36401 HW #11

Sylvia (Shuyuan) Ding
Shuyuand

Problem 1

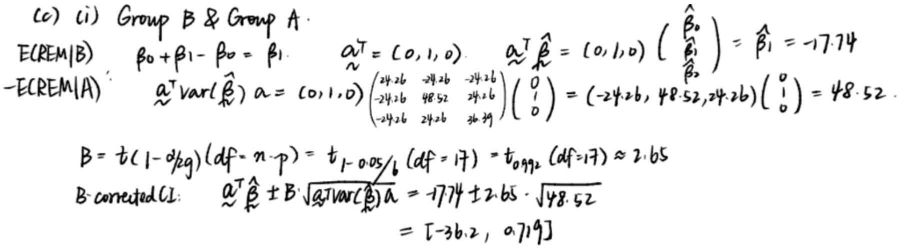
(a) X1 = 1(Low dose), 0 (control), 0 (high); X2 = 0 (Low dose), 0 (control), 1 (high)

(b) (i) beta0: The expected average REM sleep time for control group (no alcohol) is 79.28

(ii) beta1: The REM sleep time for low dose group is 17.74 less than control group

(iii) beta2: The REM sleep time for high dose group is 38.94 less than control group

(c) (i) With Bonferroni correction, we are 95% confident that the difference value of mean REM time between Group B and Group A falls in the range between -36.2 and 0.719



(c) (i) Group B & Group A.

$E(REM|B) - E(REM|A)$: $\beta_0+\beta_1-\beta_0=\beta_1$. $\underset{\sim}{a}^T=(0,1,0)$. $\underset{\sim}{a}^T\hat{\underset{\sim}{\beta}}=(0,1,0)\begin{pmatrix}\hat\beta_0\\\hat\beta_1\\\hat\beta_2\end{pmatrix}=\hat\beta_1=-17.74$

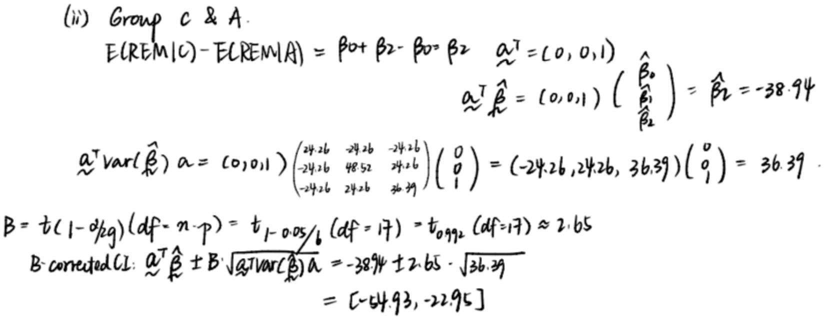
$$\underset{\sim}{a}^T var(\hat{\underset{\sim}{\beta}})\,a=(0,1,0)\begin{pmatrix}24.26 & -24.26 & -24.26\\-24.26 & 48.52 & 24.26\\-24.26 & 24.26 & 36.39\end{pmatrix}\begin{pmatrix}0\\1\\0\end{pmatrix}=(-24.26,48.52,24.26)\begin{pmatrix}0\\1\\0\end{pmatrix}=48.52.$$

$$B=t(1-\alpha/2g)(df=n-p)=t_{1-0.05/6}(df=17)=t_{0.992}(df=17)\approx 2.65$$

B-corrected CI: $\underset{\sim}{a}^T\hat{\underset{\sim}{\beta}}\pm B\cdot\sqrt{\underset{\sim}{a}^T var(\hat{\underset{\sim}{\beta}})a}=-17.74\pm 2.65\cdot\sqrt{48.52}$

$$=[-36.2,\ 0.719]$$

(ii) With Bonferroni correction, we are 95% confident that the difference value of mean REM time between Group C and Group A falls in the range between -54.93 and -22.95



(ii) Group C & A.

