Rain Levels in California
Project III

Jeremy Morris

November 22, 2005

The aim of this project is to find a good model to predict average annual precipitation in California. We have been given data from thirty meteorological stations in the state and include altitude, latitude, distance from the coast and the average annual precipitation. We will begin by considering a linear model that has the form

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon \quad (1)$$

Where the x_i are the predictor variables of altitude, latitude and distance from the coast, y represents average precipitation. It is assumed that the error term ε is normally distributed with mean zero and variance σ^2.

1 Initial Analysis

The first thing we look at is the R^2 coefficient along with the F statistic and the p-value for the F test. These things can be seen in Table 1. Although, the R^2 is not terribly high, these numbers appear to indicate an adequate fit for the model as a whole. We also look

R^2	F	$Pr(F > \lvert F^{\alpha}_{3,26} \rvert)$
0.60	13.02	2.204e-05

Table 1: R^2 and F test

at the figures in Table 2, where we see the t statistic and the corresponding p-values. Each is well within the $\alpha = 0.05$ significance level, indicating that each term that we have in the model is significant. The numbers in Table 2 show a weak relationship (meaning the value of the coefficient is quite small) between altitude and distance from coast and the average rainfall when considering the model as whole. Lattitude shows a slightly stronger effect on the average rainfall. This seems to make sense, that as we get closer to the wet regions in the northwest, we should expect more precipitation. Likewise, we expect less rain as we get farther from the coast, meaning a negative relationship, and more rain as we get into higher elevations. The coefficients show that we do, in fact, have these relationships.

If we look at plots of the predictors individually against average rainfall, we can see whether or not the relationships appear to be linear. The plots in Figure 1 show that the

	Estimate	t value	$\Pr(>\|t\|)$
β_0 (Intercept)	-1.03e+02	-3.54e+00	1.55e-03
β_1 (Altitude)	4.06e-03	3.36e+00	2.43e-03
β_2 (Latitude)	3.47e+00	4.38e+00	1.75e-04
β_3 (Dist from Coast)	-1.42e-01	-3.93e+00	5.67e-04

Table 2: T-Tests

amount of precipitation is not really a linear function of distance from coast or altitude. This could be the reason for the weak relationship. The plot for altitude vs average rainfall, however, shows a general trend to having more rainfall in the higher altitudes, although this trend is not very strong.

Figure 2 shows various plots for the linear model considered in (1). In the Residuals vs Fitted plot and the Scale-Location plot, we notice a distinct pattern where there shold be no pattern, this indicates the model does not fit very well. We also see that the error terms may not be normally distributed due to some outliers. The question, then, becomes whether or not the indicated data points are really outliers, and what would happen to the model if the outliers were removed.

2 Removing Outliers

Here, we remove the three points with the highest Cook's distance from the initial analysis and see how the model is affected. We again look at the same statistics and plots as in the first section. The values of each of the coefficients along with a t-test testing whether each term is significant ($\beta_i = 0 \forall i$) can be seen in Table 3. This table shows that the value of the coefficients were not affected significantly by removing the outliers. There is a big difference in the p-values, meaning we gained more significance in each term.

	Estimate	t value	$\Pr(>\|t\|)$
(Intercept)	-9.58e+01	-4.92e+00	5.74e-05
Altitude	4.14e-03	5.13e+00	3.35e-05
Latitude	3.19e+00	5.94e+00	4.63e-06
Dist from Coast	-1.14e-01	-5.23e+00	2.64e-05

Table 3: Coefficients and t-test with outliers removed

Table 4 shows the R^2 and the result of an F test for all coefficients being zero. This shows significant improvement over the values in Table 1. Which would indicate that the model as a whole fits better without the outliers.

Figure 3 shows the diagnostic plots for the new model without outliers. These plots show an significant improvement in the Residuals vs Fitted and Scale-Location plots. The Normal Q-Q plot also demonstrates that the residuals are closer to normality than with the original data.

