Categorical+ Continuous(Covariates) Two-way ANCOVA, Bin GLM

SIT Wai Tang

```
rm(list=ls())
library(aod)
library(car)
```

```
## Loading required package: carData
```

```
library(data.table)
library(leaps)
```

Description: train.csv is a subset dataset of the 1912 Titanic passenger survival log. It contains 891 observations and the following 12 columns:

“PassengerId”: passenger ID “Survived”: survival, 0=No, 1=Yes “Pclass”: ticket class, 1=1st, 2=2nd, 3=3rd “Name”: passenger name “Sex”: sex “Age”: age in years “SibSp”: # of siblings/spouses aboard the Titanic “Parch”: # of parents/children aboard the Titanic “Ticket”: ticket number “Fare”: passenger fare “Cabin”: cabin number “Embarked”: Port of Embarkation (C = Cherbourg, Q = Queenstown, S = Southampton)

```
train = read.csv("train.csv")

train$Sex = factor(train$Sex)
train$Survived = factor(train$Survived)
train$Pclass = factor(train$Pclass)
str(train)
```

```
## 'data.frame':    891 obs. of  12 variables:
##  $ PassengerId: int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Survived   : Factor w/ 2 levels "0","1": 1 2 2 2 1 1 1 1 2 2 ...
##  $ Pclass     : Factor w/ 3 levels "1","2","3": 3 1 3 1 3 3 1 3 3 2 ...
##  $ Name       : chr  "Braund, Mr. Owen Harris" "Cumings, Mrs. John Bradley (Florence Briggs Thayer)" "Heikkinen, Miss. La
ina" "Futrelle, Mrs. Jacques Heath (Lily May Peel)" ...
##  $ Sex        : Factor w/ 2 levels "female","male": 2 1 1 1 2 2 2 2 1 1 ...
##  $ Age        : num  22 38 26 35 35 NA 54 2 27 14 ...
##  $ SibSp      : int  1 1 0 1 0 0 0 3 0 1 ...
##  $ Parch      : int  0 0 0 0 0 0 0 1 2 0 ...
##  $ Ticket     : chr  "A/5 21171" "PC 17599" "STON/O2. 3101282" "113803" ...
##  $ Fare       : num  7.25 71.28 7.92 53.1 8.05 ...
##  $ Cabin      : chr  "" "C85" "" "C123" ...
##  $ Embarked   : chr  "S" "C" "S" "S" ...
```

a two-sample z-test

Let p_1, p_2 donate the proportion of surviving female passengers among all females and males represently.

$H_0 : p_1 = p_2 \quad vs \quad H_1 : p_1 \neq p_2$

```
train_female = train[which(train$Sex=="female"), ]
train_male = train[which(train$Sex=="male"), ]
prop.test(c( sum(as.numeric(as.character(train_female$Survived))),
             sum(as.numeric(as.character(train_male$Survived)))),
          c(nrow(train_female), nrow(train_male)))
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  c(sum(as.numeric(as.character(train_female$Survived))), sum(as.numeric(as.character(train_male$Survived)))) out of
c(nrow(train_female), nrow(train_male))
## X-squared = 260.72, df = 1, p-value < 2.2e-16
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  0.4926894 0.6135708
## sample estimates:
##    prop 1    prop 2
## 0.7420382 0.1889081
```

As the p-value < 2.2e-16 is very small, less than 0.05, hence we reject the H_0. We can conclude that the proportion of surviving female passengers and male passengers is the same.

Investigate effects between the survived proportions of female and male:

Fit a logistic regression model which relates the proportion of survival passengers and the variable Sex (binomial)

```
fit = glm(Survived~Sex, data=train, family = "binomial")
summary(fit)
```

```
##
## Call:
## glm(formula = Survived ~ Sex, family = "binomial", data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6462  -0.6471  -0.6471   0.7725   1.8256
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   1.0566     0.1290   8.191 2.58e-16 ***
## Sexmale      -2.5137     0.1672 -15.036  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1186.7  on 890  degrees of freedom
## Residual deviance:  917.8  on 889  degrees of freedom
## AIC: 921.8
##
## Number of Fisher Scoring iterations: 4
```

The fitted logistic regression model:

$logit(p) = 1.0566 + -2.5137 I_{sex=male}$

where p is the probability of a person survived.

