Midterm

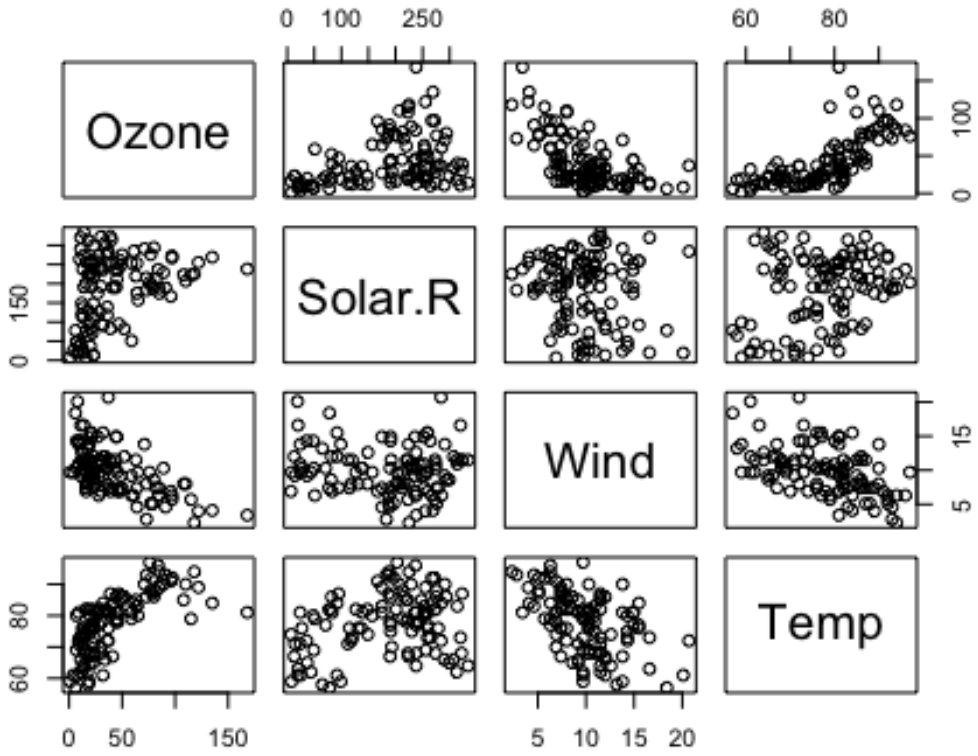
1. The airquality data set in the default R package includes New York Air Quality Measurements (ozone (ppb), solar radiation (lang), wind speed (mph) and maximum temperature (in Fahrenheit)) for 153 days. See help(airquality) for details. We call these four variables as Ozone (or O), Solar (or S), Wind (or W) and Temperature (or T) hereafter. We want to explain Ozone by Solar, Wind and Temperature.

(a) Remove all rows with one or more NAs. Report the remaining number of observations. Use this subset in the following questions.

```
aq <- na.omit(airquality)
```

(b) Make a pairwise scatter plot of the four variables: Ozone, Solar, Wind and Temperature.

```
pairs(aq[,c(1,2,3,4)])
```



(c) Calculate the correlation coefficients for each pair of the four variables.

```
cor(aq$Ozone, aq$Solar.R)
```

```
## [1] 0.3483417
```

```
cor(aq$Ozone, aq$Wind)

## [1] -0.6124966

cor(aq$Ozone, aq$Temp)

## [1] 0.6985414

cor(aq$Solar.R, aq$Wind)

## [1] -0.1271835

cor(aq$Solar.R, aq$Temp)

## [1] 0.2940876

cor(aq$Wind, aq$Temp)

## [1] -0.4971897
```

(d) Regress Ozone on the other three variables by multiple linear regression. Report the 'summary' of the linear regression, and interpret the sign and significance of estimated coefficients.

