Lab 5 BIS 505b

Maria Ciarleglio

3/22/2021

Goal of Lab 5

In **Lab 5**, we will **(1)** work with categorical predictors, **(2)** create multiple linear regression models, **(3)** explore interactions between variables and **(4)** present automated variable selection methods.

Analysis Data Set

In this lab, we will 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:

Variable Name	Definition
`fastgluc`	Fasting glucose level (mg/dL) (**Our Response**)
	88888 = Missing
`age`	Age at time of survey
`sex`	Sex
	0 = Male
	1 = Female
`oralmed`	Oral diabetes medication use
	0 = No
	1 = Yes
`race`	Race/Ethnicity

Variable Name	Definition
	1 = White
	2 = Black
	3 = Mexican-American
	4 = Other

After reviewing the Data Key, we see that missing values of `fastgluc` are coded as `88888`. Begin by re-coding this numerical value of `88888` as `NA` in **R**.

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

# Checking missing values
summary(nhanes)
```

```
##     fastgluc          age             sex            oralmed
##  Min.   : 42.2   Min.   :30.00   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:108.1   1st Qu.:56.00   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :149.2   Median :66.00   Median :1.0000   Median :0.0000
##  Mean   :177.6   Mean   :64.41   Mean   :0.5629   Mean   :0.4898
##  3rd Qu.:232.9   3rd Qu.:74.00   3rd Qu.:1.0000   3rd Qu.:1.0000
##  Max.   :594.2   Max.   :90.00   Max.   :1.0000   Max.   :1.0000
##  NA's   :264                                      NA's   :5
##       race
##  Min.   :1.000
##  1st Qu.:1.000
##  Median :2.000
##  Mean   :1.945
##  3rd Qu.:3.000
##  Max.   :4.000
##
```

- **Missing Values:**

Note: `fastgluc` contains 264 **missing values**. These observations will not be used to fit any of the models. In addition, individuals with a missing value for any of the explanatory variables included in a specific model will not be used to fit that model. This is an important consideration when thinking about which explanatory variables to include. For example, including a predictor that has a large proportion of missing values in the model will consequently exclude anyone who is missing a value for that variable from the regression model. This is called a **complete case analysis**. That is, only records with complete data on all variables included in a model, `y ~ x1 + x2 + x3`, (complete data for `y`, `x1`, `x2` and `x3`) will be analyzed.

- **Creating Factor Variables:**

There are several **categorical variables** in this data set (`sex`, `oralmed` and `race`). Use the `str()` function to see how each variable is stored in **R** (numeric, integer, factor, character).

```
str(nhanes)
```

```
## 'data.frame':    1430 obs. of  5 variables:
##  $ fastgluc: num  271.7 83.3 107.3 109.4 175.5 ...
##  $ age     : int  48 82 66 80 72 78 31 83 49 63 ...
##  $ sex     : int  0 1 1 1 0 1 0 1 1 1 ...
##  $ oralmed : int  0 0 0 1 0 0 1 1 1 1 ...
##  $ race    : int  3 1 3 1 2 2 2 1 2 1 ...
```

Notice that the variables for race, oral diabetes medication use and sex are coded as integers, but should be coded as **factor variables**. We can use either the `mutate()` function in the `dplyr` package or the `factor()` function directly to redefine these variables as factors based on the Data Key.

```
# Creating factor variables in nhanes, mutate() function in the "dplyr" package
nhanes <- mutate(nhanes,
                 sex_factor = factor(sex,
                                     levels = c(0, 1),
                                     labels = c("Male", "Female")),
                 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"
)))

str(nhanes)
```

```
## 'data.frame':    1430 obs. of  8 variables:
##  $ fastgluc      : num  271.7 83.3 107.3 109.4 175.5 ...
##  $ age           : int  48 82 66 80 72 78 31 83 49 63 ...
##  $ sex           : int  0 1 1 1 0 1 0 1 1 1 ...
##  $ oralmed       : int  0 0 0 1 0 0 1 1 1 1 ...
##  $ race          : int  3 1 3 1 2 2 2 1 2 1 ...
##  $ sex_factor    : Factor w/ 2 levels "Male","Female": 1 2 2 2 1 2 1 2 2 2 ...
##  $ oralmed_factor: Factor w/ 2 levels "No","Yes": 1 1 1 2 1 1 2 2 2 2 ...
##  $ race_factor   : Factor w/ 4 levels "White","Black",..: 3 1 3 1 2 2 2 1 2 1 ...
```

Note: When we include **factor variables** in our regression model, the **first level** specified in the `factor()` function is used as the **reference level**. For example, given the way `sex_factor` is defined above (`levels = c(0, 1)`), a linear regression model that includes `sex_factor` will assume *male* is the reference category. The effect of `sex_factor` then compares females (1) to males (0) (reference). Re-ordering the `levels=` argument of the `factor()` function allows us to specify the desired reference level by listing it first. For example, to report the average difference in fasting glucose level in males (0) vs. females (1) (reference), define `sex_factor` as `factor(sex, levels = c(1, 0), labels = c("Female", "Male"))` .

Research Questions

We are interested in studying characteristics associated with **fasting glucose levels** `fastgluc` (our response variable, y). The **research questions** include:

1. Is there evidence of an association between *age* and *fasting glucose levels*, after controlling for *sex*, *oral diabetes medication use* and *race/ethnicity*?
2. In the multiple regression model, is *race/ethnicity* an important predictor of *fasting glucose level*?
3. Does the effect of *age* on *fasting glucose levels* differ by *oral diabetes medication use*?

Multiple Linear Regression

Multiple linear regression is used to describe the relationship between a quantitative response variable y and more than one explanatory variable $x_1, \ldots, x_k$. The *population regression model*, $\mu_{y|x} = \alpha + \beta_1 x_1 + \ldots + \beta_k x_k$, is estimated using the method of **least squares**, giving a *fitted model*, $\hat{y} = a + b_1 x_1 + \ldots + b_k x_k$.

- The fitted model is used to **predict** or **estimate** the expected value of y for given values of $x_1, \ldots, x_k$ by plugging these values of x into the fitted equation, $\hat{y} = a + b_1 x_1 + \ldots + b_k x_k$.
- The estimated slope b_j is used to **describe the association** between x_j and y, controlling for or holding all other explanatory variables constant.

A test of the slope, β_j (i.e., $H_0 : \beta_j = 0$ vs. $H_1 : \beta_j \neq 0$) is used to determine if an association exists between x_j and y, holding all other explanatory variables constant. This test is performed using the t-statistic, $t = \frac{b_j}{s_{b_j}}$, which is compared to a t-distribution with $n - p$ *degrees of freedom*. Note that p is equal to the number of regression parameters estimated in the model (k slopes + 1 intercept), so $p = k + 1$.

The `lm()` function in **R** is used to estimate the regression coefficients (i.e., the intercept and slope parameter(s)) of the linear model. The result of the `lm()` function is usually saved as an object (e.g., `regobject`) and the `summary()` function is applied to that object (`summary(regobject)`) to output detailed results.

`lm()` **Function Arguments**	**Option Definition**
`formula=`	`analysis_variable ~ predictor_variable1 + predictor_variable2`
`data=`	Data frame containing sample data

A $100(1 - \alpha)$% confidence interval for the parameter β_j is equal to $b_j \pm t_{1-\alpha/2;n-p} \, s_{b_j}$. The `confint(regobject)` function is used to return 95% confidence intervals for the model parameters. By default, 95% confidence intervals (`level=0.95`) are produced.

Finally, the fitted model can be used to estimate or predict a value of y for given values of x using the `predict()` function. A new data frame must be created and specified in the `newdata=` argument of the `predict()` function that contains the values of x used to predict values of y.

Binary Predictor Variable

So far, have only considered **quantitative predictor variables**. **Categorical variables** can be included as predictors in a regression 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) 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
```

