Final Project

AMS 573

Su-Ah Kwon

Table of Content

01.

Employee Attrition

Research Problem

- What are the conditionally dependent explanatory variables for the probability of employee attrition adjusted for all other variables that gives a good fit to the model?
- What affects the probability of employee attrition the most and the least among the dependent variables?
- Which dependent variable has a linear trend against employee attrition?
- What is the association between years with current manager, performance rating, and years in current role?

Purpose of Analysis

- What are the conditionally dependent explanatory variables for the probability of employee attrition adjusted for all other variables that gives a good fit to the model?

To find a model with a good fit that explains the reason for the probability of employee attrition

- What affects the probability of employee attrition the most and the least among the dependent variables?

For a manager in a company to effectively focus on eliminating the biggest cause of employee attrition with given resources

- Which dependent variable has a linear trend against employee attrition?

To better understand the relationship between the explanatory variables and response

- What is the association between years with current manager, performance rating, years in current role?

To see how the three variables affect each other and find the cause for high performance rating

Explanation

- The dataset is a fictional data set created by IBM data scientists to find components that impact employee attrition. There are 1470 employees (or observations) and 25 variables.

```
      Age        Attrition           BusinessTravel                 Department  DistanceFromHome    Education             EducationField EnvironmentSatisfaction    Gender
 Min.   :18.00   No :1233   Non-Travel       : 150   Human Resources       : 63   Min.   : 1.000   Min.   :1.000   Human Resources : 27    Min.   :1.000           Female:588
 1st Qu.:30.00   Yes: 237   Travel_Frequently: 277   Research & Development:961   1st Qu.: 2.000   1st Qu.:2.000   Life Sciences   :606    1st Qu.:2.000           Male  :882
 Median :36.00              Travel_Rarely    :1043   Sales                 :446   Median : 7.000   Median :3.000   Marketing       :159    Median :3.000
 Mean   :36.92                                                                  Mean   : 9.193   Mean   :2.913   Medical         :464    Mean   :2.722
 3rd Qu.:43.00                                                                  3rd Qu.:14.000   3rd Qu.:4.000   Other           : 82    3rd Qu.:4.000
 Max.   :60.00                                                                  Max.   :29.000   Max.   :5.000   Technical Degree:132    Max.   :4.000

 JobInvolvement                    JobRole    JobSatisfaction  MaritalStatus MonthlyIncome   NumCompaniesWorked Over18   OverTime   PerformanceRating RelationshipSatisfaction
 Min.   :1.00   Sales Executive          :326   Min.   :1.000   Divorced:327   Min.   : 1009   Min.   :0.000      Y:1470   No :1054   Min.   :3.000     Min.   :1.000
 1st Qu.:2.00   Research Scientist       :292   1st Qu.:2.000   Married :673   1st Qu.: 2911   1st Qu.:1.000               Yes: 416   1st Qu.:3.000     1st Qu.:2.000
 Median :3.00   Laboratory Technician    :259   Median :3.000   Single  :470   Median : 4919   Median :2.000                          Median :3.000     Median :3.000
 Mean   :2.73   Manufacturing Director   :145   Mean   :2.729                  Mean   : 6503   Mean   :2.693                          Mean   :3.154     Mean   :2.712
 3rd Qu.:3.00   Healthcare Representative:131   3rd Qu.:4.000                  3rd Qu.: 8379   3rd Qu.:4.000                          3rd Qu.:3.000     3rd Qu.:4.000
 Max.   :4.00   Manager                  :102   Max.   :4.000                  Max.   :19999   Max.   :9.000                          Max.   :4.000     Max.   :4.000
                (Other)                  :215
 TotalWorkingYears WorkLifeBalance YearsAtCompany   YearsInCurrentRole YearsSinceLastPromotion YearsWithCurrManager
 Min.   : 0.00     Min.   :1.000   Min.   : 0.000   Min.   : 0.000     Min.   : 0.000          Min.   : 0.000
 1st Qu.: 6.00     1st Qu.:2.000   1st Qu.: 3.000   1st Qu.: 2.000     1st Qu.: 0.000          1st Qu.: 2.000
 Median :10.00     Median :3.000   Median : 5.000   Median : 3.000     Median : 1.000          Median : 3.000
 Mean   :11.28     Mean   :2.761   Mean   : 7.008   Mean   : 4.229     Mean   : 2.188          Mean   : 4.123
 3rd Qu.:15.00     3rd Qu.:3.000   3rd Qu.: 9.000   3rd Qu.: 7.000     3rd Qu.: 3.000          3rd Qu.: 7.000
 Max.   :40.00     Max.   :4.000   Max.   :40.000   Max.   :18.000     Max.   :15.000          Max.   :17.000
```

Explanation

- Ordinal categorical variables are already coded with numeric values given below.

- **Education** 1 ‘Below College’ 2 ‘College’ 3 ‘Bachelor’ 4 ‘Master’ 5 ‘Doctor’
- **EnvironmentSatisfaction** 1 ‘Low’ 2 ‘Medium’ 3 ‘High’ 4 ‘Very High’
- **JobInvolvement** 1 ‘Low’ 2 ‘Medium’ 3 ‘High’ 4 ‘Very High’
- **JobSatisfaction** 1 ‘Low’ 2 ‘Medium’ 3 ‘High’ 4 ‘Very High’
- **PerformanceRating** 1 ‘Low’ 2 ‘Good’ 3 ‘Excellent’ 4 ‘Outstanding’
- **RelationshipSatisfaction** 1 ‘Low’ 2 ‘Medium’ 3 ‘High’ 4 ‘Very High’
- **WorkLifeBalance** 1 ‘Bad’ 2 ‘Good’ 3 ‘Better’ 4 ‘Best’

Logistic Regression

- Consider that we have 1470 independent binary data, a binomial random component regarding whether there is employee attrition or not.
- We will construct a logistic regression model using logit link function.
- First, the 5 explanatory variables, **Job Satisfaction, Marital Status, Monthly Income, Relationship Satisfaction,** and **Work Life Balance** are assumed to be the dependent variables on the employee attrition.

