Homework 5

Piotr Mankowski

Winter 2018

Question 1

Perform a Poisson regression analysis to evaluate an association between 5 year all-cause mortality and creatinine by comparing the relative risk of death (or risk ratio of death) across groups defined by continuous serum creatinine level. (Only provide a formal report of inference when asked to.)

a.

Provide an interpretation of the slope and the intercept in the Poisson regression model, and include the numerical values of the slope and intercept in your interpretation.
The intercept of the model is -2.860; When exponentiated, this intercept equals 0.0573. This intercept represents the estimated rate of 5-year all-cause mortality for a population where serum creatine levels $= 0$.

The slope of the model is 0.942. The exponentiated slope of 2.56 represents the estimated ratio of rates of 5-year mortality between populations differing by one unit of blood creatinine.

b.

Give full inference for an association between 5 year all-cause mortality and serum creatinine levels from the Poisson regression model.
Methods: We assessed the association between 5-year all-cause mortality and serum creatinine levels using a Poisson regression model. We tested a null hypothesis that the rate of death does not vary across creatine levels against the alternative that an association exists between serum creatinine and 5-year-mortality (meaning the slope != 1) using the defined significance of $p < 0.05$. We report a point estimate and 95% Confidence Interval.

Results: Based on the Poisson regression of 5-year vital status on serum creatinine level, we estimate that the rate for death within five years increases around 156% higher for each 1 mg/dl difference in creatinine serum levels. The 95% confidence interval of the point estimate suggests that our observation would not be unusual if the true percent increase in the rate of 5-year mortality were between 194% and 239%. This increase is highly significant at the 0.05 level, since the test of the null hypothesis that no association exists between 5-year mortality and serum creatine levels (aka the exponentiated slope == 1) against the general alternative returned a p-value of < 0.00005, suggesting we can reject the null hypothesis.

c.

Compare the association results in part b that are based on risk ratios to using a logistic regression model where odds ratios of death within 5 years are used as the summary measure for an association with serum creatinine level (i.e., question 3 in homework 4). Briefly describe any similarities or differences in the association results.
The odds ratio estimate for the logistic regression in HW 4 was found to be 5.986, with a 95% CI of (3.116, 11.50). Using the Poisson regression, we found the risk ratio estimate to be 2.564, with a 95% CI of (1.938, 3.392). Both results suggest a highly significant, positive association between serum creatinine levels and

