Colin White
February 10, 2021

HOMEWORK ANALYSIS #3 - WATER AVAILABILITY

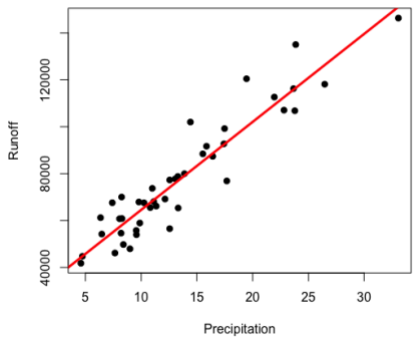
1. In your own words, summarize the overarching problem and any specific questions that need to be answered using the water data. Discuss how statistical modeling will be able to answer the posed questions.

City water planners need to accurately predict the amount of runoff to plan for the city's water needs. We are trying to determine if snowfall (precipitation) can be an accurate predictor of runoff. To be able to do this we need to know if the data will pass the assumptions for a SLR model (LINE assumptions). If the data passes the LINE assumptions, we can fit a SLR model to the data and be able to predict runoff with precipitation.

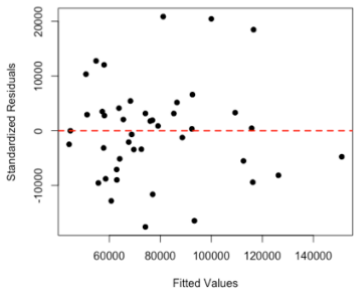
2. Using exploratory techniques (don't actually fit a model), explore the data to assess if a simple linear regression (SLR) model is suitable to analyze the water data. Justify your answer using any necessary graphics and relevant summary statistics that would suggest a SLR model would be successful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

For any data to be suitable to analyze with a SLR model, it needs to pass 4 assumptions. These assumptions are that there is a linear relationship between covariate (precipitation) and response (runoff), that errors are independent of one another and of the covariates (precipitation), that errors are normally distributed, and that there is equal variance for all ϵi.

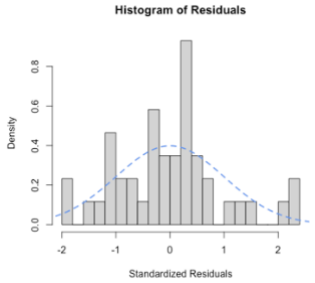
A cursory look at the scatter plot below seems to indicate that there is linear relationship between covariate of precipitation and runoff.



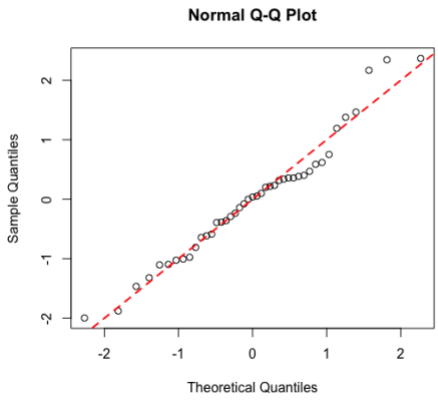
The residuals vs. fitted values scatterplot bellow suggests that the data passes the Linear, Independent and Equal Variance Assumptions.



The histogram of standardized residuals to assess suggests that the data passes the normality assumption. Although the histogram of standardized residuals does not perfectly align with a normal curve, I propose this is because of the relatively small number of observations, and if more observations were made that histogram would normalize.



The normal quantile-quantile (QQ) plot to also suggests that the data meets the normality assumption.



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 (_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.

$Y = ß_0 + x * ß_1 + \epsilon_i$ where $\epsilon_i \sim {}^{iid}N(0, \sigma^2)$

Y = Runoff in acre-feet of a river near Bishop, California
x = Precipitation (snowfall) in inches
$ß_0$ = The predicted runoff if the precipitation is zero
$ß_1$ = Amount the predicted runoff increase per unit increase x (Precipitation (snowfall) in inches)
ϵ_i = Residuals, distance to mean about line

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.