Estimate	Std. Error
(Intercept)	1.0566
Sexmale	-2.5137

Test significance of variable Sex

$H_0 : \beta_1 = 0 \quad vs \quad H_1 : \beta_1 \neq 0$

where β_1 denotes the coefficient of $I_{sex=male}$.

```
drop1(fit, test = "LR")
```

```
## Single term deletions
##
## Model:
## Survived ~ Sex
##        Df Deviance    AIC    LRT  Pr(>Chi)
## <none>       917.8  921.8
## Sex     1   1186.7 1188.7 268.85 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
wald.test(b = coef(fit), Sigma = vcov(fit), Terms = 2)
```

```
## Wald test:
## ----------
##
## Chi-squared test:
## X2 = 226.1, df = 1, P(> X2) = 0.0
```

The p-values of both Liklihood-Ratio Test and wald test are < 0.05, indicating that β_1 is significant. Hence, there is a significant difference between the survived proportions of female and male.

Full Logistic regression model

$log(\frac{p}{1-p}) = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{i,j} + (\beta\gamma)_{j,k} + (\alpha\gamma)_{i,k}$ for $i = 1, 2, 3; j = 1, 2; k = 1$ where p is the proportion of surviving passengers, μ_is a constant, $\alpha_i$represents *i*th level of *Pclass*, $\beta_j$represents *j*th level of *Sex*, $\gamma_k$represents the continuous variable *Age*, $(\alpha\beta)_{i,j}$, $(\beta\gamma)_{j,k}$, $(\alpha\gamma)_{i,k}$ represent the two-way interaction effects.

Test adequatcy of Main effects *only* and with *any* Interaction effects in fitted Two-way Logistic regression model

```
logit1 = glm(Survived~ Pclass+Sex+Age , data=train, family="binomial")
anova(logit1, type="II")
```

```
## Warning in anova.glm(logit1, type = "II"): the following arguments to
## 'anova.glm' are invalid and dropped: list(type = "II")
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: Survived
##
## Terms added sequentially (first to last)
##
##
##        Df Deviance Resid. Df Resid. Dev
## NULL                     713     964.52
## Pclass  2   94.706       711     869.81
## Sex     1  197.380       710     672.43
## Age     1   25.148       709     647.28
```

```
summary(logit1)
```

```
##
## Call:
## glm(formula = Survived ~ Pclass + Sex + Age, family = "binomial",
##     data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.7303  -0.6780  -0.3953   0.6485   2.4657
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  3.777013   0.401123   9.416  < 2e-16 ***
## Pclass2     -1.309799   0.278066  -4.710 2.47e-06 ***
## Pclass3     -2.580625   0.281442  -9.169  < 2e-16 ***
## Sexmale     -2.522781   0.207391 -12.164  < 2e-16 ***
## Age         -0.036985   0.007656  -4.831 1.36e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 964.52  on 713  degrees of freedom
## Residual deviance: 647.28  on 709  degrees of freedom
##   (177 observations deleted due to missingness)
## AIC: 657.28
##
## Number of Fisher Scoring iterations: 5
```

```
logit_full = glm(Survived~ Pclass+Sex+Age+Pclass*Sex+Pclass*Age+Sex*Age, data=train, family="binomial")

anova(logit1, logit_full,test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: Survived ~ Pclass + Sex + Age
## Model 2: Survived ~ Pclass + Sex + Age + Pclass * Sex + Pclass * Age +
##     Sex * Age
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1       709     647.28
## 2       704     604.34  5   42.948 3.786e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

H_0 : all coefficients of the interaction term = 0

H_1 : H_0 is not true.

The p-value of the LRT(Liklihood-Ratio Test) is too small, 3.786e-08 < 0.05. Hence the interation terms are significant, indicating that the mean function with *some* Interaction effects is adequate, *only main effects* will be inadequate to explain these data.

Any Interaction effects can be removed?

