Linear Regression Model for Annual Average Precipitation in California, USA

Final Project, Case Study 4

University of Waterloo
STAT 331
Spring 2016
Joslin Goh

Group 24

YAOTIAN HOU 20552069
DEZHAO OU 20574321
ZHIYANG ZHANG 20553484
GUORUI LIU 20551266

Summary

Objective:
The main goal of this project is to investigate the variation in average annual precipitation in California and find a model to fit in the observed data. The study was originally written in the journal Geography (July, 1980) by P. J. Taylor with the method of multiple regression. A sample of 30 meteorological stations data including their altitude, latitude, distances to Pacific Ocean, slope face direction along with the observed average annual precipitation value was used in the analysis.

Overview of the statistical analysis:
In order to find a suitable linear regression model, we implemented forward,stepwise and backward selection methods and narrow down all plausible models given in the project to two candidate models, M3 and M4. Then, we found the ultimate model by comparing these two models via various diagnostics. For example, we used studentized residual plots to view the relation between the residuals and the predicted values, the leverage to detect the outliers, and cross validation to manually finalize the better model between M1 and M2.

Summary of the main result:
By conducting above analysis with the comparison of the three given models, we found that M4 fits the data best among all models.

Model Selection

Two candidate models were found by model selection technique and the comparison of the given models. (Please see the Appendix for the coding in R):

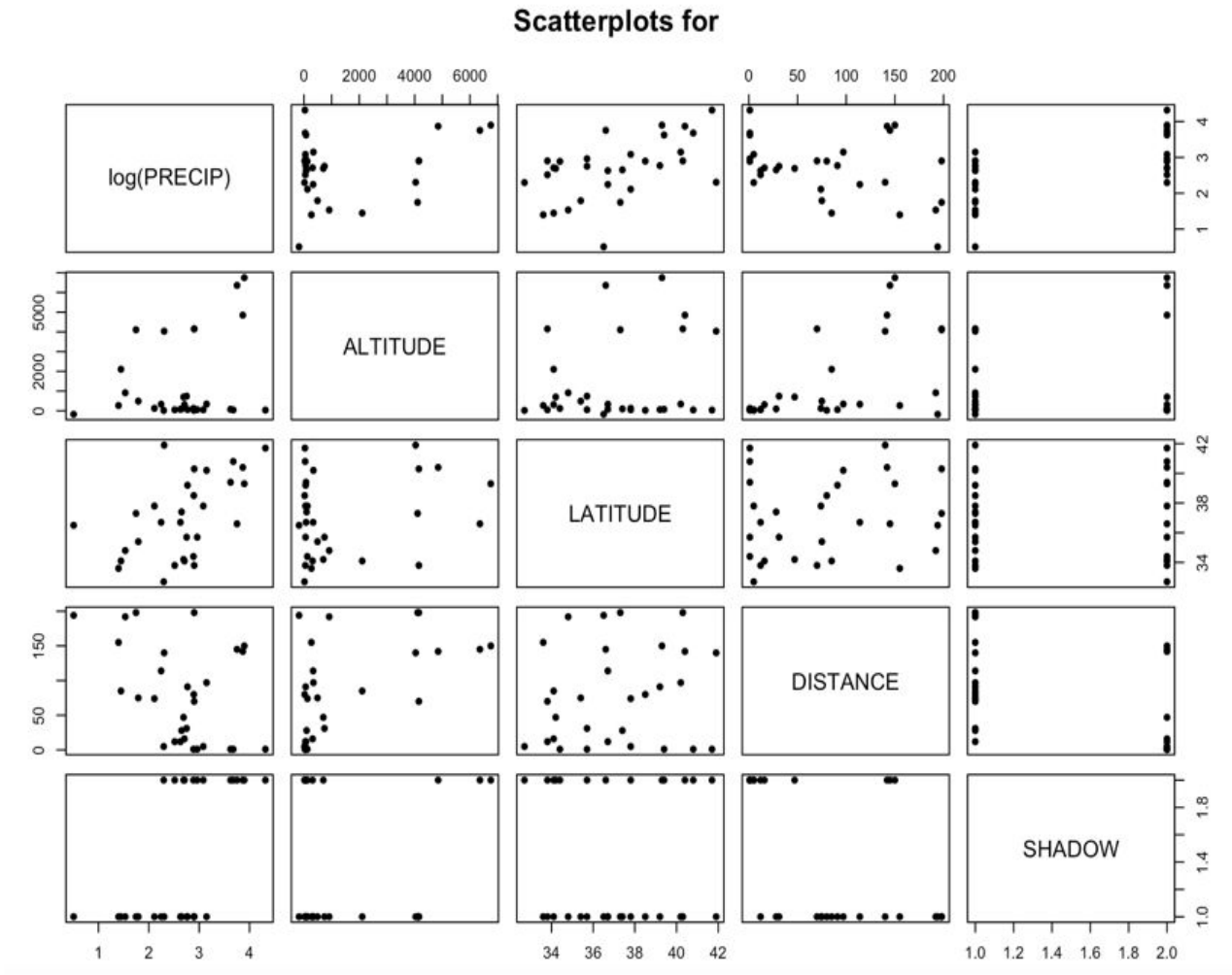
Model 3:
PRECIP~
ALTITUDE+LATITUDE+DISTANCE+SHADOW+ALTITUDE*SHADOW+LATITUDE*SHADOW+DISTANCE*SHADOW