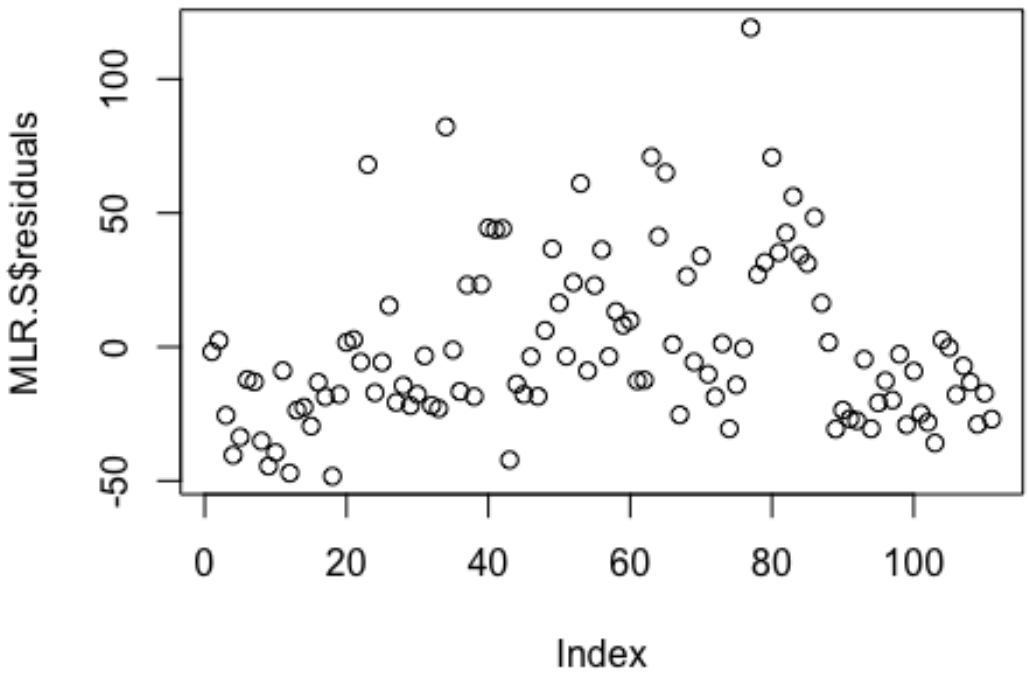
As Ozone Increases by 1 ppb, Solar Radiation increases by .06 lang, Tempurature increases by 1.65 degrees fahrenheit, and Wind decreases by 3.33 mph all the coefficients are significant falling well below .05

```
MLR <- lm(aq$Ozone ~ aq$Solar.R + aq$Wind + aq$Temp)
summary(MLR)

##
## Call:
## lm(formula = aq$Ozone ~ aq$Solar.R + aq$Wind + aq$Temp)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -40.485 -14.219  -3.551  10.097  95.619
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -64.34208   23.05472  -2.791  0.00623 **
## aq$Solar.R    0.05982    0.02319   2.580  0.01124 *
## aq$Wind      -3.33359    0.65441  -5.094 1.52e-06 ***
## aq$Temp       1.65209    0.25353   6.516 2.42e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 21.18 on 107 degrees of freedom
## Multiple R-squared:  0.6059, Adjusted R-squared:  0.5948
## F-statistic: 54.83 on 3 and 107 DF,  p-value: < 2.2e-16
```