```
summary(logit_full)
```

```
##
## Call:
## glm(formula = Survived ~ Pclass + Sex + Age + Pclass * Sex +
##     Pclass * Age + Sex * Age, family = "binomial", data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6219  -0.6405  -0.3971   0.3376   3.2713
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)      3.41025    0.96744   3.525 0.000423 ***
## Pclass2          1.24535    1.20069   1.037 0.299644
## Pclass3         -3.17259    0.91924  -3.451 0.000558 ***
## Sexmale         -2.48345    0.91908  -2.702 0.006890 **
## Age             -0.00293    0.02188  -0.134 0.893489
## Pclass2:Sexmale -1.45924    0.89989  -1.622 0.104892
## Pclass3:Sexmale  1.68959    0.73022   2.314 0.020678 *
## Pclass2:Age     -0.06445    0.02621  -2.458 0.013954 *
## Pclass3:Age     -0.01534    0.01972  -0.778 0.436844
## Sexmale:Age     -0.03034    0.01887  -1.608 0.107767
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 964.52  on 713  degrees of freedom
## Residual deviance: 604.34  on 704  degrees of freedom
##   (177 observations deleted due to missingness)
## AIC: 624.34
##
## Number of Fisher Scoring iterations: 6
```

```
anova(logit_full, type="II")
```

```
## Warning in anova.glm(logit_full, type = "II"): the following arguments to
## 'anova.glm' are invalid and dropped: list(type = "II")
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: Survived
##
## Terms added sequentially (first to last)
##
##
##            Df Deviance Resid. Df Resid. Dev
## NULL                         713     964.52
## Pclass      2   94.706       711     869.81
## Sex         1  197.380       710     672.43
## Age         1   25.148       709     647.28
## Pclass:Sex  2   33.856       707     613.43
## Pclass:Age  2    6.502       705     606.92
## Sex:Age     1    2.589       704     604.34
```

From the LRT(Liklihood-Ratio Test), the p-value of Sex:Age is 0.10759 (> 0.05), we don't rejected the null $H_0 : \beta\gamma = 0$. The two-way interaction effect of Sex and Age can be removed.

Test if removing two-way interaction (Sex:Age)

```
# remove Sex:Age and fit the model
logit2 = glm(Survived~ Pclass+Sex+Age + Pclass*Sex +Pclass*Age, data=train, family="binomial")
summary(logit2)
```

```
##
## Call:
## glm(formula = Survived ~ Pclass + Sex + Age + Pclass * Sex +
##     Pclass * Age, family = "binomial", data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.9609  -0.6292  -0.4464   0.3422   3.1508
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)      4.431327   0.802974   5.519 3.42e-08 ***
## Pclass2          1.012013   1.243331   0.814  0.41567
## Pclass3         -3.869847   0.863720  -4.480 7.45e-06 ***
## Sexmale         -3.627526   0.625297  -5.801 6.58e-09 ***
## Age             -0.030192   0.013703  -2.203  0.02757 *
## Pclass2:Sexmale -1.307750   0.908915  -1.439  0.15021
## Pclass3:Sexmale  2.164894   0.681396   3.177  0.00149 **
## Pclass2:Age     -0.057948   0.025526  -2.270  0.02320 *
## Pclass3:Age     -0.003256   0.017934  -0.182  0.85594
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 964.52  on 713  degrees of freedom
## Residual deviance: 606.92  on 705  degrees of freedom
##   (177 observations deleted due to missingness)
## AIC: 624.92
##
## Number of Fisher Scoring iterations: 6
```

