BST 210 Lab: Week 7
Logistic Regression

In public health in particular, we are often interested in drawing conclusions about a binary response variable Y as a function of several predictor variables $X_1, X_2, \ldots, X_p$. However, our standard linear regression approach—in which we would model $p = E[Y|X] = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p$—isn't sufficient:

- Our predicted probabilities $\hat{p}$ may be greater than one or less than zero
- $Var(Y|X) = p(1-p) = E[Y|X](1 - E[Y|X])$, which violates the equal variance assumption

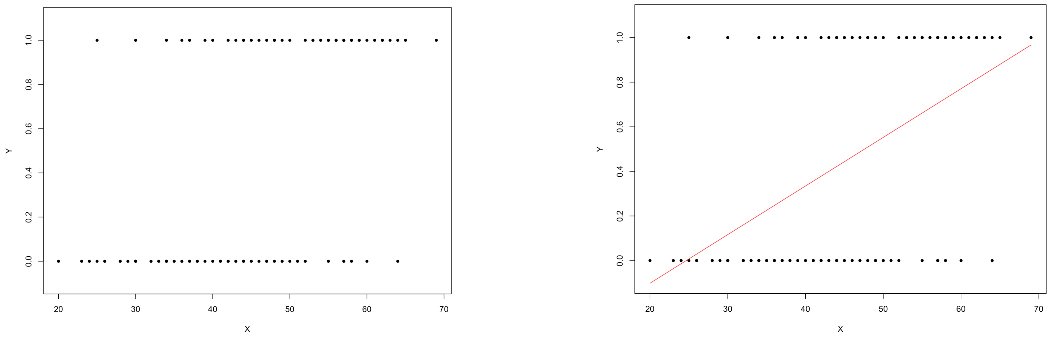


Figure 1: Observed outcomes Y as a function of X (left) with fitted linear regression model overlaid (right).

Logistic regression addresses these issues by transforming our outcome. Instead of modeling $p = E[Y|X]$ as a linear function of $X_1, \ldots, X_p$, we use $\text{logit}(p) = \log\left(\frac{p}{1-p}\right)$. So the model we fit is:

$$\log\Big(\frac{p}{1-p}\Big) = \beta_0 + \beta_1 \cdot X_1 + \ldots + \beta_p \cdot X_p \qquad \Longrightarrow \qquad p = \frac{\exp\{\beta_0 + \beta_1 \cdot X_1 + \ldots + \beta_p \cdot X_p\}}{1 + \exp\{\beta_0 + \beta_1 \cdot X_1 + \ldots + \beta_p \cdot X_p\}}.$$

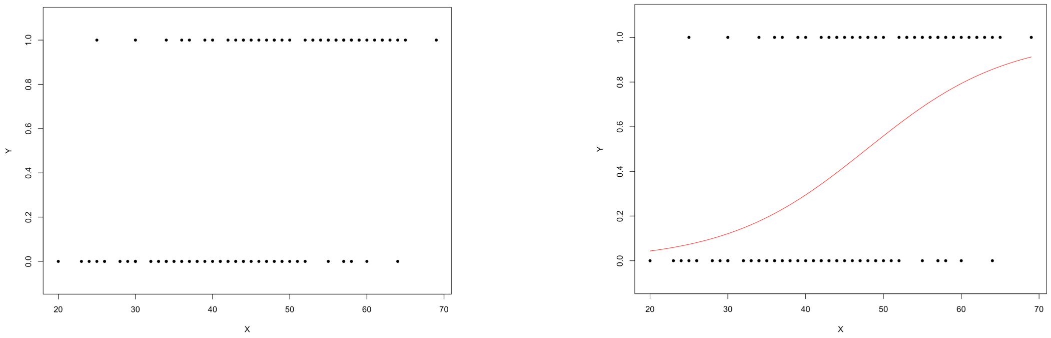


Figure 2: Observed outcomes Y as a function of X (left) with fitted logistic regression model overlaid (right).

