STAT 675 – APPLIED DISCRETE DATA ANALYSIS

Longitudinal study: Effect of tobacco use on mortality

Contents

1. Introduction

It is estimated that around 1.22 billion people worldwide smoke as of 2000 [1]. In this analysis, we aim to understand the effect of tobacco on mortality. Specifically, we seek to evaluate the effect of different level of tobacco use on mortality during a specific follow-up period. For that purpose, we will use the National Longitudinal Mortality Study (NLMS) dataset, which is described in more details in section 2. Section 3 provides methodologies used in our analysis and discusses the results. We summarize our findings in Section 4.

2. Dataset

The National Longitudinal Mortality Study (NLMS) is series of longitudinal, observational studies sponsored by the National Cancer Institute, the National Heart, Lung, and Blood Institute, the National Institute on Aging, the National Center for Health Statistics and the U.S. Census Bureau for the purpose of studying the effects of differentials in demographic and socio-economic characteristics on mortality [2]. The series of studies began in 1983, with subjects randomly selected from Census Bureau population samples. The subjects are matched to the National Death Index (NDI) maintained by the National Center for Health Statistics to determine which individuals have died.

One study in the series began in 1993 and included tobacco use information collected periodically between 1993 and 2005 with 5 years of follow-up for mortality. Full description of the dataset is available on National Heart, Lung and Blood Institute website [3]. A summary of variables used in our analysis are described below:

- inddea: indicator whether the subject is dead at the end of follow-up period (1-Yes, 0-No).
- follow: The length of follow-up period in days (1827 for subjects who survived after 5 years)
- smokstat: Cigarette smoking Status at study point (1 - Never smoked cigarettes, 2 - An everyday smoker of cigarettes, 3 - A smoker of cigarettes on some days, 4 - A former smoker of cigarettes)
- eversmoke: A derived variable to indicate whether subject has ever smoked at study point (0 – Never smoked, 1 – Smoked previously)
- age: subject's age at study point. Value greater than 90 is coded as 90.
- race: subject's race, treated as nominal variable.
- sex: subject' sex (1-Male, 2-Female). "sex" should be treated nominal variable, but since there are only two levels, we can keep the variable as it is.
- ms: subject's marital status at study point, treated as nominal variable.
- hisp: subject's Hispanic origin, treated as nominal variable.
- pob: subject's place of birth, treated as nominal variable.
- educ: subject's education level, treated as nominal variable.
- esr: subject's employment status at study point, treated as nominal variable.
- adjinc: subject's inflation adjusted income at study point, treated as ordinal variable with higher number indicating higher income.

- histatus: subject's health insurance status at study point, treated as nominal variable.
- stater: subject's interview location at study point

We simply remove records with missing values, leaving 377,750 records for our analysis.

3. Methodology and Analysis

3.1. Tobacco Use Summary Statistics

About 41% of subjects have ever smoked. The subjects have age range from 15 to 90 with median 42 and mean 44. We want to focus our analysis on groups with high percentage of people smoking.

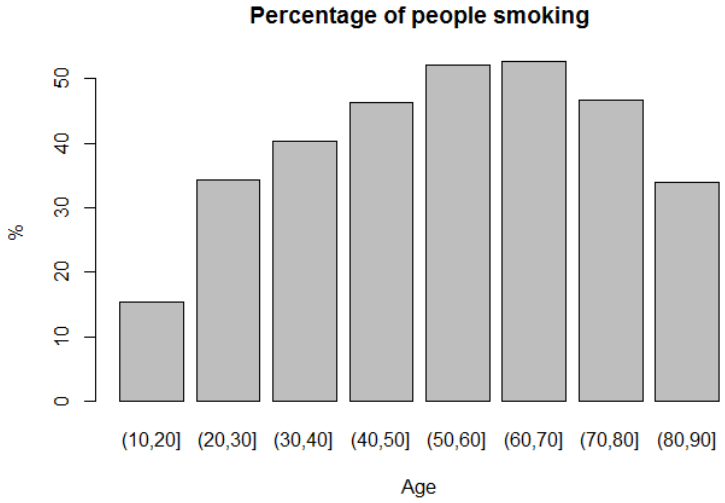

Figure 3.1.1 – Percentage of people smoking by age group

Figure 3.1.1 shows subjects in age group 50-70 have the highest percentage of people ever smoking (about 50%). In our analysis, we will look at this specific age group 50-70, which includes 89,889 records.

