Lab 6 BIS 505b

Maria Ciarleglio

4/5/2021

Goal of Lab 6

In **Lab 6**, we will analyze a **binary endpoint** using **logistic regression.** We will begin by **(1)** creating a 2x2 contingency table that describes the association between a binary exposure and a binary endpoint and **(2)** see how the 2x2 table connects to a simple logistic regression model with a binary predictor. Next, we will **(3)** use a fitted logistic regression model to estimate the probability of the event of interest and explore the effects of **(4)** a continuous predictor and a **(5)** a categorical predictor with > 2 levels in a logistic regression model. Finally, we will **(6)** build a multiple logistic regression model, **(7)** perform a likelihood ratio test, and **(8)** describe and assess the fit of the multiple logistic regression model.

Analysis Data Set

In this lab, we will again analyze a subset of data from the **National Health and Nutrition Examination Survey** (NHANES) (n = 1430) `nhanes.csv` imported as the data frame `nhanes` in code chunk 3 above). The **Data Key** is provided below. In this lab, our endpoint of interest is oral diabetes medication use (`oralmed`).

Variable Name	Definition
`age`	Age at time of survey
`sex`	Sex
	0 = Male
	1 = Female
`race`	Race/Ethnicity
	1 = White
	2 = Black

Variable Name	Definition
	3 = Mexican-American
	4 = Other
`insulin`	Insulin use
	0 = No
	1 = Yes
`oralmed`	Oral diabetes medication use (**Our Response**)
	0 = No
	1 = Yes
`fastgluc`	Fasting glucose level (mg/dL)
	88888 = Missing

- **Missing Values:**

We begin by re-coding missing values of `fastgluc` as `NA`. Recall that missing values are coded numerically as `88888` in this data set.

```
# Re-code a `fastgluc` value of 88888 as NA
nhanes$fastgluc[nhanes$fastgluc == 88888] <- NA
```

- **Creating Factor Variables:**

Next, we create factor variable versions of the **categorical variables** in this data set (`sex`, `race`, `insulin`, and `oralmed`). The **first level** specified in the `factor()` function is used as the **reference level**.

```
# Creating factor variables in nhanes using mutate() function in "dplyr" package
nhanes <- mutate(nhanes,
                 sex_factor = factor(sex,
                                     levels = c(0, 1),
                                     labels = c("Male", "Female")),
                 insulin_factor = factor(insulin,
                                         levels = c(0, 1),
                                         labels = c("No", "Yes")),
                 oralmed_factor = factor(oralmed,
                                         levels = c(0, 1),
                                         labels = c("No", "Yes")),
                 race_factor = factor(race,
                                      levels = c(1, 2, 3, 4),
                                      labels = c("White", "Black", "Mexican-American", "Other"
)))
```

Research Questions

We are interested in determining the characteristics that associated with **oral diabetes medication use** `oralmed` (our response variable, y). The main explanatory variables (Q1-Q3) and questions of interest are:

1. **Question 1:** Sex (`sex`)
2. **Question 2:** Age (`age`)
3. **Question 3:** Race (`race`)
4. **Question 4:** Estimate the age- and race-adjusted effect of sex on oral diabetes medication use

2x2 Table

In **Question 1**, we explore the association between **oral diabetes medication use** (`oralmed_factor` , **binary outcome variable**) and **sex** (`sex_factor` , **binary explanatory variable**). A 2x2 **contingency table** is used to summarize the relationship between two binary variables. We can also report the **odds ratio** to quantify the association between the variables. The `table(rowvar, colvar)` function creates basic tables, and the `epi.2by2()` function in the `epiR` package reports summary measures such as the **risk**, the **odds** and the **odds ratio** of a 2x2 table object.

By default, the `epi.2by2()` function assumes that the **column variable** is the outcome variable. The **first column** should correspond to the *outcome* of interest (success) and the **first row** should correspond to the *exposure level* of interest (exposed). Because **R** orders the factor variables with the reference category *first*, we need to `relevel()` the factors when creating the contingency table object (`tab`) and specify the exposure and outcome levels of interest as the reference levels (i.e., `"Female"` and `"Yes"`) so that they appear as the first row and column levels, respectively. We can interpret the *risk* (`Inc risk *`) and *odds* of oral diabetes medication use (`Outcome +`) in females (`Exposed +`) and in males (`Exposed -`) and the *odds ratio* of oral medication use in females vs. males (reference).

```
# ref = exposure level of interest and outcome of interest
tab <- table(relevel(nhanes$sex_factor, ref = "Female"), relevel(nhanes$oralmed_factor, ref = "Y
es"))
tab
```

```
##
##          Yes  No
##   Female 371 431
##   Male   327 296
```

```
# Summary measures using epi.2by2() function in "epiR" package
res <- epi.2by2(tab, method = "cohort.count", units = 1)
res
```

```
##              Outcome +    Outcome -      Total        Inc risk *        Odds
## Exposed +          371          431        802             0.463       0.861
## Exposed -          327          296        623             0.525       1.105
## Total              698          727       1425             0.490       0.960
##
## Point estimates and 95% CIs:
## -------------------------------------------------------------------
## Inc risk ratio                                0.88 (0.79, 0.98)
## Odds ratio                                    0.78 (0.63, 0.96)
## Attrib risk *                                 -0.06 (-0.11, -0.01)
## Attrib risk in population *                   -0.04 (-0.08, 0.01)
## Attrib fraction in exposed (%)                -13.46 (-26.10, -2.10)
## Attrib fraction in population (%)             -7.16 (-13.35, -1.30)
## -------------------------------------------------------------------
##  Test that OR = 1: chi2(1) = 5.443 Pr>chi2 = 0.02
##  Wald confidence limits
##  CI: confidence interval
##  * Outcomes per population unit
```