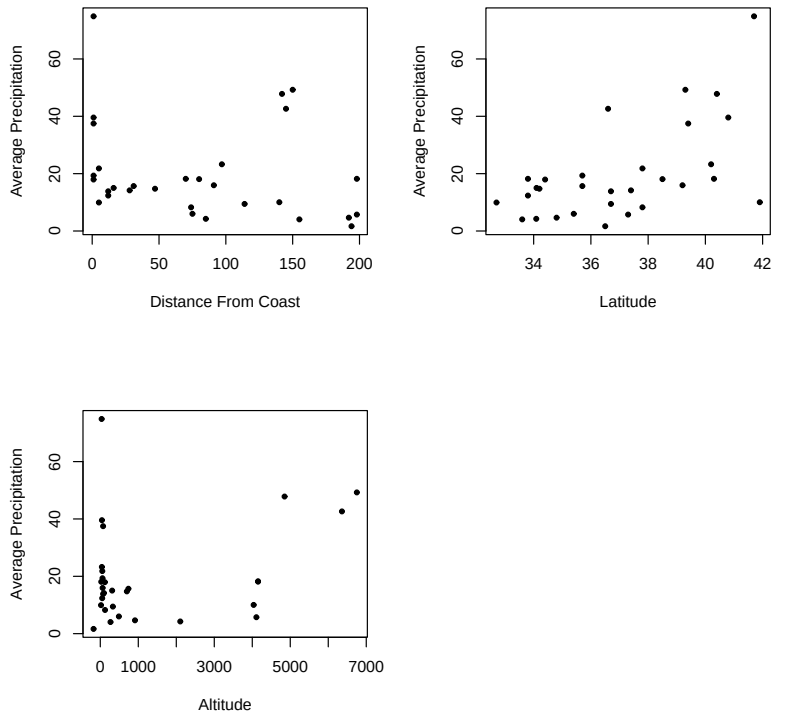

Figure 1: Plots of each variable against Avg Rainfall

R^2	F	$Pr(F > \lvert F^{\alpha}_{3,26} \rvert)$
0.77	25.37	1.775e-07

Table 4: R^2 and F test with outliers removed

This analysis leads to the conclusion that the model can be improved by removing the outliers detected during the initial analysis. Of course, the Cook's distance plot in Figure 3 indicates more data points as outliers, which means we should repeat this analysis removing those points. This may end up leading us to eliminate almost all of the data points. We will next do a more close analysis of the outliers to determine if the outliers should have been removed.

3 Outlier Analysis

An outlier is generally thought of as a data point that severly effects the accuracy of a model. There are many ways of defining and dealing with outliers. Our main concern is whether or not any of the rainfall data qualify as outliers. Meaning that certain data points may appear to be outliers, but actually define the model as opposed to negatively

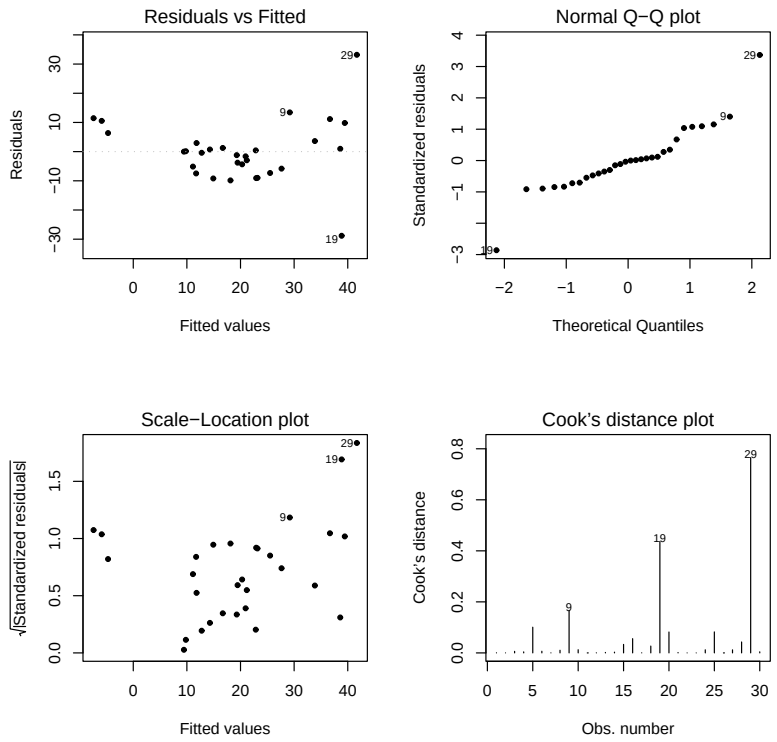

Figure 2: Linear Model Plots

affecting it. Or if those points are valid and the model needs to be altered to ensure greater accuracy.

The Cook's distance plot in Figure 2 shows that there could be at least three significant outliers in the initial analysis. One way of determining which data points are outliers is to look at a boxplot. A boxplot shows the median as a solid line within a box representing the distance between the first and third quartiles. The outer fences are calculated as 1.5 times the distance between the first and third quartiles.

Figure 4 shows boxplots for each of the variables in our model, including the dependent variable of average precipitation. The plots indicate four possible outliers based on average precipitation and altitude. Outliers are not indicated for latitude or distance from coast, most likely because these two variables were sampled evenly throughout the state.