Model 4:
PRECIP~
ALTITUDE+LATITUDE+DISTANCE+SHADOW+ALTITUDE*SHADOW+LATITUDE*SHADOW+ALTITUDE*LATITUDE

Model Diagnostic

Scatterplots

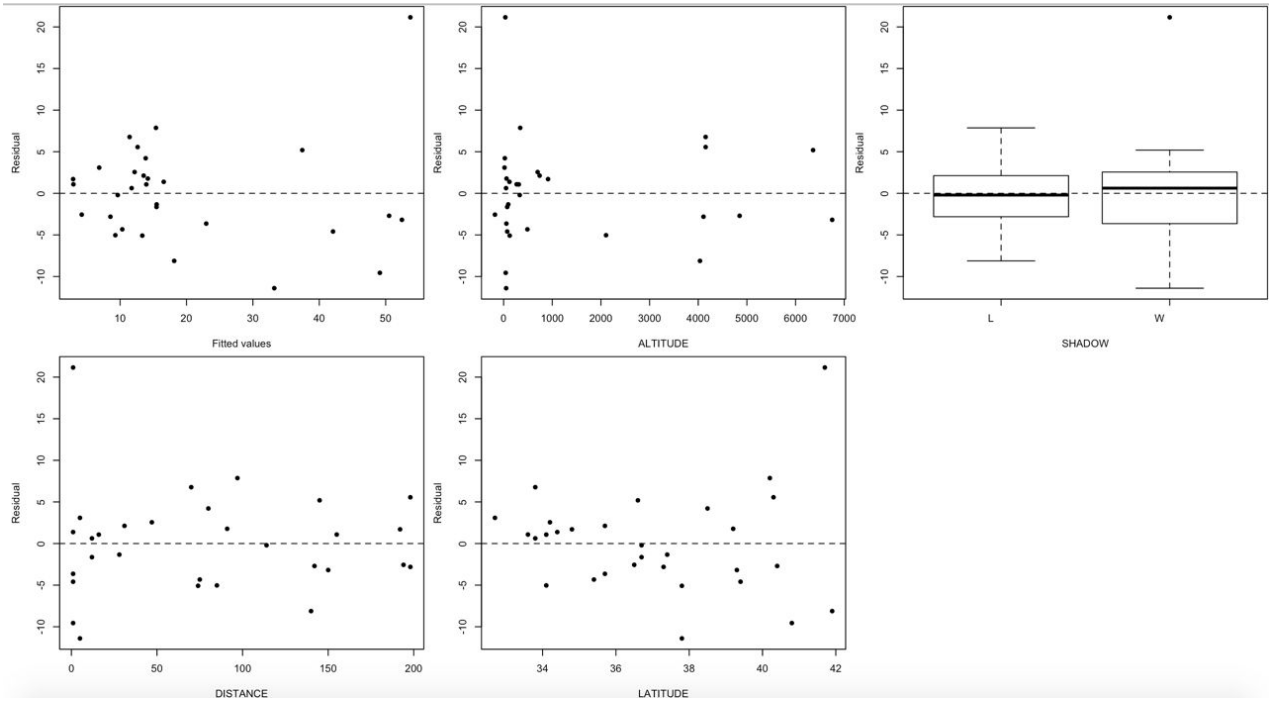
First of all,we would like to have an overview of the relationships between every two variables.