- The estimated **risk** of oral diabetes medication use in **females** is equal to $\hat{p}_1$ = 0.463, which gives an estimated odds $\frac{\hat{p}_1}{1-\hat{p}_1}$ = 0.861.
- The estimated **risk** of oral diabetes medication use in **males** (ref) is equal to $\hat{p}_0$ = 0.525 which gives an estimated odds $\frac{\hat{p}_0}{1-\hat{p}_0}$ = 1.105. Notice that the odds are > 1 when $p > 0.5$ since the probability of observing the outcome is greater than the probability of not observing the outcome.
- The estimated **odds ratio** of oral diabetes medication use in females vs. males (ref) is equal to $\hat{OR}$ = 0.78 [95% CI (0.63, 0.96)]. The odds that a female uses oral diabetes medication is 0.78 times the odds that a male uses oral diabetes medication. When the odds ratio is < 1, we can also say that the odds is $100 \times (1 - \hat{OR})\%$ lower (22% lower, here) in females compared to males.

Logistic Regression

Logistic regression or a logit model can be used to model a binary $Y = \{0, 1\}$ response, where $Y = 1$ is the event of interest, or a "success" (oral diabetes medication use = Yes, in our example). The mean of Y or $P(Y = 1) = p$ is related to the linear function of the covariates $\alpha + \beta\, x$ (a.k.a., the *linear predictor*) through the **logit function**, $\log\left(\frac{p}{1-p}\right)$, giving the logistic regression model:

$$\log\left(\frac{p}{1-p}\right) = \alpha + \beta_1\, x_1 + \beta_2\, x_2 + \ldots + \beta_k\, x_k$$

The logit function is the log-odds of "success". In logistic regression, the logit function is called the *link function*, or the function that links p to $\alpha + \beta\, x$. The model above is estimated to give $\log\left(\frac{\hat{p}}{1-\hat{p}}\right) = a + b_1\, x_1 + b_2\, x_2 + \ldots + b_k\, x_k$.

- The estimated intercept a is equal to the **log-odds of "success"** (i.e., oral diabetes medication use) when all values of $x = 0$.

- The estimated slope b_j is equal to the **log-odds ratio** associated with a 1-unit increase in x_j controlling for or holding all other predictors constant. We must **exponentiate** the slope e^{b_j} to find the **odds ratio**.

A **hypothesis test** of the slope parameter $H_0 : \beta_j = 0$ vs. $H_1 : \beta_j \neq 0$ is performed using a **Wald test**, $z = \frac{b_j}{s_{b_j}}$, which is compared to a **standard Normal distribution**. Under H_0, $\beta_j = 0$, there is no association between x_j and the outcome. When the $\log(OR) = 0$, the $OR = e^0 = 1$.

The **probability of success**, p, (i.e., the probability of oral diabetes medication use) is assumed to be a function of the predictor variables. The estimated or predicted probability of success is equal to:

$$\hat{p} = \frac{e^{a+b_1\,x_1+b_2\,x_2+\ldots+b_k\,x_k}}{1+e^{a+b_1\,x_1+b_2\,x_2+\ldots+b_k\,x_k}}$$

Using this function, we can estimate the probability of the event of interest given the values of different risk factors (x).

We use the `glm()` function in **R** to run a logistic regression model. `glm` stands for **generalized linear model** and expands on the general linear model that we discussed in Lessons 4 and 5. To run a **logistic regression** generalized linear model in **R**, we must specify which type of generalized linear model to run. A logistic regression model is specified through the function argument `family = binomial(link = "logit")`.

`glm()` Function Arguments	Option Definition
`formula=`	`analysis_variable ~ predictor_variable1 + predictor_variable2`
`data=`	Data frame containing sample data
`family=`	Error distribution and link function
	- Logistic regression `=binomial(link="logit")`
	- Poisson regression `=poisson(link="log")`
	- Linear regression `=gaussian(link="identity")`

In **Lab 7**, we will discuss a Poisson regression model, which is another generalized linear model.