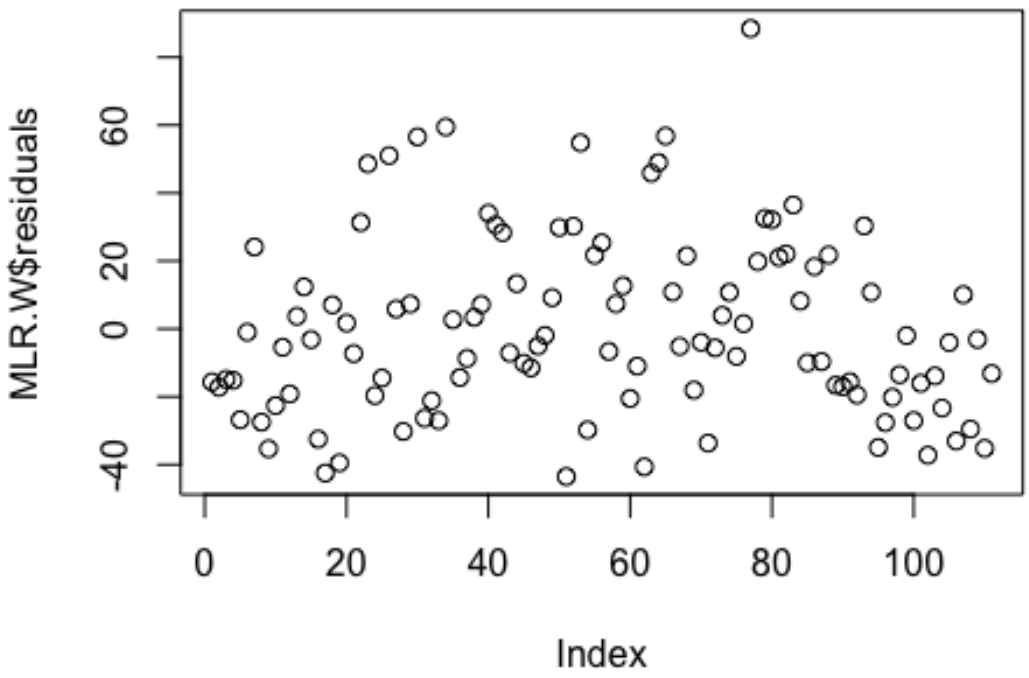
(e) Make a residual plot against each of the predictors (Solar, Wind and Temperature). (Confirm that the plots indicate that errors are roughly IID.)

Each of the residual plots are roughly IID with no perceivable pattern

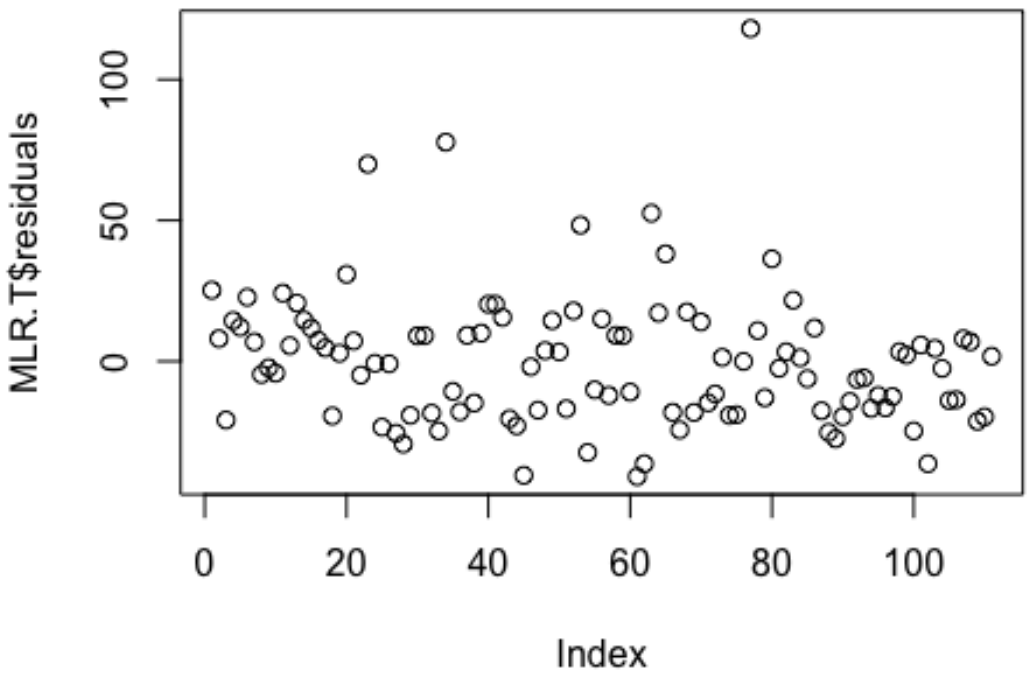
```
MLR.S <- lm(aq$Ozone ~ aq$Solar.R)
MLR.W <- lm(aq$Ozone ~ aq$Wind)
MLR.T <- lm(aq$Ozone ~ aq$Temp)

plot(MLR.S$residuals)
```




```
plot(MLR.W$residuals)
```




```
plot(MLR.T$residuals)
```



