BST210 Lab 01

January 29-30, 2026

In this lab, we will examine the relationship of systolic blood pressure (SBP) with demographic variables and other measures of health and function among a sample of older people in East Boston. Data are in the file `ebchf3.csv`. The codebook is also available on the website (`ebchf3_codebook.csv`).

Question 1

Clear your workspace, read in the `tidyverse` and `ggplot2` packages, and read the `ebchf3.csv` dataset into R.

Install the packages first, if necessary. Go to Canvas and download the dataset `ebchf3.csv` from Week 1. The codebook is also available on Canvas. Do not use Safari to download data from Canvas; use another browser or else it will not work. Save the dataset in a location that makes sense for you. Then, read in the data using the `read.csv` function and save it as a data frame called `d`. On a Mac, you can right click the file > Get Info > Select the text after "Where:" and then copy that to the clipboard and paste into your code. On a PC, you can right click the file > Copy as Path to copy it, and then paste it yourself.

```
rm(list=ls())
library(tidyverse)
library(ggplot2)
d <- read.csv("/Users/tt469/Library/CloudStorage/GoogleDrive-tthaweethai@mgh.harvard.edu/My Drive/BST210 Spring 2026/Data/ebchf3/ebchf3.csv") # specify local filepath, in quotations
```

Question 2

View the first few rows of the data.

You can use `head()` to show the first 5 rows of the data. Alternatively, click on the `d` variable name in the top right "environment" panel to bring up a view of the data in the top left panel. Scroll around a bit and see what you see! In this example, what does each row represent? Each column? Refer to the codebook for the meaning of the different variables.

```
head(d)
```

```
##   id age male sbp dbp phys cogerr  pefr mile stair prioraf priorchd valve antih
## 1  1  87    1  96  59    0      7 173.3    0     1       0        0     0     0
## 2  2  72    0 160  75    5      0 390.0    0     1       0        1     1     1
## 3  3  72    0 125  67    9      1 250.0    0     1       0        1     0     1
## 4  4  84    0 125  71    1      3 100.0    1     1       0        0     0     1
## 5  5  73    1 122  68   10      3 320.0    1     1       0        0     0     0
## 6  6  74    1 120  78    2      4 240.0    0     0       0        0     0     1
##   psmok csmok dm priorchf dig loop
## 1     1     0  0        0   0    0
## 2     0     0  1        0   0    0
## 3     1     0  1        1   0    1
## 4     0     0  0        0   0    0
## 5     1     0  0        0   0    0
## 6     1     0  0        0   0    0
```

Each row represents a different participant and each column represents a different variable.

Question 3

How many current smokers and current non-smokers are there in the data? What % of the overall population is a current smoker?

```
table(d$csmok)
```

```
##
##    0    1
## 1472  191
```

```
d %>%
  group_by(csmok) %>%
  summarise(n=n()) %>%
  ungroup()
```

```
## # A tibble: 2 × 2
##   csmok     n
##   <int> <int>
## 1     0  1472
## 2     1   191
```

Referring to the codebook, the variable `csmok` equals 1 for current smokers and equals 0 for current nonsmokers. Above are two ways of counting the number of smokers and nonsmokers in the population. Note that the `table()` function will not show if there are individuals with missing smoking data. Instead, you would use `table(d$csmok, useNA="always")` to show the count of those with missing data. In this case, there are no missing values. The second approach, using tidyverse, would automatically count those with missing data.

There are 191 current smokers and 1472 current nonsmokers. You can calculate the % by hand, or you can use R:

```
d %>%
  group_by(csmok) %>%
  summarise(n=n(),
            pct = n()/nrow(d)) %>%
  ungroup()
```

```
## # A tibble: 2 × 3
##   csmok     n   pct
##   <int> <int> <dbl>
## 1     0  1472 0.885
## 2     1   191 0.115
```

```
table(d$csmok)/nrow(d)
```

```
##
##         0         1
## 0.8851473 0.1148527
```

We divided by `nrow(d)`, the total number of participants in the dataset, to get these percentages. 88.5% of participants are current nonsmokers and 11.5% are current smokers.

