# Logistic Regression

We will explore data about smoking and mortality. These data are available from Appleton et. al (The American Satistician, 1996). These data were collectected in 1972 and 20 years later. They consist of a subset of a random sample of women in New Castle, UK. In 1972 these women were either current smokers or never smokers. We have their age in 1972 and whether or not they were alive in 1992.

We want to study the effect of smoking on mortality.

## Exploratory Data Analysis

```
sdata = read.csv("~/Dropbox/Stat151/2014Fall/Data/Kaplan/whickham.csv")
```

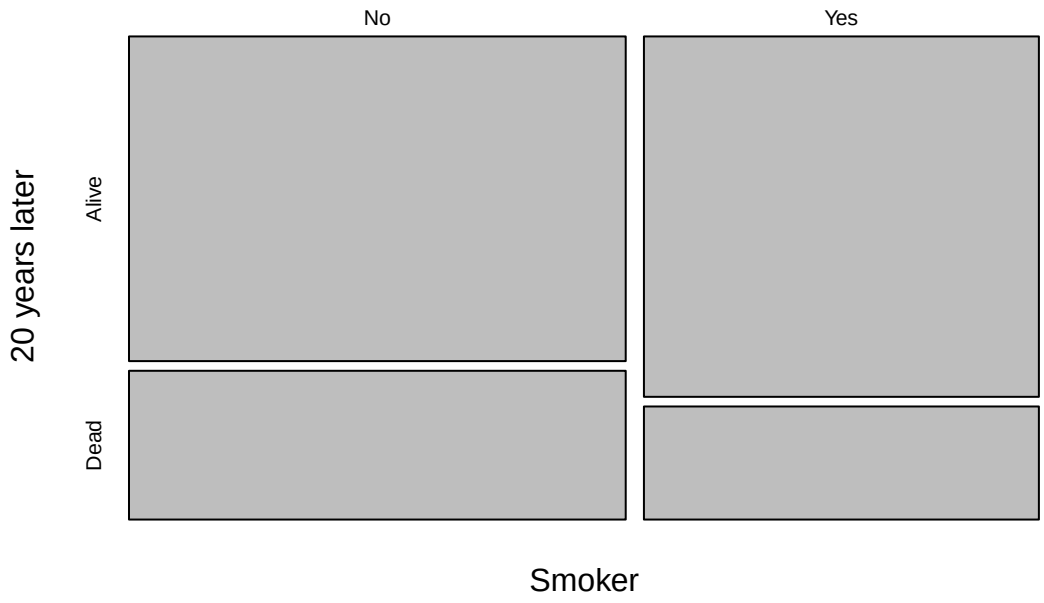
```
summary(sdata)
```

```
##   outcome    smoker          age
##  Alive:945   No :732   Min.   :18.0
##  Dead :369   Yes:582   1st Qu.:32.0
##                        Median :46.0
##                        Mean   :46.9
##                        3rd Qu.:61.0
##                        Max.   :84.0
```

```
table(sdata$outcome, sdata$smoker)
```

```
##
##          No Yes
##   Alive 502 443
##   Dead  230 139
```

```
mosaicplot(table(sdata$smoker, sdata$outcome),
           main = "", ylab = "20 years later", xlab = "Smoker")
```




What do you notice in the mosaic plot?

Let's look at the age distributions for each of the outcome x smoker combinations.

```
with(sdata, boxplot(age ~ smoker + outcome))
```


What do you notice here?

Another way to look at the relationship between age and mortality. We convert the variable outcome into a dummy variable with

