

Project1

Selin Karabulut

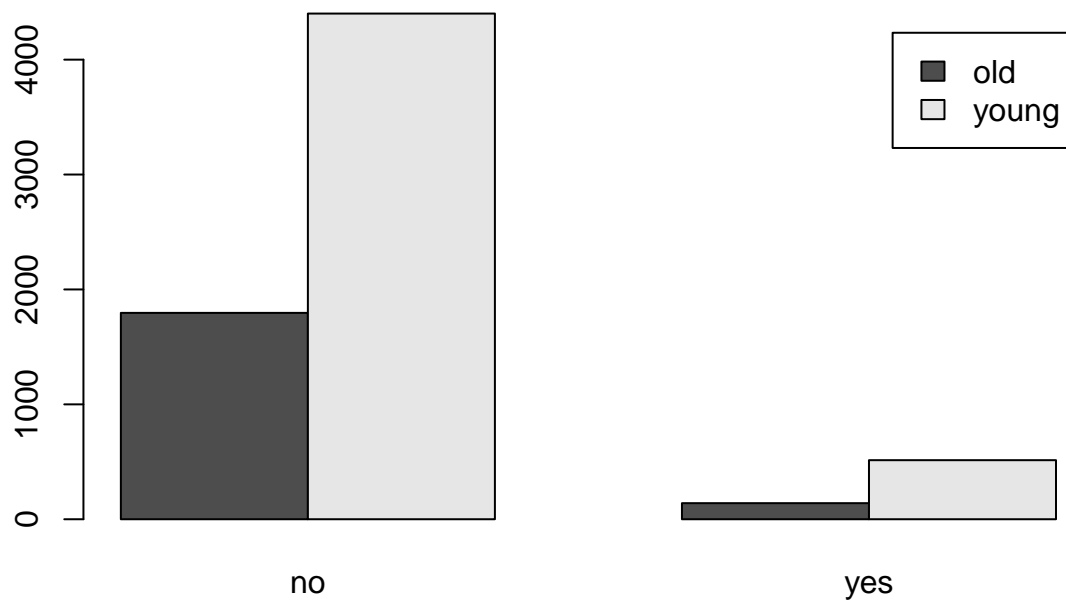
a)

```
TAB = table(dt2$age, dt2$Smoker)
TAB
```

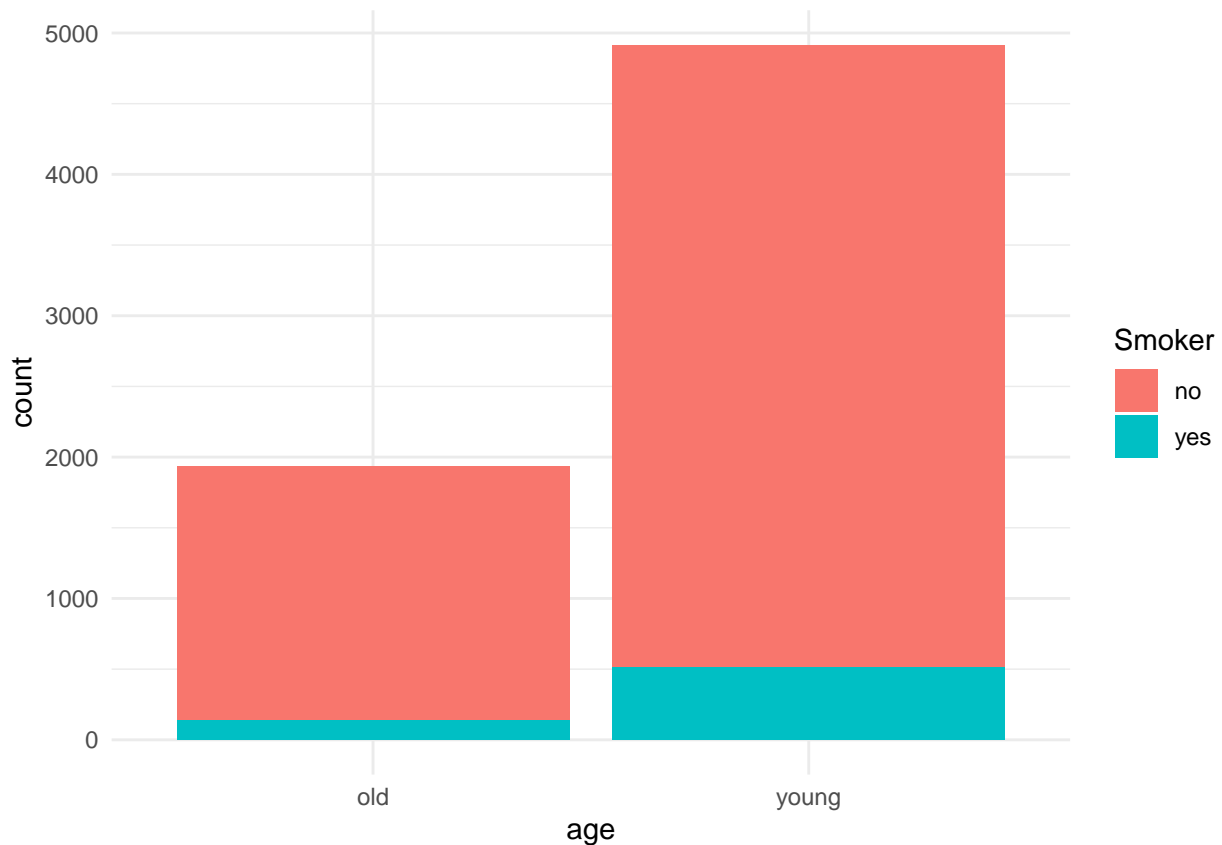
```
##
##           no  yes
##    old  1796  140
##   young 4401  514
```

```
barplot(TAB, main="Mother's age (young/old) vs Smoking habits", beside = T, legend=T)
library(ggplot2)
```

Mother's age (young/old) vs Smoking habits



```
ggplot(dt2) +
  aes(x = age, fill = Smoker) +
  geom_bar() +
  scale_fill_hue() +
  theme_minimal()
```



```
test <- chisq.test(table(dt2$age, dt2$Smoker))
test
```

```
##
## Pearson's Chi-squared test with Yates' continuity correction
##
## data: table(dt2$age, dt2$Smoker)
## X-squared = 16.372, df = 1, p-value = 5.204e-05
```

```
test$p.value
```

```
## [1] 5.204212e-05
```

```
test$expected
```

```
##
##          no      yes
## old  1751.188 184.8116
## young 4445.812 469.1884
```

```
library(vcd)
```