3.2. Contingency Table Analysis

We conduct 2-way contingency table to evaluate whether overall morality is associated with smoking tobacco.

		inddea (1-Death, 0-Survive)	
		0	**1**
eversmoke	**0**	41,381	1,431
(1-Yes, 0-No)	**1**	43,815	3,262

Table 3.2.1 – Contingency table inddea - eversmoke

The sample odd ratio of death among subjects smoked previously vs those never smoke is 2.15.

Since n = 89889 is sufficiently large, we can use normal approximation to estimate confidence interval for log-odd ratio and odd ratio. Wald confidence interval at 5% significant level for the odd ratio is (2.02, 2.29). As both

lower bound and upper bound are greater than 1, we can conclude at 95% confidence level that smoking is associated with higher mortality rate (more than two times).

3.3. Logistics Regression Analysis

Model and Variable Selection:

Although contingency table provides us with overview of how mortality associates with smoking, mortality might be influenced by other variables as well. We can use logistics regression to include other explanatory variables and evaluate the effect contributed by smoking status when other variables also present. Specifically, we use logistics regression to model the relationship between probability of death during 5-year follow-up period and explanatory variables (age, race, sex, marital status, Hispanic origin, education level, place of birth, employment status, income, health insurance status, smoking status).

We will perform alternative step-wise variable selection with Bayesian Information Criteria (BIC) to select only important variables in the model. BIC tends to select smaller model and it has "consistency" property, i.e. select the "right" model with probability approaching 1 given the right model is among those that examined in step-wise selection process [4]. We could use a binary representation, an ordinal representation, or a nominal representation with 4 levels for smoking status in the model. The best logistics regression model is shown below:

```
Call:
glm(formula = inddea ~ esr + age + as.factor(df$smokstat) + sex +
    ms + hisp + adjinc, family = binomial(link = "logit"), data = df)
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.2475  -0.3539  -0.2498  -0.1791   3.2916

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)             -8.951486   0.228184 -39.229  < 2e-16 ***
esr2                     0.514484   0.096080   5.355 8.57e-08 ***
esr3                     0.408593   0.129330   3.159  0.00158 **
esr4                     1.601147   0.050864  31.479  < 2e-16 ***
esr5                     0.561693   0.042531  13.207  < 2e-16 ***
age                      0.084898   0.003251  26.114  < 2e-16 ***
as.factor(df$smokstat)2  0.891140   0.042475  20.980  < 2e-16 ***
as.factor(df$smokstat)3  0.779054   0.080745   9.648  < 2e-16 ***
as.factor(df$smokstat)4  0.456106   0.037712  12.094  < 2e-16 ***
sex2                    -0.494292   0.033188 -14.894  < 2e-16 ***
ms2                      0.295364   0.050297   5.872 4.29e-09 ***
ms3                      0.303006   0.046928   6.457 1.07e-10 ***
ms4                      0.338751   0.100279   3.378  0.00073 ***
ms5                      0.350706   0.066204   5.297 1.17e-07 ***
hisp2                   -0.187134   0.149353  -1.253  0.21022
hisp3                    0.422677   0.101229   4.175 2.97e-05 ***
adjinc                  -0.027278   0.004770  -5.718 1.08e-08 ***
---
(Dispersion parameter for binomial family taken to be 1)
    Null deviance: 36849  on 89888  degrees of freedom
Residual deviance: 33089  on 89872  degrees of freedom
AIC: 33123
```

Model Checking:

With model containing continuous variables, the Hosmer and Lemeshow test [5] would be an appropriate Goodness of Fit test. Hosmer and Lemeshow test statistics has p-value = 0.182 > 0.05, suggesting the model fit is not bad. As an alternative, we can also look at Standardized Residuals, which have 4.6% < 5% residuals beyond 2, indicating no obvious problem with model fit.