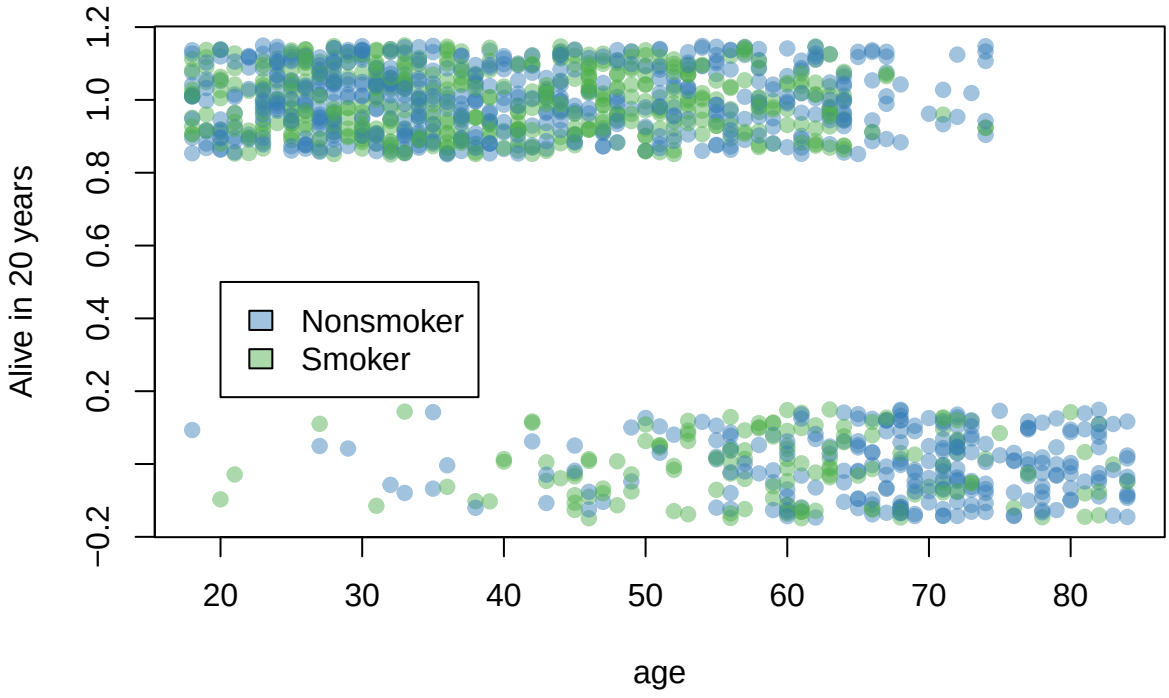
```
Dout = 0 + (sdata$outcome == "Alive")
```

Then, we jitter the value for outcome because we have 1300 observations and we want to spread them out a bit to avoid overplotting. We plot each observation with a colored dot, where the color represents smoking status -

```
require(RColorBrewer)
```

```
## Loading required package: RColorBrewer
```

```
tcol = paste( brewer.pal(9, "Set1")[ 2:3], "76", sep = "")
plot(jitter(Dout, 0.75) ~ age, data  = sdata,
      ylab = "Alive in 20 years",
     col = tcol[smoker], pch = 19)
legend(x = 20, y = 0.5, legend = c("Nonsmoker", "Smoker"), fill = tcol)
```




## Modeling outcome as a function of age and smoking status

We consider three approaches to modeling outcome. The first is a nonparametric averaging approach. The second is simple linear regression, and the third is logistic regression. We discuss each in turn.