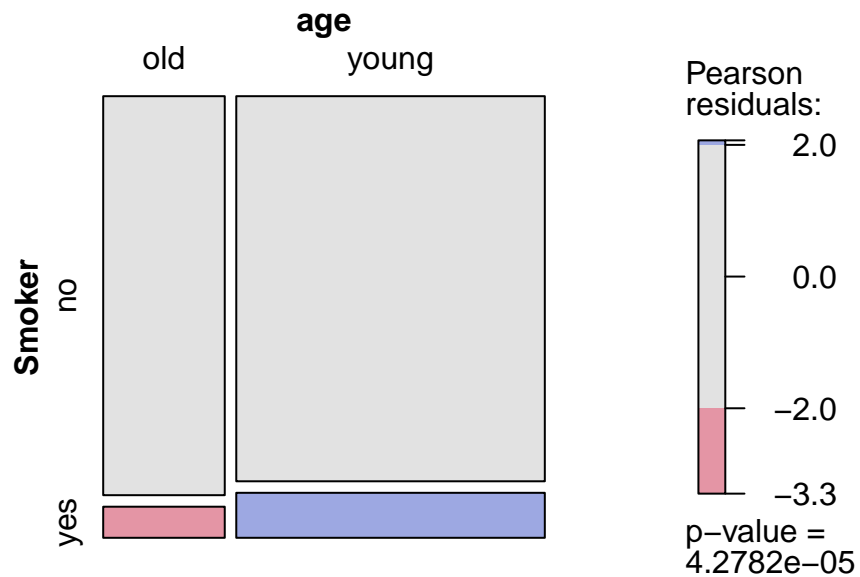
```
## Loading required package: grid
```

```
assocstats(table(dt2$age, dt2$Smoker))
```

```
##              X^2 df    P(> X^2)
## Likelihood Ratio 17.609  1 2.7134e-05
## Pearson          16.744  1 4.2782e-05
##
```

```
## Phi-Coefficient      : 0.049
## Contingency Coeff.: 0.049
## Cramer's V          : 0.049
```

```
library(vcd)
mosaic(~ age + Smoker,
       direction = c("v", "h"),
       data = dt2,
       shade = TRUE)
```

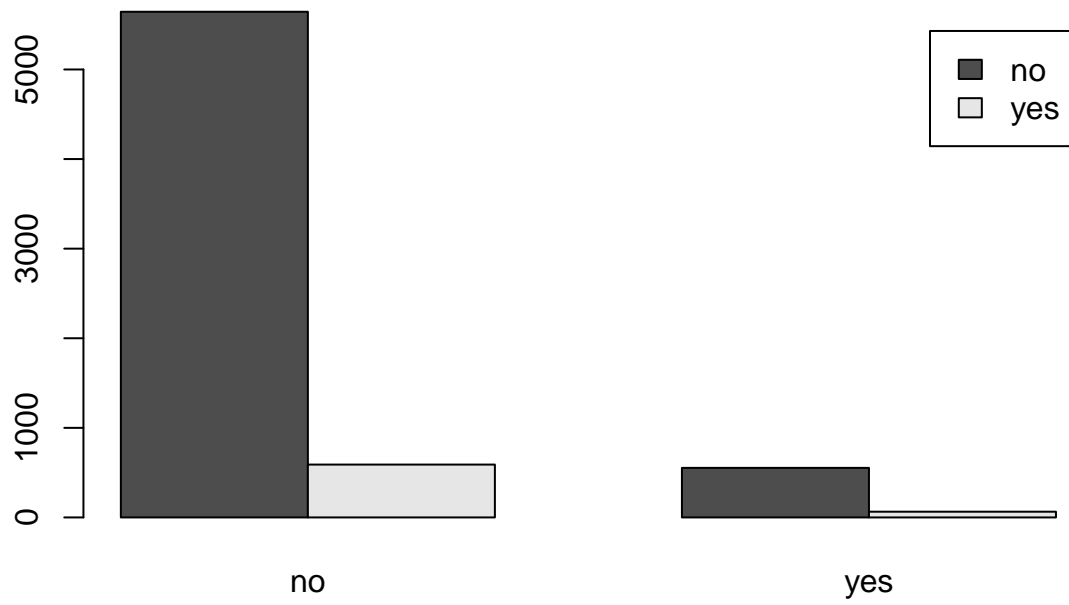


b)

```
TAB1 = table(dt2$Smoker, dt2$premature)
TAB1
```

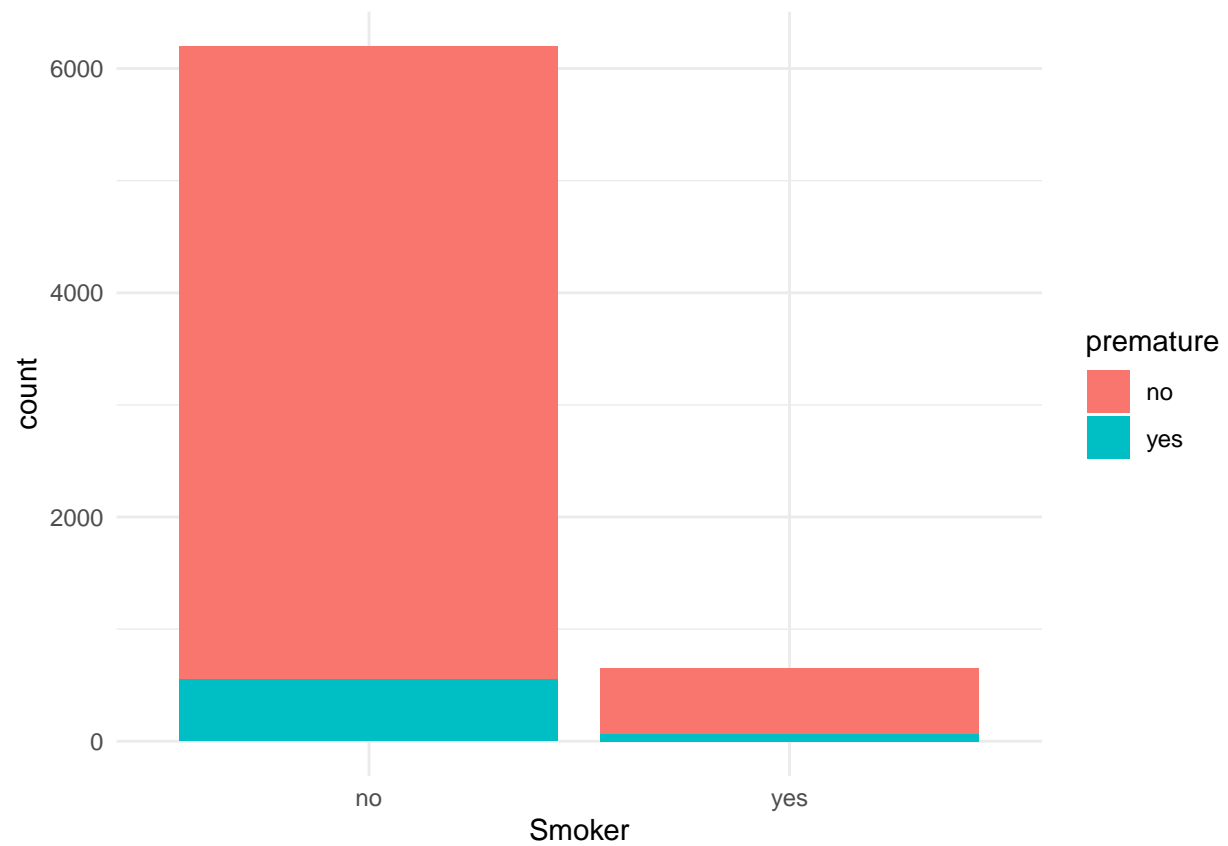
```
##
##      no  yes
## no 5644 553
## yes 590  64
```

```
barplot(TAB1, beside = T, legend=T)
```



```
library(ggplot2)

ggplot(dt2) +
  aes(x = Smoker, fill = premature) +
  geom_bar() +
  scale_fill_hue() +
  theme_minimal()
```



```
test1 <- chisq.test(table(dt2$Smoker, dt2$premature))
test1
```

```
##
## Pearson's Chi-squared test with Yates' continuity correction
##
## data: table(dt2$Smoker, dt2$premature)
## X-squared = 0.43665, df = 1, p-value = 0.5087
```

```
test1$p.value
```

```
## [1] 0.508746
```