```
anova(logit2, type="II")
```

```
## Warning in anova.glm(logit2, type = "II"): the following arguments to
## 'anova.glm' are invalid and dropped: list(type = "II")
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: Survived
##
## Terms added sequentially (first to last)
##
##
##            Df Deviance Resid. Df Resid. Dev
## NULL                         713     964.52
## Pclass      2   94.706       711     869.81
## Sex         1  197.380       710     672.43
## Age         1   25.148       709     647.28
## Pclass:Sex  2   33.856       707     613.43
## Pclass:Age  2    6.502       705     606.92
```

All the remaining two-way interaction terms are significant.(Pclass:Sex , Pclass:Age)

Model reduced: $log(\frac{p}{1-p}) = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{i,j} + (\beta\gamma)_{j,k}$ for $i = 1, 2, 3; j = 1, 2; k = 1$

Difference between the survival female adults and male adults

```
exp(coef(logit2))
```

```
##     (Intercept)         Pclass2         Pclass3         Sexmale             Age
##     84.04286813      2.75113474      0.02086157      0.02658187      0.97025894
## Pclass2:Sexmale Pclass3:Sexmale     Pclass2:Age     Pclass3:Age
##      0.27042789      8.71368169      0.94369863      0.99674948
```

The estimated odds ratio is (female as base level)

$$OddRatio(Male|Female) = e^{\hat{\beta}_1} = e^{-2.48345} = 0.08345462$$

The odds of the survival of the male adults are 91.7% lower than the odds of the survival of the female adults.

Predicte 95% confidence interval probability of survival for 30 years old female in Pclass 1

```
pred1 = predict(logit_full, newdata=data.frame(Age=30,Pclass="1",
                                  Sex="female"), se.fit=T,
               interval="confidence")
pred1
```

```
## $fit
##        1
## 3.322351
##
## $se.fit
## [1] 0.5983816
##
## $residual.scale
## [1] 1
```

Probability in Logistic

```
# predicted probability
list("fit"=plogis(pred1$fit),
"lower"=plogis(pred1$fit-1.96*pred1$se.fit),
"upper"=plogis(pred1$fit+1.96*pred1$se.fit))
```

```
## $fit
##         1
## 0.9651877
##
## $lower
##         1
## 0.8956242
##
## $upper
##         1
## 0.9889606
```

the predicted probability of survival for a female of 30 years old in Pclass 1 is 0.965.

95% confidence interval: (0.8956242, 0.9889606)

A case-control study of esophageal cancer in France. This dataset contains 1175 observations on the following 4 columns: #### AgeGroup: age group (levels: 25-34, 35-44, 45-54, 55-64, 65-74, 75+) #### Alcohol: alcohol consumption (levels: 0-39g/day, 40-79, 80-119, 120+) #### Tobacco: tobacco consumption (levels: 0-9g/day, 10-19, 20-29, 30+) #### Disease: esophageal cancer disease status (levels: 0, 1)

```
# import the dataset
cancer = read.csv("cancer.csv", header=T)
cancer$AgeGroup <- as.factor(cancer$AgeGroup)
cancer$Alcohol <- as.factor(cancer$Alcohol)
cancer$Tobacco <-as.factor(cancer$Tobacco)
head(cancer)
```

```
##   AgeGroup   Alcohol  Tobacco Disease
## 1    25-34 0-39g/day 0-9g/day       0
## 2    25-34 0-39g/day 0-9g/day       0
## 3    25-34 0-39g/day 0-9g/day       0
## 4    25-34 0-39g/day 0-9g/day       0
## 5    25-34 0-39g/day 0-9g/day       0
## 6    25-34 0-39g/day 0-9g/day       0
```