Data for the four outliers can be found in Table 5. With the exception of Crescent City, all the possible outliers come from the higher elevations that are further from the coast. Crescent City can be explained simply because it lies on the coast and very close to Oregon which is known for its high levels of precipitation. It is quite interesting that the data for Tulelake area does not show an equally high precipitation rate, this could be another reason for the high Cook's distance for Crescent City. In order to see what is going on with the other three data points, it is instructive to look at a map.

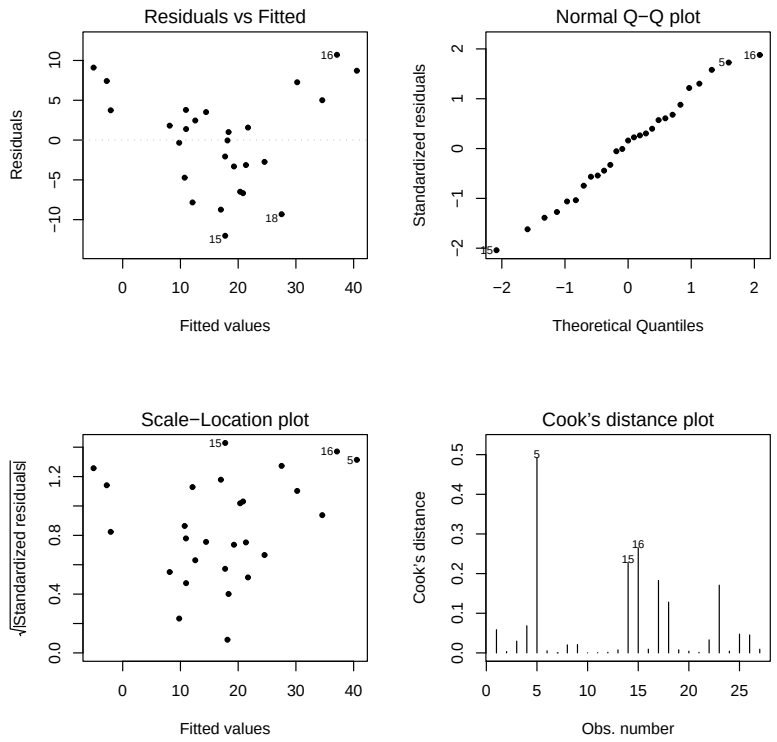

Figure 3: Diagnostic plots with no outliers.

Figure 5 shows a map of California with the data points superimposed. It is striking how well looking at this map describes exactly what is going on with the data. The ranges for the different colors were calculated by dividing up the data into four evenly spaced groups around the average precipitation. Then, given that most of the data fell into the lowest category, that category was split into two groups. This may have been a rather arbitrary way to split the data up, but it still gives a general idea of what is going on with the weather in California.

The first thing we notice is that it appears that the amount of rain does actually increase as we go further north in the state. The second is that all three of the major outliers are located in the mountains. This is not surprising since the elevation is one of the main factors in indicating their being outliers. What is interesting is if we take a pick one of the outliers and look at two of the other points along the same latitude (Table 6). What the data shows here is that we get a small amount of rain before the mountains, a large amount in the mountains and an even smaller amount of rain on the other side of the mountains. Meteorologists call this orographic lift. Orographic lift occurs when a system of clouds is forced from a low elevation to a high one due to a change in terrain. What typically happens is that as the air raises, it is cooled and the relative humidity goes up causing precipitation. This happens in many areas of the world, one of which is here in the

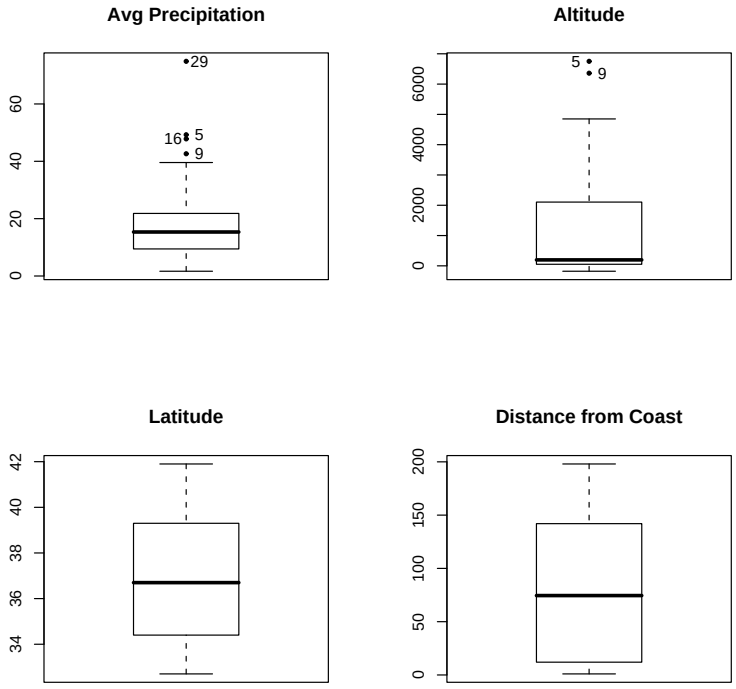

Figure 4: Boxplots

Wasatch front.

