

Assignment 2

BIOSTAT 213: Introduction to Computing in the R Software Environment

Due: Wednesday, October 23, 2019

Collaboration with other students in this course is permitted on this assignment. However, you must write all work—including R code—in your own voice, as much as possible, and acknowledge your collaborators as thoroughly as possible.

The assignment is due online, in the [Collaborative Learning Environment](#), by 11:59 pm, Wednesday, October 23, 2019. You must submit a Word-compatible document or a PDF file.

For each of the below, share your code as well as the output (if there is any output). If your answer to a question is unclear from your R output, or difficult to identify, include an answer in text form (in addition to the code and output).

You will be using the `nhanes.csv` data file for the following problems.

1. For this problem, we will be examining the association between blood pressure (`bpcat`) and history of smoking (`hs`).

- (a) Re-order the levels of `bpcat` so that they are in scientific order: `Normal`, `Elevated`, and `Hypertension`. Cross-tabulate the 2 variables, with `bpcat` along the rows and `hs` along the columns.

```
> ii$bpcat<-relevel(ii$bpcat,'Normal')
> table(ii$bpcat,ii$hs)
```

	History of smoking	No
Normal	612	1148
Elevated	305	438
Hypertension	704	813

- (b) Compute the row-wise percents (your numbers should be on a percent scale, not a 0–1 scale).

```
> prop.table(table(ii$bpcat,ii$hs),1)*100
```

	History of smoking	No
Normal	34.77273	65.22727
Elevated	41.04980	58.95020
Hypertension	46.40738	53.59262

- (c) Use a χ^2 -test to test the association between blood pressure (`bpcat`) and history of smoking (`hs`).

```
> chisq.test(table(ii$bpcat,ii$hs))
```

Pearson's Chi-squared test

data: table(ii\$bpcat, ii\$hs)

X-squared = 46.032, df = 2, p-value = 1.01e-10

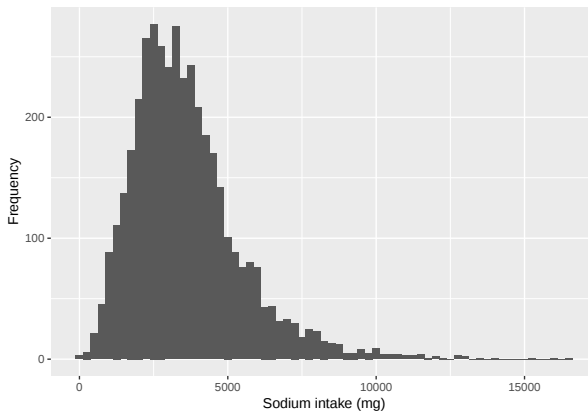
2. Fit a simple linear regression for diastolic blood pressure (`bpxdi1`), with age (`ridageyr`) as the x variable. Output only the coefficient estimates, the standard errors, the test statistics, and the p -values for this model.

```
> mm<-lm(ii$bpxdi1~ii$ridageyr)
> summary(mm)$coefficients
```

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	64.3266208	0.57066283	112.72264	0.000000e+00
ii\$ridageyr	0.1436185	0.01256436	11.43063	8.441729e-30

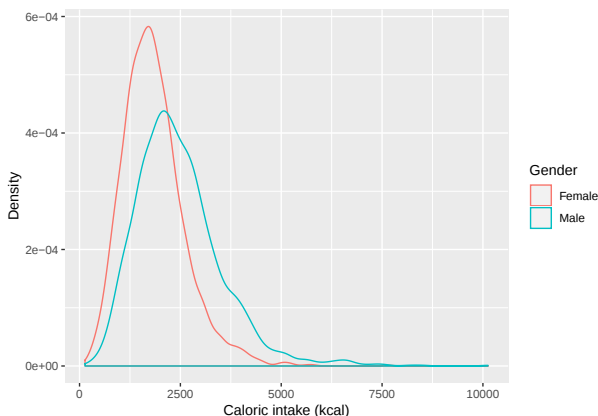
3. Use ggplot2 to create a histogram of daily sodium intake (`dr1tsodi`), using a bin width of 250.

```
> gg<-ggplot(data=ii,mapping=aes(x=dr1tsodi))
> gg+
+   geom_histogram(binwidth=250)+
+   xlab('Sodium intake (mg)')+
+   ylab('Frequency')
Warning: Removed 137 rows containing non-finite values (stat_bin).
```