Binary Predictor Variable

We can address **Question 1** using a simple logistic regression model. Just as in linear regression, **categorical variables** (including binary variables) can be included as predictors in the model through the use of numeric 0/1 **dummy** or **indicator variables** z_j, where the reference level of the variable equals `0`. When we include a **factor variable** in a regression model, **R** will automatically create the dummy variable(s) (a.k.a., "design variables") necessary to represent that categorical variable. The `contrasts()` function returns the dummy variable coding that **R** uses to represent a factor variable. For example, `sex_factor` is a dummy variable (z_1) that equals `1` for females and `0` for males (the reference category).

```
contrasts(nhanes$sex_factor)
```

```
##        Female
## Male        0
## Female      1
```

To estimate the association between **oral diabetes medication use** (`oralmed`) and **sex** (`sex_factor`), fit the logistic regression model, $\log\left(\frac{p}{1-p}\right) = \alpha + \beta_1 \text{Female}$. The result of the `glm()` function is usually saved as an object (`mod.sex` , below) and the `summary()` function is applied to that object (`summary(mod.sex)`) to output detailed results.

```
mod.sex <- glm(oralmed ~ sex_factor, data = nhanes, family = binomial(link = "logit"))
summary(mod.sex)
```

```
##
## Call:
## glm(formula = oralmed ~ sex_factor, family = binomial(link = "logit"),
##     data = nhanes)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.220  -1.114  -1.114   1.242   1.242
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)       0.09960    0.08023   1.241   0.2144
## sex_factorFemale -0.24951    0.10701  -2.332   0.0197
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1974.9  on 1424  degrees of freedom
## Residual deviance: 1969.4  on 1423  degrees of freedom
##   (5 observations deleted due to missingness)
## AIC: 1973.4
##
## Number of Fisher Scoring iterations: 3
```

We can extract the **model coefficients** (a, b_1) using the `coef()` function and the **confidence intervals** of the model parameters (α, β_1) using the `confint.default()` function. Remember that we must exponentiate b_1 to give an odds ratio. Similarly, we must exponentiate the CI for β_1 to give and the confidence interval of the odds ratio.

```
# Exponentiated slope coefficient = OR and 95% CI for OR
exp(cbind(OR = coef(mod.sex), confint.default(mod.sex)))
```

```
##                        OR     2.5 %    97.5 %
## (Intercept)      1.104730 0.9439873 1.2928435
## sex_factorFemale 0.779185 0.6317567 0.9610176
```

- The **fitted model** is given by the equation, $\log\left(\frac{\hat{p}}{1-\hat{p}}\right) = 0.0996 - 0.25$ Female.

- The **estimated intercept** $a = 0.0996$ is equal to the log-odds of oral diabetes medication use when $z_1 = 0$ (i.e., the *log-odds of oral diabetes medication use* in the reference category (males)). The exponentiated intercept $e^a = 1.105$ is equal to the *odds of oral diabetes medication use* in males. [*Can you find this value in the* `epi.2by2()` *result above?*]
- The **estimated slope** of `sex_factor` $b_1 = -0.25$ is equal to the *log-odds ratio* of oral diabetes medication use in females vs. males (ref). The exponentiated slope e^{b_1} gives the estimated **odds ratio**, $\hat{OR} = e^{b_1} = 0.78$ [95% CI (0.63, 0.96)]. Equivalently, the odds of oral diabetes medication use is 22% lower in females compared to males. [*Can you find the odds ratio and its confidence interval in the* `epi.2by2()` *result above?*]
- A **significance test of the slope** ($H_0 : \beta_1 = 0$ vs. $\beta_1 \neq 0$) reports a z-statistic z = -2.33, which is compared to a standard Normal distribution. We have evidence to reject H_0 and conclude that the odds of using oral diabetes medication is significantly different in females and males (p-value = 0.02).

The odds ratio comparing males to females (ref) is equal to 1/(odds ratio of females vs. males). We can also use the `relevel()` function to change the reference category (`ref=`) of an existing factor variable and re-run the model:

```
# Female will be the reference category (=0) of sex_factorv2
nhanes$sex_factorv2 <- relevel(nhanes$sex_factor, ref = "Female")
contrasts(nhanes$sex_factorv2)
```

```
##        Male
## Female    0
## Male      1
```

```
# Comparing males to females (ref)
mod.sexv2 <- glm(oralmed ~ sex_factorv2, data = nhanes, family = binomial(link = "logit"))
summary(mod.sexv2)
```

```
##
## Call:
## glm(formula = oralmed ~ sex_factorv2, family = binomial(link = "logit"),
##     data = nhanes)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.220  -1.114  -1.114   1.242   1.242
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -0.14991    0.07082  -2.117   0.0343
## sex_factorv2Male  0.24951    0.10701   2.332   0.0197
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1974.9  on 1424  degrees of freedom
## Residual deviance: 1969.4  on 1423  degrees of freedom
##   (5 observations deleted due to missingness)
## AIC: 1973.4
##
## Number of Fisher Scoring iterations: 3
```

```
exp(cbind(OR = coef(mod.sexv2), confint.default(mod.sexv2)))
```

```
##                        OR     2.5 %    97.5 %
## (Intercept)      0.8607889 0.7492278 0.9889616
## sex_factorv2Male 1.2833922 1.0405636 1.5828879
```

- The **fitted model** is given by the equation, $\log\left(\frac{\hat{p}}{1-\hat{p}}\right) =$ -0.1499 $+$ 0.25 Male.
- The **estimated intercept** $a =$ -0.1499 is equal to the log-odds of oral diabetes medication use in females. The exponentiated intercept $e^a =$ 0.861 is equal to the *odds of oral diabetes medication use* in females. [*Can you find this value in the* `epi.2by2()` *result above?*]
- The **estimated slope** of `sex_factorv2` $b_1 =$ 0.25 is equal to the *log-odds ratio* of oral diabetes medication use in males vs. females (ref). The estimated **odds ratio**, $\hat{OR} = e^{b_1} =$ 1.28 [95% CI (1.04, 1.58)], indicates that the odds of oral diabetes medication use is 28% higher in males than in females.