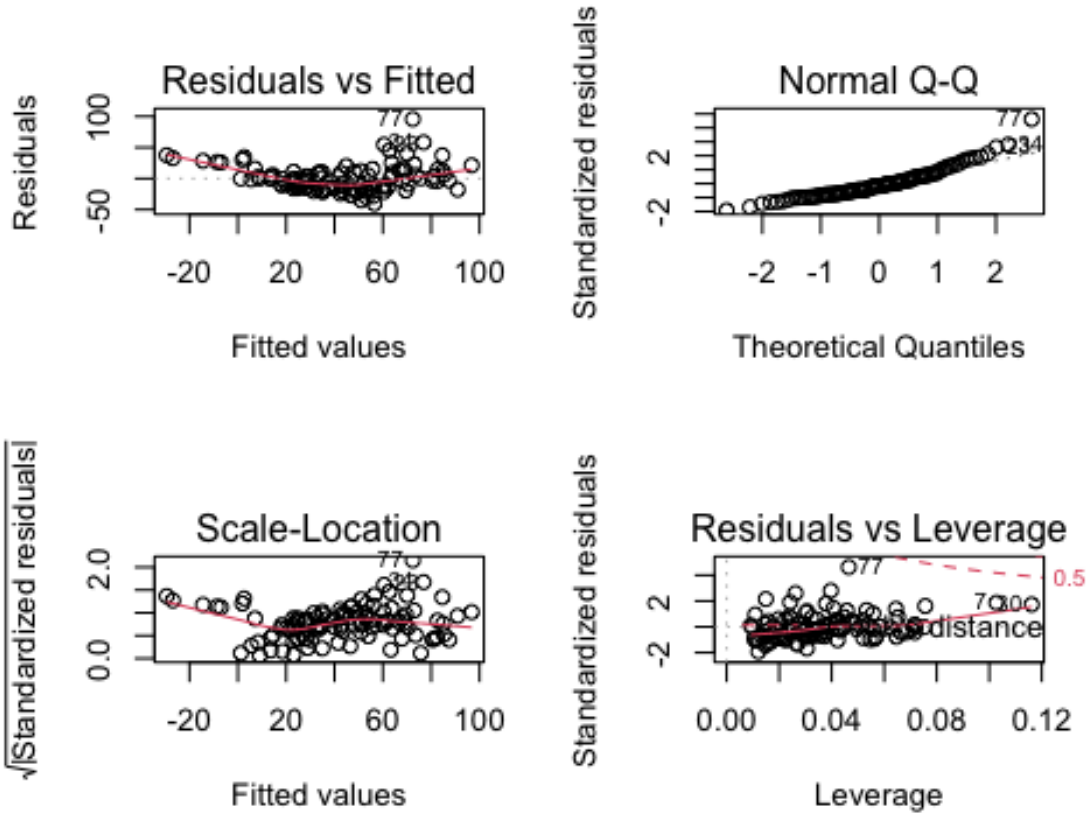
(f) Run the following code to generate the diagnostic plots: par(mfrow=c(2,2)) # this aligns four plots in 2 by 2 panes plot(name_of_your_model_in_d) par(mfrow=c(1,1)) Discuss possible improvements of the model by seeing the Residuals vs Fitted plot. Discuss possible improvements of the model by seeing the Normal Q-Q plot. Make a remark/mention limitation indicated by Residuals vs Leverage plot.