Condition for Goodness of Fit Test

- To perform the goodness of fit test, we will discretize the continuous variable, **Monthly Income**.
- MonthlyIncome=cut(MonthlyIncome, **breaks=5**, include.lowest = TRUE, **labels = c(1,2,3,4,5)**)
- **Yes=1 and No=0** is encoded in Attrition.
- Each fitted cell in the contingency table should be bigger than 5.

```
fit1=glm(Attrition~factor(JobSatisfaction)+MonthlyIncome+MaritalStatus+factor(RelationshipSatisfaction)+factor(WorkLifeBalance),family=binomial,data=dat1)
n=nrow(dat1)
fit.yes=n*fitted(fit1);fit.no=n*(1-fitted(fit1))
sum(fit.yes<5);sum(fit.no<5)
```

Statistical Analysis

```
Coefficients:
                                Estimate Std. Error z value Pr(>|z|)
(Intercept)                      -0.2344     0.3650  -0.642 0.520719
factor(JobSatisfaction)2         -0.4373     0.2238  -1.954 0.050716 .
factor(JobSatisfaction)3         -0.4414     0.1995  -2.213 0.026921 *
factor(JobSatisfaction)4         -0.9470     0.2128  -4.450 8.59e-06 ***
MonthlyIncome2                   -0.7824     0.1892  -4.136 3.53e-05 ***
MonthlyIncome3                   -0.2768     0.2390  -1.158 0.246919
MonthlyIncome4                   -1.2680     0.4818  -2.632 0.008498 **
MonthlyIncome5                   -1.7978     0.4716  -3.812 0.000138 ***
MaritalStatusMarried              0.2384     0.2230   1.069 0.284973
MaritalStatusSingle               1.1366     0.2187   5.196 2.03e-07 ***
factor(RelationshipSatisfaction)2 -0.3800     0.2326  -1.634 0.102322
factor(RelationshipSatisfaction)3 -0.3439     0.2079  -1.654 0.098083 .
factor(RelationshipSatisfaction)4 -0.4150     0.2116  -1.961 0.049884 *
factor(WorkLifeBalance)2          -0.7007     0.2987  -2.346 0.018983 *
factor(WorkLifeBalance)3          -1.0122     0.2769  -3.656 0.000256 ***
factor(WorkLifeBalance)4          -0.6845     0.3403  -2.012 0.044250 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1298.6  on 1469  degrees of freedom
Residual deviance: 1173.6  on 1454  degrees of freedom
AIC: 1205.6

Number of Fisher Scoring iterations: 5
```

- **Goodness of Fit Test**

1-pchisq(1173.6,1454) -> **1 > 0.05**

- **Likelihood Ratio Test of all coefficients being 0**

1-pchisq(1298.6-1173.6,1469-1454)< **0.05**

-> **TRUE**

- **Likelihood Ratio Test for each coefficient being 0**

```
                                Pr(>Chisq)
factor(RelationshipSatisfaction)     0.2167
```

-> conditionally independent

Statistical Analysis

```
Coefficients:
                                Estimate Std. Error z value Pr(>|z|)
(Intercept)                      -0.2344     0.3650  -0.642 0.520719
factor(JobSatisfaction)2         -0.4373     0.2238  -1.954 0.050716 .
factor(JobSatisfaction)3         -0.4414     0.1995  -2.213 0.026921 *
factor(JobSatisfaction)4         -0.9470     0.2128  -4.450 8.59e-06 ***
MonthlyIncome2                   -0.7824     0.1892  -4.136 3.53e-05 ***
MonthlyIncome3                   -0.2768     0.2390  -1.158 0.246919
MonthlyIncome4                   -1.2680     0.4818  -2.632 0.008498 **
MonthlyIncome5                   -1.7978     0.4716  -3.812 0.000138 ***
MaritalStatusMarried              0.2384     0.2230   1.069 0.284973
MaritalStatusSingle               1.1566     0.2187   5.196 2.03e-07 ***
factor(RelationshipSatisfaction)2 -0.3800     0.2326  -1.634 0.102322
factor(RelationshipSatisfaction)3 -0.3439     0.2079  -1.654 0.098083 .
factor(RelationshipSatisfaction)4 -0.4150     0.2116  -1.961 0.049884 *
factor(WorkLifeBalance)2          -0.7007     0.2987  -2.346 0.018983 *
factor(WorkLifeBalance)3          -1.0122     0.2769  -3.656 0.000256 ***
factor(WorkLifeBalance)4          -0.6845     0.3403  -2.012 0.044250 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1298.6  on 1469  degrees of freedom
Residual deviance: 1173.6  on 1454  degrees of freedom
AIC: 1205.6

Number of Fisher Scoring iterations: 5
```

- The estimated odds ratio of having employee attrition for having high income level is exp(-1.7978)= 0.17 times that for not having high income level.
- The estimated odds ratio of having employee attrition for married people is exp(0.2384)= 1.27 times that for non-married people.
- It is not significant that the coefficient of Marital Status Married is 0 while not true for high monthly income

Stepwise Algorithm: AIC

```
Step:  AIC=1204.03
Attrition ~ factor(JobSatisfaction) + MonthlyIncome + MaritalStatus +
    factor(WorkLifeBalance)

                          Df Deviance    AIC
<none>                         1178.0 1204.0
- factor(WorkLifeBalance)  3   1192.6 1212.6
- factor(JobSatisfaction)  3   1198.5 1218.5
- MonthlyIncome            4   1219.6 1237.6
- MaritalStatus            2   1220.1 1242.1
```