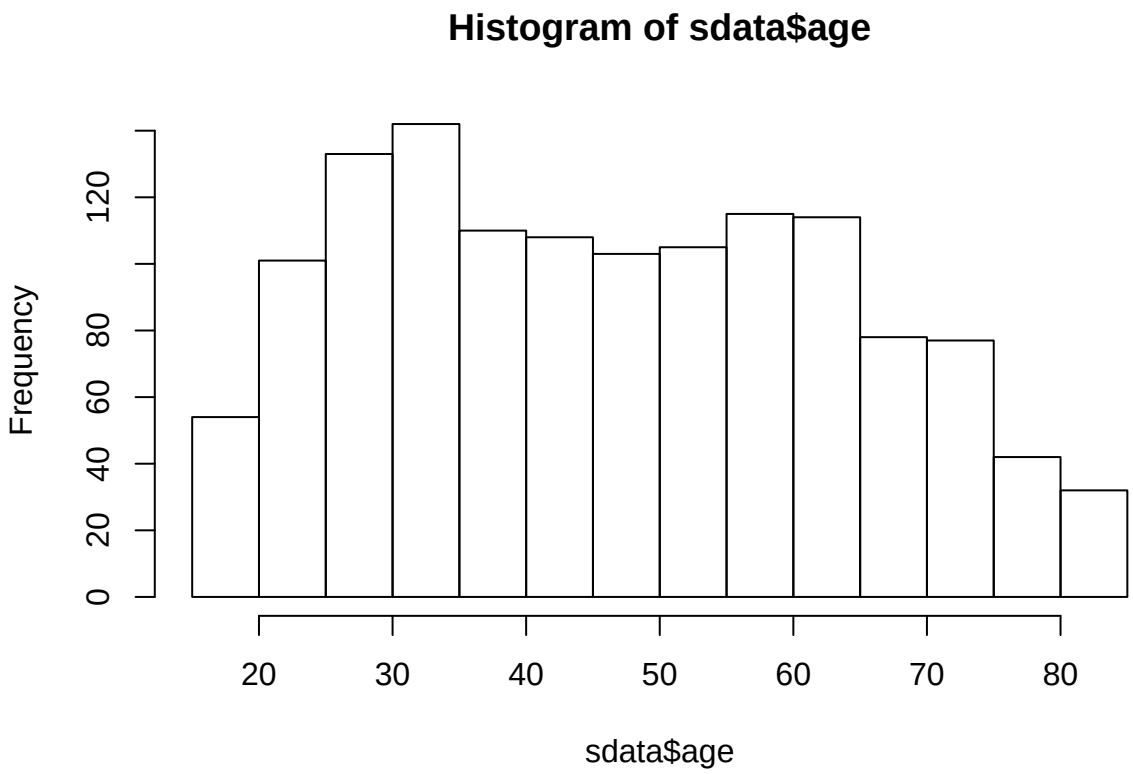
### Local averaging

Recall that the simple linear model was motivated by the conditional averaage of the y-values for a given x-value. We can do this averaging here too.

We break the age variable up into categories and find the proportion of those in the same age category who are alive 20 years later. We
want to estimate this proportion separately for smokers and nonsmokers.

We start by examining the age distribution to determine the cut points.

```
hist(sdata$age)
```




```
ageCat = cut(sdata$age, breaks = c(15, 25, 35, 45, 55, 65, 75, 85))
```

Now we can tally the number of women in each of the categories with

```
tabs = table(sdata$outcome, ageCat, sdata$smoker)
tabs
```

```
## , ,  = No
##
##        ageCat
##         (15,25] (25,35] (35,45] (45,55] (55,65] (65,75] (75,85]
##   Alive      85     151      99      70      72      25       0
##   Dead        1       6       7      15      46      93      62
##
## , ,  = Yes
##
##        ageCat
##         (15,25] (25,35] (35,45] (45,55] (55,65] (65,75] (75,85]
##   Alive      67     115      97      98      60       6       0
##   Dead        2       3      15      25      51      31      12
```

```
tots =  apply(tabs, c(2, 3), sum)
prop = tabs[1, ,] / tots
prop
```

```
##
## ageCat        No    Yes
```

```
##   (15,25] 0.9884 0.9710
##   (25,35] 0.9618 0.9746
##   (35,45] 0.9340 0.8661
##   (45,55] 0.8235 0.7967
##   (55,65] 0.6102 0.5405
##   (65,75] 0.2119 0.1622
##   (75,85] 0.0000 0.0000
```