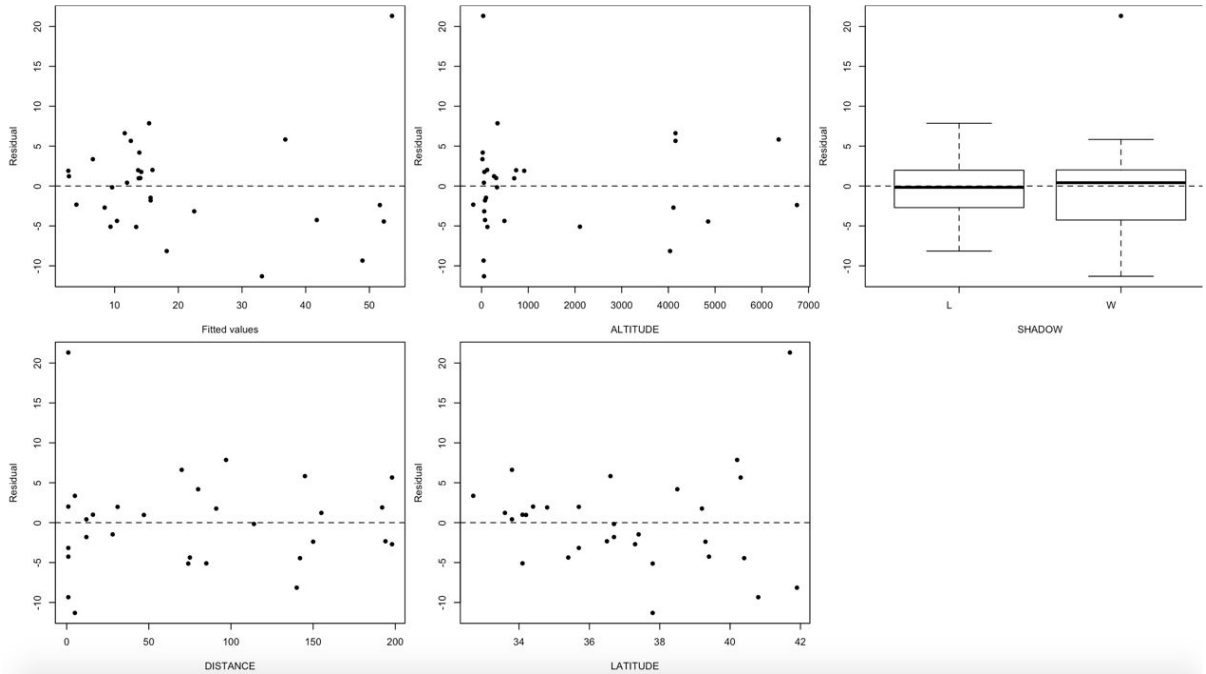
[Figure 1: Scatterplot for data]

In general, all the explanatory variables have a relationship with average annual precipitation and there might be some interaction facts between explanatory variables.

Residual Plots

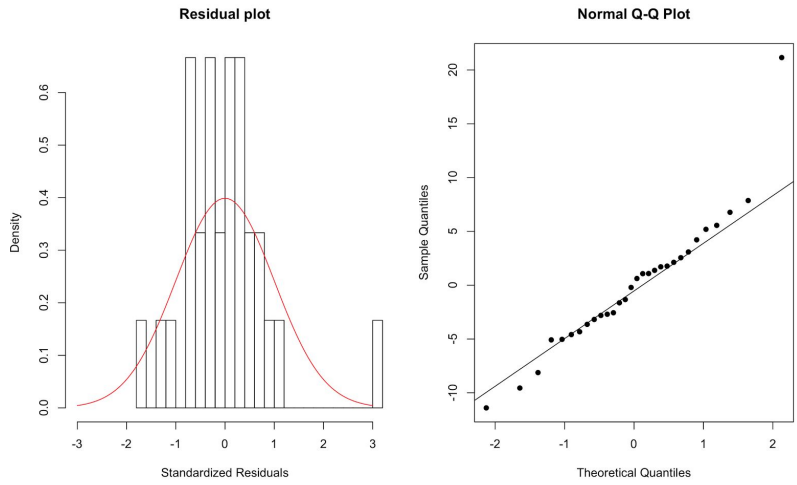


[Figure 2: Residual plot of Model 3]

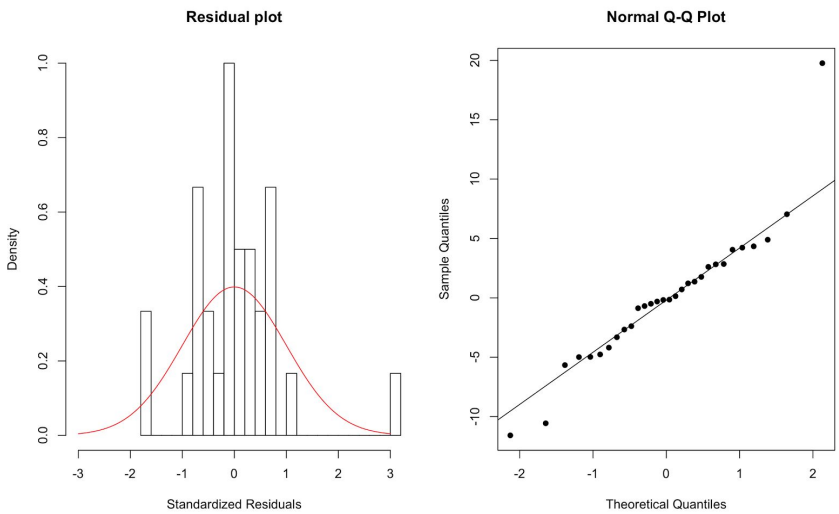


[Figure 3: Residual plot of Model 4]

Residual Plot and Normal Q-Q plot



[Figure 4: Residual plot of Model 3]



[Figure 5: Residual plot of Model 4]

From the data above, we can see that the residual plot of Model 4 looks better than the Model 3 residual plot. However, there are three observations off the track. It may be outliers or influential observations.