Model Interpretation:

Age, Sex, Marital Status, Hispanic Origin, Employment Status, Income and Smoking Status are important in the model.

Looking specifically at Smoking Status, all p-values << 0.05, indicating strong evidence that it still has significant association with probability of death, even when other variables present in the model.

As smokstat level increases (non-smoker < former smoker < someday smoker < everyday smoker), the estimated odd of death during 5-year follow up period increases, given no changes in other variables.

- Smokstat changing from 0 (Never Smoke) to 4 (Former Smoker) increase the estimated odd of death during 5-year follow-up period by 1.58 times. 95% Wald confidence interval for odd ratio is (1.46, 1.70)
- Smokstat changing from 0 (Never Smoke) to 3 (Smoke some days) increase the estimated odd of death during 5-year follow-up period by 2.18 times. 95% Wald confidence interval for odd ratio is (1.86, 2.55)
- Smokstat changing from 0 (Never Smoke) to 2 (Smoke everyday) increase the estimated odd of death during 5-year follow-up period by 2.44 times. 95% Wald confidence interval for odd ratio is (2.24, 2.65)

3.4. Mixed Effect Model

In the above logistics regression model, we assume that all observations are independent. However, such assumption might not be appropriate because the study was carried out at 51 different sites, in which subjects might form a cluster and have certain association. We can use mixed effect model to evaluate whether the random effect caused by association among subjects in the same site is significant. We extend the logistics regression model to include random effect on intercept, i.e. the linear predictor with random effect is $\beta_0 + b_{0i}$, where β_0 is the value of linear predictor at the average site, and b_{0i} across 51 sites are random samples from $N(0, \sigma_{b0}^2)$.

Model fitting using Adaptive Gaussian quadrature seems to stabilize at k = 5 quadrature points. We can see the same result that as smokstat level increases (non-smoker < former smoker < someday smoker < everyday smoker), the estimated odd of death during 5-year follow up period increases, given no changes in other variables.

```
Generalized linear mixed model fit by maximum likelihood (Adaptive
  Gauss-Hermite Quadrature, nAGQ = 5) [glmerMod]
 Family: binomial  ( logit )
Formula: inddea ~ age + esr + as.factor(smokstat) + sex + ms + hisp +
    adjinc + (1 | stater)
   Data: df

     AIC      BIC   logLik deviance df.resid
 33110.2  33279.5 -16537.1  33074.2    89871

Scaled residuals:
    Min      1Q  Median      3Q     Max
-1.1187 -0.2538 -0.1775 -0.1268 15.4877

Random effects:
 Groups Name        Variance Std.Dev.
 stater (Intercept) 0.01005  0.1003
Number of obs: 89889, groups:  stater, 51

Fixed effects:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)          -8.872467   0.230701  -38.46  < 2e-16 ***
age                   0.084991   0.003255   26.11  < 2e-16 ***
esr2                  0.513671   0.096146    5.34 9.16e-08 ***
esr3                  0.414220   0.129464    3.20 0.001377 **
esr4                  1.601529   0.051034   31.38  < 2e-16 ***
esr5                  0.564437   0.042606   13.25  < 2e-16 ***
as.factor(smokstat)2  0.888638   0.042537   20.89  < 2e-16 ***
as.factor(smokstat)3  0.781370   0.080811    9.67  < 2e-16 ***
as.factor(smokstat)4  0.452316   0.037775   11.97  < 2e-16 ***
sex2                 -0.495560   0.033211  -14.92  < 2e-16 ***
ms2                   0.298865   0.050383    5.93 3.00e-09 ***
ms3                   0.306675   0.047074    6.51 7.28e-11 ***
ms4                   0.344468   0.100575    3.42 0.000615 ***
ms5                   0.353825   0.066460    5.32 1.02e-07 ***
hisp2                -0.237281   0.152958   -1.55 0.120836
hisp3                 0.339827   0.105747    3.21 0.001311 **
adjinc               -0.026709   0.004807   -5.56 2.76e-08 ***
```

Fixed Effect:

We can evaluate the significance of fixed effects by removing one effect and perform LRT to compare each resulting model with the full model.