Near the end of the RvsF plot the there are a few outliers bringing the regression line up if we eliminated the outliers that would significantly improve the model.

Same could be said about the Normal QQ Plot the observations are going at about a straight line but a few start to taper off near the last Quantile.

Residuals vs leverage fits the data fairly well but the bounding does not fit.

```
par(mfrow=c(2,2)) # this aligns four plots in 2 by 2 panes
plot(MLR)
```




```
par(mfrow=c(1,1))
```

(g) Regress Ozone on all two-way interaction and quadratic terms (i.e., S2, S · W , S · T , W 2, W · T , T2) in addition to S, W and T. Show the 'summary' of the regression result.

```
MLR2 <- lm(Ozone ~ (Solar.R^2) + (Solar.R*Wind) + (Solar.R*Temp) + (Wind^2) +
(Wind*Temp) + (Temp^2),data = aq)
summary(MLR2)

##
## Call:
## lm(formula = Ozone ~ (Solar.R^2) + (Solar.R * Wind) + (Solar.R *
##     Temp) + (Wind^2) + (Wind * Temp) + (Temp^2), data = aq)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -38.685 -11.727  -2.169   7.360  91.244
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -1.408e+02  6.419e+01  -2.193  0.03056 *
## Solar.R      -2.260e-01  2.107e-01  -1.073  0.28591
## Wind          1.055e+01  4.290e+00   2.460  0.01555 *
```

```
## Temp          2.322e+00  8.330e-01   2.788  0.00631 **
## Solar.R:Wind -7.231e-03  6.688e-03  -1.081  0.28212
## Solar.R:Temp  5.061e-03  2.445e-03   2.070  0.04089 *
## Wind:Temp    -1.613e-01  5.896e-02  -2.735  0.00733 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 19.17 on 104 degrees of freedom
## Multiple R-squared:  0.6863, Adjusted R-squared:  0.6682
## F-statistic: 37.93 on 6 and 104 DF,  p-value: < 2.2e-16
```

(h) Calculate the AIC of the model in (d) and the model in (g). Which model is considered as a better model to predict Ozone for new observations?

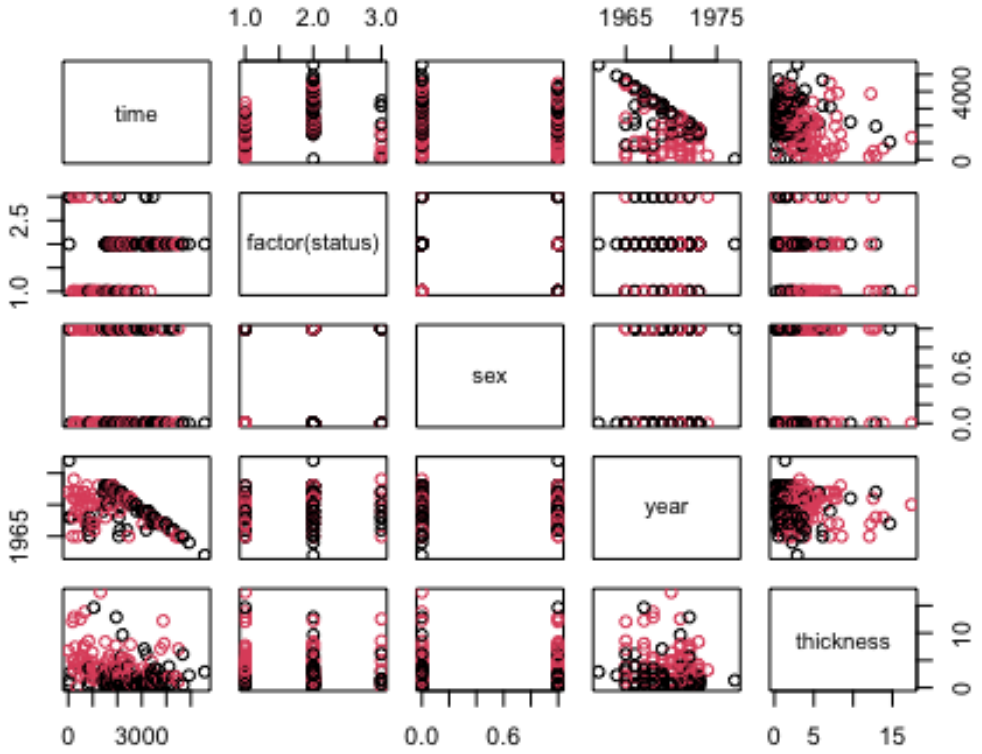
Regression from part g is the better fit.

```
AIC(MLR)

## [1] 998.7171

AIC(MLR2)

## [1] 979.3775
```