- **Goodness of Fit Test**

  1-pchisq(1178,1457) -> **1** > **0.05**

- **No complete or quasi-complete separation in the data**

- The coefficient or its standard error is not too large.
- The number of Fisher Scoring Iteration is not too large

- **Diagnostic Investigation**

```
sum(abs(rstandard(fit2,type="pearson"))>3)/n
```

-> n is the number of observations

-> 0.03; About **3%** of the cells show lack of fit based on Pearson standardized residual

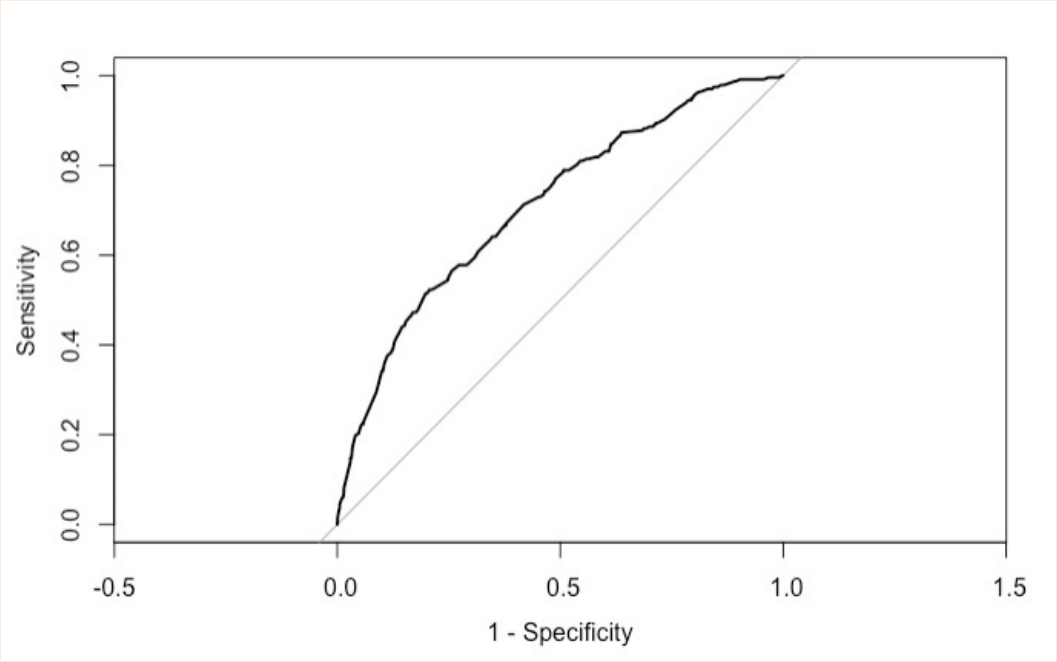
-> This result could have happened by chance

Statistical Analysis

```
Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)               -0.5141     0.3389  -1.517 0.129306
factor(JobSatisfaction)2  -0.4430     0.2233  -1.985 0.047197 *
factor(JobSatisfaction)3  -0.4396     0.1990  -2.209 0.027159 *
factor(JobSatisfaction)4  -0.9474     0.2124  -4.461 8.17e-06 ***
MonthlyIncome2            -0.7790     0.1886  -4.131 3.62e-05 ***
MonthlyIncome3            -0.3017     0.2382  -1.267 0.205319
MonthlyIncome4            -1.2659     0.4804  -2.635 0.008403 **
MonthlyIncome5            -1.8246     0.4715  -3.870 0.000109 ***
MaritalStatusMarried       0.2577     0.2224   1.159 0.246556
MaritalStatusSingle        1.1386     0.2182   5.218 1.81e-07 ***
factor(WorkLifeBalance)2  -0.7313     0.2976  -2.457 0.014011 *
factor(WorkLifeBalance)3  -1.0378     0.2760  -3.761 0.000170 ***
factor(WorkLifeBalance)4  -0.7152     0.3387  -2.111 0.034748 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1298.6  on 1469  degrees of freedom
Residual deviance: 1178.0  on 1457  degrees of freedom
AIC: 1204

Number of Fisher Scoring iterations: 5
```

- The estimated odds ratio of having employee attrition for having high income level is exp (-1.8246)= 0.17 times that for not having high income level.
- The estimated odds ratio of having employee attrition for married people is exp(0.2577)= 1.3 times that for non-married people.
- It is not significant that the coefficient of Marital Status Married is 0 while not true for high monthly income.

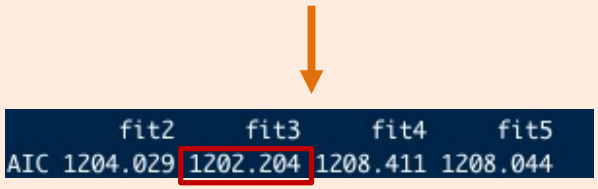
Predictive Power



-> The area under the curve: 0.7117 (better than random guessing)
-> The correlation between the observation and fitted value is 0.300759.

Linear Trend

```
fit2=glm(Attrition~factor(JobSatisfaction)+MonthlyIncome+MaritalStatus+factor(WorkLifeBalance),family=binomial,data=dat1)
# Original
fit3=glm(Attrition~JobSatisfaction+MonthlyIncome+MaritalStatus+factor(WorkLifeBalance),family=binomial,data=dat1)
# Check linear trend for Job Satisfaction
fit4=glm(Attrition~factor(JobSatisfaction)+as.numeric(MonthlyIncome)+MaritalStatus+factor(WorkLifeBalance),family=binomial,data=dat1)
# Check linear trend for Monthly Income
fit5=glm(Attrition~factor(JobSatisfaction)+MonthlyIncome+MaritalStatus+WorkLifeBalance,family=binomial,data=dat1)
# Check linear trend for Work Life Balance
```



- **Conclusion:** Job Satisfaction has a linear trend against the probability of having employee attrition.

