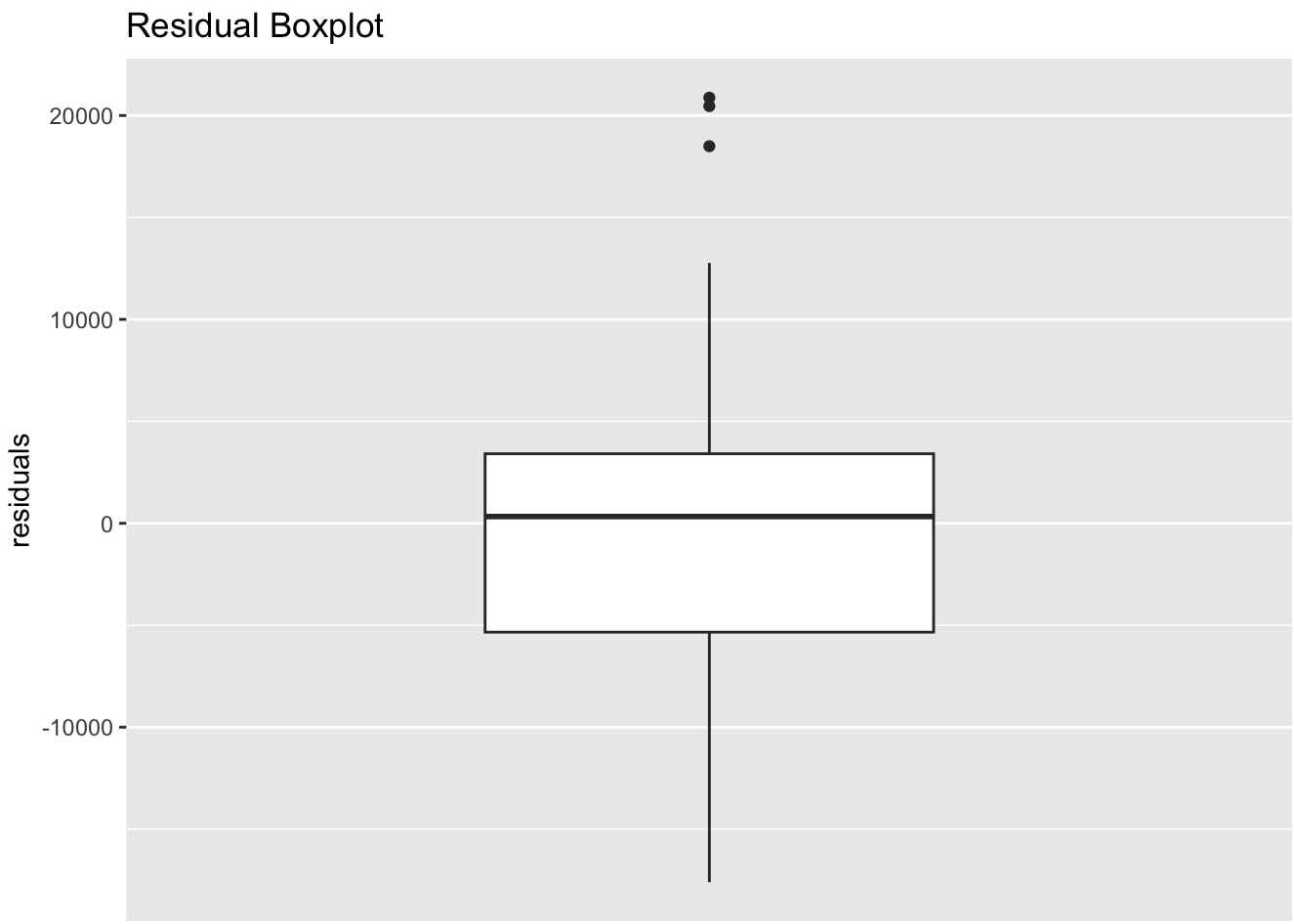
We can look at the residuals vs fitted plot given by the autplot plot function. It looks like our data is ALMOST linear, but there is some major curvature around the 0 mark, especially with higher x values, so this assumption that that X vs Y is linear is violated

7. (I) The residuals are independent (no diagnostic tools - just think about how the data was collected and briefly write your thoughts)

If we were given more specific details about the data collection, specifically if the samples were random or not, we could more easily make this prediction. If random samples are taken, or possibly, if each sample is taken independently of each other, it can more easily accept independence. Because no such statement is made about the data, we would assume that independence is violated in the real world. For this assignement, we will continue working with the data.

8. (N) The residuals are normally distributed (use at least three diagnostic tools)