H_0: A model with one fixed effect dropped

H_a: Full model

```
Single term deletions

Model:
inddea ~ age + esr + as.factor(smokstat) + sex + ms + hisp +
    adjinc + (1 | stater)
                    Df   AIC    LRT   Pr(Chi)
<none>                  33110
age                  1 33815 707.18 < 2.2e-16 ***
esr                  4 34019 916.76 < 2.2e-16 ***
as.factor(smokstat)  3 33563 458.98 < 2.2e-16 ***
sex                  1 33334 225.92 < 2.2e-16 ***
ms                   4 33178  75.48 1.580e-15 ***
hisp                 2 33145  38.89 3.595e-09 ***
adjinc               1 33139  30.94 2.668e-08 ***
```

With large n = 89889, the transformed LRT statistics is approximated with χ^2_{df} under null hypothesis. P-values << 0.05 indicating Age, Sex, Marital Status, Hispanic Origin, Employment Status, Income and Smoking Status are important in the model, given other variables in the model.

Random Effect:

The estimated variance σ_{b0}^2 is 0.10008 and sample random effect on intercept for each interview location is shown below:

(Intercept)					
11	0.1020777826	44	-0.0577599561	72	-0.0492788214
12	0.1608164835	45	-0.0028408864	73	0.0267377881
13	-0.0365513045	46	-0.0770930789	74	-0.1141859362
14	0.1969534701	47	0.0371463361	81	0.0236785570
15	-0.0589549708	51	0.0053522617	82	0.0162222813
16	0.1274710267	52	0.0001299361	83	0.0072855948
21	-0.1071572479	53	0.0666777078	84	-0.1036569953
22	0.0136054639	54	0.0171306974	85	0.0632772655
23	0.0110955358	55	0.0347273728	86	-0.0719340541
31	0.0126816895	56	0.0617829906	87	0.0231119179
32	0.0259470742	57	0.0315073947	88	0.0156424698
33	0.0075704046	58	0.0294044299	91	-0.0026325338
34	0.0081764708	59	-0.0576205114	92	-0.0107673357
35	-0.0254589790	61	0.0763002477	93	-0.1577158859
41	-0.0046297165	62	-0.0061499350	94	-0.0560396063
42	0.0113334952	63	-0.0317801353	95	-0.0877819206
43	-0.0109682815	64	0.0036245948		
		71	-0.0389381024		

For example, subjects interviewed in New York (stater=21) would have estimated odd of death 1.11 times less than the average site.

We can test the significance of variance component by comparing model with and without random effect on intercept:

H_0: Model without random effect $\sigma_{b0}^2 = 0$

H_a: Model with random effect $\sigma_{b0}^2 > 0$

With large n = 89889, the transformed LRT statistics is approximated with χ^2_{df} under null hypothesis. p-value is close to 0 << 0.05. Therefore, we can reject the null hypothesis and conclude that the variable component is significant.

3.5. Survival Analysis

As an alternative approach to understand the effect of tobacco use on mortality, we could also investigate how the odd of survival, which equals 1 – odd of death, associates with tobacco use overtime during the follow-up period. We can conduct Log-Rank test to evaluate if such survival functions, i.e. the probability of survival beyond a specific time, are the same among subjects with different level of smoking:

H_0: survival functions are the same across different level of smoking

H_a: survival functions are different among different level of smoking

With large n = 89889, the log-rank statistic is approximately standard normal under null hypothesis. Test statistics = 724 with df=3, and p-value << 0.05, suggesting strong evidence that there is difference in survival distributions among subjects with different level of smoking.