$E(REM|C)-E(REM|A)=\beta_0+\beta_2-\beta_0=\beta_2$ $\underset{\sim}{a}^T=(0,0,1)$

$$\underset{\sim}{a}^T\hat{\underset{\sim}{\beta}}=(0,0,1)\begin{pmatrix}\hat\beta_0\\\hat\beta_1\\\hat\beta_2\end{pmatrix}=\hat\beta_2=-38.94$$

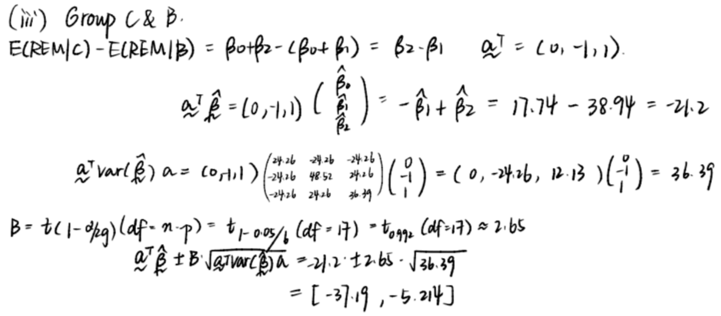
$$\underset{\sim}{a}^T var(\hat{\underset{\sim}{\beta}})\,a=(0,0,1)\begin{pmatrix}24.26 & -24.26 & -24.26\\-24.26 & 48.52 & 24.26\\-24.26 & 24.26 & 36.39\end{pmatrix}\begin{pmatrix}0\\0\\1\end{pmatrix}=(-24.26,24.26,36.39)\begin{pmatrix}0\\0\\1\end{pmatrix}=36.39.$$

$$B=t(1-\alpha/2g)(df=n-p)=t_{1-0.05/6}(df=17)=t_{0.992}(df=17)\approx 2.65$$

B-corrected CI: $\underset{\sim}{a}^T\hat{\underset{\sim}{\beta}}\pm B\cdot\sqrt{\underset{\sim}{a}^T var(\hat{\underset{\sim}{\beta}})a}=-38.94\pm 2.65\cdot\sqrt{36.39}$

$$=[-54.93,-22.95]$$

(iii) With Bonferroni correction, we are 95% confident that the difference value of mean REM time between Group C and Group B falls in the range between -37.19 and -5.21



(iii) Group C & B.

$E(REM|C)-E(REM|B)=\beta_0+\beta_2-(\beta_0+\beta_1)=\beta_2-\beta_1$ $\underset{\sim}{a}^T=(0,-1,1).$

$$\underset{\sim}{a}^T\hat{\underset{\sim}{\beta}}=(0,-1,1)\begin{pmatrix}\hat\beta_0\\\hat\beta_1\\\hat\beta_2\end{pmatrix}=-\hat\beta_1+\hat\beta_2=17.74-38.94=-21.2$$

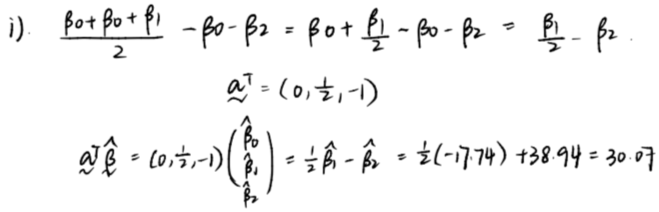
$$\underset{\sim}{a}^T var(\hat{\underset{\sim}{\beta}})\,a=(0,-1,1)\begin{pmatrix}24.26 & -24.26 & -24.26\\-24.26 & 48.52 & 24.26\\-24.26 & 24.26 & 36.39\end{pmatrix}\begin{pmatrix}0\\-1\\1\end{pmatrix}=(0,-24.26,12.13)\begin{pmatrix}0\\-1\\1\end{pmatrix}=36.39$$

$$B=t(1-\alpha/2g)(df=n-p)=t_{1-0.05/6}(df=17)=t_{0.992}(df=17)\approx 2.65$$

$\underset{\sim}{a}^T\hat{\underset{\sim}{\beta}}\pm B\cdot\sqrt{\underset{\sim}{a}^T var(\hat{\underset{\sim}{\beta}})a}=-21.2\pm 2.65\cdot\sqrt{36.39}$

$$=[-37.19,-5.214]$$

(d) (i) The linear combination of regression coefficients that corresponds to the model is (0, 1/2, -1)