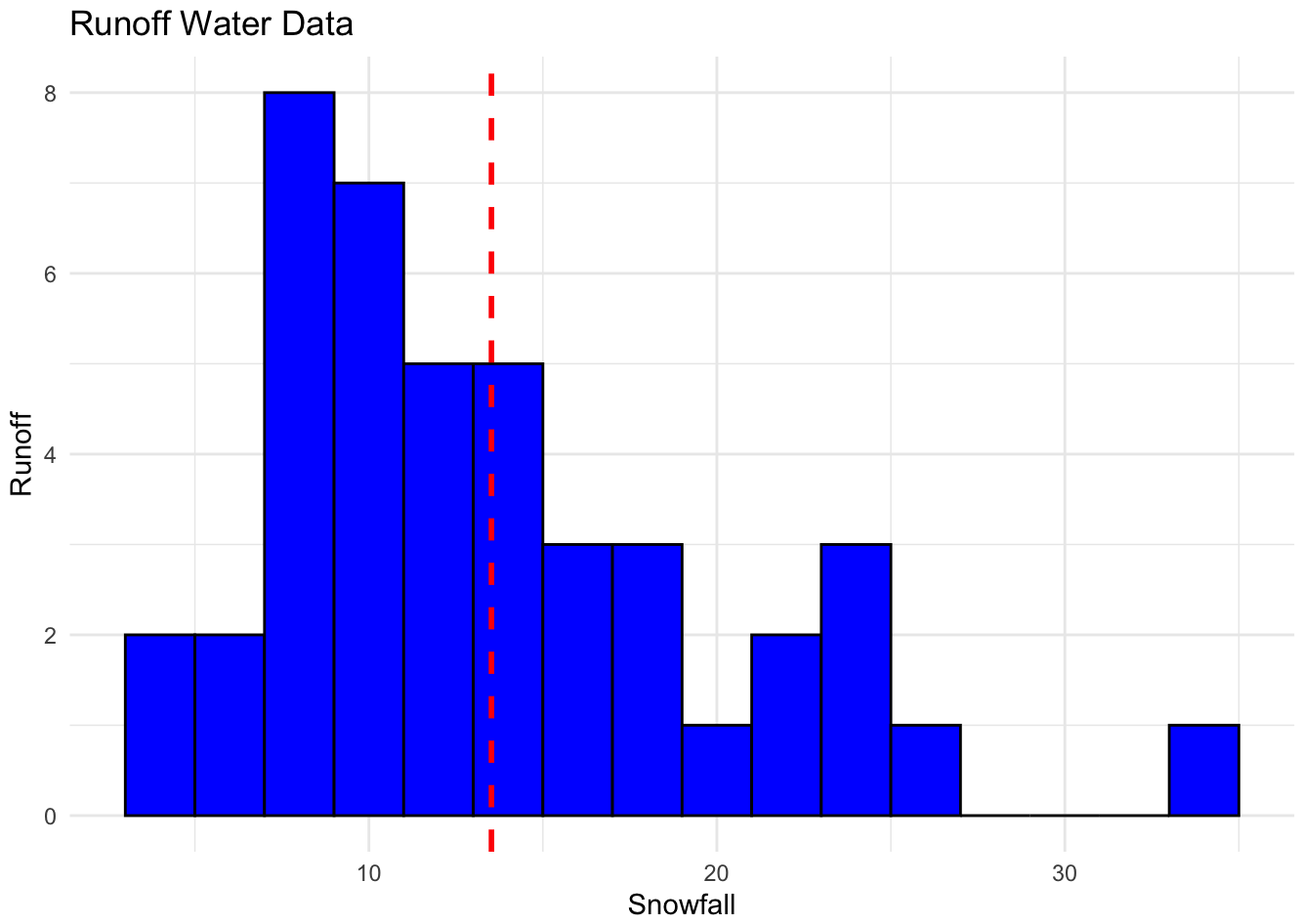
```
#1 Box plot

water$residuals <- water_lm$residuals
# The code below produces a basic boxplot...This is modified from the in class completed #2
water_residual_box <- ggplot(data = water, mapping = aes(y = residuals)) +
  geom_boxplot() +
  scale_x_discrete() +
  labs(title = "Residual Boxplot")

print(water_residual_box)
```




```
#2 Histogram...This is a combination of the histogram from completed #2
#and modification help from chat GPT...I'm still working on getting better at coding plots

# Calculate the mean of Precip
mean_precip <- mean(water$Precip)

# Create the histogram with a line at the mean of Precip
water_histogram <- ggplot(data = water, aes(x = Precip)) +
  geom_histogram(binwidth = 2, fill = "blue", color = "black") +
  geom_vline(xintercept = mean_precip, color = "red", linetype = "dashed", size = 1) +
  labs(x = "Snowfall", y = "Runoff", title = "Runoff Water Data") +
  theme_minimal()
```

```
Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
ℹ Please use `linewidth` instead.
```

```
print(water_histogram)
```