- In an **unadjusted model** that contains `sex_factor`, $\hat{y} = a + b_1 z_1$, the estimated slope of `sex_factor` b_1 equals the estimated *difference* in mean fasting glucose in females vs. males (ref) ($\bar{y}_f - \bar{y}_m$). Including additional variables in the model will give *adjusted* differences in mean fasting glucose.

```
# SLR model including sex_factor
mod.sex <- lm(fastgluc ~ sex_factor, data = nhanes)
summary(mod.sex)
```

```
##
## Call:
## lm(formula = fastgluc ~ sex_factor, data = nhanes)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -136.27  -69.36  -28.35   56.20  415.13
##
## Coefficients:
##                  Estimate Std. Error t value            Pr(>|t|)
## (Intercept)       175.724      3.934  44.672 <0.0000000000000002
## sex_factorFemale    3.348      5.260   0.637               0.525
##
## Residual standard error: 89.18 on 1164 degrees of freedom
##   (264 observations deleted due to missingness)
## Multiple R-squared:  0.0003479,  Adjusted R-squared:  -0.0005109
## F-statistic: 0.4052 on 1 and 1164 DF,  p-value: 0.5246
```

```
# Mean of fastgluc by sex_factor
mn <- aggregate(x = list(fastgluc.mean = nhanes$fastgluc),
                by = list(sex = nhanes$sex_factor),
                FUN = mean,
                na.rm = TRUE)
mn
```

```
##      sex fastgluc.mean
## 1   Male      175.7237
## 2 Female      179.0721
```

- The **fitted model** is given by the equation, $\hat{y} = 175.72 + 3.35$ Female. To make the interpretation of the fitted model easier, instead of using the variable name `sex_factor` in the written fitted model, I specified the level of the dummy variable that is being compared to the reference level (i.e., female vs. male).

- The **estimated intercept** $a =$ 175.72 is equal to the mean `fastgluc` when $z_1 = 0$ (i.e., the mean fasting glucose in the reference category (males)).
- The **estimated slope** of `sex_factor` $b_1 =$ 3.35 is equal to the *difference* in mean `fastgluc` in females (179.07) minus males (175.72).
- A **significance test of the slope** ($H_0 : \beta_1 = 0$ vs. $\beta_1 \neq 0$) reports a t-statistic t =0.64, which is compared to a t-distribution with 1164 degrees of freedom. This test does not support a significant difference in the mean fasting glucose in males vs. females (p-value = 0.525).

You can specify the reference category when creating a factor variable by listing that category as the **first level** in the `levels=` argument of the `factor()` function. For example, `sex_factorv2` will set *female* as the reference category. Notice that the dummy variable for `sex_factorv2` equals `1` for males and `0` for females.

```
# Female will be the reference category (=0) of sex_factorv2
nhanes$sex_factorv2 <- factor(nhanes$sex,
                              levels = c(1, 0),
                              labels = c("Female", "Male"))
contrasts(nhanes$sex_factorv2)
```

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

The estimated slope of `sex_factorv2` is equal to the estimated difference in mean fasting glucose in males vs. females (ref) ($\bar{y}_m - \bar{y}_f$).

```
# SLR model including sex_factorv2
mod.sexv2 <- lm(fastgluc ~ sex_factorv2, data = nhanes)
summary(mod.sexv2)
```

```
## 
## Call:
## lm(formula = fastgluc ~ sex_factorv2, data = nhanes)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -136.27  -69.36  -28.35   56.20  415.13 
## 
## Coefficients:
##                  Estimate Std. Error t value            Pr(>|t|)
## (Intercept)       179.072      3.493  51.271 <0.0000000000000002
## sex_factorv2Male   -3.348      5.260  -0.637               0.525
## 
## Residual standard error: 89.18 on 1164 degrees of freedom
##   (264 observations deleted due to missingness)
## Multiple R-squared:  0.0003479,  Adjusted R-squared:  -0.0005109 
## F-statistic: 0.4052 on 1 and 1164 DF,  p-value: 0.5246
```

We can also use the `relevel()` function to change the reference category (`ref=`) of an existing factor variable:

```
# Female will be the reference category (=0) of sex_factorv3
nhanes$sex_factorv3 <- relevel(nhanes$sex_factor, ref = "Female")
contrasts(nhanes$sex_factorv3)
```