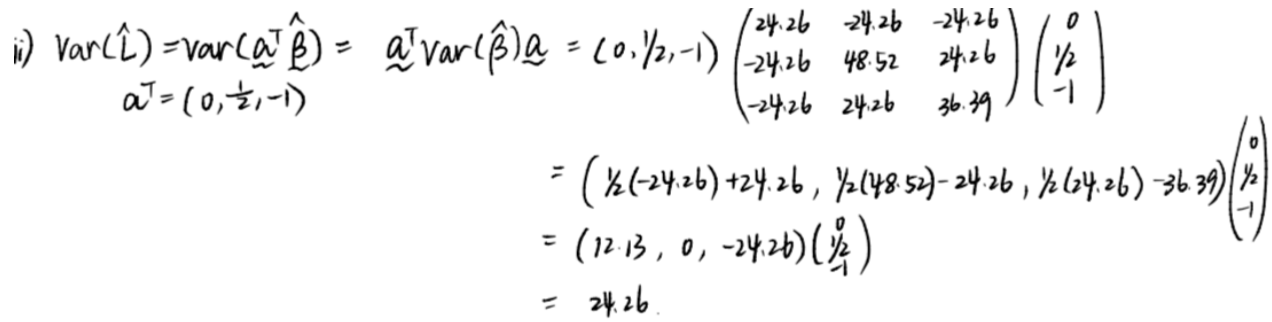


i). $\frac{\beta_0+\beta_0+\beta_1}{2} - \beta_0 - \beta_2 = \beta_0 + \frac{\beta_1}{2} - \beta_0 - \beta_2 = \frac{\beta_1}{2} - \beta_2$.

$$\underset{\sim}{a}^T = (0, \tfrac{1}{2}, -1)$$

$$\underset{\sim}{a}^T \underset{\sim}{\hat{\beta}} = (0, \tfrac{1}{2}, -1)\begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} = \tfrac{1}{2}\hat{\beta}_1 - \hat{\beta}_2 = \tfrac{1}{2}(-17.74) + 38.94 = 30.07$$

(ii)



ii) $Var(\hat{L}) = var(\underset{\sim}{a}^T \hat{\beta}) = \underset{\sim}{a}^T var(\hat{\beta}) \underset{\sim}{a} = (0, \tfrac{1}{2}, -1)\begin{pmatrix} 24.26 & -24.26 & -24.26 \\ -24.26 & 48.52 & 24.26 \\ -24.26 & 24.26 & 36.39 \end{pmatrix}\begin{pmatrix} 0 \\ \tfrac{1}{2} \\ -1 \end{pmatrix}$

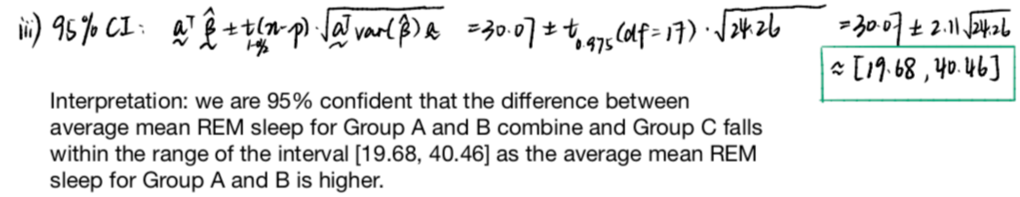
$a^T = (0, \tfrac{1}{2}, -1)$

$$= \left(\tfrac{1}{2}(-24.26) + 24.26,\ \tfrac{1}{2}(48.52) - 24.26,\ \tfrac{1}{2}(24.26) - 36.39\right)\begin{pmatrix} 0 \\ \tfrac{1}{2} \\ -1 \end{pmatrix}$$

$$= (12.13,\ 0,\ -24.26)\begin{pmatrix} 0 \\ \tfrac{1}{2} \\ -1 \end{pmatrix}$$

$$= 24.26.$$

(iii)



iii) 95% CI: $\underset{\sim}{a}^T \underset{\sim}{\hat{\beta}} \pm t_{1-\alpha/2}(n-p) \cdot \sqrt{\underset{\sim}{a}^T var(\hat{\beta}) \underset{\sim}{a}} = 30.07 \pm t_{0.975}(df=17) \cdot \sqrt{24.26} = 30.07 \pm 2.11\sqrt{24.26}$

$\boxed{\approx [19.68, 40.46]}$

Interpretation: we are 95% confident that the difference between average mean REM sleep for Group A and B combine and Group C falls within the range of the interval [19.68, 40.46] as the average mean REM sleep for Group A and B is higher.

Problem #2 IMRAD

Abstract

We are interested in what contributes to the variability in California rainfall, specifically on finding if there can be one model that predicts the Rainfall best with the given data. We also want a model that has explanatory variables that are most helpful in understanding variability in California Rainfall. We want to know if we can have good prediction of the amount of rainfall from the stations' altitude, latitude, distance to Pacific coastline and which side of the slope the station is on. A sample of average rainfall (in inches) are obtained from 30 meteorological stations scattered throughout the state. A linear regression of log(rainfall) vs different combinations of explanatory variables and interaction terms are fitted in order to find the best one. After evaluating these models by using R^2, R^2(ADJ), AIC, BIC and PRESS values, we found the best model to be:

$$\log(Rainfall) = -2.74 + 0.00026 \times Altitude + 0.16 \times Latitude - 0.008 \times Distance - 0.28 \times Shadow - 0.000126 \times Altitude{:}Shadow$$