ANOVA Analysis

```
> anova(model3)
Analysis of Variance Table

Response: PRECIP
                Df  Sum Sq Mean Sq F value    Pr(>F)
ALTITUDE         1  730.73  730.73 14.7484 0.0008895 ***
SHADOW           1 2728.65 2728.65 55.0729 2.007e-07 ***
LATITUDE         1 2293.92 2293.92 46.2987 7.764e-07 ***
DISTANCE         1  160.53  160.53  3.2399 0.0855935 .
ALTITUDE:SHADOW  1  292.55  292.55  5.9046 0.0237156 *
SHADOW:LATITUDE  1  707.48  707.48 14.2792 0.0010329 **
SHADOW:DISTANCE  1    7.80    7.80  0.1574 0.6953775
Residuals       22 1090.01   49.55
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
```

[Figure 6: ANOVA of Model 3]

	Sum of squares	df	Mean squares	F

Regression	6921.66	7	988.81	19.96
Residuals	1090.01	22	49.55	
Total	8011.67	29		

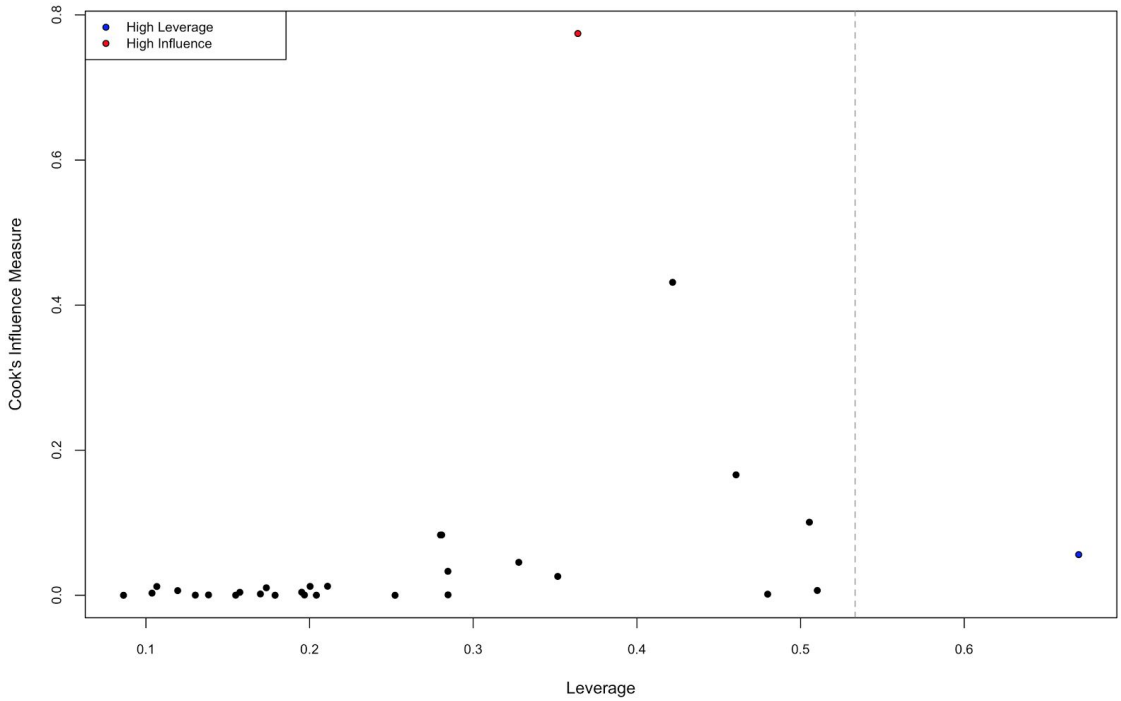
```
> model4<-lm(formula = PRECIP ~ SHADOW + LATITUDE + ALTITUDE + DISTANCE +
+     SHADOW:LATITUDE + LATITUDE:ALTITUDE + SHADOW:ALTITUDE, data = data)
> anova(model4)
Analysis of Variance Table

Response: PRECIP
                  Df  Sum Sq Mean Sq F value    Pr(>F)
SHADOW             1 2865.68 2865.68 67.0381 4.008e-08 ***
LATITUDE           1 2728.96 2728.96 63.8397 6.027e-08 ***
ALTITUDE           1  158.64  158.64  3.7112 0.0670648 .
DISTANCE           1  160.53  160.53  3.7552 0.0655829 .
SHADOW:LATITUDE    1  883.16  883.16 20.6602 0.0001592 ***
LATITUDE:ALTITUDE  1  130.80  130.80  3.0600 0.0941828 .
SHADOW:ALTITUDE    1  143.43  143.43  3.3553 0.0805629 .
Residuals         22  940.44   42.75
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
```

[Figure 7: ANOVA of Model 4]