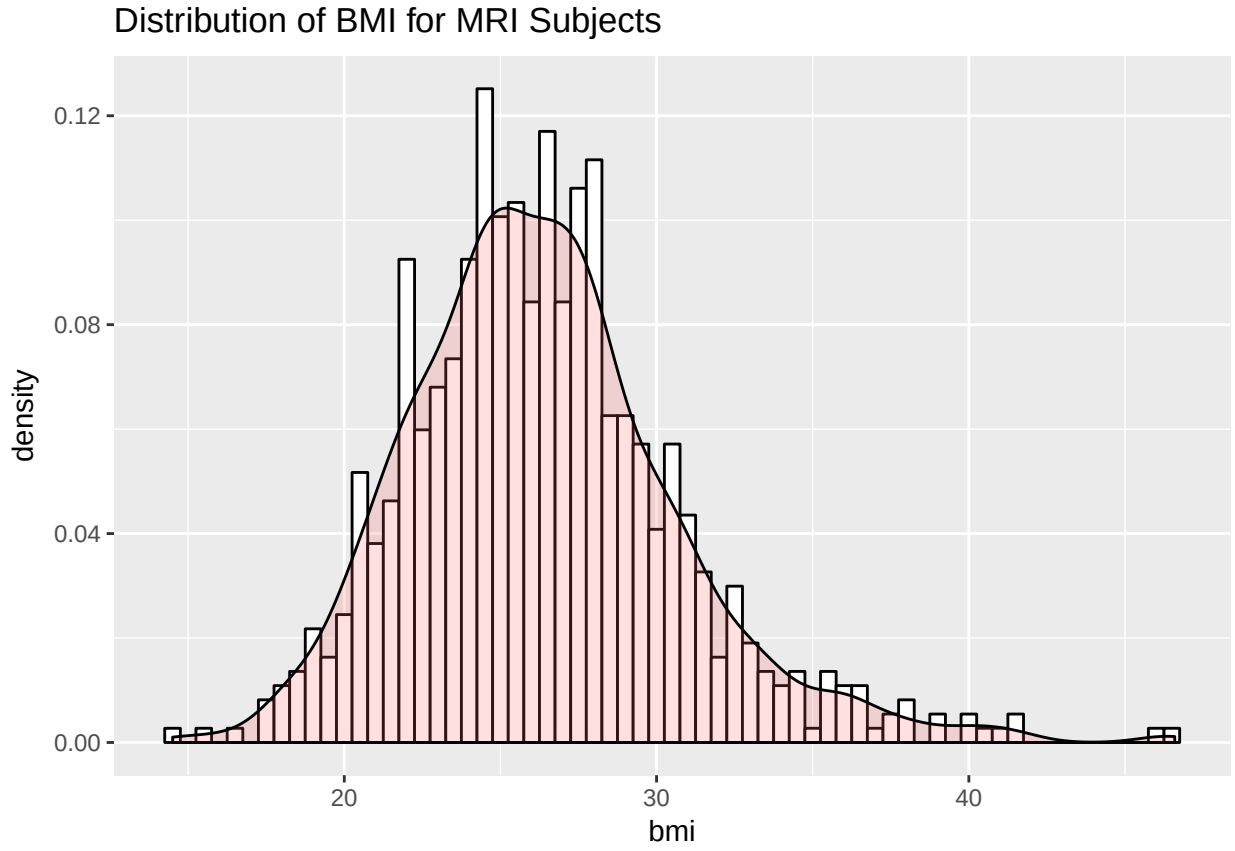
5-year mortality; they only differ in the interpretation of the point estimate. A Poisson regression allows inference on risk ratios, while the logistic regression allows inference on odds ratios.

Question 2

Questions 3 and 4 below investigate associations between serum cholesterol level, age, sex, and body mass index (BMI). In this question we will obtain some summary statistics for these variables.

a.

Create a variable for BMI using the height and weight measurements on the subjects. [Hint: Make sure that appropriate conversions of the weight and height measures are used in the calculation of BMI]. Provide a figure illustrating the distribution of BMI in the sample.



b.

Provide suitable descriptive statistics for serum creatinine levels, age, sex, and BMI.

Variable	Females	Males	All subjects
Sample size	n = 369	n = 366	n = 735
Age (years)[1]	74.4 (5.3); 65-91	74.7 (5.6); 66-99	74.6 (5.5); 65-99
Serum Creatinine Levels (mg/dl)[1]	0.9 (0.3); 0.5-3.2	1.2 (0.3); 0.7-4	1.1 (0.3); 0.5-4
BMI (kg/m^2)[1]	26.4 (4.8); 14.5-46.6	26.3 (3.7); 16.3-41.5	26.3 (4.3); 14.5-46.6
[1] mean (sd); min-max are reported	**		

