ANSWERS TO EXAMPLE ASSIGNMENT
(For illustration only)

In the following answer to the Example Assignment, it should be emphasized that often there may be no "right" answer to a given question. Data are usually subject to more than one interpretation. The important point to keep in mind is that you should always try to make a convincing *argument* for your conclusions. Note also that the following answer is not structured as a "line by line" response to the question. Rather, it is organized as a *short report* that implicitly addresses all the questions, while at the same time maintaining a logical flow of the discussion.

1. Introduction. The objective of this study is to explain the spatial pattern of rainfall in California. A map of average rainfall levels from 1961-1990 is shown in Figure 1a below.[1] This study will involve a regression analysis based on a data set collected by Taylor (1980) for 30 cities in California (shown in Figure 1b below). The data consists of rainfall levels (*Percip*) from weather stations in these cities, together with a number of geographical attributes for each city. As Taylor points out, it is clear from special geographic features of California that rainfall patterns should be influenced by altitude, latitude, and distance from the coast. Hence the attribute variables of primary interest here are the altitude (*Alt*) and latitude (*Lat*) of each city, together with its distance from the coastline (*Dist*).

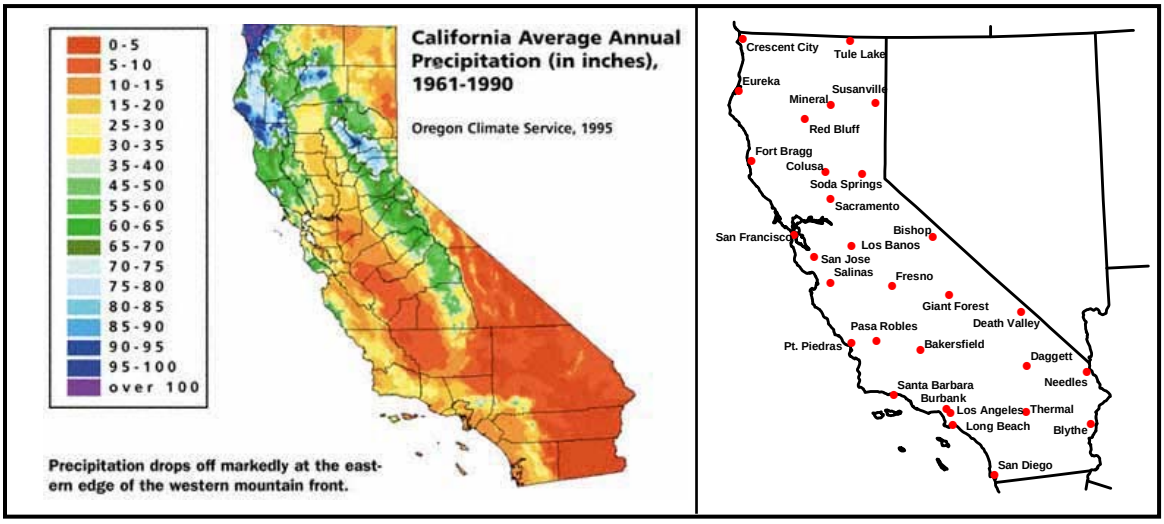

Figure 1a. Average Rainfall in California **Figure 1b.** California Cities

The rainfall data for this set of 30 cities is shown in Figure 2 below, together with a graphical depiction of their relative elevations. As seen from the legend, three of the wettest cities are located along the central ridge of the (Sierra) mountains, and four of the driest are located in the (Mojave) desert region in the southeast. This suggests that *higher elevation* may serve as a good predictor of *higher rainfall levels*. Next observe that five

[1] This map image is taken from *http://www.forester.net/ec_0003_holding.html* .

of the wettest cities are in the north, and five of the driest are in the south, suggesting that *higher latitude* may also be a good predictor of *higher rainfall levels*. Finally, note that three of the wettest cities are on the coast, and that five of the driest are far from the coast, suggesting that *greater distance* from the coast may be a good predictor of *lower rainfall levels*. Hence one can reasonably hypothesize that a regression of Precipitation on theses three variables should yield an *positive* relation with both Altitude and Latitude, and a *negative* relation with Distance from the coast.