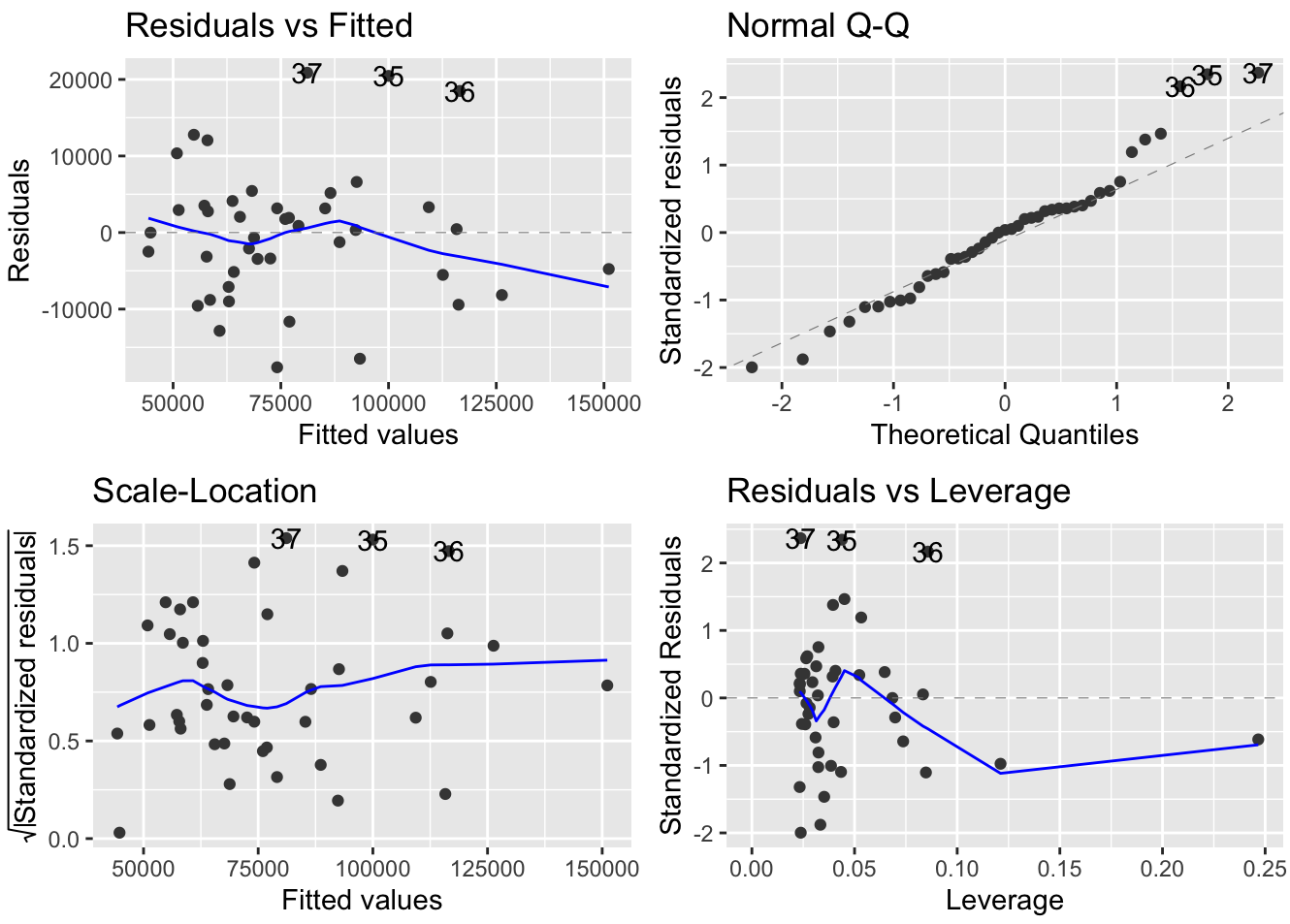

```
#3 Shapiro Wilk Test

shapiro_p_value <- shapiro.test(water$residuals)

print(shapiro_p_value)
```

```
    Shapiro-Wilk normality test

data:  water$residuals
W = 0.96775, p-value = 0.2639
```

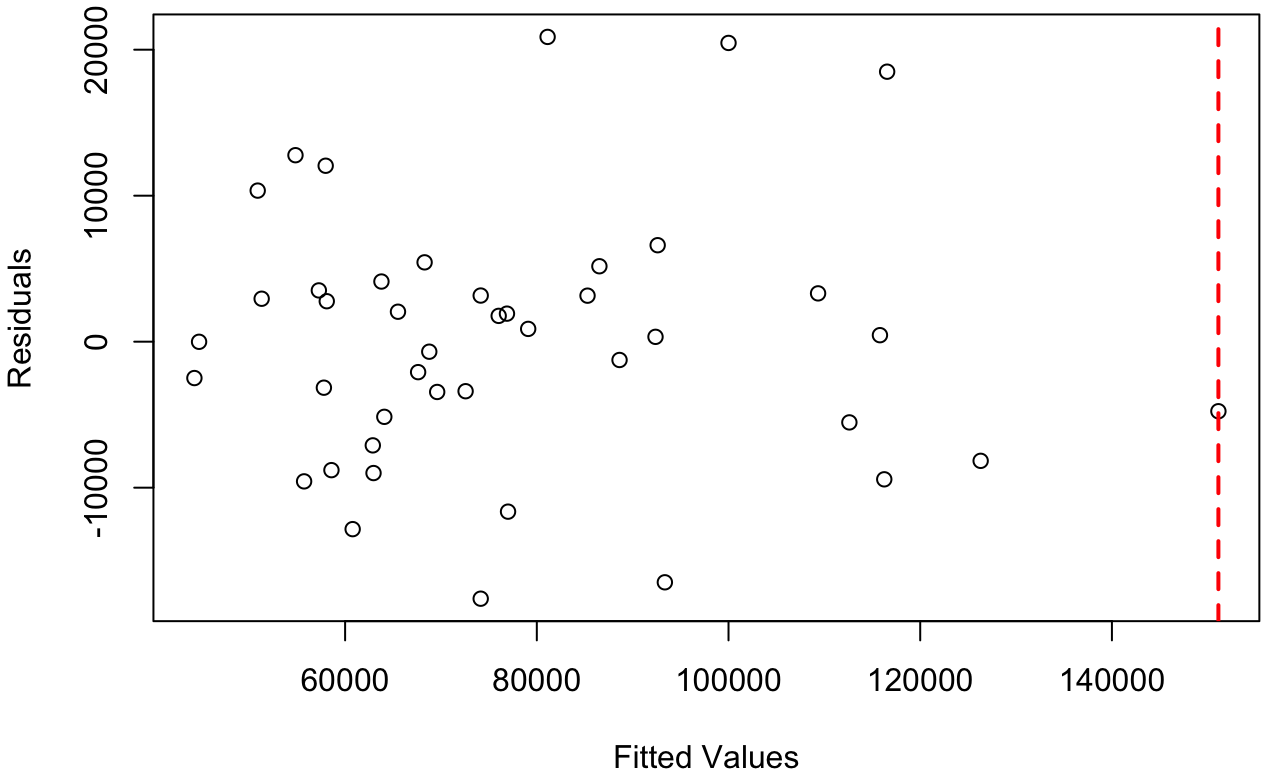
Looking at our three plots we created, I would feel okay with this assumption. Looking at the histogram, there is a bit of right skewness, along with our box plot but I believe that it is close enough to be able to work with this assumption. This is confirmed after running the Shapiro Wilk test with our residuals and getting a large p value of 0.2639. So we can accept that normal distribution is met.