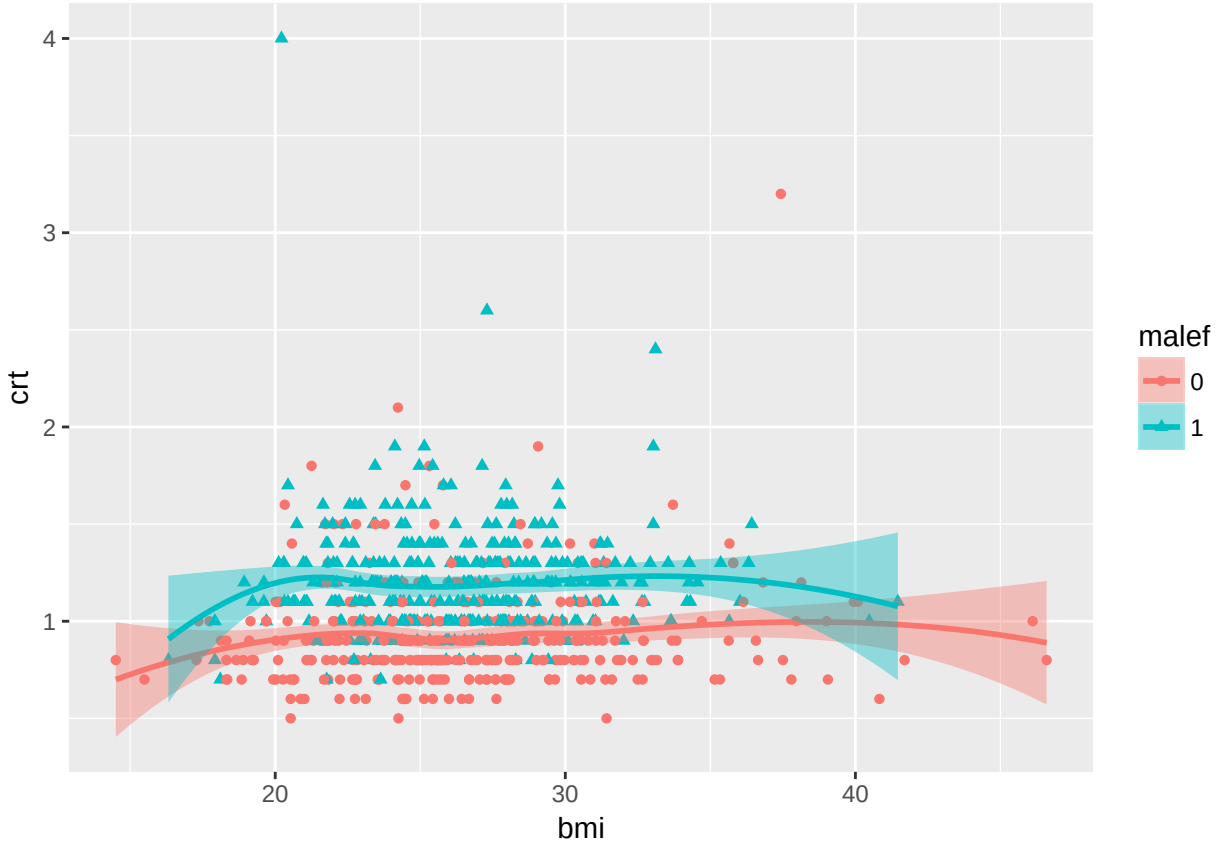
Question 3

We are interested in examining how mean serum creatinine levels vary by BMI and sex. In the questions below, you do not need to provide full statistical inference. Instead, just answer the following questions.

a.

Create a scatterplot of serum creatinine levels versus BMI. 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.

What observations do you make from the scatterplot in part a regarding the association between serum creatinine levels and BMI?

Disregarding the values for BMI < 17 and BMI > 40 due to the sparseness of the data, the association between serum creatinine levels and BMI seems to be almost non-existant. For males, the Lowess line seems to have a slope of around 0 across the whole applicable range. For females, there does seem to be a very slight positive association between BMI and creatinine levels, since the slope of the Lowess line looks slightly positive.

c.

Is there evidence from descriptive statistics (question 2) and the scatterplot in part a that sex modifies the association between serum creatinine level and BMI? Explain your reasoning. There does not seem to be evidence for sex modifying the association between serum creatinine and BMI. The Lowess lines for both males and females both have a slope of around 0; Sex does not seem to affect this slope. Although serum creatinine levels seem to be higher in males than females (as evidenced by the summary statistics and the fact that the Lowess line for males is higher than for females), the association between BMI and crt in each group seems to be simmilar.

d.

Is there evidence from descriptive statistics (question 1) and the scatterplot in part a that sex confounds the association between serum creatinine level and BMI? Explain your reasoning.

There does not seem to be evidence for sex confounding the association between serum creatinine and BMI. For sex to confound this association, sex must be associated with both BMI and crt. The BMI values for males and females seem to be simmilar (based on the descriptive statistics). Although sex does seem to affect crt levels, with males having higher serum creatinine, this is not enough to suggest that sex confounds the association between BMI and creatinine.

e.