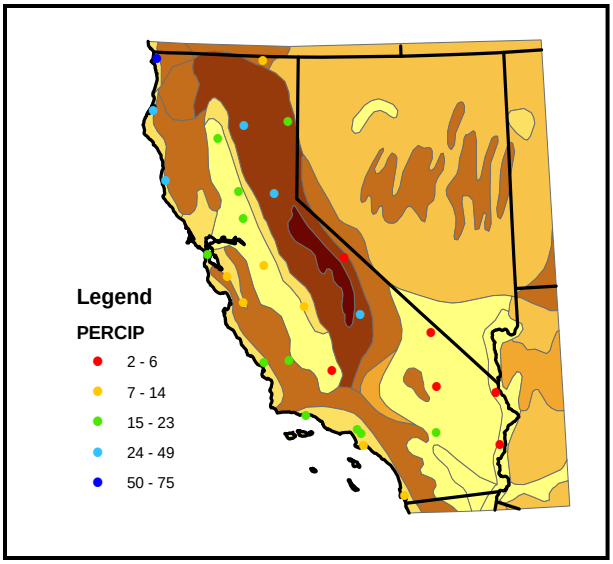

Figure 2. Rainfall Data

2. First Regression Model. These observations can be formalized in terms of the following *linear regression model*,

$$(1) \qquad Percip_i = \beta_0 + \beta_1 Alt_i + \beta_2 Lat_i + \beta_3 Dist_i + \varepsilon_i \ , \ i = 1,..,30$$

where by assumption the unobserved residuals, ε_i, are independent normal random variables with zero means and common variances, i.e., $\varepsilon_i \sim N(0,\sigma^2)$, $i = 1,..,30$. In this modeling context, one would thus expect to observe significant *positive* estimates of β_1 and β_2, and a significant *negative* estimate of β_3. The results of such a regression (in JMPIN) are shown in Table 1 below [where the coefficient estimates, $\left(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3\right)$, correspond respectively to the rows **Intercept**, **Alt**, **Lat**, and **Dist** below]:

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-102.3666	29.19894	-3.51	0.0017
Alt	0.0040884	0.001218	3.36	0.0024
Lat	3.4517572	0.79469	4.34	0.0002
Dist	-0.142945	0.036338	-3.93	0.0006

RSquare	0.600398
RSquare Adj	0.55429
Root Mean Square Error	11.09735
Mean of Response	19.81233
Observations (or Sum Wgts)	30

Table 1. Regression Results for Model (1)

Observe first that each of the slope coefficients has the anticipated sign. In addition, every *P-value* (**Prob>|t|**) is less than .003, indicating strong statistical significance of each estimate. However, the adjusted R-square value (**RSquare Adj)** shows that almost half the variation in the data remains to be explained. As one possible explanation here, a plot of

the estimated regression residuals against the predicted rainfall values indicates that there are two *outliers* in this data set, as shown in Figure 3 below.

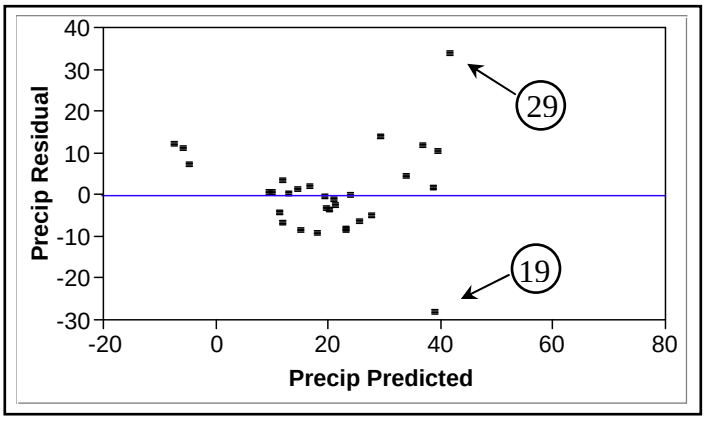

Figure 3. Regression Outliers for Model (1)

These two points correspond to the two cities on the northern border of California (29 = **Crescent City**, 19 = **Tulelake**). The extreme nature of these outliers suggests that very special types of local factors may be at work. For purposes of the present study, this will be assumed to be the case. [INSTRUCTOR NOTE: *Further research shows that the unusually heavy rainfall in the coastal region around Crescent City is due to a special interaction between the east-west Cascadia Channel directly off shore and the north-south Japanese Current. In addition, the extreme dryness of the inland region around Tulelake is known to be due to the "Rain Shadow" effect created by Mt. Shasta (as discussed below).*] Hence it is appropriate to remove these outliers in order to obtain a better estimate of the overall relations among the variables of interest. The removal of these two outliers yields a new set of regression results shown in Figure 4 below:

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-91.2635	19.90289	-4.59	0.0001
Alt	0.0047439	0.00074	6.41	<.0001
Lat	3.0658172	0.548538	5.59	<.0001
Dist	-0.117438	0.02232	-5.26	<.0001