Loglinear Regression

- Consider that we have 1470 observations about years with current manager, performance rating, and years in current role. Each count for different combination of explanatory variables are assumed to follow poisson distribution.

- We will construct a loglinear regression model using log link function.

- To avoid sparse contingency table, we discretize the continuous variables, which are years with current manager, and years in current role.

```
dat1$YearsInCurrentRole=cut(dat1$YearsInCurrentRole, breaks = 3,labels = c(1,2,3),include.lowest = T)
dat1$YearsWithCurrManager=cut(dat1$YearsWithCurrManager, breaks = 3,labels = c(1,2,3),include.lowest = T)
```

Statistical Analysis

	YearsInCurrentRole	YearsWithCurrManager	PerformanceRating	Freq
1	1	1	3	739
2	2	1	3	73
3	3	1	3	3
4	1	2	3	89
5	2	2	3	271
6	3	2	3	24
7	1	3	3	10
8	2	3	3	26
9	3	3	3	9
10	1	1	4	130
11	2	1	4	8
12	3	1	4	1
13	1	2	4	16
14	2	2	4	58
15	3	2	4	7
16	1	3	4	1
17	2	3	4	3
18	3	3	4	2

- Condition for Goodness of Fit Test

sum(fitted(fit1)*n<5)=0 -> fitted cells> 5 -> **satisfied**

Statistical Analysis

```
Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)          6.29336    0.03882  162.13   <2e-16 ***
YearsInCurrentRole2 -0.80814    0.05739  -14.08   <2e-16 ***
YearsInCurrentRole3 -3.06400    0.15082  -20.32   <2e-16 ***
YearsWithCurrManager2 -0.71863    0.05656  -12.71   <2e-16 ***
YearsWithCurrManager3 -2.92884    0.14372  -20.38   <2e-16 ***
PerformanceRating4  -1.70555    0.07231  -23.59   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 3589.92  on 17  degrees of freedom
Residual deviance:  760.64  on 12  degrees of freedom
AIC: 856.89
```

- **Fit the Loglinear Model for Independence**
- **Goodness of Fit Test**

1-pchisq(760.64,12) < **0.05**

-> **not well fitted**

-> **The three variables are not mutually independent**

Statistical Analysis

```
Coefficients:
                                          Estimate Std. Error z value Pr(>|z|)
(Intercept)                                6.60921    0.03658 180.674  < 2e-16 ***
YearsInCurrentRole2                       -2.36202    0.11994 -19.693  < 2e-16 ***
YearsInCurrentRole3                       -5.44744    0.50780 -10.728  < 2e-16 ***
YearsWithCurrManager2                     -2.15092    0.10968 -19.611  < 2e-16 ***
YearsWithCurrManager3                     -4.33042    0.30818 -14.052  < 2e-16 ***
PerformanceRating4                        -1.76420    0.09360 -18.849  < 2e-16 ***
YearsInCurrentRole2:YearsWithCurrManager2  3.51735    0.16164  21.760  < 2e-16 ***
YearsInCurrentRole3:YearsWithCurrManager2  4.14736    0.54148   7.659 1.87e-14 ***
YearsInCurrentRole2:YearsWithCurrManager3  3.33976    0.37274   8.960  < 2e-16 ***
YearsInCurrentRole3:YearsWithCurrManager3  5.39608    0.65877   8.191 2.59e-16 ***
YearsInCurrentRole2:PerformanceRating4    -0.07682    0.21657  -0.355    0.723
YearsInCurrentRole3:PerformanceRating4     0.38477    0.40877   0.941    0.347
YearsWithCurrManager2:PerformanceRating4   0.23233    0.21136   1.099    0.272
YearsWithCurrManager3:PerformanceRating4  -0.30337    0.47616  -0.637    0.524
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 3589.9163  on 17  degrees of freedom
Residual deviance:    2.0587  on  4  degrees of freedom
AIC: 114.31

Number of Fisher Scoring iterations: 4
```

- **Fit the Loglinear Model for Homogeneous association**
- **Goodness of Fit Test**

1-pchisq(2.0587,4)>0.05

-> well fitted
-> The three variables have a homogeneous association

Statistical Analysis

```
Coefficients:
                                             Estimate Std. Error z value Pr(>|z|)
(Intercept)                                   6.60921    0.03658 180.674  < 2e-16 ***
YearsInCurrentRole2                          -2.36202    0.11994 -19.693  < 2e-16 ***
YearsInCurrentRole3                          -5.44744    0.50780 -10.728  < 2e-16 ***
YearsWithCurrManager2                        -2.15092    0.10968 -19.611  < 2e-16 ***
YearsWithCurrManager3                        -4.33042    0.30818 -14.052  < 2e-16 ***
PerformanceRating4                           -1.76420    0.09360 -18.849  < 2e-16 ***
YearsInCurrentRole2:YearsWithCurrManager2     3.51735    0.16164  21.760  < 2e-16 ***
YearsInCurrentRole3:YearsWithCurrManager2     4.14736    0.54148   7.659 1.87e-14 ***
YearsInCurrentRole2:YearsWithCurrManager3     3.33976    0.37274   8.960  < 2e-16 ***
YearsInCurrentRole3:YearsWithCurrManager3     5.39608    0.65877   8.191 2.59e-16 ***
YearsInCurrentRole2:PerformanceRating4       -0.07682    0.21657  -0.355    0.723
YearsInCurrentRole3:PerformanceRating4        0.38477    0.40877   0.941    0.347
YearsWithCurrManager2:PerformanceRating4      0.23233    0.21136   1.099    0.272
YearsWithCurrManager3:PerformanceRating4     -0.30337    0.47616  -0.637    0.524
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 3589.9163  on 17  degrees of freedom
Residual deviance:    2.0587  on  4  degrees of freedom
AIC: 114.31

Number of Fisher Scoring iterations: 4
```

- The effect of conditional association between any two variables are the same at each category of the third variable.
- The association term related to performance rating is not significant.