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

Categorical Predictor 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

- In an **unadjusted model**, $\hat{y} = a + b_1 z_1 + b_2 z_2 + b_3 z_3$, the estimated slope of the first dummy variable b_1 equals the estimated difference in mean fasting glucose in Blacks vs. Whites (ref) ($\bar{y}_b - \bar{y}_w$). The estimated slope of the second dummy variable b_2 equals the estimated difference in mean fasting glucose in Mexican-Americans vs. Whites (ref) ($\bar{y}_m - \bar{y}_w$). The estimated slope of the third dummy variable b_3 equals the estimated difference in mean fasting glucose in Others vs. Whites (ref) ($\bar{y}_o - \bar{y}_w$). Including additional variables in the model will give *adjusted* differences in mean fasting glucose.

```
# Model including race_factor
mod.race <- lm(fastgluc ~ race_factor, data = nhanes)
summary(mod.race)
```

```
##
## Call:
## lm(formula = fastgluc ~ race_factor, data = nhanes)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -139.60  -67.91  -28.15   54.01  405.30
##
## Coefficients:
##                             Estimate Std. Error t value            Pr(>|t|)
## (Intercept)                  168.222      4.220  39.860 < 0.0000000000000002
## race_factorBlack              20.677      6.547   3.158             0.00163
## race_factorMexican-American   12.454      6.233   1.998             0.04595
## race_factorOther              -7.674     16.029  -0.479             0.63221
##
## Residual standard error: 88.83 on 1162 degrees of freedom
##   (264 observations deleted due to missingness)
## Multiple R-squared:  0.009968,   Adjusted R-squared:  0.007412
## F-statistic:   3.9 on 3 and 1162 DF,  p-value: 0.008699
```

- The **fitted model** is given by the equation, $\hat{y} =$ 168.22 $+$ 20.68 Black $+$ 12.45 Mexican-American $-$ 7.67 Other.
- The average fasting glucose in Blacks is $b_1 =$ 20.68 [95% CI (7.83, 33.52)] units higher than in Whites. A **significance test** of β_1 shows that there is a significant difference in the mean fasting glucose in Blacks vs. Whites (p-value = 0.002).
- The average fasting glucose in Mexican-Americans is $b_2 =$ 12.45 [95% CI (0.22, 24.68)] units higher than in Whites. A **significance test** of β_2 shows that there is a significant difference in the mean fasting glucose in Mexican-Americans vs. Whites (p-value = 0.046).
- The average fasting glucose in Other races/ethnicities is $b_3 =$ -7.67 [95% CI (-39.12, 23.77)] units different (lower) than in Whites. A **significance test** of β_3 shows that there is not a significant difference in the mean fasting glucose in Other races/ethnicities vs. Whites (p-value = 0.632).

The Unadjusted Model

The first research question asks if there is an association between *age* (`age`) and *fasting glucose levels* (`fastgluc`), after controlling for *sex* (`sex_factor`), *oral diabetes medication use* (`oralmed_factor`) and *race/ethnicity* (`race_factor`). Let's begin by estimating the **unadjusted effect** of `age` on `fastgluc` using a **simple linear regression model**. We call this the *unadjusted effect* because no other variables are being controlled for or adjusted for in the regression model.

- **Unadjusted model**: $\mu_{y|x} = \alpha + \beta_1$ Age

We fit the model using the `lm()` function and save the fitted model to the object `mod.age` . We output a summary of the results using `summary(mod.age)` and the 95% CIs of the model parameters using `confint(mod.age)` .

```r
# SLR model including age
mod.age <- lm(fastgluc ~ age, data = nhanes)

# Output results of fitted model
summary(mod.age)
```

```
##
## Call:
## lm(formula = fastgluc ~ age, data = nhanes)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -162.13  -67.00  -26.35   52.64  414.70
##
## Coefficients:
##             Estimate Std. Error t value             Pr(>|t|)
## (Intercept) 228.3632    12.7491  17.912 < 0.0000000000000002
## age          -0.8011     0.1970  -4.067            0.0000508
##
## Residual standard error: 88.57 on 1164 degrees of freedom
##   (264 observations deleted due to missingness)
## Multiple R-squared:  0.01401,    Adjusted R-squared:  0.01316
## F-statistic: 16.54 on 1 and 1164 DF,  p-value: 0.00005081
```

```r
# Confidence intervals for model parameters (intercept and slope)
confint(mod.age)
```

```
##                  2.5 %      97.5 %
## (Intercept) 203.349349 253.3769563
## age          -1.187535  -0.4146312
```

- The **fitted model** is given by the equation, $\hat{y} =$ 228.36 $-$ 0.8 Age.
- The **estimated slope** of `age` $b_1 =$ -0.8 [95% CI (-1.19, -0.41)] indicates that a 1-unit increase in age is associated with a -0.8-unit average change (a decrease) in fasting glucose.
- A **significance test of the slope** ($H_0 : \beta_1 = 0$ vs. $\beta_1 \neq 0$) reports a t-statistic t =-4.07, which is compared to a t-distribution with 1164 degrees of freedom. This yields a highly significant p-value <.001.
- The **R-squared** of this model is low at 0.014, indicating that `age` only explains 1.4% of the total variability in fasting glucose. The variability about the regression line $\sigma_{y|x}$ is estimated by the "**residual standard error**" in the output above and is equal to $s_{y|x}$ = 88.57.

The Adjusted Model

To **control** or **adjust for** *sex*, *oral diabetes medication use* and *race/ethnicity* when estimating the effect of *age* on *fasting glucose*, we will fit a **multiple linear regression model** that additionally includes the variables `sex_factor`, `oralmed_factor` and `race_factor`.