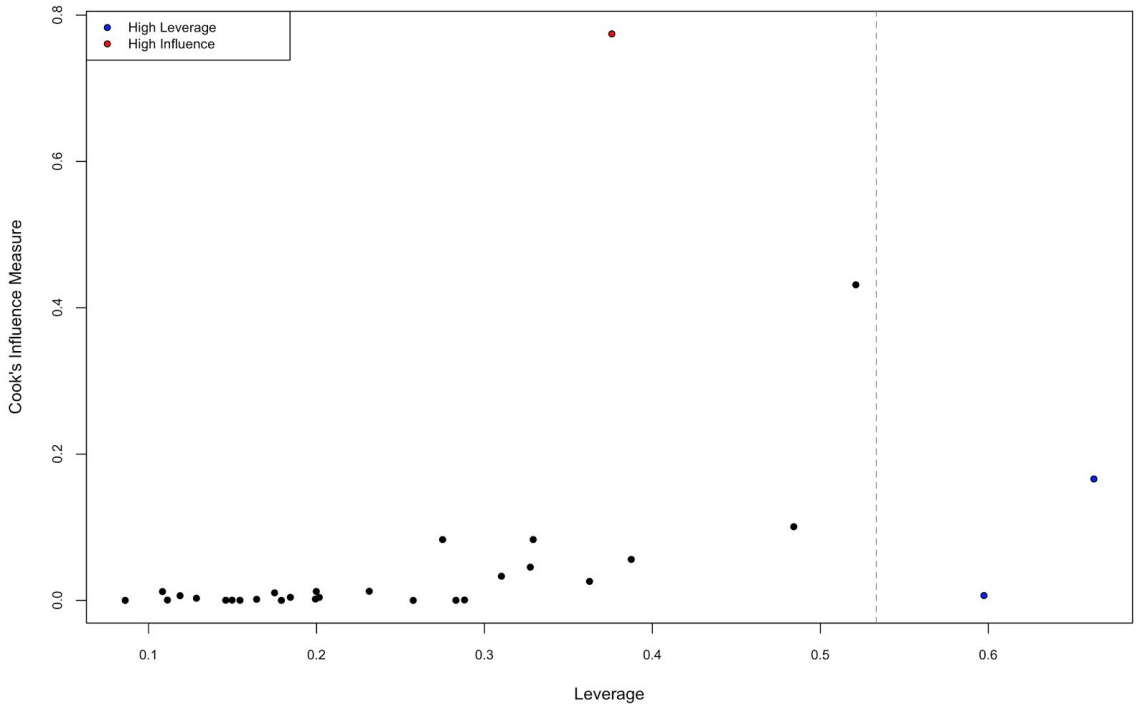
	Sum of squares	df	Mean squares	F
Regression	7071.2	7	1010.17	23.63
Residuals	940.44	22	42.75	
Total	8952.08	29		

Cook’s Distance

In order to check outliers and influential observations, we now have a look at Cook’s Distance plot.

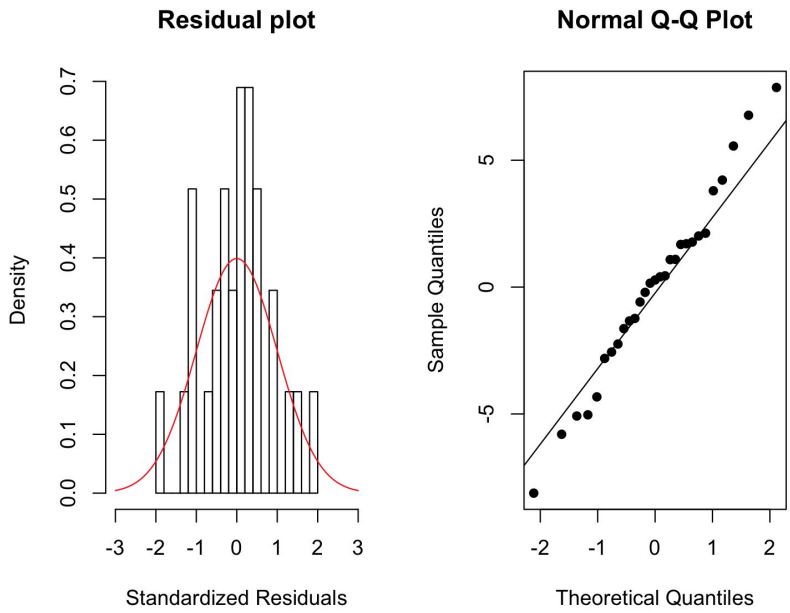


[Figure 8: cook's distance of Model 3]