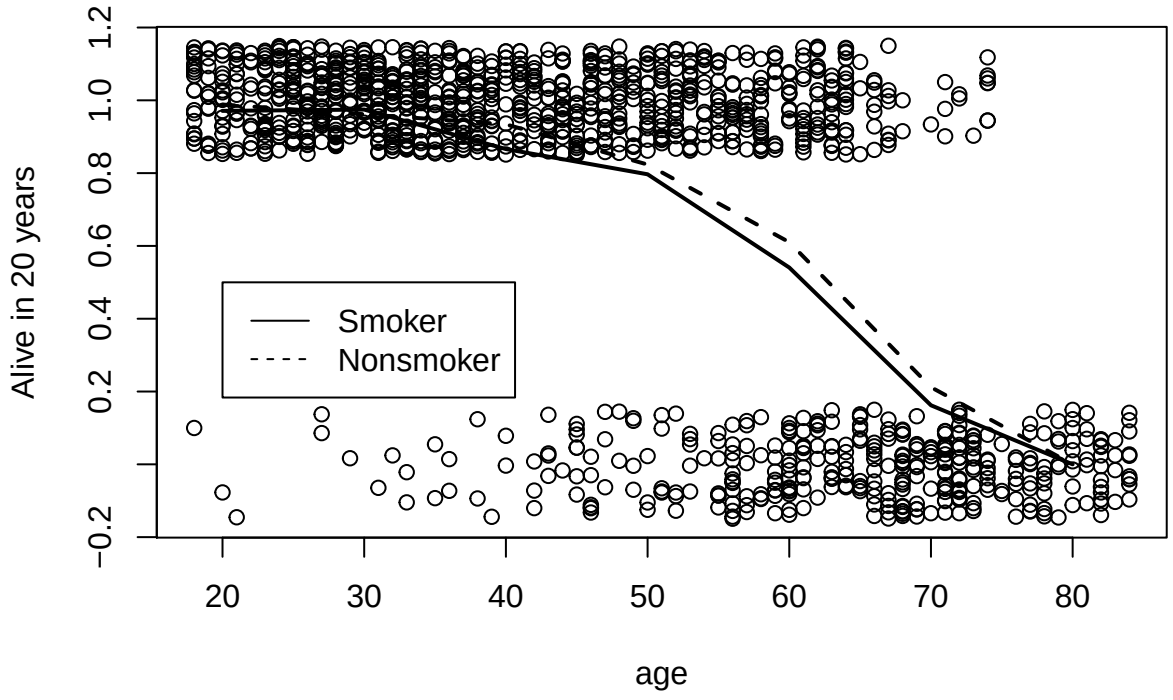
Notice that for a woman who is 35-45 years old at the start of the survey, we would estimate the chance that she is alive 20 years later is 0.93 if she is not a smoker and 0.87 if she is a smoker.

Make a plot with these proportions

```
ageMid = seq(20, 80, by = 10)

plot(jitter(Dout, 0.75) ~ age, data  = sdata,
     ylab = "Alive in 20 years")
points(x = ageMid, y = prop[ , 2], type = "l", lwd = 2, lty = 1)
points(x = ageMid, y = prop[ , 1], type = "l", lwd = 2, lty = 2)

legend(x = 20, y = 0.5, legend = c("Smoker", "Nonsmoker"), lty = 1:2)
```




**Least squares**

We can also fit age and smoking status to the response using least squares.

Then, we regress Dout on age and smoking status

```
lm.out = lm(Dout ~ age + smoker, data = sdata)
summary(lm.out)
```

```
##
## Call:
## lm(formula = Dout ~ age + smoker, data = sdata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.1818 -0.1922  0.0178  0.2601  0.7229
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.472554   0.030102   48.92   <2e-16 ***
## age         -0.016155   0.000558  -28.95   <2e-16 ***
## smokerYes    0.010474   0.019577    0.54     0.59
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.35 on 1311 degrees of freedom
## Multiple R-squared:  0.394,  Adjusted R-squared:  0.393
## F-statistic:  427 on 2 and 1311 DF,  p-value: <2e-16
```

How do we interpret this model? Suppose, someone is 40 years old at the start of the survey, then we estimate the probability that she is alive at age 60 to be

$1.47 - 0.016 * 40 = 0.83$ if she is not a smoker and $0.83 + 0.01 = 0.84$ if she is a smoker. The predictions are probabilities, which is why this type of modeling is called linear probability modeling.

The predictions from the linear probabilty model are quite different from the estimates we got with the local averaging. And, they are quite unsatisfactory when we make estimates for ages near the extremes of the age values. For example for a 20 year old and for a 70 year old the estimates are

```
predict(lm.out, newdata = data.frame(age = c(20,70),
                                     smoker = c("No", "No")))
```

