Homework 5

Piotr Mankowski

Winter 2018

Question 1

We are interested in examining how mean systolic blood pressure varies by age and sex.

a.

a. Create a scatterplot of systolic blood pressure versus age. Use different symbols and/or colors for each sex group, and include LOWESS (or LOESS) curves for each sex group.

```
## `geom_smooth()` using method = 'loess'
```

b.

Is there evidence from the scatterplot of an association between systolic blood pressure and age after adjusting for sex? Explain your reasoning.

There is evidence for an association between systolic blood pressure and age after adjusting for sex. The lowess lines for both males and females seem to have an a non-0 slope, with the mean blood pressure being different for differnt ages, suggesitng the mean blood pressure is associated with age.

c. *Is there evidence from the scatterplot that sex modifies the association between systolic blood pressure and age? Explain your reasoning.*

In order for sex to modify the association between systolic blood pressure and age, the slope for the blood pressure-age modification should be different in each sex stratum. In the graph above, the slop of the LOESS curves should be different for males and for females. This relationship is a bit difficult to assertain, since the average slopes look close. However, the female slope does seem to be a bit steeper, and the LOESS lines are not parallell across the range of age values, so we do find some evidence that sex modifies the association between blood pressure and age.

d.

Perform a statistical analysis to determine if sex modifies the association between systolic blood pressure and age. Provide full statistical inference.

Methods: For sex to modify the association between systolic blood pressure and age, the slope for this association should differ between age strata. We assessed whether sex modifies the association between systolic blood pressure and age by performing a multivariate linear regression with systolic blood pressure the response and age and sex the predictor variables. We inlcuded the age and sex interaction term to test for effect modification. We tested a null hypothesis that the difference between age slope of different sex groups is equal to 0 against the alternative that the age slope does differ across gender groups. We report the p-value for this test, the point estimate, and a 95% CI of the point estimate.

Results: Based on our linear regression model, we estimate that difference between sex groups in the difference of systolic blood pressure across age is -0.571, with a 95% CI of $(-1.090, -0.05200)$, suggesting the slope for the blood pressure - age association is $0.571 mmHG/year$ lower in females than males. We can reject the null hypothesis that no difference exists across the gender groups, since our p-vale equals 0.0311

