Problem 1(12.3)

(a)

```
> text12.3$failure<-ifelse(text12.3$Damaged==0,0,1)
> View(text12.3)
> fit.12.3 <- glm(failure~Temperature,binomial,text12.3)
> summary(fit.12.3)

Call:
glm(formula = failure ~ Temperature, family = binomial, data = text12.3)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.0611  -0.7613  -0.3783   0.4524   2.2175

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  15.0429     7.3786   2.039   0.0415 *
Temperature  -0.2322     0.1082  -2.145   0.0320 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 28.267  on 22  degrees of freedom
Residual deviance: 20.315  on 21  degrees of freedom
AIC: 24.315

Number of Fisher Scoring iterations: 5
```

I consider damage = 0 to be an O-ring success and damage >0 to be O-ring failure. I create a new variable "failure" that it equals 1 when damage >0 and it equals 0 when damage =0. I run an ordinary logistic regression: glm (failure~ Temperature) and the result is shown above. The coefficient of Temperature is **-0.2322**, which gives the changes in the log odds of P(failure) vs. P(success). Exp(-0.2322) = 0.7928. Hence the **odds of an O-ring Failure vs. Success decrease** by a factor of 0.7928 if Temperature is increased by one unit.

```
> text12.3<-text12.3[-18,]
> View(text12.3)
> fit.12.3.b <- glm(failure~Temperature,binomial,text12.3)
> summary(fit.12.3.b)

Call:
glm(formula = failure ~ Temperature, family = binomial, data = text12.3)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.0034  -0.6085  -0.2056   0.1060   2.0059

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  23.4033    11.8316   1.978   0.0479 *
Temperature  -0.3610     0.1755  -2.057   0.0397 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 25.782  on 21  degrees of freedom
Residual deviance: 14.377  on 20  degrees of freedom
AIC: 18.377

Number of Fisher Scoring iterations: 6
```

(b)

After deleted flight 18, I get the result shown left. The coefficient of Temperature changes to -0.3610. It gives the changes in the log odds of P(failure) vs. P(success). Exp(-0.361) = 0.697. Hence the **odds of an O-ring Failure vs. Success decrease** by a factor of 0.697 if Temperature is increased by one unit.

(c)

```
> predict(fit.12.3.b,data.frame(Temperature=31),type="resp")
       1
0.999995
```

The probability of an O-ring failure is 99.99995% (we can say it is 100%) when temperature at launch was 31 degrees Fahrenheit.

(d)

No, I do not advise the launching on that particular day wince the probability of an O-ring failure is almost 100%.

Problem 2(12.4)

(a)

```
> fit.NFL<-glm( cbind( Success, Failure ) ~ Distance+I(Distance^2), data = NFL, family = "binomial" )
> summary(fit.NFL)

Call:
glm(formula = cbind(Success, Failure) ~ Distance + I(Distance^2),
    family = "binomial", data = NFL)

Deviance Residuals:
       1         2         3         4         5
 0.11628  -0.00048  -0.40173   0.64209  -0.91465

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)    2.490203   1.018620   2.445   0.0145 *
Distance      -0.013167   0.065990  -0.200   0.8419
I(Distance^2) -0.001513   0.001008  -1.500   0.1335
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 147.7816  on 4  degrees of freedom
Residual deviance:   1.4238  on 2  degrees of freedom
AIC: 28.89

Number of Fisher Scoring iterations: 4
```

```
> fit.AFL<-glm( cbind( Success, Failure ) ~ Distance+I(Distance^2), data = AFL, family = "binomial" )
> summary(fit.AFL)

Call:
glm(formula = cbind(Success, Failure) ~ Distance + I(Distance^2),
    family = "binomial", data = AFL)

Deviance Residuals:
      6        7        8        9       10
 0.3187  -0.6829   0.7721  -0.5231   0.2853

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)    4.892466   1.189274   4.114 3.89e-05 ***
Distance      -0.197046   0.074348  -2.650  0.00804 **
I(Distance^2)  0.001604   0.001098   1.461  0.14395
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 78.7794  on 4  degrees of freedom
Residual deviance:  1.5192  on 2  degrees of freedom
AIC: 28.443

Number of Fisher Scoring iterations: 3
```