What this means for our model is immediately clear, the data points that were indicated for being outliers are not. They are valid points describing a specific weather pattern that exists in California. Unfortunately, there is nothing in the data to indicate that orographic lift is happening. Take the data in Table 6 for example. The model indicates that at a certain latitude, there should be more precipitation in the higher altitudes (especially since the elevation in Susanville is 100 times higher than that of Red Bluff), but there is actually less precipitation in Susanville than in Red Bluff. This happens for each of the indicated outliers in the highest elevations. These points counteract the effects of distance from coast and elevation since they cause higher elevations to recieve less rainfall than the model would predict because the stations lie behind the Sierra Nevada mountains. This can be seen with the data in Table 7 where we show the observed rainfall and the amount predicted by the model.

All of this is complicated by the fact that there is a huge desert region in the southern end of the state. In this southern region, most of the inputs do not really effect the model very much. None of the stations are going to recieve much rain regardless of their location in relation to the coast. Most of the stations, with only a few exceptions, in this region are also at low elevations meaning that the elevation in this region will not make a huge

	Station	Avg Precipitation	Altitude	Latitude	Dist from Coast
5	Soda Springs	49.3	6752	39.3	150
9	Giant Forest	42.6	6360	36.6	145
16	Mineral	47.8	4850	40.4	142
29	Crescent City	74.9	35	41.7	1

Table 5: Data for outliers

	Station	Avg Precipitation	Altitude	Latitude	Dist from Coast
2	Red Bluff	23.3	41	40.2	97
16	Mineral	47.8	4850	40.4	142
18	Susanville	18.2	4152	40.3	198

Table 6: Data along 40^{th} latitude

difference in the model.

4 Conclusions

The main conclusions reached here are that a simple linear model as in (1) does not fit the data well enough to be of any practical use. Nor will removing any of the apparent outliers improve the model any either, since those outliers represent conditions that affect the weather. This leads to the conclusion that there is at least one hidden variable at play in the model. One hidden variable could be the average humidity level at each station. Humidity would be a good indicator of average rainfall.

	Station	Avg Precipitation	Predicted Avg Precipitation
18	Susanville	18.20	25.5
15	Bishop	5.73	14.9

Table 7: Actual values vs Predicted values

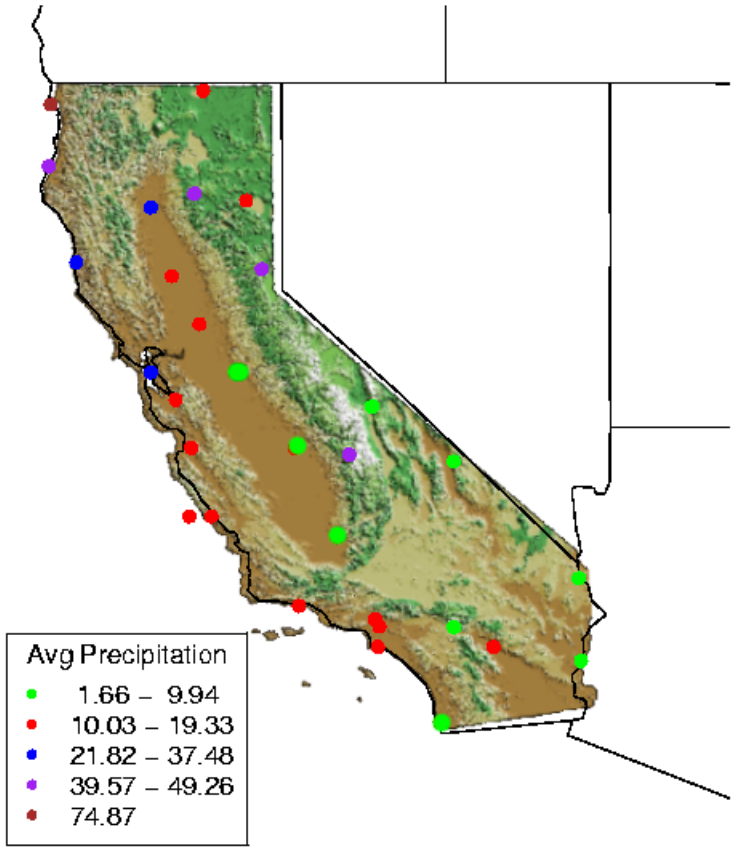

Figure 5: California Precipitation Levels