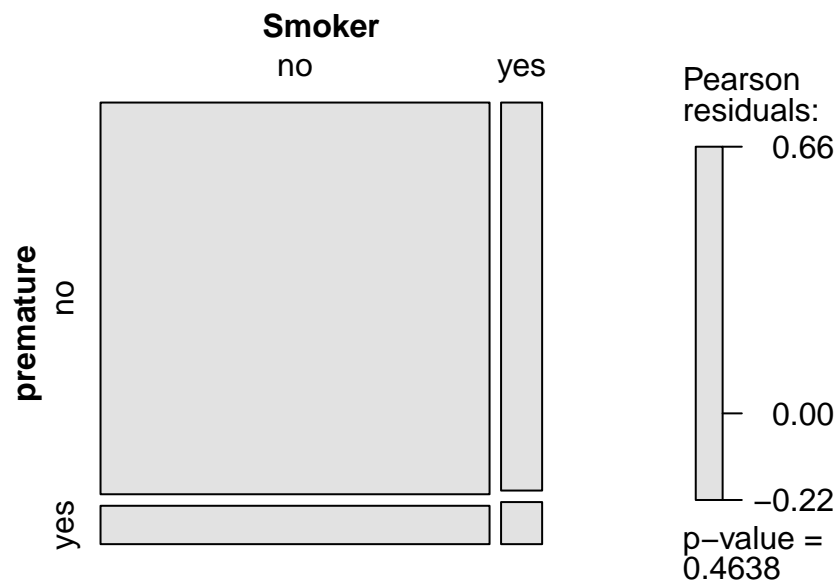
```
test1$expected
```

```
##
##           no      yes
## no 5638.8991 558.10086
## yes 595.1009  58.89914
```

```
library(vcd)
assocstats(table(dt2$Smoker, dt2$premature))
```

```
##           X^2 df P(> X^2)
## Likelihood Ratio 0.52475 1 0.46882
## Pearson          0.53671 1 0.46380
##
## Phi-Coefficient   : 0.009
## Contingency Coeff.: 0.009
## Cramer's V        : 0.009
```

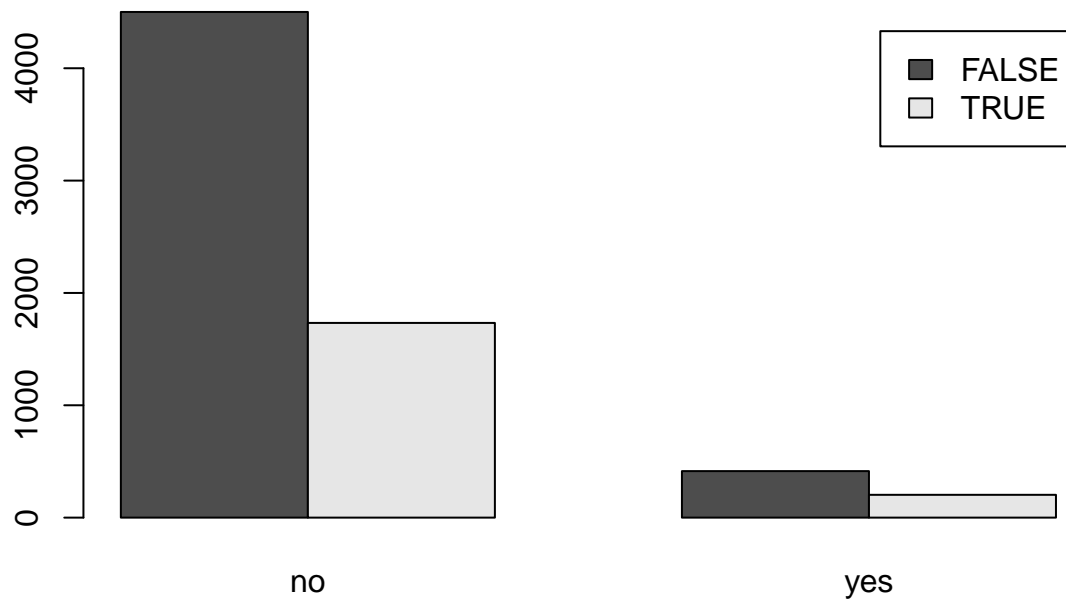
```
library(vcd)
mosaic(~ Smoker + premature,
       direction = c("v", "h"),
       data = dt2,
       shade = TRUE)
```



```
TAB2 = table(dt2$age=="old", dt2$premature)
TAB2
```