Predicted Probability

The fitted model can be used to estimate or predict the probability of a success p for given values of x using the `predict()` function. A data frame that contains the values of x for which we would like to calculate p must be specified in the `newdata=` argument of the `predict()` function. The values of x for which we would like to estimate p (i.e., `sex_factor` = “Male” and “Female”) are stored in the data frame `pred.x` :

```
# Data frame that includes desired values for prediction
levels(nhanes$sex_factor)
```

```
## [1] "Male"   "Female"
```

```
pred.x <- data.frame(sex_factor = levels(nhanes$sex_factor))

# Equivalently,
pred.x <- data.frame(sex_factor = c("Male", "Female"))

# Returns predicted probabilities
phat <- predict(mod.sex, newdata = pred.x, type = "response")
cbind(pred.x, phat)
```

```
##   sex_factor      phat
## 1       Male 0.5248796
## 2     Female 0.4625935
```

- The **estimated probability of oral diabetes medication use** in **males** is equal to $\hat{p}_0 = 0.525$.
- The **estimated probability of oral diabetes medication use** in **females** is equal to $\hat{p}_1 = 0.463$.
- [*Can you find these values in the* `epi.2by2()` *results above?*]
- As you can see, an un-adjusted logistic regression model with a **binary predictor** is equivalent to the results from the analysis of a **2x2 contingency table.**

Continuous Predictor

In **Question 2**, we explore the association between **oral diabetes medication use** (`oralmed_factor` , **binary outcome variable**) and **age** (`age` , **continuous explanatory variable**).

```
mod.age <- glm(oralmed ~ age, data = nhanes, family = binomial(link = "logit"))
summary(mod.age)
```

```
##
## Call:
## glm(formula = oralmed ~ age, family = binomial(link = "logit"),
##     data = nhanes)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.177  -1.161  -1.141   1.194   1.217
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.141479   0.263375  -0.537    0.591
## age          0.001565   0.004007   0.391    0.696
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1974.9  on 1424  degrees of freedom
## Residual deviance: 1974.7  on 1423  degrees of freedom
##   (5 observations deleted due to missingness)
## AIC: 1978.7
##
## Number of Fisher Scoring iterations: 3
```

```
# Exponentiated slope coefficient = OR and 95% CI for OR
exp(cbind(OR = coef(mod.age), confint.default(mod.age)))
```

```
##                    OR     2.5 %   97.5 %
## (Intercept) 0.8680736 0.5180495 1.454594
## age         1.0015664 0.9937316 1.009463
```

- The **fitted model** is given by the equation, $\log\left(\frac{\hat{p}}{1-\hat{p}}\right) =$ -0.1415 $+$ 0.0016 Age.
- The **estimated intercept** $a =$ -0.1415 is equal to the log-odds of oral diabetes medication use when age equals 0, which is not interpretable in this model.
- The **estimated slope** of `age` $b_1 =$ 0.0016 is equal to the *log-odds ratio* of oral diabetes medication use associated with a 1-unit increase in age The exponentiated slope e^{b_1} gives the estimated **odds ratio**, $\hat{OR} = e^{b_1} =$ 1.0016 [95% CI (0.994, 1.009)]. Equivalently, a 1-unit increase in age increases the odds of oral diabetes medication use by 0.16%.
- A **significance test of the slope** ($H_0 : \beta_1 = 0$ vs. $\beta_1 \neq 0$) reports a z-statistic z = 0.39, which is compared to a standard Normal distribution. We do not have evidence to reject H_0 and cannot conclude that the odds of using oral diabetes medication is significantly associated with age (p-value = 0.696).

```
range(nhanes$age, na.rm = TRUE)
```

```
## [1] 30 90
```

```
# Values of x used to estimate p
pred.x <- data.frame(age = c(40, 60, 80))

# Returns predicted probabilities
phat <- predict(mod.age, newdata = pred.x, type = "response")
cbind(pred.x, phat)
```

```
##   age      phat
## 1  40 0.4802923
## 2  60 0.4881102
## 3  80 0.4959340
```

- The **estimated probability of oral diabetes medication use** in **40**-year olds is equal to $\hat{p} = 0.48$.
- The **estimated probability of oral diabetes medication use** in **60**-year olds is equal to $\hat{p} = 0.488$.
- The **estimated probability of oral diabetes medication use** in **80**-year olds is equal to $\hat{p} = 0.496$.
- Notice that the estimated probability of oral diabetes medication use increases with age. This agrees with an observed estimated $\hat{OR} > 1$.