(b)

```
> fit.12.4.b<-glm( cbind( Success, Failure ) ~ Distance+I(Distance^2)+Z, data = text12.4, family = "binomia
l" )
> summary(fit.12.4.b)

Call:
glm(formula = cbind(Success, Failure) ~ Distance + I(Distance^2) +
    Z, family = "binomial", data = text12.4)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.86350  -0.20086   0.03301   0.55505   1.60112

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)    3.5241844  0.7747832   4.549  5.4e-06 ***
Distance      -0.0958710  0.0490210  -1.956   0.0505 .
I(Distance^2) -0.0001086  0.0007365  -0.147   0.8828
Z              0.1037533  0.1698311   0.611   0.5413
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 228.5180  on 9  degrees of freedom
Residual deviance:   8.9776  on 6  degrees of freedom
AIC: 59.367

Number of Fisher Scoring iterations: 4
```

(c)

The quadratic term is **not significant** in all models above.

(d)

From the model in (b), added Z =1 if league = NFL & Z=0 if league = 0 in the model, the p-value of Z is 0.5413 which is larger than 0.05. So **Z is not significant** and should not be considered in the model. Therefore, the probabilities of scoring field goals from a given distance the same for each league.

Problem 3(12.5)

(a)

RURAL is the response variable for the simple logistic regression. First, I consider all other variables into the model as predictive variables. Summary shows below. Only NSAL is significant under 95% CI.

```
Call:
glm(formula = RURAL ~ BED + MCDAYS + TDAYS + PCREV + NSAL, family = binomial,
    data = text12.5)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0666  -0.7712   0.4921   0.6825   1.4456

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.432e+00  1.259e+00   2.725  0.00643 **
BED         -1.749e-02  2.264e-02  -0.773  0.43978
MCDAYS       1.538e-02  8.689e-03   1.770  0.07678 .
TDAYS       -1.001e-02  8.976e-03  -1.115  0.26480
PCREV        6.917e-05  1.266e-04   0.546  0.58496
NSAL        -5.426e-04  2.759e-04  -1.967  0.04921 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 67.083  on 51  degrees of freedom
Residual deviance: 50.155  on 46  degrees of freedom
AIC: 62.155
```

Then I do backward selection to find a better model. Summary shows below.

It give a model :RURAL ~ BED+MCDAYS+NSAL+FEXP. However, only NSAL is significant in the model.

```
Step:  AIC=59.36
RURAL ~ BED + MCDAYS + NSAL + FEXP

         Df Deviance    AIC
<none>        49.358 59.358
- FEXP    1   51.436 59.436
- BED     1   52.911 60.911
- MCDAYS  1   53.466 61.466
- NSAL    1   56.988 64.988
```

```
Call:
glm(formula = RURAL ~ BED + MCDAYS + NSAL + FEXP, family = binomial,
    data = text12.5)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9991  -0.5890   0.4532   0.7337   1.4386

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.6442709  1.3127936   2.776   0.0055 **
BED         -0.0366403  0.0224695  -1.631   0.1030
MCDAYS       0.0126199  0.0070877   1.781   0.0750 .
NSAL        -0.0007526  0.0003165  -2.378   0.0174 *
FEXP         0.0003439  0.0002539   1.355   0.1755
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 67.083  on 51  degrees of freedom
Residual deviance: 49.358  on 47  degrees of freedom
AIC: 59.358
```

I exclude non-significant predictive variables from the model; finally, I got the **best fitting model: RURAL ~ NSAL** (shows below).

```
Call:
glm(formula = RURAL ~ NSAL, family = binomial, data = text12.5)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0661  -0.8326   0.5184   0.8419   1.4986

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.3126144  0.9695332   3.417 0.000634 ***
NSAL        -0.0006671  0.0002203  -3.028 0.002463 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 67.083  on 51  degrees of freedom
Residual deviance: 55.424  on 50  degrees of freedom
AIC: 59.424

Number of Fisher Scoring iterations: 4
```

It tells us only difference between rural facilities and non-rural facilities is **Annual nursing salaries**.