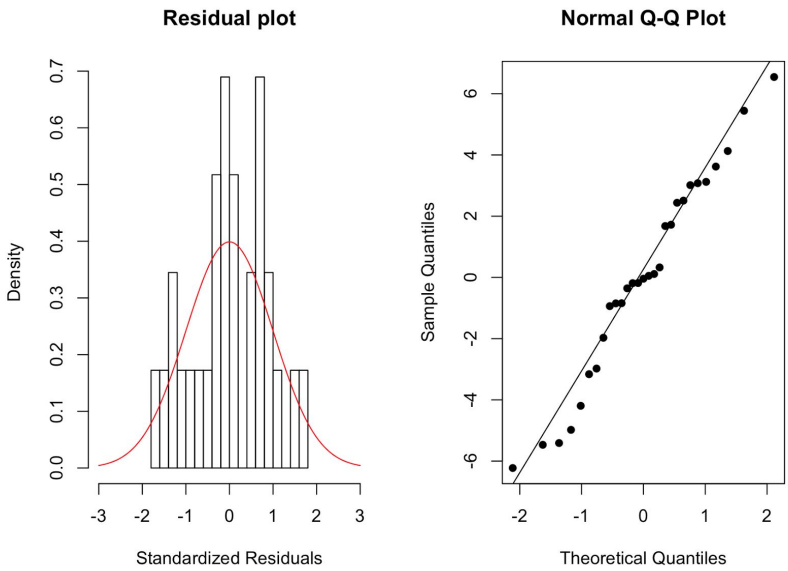


[Figure 9: Cook distance of Model 4]

We found that the 29th observation is an outlier, thus we remove it from our data. After removing the outliers we checked the QQ-plot again.



[Figure 10: residual plot of model 3 after removing outliers]



[Figure 11: residual plot of model 4 after removing outliers]

Cross Validation

To evaluate the models, we did cross validation.

PRESS statistics
Model 3
1081.309
Model 4
1088.601

AIC(Model 3)
210.9187

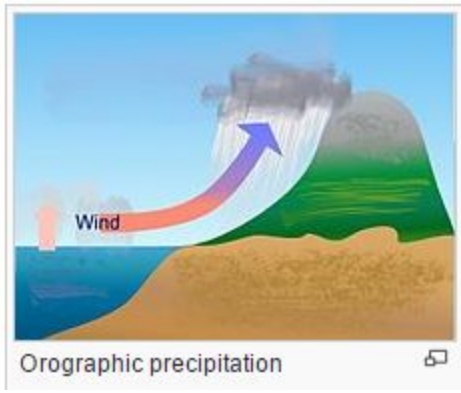
AIC(Model 4)
209.1326

Conclusion:

From the analysis above, the best fitted model we obtained is M4, which is PRECIP~ALTITUDE+LATITUDE+DISTANCE+SHADOW+ALTITUDE*SHADOW+LATITUDE*SHADOW+ALTITUDE*LATITUDE. It concludes all the single factors and three interaction factors.
In this project, we compared the residual plots of the two models to view the relation between the residuals and the predicted values. The QQ-normal plots shows that M4 fits better than M3. Then we used the leverage to detect the outliers and influential observations. After removing outliers, we checked residual plots again. Also we did cross validation to manually finalize the better model between M3 and M4. We suggest M4 as the final model as it considered interaction factors and has better residual plots.

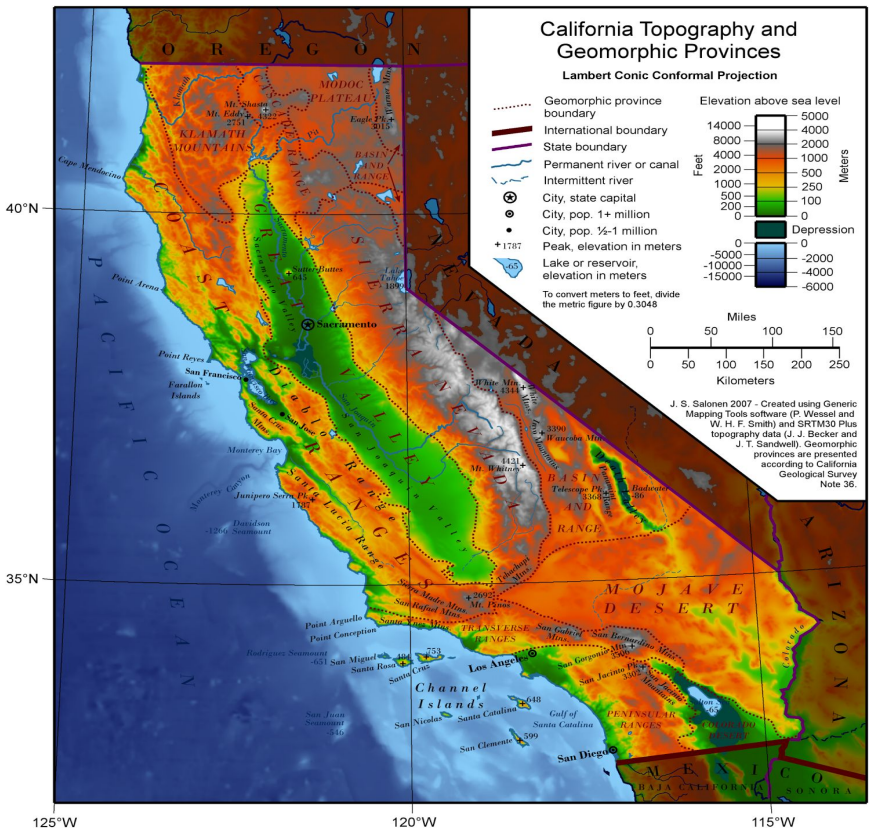
Reflection:

In the model M4, we use all the single vectors as well as the interaction effect between ALTITUDE and SHADOW, LATITUDE and SHADOW, and ALTITUDE and LATITUDE. Besides all the single vectors, it is reasonable to look at the realistic meaning of those interaction effects.