Previously On. . . Contingency Tables

Last week in lab, we began looking at data assessing the relationship between parental and student smoking habits (taken from Agresti 1990). Our outcome of interest was whether or not a student smoked (Y) modeled as a function of whether or not at least one of their parents smoked (X):

	At Least One Parent Smokes	Neither Parent Smokes	Total
Student Smokes	816	188	1004
Student Does Not Smoke	3203	1168	4371
Total	4019	1356	5375

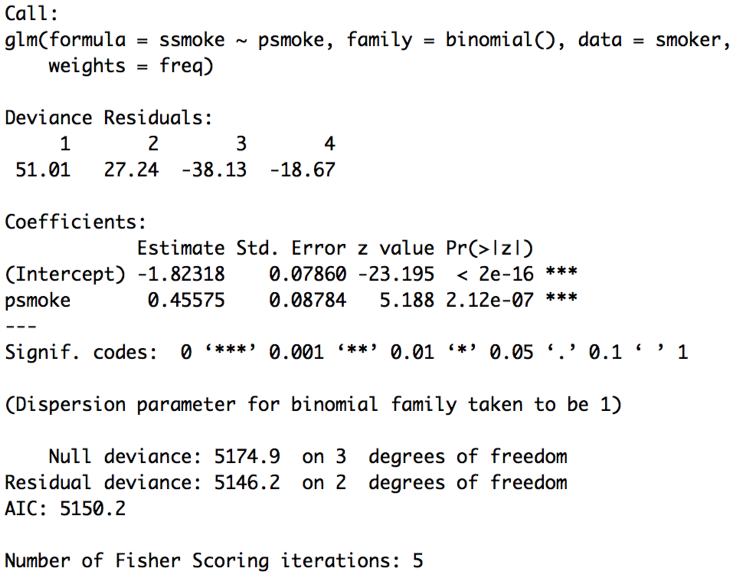
The data can be found in the file `smoker.dta` under Lab Week 7.

We previously calculated that the odds of a student smoking given that neither of their parents smoked was approximately 0.161, while the odds of a student smoking given that at least one of their parents smoked was 0.255. So our estimated odds ratio was 1.58, with a 95% confidence interval of (1.33, 1.88).

Modeling Outcomes with Logistic Regression

Let's try using logistic regression to model this association instead! We can fit the regression model in R by typing:

```
# Reading in the dataset
library(foreign)
smoker <- read.dta(file=file.choose())

# Fitting a logistic regression model
smoke.glm <- glm(ssmoke ˜ psmoke, family=binomial(), weights=freq, data=smoker)
summary(smoke.glm)
```



```
Call:
glm(formula = ssmoke ~ psmoke, family = binomial(), data = smoker,
    weights = freq)

Deviance Residuals:
     1       2       3       4
 51.01   27.24  -38.13  -18.67

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.82318    0.07860 -23.195  < 2e-16 ***
psmoke       0.45575    0.08784   5.188 2.12e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 5174.9  on 3  degrees of freedom
Residual deviance: 5146.2  on 2  degrees of freedom
AIC: 5150.2

Number of Fisher Scoring iterations: 5
```

What is the fitted regression model?

What are the estimated odds of a student smoking, given that both of their parents are non-smokers?

What is your estimate of the odds ratio comparing the odds of a student smoking given that at least one of their parents smokes to the odds given that both are non-smokers? Report both the point estimate and a 95% confidence interval in a manner appropriate for a journal.

How can you use your fitted model to calculate the odds of a student smoking, given that at least one of their parents is a smoker?

Compare the results from the logistic regression model to the results found from the contingency table analysis. What do you notice?

Hypothesis Testing

As was the case for linear regression, we want to be able to use our fitted logistic regression model to draw statistical conclusions about the relationship between Y and our predictors $X_1, \ldots, X_p$. In linear regression, we did this using two main types of tests: t-tests and F-tests. For logistic regression, we instead use Wald tests and Likelihood Ratio tests.

Wald tests

Intuitively, Wald tests function similarly to t-tests: they allow us to test whether a single coefficient or a linear combination of coefficients is equal to zero!

Formal Hypothesis: $H_0 : \beta_j = 0$ versus $H_1 : \beta_j \neq 0$

Test Statistic: $Z = \frac{\hat{\beta}_j - 0}{\text{s.d.}(\hat{\beta}_j)} \sim N(0, 1)$

Confidence Interval: $\hat{\beta}_j \pm z_{1-\frac{\alpha}{2}} \cdot \text{s.d.}(\hat{\beta}_j)$ (log odds ratio scale)
$\exp\{\hat{\beta}_j \pm z_{1-\frac{\alpha}{2}} \cdot \text{s.d.}(\hat{\beta}_j)\}$ (odds ratio scale)

Likelihood Ratio tests

Likelihood Ratio tests function similarly to F-tests: they allow us to compare the relative fits of two nested models. Suppose we have two nested models, where the reduced model is given by

$$\text{logit}(p) = \beta_0 + \beta_1 \cdot X_1 + \ldots + \beta_q \cdot X_q,$$

and the full model is given by

$$\text{logit}(p) = \beta_0 + \beta_1 \cdot X_1 + \ldots + \beta_p \cdot X_p.$$

The number of predictors in the full model is greater than the number in the reduced model, $p > q$.