The model is statistically significant in terms of explaining the variability in log(rainfall) ($p < 0.001$). The model also has R^2(ADJ) = 0.8046, AIC = 33.72, BIC = 43.52 and PRESS = 0.2123, which means the model explains 80.46% of variability of log(Rainfall) with the consideration of all explanatory variables in the model. Moreover, there seems to be no significant difference of expected mean of log(Rainfall) for different sides of the slope of the stations (95% CI: [-0.73, 0.17]). For further investigation, Several other linear models with the interaction terms of explanatory variables of this model are also fitted to analyze and compare the models by ANOVA Tests and variables that indicate the fit of the model (R^2(adj), AIC, BIC, PRESS etc.)

Introduction

The state of California operates numerous meteorological stations, and the main function of these stations is to monitor rainfall on a daily basis. In states with dry weather like California, it is important to keep temperature conditions monitored, and rainfall is an important part of the monitoring process since rainfall evaporation is part of what generates tropical humidity and induces temperature change. We are interested in the rainfall amount on a daily basis to produce an annual average precipitation level for each station, particularly on what are some of the variables that can best predict California Rainfall amount and if we can develop a best model to predict. With a prediction model, we can also possibly better predict the amount of rainfall in other states. Other than prediction, we are interested in whether a station is on the leeward will contribute statistically significant difference on the rainfall amount.

Methods

This dataset contains 30 stations scattered throughout the state, and the variables included in the dataset are as follows:

- Rainfall: average annual rainfall (in inches)
- Altitude: altitude of the station (in feet)
- Latitude: latitude of the station (in degrees)
- Distance: distance of the station from the Pacific coast (miles)
- Shadow: 1 = if station is on the leeward (eastern slope), 0 = if the station is on the westward-facing slope
- Station: station name
- Station.ID: an identification number of each station

Initial EDA is performed to investigate the univariate distribution of the variables. Before doing that, we exclude variables that we do not want to perform analysis on. These are Station and Station.ID, because the names of the station and the identification number of the station are not the variables we want to analyze on in terms of how it impacts the annual precipitation level of the station. They are rather the indices of the dataset and names of stations. Upon evaluation of our variables, we found that the response variable Rainfall is highly right skewed. Therefore, we consider doing some transformation by creating variables log.rainfall = log(Rainfall). After transformation, we use pairwise matrix plot to evaluate the relationship among the variables. Figure 1 shows all pairwise relationship between different variables.