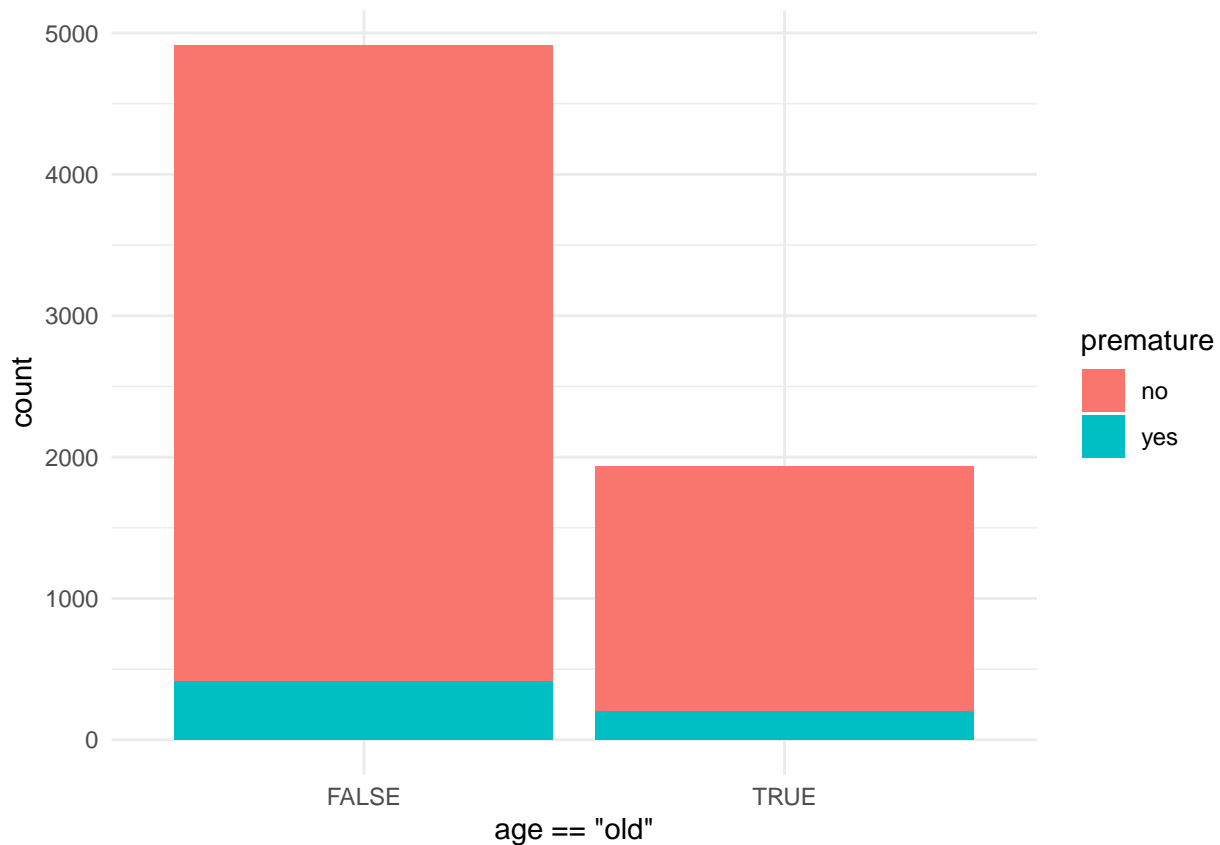
```
##
##           no  yes
## FALSE 4501  414
##  TRUE  1733  203
```

```
barplot(TAB2, beside = T, legend=T)
```



```
library(ggplot2)

ggplot(dt2) +
  aes(x = age=="old", fill = premature) +
  geom_bar() +
  scale_fill_hue() +
  theme_minimal()
```



```
test2 <- chisq.test(table(dt2$age=="old", dt2$premature))
test2
```

```
##
## Pearson's Chi-squared test with Yates' continuity correction
##
## data: table(dt2$age == "old", dt2$premature)
## X-squared = 6.9592, df = 1, p-value = 0.008339
```

```
test2$p.value
```

```
## [1] 0.008339103
```

```
test2$expected
```

```
##
##          no      yes
## FALSE 4472.356 442.6441
## TRUE  1761.644 174.3559
```

```
library(vcd)
assocstats(table(dt2$age=="old", dt2$premature))
```

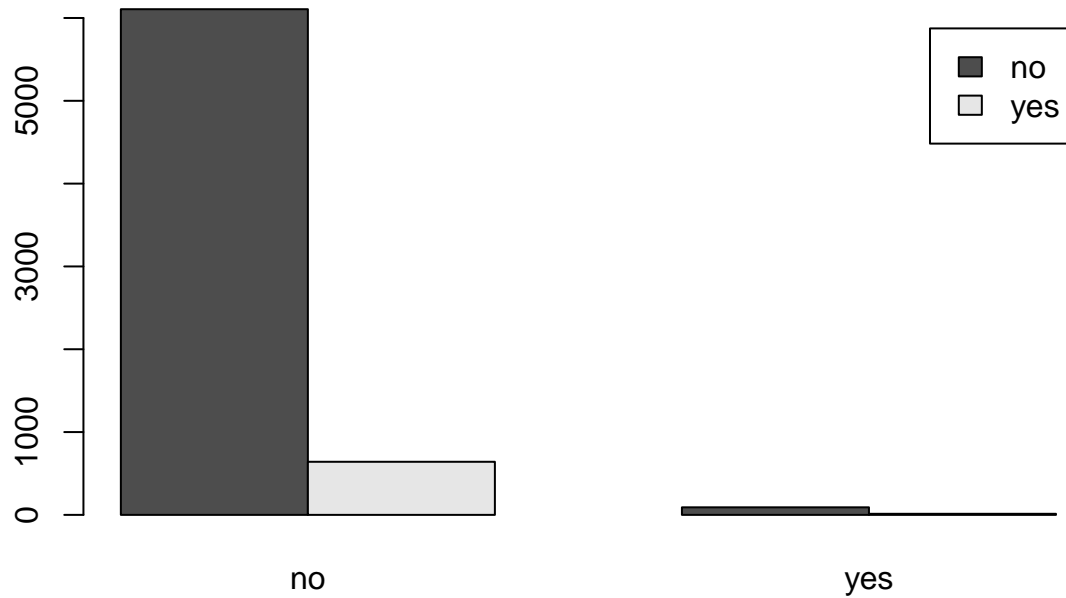
```
##              X^2 df  P(> X^2)
## Likelihood Ratio 7.0137 1 0.0080888
## Pearson          7.2086 1 0.0072554
##
## Phi-Coefficient   : 0.032
## Contingency Coeff.: 0.032
## Cramer's V       : 0.032
```

c)

```
TAB3 = table(dt2$Smoker, dt2$P_d1)
TAB3
```

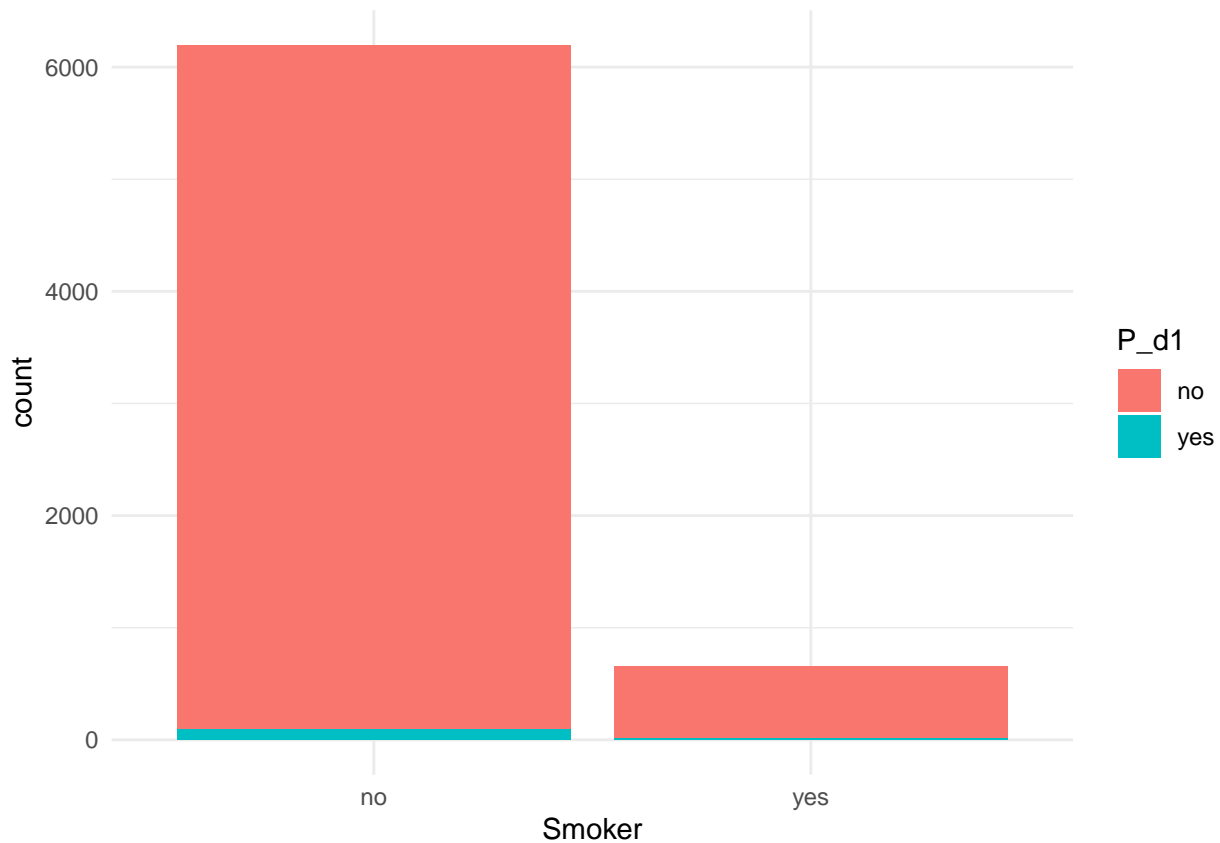
```
##
##      no  yes
## no 6106  91
## yes 641  13
```

```
barplot(TAB3, beside = T, legend=T)
```



```
library(ggplot2)
```

```
ggplot(dt2) +  
  aes(x = Smoker, fill = P_d1) +  
  geom_bar() +  
  scale_fill_hue() +  
  theme_minimal()
```

```
test3 <- chisq.test(table(dt2$Smoker, dt2$P_d1))
test3
```

```
##
## Pearson's Chi-squared test with Yates' continuity correction
##
## data: table(dt2$Smoker, dt2$P_d1)
## X-squared = 0.74806, df = 1, p-value = 0.3871
```

```
test3$p.value
```

```
## [1] 0.3870911
```