Formal Hypothesis: $\big(H_0 : \text{the reduced model is sufficient}\big)$ versus $\big(H_1 : \text{the full model is preferred}\big)$

Test Statistic: $-2 \cdot \log\big(\frac{L_{reduced}}{L_{full}}\big) = -2 \cdot \log(L_{reduced}) + 2 \cdot \log(L_{full}) \sim \chi^2_{p-q}$

Note: The quantity $-2 \cdot \log(L)$ is sometimes called the deviance—this is what is usually reported in R. So in terms of the deviances of the reduced and full model, our Likelihood Ratio test statistic becomes

$$-2 \cdot \log(L_{reduced}) - (-2) \cdot \log(L_{full}) = \text{Deviance}_{reduced} - \text{Deviance}_{full} \sim \chi^2_{p-q}.$$

Example: Hypothesis Testing with Logistic Regression

To actually see these sorts of tests in practice, we'll analyze data from the Global Longitudinal Study of Osteoporosis in Women (GLOW). The dataset contains demographic information on 500 female subjects aged 55 or older, as well as information on the primary outcome of interest: whether or not a fracture occurs in the first year of study follow-up. A full list of the variables can be found in Table 1, and the data can be found on Canvas in the file `glow.csv`!

```
# Reading in the dataset & re-naming covariates in lowercase
glow <- read.csv(file=file.choose())
names(glow) <- tolower(names(glow))
```

Table 1: Global Longitudinal Study of Osteoporosis in Women - Relevant Variables

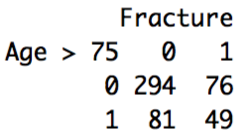
Variable	Description
SUB_ID	Subject identification code (numbered 1 through 500)
SITE_ID	Study site
PHY_ID	Physician ID code
PRIORFRAC	Indicator of history of prior fracture
AGE	Age at enrollment in the study
WEIGHT	Weight at enrollment in the study
HEIGHT	Height at enrollment in the study
BMI	BMI at enrollment in the study
PREMENO	Whether menopause occurred before age 45 (= 1) or after (= 0)
MOMFRAC	Whether the subject's mother had a hip fracture (= 1) or did not (= 0)
ARMASSIST	Whether arms are needed to stand from a chair (= 1) or not (= 0)
SMOKE	Whether a subject is a former or current smoker (= 1) or not (= 0)
RATERISK	Self-reported risk of fracture, recorded as 1 (less than others of the same age), 2 (same as others of the same age), or 3 (greater than others of the same age)
FRACSCORE	Composite fracture risk score
FRACTURE	Indicator for whether any fracture occurred in the first year of follow-up

Suppose we're interested in characterizing the association between age (X) and the risk of a woman experiencing a fracture within the first year of follow-up (Y). There are several different ways that we could go about representing age: as a binary predictor, as a categorical predictor, or as a continuous predictor. Let's explore each of these different approaches!

Age as a Binary Predictor

Let's first suppose that we're simply interested in whether or not being over the age of 75 (`age` > 75) is a significant predictor of bone fracture. In R, we can turn `age` into a binary predictor by calling:

```
# Creating an indicator for Age > 75 years
glow$age.75 <- ifelse(glow$age > 75, 1, 0)

# Contingency table for Age > 75 years and bone bracture
table(glow$age.75, glow$fracture, dnn=c("Age > 75", "Fracture"))
```



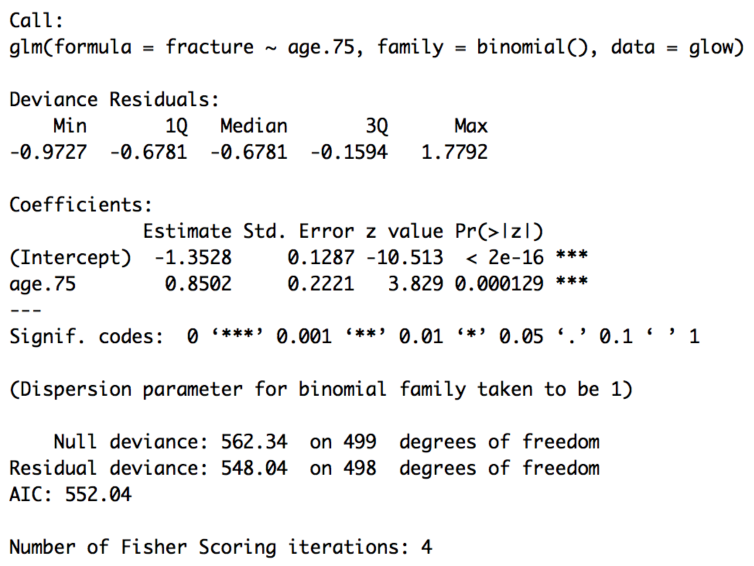
```
          Fracture
Age > 75   0   1
       0 294  76
       1  81  49
```

What does this contingency table seem to suggest about the relationship between bone fracture risk and being over the age of 75?