Statistical Analysis

```
Coefficients:
                                            Estimate Std. Error z value Pr(>|z|)
(Intercept)                                  6.60041    0.03570 184.897  < 2e-16 ***
YearsInCurrentRole2                         -2.37289    0.11617 -20.425  < 2e-16 ***
YearsInCurrentRole3                         -5.38105    0.50115 -10.737  < 2e-16 ***
YearsWithCurrManager2                       -2.11338    0.10332 -20.455  < 2e-16 ***
YearsWithCurrManager3                       -4.36945    0.30341 -14.401  < 2e-16 ***
PerformanceRating4                          -1.70555    0.07231 -23.587  < 2e-16 ***
YearsInCurrentRole2:YearsWithCurrManager2   3.51499    0.16143  21.774  < 2e-16 ***
YearsInCurrentRole3:YearsWithCurrManager2   4.16108    0.54123   7.688 1.49e-14 ***
YearsInCurrentRole2:YearsWithCurrManager3   3.34229    0.37268   8.968  < 2e-16 ***
YearsInCurrentRole3:YearsWithCurrManager3   5.38105    0.65800   8.178 2.89e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 3589.9163  on 17  degrees of freedom
Residual deviance:    5.8915  on  8  degrees of freedom
AIC: 110.14

Number of Fisher Scoring iterations: 4
```

- **Goodness of Fit Test**

  1-pchisq(5.8915,8)>0.05

-> well fitted
-> Performance Rating is independent of Years In Current Role and Years with Current Manager
-> Years In Current Role and Years with Current Manager are positively correlated for level 2 and 3.

Statistical Analysis

	YearsInCurrentRole	YearsWithCurrManager	PerformanceRating	fit
1	1	1	3	735.398639
2	2	1	3	68.546939
3	3	1	3	3.385034
4	1	2	3	88.857143
5	2	2	3	278.419048
6	3	2	3	26.234014
7	1	3	3	9.308844
8	2	3	3	24.541497
9	3	3	3	9.308844
10	1	1	4	133.601361
11	2	1	4	12.453061
12	3	1	4	0.614966
13	1	2	4	16.142857
14	2	2	4	50.580952
15	3	2	4	4.765986
16	1	3	4	1.691156
17	2	3	4	4.458503
18	3	3	4	1.691156

Suggestion & Conclusion

- What are the conditionally dependent explanatory variables for the probability of employee attrition adjusted for all other variables that gives a good fit to the model?

Job Satisfaction, Marital Status, Monthly Income, and **Work Life Balance**

- What affects the probability of employee attrition the most and the least among the dependent variables?

Monthly Income (Level 5), and Marital Status(Married)

- Which dependent variable has a linear trend against employee attrition?

Job Satisfaction

- What is the association between years with current manager, performance rating, years in current role?

Independent Performance Rating

Correlated years with current manager and years in current role

02.

Happiness 1972-2006

Research Problem

- Can we fit the given data using only finrela (relative financial status),year,or health?
- What is the association between health, and relative financial status?

Purpose of Analysis

- Can we fit the given data using only finrela (relative financial status), year, or health?

To find a model that fits the data well with only the info of finrela,year, or health

- What is the association between health, and relative financial status?

To see any possible trend between health and relative financial status

Explanation

- The data is a small sample of variables related to happiness from the general social survey (GSS). The GSS is a yearly cross-sectional survey of Americans, run from 1972 to 2006. There are 51,020 observations, and of the over 5,000 variables, nine were selected related to happiness:

```
       id                  happy             year           age            sex               marital                  degree                      finrela             health     
 Min.   :   1   not too happy: 5629   Min.   :1972   Min.   :18.00   male  :22439   married      :27998   lt high school:11777   far below average: 2438   poor     : 2164  
 1st Qu.: 491   pretty happy :25874   1st Qu.:1982   1st Qu.:31.00   female:28581   never married:10064   high school   :26307   below average    :10909   fair     : 7149  
 Median :1002   very happy   :14800   Median :1990   Median :43.00                  divorced     : 6131   junior college: 2601   average          :23363   good     :17227  
 Mean   :1146   NA's         : 4717   Mean   :1990   Mean   :45.43                  widowed      : 5032   bachelor      : 6918   above average    : 8536   excellent:11951  
 3rd Qu.:1504                         3rd Qu.:2000   3rd Qu.:58.00                  separated    : 1781   graduate      : 3253   far above average:  898   NA's     :12529  
 Max.   :4510                         Max.   :2006   Max.   :89.00                  NA's         :   14   NA's          :  164   NA's             : 4876                    
                                                     NA's   :184                                                                                                           
    wtssall      
 Min.   :0.4297  
 1st Qu.:0.5501  
 Median :1.0116  
 Mean   :1.0000  
 3rd Qu.:1.0985  
 Max.   :6.4287  
```

Explanation

- We will deal with independent 34823 obs after eliminating rows with NA values.