2. The 'melanoma' dataset in 'boot' package includes 7 variables: time, status, sex, age, year , (tumor) thickness , ulceration (1: present; 0: absent) for 205 patients with malignant melanoma. See 'help(melanoma)' for the definition of variables. We want to make a model to explain the presence of ulceration in patients.

(a) Make a scatter plot for the six variables time, status, sex, age, year, thickness. Differentiate points by the ulceration by color. (Note: Use the JPEG format for the plot to make the file small.) 'status' should be a factor variable, so transform it from numeric to factor.

```
library(boot)
pairs(~time + factor(status) + sex + year + thickness, col=factor(melanoma$ul
cer), data=melanoma)
```



(b) Regress ulceration ('ulcer') on all other variables with logistic regression, and report the sum- mary of the regression. Determine what predictor(s) are statistically significant at a 5% sig- nificance level and in which direction (e.g., higher age is associated with a higher chance of ulceration).

The only predictor that passes the .05 significance level is thickness but status and time are close to being significant.

Higher thickness is associated with higher chance of ulceration.

```
GLM <- glm(ulcer ~ time + status + sex + age + year + thickness,family=binomi
al, data=melanoma)
summary(GLM)

##
## Call:
## glm(formula = ulcer ~ time + status + sex + age + year + thickness,
##     family = binomial, data = melanoma)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.3247  -0.7934  -0.6089   0.8992   1.9623
##
## Coefficients:
```

```
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  7.380e+01  1.586e+02   0.465   0.6416
## time        -3.422e-04  2.022e-04  -1.693   0.0905 .
## status      -6.261e-01  3.360e-01  -1.863   0.0624 .
## sex          2.437e-01  3.432e-01   0.710   0.4776
## age          2.196e-03  1.044e-02   0.210   0.8334
## year        -3.728e-02  8.047e-02  -0.463   0.6431
## thickness    3.851e-01  8.798e-02   4.378  1.2e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 281.13  on 204  degrees of freedom
## Residual deviance: 223.18  on 198  degrees of freedom
## AIC: 237.18
##
## Number of Fisher Scoring iterations: 5
```

(c) We want to drop unnecessary predictors by 'backward deletion'. Backward deletion iteratively drops one of the predictors from the model so that the AIC is minimized, the iteration stops when dropping one more predictor always makes AIC larger. Use the 'step' function for backward deletion. The format is 'step(name_of_glm_object)'. Report the final model and interpret statistical significance of each of predictors.