(b)

Variables which are relative to hospital characteristics are : RURAL+BED + MCDAYS + TDAYS + NSAL +FEXP. I include all of them into a multiple linear regression with the response variable PCREV. But most of these predictors are non significant. (left graph)

```
Call:
lm(formula = PCREV ~ factor(RURAL) + BED + MCDAYS + TDAYS + NSAL +
    FEXP, data = text12.5)

Residuals:
     Min       1Q   Median       3Q      Max
-11886.0   -547.4    138.3   1179.3   7554.0

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)    -2839.5853  1542.1909  -1.841 0.072180 .
factor(RURAL)1   343.7008   940.9420   0.365 0.716619
BED               43.7229    18.6096   2.349 0.023248 *
MCDAYS             3.2157     8.6774   0.371 0.712683
TDAYS             33.3823     9.1058   3.666 0.000648 ***
NSAL               0.5374     0.3246   1.656 0.104690
FEXP               0.2647     0.2440   1.085 0.283719
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2699 on 45 degrees of freedom
Multiple R-squared:  0.8678,    Adjusted R-squared:  0.8502
F-statistic: 49.25 on 6 and 45 DF,  p-value: < 2.2e-16
```

```
Call:
lm(formula = PCREV ~ BED + TDAYS + NSAL, data = text12.5)

Residuals:
     Min       1Q   Median       3Q      Max
-11879.3   -706.2    -26.6   1174.4   7192.8

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -2187.5906  1072.5548  -2.040   0.0469 *
BED            48.0237    16.1960   2.965   0.0047 **
TDAYS          34.8066     5.4979   6.331 7.81e-08 ***
NSAL            0.5683     0.2674   2.125   0.0388 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2673 on 48 degrees of freedom
Multiple R-squared:  0.8618,    Adjusted R-squared:  0.8531
F-statistic: 99.75 on 3 and 48 DF,  p-value: < 2.2e-16
```

Then, I use backward selection and it gives the model: lm(formula = PCREV ~ BED + TDAYS + NSAL, data = text12.5) (right graph) BED, TDAYS and NSAL are predictors in the model and all of them are significant. Adjusted-R^2 is 85.31%. So lm(**PCREV ~ BED + TDAYS + NSAL**) is the **best model**. Therefore, number of beds in home, annual total patient days and annual nursing salaries affect the annual total patience care revenue.

Problem 4(12.6)