- **age** - age in years: 18–89.
- **degree** - highest education: lt high school, high school, junior college, bachelor, graduate.
- **finrela** - relative financial status: far above, above average, average, below average, far below.
- **happy** - happiness: very happy, pretty happy, not too happy.
- **health**- health: excellent, good, fair, poor.
- **marital**- marital status: married, never married, divorced, widowed, separated.
- **sex**- sex: female, male.
- **wtsall**- probability weight. 0.43–6

Logistic Regression

- The response variable is happy with 3 categories:

  very happy, pretty happy, not too happy

- We assume that the response follows a multinomial distribution.
- We will construct **a Baseline-Category Logit Model and Cumulative Logit Model** using logit link function.
- The 3 explanatory variables, **finrela, health, and year** are considered to be the dependent variables for happiness degree.
- To avoid sparse contingency table, we will discretize the continuous variable, **year**.

```
dat2$year=cut(dat2$year,breaks=4, include.lowest = TRUE,labels = c("earlier","early","mid","late"))
```

Statistical Analysis

	fin	year	health	not	pretty	very
1	far below average	earlier	poor	40	32	9
2	below average	earlier	poor	89	131	49
3	average	earlier	poor	77	140	51
4	above average	earlier	poor	10	22	8
5	far above average	earlier	poor	5	4	1
16	far below average	early	poor	40	16	14
17	below average	early	poor	91	80	38
18	average	early	poor	38	122	40
19	above average	early	poor	5	9	5
20	far above average	early	poor	4	0	1
31	far below average	mid	poor	26	22	4
32	below average	mid	poor	49	66	22
33	average	mid	poor	26	63	23
34	above average	mid	poor	5	15	6
35	far above average	mid	poor	1	0	0
46	far below average	late	poor	57	25	10
47	below average	late	poor	52	90	32
48	average	late	poor	46	76	23
49	above average	late	poor	10	14	3
50	far above average	late	poor	3	4	2

- Condition for Goodness of Fit Test

sum(fitted(fit1)*n<5)=0 -> fitted cells> 5 -> **satisfied**

Cumulative Logit Model

```
Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept):1       -0.23110    0.06321  -3.656 0.000256 ***
(Intercept):2        2.68206    0.06557  40.904  < 2e-16 ***
yearearly            0.07888    0.02893   2.727 0.006395 **
yearmid              0.12118    0.03029   4.001 6.31e-05 ***
yearlate             0.05507    0.02871   1.918 0.055087 .
finbelow average    -0.50643    0.05219  -9.704  < 2e-16 ***
finaverage          -1.03858    0.05036 -20.622  < 2e-16 ***
finabove average    -1.24369    0.05454 -22.802  < 2e-16 ***
finfar above average -1.24016    0.09275 -13.372  < 2e-16 ***
healthfair          -0.45652    0.05156  -8.854  < 2e-16 ***
healthgood          -0.95038    0.04872 -19.507  < 2e-16 ***
healthexcellent     -1.59838    0.05049 -31.660  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Names of linear predictors: logitlink(P[Y<=1]), logitlink(P[Y<=2])

Residual deviance: 527.9542 on 148 degrees of freedom
```

- **Goodness of Fit Test**

  1-pchisq(527.9542,148)<0.05

-> TRUE
-> not well fitted

Baseline-Category Logit Model

```
Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept):1         1.647755   0.098049  16.805  < 2e-16 ***
(Intercept):2         1.100880   0.089821  12.256  < 2e-16 ***
yearearly:1           0.005481   0.051064   0.107   0.9145
yearearly:2           0.172796   0.033187   5.207 1.92e-07 ***
yearmid:1             0.071542   0.053810   1.330   0.1837
yearmid:2             0.234761   0.034896   6.728 1.73e-11 ***
yearlate:1           -0.007689   0.050610  -0.152   0.8792
yearlate:2            0.133266   0.032856   4.056 4.99e-05 ***
finbelow average:1   -0.643234   0.079867  -8.054 8.03e-16 ***
finbelow average:2    0.007608   0.071250   0.107   0.9150
finaverage:1         -1.568597   0.077271 -20.300  < 2e-16 ***
finaverage:2         -0.356732   0.067900  -5.254 1.49e-07 ***
finabove average:1   -1.923852   0.089827 -21.417  < 2e-16 ***
finabove average:2   -0.562806   0.071200  -7.905 2.69e-15 ***
finfar above average:1 -1.556327   0.159557  -9.754  < 2e-16 ***
finfar above average:2 -0.736489   0.108691  -6.776 1.24e-11 ***
healthfair:1         -0.481861   0.078582  -6.132 8.68e-10 ***
healthfair:2          0.068022   0.071199   0.955   0.3394
healthgood:1         -1.372106   0.074716 -18.364  < 2e-16 ***
healthgood:2         -0.151807   0.066548  -2.281   0.0225 *
healthexcellent:1    -2.224899   0.079802 -27.880  < 2e-16 ***
healthexcellent:2    -0.851415   0.067184 -12.673  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Names of linear predictors: log(mu[,1]/mu[,3]), log(mu[,2]/mu[,3])

Residual deviance: 207.0248 on 138 degrees of freedom
```

- **Goodness of Fit Test**

  1-pchisq(**207.0248**,138)= 0.0001302404 <0.05

-> TRUE
-> not well fitted
-> However, p-value is larger than that of cumulative logit model
-> The coefficient related to year has the most non-significant p-values.

Likelihood Test

```r
fit.nom2=vglm(as.matrix(BB[,-c(1,2,3)])~1,family=multinomial,data=BB)
fit.nominal=vglm(as.matrix(BB[,-c(1,2,3)])~year+fin+health,family=multinomial,data=BB)
fit3=vglm(as.matrix(BB[,-c(1,2,3)])~year+fin,family=multinomial,data=BB)
fit4=vglm(as.matrix(BB[,-c(1,2,3)])~year+health,family=multinomial,data=BB)
fit5=vglm(as.matrix(BB[,-c(1,2,3)])~health+fin,family=multinomial,data=BB)
lrtest(fit.nominal,fit.nom2) # P-value < 0.05
lrtest(fit.nominal,fit3) # P-value < 0.05
lrtest(fit.nominal,fit4) # P-value < 0.05
lrtest(fit.nominal,fit5) # P-value < 0.05
```

- Year, finrela,and health are dependent variables to the degree of happiness.
- Based on the overall test, at least one of the 3 explanatory has a nonzero coefficient.