- **Adjusted model**: $\mu_{y|x} = \alpha + \beta_1$ Age $+\beta_2$ Female $+\beta_3$ MedUse $+\beta_4$ Black $+\beta_5$ Mexican-American $+\beta_6$ Other

```
# MLR model including age, sex_factor, oralmed_factor, race_factor
mod.mlr <- lm(fastgluc ~ age + sex_factor + oralmed_factor + race_factor, data = nhanes)
summary(mod.mlr)
```

```
##
## Call:
## lm(formula = fastgluc ~ age + sex_factor + oralmed_factor + race_factor,
##     data = nhanes)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -153.65  -65.29  -27.40   49.84  399.79
##
## Coefficients:
##                            Estimate Std. Error t value            Pr(>|t|)
## (Intercept)                211.0088    15.1442  13.933 < 0.0000000000000002
## age                         -0.7065     0.2061  -3.428            0.000629
## sex_factorFemale             2.0178     5.2643   0.383            0.701578
## oralmed_factorYes            9.6060     5.2393   1.833            0.066994
## race_factorBlack            14.8757     6.8005   2.187            0.028910
## race_factorMexican-American   6.0303     6.4286   0.938            0.348418
## race_factorOther           -15.4705    16.1097  -0.960            0.337094
##
## Residual standard error: 88.45 on 1156 degrees of freedom
##   (267 observations deleted due to missingness)
## Multiple R-squared:  0.02231,    Adjusted R-squared:  0.01723
## F-statistic: 4.396 on 6 and 1156 DF,  p-value: 0.0002109
```

```
confint(mod.mlr)
```

```
##                                  2.5 %      97.5 %
## (Intercept)                181.2955887 240.7219343
## age                         -1.1108063  -0.3021315
## sex_factorFemale            -8.3109948  12.3464959
## oralmed_factorYes           -0.6736602  19.8856553
## race_factorBlack             1.5330274  28.2184219
## race_factorMexican-American  -6.5827490  18.6433961
## race_factorOther           -47.0779068  16.1369944
```

- The **fitted model** is given by the equation, $\hat{y} =$ 211.01 $-$ 0.71 Age $+$ 2.02 Female $+$ 9.61 MedUse $+$ 14.88 Black $+$ 6.03 Mexican-American $-$ 15.47 Other.
- Sex, medication use, and race-adjusted effect of **age** on fasting glucose:
  - The **estimated slope** of `age` $b_1 =$ -0.71 [95% CI (-1.11, -0.3)] in the multiple linear regression model indicates that a 1-unit increase in age is associated with a -0.71-unit average change (a decrease) in fasting glucose, controlling for sex, oral diabetes medication use and race.
  - A **significance test of the slope** ($H_0 : \beta_1 = 0$ vs. $\beta_1 \neq 0$) shows a highly significant association between age and fasting glucose when controlling for sex, oral diabetes medication use and race (p-value <.001).

- Age, medication use, and race-adjusted effect of **sex** on fasting glucose:
    - The **estimated slope** of `sex_factor` $b_2 =$ 2.02 [95% CI (-8.31, 12.35)] estimates the average difference fasting glucose in females vs. males (reference), controlling for age, oral diabetes medication use and race. The adjusted average fasting glucose level is 2.02-units higher in females than in males.
    - A **significance test of the slope** ($H_0: \beta_2 = 0$ vs. $\beta_2 \neq 0$) shows there is not a statistically significant difference in average fasting glucose in females and males when controlling for age, oral diabetes medication use and race (p-value = 0.702). Notice that the 95% confidence interval for β_2 supports this conclusion since it includes 0 (i.e., the value of β_2 hypothesized under H_0).
    - Since sex is not a significant predictor in the presence of age, oral diabetes medication use and race, we may want to remove this variable from the regression model. However, if **confounding** is a concern, we can retain the variable regardless of statistical significance.

- The **Adjusted R-squared** (R^2_a) of this model remains low at 0.017. The **residual standard error** is equal to $s_{y|x}$ = 88.45, and is only slightly smaller than the estimate from the unadjusted model.

Exercise: Interpret the effect of oral diabetes medication use in the model above.

▶ Answer:

Exercise: Interpret the effect of race/ethnicity in the model above.

▶ Answer:

Overall F-Test

The **Overall F-Test** tests whether the explanatory variables collectively have an effect on the response variable. Under H_0, $\beta_1 = \beta_2 = \ldots = \beta_k = 0$. Under H_1, at least one $\beta_j \neq 0$ for $j = 1, \ldots, k$. The **F-test statistic** is equal to the ratio of the variability explained by the *model* (MSM) to the mean squared *error* (MSE), giving $F = \frac{MSM}{MSE}$. The **F-test statistic** is compared to an F-distribution with *numerator degrees of freedom* k and *denominator degrees of freedom* $n - p$.

- **Option 1:** The overall F-test is presented in the last line of the `summary()` of model results.

```
# Overall F-test on last line of output
summary(mod.mlr)
```

```
##
## Call:
## lm(formula = fastgluc ~ age + sex_factor + oralmed_factor + race_factor,
##     data = nhanes)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -153.65  -65.29  -27.40   49.84  399.79
##
## Coefficients:
##                              Estimate Std. Error t value             Pr(>|t|)
## (Intercept)                  211.0088    15.1442  13.933 < 0.0000000000000002
## age                           -0.7065     0.2061  -3.428             0.000629
## sex_factorFemale               2.0178     5.2643   0.383             0.701578
## oralmed_factorYes              9.6060     5.2393   1.833             0.066994
## race_factorBlack              14.8757     6.8005   2.187             0.028910
## race_factorMexican-American    6.0303     6.4286   0.938             0.348418
## race_factorOther             -15.4705    16.1097  -0.960             0.337094
##
## Residual standard error: 88.45 on 1156 degrees of freedom
##   (267 observations deleted due to missingness)
## Multiple R-squared:  0.02231,    Adjusted R-squared:  0.01723
## F-statistic: 4.396 on 6 and 1156 DF,  p-value: 0.0002109
```

The **overall F-test** F-statistic, F = 4.396, is compared to an F-distribution with 6 and 1156 degrees of freedom, giving a p-value <.001. There is evidence to conclude that a significant association exists between fasting glucose level and at least one explanatory variable in the MLR model.