Categorical Predictor Variable

In **Question 3**, we explore the association between **oral diabetes medication use** (`oralmed_factor` , **binary outcome variable**) and **race** (`race_factor` , 4-level **categorical explanatory variable**).

Categorical variables with C levels are represented by a set of $C - 1$ dummy variables. Again, when using factor versions of our categorical variables, **R** automatically creates the dummy variables needed to represent the categorical variable in a regression model. Be sure **not** to use `ordered = TRUE` when creating factor variables for inclusion in a regression model.

`race_factor` contains **4** levels (White, Black, Mexican-American, and Other) and must be represented by **3** dummy variables (z_1, z_2 and z_3). When we created `race_factor` at the beginning of this Lab, White was specified as the first level (`levels=c(1,2,3,4)` corresponding to `labels=c("White", "Black", "Mexican-American", "Other")`), thus `"White"` will be the reference category. All dummy variables will equal `0` for the reference level of the categorical variable. Below, we see the 3 dummy variables that describe race:

```
contrasts(nhanes$race_factor)
```

```
##                 Black Mexican-American Other
## White               0                0     0
## Black               1                0     0
## Mexican-American    0                1     0
## Other               0                0     1
```

1. z_1 equals `1` when `race_factor == "Black"` and equals `0` otherwise
2. z_2 equals `1` when `race_factor == "Mexican-American"` and equals `0` otherwise
3. z_3 equals `1` when `race_factor == "Other"` and equals `0` otherwise

```
mod.race <- glm(oralmed ~ race_factor, data = nhanes, family = binomial(link = "logit"))
summary(mod.race)
```

```
##
## Call:
## glm(formula = oralmed ~ race_factor, family = binomial(link = "logit"),
##     data = nhanes)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.297  -1.145  -1.029   1.210   1.334
##
## Coefficients:
##                              Estimate Std. Error z value Pr(>|z|)
## (Intercept)                 -0.07669    0.08354  -0.918   0.3586
## race_factorBlack            -0.28375    0.13287  -2.136   0.0327
## race_factorMexican-American  0.35351    0.12838   2.753   0.0059
## race_factorOther             0.24854    0.34944   0.711   0.4769
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1974.9  on 1424  degrees of freedom
## Residual deviance: 1953.9  on 1421  degrees of freedom
##   (5 observations deleted due to missingness)
## AIC: 1961.9
##
## Number of Fisher Scoring iterations: 4
```

```
# Exponentiated slope coefficient = OR and 95% CI for OR
exp(cbind(OR = coef(mod.race), confint.default(mod.race)))
```

```
##                                 OR     2.5 %    97.5 %
## (Intercept)                 0.9261745 0.7862916 1.0909428
## race_factorBlack            0.7529558 0.5803249 0.9769397
## race_factorMexican-American 1.4240501 1.1072455 1.8314988
## race_factorOther            1.2821558 0.6463895 2.5432397
```

- The **fitted model** is given by the equation, $\log\left(\frac{\hat{p}}{1-\hat{p}}\right)$ = -0.0767 − 0.284 Black + 0.354 Mexican-American + 0.249 Other.
- The **estimated intercept** a = -0.077 is equal to the log-odds of oral diabetes medication use when $z_1 = z_2 = z_3 = 0$ (i.e., the *log-odds of oral diabetes medication use* in the reference category (Whites)). The exponentiated intercept e^a = 0.926 is equal to the *odds of oral diabetes medication use* in Whites.
- The **estimated slope** of the first dummy variable z_1 (Black), b_1 = -0.28 is equal to the *log-odds ratio* of oral diabetes medication use in Blacks vs. Whites (ref). The exponentiated slope e^{b_1} gives the estimated **odds ratio**, $\hat{OR} = e^{b_1}$ = 0.75 [95% CI (0.58, 0.98)]. Equivalently, the odds of oral diabetes medication use is 25% lower in Blacks compared to Whites.

    - A **significance test of the slope** ($H_0: \beta_1 = 0$ vs. $\beta_1 \neq 0$) reports a z-statistic z = -2.14, which is compared to a standard Normal distribution. We have evidence to reject H_0 and conclude that the odds of using oral diabetes medication is significantly different in Blacks and Whites (p-value = 0.033).

- The **estimated slope** of the second dummy variable z_2 (Mexican-American), $b_2 = 0.35$ is equal to the *log-odds ratio* of oral diabetes medication use in Mexican-Americans vs. Whites (ref). The exponentiated slope e^{b_2} gives the estimated **odds ratio**, $\hat{OR} = e^{b_2} = 1.42$ [95% CI (1.11, 1.83)]. Equivalently, the odds of oral diabetes medication use is 42% higher in Mexican-Americans compared to Whites.
    - A **significance test of the slope** ($H_0: \beta_2 = 0$ vs. $\beta_2 \neq 0$) reports a z-statistic z = 2.75. We have evidence to reject H_0 and conclude that the odds of using oral diabetes medication is significantly different in Mexican-Americans and Whites (p-value = 0.006).