Ordinal Variables

- Now, we assume that relative financial status and health have a linear trend to the degree of happiness.
- Each variable will have scores **(1,2,3,4,5)**, and **(1,2,3,4)** respectively.
- We will now compare baseline and cumulative logit models again.
- **Result:** Both models give strong significant p-value.

Ordinal Variables

Cumulative Logit Model

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept):1  3.56660    0.08938  39.905  < 2e-16 ***
(Intercept):2  2.45153    0.06621  37.026  < 2e-16 ***
yearearly:1    0.01648    0.05093   0.324   0.7462
yearearly:2    0.17930    0.03304   5.427 5.73e-08 ***
yearmid:1      0.09412    0.05361   1.756   0.0791 .
yearmid:2      0.25247    0.03469   7.278 3.40e-13 ***
yearlate:1     0.02105    0.05035   0.418   0.6759
yearlate:2     0.15284    0.03264   4.683 2.83e-06 ***
fin:1         -0.68621    0.02370 -28.951  < 2e-16 ***
fin:2         -0.23697    0.01544 -15.345  < 2e-16 ***
health:1      -0.89427    0.02248 -39.777  < 2e-16 ***
health:2      -0.42714    0.01590 -26.863  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Names of linear predictors: log(mu[,1]/mu[,3]), log(mu[,2]/mu[,3])

Residual deviance: 539.5048 on 148 degrees of freedom
```

Baseline-Category Logit Model

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept):1  0.61954    0.05134  12.069  < 2e-16 ***
(Intercept):2  3.52344    0.05543  63.564  < 2e-16 ***
yearearly      0.08734    0.02890   3.022  0.00251 **
yearmid        0.13836    0.03023   4.577 4.71e-06 ***
yearlate       0.07672    0.02862   2.680  0.00735 **
fin           -0.37655    0.01337 -28.172  < 2e-16 ***
health        -0.55740    0.01318 -42.275  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Names of linear predictors: logitlink(P[Y<=1]), logitlink(P[Y<=2])

Residual deviance: 647.8967 on 153 degrees of freedom
```

Ordinal Variables

```
fit8=vglm(as.matrix(BB[,-c(1,2,3)])~year+factor(fin)+health,family=multinomial,data=BB)
fit9=vglm(as.matrix(BB[,-c(1,2,3)])~year+fin+factor(health),family=multinomial,data=BB)
summary(fit8)
summary(fit9)
1-pchisq(408.0058,142)
1-pchisq(334.5304,144)
deviance(fit8) # 408.0048 > 207.0248
deviance(fit9) # 334.5304 > 207.0248
```

```
fit1.int01=vglm(as.matrix(BB[,-c(1,2,3)])~year+factor(fin)+factor(health)+factor(fin):factor(health),family=multinomial,data=BB)
fit1.int02=vglm(as.matrix(BB[,-c(1,2,3)])~year+factor(fin)+factor(health)+factor(fin):year,family=multinomial,data=BB)
summary(fit1.int01)
summary(fit1.int02)
1-pchisq(169.5027,114)
1-pchisq(159.9879,114) # Correlated vs. 0.0001302404
```

- The baseline category logit model with interaction (fit1.int02) only gives bigger p-value than the original one. **(0.002938662)**

Statistical Analysis

- Adding another explanatory variable to **finrela, health, and year** can create a sparse contingency table.
- The coefficient related to year has the most non-significant p-values.
- **Eliminate years and fit with the finrela and happy**.
- As a result, the baseline category logit model produces bigger p-value than the original for the goodness of fit test **0.003494556**.

Statistical Analysis

```
Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept):1          1.65872    0.09393  17.659  < 2e-16 ***
(Intercept):2          1.22305    0.08787  13.920  < 2e-16 ***
finbelow average:1    -0.64318    0.07982  -8.057 7.79e-16 ***
finbelow average:2     0.00139    0.07119   0.020   0.9844
finaverage:1          -1.56861    0.07716 -20.328  < 2e-16 ***
finaverage:2          -0.36979    0.06782  -5.453 4.97e-08 ***
finabove average:1    -1.92382    0.08980 -21.423  < 2e-16 ***
finabove average:2    -0.56963    0.07114  -8.007 1.18e-15 ***
finfar above average:1 -1.55559    0.15953  -9.751  < 2e-16 ***
finfar above average:2 -0.73681    0.10858  -6.786 1.15e-11 ***
healthfair:1          -0.48074    0.07856  -6.119 9.39e-10 ***
healthfair:2           0.07199    0.07115   1.012   0.3116
healthgood:1          -1.36895    0.07461 -18.348  < 2e-16 ***
healthgood:2          -0.13706    0.06646  -2.062   0.0392 *
healthexcellent:1     -2.22143    0.07976 -27.853  < 2e-16 ***
healthexcellent:2     -0.84011    0.06712 -12.517  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Names of linear predictors: log(mu[,1]/mu[,3]), log(mu[,2]/mu[,3])

Residual deviance: 46.8493 on 24 degrees of freedom
```

- The model that treated finrela or health as ordinal is not a good fit.
- The model with interaction term of nominal health and finrela did not fit well, too.
- **Conclusion:** We cannot fit the data with only any combination of year,finrela, and health possibly unless we reduce the # of categories of years, finrela, or health.