```
str(cancer)
```

```
## 'data.frame':    1175 obs. of  4 variables:
##  $ AgeGroup: Factor w/ 6 levels "25-34","35-44",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ Alcohol : Factor w/ 4 levels "0-39g/day","120+",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ Tobacco : Factor w/ 4 levels "0-9g/day","10-19",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ Disease : int  0 0 0 0 0 0 0 0 0 0 ...
```

Logistic regression model

$log () = μ + α_i + β_j + γ_k + ($\alpha\beta$)_i,_j+ ($\alpha\gamma$)_i,_k + ($\beta\gamma$)_j,_k $ for $i = 1, 2, \ldots, 6; j = 1, 2, 3, 4; k = 1, 2, 3, 4$ where p is the proportion of esophageal cancer patients,
μ_is a constant, $\alpha_i$represents *i*th level of *AgeGroup*, $\beta_j$represents *j*th level of *Alcohol*, $\gamma_k$represents the continuous variable *Tobacco*,
$(\alpha\beta)_{i,j}$, $(\beta\gamma)_{j,k}$, $(\alpha\gamma)_{i,k}$ represent the two-way interaction effects(with constraints
$\boldsymbol{\alpha_1} = \boldsymbol{\beta_1} = \boldsymbol{\gamma_1} = (\boldsymbol{\alpha\beta})_{1,1} = (\boldsymbol{\alpha\gamma})_{1,1} = (\boldsymbol{\beta\gamma})_{1,1} = \mathbf{0}$).

Test adequatcy of Main effects *only* and with *any* Interaction effects in fitted Two-way Logistic regression model

```
fit_1 <- glm(Disease ~ AgeGroup + Alcohol + Tobacco, family="binomial", data = cancer)
summary(fit_1)
```

```
##
## Call:
## glm(formula = Disease ~ AgeGroup + Alcohol + Tobacco, family = "binomial",
##     data = cancer)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5301  -0.6548  -0.3873  -0.1525   2.8211
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -5.9108     1.0302  -5.738 9.59e-09 ***
## AgeGroup35-44   1.6095     1.0675   1.508 0.131631
## AgeGroup45-54   2.9752     1.0242   2.905 0.003673 **
## AgeGroup55-64   3.3584     1.0198   3.293 0.000991 ***
## AgeGroup65-74   3.7270     1.0252   3.635 0.000278 ***
## AgeGroup75+     3.6818     1.0644   3.459 0.000542 ***
## Alcohol120+     2.1154     0.2876   7.356 1.90e-13 ***
## Alcohol40-79    1.1216     0.2384   4.704 2.55e-06 ***
## Alcohol80-119   1.4471     0.2628   5.506 3.68e-08 ***
## Tobacco10-19    0.3407     0.2054   1.659 0.097159 .
## Tobacco20-29    0.3962     0.2456   1.613 0.106708
## Tobacco30+      0.8677     0.2765   3.138 0.001701 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1072.13  on 1174  degrees of freedom
## Residual deviance:  898.86  on 1163  degrees of freedom
## AIC: 922.86
##
## Number of Fisher Scoring iterations: 7
```

```
anova(fit_1,test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: Disease
##
## Terms added sequentially (first to last)
##
##
##          Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                      1174    1072.13
## AgeGroup  5   88.128      1169     984.00 < 2.2e-16 ***
## Alcohol   3   74.541      1166     909.46 4.545e-16 ***
## Tobacco   3   10.599      1163     898.86   0.01411 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

H_0 : all coefficients of main effects term = 0

H_1 : H_0 is not true.

all p-value < 0.05. Hence the main terms are significant, indicating that the mean function *only main effects* will be adequate to explain these data.

Any Interaction effects can be removed?

```
fit_full = glm(Disease ~ AgeGroup + Alcohol + Tobacco+ AgeGroup:Alcohol +AgeGroup:Tobacco+Alcohol:Tobacco, family="binomia
l", data = cancer)
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
anova(fit_full, test="Chisq")
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: Disease
##
## Terms added sequentially (first to last)
##
##
##                  Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                              1174    1072.13
## AgeGroup          5   88.128      1169     984.00 < 2.2e-16 ***
## Alcohol           3   74.541      1166     909.46 4.545e-16 ***
## Tobacco           3   10.599      1163     898.86   0.01411 *
## AgeGroup:Alcohol 15   20.177      1148     878.68   0.16525
## AgeGroup:Tobacco 15   10.020      1133     868.66   0.81845
## Alcohol:Tobacco   9    7.667      1124     860.99   0.56806
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
drop1(fit_full, test = "LR")
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
## Single term deletions
##
## Model:
## Disease ~ AgeGroup + Alcohol + Tobacco + AgeGroup:Alcohol + AgeGroup:Tobacco +
##     Alcohol:Tobacco
##                  Df Deviance    AIC     LRT Pr(>Chi)
## <none>                860.99 962.99
## AgeGroup:Alcohol 15   879.25 951.25 18.2547   0.2495
## AgeGroup:Tobacco 15   872.03 944.03 11.0363   0.7500
## Alcohol:Tobacco   9   868.66 952.66  7.6666   0.5681
```