We can now fit a logistic regression to formally assess the relationship between the risk of bone fracture and being over 75 years of age:

$$\text{logit}(p) = \beta_0 + \beta_1 \cdot I(Age > 75)$$

```
# Fitting a simple logistic regression model
age.binary <- glm(fracture ~ age.75, family=binomial(), data=glow)
summary(age.binary)
```



```
Call:
glm(formula = fracture ~ age.75, family = binomial(), data = glow)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.9727  -0.6781  -0.6781  -0.1594   1.7792

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.3528     0.1287 -10.513  < 2e-16 ***
age.75        0.8502     0.2221   3.829 0.000129 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 562.34  on 499  degrees of freedom
Residual deviance: 548.04  on 498  degrees of freedom
AIC: 552.04

Number of Fisher Scoring iterations: 4
```

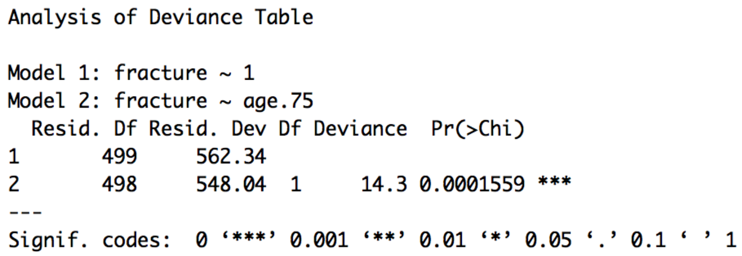
Using the output above, interpret both the intercept and slope coefficients on the odds/odds ratio scale.

Suppose we want to test for whether or not being over 75 years of age is a statistically significant predictor of bone fracture risk. *What are the null and alternative hypotheses for this test?*

Conduct a Wald test for this hypothesis. What are your conclusions, and how would you report them?

Alternatively, we could conceptualize testing $\beta_1 = 0$ as actually comparing two different nested models: the full model including the indicator for whether `age`> 75, and the reduced model including only the intercept term. *What are the null and alternative hypotheses for this Likelihood Ratio test?*

```
# Fitting the reduced model with only the intercept
age.binary.red <- glm(fracture ˜ 1, family=binomial(), data=glow)

# Performing the Likelihood Ratio test
anova(age.binary.red, age.binary, test="Chisq")
```



```
Analysis of Deviance Table

Model 1: fracture ~ 1
Model 2: fracture ~ age.75
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1       499     562.34
2       498     548.04  1     14.3 0.0001559 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

What are our conclusions from the Likelihood Ratio test? How do they compare to the results from the Wald test?

Age as a Categorical Predictor

We could also choose to model age as a categorical predictor. For example, suppose we thought there was some sort of meaningful difference in risk for women between the ages of 55 and 65, women between the ages of 65 and 75, and women over 75.

```
# Creating categorical age
glow$age.cat <- rep(NA, nrow(glow))
for (i in 1:nrow(glow)){
        if (glow$age[i] <= 65) {
                glow$age.cat[i] <- 0
        } else if (glow$age[i] <= 75) {
                glow$age.cat[i] <- 1
        } else{
                glow$age.cat[i] <- 2
        }
}
```

There are two different ways we could model this association between bone fracture risk and categorical age:

- By estimating the association for each category separately using indicator variables
- By assuming that the association changes in a linear fashion from one category to the next

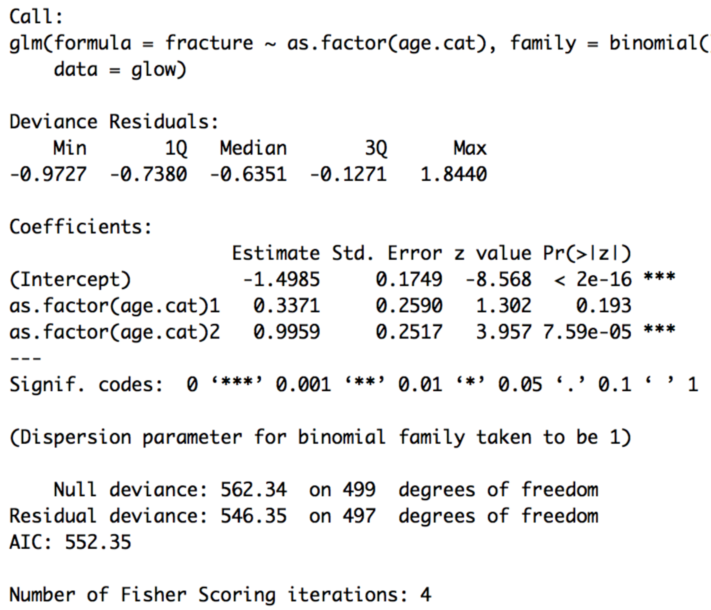
Estimating the Associations Separately

This approach makes fewer assumptions: we simply estimate the association separately for each age category! In order to do so, we have to use indicator variables (like we did in the binary case) and choose one category to be the "reference" level. For the purposes of this analysis, let's assume that our reference category is the population of women who are 65 or younger. Then our categorical model is given by

$$\text{logit}(p) = \beta_0 + \beta_1 \cdot I(65 < Age \leq 75) + \beta_2 \cdot I(Age > 75)$$

We can fit this model in R by treating `age.cat` as a factor variable:

```
# Fitting a logistic regression model with categorical age
age.category <- glm(fracture ~ as.factor(age.cat), family=binomial(), data=glow)
summary(age.category)
```



```
Call:
glm(formula = fracture ~ as.factor(age.cat), family = binomial(),
    data = glow)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.9727  -0.7380  -0.6351  -0.1271   1.8440

Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)          -1.4985     0.1749  -8.568  < 2e-16 ***
as.factor(age.cat)1   0.3371     0.2590   1.302    0.193
as.factor(age.cat)2   0.9959     0.2517   3.957 7.59e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 562.34  on 499  degrees of freedom
Residual deviance: 546.35  on 497  degrees of freedom
AIC: 552.35

Number of Fisher Scoring iterations: 4
```

How would you interpret the slope term for the third age category, $I(Age > 75) = 1$, on the odds ratio scale?

Assuming the Association Changes Linearly

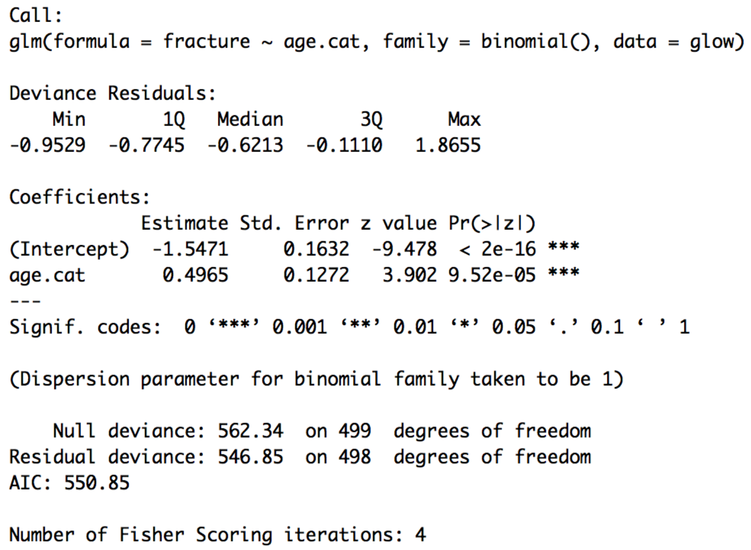
Alternatively, we could choose to model categorical age as if it were a continuous variable. In this approach, we are, however, making an additional assumption—we are assuming that the association between categorical age and the log odds of bone fracture changes linearly from one age category to the next:

$$\text{logit}(p) = \beta_0 + \beta_1 \cdot AgeCategory$$

Why might we want to make this assumption?

If we omit `as.factor()` in R, the software will treat `age.cat` as a continuous variable:

```
# Fitting a logistic regression model with categorical age (treated as continuous)
age.category2 <- glm(fracture ~ age.cat, family=binomial(), data=glow)
summary(age.category2)
```



```
Call:
glm(formula = fracture ~ age.cat, family = binomial(), data = glow)

Deviance Residuals:
    Min       1Q   Median       3Q      Max  
-0.9529  -0.7745  -0.6213  -0.1110   1.8655  

Coefficients:
            Estimate Std. Error z value Pr(>|z|)    
(Intercept)  -1.5471     0.1632  -9.478  < 2e-16 ***
age.cat       0.4965     0.1272   3.902 9.52e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 562.34  on 499  degrees of freedom
Residual deviance: 546.85  on 498  degrees of freedom
AIC: 550.85