Perform an analysis to determine whether mean serum creatinine levels differ across sex groups. Briefly describe the analysis that you performed and clearly state the basis of your conclusion regarding an association.

After performing a linear regression of creatinine on sex using robust SE, we do find evidence for mean creatinine levels differing across sex groups. The slope is estimated to be 0.269, suggesting creatinine levels in males are on average $0.269 mg/dl$ higher than in females; with a p-value of <0.00005, this result is highly significant at $\alpha = 0.05$.

f.

Perform an analysis to determine whether there is a linear trend in mean serum creatinine levels by BMI. Briefly describe the analysis that you performed and clearly state the basis of your conclusion regarding an association.

There does not seem to be a significant linear trend in mean serum creatinine levels by BMI. A linear regression of creatinine levels on BMI gives a slope estimate of 0.00319, which is close to 0. The 95% CI of the slope includes 0, and the p-value for the hypothesis test of whether this slope differs from 0 has a p-value of 0.311, suggesting a lack of evidence for this linear trend at $\alpha = 0.05$.

g.

Perform an analysis to determine whether mean serum creatinine levels differ across sex groups after adjustment for BMI. Briefly describe the analysis that you performed and clearly state the basis of your conclusion regarding an association.

After performing a multivariate regression of creatinine levels on sex adjusted for bmi, we find that the estimated difference in creeatinine between males and females with similar bmi is $0.270 mg/dl$; this result is highly significant at $\alpha = 0.05$, with the p-value $= < 0.0005$, suggesting mean creatinine levels are higher in males than in females after adjusting for bmi.

h.

Perform an analysis to determine whether there is a linear trend in mean serum creatinine levels by BMI after adjustment for sex. Briefly describe the analysis that you performed and clearly state the basis of your conclusion regarding an association.

Using the same mulitvariate linear regression as in part g, we do not find evidence for a linear trend in mean creatinine levels by BMI after adjusting for sex. The estimated difference in mean creatinine levels for two populations with the same sex but with a 1 kg/m^2 difference in BMI is 0.00381, which is close to 0. The hypothesis test of whether this difference is not 0 has a p-value of 0.193, which means we cannot reject the null hypothesis at $\alpha = 0.05$.

i.

Perform an analysis to determine if sex modifies the association between mean serum creatinine levels and BMI. Briefly describe the analysis that you performed and clearly state the basis of your conclusion regarding an association.

Since the adjusted for sex(part h) and unajusted for sex(part f) analyses of the association between mean serum creatinine levels and BMI have similar results, there does not seem to be evidence for sex modifying this association.

j.

How would you summarize the association between serum creatinine levels and BMI and sex? Provide a summary of your findings that is suitable for inclusion in a manuscript.

We examined the association between serum creatinine levels and bmi and sex by performing a multivariate linear regression using Huber-White robust standard errors. We tested the null hypotheses that creatinine levels do not vary with BMI and sex at the significance level of 0.05.
We found that mean creatinine levels for popluations of males and females with teh same bmi differ by an estimated value of $0.270 mg/dl$, with males having the higher level; this estimate has a confidence interval of $(0.2305, 0.3087)$. We reject the null hypothesis that no difference exists between males and females of the same bmi, since our p-value is < 0.0005.

We estimate the difference in creatinine levels for populations of the same sex but varying in BMI by 1 kg/m^2 to be $0.00382 mg/dl$, where higher bmi populations have higher creatinine levels. This estimate has a 95% confidence interval of (-0.00193, 0.00958). We cannot reject the null hypothesis of no association between creatinine levels and bmi when adjusted for sex, since the p-value equals 0.1929.

Question 4

4. Now consider a multivariate linear regression analysis with serum creatinine level as the response and the variables age, sex, and BMI as predictors.

a.

Provide an interpretation of the intercept in the regression model. Is the intercept estimate scientifically useful?

The intercept for this multivariate linear regression is 0.355. This intercept represents the estimated mean creatinine level for a population where the predictors are all 0: newborn females who have a bmi of 0. This theoretical population does not exist, and this estimate has no real scientific value.

b.