- The **estimated slope** of the third dummy variable z_3 (Other), $b_3 = 0.25$ is equal to the *log-odds ratio* of oral diabetes medication use in Others vs. Whites (ref). The exponentiated slope e^{b_3} gives the estimated **odds ratio**, $\hat{OR} = e^{b_3} = 1.28$ [95% CI (0.65, 2.54)]. Equivalently, the odds of oral diabetes medication use is 28% higher in Others compared to Whites.
    - A **significance test of the slope** ($H_0: \beta_3 = 0$ vs. $\beta_3 \neq 0$) reports a z-statistic z = 0.71. We do not have evidence to reject H_0 and cannot conclude that the odds of using oral diabetes medication is significantly different in Others and Whites (p-value = 0.477).

```
# Values of x used to estimate p
pred.x <- data.frame(race_factor = levels(nhanes$race_factor))

# Returns predicted probabilities
phat <- predict(mod.race, newdata = pred.x, type = "response")
cbind(pred.x, phat)
```

```
##        race_factor      phat
## 1            White 0.4808362
## 2            Black 0.4108527
## 3 Mexican-American 0.5687646
## 4            Other 0.5428571
```

- The **estimated probability of oral diabetes medication use** in **Whites** is equal to $\hat{p} = 0.481$.
- The **estimated probability of oral diabetes medication use** in **Blacks** is equal to $\hat{p} = 0.411$.
- The **estimated probability of oral diabetes medication use** in **Mexican-Americans** is equal to $\hat{p} = 0.569$.
- The **estimated probability of oral diabetes medication use** in **Others** is equal to $\hat{p} = 0.543$.

Exercise: Verify that the estimated probabilities from this unadjusted logistic regression model match the raw proportions of individuals who use oral diabetes medication within each level of race.

▶ Answer:

Multiple Logistic Regression

In **Question 4**, we are interested in estimating the age- (`age`) and race-adjusted (`race_factor`) effect of **sex** (`sex_factor`) on **oral diabetes medication use** (`oralmed`). To control for age and race, we extend the simple logistic regression model containing `sex_factor` to also contain `age` and `race_factor` :

```
mod.adjsex <- glm(oralmed ~ sex_factor + age + race_factor, data = nhanes, family = binomial(lin
k = "logit"))
summary(mod.adjsex)
```

```
##
## Call:
## glm(formula = oralmed ~ sex_factor + age + race_factor, family = binomial(link = "logit"),
##     data = nhanes)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.3813  -1.1068  -0.9732   1.1575   1.4041
##
## Coefficients:
##                               Estimate Std. Error z value Pr(>|z|)
## (Intercept)                  -0.081148   0.310389  -0.261  0.79375
## sex_factorFemale             -0.251824   0.108186  -2.328  0.01993
## age                           0.001937   0.004243   0.456  0.64813
## race_factorBlack             -0.245813   0.138224  -1.778  0.07534
## race_factorMexican-American   0.385906   0.132889   2.904  0.00368
## race_factorOther              0.302617   0.352675   0.858  0.39086
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1974.9  on 1424  degrees of freedom
## Residual deviance: 1948.2  on 1419  degrees of freedom
##   (5 observations deleted due to missingness)
## AIC: 1960.2
##
## Number of Fisher Scoring iterations: 4
```

```
# Exponentiated slope coefficient = OR and 95% CI for OR
exp(cbind(OR = coef(mod.adjsex), confint.default(mod.adjsex)))
```

```
##                                    OR     2.5 %    97.5 %
## (Intercept)                 0.9220569 0.5018273 1.6941864
## sex_factorFemale            0.7773814 0.6288490 0.9609968
## age                         1.0019384 0.9936398 1.0103063
## race_factorBlack            0.7820687 0.5964694 1.0254198
## race_factorMexican-American 1.4709471 1.1336573 1.9085884
## race_factorOther            1.3533964 0.6779967 2.7016089
```