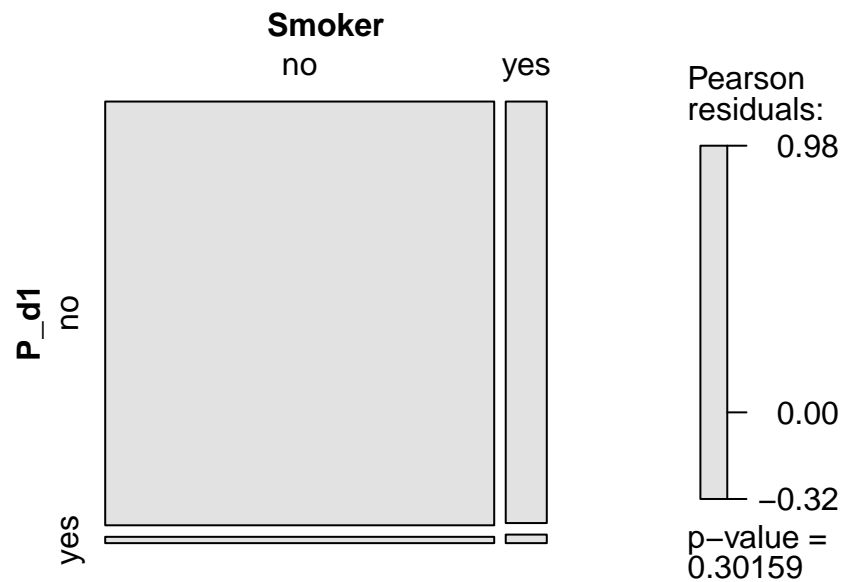
```
test3$expected
```

```
##
##           no      yes
## no  6102.9279  94.072106
## yes   644.0721   9.927894
```

```
library(vcd)
assocstats(table(dt2$Smoker, dt2$P_d1))
```

```
##           X^2 df P(> X^2)
## Likelihood Ratio 0.98307 1 0.32144
## Pearson          1.06716 1 0.30159
##
## Phi-Coefficient   : 0.012
## Contingency Coeff.: 0.012
## Cramer's V        : 0.012
```

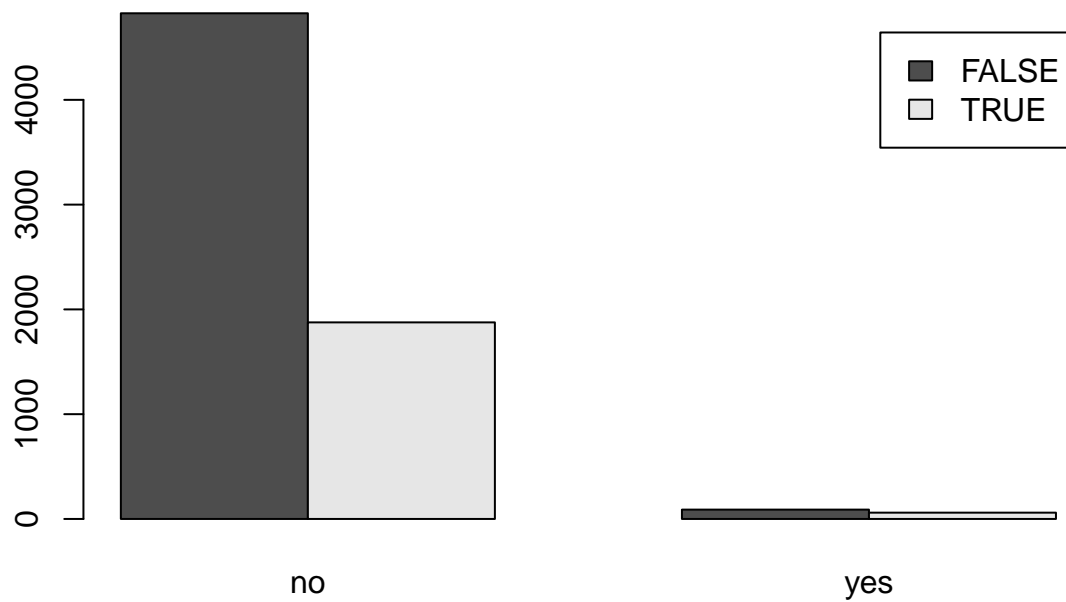
```
library(vcd)
mosaic(~ Smoker + P_d1,
       direction = c("v", "h"),
       data = dt2,
       shade = TRUE)
```



```
TAB4 = table(dt2$age=="old", dt2$d1)
TAB4
```