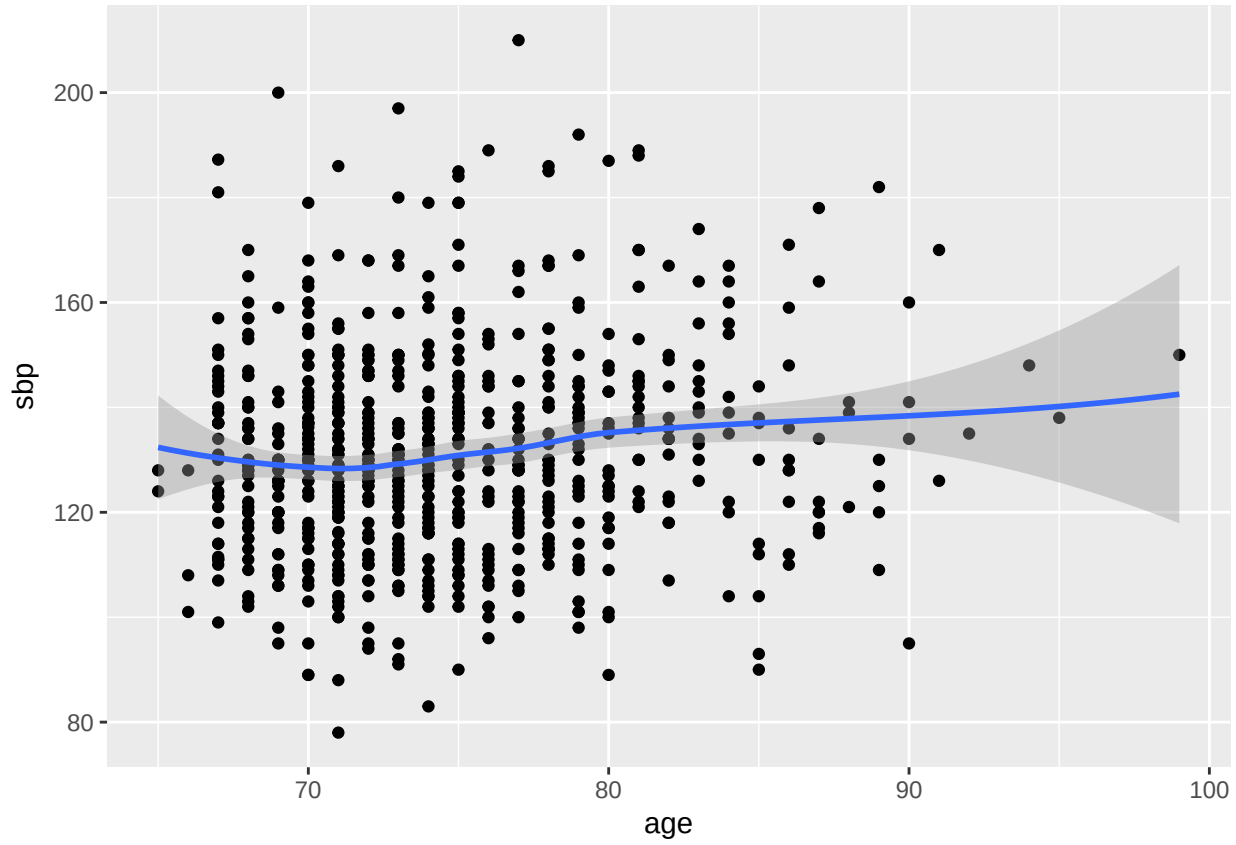
Question 2

2. Now suppose we are interested in examining how mean systolic blood pressure varies by race and age.

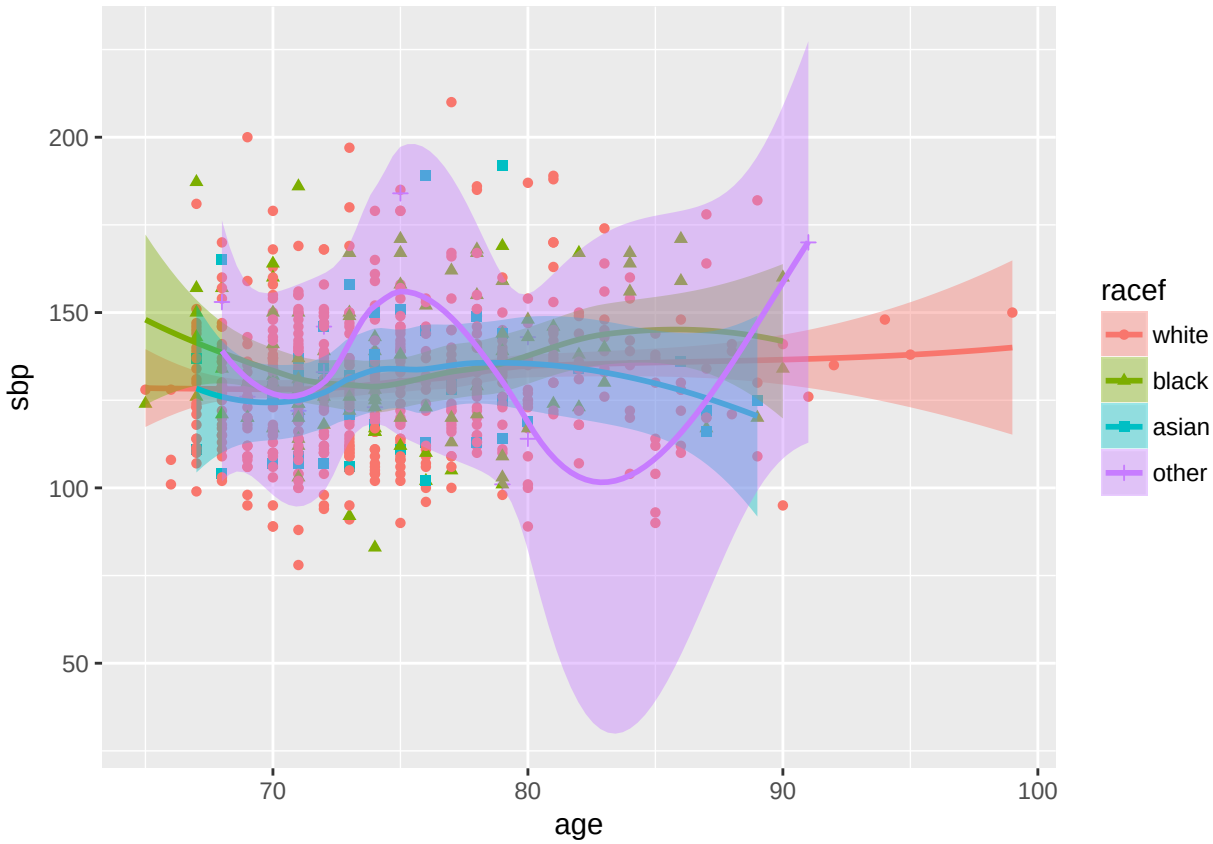
a.