The Kaplan–Meier estimator can be used to estimate survival function. The estimated survival functions (solid lines), together with 95% confidence intervals (dotted lines) of subjects who never smoke and those who do are shown in Figure 3.5.1:

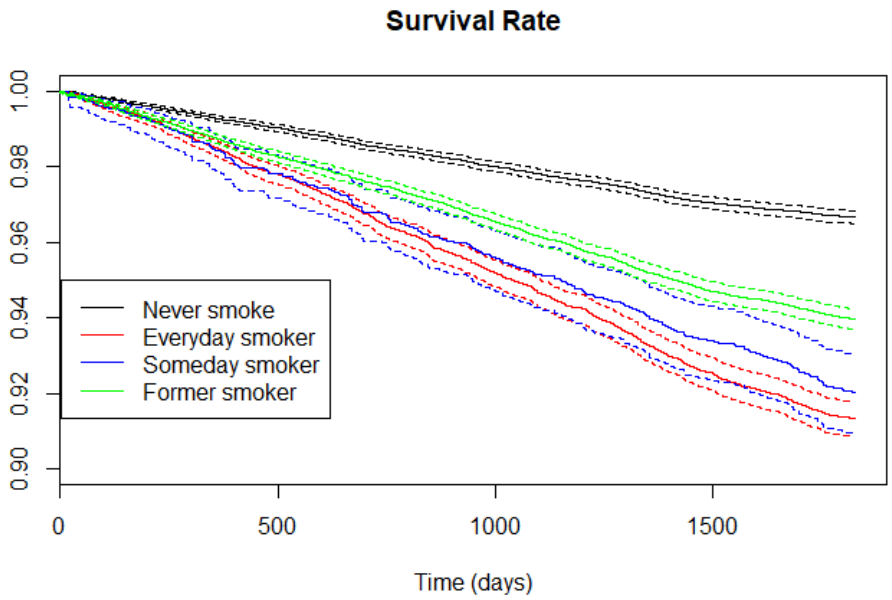

Figure 3.5.1 – Expected survival rate and 95% confidence interval vs Time

The Kaplan–Meier estimator shows as smokstat level increases, the estimated probability of survival during 5-year follow up period decreases. Looking at 95% confidence interval:

- The probability of survival is higher for those who never smoke compared to those smoking
- Near the end of follow-up period, the probability of survival of former smokers is higher compared to those smoking at study point
- There is not enough evidence to conclude on differences in the probability of survival between everyday smokers and some-day smokers.

4. Conclusion

In summary, more than 50% of subjects in age range 50-70 have smoked, and smoking tobacco clearly associates with higher mortality (or lower survival) during 5-year follow up period in this segment. The data shows at 95% confidence level that the mortality rate at the end of follow up period increases with the smoking levels: non-smokers < former smokers < current smokers. However, since the dataset is not originated from a controlled experiment, we can't draw any inference on causal relationship between tobacco use and mortality.

REFERENCES

1. *Guindon G. Emmanuel, Boisclair, David (2003),* Past, current and future trends in tobacco use, *Washington DC: The International Bank for Reconstruction and Development / The World Bank: 13–16.*
2. National Longitudinal Mortality Study, *National Heart, Lung and Blood Institute. Retrieved on 10 Apr 2017* [link]
3. Code Manuals and Forms, *National Heart, Lung and Blood Institute. Retrieved on 10 Apr 2017* [link]
4. *Christopher R. Bilder, Thomas M. Loughin*, Analysis of Categorical Variable with R, *Page 268.*
5. *Christopher R. Bilder, Thomas M. Loughin*, Analysis of Categorical Variable with R, *Goodness of Fit Test* [link]