Give full inference for the age slope in the regression model.

We assessed the association between serum creatinine levels and age, adjusted for bmi and sex, by perfoming a a multivariate linear regression with Huber-White robust standard errors. We tested the null hypothesis that no difference in mean creatinine levels exists for different ages using a defined significance as $p < 0.05$. We report a point estimate and 95% confidence interval calculated with robust standard errors. Based on this analysis, we estimate the difference in mean creatinine levels between populations one year apart and with the same sex and bmi to be $0.00582 md/dl$, with the older population having higher creatinine levels, and a 95% confidence interval of (0.00140, 0.01024). We reject the null hypothesis at the 0.05 significance level, since the p-value for the hypothesis test is 0.0099. Age controlled for bmi and sex has a slight positive association with creatinine levels.

c.

Give full inference for the sex slope in the regression model.

We assessed the association between serum creatinine levels and sex, adjusted for bmi and age, by perfoming a a multivariate linear regression with Huber-White robust standard errors. We tested the null hypothesis that no difference in mean creatinine levels exists for males and females of the same bmi and age using a defined significance as $p < 0.05$. We report a point estimate and 95% confidence interval calculated with robust standard errors. Based on this analysis, we estimate the difference in mean creatinine levels between populations of males and females of the same age and bmi to be $0.268 mg/dl$, with males having the higher creatinine. The 95% confidence interval for this estimate is of $(0.00140 mg/dl, 0.01024 mg/dl)$. We reject the null hypothesis at the 0.05 significance level, since the p-value for the hypothesis test is < 0.0005. Creatinine levels and sex have a postivive association when controlled for bmi and age.

d.

Give full inference for the BMI slope in the regression model.

We assessed the association between serum creatinine levels and bmi, adjusted for sex and age, by perfoming a a multivariate linear regression with Huber-White robust standard errors. We tested the null hypothesis that no difference in mean creatinine levels exists populations of the same bmi and age using a defined significance as $p < 0.05$. We report a point estimate and 95% confidence interval calculated with robust standard errors. Based on this analysis, we estimate the difference in mean creatinine levels between populations 1 kg/m^2 apart with the same age and sex to be $0.00536 mg/dl$, with the higher bmi group having the higher creatinine.

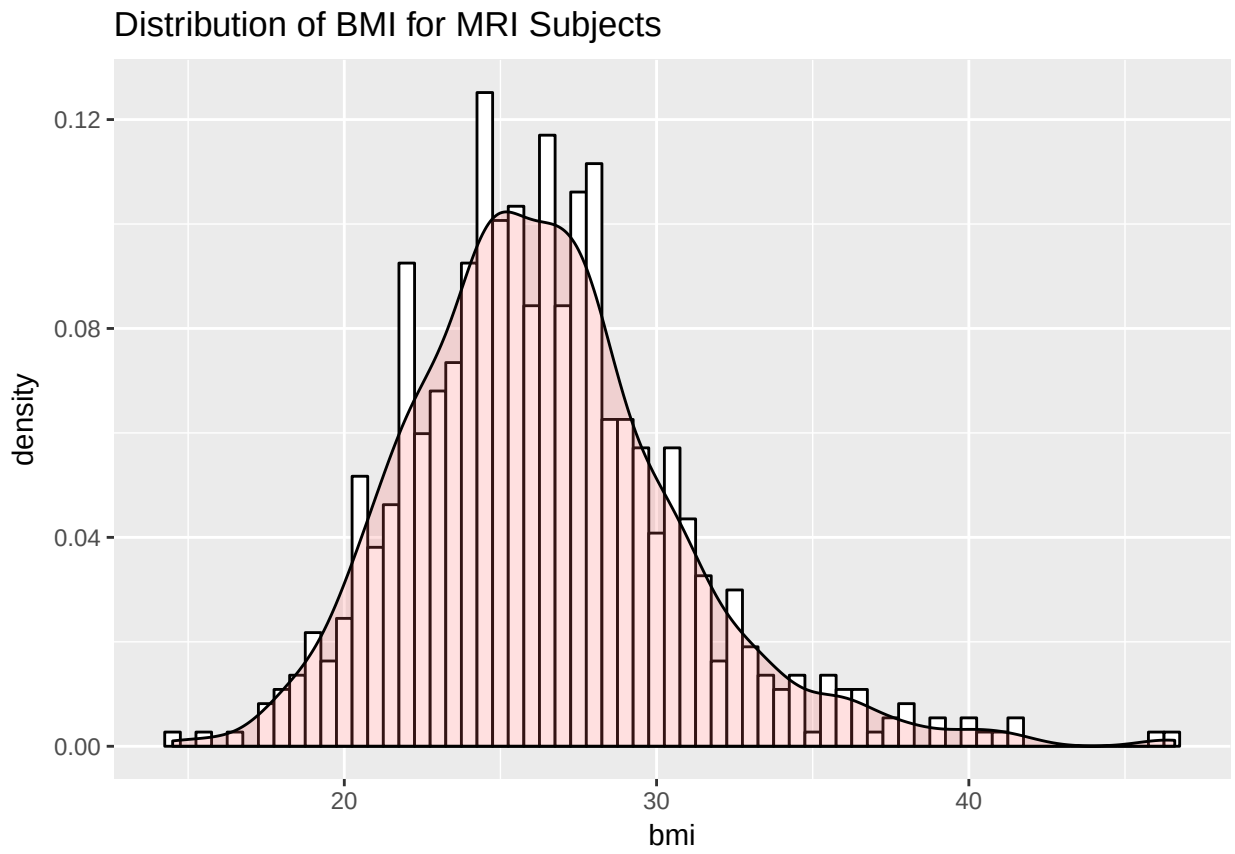
The 95% confidence interval for this estimate is of $(-6.44e-05, 0.0108)$. We cannot reject the null hypothesis at the 0.05 significance level, since the p-value for the hypothesis test is 0.0528. Creatinine levels and bmi do not have a significant association.