9. (E) The residuals have equal (constant) variance across all values of X (homoscedastic) (use two diagnostic tools)

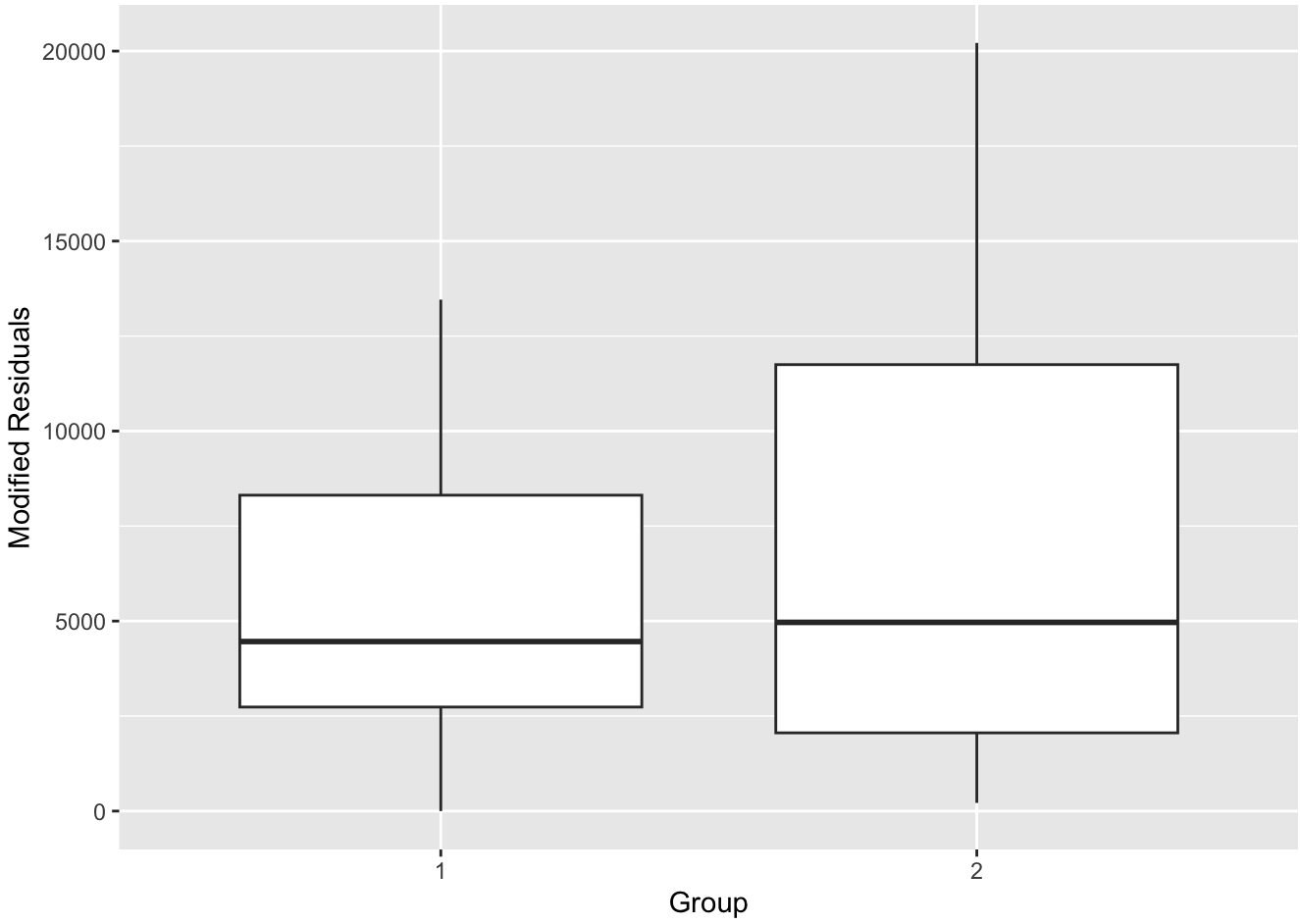
```
#sqrt of the standardized residuals with the fitted values

autoplot(water_lm) #does this also help with constant variance? Pg 11 of lecture 2 module 2
```




```
#brown forsithe

water$resids <- water_lm$residuals
water$fits <- water_lm$fitted.values

water_bw <- water |>
arrange(fits) |>
mutate(index = 1:nrow(water),
e_group = ifelse(index < nrow(water)/2, 1, 2) |>
as.factor()) |>
mutate(vals = abs(resids - median(resids)), .by = "e_group")
plot(water$fits, water$resids, xlab = "Fitted Values", ylab = "Residuals")
abline(v = water$fits[round(nrow(water_bw)/2)],
col = "red", lty = 2, lwd = 2)
```




```
water_bw |>
ggplot(aes(e_group, vals)) +
geom_boxplot() +
labs(x = "Group", y = "Modified Residuals")
```




```
bf.test(vals ~ e_group, water_bw)
```