APPENDIX – R CODE AND KEY OUTPUT

Data Pre-processing

```
## Read Data

tu <- read.csv(file = "tu.csv")

df <- subset(tu, select = c("age", "race", "sex", "ms", "hisp", "educ", "pob", "adjinc", "esr",
"histatus", "stater", "smokstat", "inddea", "follow"))

df <- df[complete.cases(df),]


## Data Pre-Processing

df$age_group = cut(df$age, breaks=c(10, 20, 30, 40, 50, 60, 70, 80, 90))

df$inddea = factor(df$inddea)

df$race = factor(df$race)

df$sex = factor(df$sex)

df$ms = factor(df$ms)

df$hisp = factor(df$hisp)

df$educ = factor(df$educ)

df$pob = factor(df$pob)

df$esr = factor(df$esr)

df$histatus = factor(df$histatus)

df$stater = factor(df$stater)

df$eversmoke <- ifelse((df$smokstat==1), yes=0, no=1)

head(df)head(df)
```

```
       age race sex ms hisp educ pob adjinc esr histatus stater smokstat
102038  68    1   2  1    3    4 917      5   5         1     33         2
102041  62    1   1  5    3    2 917      1   5         1     33         2
102042  53    1   1  1    3    4 917      5   1         0     33         4
102044  67    1   2  2    3    8 917      5   5         1     33         1
102045  70    1   1  2    3    4 917      4   5         1     33         1
102046  51    1   1  1    3   14 917     13   1         1     33         4
       inddea follow age_group eversmoke
102038      0   1827   (60,70]         1
102041      0   1827   (60,70]         1
102042      0   1827   (50,60]         1
102044      0   1827   (60,70]         0
102045      0   1827   (60,70]         0
102046      0   1827   (50,60]         1
```

Summary Statistics

```
summary(df$eversmoke)

summary(df$age)

library(reshape2)

smoke.agegroup <- dcast(df, age_group ~ ., function(eversmoke) mean(eversmoke))

barplot(100*smoke.agegroup$., main="Percentage of people smoking", xlab="Age", ylab="%",

      names.arg=smoke.agegroup$age_group)
```

Contingency Table Analysis

```
df <- df[df$age > 50 & df$age <= 70,]

c.table <- xtabs(formula = ~ eversmoke + inddea, data = df)

c.table

OR.hat <- 1.0*c.table[1,1]*c.table[2,2]/(c.table[2,1]*c.table[1,2])

paste("sample OR:", round(OR.hat,4))

var.log.or <- 1/c.table[1,1] + 1/c.table[1,2] + 1/c.table[2 ,1] + 1/c.table [2,2]

alpha = 0.05

OR.CI <- exp(log(OR.hat) + qnorm(p=c(alpha/2, 1-alpha/2))*sqrt(var.log.or))

paste("CI OR:", round (OR.CI , 4))
```

```
         inddea
eversmoke      0      1
        0 214997   6802
        1 147122   8829
"sample OR: 1.8968"
"CI OR: 1.7911" "CI OR: 2.0089"
```