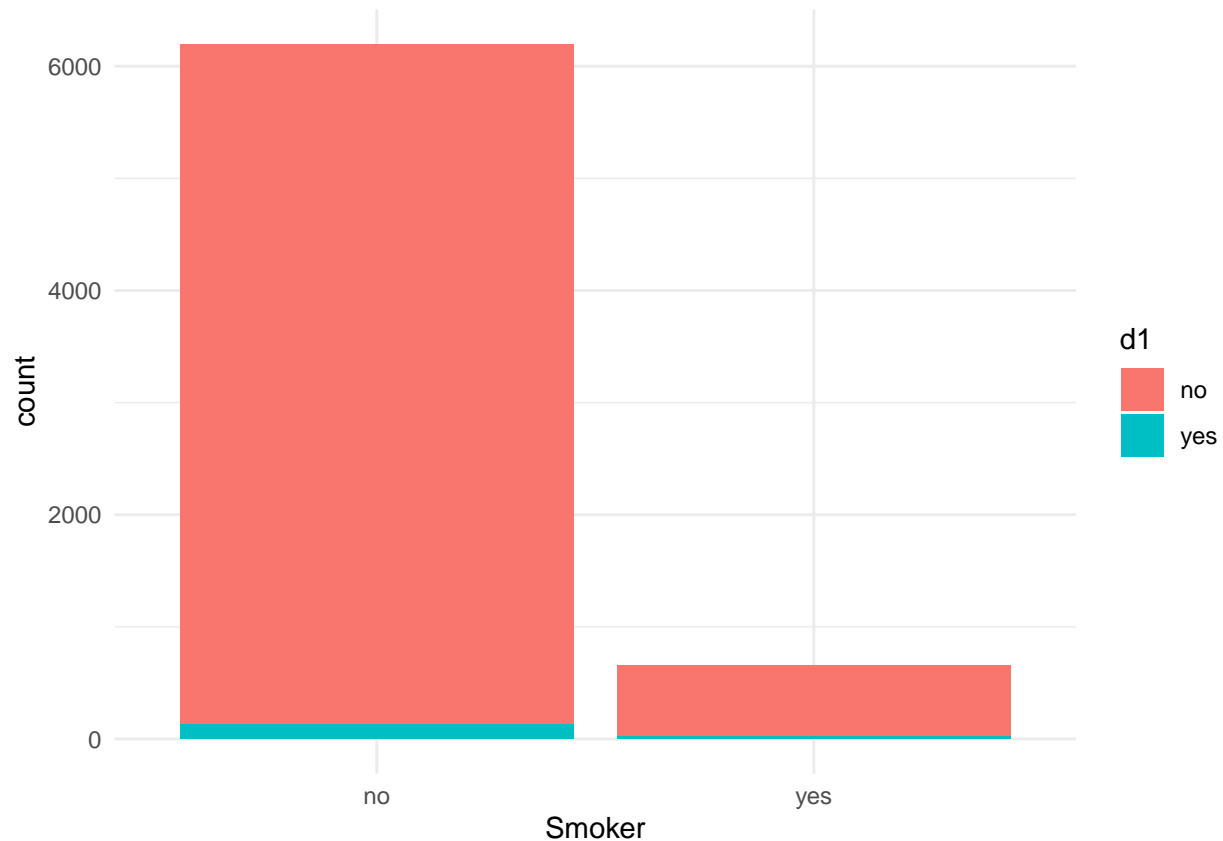
```
##
##      no  yes
## FALSE 4826 89
## TRUE  1876 60
```

```
barplot(TAB4, beside = T, legend=T)
```



```
library(ggplot2)
```

```
ggplot(dt2) +  
  aes(x = Smoker, fill = d1) +  
  geom_bar() +  
  scale_fill_hue() +  
  theme_minimal()
```



```
test4 <- chisq.test(table(dt2$age=="old", dt2$d1))  
test4
```

```
##  
## Pearson's Chi-squared test with Yates' continuity correction  
##  
## data: table(dt2$age == "old", dt2$d1)  
## X-squared = 10.239, df = 1, p-value = 0.001375
```

```
test4$p.value
```

```
## [1] 0.001374768
```

```
test4$expected
```

```
##  
##           no           yes  
## FALSE 4808.105 106.89461  
## TRUE  1893.895  42.10539
```

d) How do age and smoking habits affect survival?

```
dt3 <- read_excel("~/Desktop/FALL 2020/PSTAT220A/Project1/Project1_2.xlsx")
```

```
aa <- lm(dt3$a1 ~ dt3$age + dt3$Smoker)
summary(aa)
```

```
##
## Call:
## lm(formula = dt3$a1 ~ dt3$age + dt3$Smoker)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.97567  0.00057  0.00057  0.02979  0.05355
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.970210   0.002378  408.014 < 2e-16 ***
## dt3$age        0.029223   0.002787   10.485 < 2e-16 ***
## dt3$Smoker    -0.023758   0.004270   -5.563 2.75e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1037 on 6848 degrees of freedom
## Multiple R-squared:  0.0194, Adjusted R-squared:  0.01911
## F-statistic: 67.73 on 2 and 6848 DF,  p-value: < 2.2e-16
```

```
aa_1<- glm(dt3$d1~dt3$age*dt3$Smoker, family=binomial(link=logit))
summary(aa_1)
```

```
##
## Call:
## glm(formula = dt3$d1 ~ dt3$age * dt3$Smoker, family = binomial(link = logit))
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -0.2697  -0.2494  -0.1842  -0.1842   2.8585
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -3.4549     0.1370 -25.227 < 2e-16 ***
## dt3$age        -0.6137     0.1803  -3.404 0.000663 ***
## dt3$Smoker       0.1590     0.4756   0.334 0.738053
## dt3$age:dt3$Smoker  0.4050     0.5555   0.729 0.465992
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##      Null deviance: 1435.5  on 6850  degrees of freedom
## Residual deviance: 1421.9  on 6847  degrees of freedom
## AIC: 1429.9
##
## Number of Fisher Scoring iterations: 6
```

```

ab <- lm(dt3$a1 ~ dt3$age*dt3$Smoker)
summary(ab)

##
## Call:
## lm(formula = dt3$a1 ~ dt3$age * dt3$Smoker)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.97082  0.00000  0.00000  0.03118  0.03571
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    0.968820   0.002447  395.910  <2e-16 ***
## dt3$age        0.031180   0.002904  10.738  <2e-16 ***
## dt3$Smoker     -0.004534   0.009100  -0.498   0.6183
## dt3$age:dt3$Smoker -0.024649  0.010304  -2.392   0.0168 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1037 on 6847 degrees of freedom
## Multiple R-squared:  0.02022,    Adjusted R-squared:  0.01979
## F-statistic: 47.09 on 3 and 6847 DF,  p-value: < 2.2e-16

ac <- glm(dt3$a1 ~ dt3$age*dt3$Smoker, family=binomial(link=logit))
summary(ac)

##
## Call:
## glm(formula = dt3$a1 ~ dt3$age * dt3$Smoker, family = binomial(link = logit))
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -2.65864  0.00003  0.00003  0.25170  0.26969
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)    3.4363    0.1358  25.311  <2e-16 ***
## dt3$age       18.1298   440.6455   0.041   0.967
## dt3$Smoker     -0.1405    0.4752  -0.296   0.768
## dt3$age:dt3$Smoker -17.9211  440.6459  -0.041   0.968
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##      Null deviance: 835.37  on 6850  degrees of freedom
## Residual deviance: 677.37  on 6847  degrees of freedom
## AIC: 685.37
##
## Number of Fisher Scoring iterations: 20

dt3$yhat<- fitted.values(ac)
dt3$u <- residuals(ac)
data <- data.frame(dt3, yhat=fitted(ac), u=resid(ac))

```

```
par(mfrow=c(1,3))

hist(dt3$u, main="Figure 1. Residual distribution")
boxplot(dt3$u, main="Figure 2. Boxplot for residuals")
qqnorm(dt3$u, main="Figure 3. Q-Q plot")
qqline(dt3$u)
```