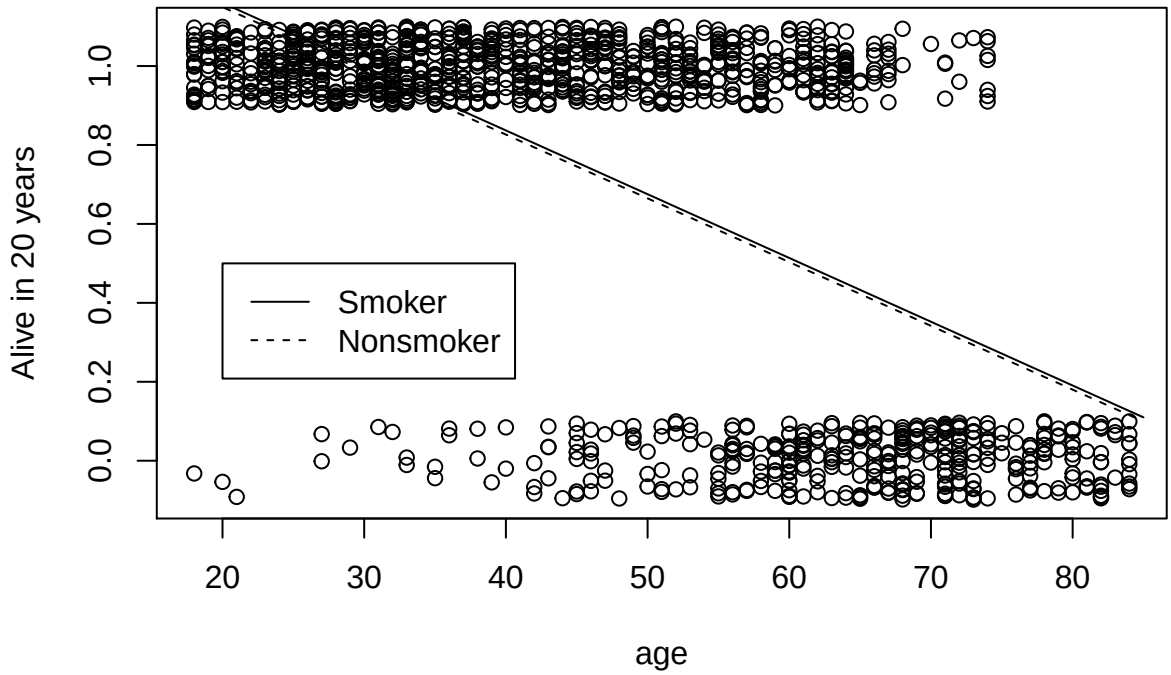
```
##      1      2
## 1.1494 0.3417
```

The propobability is greater than 1 for the 20-year old! And, the 70-year old has better than a 1 in 3 chance of surviving to 90! (Recall these data were collected in 1972). With some linear probability models, negative values for the probability can be produced.

We examine a plot of the data with the fitted lines superposed.

```
plot( jitter(Dout, 0.5) ~ age, data  = sdata,
      ylab = "Alive in 20 years")
points(x = 15:85, predict(lm.out, data.frame(age = 15:85,
                                             smoker = rep("Yes", 71))),
       type = "l", lty = 1)
points(x = 15:85, predict(lm.out, data.frame(age = 15:85,
                                             smoker = rep("No", 71))),
       type = "l", lty = 2)

legend(x = 20, y = 0.5, legend = c("Smoker", "Nonsmoker"), lty = 1:2)
```




Why is this happening? When we compare the shape of the local smoothed curve to the line, we see that they are quite different in shape. The line is not able to reflect that most of the 40 - 50 year olds in the study were still alive 20 years later without over estimating the chance that a 20-year old is still alive.

We might argue that this is not a problem, if the chance is above 1, we can treat it as 1. However, we can't tell the influence of the 20-year olds on the line, i.e., how much it has been pulled down to fit to the 20-year olds and therefor underestimate the chance for the 40-year olds. Actually, our local smooth gives us an idea of the extent of the problem. There we saw that 40-year olds had a chance of 0.93 if she is not a smoker and 0.87 so the impact is worse for the non-smoker.

## Logistic Regression

Our observations about the linear probabiliy fit indicate that we need a mapping from the linear predictor into [0, 1]. That is, we would like a function $P$ such that

$$\pi_i = P(\alpha + \beta x_i)$$

And, we would like $P$ to be invertible. One good choice is the logistic function

$$P(z) = \frac{1}{1 + \exp(-z)}$$

which implies,

$$\pi_i = \frac{1}{1 + \exp(-(\alpha + \beta x_i))}$$

.