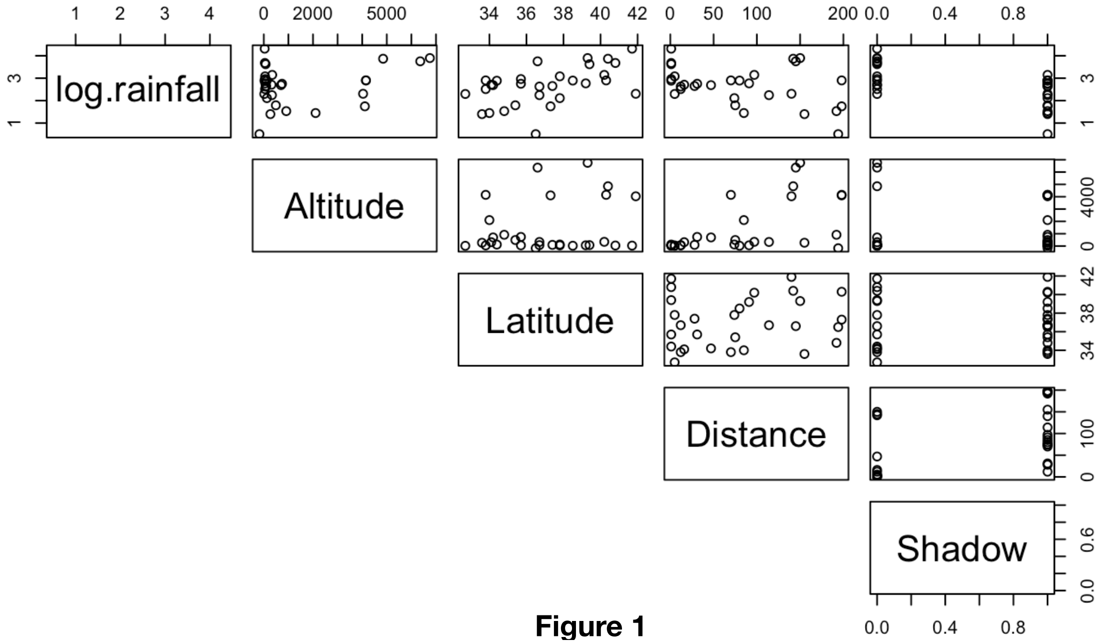


Figure 1

As Figure 1 shows, the positive correlation between Latitude and log(Rainfall) is the most obvious. Other seemingly positive correlations are Distance and Altitude, and Altitude and log(Rainfall). Other than that, we see there is some negative relationship between Distance and log(Rainfall). Currently the data pattern of correlation between Distance and Latitude, and Latitude and Altitude still seem somehow random with no clear trend, and we cannot draw conclusion from only the graphs in this matrix.

To start finding a best model that can explain the variability in California rainfall, we make our initial model to be a reduced model that includes all variables we just discussed but without considering the interactions between the pairs of explanatory variables:

M1 = lm(log.rainfall~Altitude +Latitude+Distance+shadow

After having M1, we want to build on top of it by attempting to add in all the interaction terms between the explanatory variables. The new model then comes to be:

M2 <-lm(log.rainfall~Altitude + Latitude + Distance + Shadow + Altitude:Latitude + Altitude:Distance + Altitude:Shadow + Latitude:Distance+Latitude:Shadow+Distance:Shadow, data = Calrain.

Then we examine the VIFs (Variation Inflation Factors) of all the interactions to inspect the potential multicollinearity among explanatory variables and eliminate the interaction terms that have really high VIF values. After elimination, we have reached our newer Model:

M3=log.rainfall~Altitude+Latitude+Distance+Shadow+Altitude:Latitude+Altitude:Shadow, data = Calrain.

Again, we examine the VIFs and see if all VIFs now are not exceptionally high for us to go further. After this elimination, now it left us with Model 4 (M4):

M4=log.rainfall~Altitude+Latitude+Distance+Shadow+Altitude:Shadow, data = Calrain.

This model has all the explanatory variables from Model 1 (M1) with an additional interaction term Altitude:Shadow. To make sure we this is the model we want, we examine the VIF again.

Table 1

	Altitude	Latitude	Distance	Shadow	Altitude:Latitude	Altitude:Distance	Altitude:Shadow	Latitude:Distance	Latitude:Shadow	Distance:Shadow	F* Global
M1	x	x	x	x							27.93 (p<0.01)

M2	x	x	x	x	x	x	x	x	x	x	12.59 (p<0.01)
M3	x	x	x	x	x		x				23.34 (p<0.01)
M4	x	x	x	x			x				24.88 (p<0.01)

Table 1 above illustrates the process of adding in the interactions into the model and later eliminate the interactions from the models based on VIF values. The table shows what variables are included in each of the four models during our analysis. After having Model 4 (M4), we want to inspect some of its statistics to see if this model explains the variability of the response variable well and if its residual plot violates the normality assumptions. We will compare the statistics with other models and with our initial model (M1) to see which is the best model that we want to best predict the response variable-log(Rainfall). Moreover, we want to explore whether a station is on the eastern slope will create any difference in terms of the expected mean of California log(rainfall) (Shadow = 1) under this model. To explore that, we calculate the 95% confidence interval of the comparison between stations on eastern slope and stations not on eastern slope again. Then we will conclude if they are statistically significant in terms of explaining the variation of California Rainfall by seeing if the confidence interval includes 0.

Results

The summary of Model 1 (M1) shows all explanatory variables are statistically significant in terms of explaining the variance of the response variable (all p < 0.05) and together they explain about 78.8% of total variation of log(Rainfall). Also with this model the 95% confidence interval for shadow is [-0.88, -0.11]. The interpretation here is that under this reduced model, we are 95% confident the difference of expected mean of log(Rainfall) between leeward station and stations at westward-facing slope falls in the range within [-0.88, -0.11]. And since the confidence interval does not include 0, under this model, the expected mean of log(Rainfall) in leeward station (Shadow = 1) is statistically different from stations at westward-facing slope (Shadow = 0). Also the residual plot of this M1 does not violate the normality assumptions. Figure 2 on the next page shows the residual plot of model1, which the residuals are randomly scattered above and below the 0-line as the lowess plot shows and the assumptions hold.

For Model 2 (M2), it is also statistically significant by itself (p <0.01), but wee eliminate all interaction variables that have really high VIFs-the ones with VIF values greater than 100 (

(VIF(Altitude:Latitude) = 839.88, VIF(Altitude:Distance) = 31.23, VIF(Altitude:Shadow) = 26.49, VIF(Latitude:Distance) =1645.35, VIF(Latitude:Shadow) = 774.41, VIF(Distance:Shadow) = 117.65).