- **Option 2:** We can request the ANOVA table of a model using the `anova()` function on the model object (e.g., `mod.mlr`). Note that the ANOVA table has the contribution to the sum of squares broken down by predictor ("Sequential SS" or "Type I SS") and does **not** provide the overall F-statistic. To find the Model SS (SSM), add the sum of squares for all predictors. To find the MSM, divide SSM by the number of independent variables, k, which is the number of parameters tested under H_0 (i.e., 6 in this case; recall that race is represented by 3 dummy variables).

```
# ANOVA table
anova(mod.mlr)
```

```
## Analysis of Variance Table
##
## Response: fastgluc
##                  Df  Sum Sq Mean Sq F value     Pr(>F)
## age               1  129225  129225 16.5177 0.00005146
## sex_factor        1    1574    1574  0.2012    0.65382
## oralmed_factor    1   22324   22324  2.8535    0.09145
## race_factor       3   53238   17746  2.2683    0.07898
## Residuals      1156 9043883    7823
```

The **overall F-test** F-statistic, F = [(129225 + 1574 + 22324 + 53238)/(1 + 1 + 1 + 3)]/7823 = 4.396 agrees with the F-statistic reported in the `summary()` output.

- **Option 3:** Finally, we can perform the overall F-test using the `anova()` function to compare two **nested models** using the syntax `anova(reducedmodel, fullmodel)`.
    - **Full model**: $\mu_{y|x} = \alpha + \beta_1$ Age $+\beta_2$ Female $+\beta_3$ MedUse $+\beta_4$ Black $+\beta_5$ Mexican-American $+\beta_6$ Other
    - **Reduced model** (i.e., model under H_0): $\mu_{y|x} = \alpha$

Under the reduced model, H_0 is assumed to be true (i.e., $\beta_1 = \beta_2 = \ldots = \beta_6 = 0$). When performing the overall F-test, the reduced model is also known as the **null model** since it contains only the intercept term and no explanatory variables. The **R** syntax for fitting a model that contains only the intercept is `yvariable ~ 1`, where `1` represents the intercept.

Note: Recall that each regression model only includes records with **complete data** on all variables included in that model (**complete case analysis**). Since there are some individuals with missing values for `oralmed_factor`, the sample size used to fit `mod.full` is *smaller* than the sample size used to fit `mod.null`. To compare the same subset of observations under both the full and reduced models, we must specify that the analysis data set used to fit `mod.full` should also be used to fit `mod.null`. The "complete case" data frame used to fit `mod.full` is `data = mod.full$model` and is specified as the data source for `mod.null`:

```
# Full model
mod.full <- lm(fastgluc ~ age + sex_factor + oralmed_factor + race_factor, data = nhanes)

# Reduced model (null model) fit using the same observations used to fit full model
mod.null <- lm(fastgluc ~ 1, data = mod.full$model)  # note **data=** here!

# F-test comparing full and reduced models
anova(mod.null, mod.full)
```

```
## Analysis of Variance Table
##
## Model 1: fastgluc ~ 1
## Model 2: fastgluc ~ age + sex_factor + oralmed_factor + race_factor
##   Res.Df     RSS Df Sum of Sq      F    Pr(>F)
## 1   1162 9250243
## 2   1156 9043883  6    206360 4.3962 0.0002109
```

The **overall F-test** F-statistic, F = [206360/6]/[9043883/1156] = 4.396 agrees with the F-statistic reported in the `summary()` output.

Partial F-Test

The **Partial F-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:
$H_0 : \beta_4 = \beta_5 = \beta_6 = 0$ vs. $H_1 : \beta_4, \beta_5, \beta_6$ not all 0.

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

- **Full model**: $\mu_{y|x} = \alpha + \beta_1$ Age $+\beta_2$ Female $+\beta_3$ MedUse $+\beta_4$ Black $+\beta_5$ Mexican-American $+\beta_6$ Other

- **Reduced model** (i.e., model under H_0, without `race_factor`): $\mu_{y|x} = \alpha + \beta_1$ Age $+\beta_2$ Female $+\beta_3$ MedUse

The **partial F-test statistic** is equal to $\mathrm{F}_0 = \frac{\frac{SSM(F)-SSM(R)}{\text{Number parameters tested under } H_0}}{\frac{SSE(F)}{df_2(F)}}$.

The F-statistic is compared to an F-distribution with *numerator degrees of freedom* equal to the number of parameters tested under H_0 and *denominator degrees of freedom* $n - p$.

- **Option 1:** We can perform the partial F-test using the `anova()` function to compare two **nested models** using the syntax `anova(reducedmodel, fullmodel)` .

As above, 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` .

```
# Full model
mod.full <- lm(fastgluc ~ age + sex_factor + oralmed_factor + race_factor, data = nhanes)

# Reduced model (under H0, does not include race_factor)
# Fit using the same observations included in the full model
mod.red <- lm(fastgluc ~ age + sex_factor + oralmed_factor, data = mod.full$model)

# F-test comparing full and reduced models
anova(mod.red, mod.full)
```

```
## Analysis of Variance Table
##
## Model 1: fastgluc ~ age + sex_factor + oralmed_factor
## Model 2: fastgluc ~ age + sex_factor + oralmed_factor + race_factor
##   Res.Df     RSS Df Sum of Sq      F  Pr(>F)
## 1   1159 9097120
## 2   1156 9043883  3     53238 2.2683 0.07898
```

The **partial F-test** F-statistic, F = [53238/3]/[9043883/1156] = 2.268 is compared to an F-distribution with 3 and 1156 degrees of freedom. The effect of race is not statistically significant in the full model (p = 0.079), thus we cannot reject H_0. Although we did see that β_4 (coefficient for dummy variable of Black vs. White) was significantly different from 0 in the individual t-tests of the slopes in `mod.mlr` , perhaps the effect was not strong enough to outweigh the lack of statistical significance seen in the other two dummy variables that make up `race_factor` .