Create a scatterplot of systolic blood pressure versus age. Use different symbols and/or colors for each race group, and include LOWESS (or LOESS) curves for each race group.

```
## `geom_smooth()` using method = 'loess'
```




```
## `geom_smooth()` using method = 'loess'
```



b.

What observations do you make from the scatterplot regarding an association between systolic blood pressure and race.

There's evidence for a positive association between systolic blood pressure and age, as seen in the first scatterplot. When stratified by race, this association does seem to effects of race are This association does seem to be modified by race category; there is a some evidence for the slopes of the LOESS lines for white, black, and asian participants to be different, and the lines are not parallel across the age range. However, the small numbers in each race category group - and especially in the `other` group - do make such observations difficult; the SE intervals for the LOESS lines illustrate this point, with the intervals overlapping significantly across the range of ages.

c.

Perform a multivariate linear regression analysis with systolic blood pressure as the response and with race and age as predictors. What is the baseline group for race in your regression model. Provide an interpretation of the intercept in your regression model and include the numerical value of the intercept in your interpretation. Is the intercept scientifically useful? Briefly explain.

The baseline group for race in our model is the race represented by a value of 0 - which is not interpretable in our original model, since race is actually an categorical variable that is not ordered, and we do not have a race corresponding to the value 0. However, we can rescale this variable to make the baseline race group correspond to one of the actual race groups. In that case, the intercept would represent the mean blood pressure for newborns of whatever race was set to 0. From our model, we estimate the intercept - the mean

systolic blood pressure for newborns of race 0 (which does not exist) - to be $97.51 mmHg$. This intercept is not sceintifically useful.

d.

Provide an interpretation of the age slope in your regression model in part c, and include the numerical value of the age slope in your interpretation. Is the age slope scientifically useful? Briefly explain.

The slope for age in the regression model, estimated to be $0.425 mmHg/year$, represents the difference in mean blood pressure in groups of similar race but age differing by 1 year. Our slope estimate suggests the mean blood pressure is higher in individuals of the same race but 1 year older by $0.425 mmHg$.

e.

Is race a confounder, precision variable, or neither for the association between systolic blood pressure and age? Explain and provide evidence to support your reasoning.

Race does not seem to be a precision variable; adjusting for race does not seem to decrease the variability in each race stratum, which is suggested by looking at the LOESS curves and evident when comparing regression model residual stanard error between adjusted and unadjusted models.

Race does not seem to be a confounder of the blood pressure-age association: race is not significantly associated with blood pressure itself, and this is a requirement for a confounder.

f.

Perform a statistical analysis using the multivariate regression model in part c to determine if race is associated with systolic blood pressure after adjusting for age. Provide full statistical inference.

Methods: We performed a multivariate linear regression with systolic blood pressure as the response variable and age and race as the predictors. We report the slope for race to estimate the difference in systolic blood pressure for groups of the same age but differing by race category. We compute a 95% Wald confidence interval and p-value from the t-test ofthe hypothesis that the slope is equal to zero.

Results: We estimate that mean systolic blood pressure is $1.43 mmHg$ different between populations of different racial categories but similar ages. This estimate has a 95% CI of $(-0.800, 3.67) mmHg$. We cannot reject the null hypothesis at $\alpha = 0.05$ that mean systolic blood pressure does not vary with racial category in populations of the same age, since the p-value for this hypothesis test equals 0.2082 *****

Question 3

Perform a multivariate linear regression analysis with systolic blood pressure as the response and with race, sex, age, and an interaction for sex and age as predictors.

a.

What is the baseline group for race in your regression model. Provide an interpretation of the intercept in your regression model and include the numerical value of the intercept in your interpretation. What, if any, scientific use would you make of the intercept?