Logistics Regression Model

```
# Step-wise Variable Selection

empty.mod = glm(formula = inddea ~ 1, family = binomial(link="logit"), data = df)


# with binary smoking status (Yes/No)

full.mod = glm(formula = inddea ~ age + race + sex + ms + hisp + educ + pob + esr + adjinc + histatus
+ eversmoke, family = binomial(link="logit"), data = df)

step.sel <- step(object = empty.mod, scope = list(upper = full.mod), k = log(nrow(df)), trace = TRUE)

summary(step.sel)


# with ordinal smoking status (1 < 4 < 3 < 2)

df$smokstat_order <- ifelse((df$smokstat==1), yes=1,

                    no=ifelse((df$smokstat==4), yes=2,

                    no=ifelse((df$smokstat==3), yes=3, no=2)))

full.mod = glm(formula = inddea ~ age + race + sex + ms + hisp + educ + pob +esr + adjinc + histatus
+ smokstat_order, family = binomial(link="logit"), data = df)

step.sel <- step(object = empty.mod, scope = list(upper = full.mod), k = log(nrow(df)), trace = TRUE)

summary(step.sel)


# with nominal smoking status (1, 2, 3, 4)

full.mod = glm(formula = inddea ~ age + race + sex + ms + hisp + educ + pob +esr + adjinc + histatus
+ as.factor(df$smokstat), family = binomial(link="logit"), data = df)

step.sel <- step(object = empty.mod, scope = list(upper = full.mod), k = log(nrow(df)), trace = TRUE)

summary(step.sel)
```

```
Call:
glm(formula = inddea ~ esr + age + eversmoke + sex + ms + adjinc +
    hisp, family = binomial(link = "logit"), data = df)
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.1479  -0.3541  -0.2503  -0.1816   3.2894

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -8.652316   0.225624 -38.348  < 2e-16 ***
esr2         0.514719   0.096016   5.361 8.29e-08 ***
esr3         0.416769   0.129214   3.225 0.001258 **
esr4         1.615477   0.050826  31.785  < 2e-16 ***
esr5         0.564697   0.042536  13.276  < 2e-16 ***
age          0.080256   0.003213  24.978  < 2e-16 ***
eversmoke    0.617000   0.033980  18.158  < 2e-16 ***
sex2        -0.484482   0.033165 -14.608  < 2e-16 ***
ms2          0.322945   0.050171   6.437 1.22e-10 ***
ms3          0.337020   0.046736   7.211 5.55e-13 ***
ms4          0.373779   0.099938   3.740 0.000184 ***
ms5          0.363266   0.066087   5.497 3.87e-08 ***
adjinc      -0.031575   0.004753  -6.643 3.07e-11 ***
hisp2       -0.183354   0.149189  -1.229 0.219070
hisp3        0.426731   0.101152   4.219 2.46e-05 ***
---
(Dispersion parameter for binomial family taken to be 1)
    Null deviance: 36849  on 89888  degrees of freedom
Residual deviance: 33205  on 89874  degrees of freedom
AIC: 33235
```

```
Call:
glm(formula = inddea ~ esr + age + smokstat_order + sex + ms +
    hisp + adjinc, family = binomial(link = "logit"), data = df)
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.3511  -0.3553  -0.2503  -0.1827   3.2870

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)    -9.174127   0.231136 -39.691  < 2e-16 ***
esr2            0.517465   0.096007   5.390 7.05e-08 ***
esr3            0.418159   0.129221   3.236 0.001212 **
esr4            1.616915   0.050845  31.801  < 2e-16 ***
esr5            0.566964   0.042530  13.331  < 2e-16 ***
age             0.080762   0.003213  25.138  < 2e-16 ***
smokstat_order  0.513532   0.028518  18.007  < 2e-16 ***
sex2           -0.503542   0.032984 -15.266  < 2e-16 ***
ms2             0.319796   0.050169   6.374 1.84e-10 ***
ms3             0.338544   0.046746   7.242 4.41e-13 ***
ms4             0.354020   0.100134   3.535 0.000407 ***
ms5             0.352298   0.066053   5.334 9.63e-08 ***
hisp2          -0.170667   0.149223  -1.144 0.252746
hisp3           0.455244   0.101139   4.501 6.76e-06 ***
adjinc         -0.030934   0.004749  -6.513 7.35e-11 ***
---
(Dispersion parameter for binomial family taken to be 1)
    Null deviance: 36849  on 89888  degrees of freedom
Residual deviance: 33227  on 89874  degrees of freedom
AIC: 33257
```

```
Call:
glm(formula = inddea ~ esr + age + as.factor(df$smokstat) + sex +
    ms + hisp + adjinc, family = binomial(link = "logit"), data = df)

# Result shown in Section 3-3
```

Model Checking:

```
s.res1 = rstandard(step.sel, type="pearson")

(length(s.res1[s.res1>2])+length(s.res1[s.res1<-2]))

length(s.res1)

source("AllGOFTests.R")

HL <- HLTest(obj=step.sel, g = 10)

HL <- HLTest(obj=step.sel, g = 100)
```

```
> 4209/89889
[1] 0.04682442

Hosmer and Lemeshow goodness-of-fit test with 10 bins
data:  step.sel
X2 = 12.552, df = 8, p-value = 0.1282

Hosmer and Lemeshow goodness-of-fit test with 100 bins
data:  step.sel
X2 = 110.55, df = 98, p-value = 0.182
```