- **Option 2:** Option 1 is a more flexible option for carrying out a partial F-test and be used to simultaneously test many different slope parameters involving *different* variables (e.g., simultaneously test the effect of `age` and `sex_factor` by testing $H_0 : \beta_1 = \beta_2 = 0$). However, if the goal of the partial F-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 F-tests for each variable in the model. A reduced model does not need to be explicitly specified in Option 2.

To test the effect of `race_factor` in a model containing age, sex and oral diabetes medication use, $\mu_{y|x} = \alpha + \beta_1$ Age $+\beta_2$ Female $+\beta_3$ MedUse $+\beta_4$ Black $+\beta_5$ Mexican-American $+\beta_6$ Other, we would test $H_0 : \beta_4 = \beta_5 = \beta_6 = 0$ vs. $H_1 : \beta_4, \beta_5, \beta_6$ not all 0.

```
mod.full <- lm(fastgluc ~ age + sex_factor + oralmed_factor + race_factor, data = nhanes)

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

```
## Anova Table (Type II tests)
##
## Response: fastgluc
##                 Sum Sq   Df F value    Pr(>F)
## age              91939    1 11.7518 0.0006292
## sex_factor        1149    1  0.1469 0.7015782
## oralmed_factor   26298    1  3.3615 0.0669941
## race_factor      53238    3  2.2683 0.0789751
## Residuals      9043883 1156
```

Based on the output above, the **partial F-test** of all dummy variables that make up `race_factor` has an F-statistic = [53238/3]/[9043883/1156] = 2.268, which is compared to an F-distribution with 3 and 1156 degrees of freedom. As we saw above, the effect of race is not statistically significant in the presence of the other variables (p = 0.079), thus we cannot reject H_0.

Interactions

Next, we would like to determine if the effect of *age* on *fasting glucose level* differs by *oral diabetes medication use*. Answering this question involves examining the **interaction** between `age` and `oralmed_factor`. The model that includes the interaction `age*oralmed_factor` also includes the main effects of `age` and `oralmed_factor`,

- **Interaction model**: $\mu_{y|x} = \alpha + \beta_1$ Age $+\beta_2$ MedUse $+\beta_3$ Age $\times$ MedUse
- Model in those who use oral diabetes medication (MedUse = 1): $\mu_{y|x} = (\alpha + \beta_2) + (\beta_1 + \beta_3)$ Age
- Model in those who do not use oral diabetes medication (reference) (MedUse = 0): $\mu_{y|x} = \alpha + \beta_1$ Age

Thus, a test of β_3 will determine if there is a significant interaction between age and oral diabetes medication use. If a significant interaction exists, then we can estimate the slope of age (i.e., the effect of age on fasting glucose level) separately in the two medication use categories. Note that for simplicity, I am examining this interaction in a model that does not control for sex or race. However, you could easily also include these variables in the model to examine the interaction while controlling for sex and race.

```
# Interaction model of age*oralmed_factor
mod.intx <- lm(fastgluc ~ age + oralmed_factor + age*oralmed_factor, data = nhanes)
summary(mod.intx)
```

```
##
## Call:
## lm(formula = fastgluc ~ age + oralmed_factor + age * oralmed_factor,
##     data = nhanes)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -146.12  -66.85  -26.22   51.04  409.13
##
## Coefficients:
##                        Estimate Std. Error t value            Pr(>|t|)
## (Intercept)            201.6624    17.1431  11.763 <0.0000000000000002
## age                     -0.4448     0.2655  -1.675              0.0941
## oralmed_factorYes       59.5417    25.6102   2.325              0.0202
## age:oralmed_factorYes   -0.8033     0.3957  -2.030              0.0426
##
## Residual standard error: 88.45 on 1159 degrees of freedom
##   (267 observations deleted due to missingness)
## Multiple R-squared:  0.01979,    Adjusted R-squared:  0.01726
## F-statistic: 7.801 on 3 and 1159 DF,  p-value: 0.00003703
```

```
confint(mod.intx)
```

```
##                              2.5 %       97.5 %
## (Intercept)            168.0274868 235.29740684
## age                     -0.9658059   0.07613036
## oralmed_factorYes        9.2940930 109.78933801
## age:oralmed_factorYes   -1.5797124  -0.02697211
```

- The **fitted model** is given by the equation, $\hat{y} =$ 201.66 $-$ 0.44 Age $+$ 59.54 MedUse $-$ 0.8 Age $\times$ MedUse.

- The **estimated slope** of the interaction `age*oralmed_factor` $b_3 =$ -0.8 [95% CI (-1.58, -0.03)] is equal to the *difference* in the slope of age in those who use oral diabetes medication vs. those who do not (ref).

- A **significance test of the interaction term** ($H_0 : \beta_3 = 0$ vs. $\beta_3 \neq 0$) reports a t-statistic t =-2.03, which is compared to a t-distribution with 1159 degrees of freedom. This test supports a significant difference in the effect of age on fasting glucose level in those who use oral diabetes medication vs. those who do not (p-value = 0.043).

The effect (slope) of age on fasting glucose level in those who use oral diabetes medication is estimated by $b_1 + b_3$; the effect (slope) of age in those who do not use oral diabetes medication is estimated by b_1. We can use **R** to compute the slope in the medication use group and perform a hypothesis test to determine if age significantly affects fasting glucose level in those who use oral diabetes medication by testing $H_0 : \beta_1 + \beta_3 = 0$ vs. $H_1 : \beta_1 + \beta_3 \neq 0$.

We can estimate this slope and test this **linear contrast** by using the `glht()` function in the `multcomp` package. We begin by specifying `K`, which identifies the coefficients that are involved in the estimation (i.e., $b_1 + b_3$), or `c(0, 1, 0, 1)`. We then specify `K` in the `linfct=` argument of the `glht()` function to specify the linear hypothesis to be tested. `summary()` returns the estimate of the effect and the hypothesis test results; `confint()` returns a confidence interval for the effect.