Question 1 code

```
##############
# Question 1
##############
mod1 <- regress("rate", dead_at_5yr ~ crt, data=mri)
```

Question 2 code

```
ggplot(mri, aes(x=bmi)) + geom_histogram(aes(y=..density..), binwidth=.5, colour="black", fill="white")
geom_density(alpha=.2, fill="#FF6666") + ggtitle("Distribution of BMI for MRI Subjects")
```



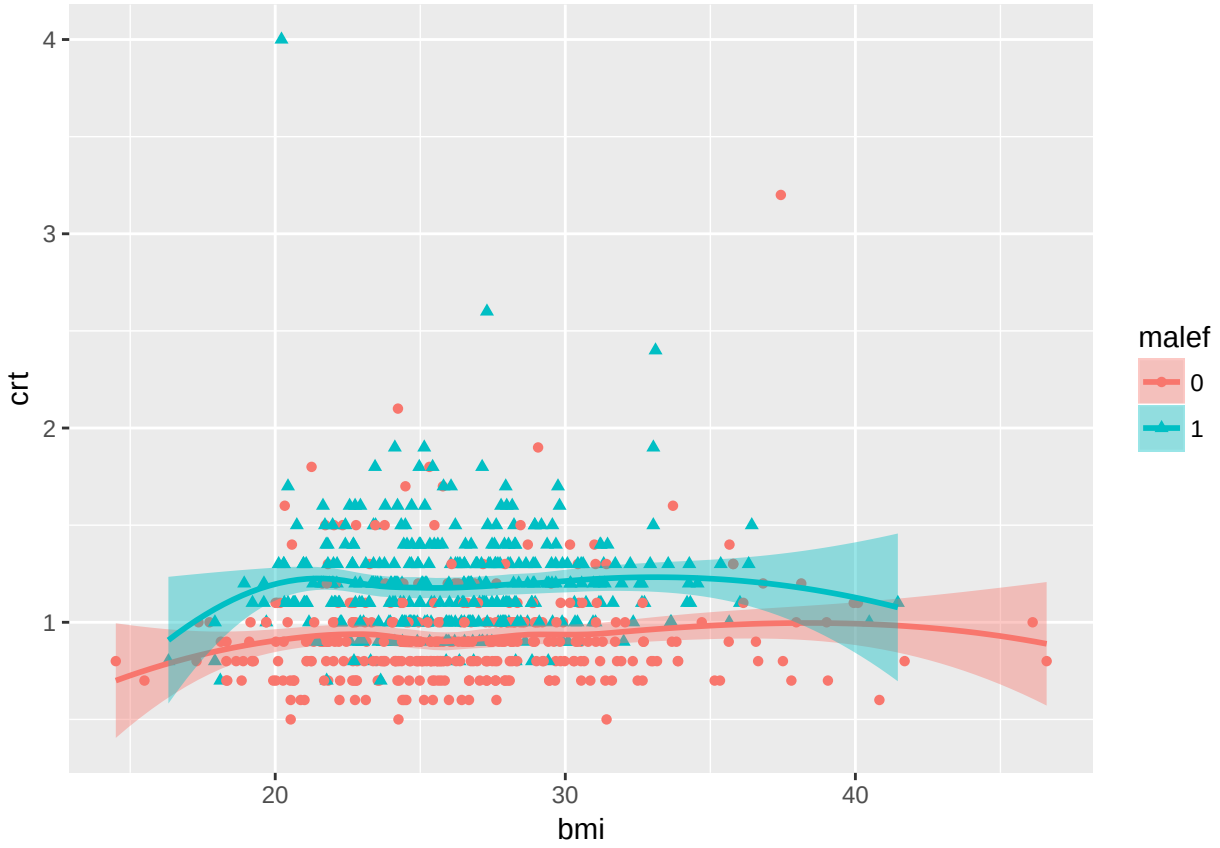

```
# function to summarize continuous variables
cont.summary <- function(x, digits=1){
  fmt <- paste0("%0.", digits, "f")
  m <- mean(x, na.rm = TRUE)
  m <- sprintf(fmt = fmt, m)
  s <- sd(x, na.rm = TRUE)
  s <- sprintf(fmt = fmt, s)
  mini <- round(min(x, na.rm = TRUE), digits=digits)
  maxi <- round(max(x, na.rm = TRUE), digits=digits)
  paste0(m, " (", s, "); ", mini, "-", maxi)
}
```

```r
# function to summarize discrete variables
disc.summary <- function(x, digits=1){
  fmt <- paste0("%0.", digits, "f")
  pct <- mean(x, na.rm = TRUE) * 100
  pct <- sprintf(fmt = fmt, pct)
  paste0(pct, "%")
}

# names of continuous variables
cont.covars <- c("age", "crt", "bmi")
disc.covars <- c()

tab1col1 <- c(
  paste0("n = ", length(mri[mri$male == 0,]$ptid)),