Similar result shows with Model 3(M3). It is again a statistically significant model by itself (p<0.01), and it has an interaction term with really high VIF: the VIF value of Altitude:Latitude is exceptionally high (VIF(Latitude:Altitude) = 330.34) in this model, so we eliminate from this model (M3) and we reach our Model 4 (M4). Model 4 has no high VIF interaction terms (VIF(Altitude:Shadow) = 2.07).

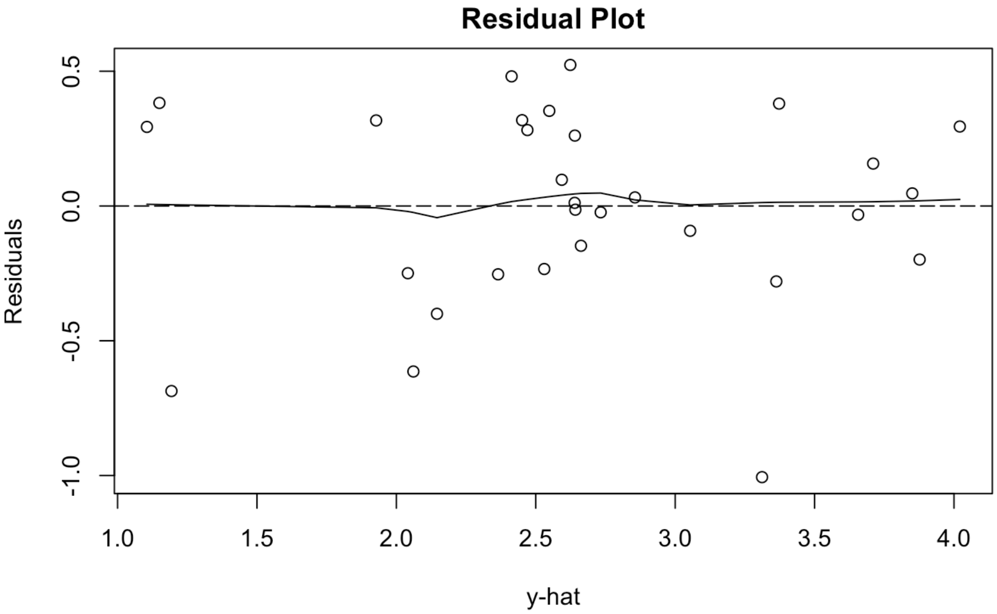


Figure 2

Table 2					
Model	R^2	R^2(ADJ)	PRESS	AIC	BIC
M1	0.8171	0.7879	0.2021	35.4022	43.8094
M2	0.8688	0.7998	0.8083	36.3947	49.0055
M3	0.8589	0.8221	0.2055	31.6167	42.8262
M4	0.8383	0.8046	0.2123	33.7161	43.5245

Interestingly, For the models that we have built, none of them (M2, M3 and M4) shows a p-value less than 0.05 in the bi-model anova test with M1, which means none of the model improves fit of the earliest model (M1). However, our final model (M4) does have a higher R^2(ADJ) value and lower AIC and BIC values as Table 2 above shows. It means that the final model we developed (V4) explains more variation of California rainfall when accounting for the number of predictors in the model, and it is the model that we want to choose. Also all models that we have built by themselves have p-values less than 0.05, which means all the models from M1 to M4 are statistically significant in terms of explaining the variation of California Rainfall by their fit. Something interesting to notice here is that although M3 is not the best model because one of the interaction term it includes (Latitude:Altitude) has really high VIF, but it has the highest R^2(ADJ) among all the models we developed and the lowest AIC and BIC from Table 2.

And upon evaluating the boxplot of residuals for M4 (Figure 3) and the norm-QQ plot (Figure 4), we see that there is no outlier for residuals, and the distribution of residual seems to be randomly distributed. Also the normal-QQ plot shows the trend that the dots are mostly close to the line with only some deviations at the tails, so the residual normality assumptions are not violated for Model 4 (M4).