$ß_0$ = 27,014.6
$ß_1$ = 3,752.5

$Y = 27{,}014.6 + x * 3{,}752.5 + \epsilon_i$

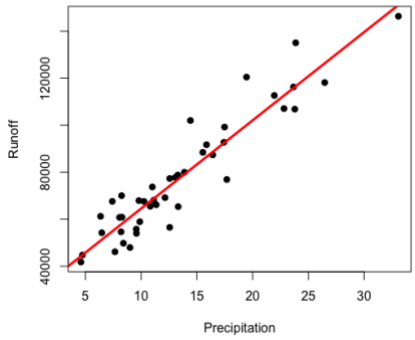
This means that in a year with no precipitation the expected runoff will be 27014.6 acre-feet of water. And for every unit increase in precipitation, there will be 3,752.5 more acre-feet of runoff.

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

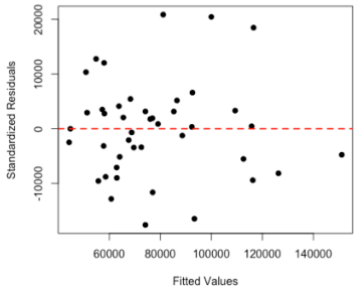
Assumptions: I am assuming that, there is a linear relationship between covariate (precipitation) and response (runoff), that errors are independent of one another and of the covariates (precipitation), that errors are normally distributed, and that there is equal variance for all ϵi.

Justification:

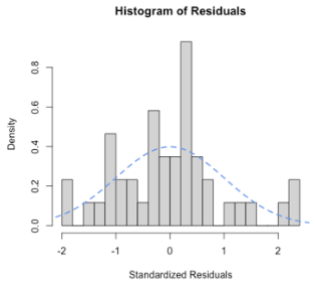
A cursory look at the scatter plot below seems to indicate that there is linear relationship between covariates of precipitation and runoff.



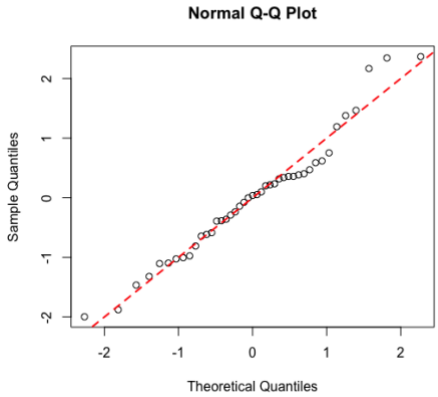
The residuals vs. fitted values scatterplot bellow suggests that the data passes the Linear, Independent and Equal Variance Assumptions.



The histogram of standardized residuals to assess suggests that the data passes the normality assumption. Although the histogram of standardized residuals does not perfectly align with a normal curve, I propose this is because of the relatively small number of observations, and if more observations were made that histogram would normalize.



The normal quantile-quantile (QQ) plot to also suggests that the data meets the normality assumption.



6. Assess the fit and predictive capability of your model. Discuss on the level of your target audience (e.g. interpret your model R2). Draw a conclusion about how "good" your predictions are relative to the range of the response variable.

For the SLM above the R-squared value is 0.8807. This means the 88.07% of the variation of runoff is explained by precipitation.

Cross validation;
mean(RPMSE) = 8,947.72
range(climate$Runoff) 41,785 146345

mean(bias) = 297.5565
sd(climate$Runoff) = 25,518.91

The RPMSE spans hardly any of the range and it is much smaller than the standard deviation indicating that the predictive accuracy of the model is "good".

The bias is also much smaller than the standard deviation confirming that the predictive accuracy of the model is good.

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).

We conducted a p-test to test whether there is a relationship between runoff and precipitation. The p-value obtained is zero. Because the p-value is low (zero), we reject the null hypothesis and conclude that there is a relationship between runoff and precipitation

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 the true value of $ß_0$ is between 20,513.978 – 33,515.197 and also 95% confident that $ß_1$ is between 3,316.809 - 4,188.162.

This means we are 95% confident that the true amount of runoff with no precipitation is between 20,513.978 – 33,515.197 acre-feet. We are also 95% confident that for every unit increase in precipitation there is between 3,316.809 - 4,188.162 acre-feet increase in runoff.

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.

The SLR model predicts that the runoff with 4.5 inches of precipitation will be 43,900.77 acre-feet. We are 95% confident that the actual runoff will be between 25,254.2 and 62,547.34 acre-feet of runoff. I hesitate to make this prediction because there are no observations of precipitation this low. The model is not built to predict runoff in this case with such low precipitation. This may cause the prediction to be incorrect.