As far as we know, the SHADOW(slope face) has a huge impact on precipitation due to the orographic effects(source: Wikipedia). Therefore, it is reasonable to put in the interaction effect ALTITUDE*SHADOW and LATITUDE*SHADOW since the slope face of a location scale up the precipitation a lot.

For the interaction effect ALTITUDE*LATITUDE, it is related to California's topography(see below), the California current and the coastal storm(low altitude) that happened mostly in north California(high latitude), which leads to a huge difference in precipitation.



California Topographic Map(source: Wikipedia)

In our code and analysis above, we didn't consider interaction effect of more than two factors to avoid overfitting the data. Still, it may be reasonable to put in those effects in the model since the precipitation of a point is always affected by different factor simultaneously by looking at the topographic map that we have.

In our previous analysis, we only consider the linear models with the interaction effect to fit in the dataset that we are provided. However, it is possible that the data will fit better with some nonlinear models, such as models with polynomial functions, or some more complicated functions such as logarithmic and trigonometric functions.

For quadratic or even higher degree functions, we didn't use them to avoid overfitting the dataset that we have. For 30 samples(29 after removing outliers), it is easy to overfit, although we may get a more fitted curves. Also, it would be hard to understand the realistic meaning behind the model if using non-linear polynomial regression or trigonometric regression.

However, logistic regression is in our consideration. In theory, it is possible that the quantitative factors need to be logged to avoid the large impact on a small change in such as DISTANCE, which doesn't really make much difference in precipitation. So in further study, we may try logistic regression to improve the model, but still, linear regression model should be suitable for fitting the dataset that we have in the case.

Appendix

****

1) Data settlement given by the sample code

```
data<-read.csv("caliRain.csv")
# Scatterplots
```

```
pairs(~log(PRECE)+ALTITUDE+LATITUDE+DISTANCE+SHADOW,data=caliRain,
      main="Scatterplots for ")
```

****

2) R CODE OF MODEL SELECTION

```
#Model selection
M0 <- lm(PRECIP~1,data = data) #Model with only an intercept
Mfull_main <- lm(PRECIP~.,data=data) #Model with all main effects
Mfull <- lm(PRECIP~(ALTITUDE+DISTANCE+LATITUDE+SHADOW)^2,data=data)
#Forward
Mfwd <- step(object = M0, scope = list(lower = M0, upper = Mfull),direction =
"forward",trace = FALSE)
#Backward
Mback <- step(object = Mfull, scope = list(lower = M0, upper = Mfull),direction =
"backward",trace = FALSE)
#Stepwise
Mstep <- step(object = M0, scope = list(lower = M0, upper = Mfull),direction =
"both",trace = FALSE)

#Define model
M3<-lm(PRECIP~ALTITUDE*SHADOW+LATITUDE*SHADOW+DISTANCE*SHAD
OW,data=caliRain)
M4<-lm(PRECIP~ALTITUDE+LATITUDE+DISTANCE+SHADOW+SHADOW*ALTIT
UDE+SHADOW*LATITUDE+LATITUDE+ALTITUDE,data=caliRain)
```