- The **fitted model** is given by the equation, $\log\left(\frac{\hat{p}}{1-\hat{p}}\right)$ = -0.0811 − 0.252 Female + 0.002 Age − 0.246 Black + 0.386 Mexican-American + 0.303 Other.
- The **estimated slope** of `sex_factor` b_1 = -0.25 is equal to the age- and race-adjusted *log-odds ratio* of oral diabetes medication use in females vs. males (ref). The exponentiated slope e^{b_1} gives the estimated **adjusted odds ratio**, $\hat{OR} = e^{b_1}$ = 0.78 [95% CI (0.63, 0.96)].
- A **significance test of the slope** ($H_0 : \beta_1 = 0$ vs. $\beta_1 \neq 0$) reports a z-statistic z = -2.33, which is compared to a standard Normal distribution. We have evidence to reject H_0 and conclude that the odds of using oral diabetes medication is still significantly different in females and males even after controlling for age and race (p-value = 0.02).
- We can similarly interpret the sex- and race-adjusted effect of **age** and the sex- and age-adjusted effect of **race**.
- In the **unadjusted model** we saw that there was a significant difference in the odds of oral diabetes medication use in both Blacks vs. Whites and Mexican-Americans vs. Whites. In the adjusted model, we only see a significant difference between Mexican-Americans and Whites. The **estimated slope** of the second dummy variable of `race_factor` (Mexican-American), b_4 = 0.39 is equal to the age- and sex-adjusted *log-odds ratio* of oral diabetes medication use in Mexican-Americans vs. Whites (ref). The age- and sex-**adjusted odds ratio**, $\hat{OR} = e^{b_4}$ = 1.47 [95% CI (1.13, 1.91)]. The odds of oral diabetes medication use is 47% higher in Mexican-Americans compared to Whites, keeping sex and age constant.

```
# Values of x used to estimate p
pred.x <- data.frame(sex_factor = c("Female", "Male"), age = 50, race_factor = "White")

# Returns predicted probabilities
phat <- predict(mod.adjsex, newdata = pred.x, type = "response")
cbind(pred.x, phat)
```

```
##   sex_factor age race_factor      phat
## 1     Female  50       White 0.4412365
## 2       Male  50       White 0.5039196
```

- The **estimated probability of oral diabetes medication use** in a **White**, **50**-year old **female** is equal to $\hat{p}$ = 0.441.
- The **estimated probability of oral diabetes medication use** in a **White**, **50**-year old **male** is equal to $\hat{p}$ = 0.504.
- Again, we see the probability of oral diabetes medication use is lower in females than males, agreeing with the observed adjusted $\hat{OR} < 1$.

Likelihood Ratio Test

The **Likelihood Ratio Test** simultaneously tests the significance of a group or set of parameters. This test is commonly used to test the effect of categorical variables that are naturally made up of more than one dummy variable. For example, to test the significance of **race/ethnicity** in the adjusted model, we would test:
$H_0 : \beta_3 = \beta_4 = \beta_5 = 0$ vs. $H_1 : \beta_3, \beta_4, \beta_5$ not all 0.

Here, we are comparing two **nested models**,

- **Full model**: $\log\left(\frac{p}{1-p}\right) = \alpha + \beta_1$ Female $+\beta_2$ Age $+\beta_3$ Black $+\beta_4$ Mexican-American $+\beta_5$ Other
- **Reduced model** (i.e., model under H_0, without `race_factor`): $\log\left(\frac{p}{1-p}\right) = \alpha + \beta_1$ Female $+\beta_2$ Age

The **likelihood ratio test statistic** is equal to $\mathrm{G} = -2\log\text{-likelihood}(R) - (-2\log\text{-likelihood}(F))$. We can extract the log-likelihood of a model using the `logLik()` function. The test statistic is compared to an Chi-square distribution with *degrees of freedom* equal to the number of parameters tested under H_0, χ^2_{df}.

- **Option 1:** We perform the likelihood ratio test using the `anova()` function to compare two **nested models** using the syntax `anova(reducedmodel, fullmodel, test = "Chisq")`. Just as we did in when performing F-tests, we need to be sure that we are fitting both the full and reduced models using the same data set. Thus, when fitting the *reduced model*, use the observations that were included in the full model by specifying `data=mod.full$model`. Remember that observations with **missing values** for any of the variables involved in fitting the model are dropped or not included in the model. We must compare the full and reduced model on the same data, thus we specify `data=mod.full$model` when fitting `mod.red`.

```
# Full model
mod.full <- glm(oralmed ~ sex_factor + age + race_factor, data = nhanes, family = binomial(link
 = "logit"))
logLik(mod.full)                                    # log-likelihood(F)
```

```
## 'log Lik.' -974.1108 (df=6)
```

```
-2*as.numeric(logLik(mod.full))                     # -2LL(F)
```

```
## [1] 1948.222
```

```
# Reduced model (under H0, does not include race_factor)
# Fit using the same observations included in the full model (data = mod.full$model)
mod.red <- glm(oralmed ~ sex_factor + age, data = mod.full$model, family = binomial(link = "logi
t"))
logLik(mod.red)                                     # log-likelihood(R)
```

```
## 'log Lik.' -984.6616 (df=3)
```

```
-2*as.numeric(logLik(mod.red))                      # -2LL(R)
```

```
## [1] 1969.323
```

```
# Test statistic, G = -2LL(R) - (-2LL(F))
G <- (-2*as.numeric(logLik(mod.red))) - (-2*as.numeric(logLik(mod.full)))
G
```

```
## [1] 21.10159
```

The **likelihood ratio test** test statistic, G = -2(-984.7) - (-2(-974.1)) = 21.1, which is compared to a Chi-square distribution with 3 degrees of freedom, corresponding to the number of paramters tested under $H_0 : \beta_3 = \beta_4 = \beta_5 = 0$. At the $\alpha = 0.05$-level, the critical value of this test is 7.81.

```
# LRT comparing full and reduced models
anova(mod.red, mod.full, test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: oralmed ~ sex_factor + age
## Model 2: oralmed ~ sex_factor + age + race_factor
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1      1422     1969.3
## 2      1419     1948.2  3   21.102 0.0001003
```

The `Analysis of Deviance Table` reports -2 log-likelihood(R) = 1969.3; -2 log-likelihood(F) = 1948.2. As we calculated above, the difference of these two values equals our likelihood ratio test statistic G = 21.1. The overall effect of race is statistically significant in the full model that controls for sex and age (p-value <.001). Thus we have evidence to reject H_0 and conclude that at least one β_3, β_4, or β_5 is not equal to 0.