(a)

```
Call:
mlogit(formula = CC ~ 0 | IR + SSPG, data = diab, reflevel = "3",
    method = "nr", print.level = 0)

Frequencies of alternatives:
      3       1       2
0.52414 0.22759 0.24828

nr method
7 iterations, 0h:0m:0s
g'(-H)^-1g = 6.13E-05
successive function values within tolerance limits

Coefficients :
                Estimate Std. Error t-value  Pr(>|t|)
1:(intercept) -7.1106613  1.6882293 -4.2119 2.532e-05 ***
2:(intercept) -4.5484979  0.7714721 -5.8959 3.727e-09 ***
1:IR          -0.0134273  0.0046513 -2.8868  0.003892 **
2:IR           0.0032576  0.0022923  1.4211  0.155289
1:SSPG         0.0425948  0.0079735  5.3420 9.191e-08 ***
2:SSPG         0.0195105  0.0044519  4.3825 1.173e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Log-Likelihood: -72.029
McFadden R^2:  0.51364
Likelihood ratio test : chisq = 152.14 (p.value = < 2.22e-16)

Call:
mlogit(formula = CC ~ 0 | IR + SSPG + RW, data = diab, reflevel = "3",
    method = "nr", print.level = 0)

Frequencies of alternatives:
      3       1       2
0.52414 0.22759 0.24828

nr method
7 iterations, 0h:0m:0s
g'(-H)^-1g = 0.00028
successive function values within tolerance limits

Coefficients :
                Estimate Std. Error t-value  Pr(>|t|)
1:(intercept) -1.8446132  3.4634601 -0.5326  0.594316
2:(intercept) -7.6154166  2.3356317 -3.2605  0.001112 **
1:IR          -0.0133537  0.0050193 -2.6605  0.007804 **
2:IR           0.0035868  0.0023492  1.5268  0.126803
1:SSPG         0.0455039  0.0092415  4.9239 8.486e-07 ***
2:SSPG         0.0164141  0.0049819  3.2948  0.000985 ***
1:RW          -5.8674627  3.8665785 -1.5175  0.129145
2:RW           3.4727694  2.4461624  1.4197  0.155701
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Log-Likelihood: -68.415
McFadden R^2:  0.53805
Likelihood ratio test : chisq = 159.37 (p.value = < 2.22e-16)
```

```
> ctable;
     Y.hat
       1   2   3 Sum
  1   27   3   3  33
  2    1  22  13  36
  3    2   5  69  76
  Sum 30  30  85 145
> sum(diag(ctable)[-4])/diag(ctable)[4]
      Sum
0.8137931
```

```
> ctable;
     Y.hat
       1   2   3 Sum
  1   27   3   3  33
  2    0  24  12  36
  3    2   5  69  76
  Sum 29  32  84 145
> sum(diag(ctable)[-4])/diag(ctable)[4]
      Sum
0.8275862
```

Before adding RW into the multinomial logistic model (just using IR + SSPG), the result shows left. I can get the classification rate for the model is 81.38%.
After adding RW into the model, the result shows right. I can get the classification rate for the model is 82.75%.
The **increase of classification rate is very small**, just 1.37%. Therefore, **RW does not** result in a substantial improvement in the classification rate.

(b)

```
formula: CC.ordered ~ IR + SSPG
data:    diabetes

 link  threshold nobs logLik AIC    niter max.grad cond.H
 logit flexible  145  -81.75 171.50 6(0)  2.93e-12 2.6e+06

Coefficients:
      Estimate Std. Error z value Pr(>|z|)
IR    0.004058   0.001745   2.326     0.02 *
SSPG -0.028142   0.003592  -7.835   4.7e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Threshold coefficients:
    Estimate Std. Error z value
1|2  -6.7944     0.8569  -7.929
2|3  -4.1893     0.6622  -6.326

> ctable2;
     Y.hat2
         1   2   3 Sum
  1     26   5   2  33
  2      3  20  13  36
  3      0   8  68  76
  Sum   29  33  83 145
> sum(diag(ctable2)[-4])/diag(ctable2)[4]
      Sum
0.7862069
```

```
formula: CC.ordered ~ IR + SSPG + RW
data:    diabetes

 link  threshold nobs logLik AIC    niter max.grad cond.H
 logit flexible  145  -81.21 172.42 6(0)  1.53e-12 2.1e+07

Coefficients:
      Estimate Std. Error z value Pr(>|z|)
IR    0.003765   0.001782   2.113   0.0346 *
SSPG -0.029275   0.003826  -7.651 1.99e-14 ***
RW    1.920021   1.861554   1.031   0.3023
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Threshold coefficients:
    Estimate Std. Error z value
1|2   -5.155      1.773  -2.907
2|3   -2.527      1.715  -1.473

> ctable2;
     Y.hat2
         1   2   3 Sum
  1     26   5   2  33
  2      2  21  13  36
  3      0   9  67  76
  Sum   28  35  82 145
> sum(diag(ctable2)[-4])/diag(ctable2)[4]
      Sum
0.7862069
```