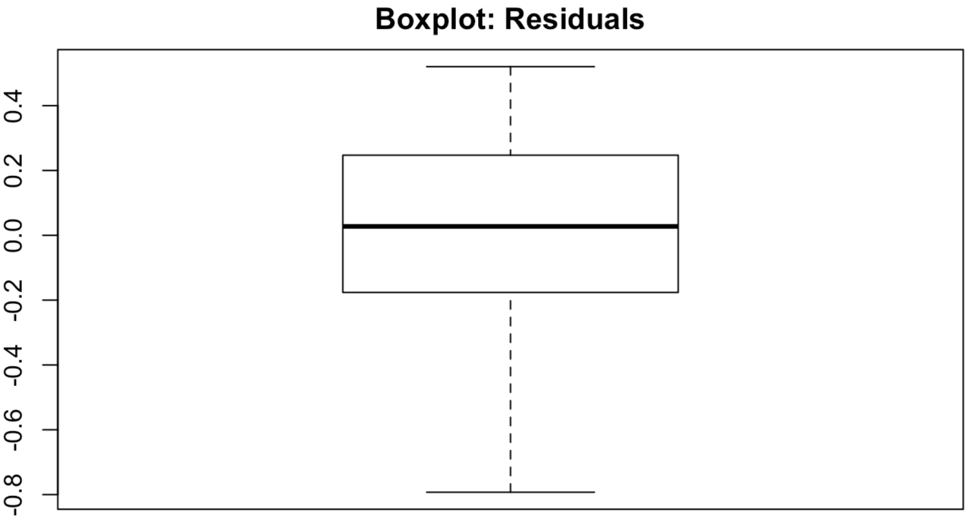


Figure 3

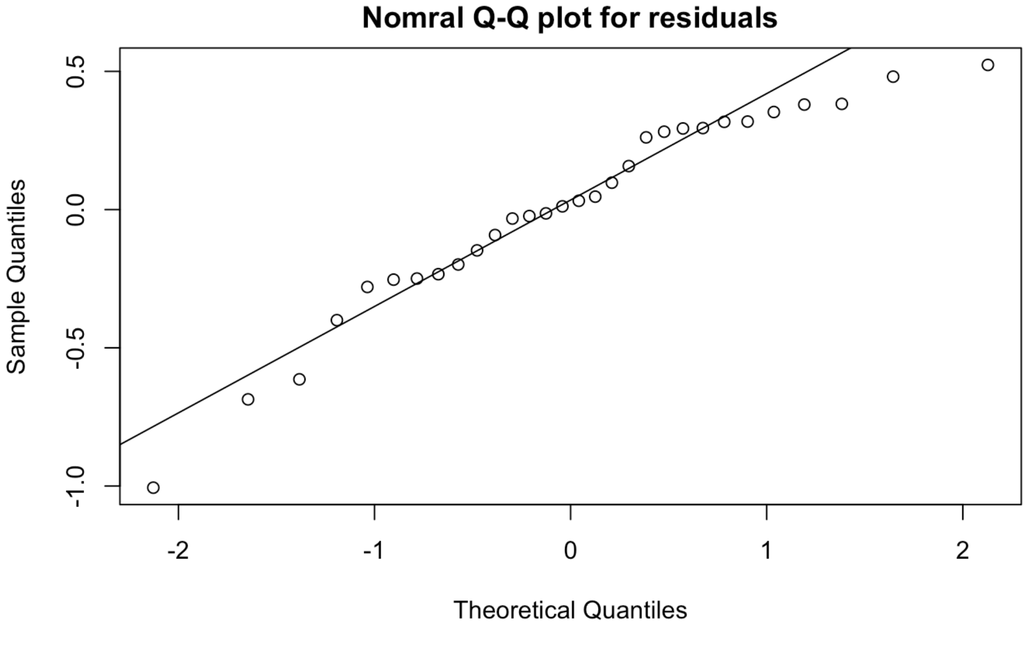


Figure 4

After considering all aspects of building and selecting the model, we conclude that the best model for us to predict log of California rainfall is Model 4 (M4), which is the one with all explanatory variables plus one interaction term between Altitude and Shadow. The reason is that although it is not the one explains the most variability of the response variable, and it does reject the null hypothesis during the ANOVA Test conducted to compare with Model 1 (M1), it has a R^2(ADJ) higher than Model 1(M1), and lower AIC and BIC than Model 1 (M1). And for all the other models (M2 and M3), they both have interaction terms that have high VIF values. Thus we choose Model 4 (M4) to be our best model to predict the response variable for this dataset.

In terms of the difference between eastern-slope and westward-facing slope stations, the 95% Confidence Interval for the comparison under this model is [-0.73, 0.17]. The interpretation here is under this new model (M4), we are 95% confident the difference of expected mean of log(Rainfall) between leeward station and stations at westward-facing slope falls in the range within [-0.73, 0.17]. And since the confidence interval does include 0 now, under this model, expected mean of log(Rainfall) in leeward station (Shadow = 1) is not statistically different from stations at westward-facing slope (Shadow = 0).