- **Option 2:** Option 1 is a more flexible option for carrying out a likelihood ratio test and be used to simultaneously test many different slope parameters involving *different* variables (e.g., simultaneously test the effect of `sex_factor` and `age` by testing $H_0 : \beta_1 = \beta_2 = 0$). However, if the goal of the likelihood ratio test is to test $C - 1$ dummy variables of a *single* C-level categorical variable, then we can use the `Anova()` function in the `car` package. The `Anova()` function applied to a model object (e.g., `mod.full`) returns individual likelihood ratio tests for each variable in the model. A reduced model does not need to be explicitly specified in Option 2.

```
# Anova() function in the "car" package
Anova(mod.full)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: oralmed
##             LR Chisq Df Pr(>Chisq)
## sex_factor    5.4287  1  0.0198087
## age           0.2083  1  0.6480721
## race_factor  21.1016  3  0.0001003
```

Based on the output above, the **likelihood ratio test** of all dummy variables that make up `race_factor` has a test statistic G = 21.1, which is compared to an Chi-square distribution with 3 degrees of freedom. As we saw in **Option 1**, the effect of race is statistically significant in the presence of the other variables (p-value <.001).

Model Diagnostics

Finally, we evaluate the fit of the fully-adjusted model.

- **McFadden's** $R^2 = R^2_m = 1 - \frac{\log\text{-likelihood}(Mod)}{\log\text{-likelihood}(Null)}$ is a pseudo-R^2 that ranges between 0 and 1. The closer the value is to 1, the better the model fits the data. We compute this metric by comparing the log-

likelihood of the model under evaluation to the log-likelihood of the null model that contains no predictor variables (intercept only). The better the model fits the data, the larger the log-likelihood of the evaluated model (`mod`) relative to the log-likelihood of the null model, giving a larger R^2_m. The `PsuedoR2(mod, which = "McFadden")` function in the `DescTools` package can be used to report R^2_m.

```
# Model being evaluated
mod.adjsex <- glm(oralmed ~ sex_factor + age + race_factor, data = nhanes, family = binomial(lin
k = "logit"))

# Null model (intercept only)
# Fit using the same observations included in mod.adjsex (data = mod.adjsex$model)
mod.null <- glm(oralmed ~ 1, data = mod.adjsex$model, family = binomial(link="logit"))

# McFadden's R2
R2m <- 1 - as.numeric(logLik(mod.adjsex))/as.numeric(logLik(mod.null))
R2m
```

```
## [1] 0.01349834
```

```
# PseudoR2() function in the "DescTools" package
PseudoR2(mod.adjsex, which = "McFadden")
```

```
##   McFadden
## 0.01349834
```

As we see, McFadden's $R^2 = 0.0135$ is rather low, indicating a poor predictive ability of the adjusted model.

- The **Hosmer-Lemeshow goodness-of-fit test** can be used to determine how well a logistic regression model fits the data and tests,
    - H_0: The model **does** fit the data well vs.
    - H_1: The model **does not** fit the data well

A *significant* Hosmer-Lemeshow test result indicates the model is not correctly specified and does not fit the data well, while a *non-significant* Hosmer-Lemeshow test result indicates there is not enough evidence to conclude that the model does not fit the data well, or the model may fit the data well. The Hosmer-Lemeshow test can be performed to test the model `mod` using the `hoslem.test(Yvariable, fitted(mod))` function in the `ResourceSelection` package. We must input the response variable (Y) and the fitted probabilities from the model in the `hoslem.test()` function. Once again, it is important that we specify the y_i values that are used to fit the model being evaluated. Below, `df.use` is the data frame that contains the observations used to fit the model of interest, `mod.adjsex` . `df.use$oralmed` are the values of Y used to fit the logistic regression model. The fitted probabilities ($\hat{p}$) of `mod.adjsex` are extracted using `fitted(mod.adjsex)` :

```
# Observations included in mod.adjsex (data = mod.adjsex$model)
df.use <- mod.adjsex$model

# Run Hosmer-Lemeshow test: input Y-variable and fitted probabilities from mod.adjsex
HL <- hoslem.test(df.use$oralmed, fitted(mod.adjsex))
HL
```

```
##
##  Hosmer and Lemeshow goodness of fit (GOF) test
##
## data:  df.use$oralmed, fitted(mod.adjsex)
## X-squared = 7.9055, df = 8, p-value = 0.4428
```

We do not have evidence to reject H_0 and cannot conclude that the model does not fit the data well (p-value = 0.4428). Remember, in hypothesis testing, we do not **prove** that H_0 is true. We simply **fail to reject** H_0.