It is perfectly fine to look at your R output and copy down the numbers when working on problem sets. This is not necessary, but if you want to be super advanced about your R Markdown output, you can have R automatically output results in your text. The following sentence looks normal, but take a look at the code itself used to generate it:

There are 191 current smokers and 1472 nonsmokers.

Question 4

Create a subset of the data with only smokers using the filter function in the tidyverse package.

```
d.smok <- d %>% filter(csmok==1)
```

Question 5

Perform a two-sample t-test (unequal variances) evaluating whether the mean SBP differs by current smoking status.

Use the `t.test` function in R and refer to the help documentation, referring to `?t.test` if necessary. You can use `$` to extract a given variable from a data frame to obtain a vector; for example: `d$sbp` will give you a vector that has the SBP for all participants in the dataset. Interpret the results.

```
d.nonsmok <- d %>% filter(csmok==0)
t.test(d.smok$sbp, d.nonsmok$sbp)
```

```
##
##  Welch Two Sample t-test
##
## data:  d.smok$sbp and d.nonsmok$sbp
## t = 0.27485, df = 241.11, p-value = 0.7837
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -2.453230  3.248813
## sample estimates:
## mean of x mean of y
##  137.1152  136.7174
```

The average SBP among current smokers is 137.1 mmHg. The average SBP among current nonsmokers is 136.7 mmHg. Using a significance level of 0.05, the p-value of 0.784 indicates that we do not have evidence to reject the null that the average SBP between current smokers and nonsmokers is the same.

Question 6

Fit a linear regression evaluating whether the mean SBP differs by smoking status. Write down the model that you are fitting. Interpret the estimated regression coefficients. According to the model, what is the average SBP among current smokers?

```
fit.csmok <- lm(sbp ~ csmok, data = d)
summary(fit.csmok)
```

```
##
## Call:
## lm(formula = sbp ~ csmok, data = d)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -44.717 -12.717  -2.717  11.283  80.283
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 136.7174     0.4875 280.466   <2e-16 ***
## csmok         0.3978     1.4384   0.277    0.782
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.7 on 1661 degrees of freedom
## Multiple R-squared:  4.604e-05,  Adjusted R-squared:  -0.000556
## F-statistic: 0.07648 on 1 and 1661 DF,  p-value: 0.7822
```

Model: $SBP_i = \beta_0 + \beta_1 Currentsmoker_i + \epsilon_i$

$\widehat{\beta}_0 = 136.7$: The average SBP among current non-smokers is 136.7 mmHg.

$\widehat{\beta}_1 = 0.40$: On average, current smokers have SBP 0.40 mmHg higher than current non-smokers.

According to the model, the average SBP among current smokers is given by someone who has $Currentsmoker_i = 1$:

$$E\left[SBP_i \middle| Currentsmoker_i = 1\right] = \widehat{\beta}_0 + \widehat{\beta}_1 = 136.7 + 0.40 = 137.1$$

The average SBP among current smokers is 137.1. Note that the two estimated means are the same as obtained via the t-test.

Question 7

Perform a hypothesis test evaluating whether the mean = SBP differs by smoking status.

The null hypothesis is that the mean SBP is the same between current nonsmokers and current smokers: $H_0 : \beta_1 = 0$.

The alternative hypothesis is that the mean SBP is different between current nonsmokers and current smokers: $H_1 : \beta_1 \neq 0$

Using a significance level of 0.05, the p-value of 0.782 indicates that we do not have evidence to reject the null that the mean SBP between current smokers and current nonsmokers is the same.

This was almost exactly the same conclusion as the t-test, with a p-value that only differs by 0.002.

Question 8

We will now look at age as the exposure. Using the `cor()` function in R, what is the correlation between age and SBP? Is this Pearson, Kendall, or Spearman correlation?

```
cor(d$age, d$sbp)
```