This function has a special advantage: the inverse has a nice interpretation:

$$\log(\frac{\pi_i}{1 - \pi_i}) = \alpha + \beta x_i$$

In other words,

$$\frac{\pi_i}{1 - \pi_i} = \exp(\alpha + \beta x_i)$$

This inverse is called the logit or log odds.

We can fit this model with logistic regression, via the glm function.

```
logit.out = glm(outcome == "Alive"~ age + smoker, data = sdata,
              family = "binomial")
logit.out
```

```
##
## Call:  glm(formula = outcome == "Alive" ~ age + smoker, family = "binomial",
##     data = sdata)
##
## Coefficients:
## (Intercept)          age    smokerYes
##       7.599       -0.124       -0.205
##
## Degrees of Freedom: 1313 Total (i.e. Null);  1311 Residual
## Null Deviance:       1560
## Residual Deviance: 945   AIC: 951
```

We can interpret the effect of smoking as follows. The estimate for the odds is

$$\frac{\hat{\pi}_i}{1-\hat{\pi}_i} = \exp -7.6 - 0.12age - 0.20smokerYes$$

For an individual, who does not smoke this reduces to

$$\frac{\hat{\pi}_i}{1-\hat{\pi}_i} = \exp -7.6 - 0.12age$$

and so we see that for for an individual who smokes, we have

$$\frac{\hat{\pi}_i}{1-\hat{\pi}_i} = \exp -7.6 - 0.12age \times \exp(-0.20)$$

Since $\exp(-0.20) = 0.82$ we see that the odds being alive decreases be a factor of 0.82 for smokers.

We can use the same functions such as fitted() and predict() with glm objects. However, predict() has more options, it can predict the response, i.e., the probabilities, or the log-odds, i.e., the logit (which is the "link" option).

```
head(fitted(logit.out))
```

```
##      1      2      3      4      5      6
## 0.9895 0.9943 0.1999 0.3346 0.4215 0.9367
```

```
head(predict(logit.out, type = "response"))
```

```
##      1      2      3      4      5      6
## 0.9895 0.9943 0.1999 0.3346 0.4215 0.9367
```

```
head(predict(logit.out, type = "link"))
```

```
##       1       2       3       4       5       6
##  4.5498  5.1682 -1.3870 -0.6876 -0.3165  2.6946
```

Similarly, there is an analogy to the sums of squares. It uses the concept of residual deviance, which is $-2 \log L$, i.e., it is based on the log-likelihood ratio.

```
logit.int = glm(outcome == "Alive"~ age + smoker + age:smoker,
                data = sdata, family = "binomial")
summary(logit.int)
```

```
##
## Call:
## glm(formula = outcome == "Alive" ~ age + smoker + age:smoker,
##     family = "binomial", data = sdata)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -3.398  -0.426   0.216   0.560   1.928
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)    8.16923    0.60660   13.47   <2e-16 ***
## age           -0.13323    0.00995  -13.39   <2e-16 ***
## smokerYes     -1.45784    0.83723   -1.74    0.082 .
## age:smokerYes  0.02223    0.01449    1.53    0.125
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1560.32  on 1313  degrees of freedom
## Residual deviance:  942.68  on 1310  degrees of freedom
## AIC: 950.7
##
## Number of Fisher Scoring iterations: 6
```

```
logit.noS = glm(outcome == "Alive"~ age, data = sdata,
                family = "binomial")
summary(logit.noS)
```

```
##
## Call:
## glm(formula = outcome == "Alive" ~ age, family = "binomial",
##     data = sdata)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -3.230  -0.428   0.229   0.554   1.895
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  7.40313    0.40352    18.4   <2e-16 ***
## age         -0.12186    0.00694   -17.6   <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: 1560.32  on 1313  degrees of freedom
## Residual deviance:  946.51  on 1312  degrees of freedom
## AIC: 950.5
##
## Number of Fisher Scoring iterations: 6
```

```r
anova(logit.int, logit.noS, test= "Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: outcome == "Alive" ~ age + smoker + age:smoker
## Model 2: outcome == "Alive" ~ age
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1      1310        943
## 2      1312        947 -2    -3.83     0.15
```

We see that there is not strong evidence in these data that smoking significantly reduces the odds of survival.