Since all the p-values >0.05 by the likelihood ratio test, all two-way interaction effects can be removed at 95% significance level.

Find the ratio of odds of having esophageal cancer for Alcohol group 120+ and Alcohol group 0-39g/day

```
exp(coef(fit_1))
```

```
##   (Intercept) AgeGroup35-44 AgeGroup45-54 AgeGroup55-64 AgeGroup65-74
##   0.002710046   5.000426419  19.592860601  28.741838714  41.554820473
##   AgeGroup75+   Alcohol120+  Alcohol40-79 Alcohol80-119  Tobacco10-19
##  39.716131696   8.292938857   3.069638047   4.250811157   1.405982889
##  Tobacco20-29    Tobacco30+
##   1.486221090   2.381435327
```

The odds ratio is 8.292938857. Having alchohol over 120g/day is 8.2929 times higher probability of getting esophageal cancer than people who have alchohol 0-39g/day.

Predict Probability of getting esophageal cancer for 60 years old, does not drink alcohol at all, but has tobacco consumption more than 30g/day.

```
newdata <- data.frame(AgeGroup = "55-64", Alcohol ="0-39g/day" ,Tobacco ="30+")
pred3 = predict(fit_1, newdata = newdata, type = "response", se = TRUE)
pred3
```

```
## $fit
##         1
## 0.1564698
##
## $se.fit
##          1
## 0.04208908
##
## $residual.scale
## [1] 1
```

probability: 0.1564698, standard error: 0.04208908

```
list("Probability"=pred3$fit,
     "Lower"=pred3$fit-1.96*pred3$se.fit,
     "Upper"=pred3$fit+1.96*pred3$se.fit)
```

```
## $Probability
##         1
## 0.1564698
##
## $Lower
##          1
## 0.07397523
##
## $Upper
##         1
## 0.2389644
```

Predict Probability of getting esophageal cancer for 60 years old, does not smoke at all, but has alcohol consumption more than 120g/day?

```
pred4 = predict(fit_1, newdata=data.frame(AgeGroup="55-64",Alcohol="120+",Tobacco="0-9g/day"), interval="confidence", type
="response", se=TRUE)
pred4
```

```
## $fit
##         1
## 0.3924485
##
## $se.fit
##          1
## 0.06091598
##
## $residual.scale
## [1] 1
```

```
list("Probability"=pred4$fit,
     "Lower"=pred4$fit-1.96*pred4$se.fit,
     "Upper"=pred4$fit+1.96*pred4$se.fit)
```

```
## $Probability
##         1
## 0.3924485
##
## $Lower
##         1
## 0.2730532
##
## $Upper
##         1
## 0.5118439
```

60 years old, does not smoke at all, but has alcohol consumption more than 120g/day will be 39.24485% getting esophageal cancer, Range from 27.30532% to 51.18439%.

Conclusion:

Drinking a lot might have more impact on getting esophageal cancer compaered with heavy smoking.

An experiment on the effect of diet on early growth of chicks.

weight: a numeric vector giving the body weight of the chick (gm).

Time: a numeric vector giving the number of days since birth when the measurement was made.

Chick: an ordered factor with levels 18 < … < 48 giving a unique identifier for the chick. The ordering of the levels groups chicks on the same diet together and orders them according to their final weight (lightest to heaviest) within diet.