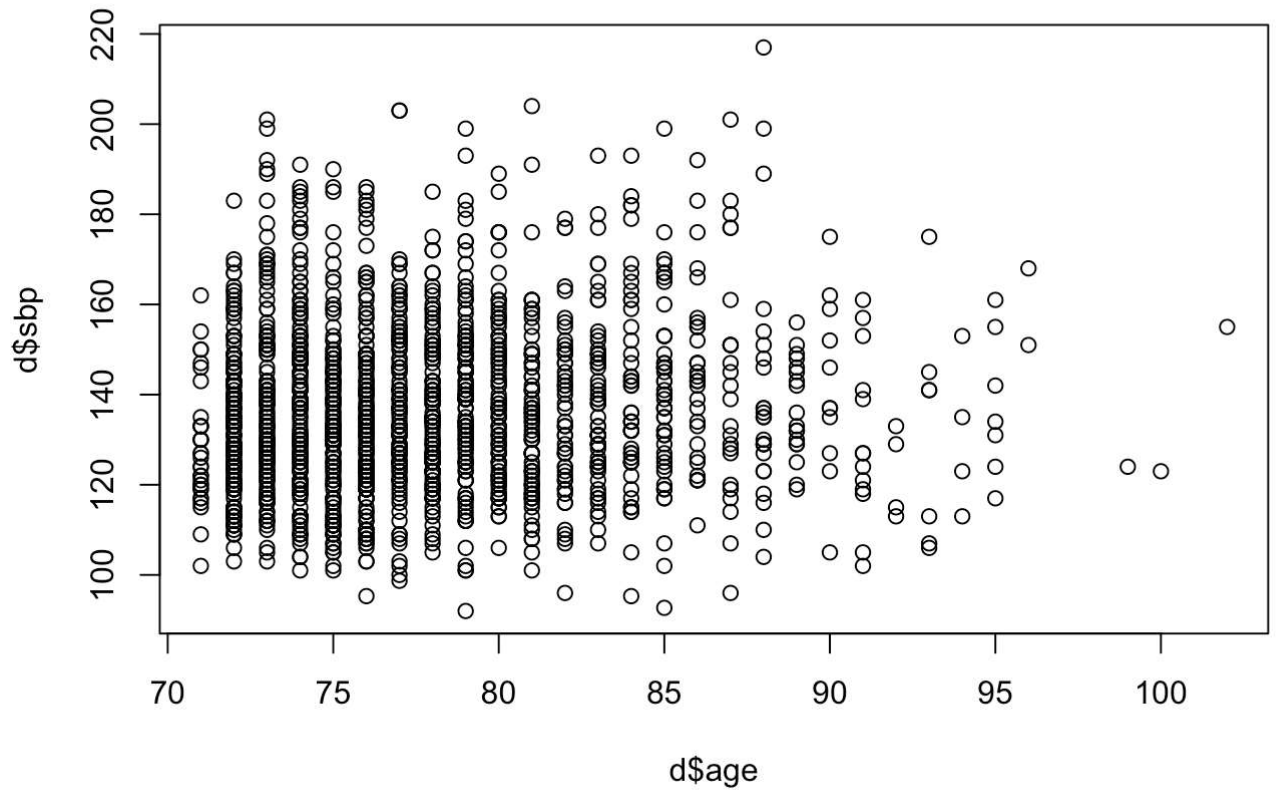
```
## [1] 0.06938894
```

The correlation between age and SBP is 0.0694. If you look at the help documentation for the function, and scroll down to “method” under the Arguments section, you can see that the Pearson correlation is default. So, this is the Pearson correlation.