With the removed predictors. The significance level of each predictor in each model improved. Time is still slightly not significant.

```
step(GLM)
```

```
## Start:  AIC=237.18
## ulcer ~ time + status + sex + age + year + thickness
##
##             Df Deviance    AIC
## - age        1   223.23 235.23
## - year       1   223.40 235.40
## - sex        1   223.69 235.69
## <none>           223.19 237.19
## - time       1   226.06 238.06
## - status     1   226.70 238.70
## - thickness  1   250.89 262.89
##
## Step:  AIC=235.23
## ulcer ~ time + status + sex + year + thickness
##
##             Df Deviance    AIC
## - year       1   223.43 233.43
## - sex        1   223.73 233.73
## <none>           223.23 235.23
## - time       1   226.31 236.31
```

```
## - status     1   226.70 236.70
## - thickness  1   252.03 262.03
##
## Step:  AIC=233.43
## ulcer ~ time + status + sex + thickness
##
##             Df Deviance    AIC
## - sex        1   223.94 231.94
## <none>           223.43 233.43
## - time       1   226.88 234.88
## - status     1   228.12 236.12
## - thickness  1   254.88 262.88
##
## Step:  AIC=231.94
## ulcer ~ time + status + thickness
##
##             Df Deviance    AIC
## <none>           223.94 231.94
## - time       1   227.62 233.62
## - status     1   228.68 234.68
## - thickness  1   258.08 264.08

##
## Call:  glm(formula = ulcer ~ time + status + thickness, family = binomial,
##     data = melanoma)
##
## Coefficients:
## (Intercept)         time       status    thickness
##   0.5179847   -0.0003055   -0.6787092    0.4080844
##
## Degrees of Freedom: 204 Total (i.e. Null);  201 Residual
## Null Deviance:       281.1
## Residual Deviance: 223.9     AIC: 231.9
```

```
GLM2 <- glm(ulcer ~ time + status + thickness,family=binomial, data=melanoma)
summary(GLM2)
```

```
##
## Call:
## glm(formula = ulcer ~ time + status + thickness, family = binomial,
##     data = melanoma)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.3190  -0.7996  -0.5980   0.9212   1.9530
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.5179847  0.6504255   0.796   0.4258
## time        -0.0003055  0.0001610  -1.897   0.0578 .
```

```
## status      -0.6787092  0.3145346  -2.158   0.0309 *
## thickness    0.4080844  0.0851858   4.791 1.66e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 281.13  on 204  degrees of freedom
## Residual deviance: 223.94  on 201  degrees of freedom
## AIC: 231.94
##
## Number of Fisher Scoring iterations: 5
```

(d) Assume that ulceration is predicted as present if its predicted probability in (c) is more than 0.5. Make a contingency table to summarize the prediction results in (c). Also calculate the error rate (i.e., the proportion of correct prediction).

```
GLM2.predict <- predict(GLM2, type="response")
c <- GLM2.predict > .5
t <- table(as.integer(c), melanoma$ulcer)
t

##
##      0  1
##   0 99 33
##   1 16 57

sum(diag(t))/sum(t)

## [1] 0.7609756
```

(e) Do 10-fold CV for 20 times to evaluate the error rate of the model in (d), and report the mean and the standard deviation of the error rate. (Hint: You have to define the error rate function first, then use it in 'cv.glm' function. Cf. Exer0222s.pdf).

```
correct.rate <- function(x,y){
  y2 <- (y > .5)
  TA <- table(x,y2)
  sum(diag(TA))/sum(TA)
}

correct.rate(melanoma$ulcer, GLM2.predict)

## [1] 0.7609756

CV.err <- numeric(20)

for (i in 1: 20)
{
  GLM.i <- glm(ulcer ~ time + status + thickness,family=binomial, data =melan
oma)
  CV.err[i] <- cv.glm(melanoma$ulcer, GLM2, cost=correct.rate, K=10)
```

```
}

sd(CV.err)
mean(CV.err)
```

(f) Implement linear discriminant analysis to predict the presence of ulceration by using the same predictors as (c). Make a contingency table in the same way as (d).

```
library(MASS)
LDA <- lda(ulcer ~ time + status + thickness, data =melanoma)
LDA.Pred <- predict(LDA)
table(LDA.Pred$class , melanoma$ulcer)

##
##       0   1
##   0 101  39
##   1  14  51
```