RSquare	0.777853
RSquare Adj	0.750085
Root Mean Square Error	6.673897
Mean of Response	18.19643
Observations (or Sum Wgts)	28

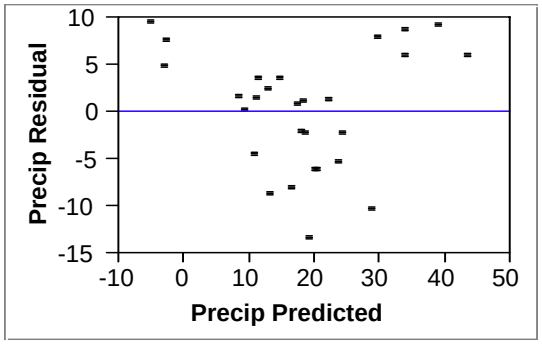

Table 2. Regression Results with Outliers Removed

Hence it is now clear that all coefficients are more significant, and that this reduced model captures 75% of the data variation. In addition the remaining residual variation looks fairly random, and suggests that the model fits reasonably well. [INSTRUCTOR NOTE: *It is important to emphasize here the removal of outliers will always make the model fit better. So one must be very careful to justify removal of outliers*!]

While the above residuals look quite random, a *spatial plot* of these residuals shows a rather distinct pattern, as can be illustrated by comparing the case of Salinas with that of Pt. Piedras (shown in Figure 4b below). Both cities are close to the coast (12 miles for Salinas and 1 mile for St. Piedras). In addition, both are quite close to sea level (74 feet above sea level for Salinas and 52 feet for St. Piedras). Hence observing that St. Piedras is considerably south of Salinas, it is not surprising that our model predicts more rainfall for Salinas than for St. Piedras (20 inches versus 18 inches). But in fact, Salinas has a significantly *lower* level of rainfall than St. Piedras (14 inches versus 19 inches). This corresponds to the blue dot for Salinas on the right hand figure below, indicating a *negative* residual (over prediction) for rainfall in Salinas.

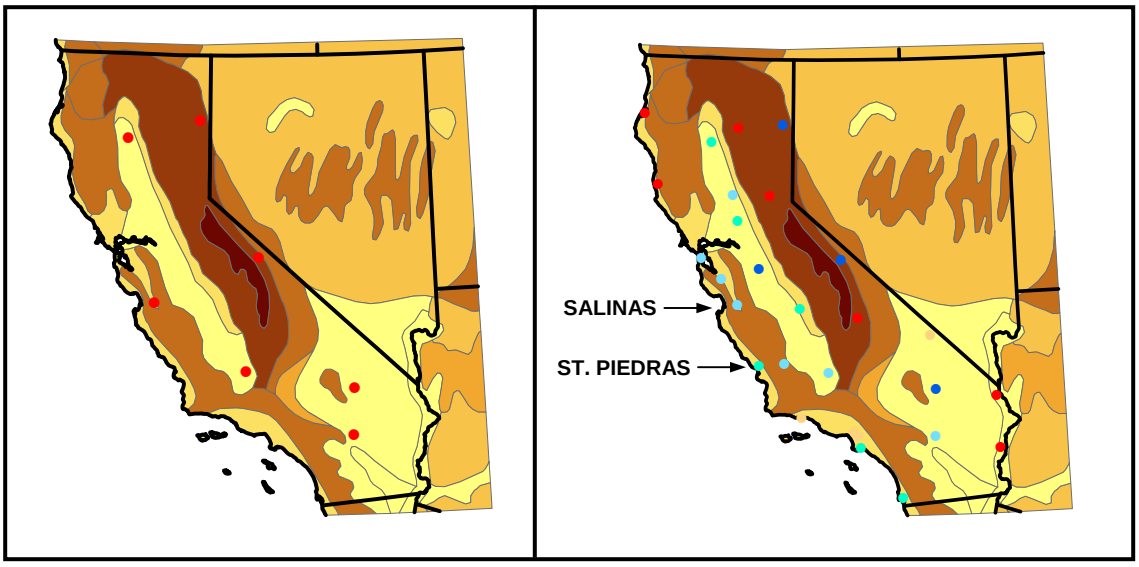

Figure 4a. Rain Shadow Cities **Figure 4b.** Spatial Residuals

One possible explanation for this difference is the relation of these two cities to the coastal mountains (Coastal Range). While St. Piedras is directly exposed to the Pacific ocean, Salinas is separated from the coast by these mountains, and hence is protected from some of the coastal rain that falls on St. Piedras. This type of "Rain Shadow" effect is in fact exhibited much more generally throughout California. Some extreme examples are the cities (Susanville, Bishop, and Daggett) corresponding to the three dark blue dots closest to the eastern border of California in the right-hand figure above. These three cities are all on the eastern slopes of the Sierra range in California, and all are predicted to have a much wetter climate than they do. Hence it seems clear that the "Rain Shadow" effect is at work here. Figure 4a shows six cities that all exhibit a potential for this type of effect.