The baseline group for race in our model is the race represented by a value of 0 - which is not interpretable in our original model, since race is actually an categorical variable that is not ordered, and we do not have a race corresponding to the value 0. However, we can rescale this variable to make the baseline race group correspond to one of the actual race groups. In that case, the intercept would represent the mean blood pressure for newborns of whatever race was set to 0. From our model, we estimate the intercept - the mean systolic blood pressure for newborn females of race 0 (which does not exist) - to be $75.5 mmHg$. This intercept is not scientifically useful.

b.

Provide an interpretation of the sex slope in your regression model, and include the numerical value of the sex slope in your interpretation. Is the sex slope scientifically useful? Briefly explain.

The sex slope in our model, 41.3, represents the estimated difference in blood pressure between

c.

Provide an interpretation of the age slope in your regression model, and include the numerical value of the age slope in your interpretation. Is the age slope scientifically useful? Briefly explain.

The age slope estimate, 0.726, represents the difference in blood pressure in females of race 0 but differeing in age by 1 year. This slope is scientifcally for insight into the association between age and blood pressure in females of race 0; however, it gets confusing since this race does not exist

d.

Perform a statistical analysis using the multivariate regression model to determine if age is associated with systolic blood pressure. Provide full statistical inference.

We performed a multivariate regression with systolic blood pressure as the response variable, and race, sex, and age as the predictors of interest. we included an sex-age interaction term. We report the p-value of the hypothesis test that the difference in blood pressure across different ages is equal to 0, and report the point estimate of the difference in bp across age, and 95% CI for the point estimate.

The age slope estimate is $.726 mmHg/year$, with a 95% CI of $(0.3653, 1.087)$. We can reject the null hypothesis that age and bp are not associated at the 0.05 level, as the p-value for the test is equal to 0.0001

e.

Perform a statistical analysis using the multivariate regression model to determine if sex is associated with systolic blood pressure. Provide full statistical inference.

We performed a multivariate regression with systolic blood pressure as the response variable, and race, sex, and age as the predictors of interest. we included an sex-age interaction term. We report the p-value of the hypothesis test that the difference in blood pressure across different sexes is equal to 0, and report the point estimate of the difference in bp across sex, and 95% CI for the point estimate.

The sex slope estimate is $41.3 mmHg$, with a 95% CI of $(2.86, 79.7)$. We can reject the null hypothesis that age and bp are not associated at the 0.05 level, as the p-value for the test is equal to 0.0352

f.

Perform a statistical analysis using the multivariate regression model to determine if race is associated with systolic blood pressure. Provide full statistical inference.

We performed a multivariate regression with systolic blood pressure as the response variable, and race, sex, and age as the predictors of interest. we included an sex-age interaction term. We report the p-value of the hypothesis test that the difference in blood pressure across different races is equal to 0, and report the point estimate of the difference in bp across race with 95% CI for the point estimate.

The estimate of difference in blood pressure between races is $1.367 mmHg$, with a 95% CI of $(-0.843, 3.58)$. We cannot reject the null hypothesis that race and bp are not associated at the 0.05 level, as the p-value for the test is equal to 0.2250.

g.

Perform a statistical analysis using the multivariate regression model for testing the null hypothesis that both age and sex are not associated with systolic blood pressure. Provide full statistical inference.

Question 1 code

```
mri[,malef:=as.factor(male)]
ggplot(mri, aes(x=age, y=sbp, color=malef, shape=malef, fill=malef)) + geom_point() + geom_smooth()
```

```
## `geom_smooth()` using method = 'loess'
```


```
mod1d <- regress("mean", sbp~age*male, data=mri)
mod1d
```