Number of Fisher Scoring iterations: 4
```

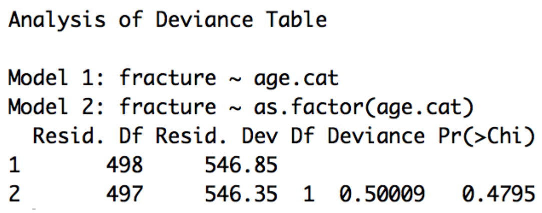
How would you interpret the slope term for categorical age on the odds ratio scale?

Comparing the Two Approaches

It’s great that we can model categorical age in two separate ways, but all of this begs the question: which method actually fits/describes our observed data better? To that end, we might want to consider performing something like a Likelihood Ratio test, which allows us to compare nested models.

Are the two different categorical age models nested in one another? Why or why not? And if they are nested, which is the full model and which is the reduced model?

```
# Conducting the Likelihood Ratio test (comparing categorical models)
anova(age.category2, age.category, test="Chisq")
```



```
Analysis of Deviance Table

Model 1: fracture ~ age.cat
Model 2: fracture ~ as.factor(age.cat)
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       498     546.85
2       497     546.35  1  0.50009   0.4795
```

What do we conclude from this Likelihood Ratio Test—which model is the preferred model?

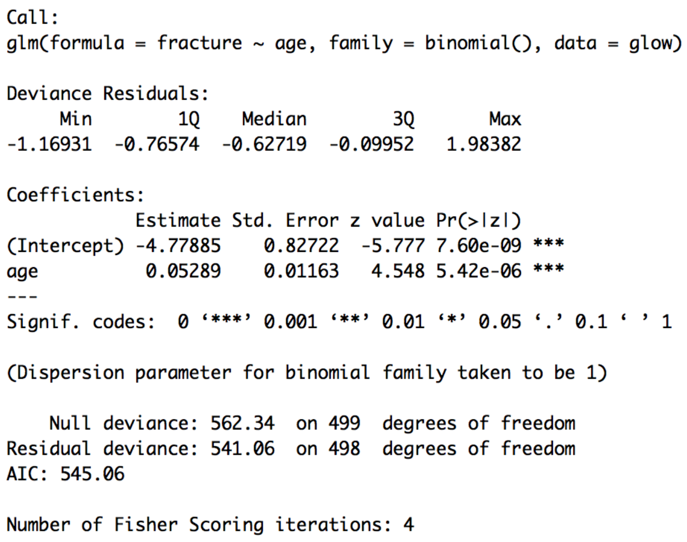
Age as a Continuous Predictor

Finally, we could also consider modeling age as a continuous predictor. In that case, the logistic regression model we fit is

$$\text{logit}(p) = \beta_0 + \beta_1 \cdot Age.$$

We can then fit this model in R:

```
# Fitting a logistic regression model with continuous age
age.continuous <- glm(fracture ~ age, family=binomial(), data=glow)
summary(age.continuous)
```



```
Call:
glm(formula = fracture ~ age, family = binomial(), data = glow)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.16931  -0.76574  -0.62719  -0.09952   1.98382

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.77885    0.82722  -5.777 7.60e-09 ***
age          0.05289    0.01163   4.548 5.42e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 562.34  on 499  degrees of freedom
Residual deviance: 541.06  on 498  degrees of freedom
AIC: 545.06