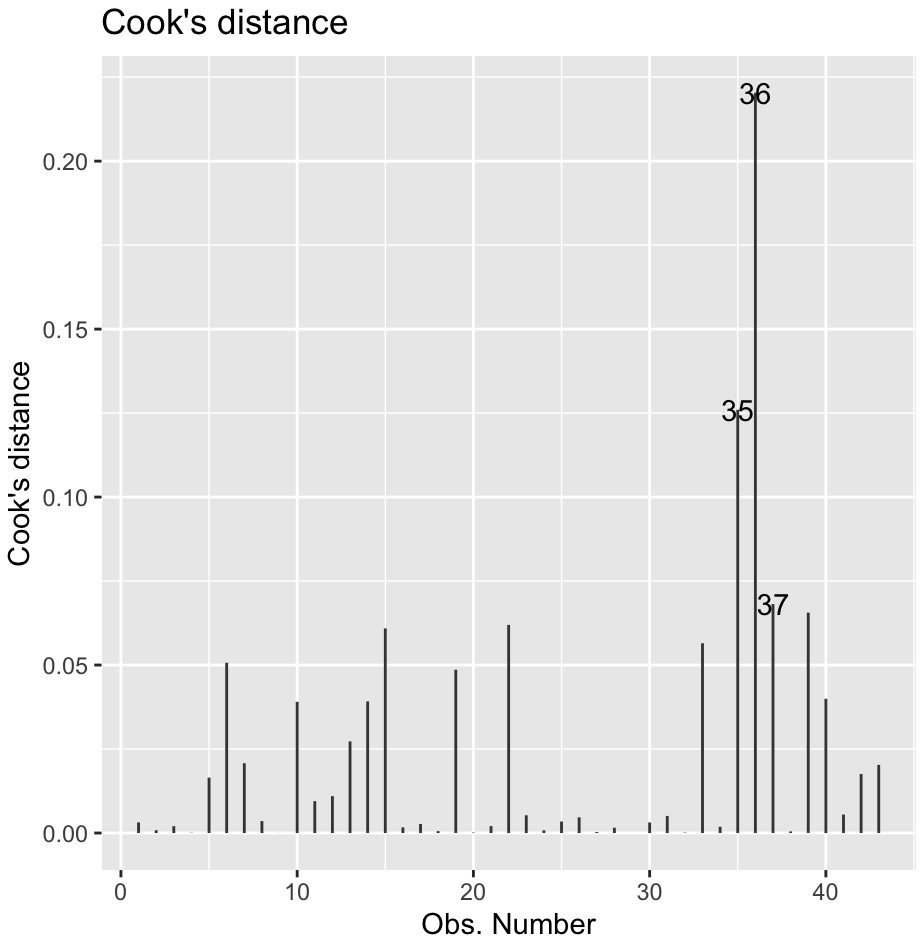
```
  Brown-Forsythe Test (alpha = 0.05)
-------------------------------------------------------------
  data : vals and e_group

  statistic  : 0.9807822
  num df     : 1
  denom df   : 34.15676
  p.value    : 0.3289699

  Result     : Difference is not statistically significant.
-------------------------------------------------------------
```

We can look at plot of the sqrt of the standardized residuals with the fitted values given to us by the autplot function and see that the line does stray from zero is some areas but nothing major. We can also run a Brown-Forsythe Test of Constant Variance and get a fairly large p value of 0.3289699 and see that the difference is not statistically significant, therefore we can assume that constant variance is NOT violated and we can proceed.

10. Check for influential points and report your findings.

```
# 'which = 4' produces a plot of Cook's distance.
# Values greater than .5 deserve attention
```

```r
# Modified code given in the in class activity #2. Thank you Dr. Sandholtz

autoplot(water_lm, which = 4, ncol = 1, nrow = 1)  +
 theme(aspect.ratio = 1)
```



We can run a test to look at Cook's distance. 'which = 4' produces a plot of Cook's distance. Values greater than .5 deserve attention, however, we do not have any such value in our data, so we can assume that there are no influential points that deserve attention in our findings.

Based on your answers to questions 6 through 10, you may (or may not) have decided a transformation to the data is needed. This was, hopefully, good practice for assessing model assumptions. For simplicity for this assignment, we will use the orignial model (no transformations) for the rest of the questions.

11. Mathematically write out the fitted simple linear regression model for this data set using the coefficients you found above (do not use betas or matrix notation). Do not use "X" and "Y" in your model - use variable names that are fairly descriptive.

$$\widehat{\text{Estimated\_Stream\_Runoff}}_i = 27014.6 + 3752.5 \times \text{Snowfall}_i$$

12. Compute a 95% confidence interval for the slope using the output from the lm() function (to get the standard error of beta_1) in tandem with the qt() function (to get

the correct critical value).

```
S_XX <- sum((water$Precip - mean(water$Precip))^2)
s_e <- sqrt(sum(water$residuals^2/(nrow(water) - 2)))

s_e <- summary(water_lm)$sigma

b1_hat_std_err <- s_e / sqrt(S_XX)


print(b1_hat_std_err)
```

```
[1] 215.7304
```

```
df <- length(water_lm$residuals) - 2


qt( 0.975,df)
```

```
[1] 2.019541
```

```
margin_of_error <- (qt( 0.975,df)) * b1_hat_std_err

print(margin_of_error)
```

```
[1] 435.6765
```

```
uppr <- margin_of_error + 3752.5
lwr <- margin_of_error - 3752.5


#this last number is what you will add and subtract from the slope to get your interval

#Interval = prediction

cat("The 95% confidence interval is (", lwr, ",", uppr, ")\n")
```

```
The 95% confidence interval is ( -3316.824 , 4188.176 )
```

13. Compute a 95% confidence interval for the slope using the confint() function in R (you should get the same answer as in 12). Interpret the confidence interval.

```
CI_slope_95 <- confint(water_lm, par = 2, level = 0.95)
CI_slope_95
```

```
          2.5 %   97.5 %
Precip 3316.809 4188.162
```

We are 95% confident that the change in average stream runoff at the California location will be between 3316.809 and 4188.162 acre feet when the snowfall at the Nevada location increases by 1 inch.

14. Based on the confidence interval, is there a statistically significant linear association between snowfall and stream water? Why or why not?

Because zero is not in our confidence interval, we can assume that there is a positive linear relationship between snowfall and stream runoff.This is because it shows us that the effect of the predictor variable, snowfall on the response of runoff is statistically significant at a 95% confidence level.

15. Print a summary of the linear model. Interpret the results from the hypothesis test output for the slope.

```
summary(water_lm)
```

```
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 1.93e-10 ***
Precip        3752.5      215.7  17.394  < 2e-16 ***
---
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: < 2.2e-16
```

```
#need to say that the p value is low so it is stat significant
```

For each inch of snowfall at the Nevada location, our estimated runoff will increase by 3752.5 Acre-Feet. Because of the extremely small p value of 2e-16 we know that these two variables have a statistically significant association.