Diet: a factor with levels 1, 2, 3, 4 indicating which experimental diet the chick received.

Fitting Two-way ANOVA mean function: $\boldsymbol{E(y_ij) = \mu + \alpha_i + \beta x_{i,j} + \beta_i x_{i,j}, i = 1,2,3,4; j = 1,2,\ldots,n_i}$

```
data(ChickWeight)
str(ChickWeight)
```

```
## Classes 'nfnGroupedData', 'nfGroupedData', 'groupedData' and 'data.frame':   578 obs. of  4 variables:
##  $ weight: num  42 51 59 64 76 93 106 125 149 171 ...
##  $ Time  : num  0 2 4 6 8 10 12 14 16 18 ...
##  $ Chick : Ord.factor w/ 50 levels "18"<"16"<"15"<..: 15 15 15 15 15 15 15 15 15 15 ...
##  $ Diet  : Factor w/ 4 levels "1","2","3","4": 1 1 1 1 1 1 1 1 1 1 ...
##  - attr(*, "formula")=Class 'formula'  language weight ~ Time | Chick
##   .. ..- attr(*, ".Environment")=<environment: R_EmptyEnv>
##  - attr(*, "outer")=Class 'formula'  language ~Diet
##   .. ..- attr(*, ".Environment")=<environment: R_EmptyEnv>
##  - attr(*, "labels")=List of 2
##   ..$ x: chr "Time"
##   ..$ y: chr "Body weight"
##  - attr(*, "units")=List of 2
##   ..$ x: chr "(days)"
##   ..$ y: chr "(gm)"
```

```
model_1 = lm(weight~Time + Diet + Time:Diet, data=ChickWeight)
model_1
```

```
##
## Call:
## lm(formula = weight ~ Time + Diet + Time:Diet, data = ChickWeight)
##
## Coefficients:
## (Intercept)         Time        Diet2        Diet3        Diet4   Time:Diet2
##     30.9310       6.8418      -2.2974     -12.6807      -0.1389       1.7673
##  Time:Diet3   Time:Diet4
##      4.5811       2.8726
```

```
Anova(model_1, type="II")
```

```
## Anova Table (Type II tests)
##
## Response: weight
##            Sum Sq  Df  F value    Pr(>F)
## Time      2016357   1 1737.367 < 2.2e-16 ***
## Diet       129876   3   37.302 < 2.2e-16 ***
## Time:Diet   80804   3   23.208 3.474e-14 ***
## Residuals  661532 570
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Fitting ANOVA mean function: $\boldsymbol{E(y_{i,j}) = \mu + \alpha_i + \beta x_{i,j}, i = 1, 2, 3, 4; j = 1, 2, \ldots, n_i}$

```
model_2 = lm(weight~Time + Diet, data=ChickWeight)
Anova(model_2, type="II")
```

```
## Anova Table (Type II tests)
##
## Response: weight
##            Sum Sq  Df  F value    Pr(>F)
## Time      2016357   1 1556.401 < 2.2e-16 ***
## Diet       129876   3   33.417 < 2.2e-16 ***
## Residuals  742336 573
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Compare if Two-way ANOVA model is better

```
anova(model_2, model_1)
```

```
## Analysis of Variance Table
##
## Model 1: weight ~ Time + Diet
## Model 2: weight ~ Time + Diet + Time:Diet
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    573 742336
## 2    570 661532  3     80804 23.208 3.474e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The P-value of Two-way ANOVA = 3.474e-14 < 0.001, The Interaction effect is significant, therefore the Two-way ANOVA mean function model is better.

Predict 95% Confidence Interval for the mean weight of chick, when diet 1 at time 22

```
predict(model_1, newdata=data.frame(Time=22, Diet="1"),
        interval="confidence",type="response")
```

```
##        fit      lwr      upr
## 1 181.4505 172.5178 190.3832
```

The predicted mean weight of chick is 181.4505, and 95% confidence interval is (172.5178, 190.3832)