****

3) R CODE OF MODEL DIAGNOSTICS

3.1) RESIDUAL PLOTS

```
res1 <- resid(M3) #residuals
sigma.hat1 <- sqrt(sum(res1^2)/M3$df)
z1<- res1/sigma.hat1 #standardized residuals
yhat1 <- predict(M3)
yhat2 <- predict(M4)
standard.res1 <- res1/sigma.hat1 #standardized residuals
par(mfrow=c(2,3 ))
plot(yhat1,res1,pch = 16, cex = 1, xlab = "Fitted values", ylab = "Residual")
abline(h = 0, lty = 2) #Add a horizontal line at y = 0
plot(data$ALTITUDE,res1,pch = 16, cex = 1, xlab = "ALTITUDE", ylab = "Residual")
abline(h = 0, lty = 2) #Add a horizontal line at y = 0
plot(data$SHADOW,res1,pch = 16, cex = 1, xlab = "SHADOW", ylab = "Residual")
```

```
abline(h = 0, lty = 2) #Add a horizontal line at y = 0
plot(data$DISTANCE,res1,pch = 16, cex = 1, xlab = "DISTANCE", ylab = "Residual")
abline(h = 0, lty = 2)
plot(data$LATITUDE,res1,pch = 16, cex = 1, xlab = "LATITUDE", ylab = "Residual")
abline(h = 0, lty = 2)

standard.res2<- res2/sigma.hat2 #standardized residuals
par(mfrow=c(2,3 ))
plot(yhat2,res2,pch = 16, cex = 1, xlab = "Fitted values", ylab = "Residual")
abline(h = 0, lty = 2) #Add a horizontal line at y = 0
plot(data$ALTITUDE,res2,pch = 16, cex = 1, xlab = "ALTITUDE", ylab = "Residual")
abline(h = 0, lty = 2) #Add a horizontal line at y = 0
plot(data$SHADOW,res2,pch = 16, cex = 1, xlab = "SHADOW", ylab = "Residual")
abline(h = 0, lty = 2) #Add a horizontal line at y = 0
plot(data$DISTANCE,res2,pch = 16, cex = 1, xlab = "DISTANCE", ylab = "Residual")
abline(h = 0, lty = 2)
plot(data$LATITUDE,res2,pch = 16, cex = 1, xlab = "LATITUDE", ylab = "Residual")
abline(h = 0, lty = 2)

par(mfrow = c(1,2))
hist(z1,breaks=20,freq=FALSE, xlim=c(-3,3), xlab = "Standardized Residuals",
main="Residual plot")
curve(dnorm, add = TRUE, col = "red") #PDF of N(0,1)
qqnorm(res1,pch = 16, cex = 1) #QQ plot
abline(qqline(res1),col = "red", lty = 2)
res2 <- resid(M4) #residuals
sigma.hat2 <- sqrt(sum(res2^2)/M4$df)
z2<- res2/sigma.hat2 #standardized residuals
par(mfrow = c(1,2))
hist(z2,breaks=20,freq=FALSE, xlim=c(-3,3), xlab = "Standardized Residuals",
main="Residual plot")
curve(dnorm, add = TRUE, col = "red") #PDF of N(0,1)
qqnorm(res2,pch = 16, cex = 1) #QQ plot
abline(qqline(res2),col = "red", lty = 2)
```

3.2)LEVERAGE AND INFLUENCE MEASURES

```
D1 <- cooks.distance(M3)
plot(D1, pch = 16, cex = 1, xlab = "Cook's Influence Measure", ylab = "Leverage")
D2 <- cooks.distance(M4)
```

```
plot(D2,pch = 16, cex = 1, xlab = "Cook's Influence Measure", ylab = "Leverage")
anova(M3)
anova(M4)

**********************************************************
3.3 CROSSVALIDATION

nreps <- 100 # number of replications
n <- nrow(data) # total number of observations
ntrain <- ceiling(0.9*n) # size of training set (90% of the data set)
ntest <- n-ntrain # size of test set
sse1 <- rep(NA, nreps) # sum-of-square errors for each CV replication
sse2 <- rep(NA, nreps)
log.likelihood1 <- rep(NA, nreps) # likelihod of of the model for the numerator for
each replication
log.likelihood2 <- rep(NA, nreps)
log.likelihood <- rep(NA, nreps)# likelihod of of the model for the denominator for
each replication

#Start of for loop
for(ii in 1:nreps) {
  if(ii%%100 == 0) message("ii = ", ii)
  # randomly select training observations
  train.ind <- sample(n, ntrain) # training observations
  M3.cv <- update(M3, subset = train.ind)
  M4.cv <- update(M4, subset = train.ind)
  # testing residuals for both models
  # that is, testing data - predictions with training parameters
  M3.res <- data$PREDIC[-train.ind] - predict(M3.cv, newdata = data[-train.ind,])
  M4.res <- data$PREDIC[-train.ind] - predict(M4.cv, newdata = data[-train.ind,])
  # total sum of square errors
  sse1[ii] <- sum((M3.res)^2)
  sse2[ii] <- sum((M4.res)^2)
  # testing likelihood ratio
  M3.sigma <- sqrt(sum(resid(M3.cv)^2)/ntrain) # MLE of sigma
  M4.sigma <- sqrt(sum(resid(M4.cv)^2)/ntrain)
  log.likelihood1[ii] <- sum(dnorm(M3.res, mean = 0, sd = M3.sigma, log = TRUE))
  log.likelihood2[ii] <- sum(dnorm(M4.res, mean = 0, sd = M4.sigma, log = TRUE))
}

#Comparison of SSE
mean(sse1)
```

```
mean(sse2)

#Likelihood ratio statistics
mean(log.likelihood1[ii] - log.likelihood2[ii]) #Average likelihood ratio (log)

#PRESS statistics
press1 <- resid(M3)/1-hatvalues(M4)
sum(press1^2)
press2 <- resid(M3)/1-hatvalues(M4)
sum(press2^2)

#AIC
AIC(M3)
AIC(M4)
```

****

4) OTHER R CODE USED IN ANALYSIS

In the Cook's Distance figure, we can also find 1 high leverage point in M3(19) and 2 high leverage point in M4(1, 6). We remove them and do the summary again, but for both models , the R^2 values have hardly changed at all (0.8767->0.9207 and 0.9406->0.9355) as well as the coefficient. Therefore, those points are neither outliers or influential, but just high leverage points, and we do not remove them.