16. Briefly describe the difference between (1) a confidence interval for the slope, (2) a confidence interval for the conditional mean of Y given x, and (3) a prediction interval for individual observations.

A confidence interval for slope will give us a range of values in which our actual or true slope will be located under a certain confidence (usually 90-95%). This can help us see where our OLS line falls within this range and help us make better judgement about our model.

For conditional mean, for our y given an x value will give us a range in which our actual y value, this case our stream runoff mean value is to be found given a specified x and within the specified confidence level.

Prediction interval for individual observations will estimate the range that a completely new value of Y (Stream Runoff) will be found given a specific x value. This method will take into account uncertainty and variability to its calculation.

17. Compute, print, *and interpret* a 95% confidence interval for the average of Y when $x_i = 30$. You may use the predict() function in R.

```
x_new <- data.frame(Precip = 30)

predicted_Runoff <- predict(water_lm, x_new, interval = "confidence", level = .95)

print(predicted_Runoff)
```

```
       fit      lwr      upr
1 139589.2 131902.2 147276.1
```

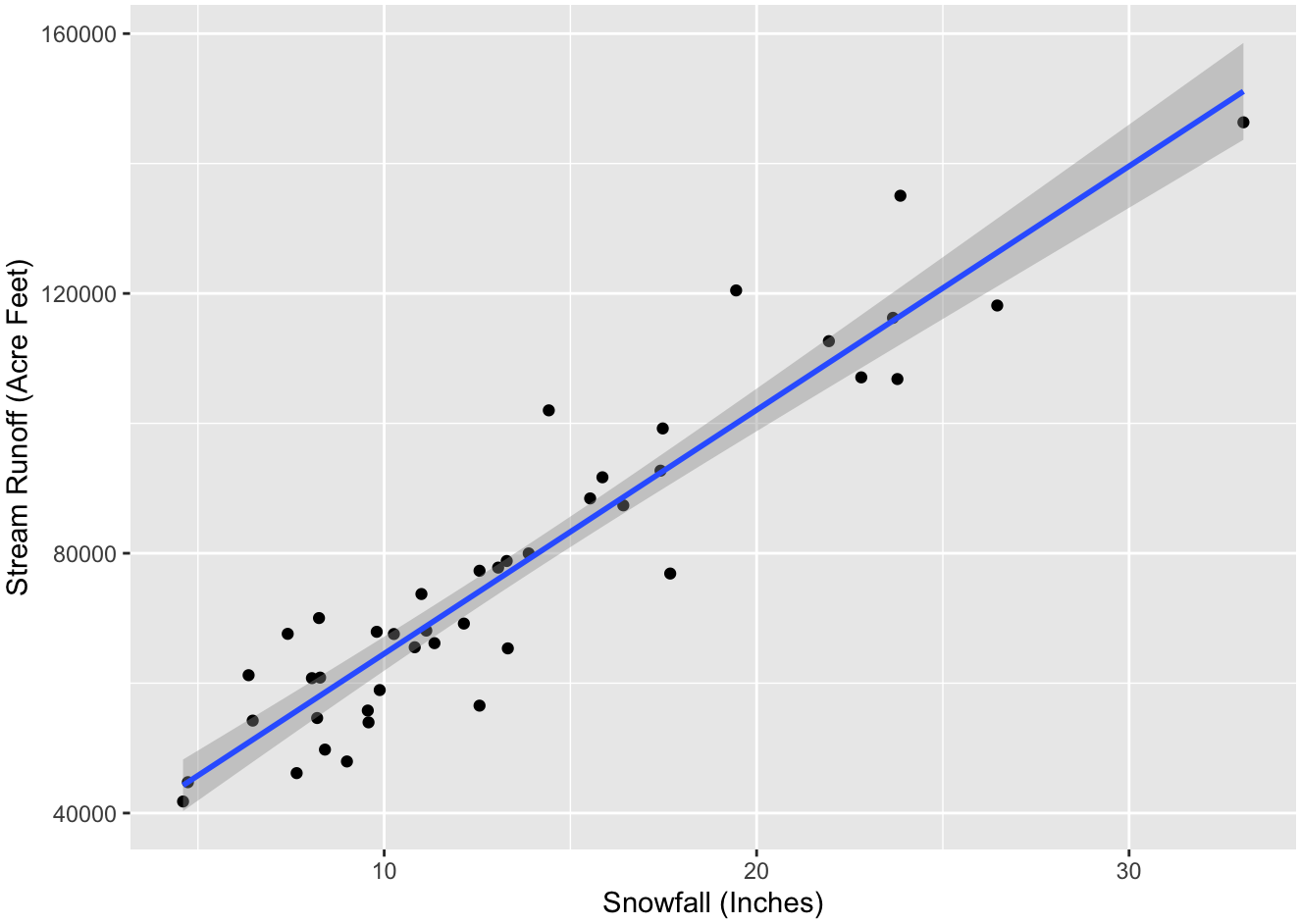
```
#average runoff here between upper and lower, when there are 30

#20 95 could take on any of these two....Prediction will be a lot wider and all values of Y,
```

We are 95% confident that the average Stream Runoff at the California location will be between 131902.2 and 147276.1 Acre Feet when snowfall is 30 inches at the Nevada location

18. Create a confidence band for the average of Y over a sequence of X values spanning the range of the data, and overlay this band (using a distinct color) on your previous scatterplot that you created in 4. Print the plot.

```
water_scatter_plot +

  geom_smooth(method = 'lm',
              se = TRUE,
              level = .9) # for a 90%  CI
```

```
`geom_smooth()` using formula = 'y ~ x'
```



19. Briefly explain why the confidence band is shaped the way that it is.

In the areas where we have less data, (where snowfall is high) our confidence decreases in where the true value is, which results in a wider band to take on more possible values. Whereas where we have more data, we are more confident of the true value, which gives us a more narrow band of possible values.