3. Second Regression Model. To determine whether this effect does add to the above explanation of California rainfall patterns, a dummy variable "Shadow" has been constructed that assigns the value '1' to these six cities and '0' elsewhere. This yields a new regression model of the form:

(2) $$Percip = \beta_0 + \beta_1 Alt + \beta_2 Lat + \beta_3 Dist + \beta_4 Shadow + \varepsilon$$

The results of this regression are shown in Table 3 below:

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-87.75772	15.00977	-5.85	<.0001
Alt	0.0051327	0.000564	9.10	<.0001
Lat	3.007349	0.41331	7.28	<.0001
Dist	-0.109574	0.016904	-6.48	<.0001
Shadow	-10.00372	2.27605	-4.40	0.0002

RSquare	0.879262
RSquare Adj	0.858264
Root Mean Square Error	5.026009
Mean of Response	18.19643
Observations (or Sum Wgts)	28

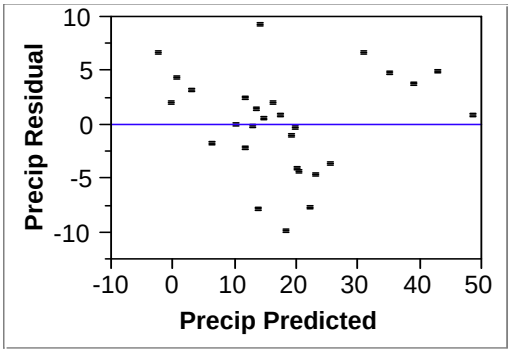

Table 3. Regression Results for Model (2)

Here it is clear that this refinement of the model adds significantly to the explanation of rainfall patterns. The adjusted R-square value now shows that this model accounts for more than 85% of the variation in rainfall levels. Moreover, the three original variables not only remain very significant (with the appropriate signs), but also the Shadow variable is now very significantly *negative* – indicating that (all else being equal) the presence of a rain shadow can be expected to lower average rainfall levels by a full *10 inches*. Finally, the spatial residuals for this new regression in Figure 5 show no clear pattern, adding further confirmation that this model accounts for most of the spatial variation in rainfall levels.

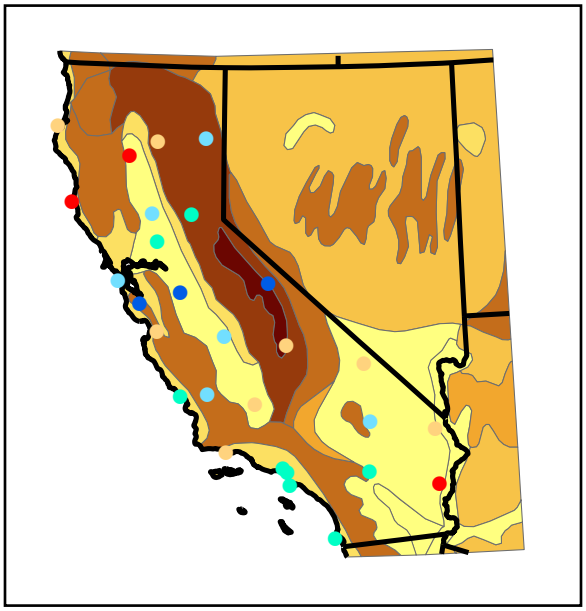

Figure 5. Regression Residuals for Model (2)

4. Concluding Remarks

The present study has produced a remarkably successful regression model of rainfall in California, in which 85% of the variation is captured by four key geographic variables. But the main intent of the study is to show that when analyzing spatial distributed phenomena by means of regression models, it generally useful to consider the *spatial pattern* of regression residuals, as well as the usual diagnostic plots of these residuals. In the present case of California rainfall, it has been shown that these residuals indicate the presence of a significant "rain shadow" effect created by the interaction of mountain ranges and prevailing wind flows. While it retrospect it can be argued that such effects are "obvious" and should have been incorporated in the original model, the present example is meant to be *pedagogical* in nature. Indeed, there may often be significant spatial effects unknown to the researcher that can be discovered by a careful analysis of spatial diagnostics.

5. References

Taylor, J.P. (1980) "A pedagogic application of multiple regression analysis: precipitation in California", *Geography*, 65: 203-212.