4. Use ggplot2 to create a single plot showing the density curve for caloric intake (`dr1tkcal`) among females and the density curve for caloric intake among males. The gender variable in `nhanes.csv` is named `gender`.

```
> ggplot(data=ii,mapping=aes(x=dr1tkcal,color=gender))+
+   geom_density()+
+   xlab('Caloric intake (kcal)')+
+   ylab('Density')+
+   labs(color='Gender')
Warning: Removed 137 rows containing non-finite values (stat_density).
```



5. Write code that will re-order the levels of high-density lipoprotein (HDL) cholesterol groups (`hdlcat`). The desired order is: Low, Medium, and High. Use `table()` to verify your work.

```
> ii$hdlcat<-factor(ii$hdlcat,c('Low','Medium','High'))
> table(ii$hdlcat)
```

Low	Medium	High
793	1976	1245

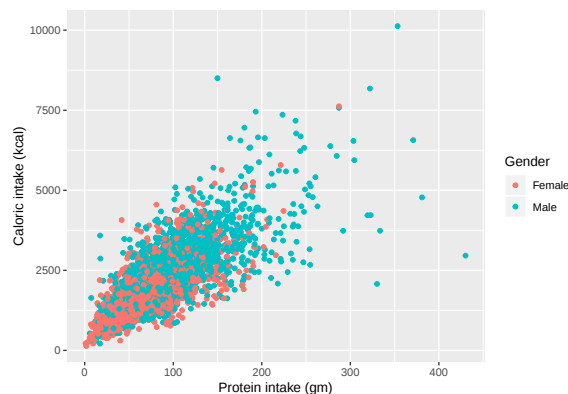
6. Fit a linear regression model for caloric intake (`dr1tkcal`), with gender (`gender`) protein intake (`dr1tprot`) as your x variables, with quadratic adjustment for protein intake. Use this model to predict caloric intake for individuals who are female and have protein intake of 25, 50, 75, 100, 125, and 150 mg.

```
> mm<-lm(dr1tkcal~gender+dr1tprot+I(dr1tprot^2),data=ii)
> nn<-data.frame(gender='Female',dr1tprot=seq(25,150,25))
> nn$predictions<-predict(mm,nn)
> nn
  gender dr1tprot predictions
1 Female      25    988.6088
2 Female      50   1484.4682
3 Female      75   1953.6557
4 Female     100   2396.1713
5 Female     125   2812.0150
6 Female     150   3201.1868
```

7. For this problem, we will consider the association between caloric intake (`dr1tkcal`) and protein intake (`dr1tprot`), by gender.

- (a) Produce a scatter plot of caloric intake (`dr1tkcal`) and protein intake (`dr1tprot`), with caloric intake (`dr1tkcal`) on the y axis and color coding for gender.

```
> gg<-ggplot(data=ii,mapping=aes(x=dr1tprot,y=dr1tkcal,color=gender))
> gg+
+   geom_point()+
+   xlab('Protein intake (gm)')+
+   ylab('Caloric intake (kcal)')+
+   labs(color='Gender')
Warning: Removed 137 rows containing missing values (geom_point).
```



- (b) An alternate way of presenting the data is to take advantage of ggplot2's *faceting* feature. Run the following code and consider the output.

```
> gg<-ggplot(data=ii,mapping=aes(x=dr1tprot,y=dr1tkcal))
> gg+
+   facet_grid(.~gender)+
+   geom_point()+
+   xlab('Protein intake (gm)')+
+   ylab('Caloric intake (kcal)')+
+   labs(color='Gender')
```

Warning: Removed 137 rows containing missing values (geom_point).