Statistical Analysis

```
fit.or2=vglm(as.matrix(CC[,-c(1,2)])~as.numeric(fin)+as.numeric(health),family=cumulative(parallel =T),data=CC) # No Linear Trend
fit.nominal2=vglm(as.matrix(CC[,-c(1,2)])~as.numeric(fin)+as.numeric(health),family=multinomial,data=CC)
summary(fit.or2);summary(fit.nominal2)
1-pchisq(450.2304,36)
1-pchisq(386.4982,34)

fit.or3=vglm(as.matrix(CC[,-c(1,2)])~health+fin+health:fin,family=cumulative(parallel =T),data=CC)
fit.nominal3=vglm(as.matrix(CC[,-c(1,2)])~fin+health+fin:health,family=multinomial,data=CC)
summary(fit.or3);summary(fit.nominal3)
1-pchisq(289.6491,19)
# Inaccurate residual deviance due to convergence at a half-step

fit.or4=vglm(as.matrix(CC[,-c(1,2)])~as.numeric(fin)+health,family=cumulative(parallel =T),data=CC)
fit.nominal4=vglm(as.matrix(CC[,-c(1,2)])~fin+as.numeric(health),family=multinomial,data=CC)
summary(fit.or4);summary(fit.nominal4)
1-pchisq(434.0577,34)
1-pchisq(256.4947,28)

fit10=vglm(as.matrix(CC[,-c(1,2)])~fin,family=multinomial,data=CC)
fit11=vglm(as.matrix(CC[,-c(1,2)])~health,family=multinomial,data=CC)
summary(fit10);summary(fit11)
1-pchisq(1041.41,32)
1-pchisq(1986.135,30)
```

Loglinear Regression

- Consider that we have 34823 observations about health and finrela. Each count for different combinations of the explanatory variables are assumed to follow poisson distribution.

- We will construct a loglinear regression model using log link function.

Loglinear Regression

	health	finrela	Freq
1	poor	far below average	295
2	fair	far below average	469
3	good	far below average	628
4	excellent	far below average	369
5	poor	below average	789
6	fair	below average	2051
7	good	below average	3578
8	excellent	below average	1844
9	poor	average	725
10	fair	average	3215
11	good	average	8352
12	excellent	average	5535
13	poor	above average	112
14	fair	above average	668
15	good	above average	2771
16	excellent	above average	2802
17	poor	far above average	25
18	fair	far above average	72
19	good	far above average	200
20	excellent	far above average	323

- **Fit the Loglinear Model for Independence**
- **Goodness of Fit Test**

  1-pchisq(2007.6,12) < **0.05**

-> not well fitted

-> The three variables are not mutually independent

Loglinear Regression

```
Coefficients:
                                  Estimate Std. Error z value Pr(>|z|)
(Intercept)                       5.264539   0.032751 160.743  < 2e-16 ***
healthfair                        0.338231   0.031759  10.650  < 2e-16 ***
healthgood                        0.272130   0.047062   5.782 7.36e-09 ***
healthexcellent                  -1.099418   0.070305 -15.638  < 2e-16 ***
finrelabelow average              0.638337   0.032605  19.578  < 2e-16 ***
finrelaaverage                    0.410441   0.049700   8.258  < 2e-16 ***
finrelaabove average             -1.696242   0.075128 -22.578  < 2e-16 ***
finrelafar above average         -5.163116   0.110357 -46.786  < 2e-16 ***
as.numeric(health):as.numeric(finrela)  0.339828   0.008225  41.318  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 42474.25  on 19  degrees of freedom
Residual deviance:   170.73  on 11  degrees of freedom
AIC: 357.66
```

- **Fit the Linear-by-Linear model**
- **Goodness of Fit Test**

  1-pchisq(170.73,11) < **0.05**

-> not well fitted

-> There is no linear trend between the health and finrela

Loglinear Regression

```
Coefficients:
                                      Estimate Std. Error z value Pr(>|z|)
(Intercept)                            5.68698    0.05822  97.677  < 2e-16 ***
healthfair                             0.46363    0.07431   6.239 4.40e-10 ***
healthgood                             0.75556    0.07058  10.704  < 2e-16 ***
healthexcellent                        0.22382    0.07810   2.866  0.00416 **
finrelabelow average                   0.98379    0.06824  14.416  < 2e-16 ***
finrelaaverage                         0.89920    0.06906  13.021  < 2e-16 ***
finrelaabove average                  -0.96848    0.11099  -8.726  < 2e-16 ***
finrelafar above average              -2.46810    0.20830 -11.849  < 2e-16 ***
healthfair:finrelabelow average        0.49169    0.08531   5.764 8.22e-09 ***
healthgood:finrelabelow average        0.75623    0.08080   9.359  < 2e-16 ***
healthexcellent:finrelabelow average   0.62510    0.08894   7.029 2.08e-12 ***
healthfair:finrelaaverage              1.02578    0.08493  12.079  < 2e-16 ***
healthgood:finrelaaverage              1.68852    0.08051  20.974  < 2e-16 ***
healthexcellent:finrelaaverage         1.80885    0.08752  20.668  < 2e-16 ***
healthfair:finrelaabove average        1.32216    0.12628  10.470  < 2e-16 ***
healthgood:finrelaabove average        2.45290    0.11946  20.533  < 2e-16 ***
healthexcellent:finrelaabove average   2.99577    0.12404  24.152  < 2e-16 ***
healthfair:finrelafar above average    0.59416    0.24374   2.438  0.01478 *
healthgood:finrelafar above average    1.32388    0.22357   5.922 3.19e-09 ***
healthexcellent:finrelafar above average 2.33496  0.22180  10.527  < 2e-16 ***
```

Conclusion:
The saturated model of the two-way contingency table

Suggestion & Conclusion

- Can we fit the given data using only finrela (relative financial status),year,or health?

We cannot fit the data with only any combination of year,finrela, or health possibly unless we reduce the # of categories of years,finrela or health.

- What is the association between health, and relative financial status?

Health and financial status are correlated and do not show linear trend

Thank you