Before adding RW into the ordinal logistic model (just using IR + SSPG), the result shows left. I can get the classification rate for the model is 78.62%.
After adding RW into the model, the result shows right. I can get the classification rate for the model is 78.62%.
The **increase of classification rate is ZERO**. Therefore, **RW does not** result in a substantial improvement in the classification rate from s model using only IR and SSPG.

```
> anova(fit.12.6,fit.12.6.b)
Likelihood ratio tests of cumulative link models:

           formula:                     link: threshold:
fit.12.6   CC.ordered ~ IR + SSPG      logit flexible
fit.12.6.b CC.ordered ~ IR + SSPG + RW logit flexible

           no.par    AIC  logLik LR.stat df Pr(>Chisq)
fit.12.6        4 171.50 -81.749
fit.12.6.b      5 172.42 -81.211  1.0761  1     0.2996
```

Comparing two models by using Anova, Pr(>Chisq) = 0.2996 is larger than 0.05 for fit.12.6.b which is included RW. Therefore, the model with RW is not good to use. So there is not a substantial improvement in fit by adding RW in the model.

Problem 5

(a)

```
library(mlogit)
prob5.train = mlogit.data(data = train, choice="ME", shape="wide",varying=NULL);
fit.prob5.a = mlogit(ME~0|HIST+PB, data = prob5.train, reflevel="2");
summary(fit.prob5.a);

prob5.test = mlogit.data(data = test, choice="ME", shape="wide",varying=NULL);
Y.prob= predict(fit.prob5.a,prob5.test,type="resp")
head(Y.prob);

# classify to the category for which it has the highest estimated probabilities
n = dim(train)[1];
Y.hat = rep(0,n);
for(i in 1:n){
  if(max(Y.prob[i,]) == Y.prob[i,1]){
    Y.hat[i]=2;
  }else if(max(Y.prob[i,]) == Y.prob[i,2]){
    Y.hat[i]=0;
  }else if(max(Y.prob[i,]) == Y.prob[i,3]){
    Y.hat[i]=1;
  }
}
Y.hat;

ctable = table(test$ME, Y.hat);
ctable = addmargins(ctable);
ctable;
```

```
> ctable;
     Y.hat
        0   1 Sum
  0   106   4 110
  1    48   7  55
  2    37   4  41
  Sum 191  15 206
> 1-(106+7+0)/(206) #misclassification rate
[1] 0.4514563
```

The total misclassification rate is 45.145%.

The misclassification rate table for each category is shown:

0	1	2
3.636364%	87.27%	100%

(b)

```
> library(ordinal)
> train$ME.ordered = factor(train$ME,levels=c(0,2,1));
> pro5.b <- clm(ME.ordered~HIST+PB, data = train)
> summary(pro5.b)
formula: ME.ordered ~ HIST + PB
data:    train

 link  threshold nobs logLik  AIC    niter max.grad cond.H
 logit flexible  206  -181.61 371.23 6(0)  9.90e-14 1.7e+03

Coefficients:
     Estimate Std. Error z value Pr(>|z|)
HIST  1.37243    0.42024   3.266 0.001091 **
PB   -0.26263    0.07729  -3.398 0.000678 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Threshold coefficients:
    Estimate Std. Error z value
0|2  -1.3439     0.5746  -2.339
2|1  -0.5102     0.5699  -0.895
>
> Y.hat2 = predict(pro5.b, newdata = test, type="class")$fit;
> ctable2 = table(test$ME, Y.hat2);
> ctable2 = addmargins(ctable2);

> ctable2;
     Y.hat2
        0   2   1 Sum
  0   106   0   4 110
  1    48   0   7  55
  2    37   0   4  41
  Sum 191   0  15 206
> 1-(106+7+0)/206 #misclassification rate
[1] 0.4514563
```

The total misclassification rate is 45.145%. We don’t get better prediction since the same misclassification rate.