Inference:

```
OR.hat <- exp(summary(step.sel)$coefficients[9][1])

var.log.or <- summary(step.sel)$coefficients[26][1]

OR.CI <- exp(log(OR.hat) + qnorm(p=c(alpha/2, 1-alpha/2))*(var.log.or))

paste("sample OR:", round(OR.hat,4))

paste("CI OR:", round (OR.CI , 4))


OR.hat <- exp(summary(step.sel)$coefficients[8][1])

var.log.or <- summary(step.sel)$coefficients[25][1]

OR.CI <- exp(log(OR.hat) + qnorm(p=c(alpha/2, 1-alpha/2))*(var.log.or))

paste("sample OR:", round(OR.hat,4))

paste("CI OR:", round (OR.CI , 4))


OR.hat <- exp(summary(step.sel)$coefficients[7][1])

var.log.or <- summary(step.sel)$coefficients[24][1]

OR.CI <- exp(log(OR.hat) + qnorm(p=c(alpha/2, 1-alpha/2))*(var.log.or))

paste("sample OR:", round(OR.hat,4))

paste("CI OR:", round (OR.CI , 4))
```

Mixed Effect Model

```
library(lme4)
k_list <- c(1,2,5,10,20)
for (k in k_list) {
  mod.glmm <- glmer(formula = inddea ~ age + esr + as.factor(smokstat) + sex + ms + hisp +
                    adjinc + (1|stater), nAGQ = k, family = binomial(link = "logit"), data = df)
  print(paste("nAGQ: ", k))
  print(summary(mod.glmm)$varcor)
}
mod.glmm <- glmer(formula = inddea ~ age + esr + as.factor(smokstat) + sex + ms + hisp +
                  adjinc + (1|stater), nAGQ = 5, family = binomial(link = "logit"), data = df)
summary(mod.glmm)
```

```
# Model Summary shown in Section 3.4
```

```
# Fixed Effect
fixef(mod.glmm)
lrt <- drop1(mod.glmm, test = "Chisq")
lrt


# Variance component
ranef(mod.glmm)$stater
LRstat.vc <- deviance(step.sel) - deviance(mod.glmm)
```

```
# LR Test Result for Fixed Effects shown in Section 3.4

> LRstat.vc <- deviance(step.sel) - deviance(mod.glmm)
> (1 - pchisq(LRstat.vc, df = 1))/2
[1] 3.330669e-16
```

Survival Analysis

```
library(survival)

df$SurvObj <- Surv(df$follow, df$inddea==1)



# Log-Rank Test

survdiff(SurvObj ~ as.factor(smokstat),data=df)



# Kaplan-Meier estimator

km.by.smokstat <- survfit(SurvObj ~ as.factor(smokstat), data = df, conf.type = "log-log")

km.by.smokstat

plot(km.by.smokstat, conf.int=TRUE , ylim=c(0.9, 1),

    col=c('black','red', 'blue', 'green'), main="Survival Rate", xlab='Time (days)')

legend(1,0.95, legend=c("Never smoke", "Everyday smoker", "Someday smoker", "Former
smoker"),

      lty=c(1,1,1,1) ,col=c('black','red', 'blue', 'green'))
```

```
Call:
survdiff(formula = SurvObj ~ as.factor(smokstat), data = df)

                           N Observed Expected (O-E)^2/E (O-E)^2/V
as.factor(smokstat)=1 42812     1431     2256     301.5     580.7
as.factor(smokstat)=2 14170     1227      728     342.5     405.4
as.factor(smokstat)=3  2562      204      132      39.2      40.3
as.factor(smokstat)=4 30345     1831     1577      40.8      61.4

 Chisq= 724  on 3 degrees of freedom, p= 0


Call: survfit(formula = SurvObj ~ as.factor(smokstat), data = df, conf.type
= "log-log")
                           n events median 0.95LCL 0.95UCL
as.factor(smokstat)=1 42812   1431     NA      NA      NA
as.factor(smokstat)=2 14170   1227     NA      NA      NA
as.factor(smokstat)=3  2562    204     NA      NA      NA
as.factor(smokstat)=4 30345   1831     NA      NA      NA
```