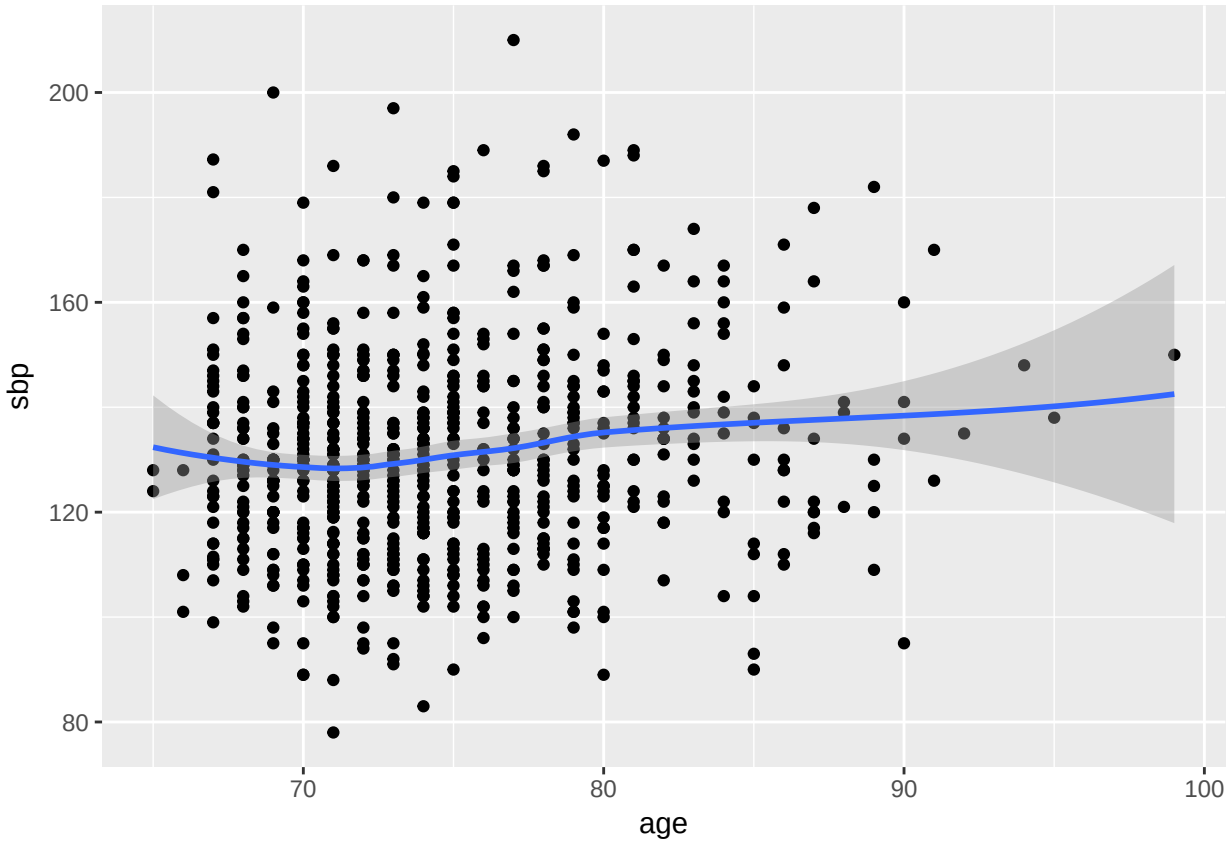
```
##
## Call:
## regress(fnctl = "mean", formula = sbp ~ age * male, data = mri)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -50.80 -13.60  -0.94  10.30  76.77
##
## Coefficients:
##                  Estimate  Naive SE  Robust SE    95%L      95%H
## [1] Intercept      76.47     14.42     13.74        49.49     103.4
## [2] age             0.7372    0.1933    0.1855       0.3730    1.101
## [3] male           42.00     19.79     19.64        3.445     80.55
## [4] age:male      -0.5709    0.2647    0.2643      -1.090    -0.05200
##                     F stat    df Pr(>F)
## [1] Intercept           30.97 1  < 0.00005
## [2] age                 15.79 1    0.0001
```

```
## [3] male                4.57 1    0.0328
## [4] age:male            4.67 1    0.0311
##
## Residual standard error: 19.5 on 731 degrees of freedom
## Multiple R-squared:  0.02073,    Adjusted R-squared:  0.01671
## F-statistic: 5.527 on 3 and 731 DF,  p-value: 0.0009362
```

