Unit 3 - Predicting the Baseball World Series Champion

```
baseball <- read.csv("baseball.csv")
```

Problem 1.1

```
nrow(baseball)
```

```
## [1] 1232
```

Problem 1.2

```
length(table(baseball$Year))
```

```
## [1] 47
```

Problem 1.3

```
baseball <- subset(baseball, Playoffs == 1)
nrow(baseball)
```

```
## [1] 244
```

Problem 1.4

```
unique(table(baseball$Year))
```

```
## [1]  2  4  8 10
```

Problem 2.1

```
PlayoffTable <- table(baseball$Year)
names(PlayoffTable)
```

```
##  [1] "1962" "1963" "1964" "1965" "1966" "1967" "1968" "1969" "1970" "1971"
## [11] "1973" "1974" "1975" "1976" "1977" "1978" "1979" "1980" "1982" "1983"
## [21] "1984" "1985" "1986" "1987" "1988" "1989" "1990" "1991" "1992" "1993"
## [31] "1996" "1997" "1998" "1999" "2000" "2001" "2002" "2003" "2004" "2005"
## [41] "2006" "2007" "2008" "2009" "2010" "2011" "2012"
```

Vector of years stored as strings (type chr)

Problem 2.2

PlayoffTable[c(“1990”, “2001”)]

Problem 2.3

```
baseball$NumCompetitors = PlayoffTable[as.character(baseball$Year)]
```

Problem 2.4

```
baseball$NumCompetitors <- PlayoffTable[as.character(baseball$Year)]
nrow(subset(baseball, NumCompetitors == 8))
```

```
## [1] 128
```

Problem 3.1

```
baseball$WorldSeries = as.numeric(baseball$RankPlayoffs == 1)
table(baseball$WorldSeries)
```

```
##
##   0   1
## 197  47
```

197

Problem 3.2

```
model1 <- glm(WorldSeries ~ Year, baseball, family = "binomial")
summary(model1)
```

```
##
## Call:
## glm(formula = WorldSeries ~ Year, family = "binomial", data = baseball)
##
```

```
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.0297  -0.6797  -0.5435  -0.4648   2.1504
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) 72.23602   22.64409    3.19  0.00142 **
## Year        -0.03700    0.01138   -3.25  0.00115 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 239.12  on 243  degrees of freedom
## Residual deviance: 228.35  on 242  degrees of freedom
## AIC: 232.35
##
## Number of Fisher Scoring iterations: 4
```

```
model2 <- glm(WorldSeries ~ RA, baseball, family = "binomial")
summary(model2)
```

```
##
## Call:
## glm(formula = WorldSeries ~ RA, family = "binomial", data = baseball)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.9749  -0.6883  -0.6118  -0.4746   2.1577
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.888174   1.483831   1.272   0.2032
## RA          -0.005053   0.002273  -2.223   0.0262 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 239.12  on 243  degrees of freedom
## Residual deviance: 233.88  on 242  degrees of freedom
## AIC: 237.88
##
## Number of Fisher Scoring iterations: 4
```

```
model3 <- glm(WorldSeries ~ RankSeason, baseball, family = "binomial")
summary(model3)
```

```
##
## Call:
## glm(formula = WorldSeries ~ RankSeason, family = "binomial",
##     data = baseball)
##
```

```
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.7805  -0.7131  -0.5918  -0.4882   2.1781
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -0.8256     0.3268  -2.527   0.0115 *
## RankSeason   -0.2069     0.1027  -2.016   0.0438 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 239.12  on 243  degrees of freedom
## Residual deviance: 234.75  on 242  degrees of freedom
## AIC: 238.75
##
## Number of Fisher Scoring iterations: 4
```

```
model4 <- glm(WorldSeries ~ NumCompetitors, baseball, family = "binomial")
summary(model4)
```

```
##
## Call:
## glm(formula = WorldSeries ~ NumCompetitors, family = "binomial",
##     data = baseball)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.9871  -0.8017  -0.5089  -0.5089   2.2643
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)     0.03868    0.43750   0.088 0.929559
## NumCompetitors -0.25220    0.07422  -3.398 0.000678 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 239.12  on 243  degrees of freedom
## Residual deviance: 226.96  on 242  degrees of freedom
## AIC: 230.96
##
## Number of Fisher Scoring iterations: 4
```

Problem 4.1

```
model <- glm(WorldSeries ~ Year + RA + RankSeason + NumCompetitors,
             baseball, family = "binomial")
summary(model)
```

```
##
## Call:
## glm(formula = WorldSeries ~ Year + RA + RankSeason + NumCompetitors,
##     family = "binomial", data = baseball)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.0336  -0.7689  -0.5139  -0.4583   2.2195
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)    12.5874376 53.6474210   0.235    0.814
## Year           -0.0061425  0.0274665  -0.224    0.823
## RA             -0.0008238  0.0027391  -0.301    0.764
## RankSeason     -0.0685046  0.1203459  -0.569    0.569
## NumCompetitors -0.1794264  0.1815933  -0.988    0.323
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 239.12  on 243  degrees of freedom
## Residual deviance: 226.37  on 239  degrees of freedom
## AIC: 236.37
##
## Number of Fisher Scoring iterations: 4
```

Problem 4.2

```
cor(baseball[c("Year", "RA", "RankSeason", "NumCompetitors")])
```

```
##                     Year        RA RankSeason NumCompetitors
## Year           1.0000000 0.4762422  0.3852191      0.9139548
## RA             0.4762422 1.0000000  0.3991413      0.5136769
## RankSeason     0.3852191 0.3991413  1.0000000      0.4247393
## NumCompetitors 0.9139548 0.5136769  0.4247393      1.0000000
```

Year/NumCompetitors

Problem 4.3

```
model <- glm(WorldSeries ~ Year + NumCompetitors, baseball, family = "binomial")
summary(model)
```

```
##
## Call:
## glm(formula = WorldSeries ~ Year + NumCompetitors, family = "binomial",
##     data = baseball)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.0050  -0.7823  -0.5115  -0.4970   2.2552
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)    13.350467  53.481896   0.250    0.803
## Year           -0.006802   0.027328  -0.249    0.803
## NumCompetitors -0.212610   0.175520  -1.211    0.226
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 239.12  on 243  degrees of freedom
## Residual deviance: 226.90  on 241  degrees of freedom
## AIC: 232.9
##
## Number of Fisher Scoring iterations: 4
```

NumCompetitors