20. Compute, print, *and interpret* a 95% prediction interval for Y when $x_i = 30$. You may use the predict() function in R.

```
x_new_prediction <- data.frame(Precip = 30)


predicted_Runoff_conf <- predict(water_lm, x_new_prediction, interval = "prediction")

print(predicted_Runoff_conf)
```

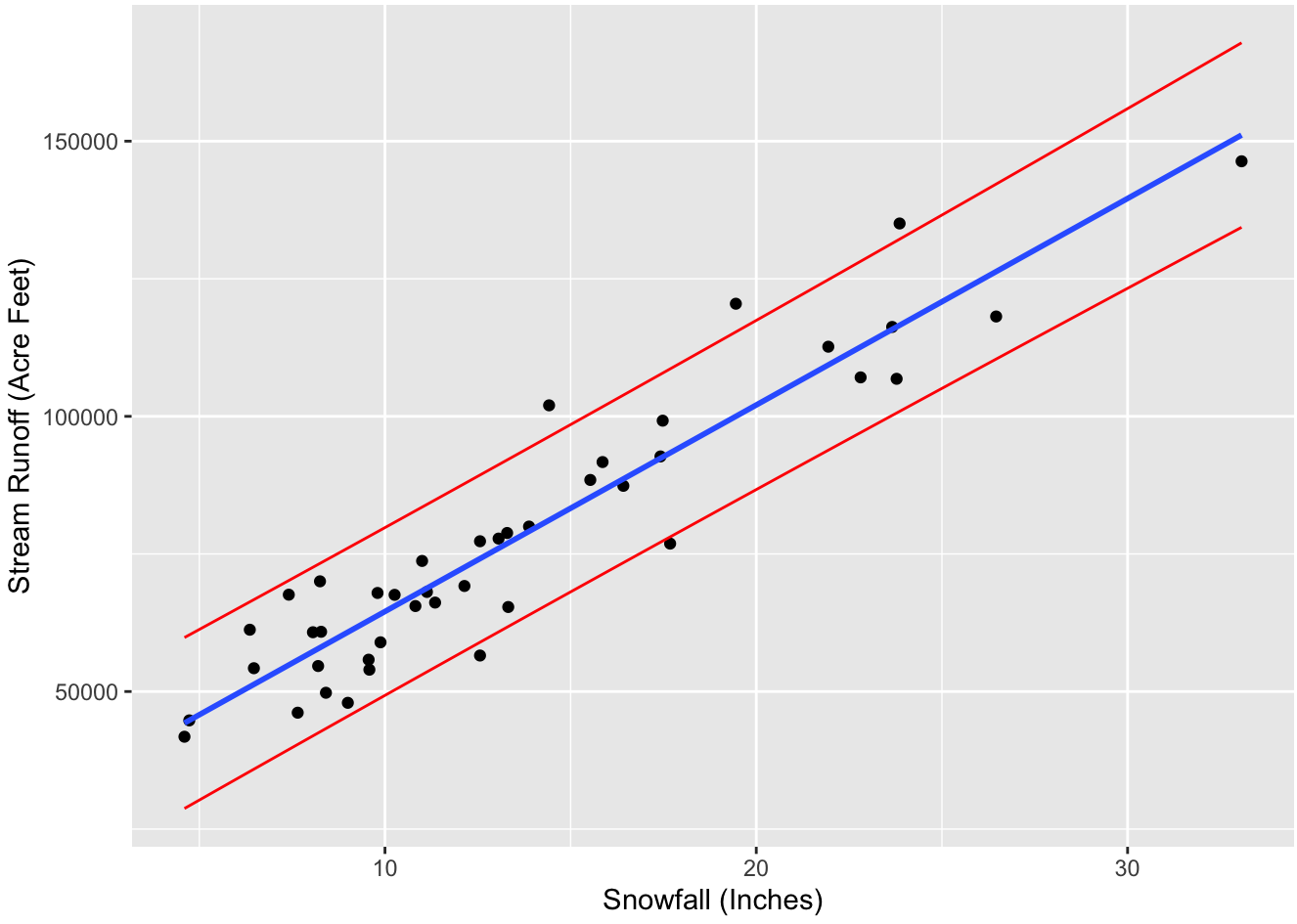
```
       fit      lwr      upr
1 139589.2 119998.8 159179.5
```

We are 95% confident that for an individual future observation of Stream runoff, that the Stream runoff at the California location for a specific prediction if the Snowfall at the Nevada location is 30 inches, will be between 119998.8 and 159179.5 Acre Feet for this specific future prediction.

21. Create a prediction band for Y over a sequence of X values spanning the range of the data, and overlay this band (using a distinct color) on your previous scatterplot that you created in 4. Print the plot.

```
# Create Precip_grid for prediction
Precip_grid <- seq(min(water$Precip), max(water$Precip), length.out = 100)

# Predict with prediction interval
water_PI <- predict(water_lm,
                    level = 0.9,
                    newdata = data.frame(Precip = Precip_grid),
                    interval = "prediction") %>%
  as_tibble() %>%
  mutate(Precip = Precip_grid)

# Base plot
water_scatter_plot +
  geom_smooth(method = 'lm', se = FALSE) +
  geom_line(data = water_PI, aes(x = Precip, y = lwr), color = "red") +
  geom_line(data = water_PI, aes(x = Precip, y = upr), color = "red")
```

```
`geom_smooth()` using formula = 'y ~ x'
```




```
#This is modified code from the in class activity #3 and help from chatGPT. Again, plots are
```

```
#find that it saves a lot of time using it for plots.
```

22. Briefly explain how/why the prediction band differs from the confidence band.

Our prediction band is looking to estimate a set of values in which individual future observations will be found based on our specified confidence of .9. Giving us a much different looking band as it is looking for future predictions based on the given x value of Snowfall.