Question 2 code

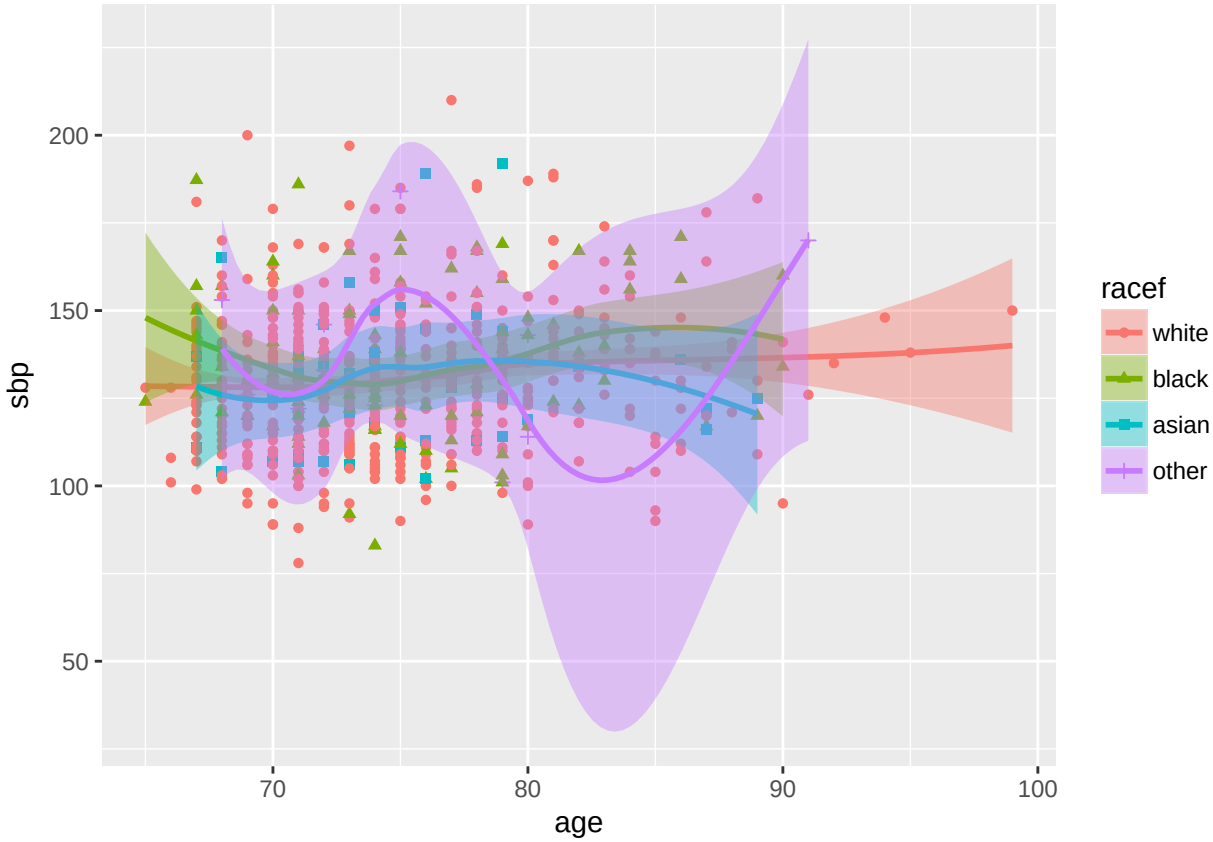
```
mri[,racef:=factor(race, levels=c(1,2,3,4), labels=c('white', 'black', 'asian', 'other'))]
ggplot(mri, aes(x=age, y=sbp)) + geom_point() + geom_smooth()
```

```
## `geom_smooth()` using method = 'loess'
```




```
mri[,racef:=factor(race, levels=c(1,2,3,4), labels=c('white', 'black', 'asian', 'other'))]
ggplot(mri, aes(x=age, y=sbp, color=racef, shape=racef, fill=racef)) + geom_point() + geom_smooth()
```

```
## `geom_smooth()` using method = 'loess'
```




```
mod2c <- regress("mean", sbp~race+age, data=mri)
mod2c
```

```
##
## Call:
## regress(fnctl = "mean", formula = sbp ~ race + age, data = mri)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -51.133 -13.471  -0.857  10.690  78.316
##
## Coefficients:
##                  Estimate  Naive SE  Robust SE    95%L      95%H
## [1] Intercept      97.51     9.944     9.883        78.11     116.9
## [2] race           1.434     1.083     1.138       -0.8008    3.668
## [3] age            0.4252    0.1323    0.1316       0.1668    0.6836
##                     F stat    df Pr(>F)
## [1] Intercept           97.35 1  < 0.00005
## [2] race                 1.59 1    0.2082
## [3] age                 10.44 1    0.0013
##
## Residual standard error: 19.53 on 732 degrees of freedom
## Multiple R-squared:  0.01665,    Adjusted R-squared:  0.01396
## F-statistic: 6.029 on 2 and 732 DF,  p-value: 0.002529
```

```
mod2e1 <- regress("mean", sbp~age, data=mri)
mod2e2 <- regress("mean", sbp~age+race, data=mri)
```

```
mod2e3 <- regress("mean", sbp~age*race, data=mri)
mod2e1
```

```
##
## Call:
## regress(fnctl = "mean", formula = sbp ~ age, data = mri)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -51.568 -13.843  -0.568  10.432  77.845
##
## Coefficients:
##                  Estimate  Naive SE  Robust SE    95%L      95%H
## [1] Intercept      98.95     9.889     9.817        79.68     118.2
## [2] age            0.4312    0.1323    0.1321       0.1718    0.6907
##                     F stat    df Pr(>F)
## [1] Intercept          101.59 1  < 0.00005
## [2] age                 10.65 1    0.0012
##
## Residual standard error: 19.54 on 733 degrees of freedom
## Multiple R-squared:  0.01429,    Adjusted R-squared:  0.01295
## F-statistic: 10.65 on 1 and 733 DF,  p-value: 0.001152
```

```
mod2e2
```

```
##
## Call:
## regress(fnctl = "mean", formula = sbp ~ age + race, data = mri)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -51.133 -13.471  -0.857  10.690  78.316
##
## Coefficients:
##                  Estimate  Naive SE  Robust SE    95%L      95%H
## [1] Intercept      97.51     9.944     9.883        78.11     116.9
## [2] age            0.4252    0.1323    0.1316       0.1668    0.6836
## [3] race           1.434     1.083     1.138       -0.8008    3.668
##                     F stat    df Pr(>F)
## [1] Intercept          97.35 1  < 0.00005
## [2] age                10.44 1    0.0013
## [3] race                1.59 1    0.2082
##
## Residual standard error: 19.53 on 732 degrees of freedom
## Multiple R-squared:  0.01665,    Adjusted R-squared:  0.01396
## F-statistic: 6.029 on 2 and 732 DF,  p-value: 0.002529
```

```
mod2e3
```

```
##
## Call:
## regress(fnctl = "mean", formula = sbp ~ age * race, data = mri)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -51.158 -13.487  -0.829  10.746  78.335
##
## Coefficients:
##                   Estimate  Naive SE  Robust SE    95%L      95%H
## [1] Intercept      99.75     21.58     21.74        57.07     142.4
## [2] age             0.3952    0.2884    0.2939      -0.1818    0.9722
## [3] race           -0.2608    14.51     14.97       -29.64     29.12
## [4] age:race        0.02264   0.1933    0.2023      -0.3746    0.4199
##                     F stat    df Pr(>F)
## [1] Intercept          21.05 1  < 0.00005
## [2] age                 1.81 1    0.1791
## [3] race                0.00 1    0.9861
## [4] age:race            0.01 1    0.9109
##
## Residual standard error: 19.54 on 731 degrees of freedom
## Multiple R-squared:  0.01667,    Adjusted R-squared:  0.01263
## F-statistic: 4.045 on 3 and 731 DF,  p-value: 0.007235
```