  apply(mri[male == 0, cont.covars, with=FALSE], 2, cont.summary)
  # apply(mri[male == 0, disc.covars, with=FALSE], 2, disc.summary)
)

tab1col2 <- c(
  paste0("n = ", length(mri[mri$male == 1,]$ptid)),
  apply(mri[male == 1, cont.covars, with=FALSE], 2, cont.summary)
  #  apply(mri[male == 1, disc.covars, with=FALSE], 2, disc.summary)
)

tab1col3 <- c(
  paste0("n = ", length(mri$ptid)),
  apply(mri[, cont.covars, with=FALSE], 2, cont.summary)
  # apply(mri[, disc.covars, with=FALSE], 2, disc.summary)
)

tab1names <- c(
  "Sample size",
  "Age (years)^1^",
  "Serum Creatinine Levels (mg/dl)^1^",
  "BMI (kg/m^2^)^1^"
  )

tab1 <- cbind(tab1names, tab1col1, tab1col2, tab1col3)
tab1col.names <- c("Variable", "Females", "Males", "All subjects")
```

Question 3 code

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

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




```
mod3e <- regress('mean', crt~male, data=mri)

mod3f <- regress('mean', crt~bmi, data=mri)
mod3g <- regress('mean', crt~male+bmi, data=mri)
```

Question 4 code

```
##############
# Question 4
##############

mod4 <- regress("mean", crt~age+male+bmi, data=mri)
```