```
# b1 + b3: Effect of age in those with oralmed_factor = 1

# Vector that specifies linear combination of coefficients interested in
K <- rbind(c(0, 1, 0, 1))   # 1 = coefficients "on" when estimating slope in oralmed_factor = 1

# Label for comparison (printed in the output)
rownames(K) <- "b1+b3 (slope in oralmed_factor = 1)"

# Estimate of slope (b1 + b3) and hypothesis test, glht() function in the "multcomp" package
summary(glht(mod.intx, linfct = K))
```

```
## 
##   Simultaneous Tests for General Linear Hypotheses
## 
## Fit: lm(formula = fastgluc ~ age + oralmed_factor + age * oralmed_factor, 
##     data = nhanes)
## 
## Linear Hypotheses:
##                                          Estimate Std. Error t value  Pr(>|t|)
## b1+b3 (slope in oralmed_factor = 1) == 0  -1.2482     0.2934  -4.254 0.0000226
## (Adjusted p values reported -- single-step method)
```

```
# Confidence interval for beta1 + beta3
confint(glht(mod.intx, linfct = K))
```

```
## 
##   Simultaneous Confidence Intervals
## 
## Fit: lm(formula = fastgluc ~ age + oralmed_factor + age * oralmed_factor, 
##     data = nhanes)
## 
## Quantile = 1.962
## 95% family-wise confidence level
##  
## 
## Linear Hypotheses:
##                                          Estimate lwr     upr    
## b1+b3 (slope in oralmed_factor = 1) == 0 -1.2482  -1.8238 -0.6726
```

- Effect of **age** on fasting glucose levels in those who *use oral diabetes medication* is estimated by $b_1 + b_3 =$ -1.25 [95% CI (-1.82, -0.67)]. We have evidence to reject $H_0: \beta_1 + \beta_3 = 0$ and conclude that there is a significant association between age and fasting blood glucose in those who use oral diabetes medication (p-value <.001).

- Effect of **age** on fasting glucose levels in those who *do not use oral diabetes medication* is estimated by $b_1 =$ -0.44 [95% CI (-0.97, 0.08)]. We do not have evidence to reject $H_0: \beta_1 = 0$ and cannot conclude that there is a significant association between age and fasting blood glucose in those who do not use oral diabetes medication (p-value = 0.094).

- The **estimated slope** of the interaction $b_3 =$ -0.8 is equal to the *difference* in the slope of age in those who use oral diabetes medication vs. those who do not (ref). The test of the interaction is telling us that there is a significant difference in the effect of age in these two groups.

Automated Variable Selection

Automated variable selection methods have been developed to choose the "best-fitting" model (i.e., the "best" subset of predictors). There are three automated variable selection procedures:

1. **Backward elimination** begins with the full model and iteratively removes predictors that contribute least to the model until all variables remaining exceed a certain threshold. Once a variable is removed, it cannot re-enter the model. Backward elimination tends to be helpful if you have a modest-sized model and would like to eliminate a few predictors.
2. **Forward selection** begins with the null model (no predictors) and iteratively adds the most important predictors, stopping when there the amount of improvement is below a certain threshold. Once a variable is entered into the model, it cannot be removed. Forward selection tends to be helpful if you have a large set of potential predictors and wish to identify a few important variables.
3. **Stepwise selection** is a combination of forward selection and backward elimination. This method begins with the null model and iteratively adds predictors. After each addition, there is the option of removing any of the variables already included if removing that variable improves the model fit.

Automated selection methods can be based variable p-value thresholds or other model fit statistics, such as R^2_a. The **Akaike Information Criterion** (**AIC**) and the **Bayes Information Criterion** (**BIC**) are other commonly used criteria. The goal is to *minimize* AIC and BIC (i.e., smaller AIC and BIC is better). Just like R^2_a, both of these statistics penalize larger models and will not automatically decrease when additional variables are added to the model.

The `stepAIC()` function in the `MASS` package can be used to conduct automated variable selection based on the AIC. The **null model** (intercept only model) and the **full model** (model that includes all candidate predictors) must be defined.

`stepAIC()` Function Arguments	**Option Definition**
`object=`	Model object (full model for backward elimination, null model for forward and stepwise selection)
`scope=`	Range of models `=list(lower = nullmodel, upper = fullmodel)`
`direction=`	`=both` (stepwise), `=backward` (backward), `=forward` (forward)

Again, to avoid error messages about missing observations resulting in different data sets used in the null model and the full model, we fit the null model using the observations included in the full model `data=mod.full$model`.

```
# Full model (contains all predictors under consideration)
mod.full <- lm(fastgluc ~ age + sex_factor + oralmed_factor + race_factor, data = nhanes)

# Null model (intercept only, notice data= here)
mod.null <- lm(fastgluc ~ 1, data = mod.full$model)
```

- **Backward elimination** stops when the AIC does not improve (i.e., does not decrease) after removing a predictor.

```
# Backward elimination, stepAIC() function in the "MASS" package
stepAIC(mod.full, scope = list(lower = mod.null, upper = mod.full),
        data = nhanes, direction = 'backward')
```

```
## Start:  AIC=10433.13
## fastgluc ~ age + sex_factor + oralmed_factor + race_factor
## 
##                  Df Sum of Sq     RSS   AIC
## - sex_factor      1      1149 9045032 10431
## <none>                        9043883 10433
## - race_factor     3     53238 9097120 10434
## - oralmed_factor  1     26298 9070181 10434
## - age             1     91939 9135822 10443
## 
## Step:  AIC=10431.28
## fastgluc ~ age + oralmed_factor + race_factor
## 
##                  Df Sum of Sq     RSS   AIC
## <none>                        9045032 10431
## - race_factor     3     54361 9099393 10432
## - oralmed_factor  1     25805 9070837 10433
## - age             1     92232 9137264 10441
```

```
## 
## Call:
## lm(formula = fastgluc ~ age + oralmed_factor + race_factor, data = nhanes)
## 
## Coefficients:
##                (Intercept)                          age  
##                   212.1288                      -0.7075  
##          oralmed_factorYes             race_factorBlack  
##                     9.5030                      15.1203  
## race_factorMexican-American             race_factorOther  
##                     6.1850                     -15.1268
```

- **Forward selection** stops when the AIC does not improve after adding a predictor.