(g) Divide the data set into a training set (with a sample size of 103) and a test set (with a sample size of 102). Then, implement k-nearest neighborhood (with k = 10) to predict the presence of ulceration by using the same predictors as (c). Make a contingency table for the test set in the same way as (d).

```
library(class)
m.train <- melanoma[(1:103),c(1,3,6)]
m.test <- melanoma[(104:205),c(1,3,6)]
KNN <- knn(m.train,m.test,melanoma$ulcer[1:103],k=10)
table(KNN,melanoma$ulcer[1:102])

##
## KNN  0  1
##   0 49 53
##   1  0  0
```

3. The Default of credit card clients data set available at: https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients includes 30,000 consumers' demographic and credit information and whether or not each person defaulted in the next month: Y (= 1 if default, = 0 if not). We want to predict Y by the other variables.

(a) Regress Y on a polynomial of PAY_0 by logistic regression. Choose an appropriate degree for the polynomial by AIC (the optimal degree may not be perfectly objective, so decide by yourself). The variable PAY_0 represents payment delay of the most recent month and is considered the most relevant variable for Y .

7 degrees

```
library("readxl")
Default <- read_excel("default_of_credit_card_clients.xls")

Model.Default <- glm(`default payment next month`~PAY_0,family=binomial, data
=Default)
```

```
for (i in 1:10)
{
  print(i)
  print(AIC(glm(`default payment next month`~poly(PAY_0,i),family=binomial, d
ata=Default)))
}
```

```
## [1] 1
## [1] 28535.57
## [1] 2
## [1] 28167.69
## [1] 3
## [1] 27625.62
## [1] 4
## [1] 27613.08
## [1] 5
## [1] 27362.84
## [1] 6
## [1] 27164.8
## [1] 7
## [1] 27159.03
## [1] 8
## [1] 27160.97
## [1] 9
## [1] 27162.53
## [1] 10
## [1] 27164.2
```

```
Model.Default <- glm(`default payment next month`~poly(PAY_0,7),family=binomi
al, data=Default)
```

(b) Calculate the AUC (area under the curve; see Chapter 4 slides 50-51 for details) of the selected model in (a). You can use the auc function in the pROC package.

```
library(pROC)
```

```
## Type 'citation("pROC")' for a citation.
```

```
##
## Attaching package: 'pROC'
```

```
## The following objects are masked from 'package:stats':
##
##     cov, smooth, var
```

```
Model.Default.pred <- predict(Model.Default,type="response")
auc(Default$`default payment next month`,Model.Default.pred)
```

```
## Setting levels: control = 0, case = 1
```

```
## Setting direction: controls < cases
```

```
## Area under the curve: 0.7124
```

(c) Calculate the AUC (area under the curve) of the model in (a) by 5-fold cross-validation.

```
cv.glm(Default$`default payment next month`,Model.Default, k=5)
```

(d) Regress the same model as (a) but by LDA. Use the same degree polynomial as (a).

```
LDA.Default <- lda(`default payment next month`~poly(PAY_0,7), data=Default)
LDA.Default.pred <- predict(LDA.Default)
```

(e) Calculate the AUC of the selected model in (d).

```
auc(Default$`default payment next month`, as.numeric(LDA.Default.pred$class))

## Setting levels: control = 0, case = 1

## Setting direction: controls < cases

## Area under the curve: 0.6429
```

(f) Calculate the AUC of the model in (d) by 5-fold cross-validation.

```
cv.glm(Default$`default payment next month`,LDA.Default, k=5)
```

(g) Find a better prediction model by using any other variables, transforming variables, interaction of variables, etc. However, you can only use logistic regression and LDA. Report the final model and its 5-fold cross-validated AUC. (The best three students will get bonus credits. You do not have to describe how to find the model, but include all transformation, etc. for reproducibility.)

```
LDA.Default2 <- lda(`default payment next month`~PAY_0+ PAY_2+ PAY_3+ PAY_4+
PAY_5+ PAY_6, data=Default)

cv.glm(Default$`default payment next month`,LDA.Default2, k=5)
```