Number of Fisher Scoring iterations: 4
```

Interpret the slope for age on the odds/odds ratio scale, and report a 95% confidence interval for your estimate.

Can we formally compare (using a Likelihood Ratio test) the model with continuous age to either the binary or the categorical models? Why or why not?

Confounding & Effect Modification

Up until this point, every example we've seen in lab has been a simple logistic regression model, where we consider the association between Y and just one predictor, X. As was the case for linear regression, we can easily extend our understanding of logistic regression to include multiple predictors: $X_1, X_2, \ldots, X_p$. Two important motivations for performing multiple logistic regression are confounding and effect modification.

A quick review:

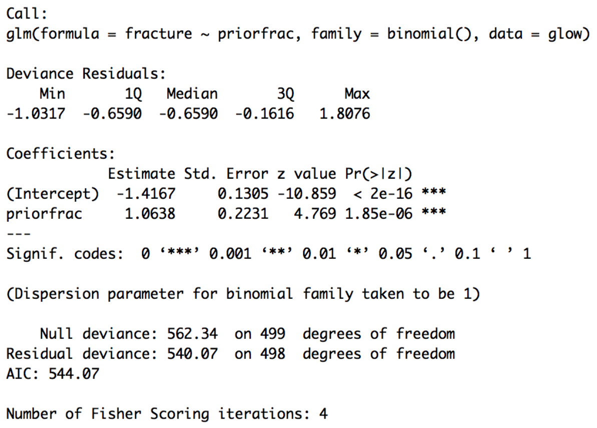
- A **confounder** is a variable that is associated with both our outcome of interest Y and the exposure X, but is not a consequence of the exposure.
  - In logistic regression, we'll consider a variable to be a meaningful confounder between X and Y if adjusting for it changes the estimated slope by 10% or more!
- A variable is considered to be an **effect modifier** if the magnitude of the association between X and Y varies across its different levels.

As we'll see in the following two examples, we adjust for and assess confounding and effect modification in logistic regression models in exactly the same way as we did for linear regression models! We'll continue to use the GLOW dataset, but now we'll also consider the PRIORFRAC covariate—an indicator for whether or not a woman has had a prior fracture.

Confounding & Logistic Regression

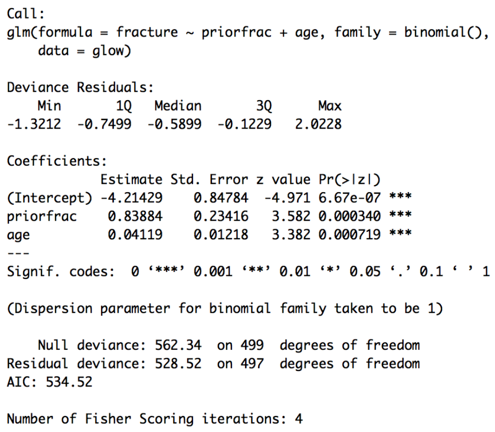
Suppose that we're now interested in assessing the relationship between having a history of prior fractures and the risk of developing a fracture within the first year of follow-up. *Is age a potential confounder of this relationship?*

```
# Fitting a logistic regression model without age
priorfrac1 <- glm(fracture ~ priorfrac, family=binomial(), data=glow)
summary(priorfrac1)

# Fitting a logistic regression model with age
priorfrac2 <- glm(fracture ~ priorfrac + age, family=binomial(), data=glow)
summary(priorfrac2)
```



```
Call:
glm(formula = fracture ~ priorfrac, family = binomial(), data = glow)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.0317  -0.6590  -0.6590  -0.1616   1.8076

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.4167     0.1305 -10.859  < 2e-16 ***
priorfrac     1.0638     0.2231   4.769 1.85e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 562.34  on 499  degrees of freedom
Residual deviance: 540.07  on 498  degrees of freedom
AIC: 544.07

Number of Fisher Scoring iterations: 4
```

(a) Regression output for the unadjusted model.



```
Call:
glm(formula = fracture ~ priorfrac + age, family = binomial(),
    data = glow)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.3212  -0.7499  -0.5899  -0.1229   2.0228

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.21429    0.84784  -4.971 6.67e-07 ***
priorfrac    0.83884    0.23416   3.582 0.000340 ***
age          0.04119    0.01218   3.382 0.000719 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 562.34  on 499  degrees of freedom
Residual deviance: 528.52  on 497  degrees of freedom
AIC: 534.52

Number of Fisher Scoring iterations: 4
```

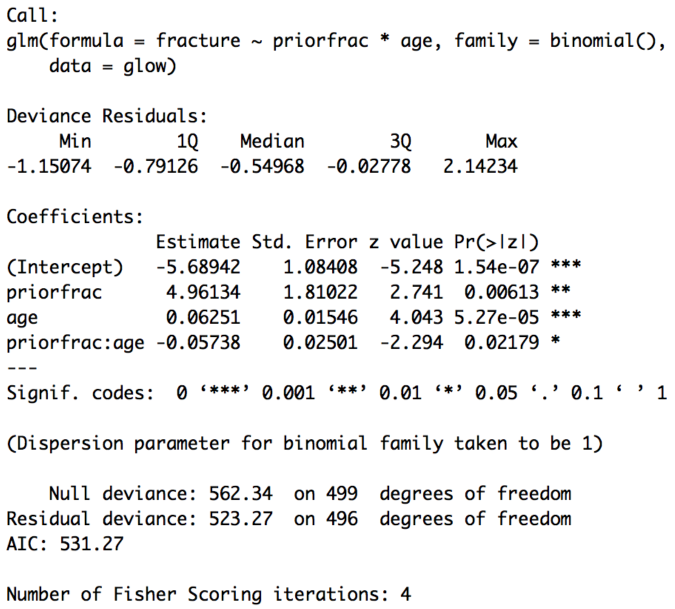
(b) Regression output for the adjusted model.

Does age appear to be a meaningful confounder? How do we interpret the coefficient for prior history of fracture in the adjusted model?

Effect Modification & Logistic Regression

Similarly, we can also use logistic regression to assess whether or not having a history of prior fractures modifies the association between age and risk of fracture.