Figure 1. Residual distribution

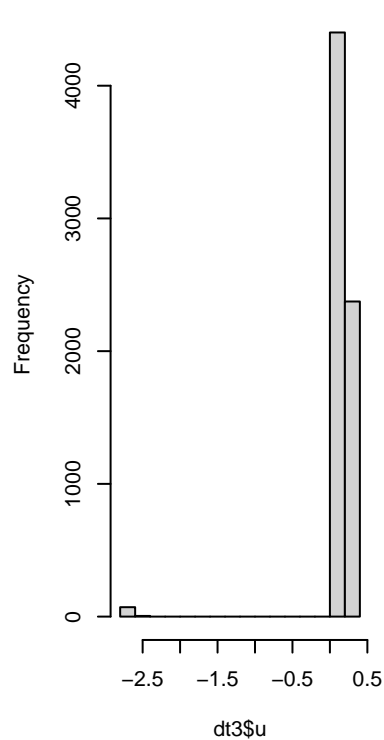


Figure 2. Boxplot for residuals

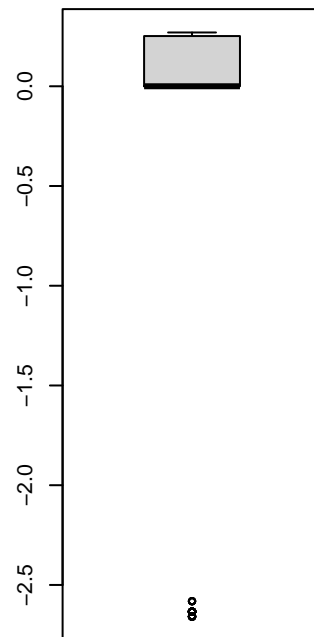
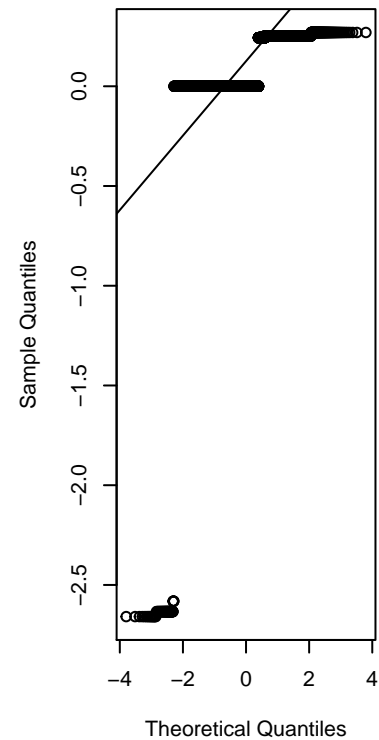


Figure 3. Q-Q plot



```
library("nortest")
ad.test(dt3$u)
```

```
##
## Anderson-Darling normality test
##
## data: dt3$u
## A = 1234.7, p-value < 2.2e-16
```

very small p-value, reject the null hypothesis. these residuals do not come from normal distr test for heteroscedasticity

```
data$resid <- resid(ac)
data$fitted <- fitted(ac)
library("lmtest")
```

```
## Loading required package: zoo
```

```
##
```

```
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
```

```
##
```

```
##      as.Date, as.Date.numeric)
bptest(ac,studentize=FALSE)

##
## Breusch-Pagan test
##
## data:  ac
## BP = 6171.8, df = 3, p-value < 2.2e-16
p value is small we reject the null of homoskedasticity.
test for multicollinearity
library("car")

## Loading required package: carData
vif(ac)

##           dt3$age           dt3$Smoker dt3$age:dt3$Smoker
##      2.268371e+06      3.225279e+00      2.268374e+06
vif<5 then no serious concern for multicollinearity
How does age and smoking habits affect gestational age?
ad <- lm(dt3$premature ~ dt3$age + dt3$Smoker)
summary(ad)

##
## Call:
## lm(formula = dt3$premature ~ dt3$age + dt3$Smoker)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.11433 -0.10412 -0.08316 -0.08316  0.91684
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.104117   0.006559  15.873 < 2e-16 ***
## dt3$age      -0.020953   0.007688  -2.725  0.00644 **
## dt3$Smoker    0.010210   0.011780   0.867  0.38613
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2862 on 6848 degrees of freedom
## Multiple R-squared:  0.001162, Adjusted R-squared:  0.0008701
## F-statistic: 3.983 on 2 and 6848 DF, p-value: 0.01868
ae <- lm(dt3$premature ~ dt3$age*dt3$Smoker)
summary(ae)

##
## Call:
## lm(formula = dt3$premature ~ dt3$age * dt3$Smoker)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.10714 -0.10468 -0.08294 -0.08294  0.91706
```

```
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    0.104677   0.006753  15.501 < 2e-16 ***
## dt3$age        -0.021741   0.008013  -2.713  0.00668 **
## dt3$Smoker      0.002466   0.025112   0.098  0.92178
## dt3$age:dt3$Smoker 0.009929   0.028435   0.349  0.72696
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2862 on 6847 degrees of freedom
## Multiple R-squared:  0.00118,    Adjusted R-squared:  0.0007419
## F-statistic: 2.695 on 3 and 6847 DF,  p-value: 0.04435
af <- glm(dt3$premature ~ dt3$age*dt3$Smoker, family=binomial(link=logit))
summary(af)

##
## Call:
## glm(formula = dt3$premature ~ dt3$age * dt3$Smoker, family = binomial(link = logit))
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -0.4761  -0.4703  -0.4161  -0.4161   2.2315
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -2.14630    0.07708 -27.846 < 2e-16 ***
## dt3$age        -0.25681    0.09449  -2.718  0.00657 **
## dt3$Smoker      0.02604    0.28391   0.092  0.92692
## dt3$age:dt3$Smoker 0.12685    0.32581   0.389  0.69702
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##      Null deviance: 4147.3  on 6850  degrees of freedom
## Residual deviance: 4139.4  on 6847  degrees of freedom
## AIC: 4147.4
##
## Number of Fisher Scoring iterations: 5
```