Discussion

The log transformation of Rainfall made us to see a clearer trend between pairwise data and build our initial models with residuals that does not violate assumptions. P-values of each model enables us to identify that all the models that we developed during the building process are statistically significant by itself. However, several models we had includes interaction terms that have exceptionally high VIF values to indicate multicollinearity. Thus, one important step during this model building process is to eliminate these interaction terms with high VIFs. Also for the later models we developed, we performed ANOVA Test on all of them in comparison with Model 1 (M1), but we found none of them fits the dataset better than our initial Model (M1). But meanwhile, we also analyze on the statistics like R^2(ADJ), AIC, BIC, and PRESS value, and after consider all aspects of the models, we think Model 4 (M4) is better than Model 1 (M1) and choose it to be our best and final model. Interestingly, in the initial model, whether a station is on the leeward (Shadow = 1) will make the expected mean fo log(Rainfall) statistically significantly different from the stations not on the leeward (Shadow = 0), but as we switch to the best model we found (M4), this statistically significant difference disappears.

Appendix

```
Calrain <- read.csv("Calrain.csv", header = T)
head(Calrain)
Calrain$log.rainfall = log(Calrain$Rainfall)
pairs(log.ainfall~Altitude+Latitude+Distance+Shadow, data = Calrain, lower.panel = NULL)
pairs(sqrt(Rainfall)~Altitude+Latitude+Distance+Shadow, data = Calrain, lower.panel = NULL)
M1 <- lm(log.rainfall~Altitude+Latitude+Distance+Shadow, data = Calrain)
summary(M1)
vif(M1)
AIC(M1)
BIC(M1)
PRESS.M1 <- mean((resid(M1)/(1-hatvalues(M1)))^2)
PRESS.M1
plot(x = fitted(M1), # fitted values on x axis
    y = residuals(M1), # residuals on y axis
    xlab = "Fitted", # x label
    ylab = "Residuals", # y label
    main = "Residuals vs Fitted" # title label
)
abline(h = 0) # draw a horizontal line at 0

y.hat1 <- fitted(M1)
ep.hat1 <- resid(M1)
plot(y.hat1, ep.hat1, main = "Residual Plot", ylab = "Residuals", xlab = "y-hat")
abline(h=0, lty = 5)
lines(lowess(ep.hat1~y.hat1, f = 8/10, iter = 3), lty = 1)

boxplot(ep.hat1, main = "Boxplot: Residuals")
qqnorm(ep.hat1, main = "Nomral Q-Q plot for residuals")
qqline(ep.hat1)

M2 <-
lm(log.rainfall~Altitude+Latitude+Distance+Shadow+Altitude:Latitude+Altitude:Distance+Altitude:Shadow+Latitude:Distance+Latitude:Shadow+Distance:Shadow, data = Calrain)

summary(M2)
vif(M2)

PRESS.M2 <- mean((resid(M2)/(1-hatvalues(M2)))^2)
PRESS.M2
anova(M1, M2)
AIC(M2)
BIC(M2)
M3 <- lm(log.rainfall~Altitude +Latitude+Distance+Shadow+Altitude:Latitude+Altitude:Shadow, data =
Calrain)
summary(M3)
vif(M3)
anova(M1, M3)
AIC(M3)
BIC(M3)
PRESS.M3 <- mean((resid(M3)/(1-hatvalues(M3)))^2)
PRESS.M3
M4 <- lm(log.rainfall~Altitude +Latitude+Distance+Shadow+Altitude:Shadow, data = Calrain)
summary(M4)
vif(M4)
anova(M1, M4)
```

```
AIC(M4)
BIC(M4)
PRESS.M4 <- mean((resid(M4)/(1-hatvalues(M4)))^2)
PRESS.M4

y.hat4 <- fitted(M4)
ep.hat4 <- resid(M4)
plot(y.hat4, ep.hat4, main = "Residual Plot", ylab = "Residuals", xlab = "y-hat")
abline(h=0, lty = 5)
lines(lowess(ep.hat4~y.hat4, f = 8/10, iter = 3), lty = 1)

boxplot(ep.hat4, main = "Boxplot: Residuals")
qqnorm(ep.hat4, main = "Nomral Q-Q plot for residuals")
qqline(ep.hat4)

M.CI <- lm(log.rainfall~Altitude+Latitude+Distance+factor(Shadow), data = Calrain)
summary(M.CI)
confint(M.CI)
confint(M1)
confint(M4)
```