```
# Assessing whether prior history of fractures is an effect modifier
priorfrac3 <- glm(fracture ~ priorfrac*age, family=binomial(), data=glow)
summary(priorfrac3)
```



```
Call:
glm(formula = fracture ~ priorfrac * age, family = binomial(),
    data = glow)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.15074  -0.79126  -0.54968  -0.02778   2.14234

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)   -5.68942    1.08408  -5.248 1.54e-07 ***
priorfrac      4.96134    1.81022   2.741  0.00613 **
age            0.06251    0.01546   4.043 5.27e-05 ***
priorfrac:age -0.05738    0.02501  -2.294  0.02179 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 562.34  on 499  degrees of freedom
Residual deviance: 523.27  on 496  degrees of freedom
AIC: 531.27

Number of Fisher Scoring iterations: 4
```

Does prior history of fracture appear to be a significant effect modifier of this relationship?

What is the final fitted model for those with a prior history of fractures? What about for those without a prior history of fractures?

Finally, note that we can plot the fitted log odds—and corresponding fitted probabilities—for those with a prior history of fractures and those without a prior history of fractures using the R code below:

```
# Plotting fitted log odds
curve(coef(priorfrac3)[1] + coef(priorfrac3)[3]*x, xlim=c(55, 90), xlab="Age",
        ylab="Logit(p)", col="dodgerblue")
curve(coef(priorfrac3)[1] + coef(priorfrac3)[2] +
        (coef(priorfrac3)[3] + coef(priorfrac3)[4])*x, xlim=c(55, 90), col="magenta",
        add=T)

# Plotting fitted probabilities
curve(exp(coef(priorfrac3)[1] + coef(priorfrac3)[3]*x)/
        (1+ exp(coef(priorfrac3)[1] + coef(priorfrac3)[3]*x)), xlim=c(55, 90),
        ylim=c(0,1), xlab="Age", ylab="Risk of Fracture", col="dodgerblue")
curve(exp(coef(priorfrac3)[1] + coef(priorfrac3)[2] +
        (coef(priorfrac3)[3] + coef(priorfrac3)[4])*x)/(1 + exp(coef(priorfrac3)[1] +
        coef(priorfrac3)[2] + (coef(priorfrac3)[3] + coef(priorfrac3)[4])*x)),
        xlim=c(55, 90), col="magenta", add=T)
```

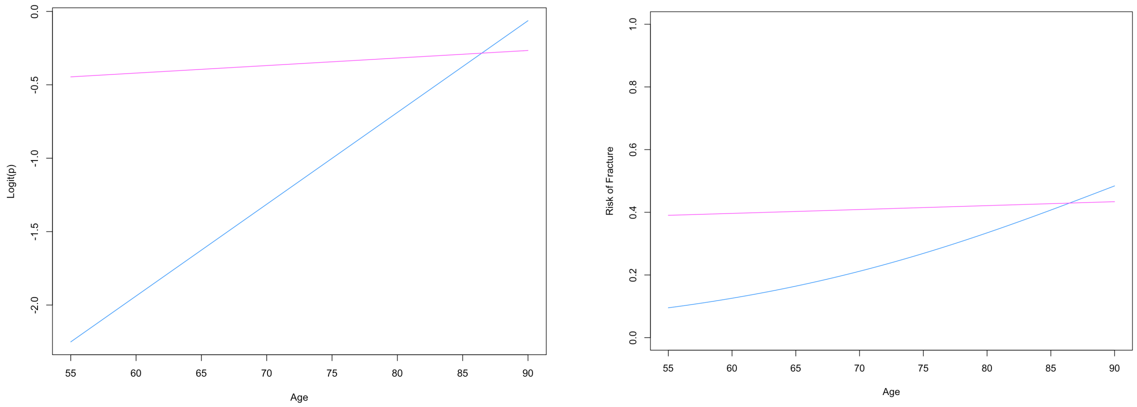


Figure 4: Fitted log odds (left) and probabilities (right) of bone fracture in the first year of follow-up as a function of age, separated out by women with a prior history of bone fracture (magenta) and no prior history of bone fracture (blue).