23. What is the MSE (Mean Square Error) for the linear model you fit? Hint: you may refer to the ANOVA table results.

```
#This is modified code found on statology.org from the page called, "How to Calculate MSE in

water_summary <- summary(water_lm)

MSE <- (mean(water_summary$residuals^2))

print("Mean Squared Error (MSE):")
```

```
[1] "Mean Squared Error (MSE):"
```

```
print(MSE)
```

```
[1] 75907220
```

```
aovWater<- aov(water_lm)

summary(aovWater)
```

```
            Df    Sum Sq   Mean Sq F value Pr(>F)
Precip       1 2.409e+10 2.409e+10   302.6 <2e-16 ***
Residuals   41 3.264e+09 7.961e+07
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#look at the mean square of the residuals in this output....

#You can also just sqaure sigma from the summary

newMSE<- summary(water_lm)

newMSE$sigma^2 #this also works
```

```
[1] 79610011
```

24. Briefly explain (1) what the MSE estimates and (2) a drawback to using it as a model evaluation metric.

MSE will measure the average squared difference between what we actually record and our predicted values for our model of Snowfall and Runoff. It will give us a good big picture view of the model by taking into account error values that stray far from the others. It does this by squaring our residuals. So in the context of our data, the MSE will be the average squared diff between the actual water runoff and the predicted value. So, on average the model deviates from our observed water runoff by 75,907,220 square units

This can be quite useful but it will also be sensitive to outliers due to giving the same weight to all residuals, causing outliers to potentially skew our model in such a way that is no longer valuable to us.

25. Calculate the RMSE (Root Mean Square Error) for the linear model you fit. Print and interpret the result.

```
RSME <- sqrt(MSE)

print("Root Mean Squared Error (RMSE):")
```

```
[1] "Root Mean Squared Error (RMSE):"
```

```
print(RSME)
```

```
[1] 8712.475
```

The RMSE is the square root of the MSE.Meaning, it often gives us a better estimate of how good our model is and giving us insight into how well our model performs with larger residual values.

This will give us an average difference between our predicted water runoff and our actual water runoff. Meaning, that on average, we have a difference of 8712.475 of acre-feet water runoff

26. Print a summary of the linear model. Briefly interpret the R-Squared (Coefficient of Determination) value.

```
summary(water_lm)
```

```
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 1.93e-10 ***
Precip        3752.5      215.7  17.394  < 2e-16 ***
---
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: < 2.2e-16
```

Looking at our summary table for our model we see that the R-Squared value of 0.8807 is the proportion of variance in the dependent variable, in our case being Runoff, that is predictable from the independent variable snowfall.

27. What is the difference between the R-Squared value and the Adjusted R-Squared (shown in the summary output above)?

NOTE: Also received help from this from a page called "R vs R-Squared: What's the Difference" page on statology.org. But the words and interpretation are my own.

R-squared gives us the proportion of variance in the dependent variable which is water runoff explained by predictors Snowfall/Precipitation, R-squared balances for model complexity by penalizing unneeded predictors, giving us a more precise assessment in our model. 0.8807 indicates that about 88.07% of water runoff variability can be explained by precipitation, while the adjusted R-squared (0.8778) considers the model's complexity, providing a slightly more controlled estimate of the explanatory variable.

28. Look at the F-Statistic and corresponding p-value from the summary of the linear model (output shown above). Do these values indicate that X has a statistically significant linear association with Y?

A high F statistic of 302.6 which suggest a significant model fit, meaning that it tells us the variance in the water runoff using our predictor variable, snowfall, meaning, they are related to each other.

This along with an extremely small p value of 2.2e-16 which tells us the probability of observing an F-Statistic as or more extreme as the 302.6 when we assume the null hypothesis is true.

Therefore, we can conclude that Snowfall and Water Runoff have a strong linear association one with another.

29. Briefly summarize what you learned, personally, from this analysis about the statistics, model fitting process, etc.

It is now starting to come together and see how the concepts of this class fit together, rather than seeing them isolated."The big picture" is starting to become more clear. Especially the importance of data inspection and transformation (if needed) before jumping to conclusions. This HW was especially helpful in learning how to compute confidence and prediction intervals in the context of regression and what the difference between the two means.It has been interesting to sort of take a step back from graphs and look more at hard values like MSE and RMSE, especially in real world contexts which has provided a lot of learning and growth in the class content.

These homeworks take a while but I do feel like I'm learning!

30. Briefly summarize what you learned from this analysis *to a non-statistician*. Write a few sentences about (1) the purpose of this data set and analysis and (2) what you learned about this data set from your analysis. Write your response as if you were addressing the mayor of city that relies on the stream runoff from the Sierras (avoid using statistics jargon) and just provide the main take-aways.

This data set provides us with information about the stream runoff in acre-feet of a river near Bishop, California and snow fall in inches at a location in the Sierra Nevada Mountains. The purpose of our work with the data is to see if we can somehow predict the runoff of water with the help of snowfall. If we know the amount of snow we're getting in inches, can we predict the amount of water runoff we will receive? If so, this will help city and state planning much more efficient with how resources are used.

After conducting our analysis, we thoroughly investigated the data to make sure it was useful and verified that we could use it to give us accurate results to our research questions. After doing this we discovered that as snowfall increases by 1 inch at the location in the Sierra Nevada Mountains, our water runoff at the Bishop, California Stream will increase by 3752.5 acre-feet. We ran a number of strict tests to verify the validity of this estimation prediction and are confident in our findings.