```
# Forward selection
stepAIC(mod.null, scope = list(lower = mod.null, upper = mod.full),
        data = nhanes, direction = 'forward')
```

```
## Start:  AIC=10447.37
## fastgluc ~ 1
## 
##                  Df Sum of Sq     RSS   AIC
## + age             1    129225 9121018 10433
## + race_factor     3     89134 9161108 10442
## + oralmed_factor  1     19536 9230707 10447
## <none>                        9250243 10447
## + sex_factor      1      3123 9247120 10449
## 
## Step:  AIC=10433.01
## fastgluc ~ age
## 
##                  Df Sum of Sq     RSS   AIC
## + oralmed_factor  1     21625 9099393 10432
## + race_factor     3     50181 9070837 10433
## <none>                        9121018 10433
## + sex_factor      1      1574 9119444 10435
## 
## Step:  AIC=10432.25
## fastgluc ~ age + oralmed_factor
## 
##               Df Sum of Sq     RSS   AIC
## + race_factor  3     54361 9045032 10431
## <none>                     9099393 10432
## + sex_factor   1      2273 9097120 10434
## 
## Step:  AIC=10431.28
## fastgluc ~ age + oralmed_factor + race_factor
## 
##              Df Sum of Sq     RSS   AIC
## <none>                    9045032 10431
## + sex_factor  1    1149.3 9043883 10433
```

```
## 
## Call:
## lm(formula = fastgluc ~ age + oralmed_factor + race_factor, data = mod.full$model)
## 
## Coefficients:
##                 (Intercept)                          age  
##                    212.1288                      -0.7075  
##           oralmed_factorYes             race_factorBlack  
##                      9.5030                      15.1203  
## race_factorMexican-American             race_factorOther  
##                      6.1850                     -15.1268
```

- **Stepwise selection** stops when the AIC does not improve after potentially adding or removing a predictor.

```
# Stepwise selection
stepAIC(mod.null, scope = list(lower = mod.null, upper = mod.full),
        data = nhanes, direction = 'both')
```

```
## Start:  AIC=10447.37
## fastgluc ~ 1
## 
##                  Df Sum of Sq     RSS   AIC
## + age             1    129225 9121018 10433
## + race_factor     3     89134 9161108 10442
## + oralmed_factor  1     19536 9230707 10447
## <none>                        9250243 10447
## + sex_factor      1      3123 9247120 10449
## 
## Step:  AIC=10433.01
## fastgluc ~ age
## 
##                  Df Sum of Sq     RSS   AIC
## + oralmed_factor  1     21625 9099393 10432
## + race_factor     3     50181 9070837 10433
## <none>                        9121018 10433
## + sex_factor      1      1574 9119444 10435
## - age             1    129225 9250243 10447
## 
## Step:  AIC=10432.25
## fastgluc ~ age + oralmed_factor
## 
##                  Df Sum of Sq     RSS   AIC
## + race_factor     3     54361 9045032 10431
## <none>                        9099393 10432
## - oralmed_factor  1     21625 9121018 10433
## + sex_factor      1      2273 9097120 10434
## - age             1    131314 9230707 10447
## 
## Step:  AIC=10431.28
## fastgluc ~ age + oralmed_factor + race_factor
## 
##                  Df Sum of Sq     RSS   AIC
## <none>                        9045032 10431
## - race_factor     3     54361 9099393 10432
## - oralmed_factor  1     25805 9070837 10433
## + sex_factor      1      1149 9043883 10433
## - age             1     92232 9137264 10441
```

```
## 
## Call:
## lm(formula = fastgluc ~ age + oralmed_factor + race_factor, data = mod.full$model)
## 
## Coefficients:
##                (Intercept)                          age  
##                   212.1288                      -0.7075  
##          oralmed_factorYes             race_factorBlack  
##                     9.5030                      15.1203  
## race_factorMexican-American             race_factorOther  
##                     6.1850                     -15.1268
```

All three selection procedures exclude sex from the selected model but retain age, oral diabetes medication use and race. We consistently saw that sex was not an important predictor of fasting glucose levels. The selected model is given by,

```
# Selected model
mod.selected <- lm(fastgluc ~ age + oralmed_factor + race_factor, data = nhanes)
summary(mod.selected)
```

```
##
## Call:
## lm(formula = fastgluc ~ age + oralmed_factor + race_factor, data = nhanes)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -154.89  -65.19  -27.71   50.56  400.61
##
## Coefficients:
##                            Estimate Std. Error t value             Pr(>|t|)
## (Intercept)                212.1288    14.8541  14.281 < 0.0000000000000002
## age                         -0.7075     0.2060  -3.435             0.000614
## oralmed_factorYes            9.5030     5.2305   1.817             0.069501
## race_factorBlack            15.1203     6.7680   2.234             0.025668
## race_factorMexican-American   6.1850     6.4136   0.964             0.335063
## race_factorOther           -15.1268    16.0788  -0.941             0.347008
##
## Residual standard error: 88.42 on 1157 degrees of freedom
##   (267 observations deleted due to missingness)
## Multiple R-squared:  0.02218,    Adjusted R-squared:  0.01796
## F-statistic:  5.25 on 5 and 1157 DF,  p-value: 0.00008964
```

- The final **fitted model** is given by the equation, $\hat{y} = 212.13 - 0.71 \text{ Age} + 9.5 \text{ MedUse} + 15.12 \text{ Black} + 6.19 \text{ Mexican-American} - 15.13 \text{ Other.}$

Notice that the variable `oralmed_factor` is included in the final model despite not being “statistically significant” at the $\alpha = 0.05$-level. Not all variables must be statistically significant to contribute to the model. In addition, you can refine the model further by again exploring the interaction of age and oral medication use or other plausible interactions in this model. I recommend not relying solely on automated variable selection methods to choose a final model. Rather, use your subject-area knowledge to help you build and refine a final model that makes sense in your application.