e) How do age and smoking habits affect survival among premature babies?

```
ba_1 <- lm(dt3$P_d1 ~ dt3$age + dt3$Smoker)
summary(ba_1)

##
## Call:
## lm(formula = dt3$P_d1 ~ dt3$age + dt3$Smoker)
##
```



```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.02886 -0.02281 -0.01137 -0.01137  0.98863
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.022806   0.002801   8.143 4.53e-16 ***
## dt3$age      -0.011435   0.003282  -3.484 0.000497 ***
## dt3$Smoker    0.006059   0.005029   1.205 0.228334
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1222 on 6848 degrees of freedom
## Multiple R-squared:  0.001925, Adjusted R-squared:  0.001633
## F-statistic: 6.603 on 2 and 6848 DF, p-value: 0.001365
```

```
bb_1 <- lm(dt3$P_d1 ~ dt3$age*dt3$Smoker)
summary(bb_1)
```

```
##
## Call:
## lm(formula = dt3$P_d1 ~ dt3$age * dt3$Smoker)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.02857 -0.02283 -0.01136 -0.01136  0.98864
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    0.0228285   0.0028832   7.918 2.8e-15 ***
## dt3$age        -0.0114675   0.0034213  -3.352 0.000807 ***
## dt3$Smoker      0.0057429   0.0107217   0.536 0.592227
## dt3$age:dt3$Smoker 0.0004058   0.0121405   0.033 0.973340
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1222 on 6847 degrees of freedom
## Multiple R-squared:  0.001925, Adjusted R-squared:  0.001488
## F-statistic: 4.402 on 3 and 6847 DF, p-value: 0.004237
```

```
bc_1 <- glm(dt3$P_d1 ~ dt3$age*dt3$Smoker, family=binomial(link=logit))
summary(bc_1)
```

```
##
## Call:
## glm(formula = dt3$P_d1 ~ dt3$age * dt3$Smoker, family = binomial(link = logit))
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -0.2408  -0.2149  -0.1512  -0.1512   2.9925
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -3.7567     0.1580 -23.778 < 2e-16 ***
## dt3$age        -0.7095     0.2126  -3.338 0.000845 ***
```

```
## dt3$Smoker          0.2303      0.5313    0.433 0.664707
## dt3$age:dt3$Smoker  0.2085      0.6447    0.323 0.746372
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##      Null deviance: 1077.5  on 6850  degrees of freedom
## Residual deviance: 1065.2  on 6847  degrees of freedom
## AIC: 1073.2
##
## Number of Fisher Scoring iterations: 7
```

f) How do smoking habits affect survival among premature babies with young mothers?

```
dt4 <- read_excel("~/Desktop/FALL 2020/PSTAT220A/Project1/Book4.xlsx")
```

```
ca <- lm(dt4$P_a1~dt4$Smoker)
summary(ca)
```

```
##
## Call:
## lm(formula = dt4$P_a1 ~ dt4$Smoker)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.07782 -0.07157 -0.07157 -0.07157  0.92843
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.071575   0.003903  18.339  <2e-16 ***
## dt4$Smoker   0.006246   0.012069   0.518   0.605
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2589 on 4913 degrees of freedom
## Multiple R-squared:  5.452e-05, Adjusted R-squared:  -0.000149
## F-statistic: 0.2679 on 1 and 4913 DF,  p-value: 0.6048
ca_1 <- lm(dt4$P_d1~dt4$Smoker)
summary(ca_1)
```

```
##
## Call:
## lm(formula = dt4$P_d1 ~ dt4$Smoker)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.01751 -0.01136 -0.01136 -0.01136  0.98864
##
```

```
## Coefficients:
##           Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.011361   0.001642   6.920 5.08e-12 ***
## dt4$Smoker  0.006149   0.005077   1.211   0.226
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1089 on 4913 degrees of freedom
## Multiple R-squared:  0.0002985, Adjusted R-squared:  9.502e-05
## F-statistic: 1.467 on 1 and 4913 DF, p-value: 0.2259

cb <- glm(dt4$P_a1 ~ dt4$Smoker, family=binomial(link=logit))
summary(cb)

##
## Call:
## glm(formula = dt4$P_a1 ~ dt4$Smoker, family = binomial(link = logit))
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -0.4025  -0.3854  -0.3854  -0.3854   2.2965
##
## Coefficients:
##           Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.56275    0.05848 -43.826  <2e-16 ***
## dt4$Smoker   0.09042    0.17473   0.518   0.605
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##      Null deviance: 2549.5  on 4914  degrees of freedom
## Residual deviance: 2549.3  on 4913  degrees of freedom
## AIC: 2553.3
##
## Number of Fisher Scoring iterations: 5

cb_1 <- glm(dt4$P_d1 ~ dt4$Smoker, family=binomial(link=logit))
summary(cb_1)

##
## Call:
## glm(formula = dt4$P_d1 ~ dt4$Smoker, family = binomial(link = logit))
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -0.1880  -0.1512  -0.1512  -0.1512   2.9925
##
## Coefficients:
##           Estimate Std. Error z value Pr(>|z|)
## (Intercept) -4.4661    0.1422 -31.401  <2e-16 ***
## dt4$Smoker   0.4388    0.3651   1.202   0.229
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## (Dispersion parameter for binomial family taken to be 1)
##
## Null deviance: 639.15 on 4914 degrees of freedom
## Residual deviance: 637.84 on 4913 degrees of freedom
## AIC: 641.84
##
## Number of Fisher Scoring iterations: 7
```

How do smoking habits affect survival among premature babies with older mothers?

```
dt5 <- read_excel("~/Desktop/FALL 2020/PSTAT220A/Project1/Book5.xlsx")
cc <- lm(dt5$P_a1~dt5$Smoker)
summary(cc)
```

```
##
## Call:
## lm(formula = dt5$P_a1 ~ dt5$Smoker)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.08185 -0.08185 -0.08185 -0.08185  0.92143
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.081849   0.006463  12.663  <2e-16 ***
## dt5$Smoker  -0.003277   0.024035  -0.136    0.892
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2739 on 1934 degrees of freedom
## Multiple R-squared:  9.612e-06, Adjusted R-squared:  -0.0005074
## F-statistic: 0.01859 on 1 and 1934 DF, p-value: 0.8916
```

```
cc_1 <- lm(dt5$P_d1~dt5$Smoker)
summary(cc_1)
```

```
##
## Call:
## lm(formula = dt5$P_d1 ~ dt5$Smoker)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.02857 -0.02283 -0.02283 -0.02283  0.97717
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.022829   0.003557   6.418 1.74e-10 ***
## dt5$Smoker   0.005743   0.013228   0.434   0.664
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1507 on 1934 degrees of freedom
## Multiple R-squared:  9.745e-05, Adjusted R-squared:  -0.0004196
## F-statistic: 0.1885 on 1 and 1934 DF, p-value: 0.6642
```

```
cd <- glm(dt5$P_a1 ~ dt5$Smoker, family=binomial(link=logit))
summary(cd)
```

```
##
## Call:
## glm(formula = dt5$P_a1 ~ dt5$Smoker, family = binomial(link = logit))
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -0.4133  -0.4133  -0.4133  -0.4133   2.2555
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -2.41749    0.08608  -28.085  <2e-16 ***
## dt5$Smoker   -0.04443    0.32568   -0.136    0.891
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##      Null deviance: 1094.6  on 1935  degrees of freedom
## Residual deviance: 1094.5  on 1934  degrees of freedom
## AIC: 1098.5
##
## Number of Fisher Scoring iterations: 5
```

```
cd_1 <- glm(dt5$P_d1 ~ dt5$Smoker, family=binomial(link=logit))
summary(cd_1)
```

```
##
## Call:
## glm(formula = dt5$P_d1 ~ dt5$Smoker, family = binomial(link = logit))
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -0.2408  -0.2149  -0.2149  -0.2149   2.7494
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -3.7567    0.1580  -23.779  <2e-16 ***
## dt5$Smoker    0.2303    0.5313   0.433    0.665
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##      Null deviance: 427.50  on 1935  degrees of freedom
## Residual deviance: 427.32  on 1934  degrees of freedom
## AIC: 431.32
##
## Number of Fisher Scoring iterations: 6
```