Question 3 code

```
#############
# Question 3
#############

mod3 <- regress("mean", sbp ~ race + male * age, data=mri)
mod3
```

```
##
## Call:
## regress(fnctl = "mean", formula = sbp ~ race + male * age, data = mri)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -50.409 -13.540  -1.135  10.582  77.234
##
## Coefficients:
##                   Estimate  Naive SE  Robust SE    95%L      95%H
## [1] Intercept      75.48     14.44     13.69        48.61     102.4
## [2] race            1.367     1.081     1.126       -0.8432    3.577
## [3] male           41.27     19.79     19.56         2.861     79.68
## [4] age             0.7262    0.1934    0.1838       0.3653    1.087
## [5] male:age       -0.5611    0.2647    0.2633      -1.078    -0.04419
##                     F stat    df Pr(>F)
## [1] Intercept          30.41 1  < 0.00005
## [2] race                1.47 1    0.2250
## [3] male                4.45 1    0.0352
## [4] age                15.61 1    0.0001
## [5] male:age            4.54 1    0.0334
##
## Residual standard error: 19.49 on 730 degrees of freedom
## Multiple R-squared:  0.02287,    Adjusted R-squared:  0.01751
## F-statistic: 4.486 on 4 and 730 DF,  p-value: 0.00138
```