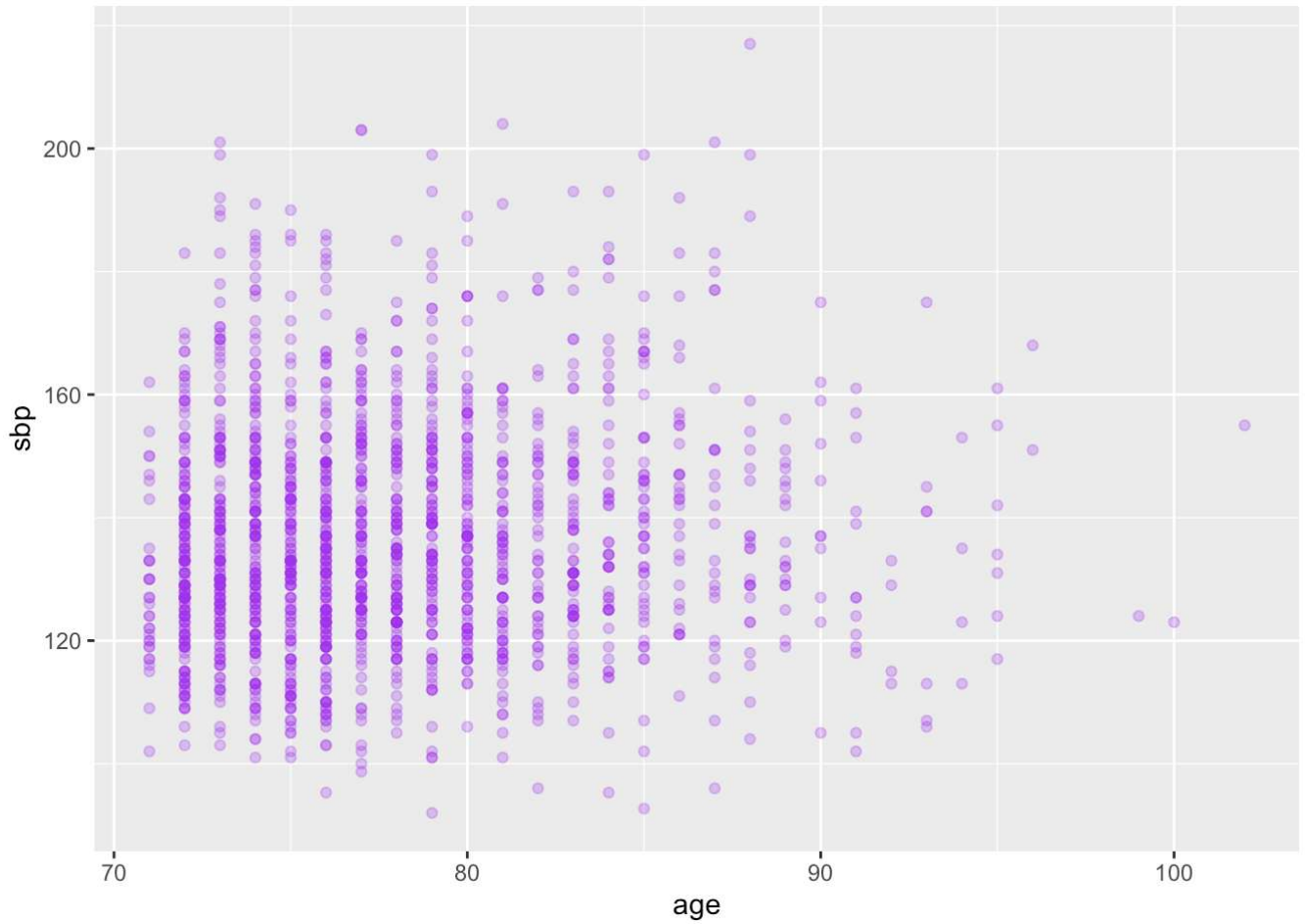
Question 9

Generate a scatter plot with exposure age and outcome SBP. You can either use base R, or ggplot2.

```
plot(d$age,d$sbp)
```




```
ggplot(data=d,aes(x=age,y=sbp)) +
  geom_point(color="purple",alpha=0.25)
```



Note that generating plots will always be extra credit and learning to generate plots is optional, but highly recommended as data visualization is an extremely useful skill. In this course we will generally use ggplot2 to plot. Here we have chosen to use a color for the dots, to make it more visually appealing, but we have also chosen alpha=0.25 which is the opacity of the dots (from 0 to 1). When you use a lower opacity than 1 (default) settings, it is easier to see where there is more overlap of data points.

Question 10

Fit a linear regression with exposure variable age (centered at 80) and outcome SBP.

```
d$age_80 <- d$age-80
fit.age <- glm(sbp~age_80,data=d)
summary(fit.age)
```

```
##
## Call:
## glm(formula = sbp ~ age_80, data = d)
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 137.27823    0.49229 278.856  < 2e-16 ***
## age_80        0.25535    0.09008   2.835  0.00464 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 348.111)
##
##     Null deviance: 581010  on 1662  degrees of freedom
## Residual deviance: 578212  on 1661  degrees of freedom
## AIC: 14456
##
## Number of Fisher Scoring iterations: 2
```

Question 11

Write down the model and interpret the estimated regression coefficients.

$SBP_i = \beta_0 + \beta_1(Age_i - 80) + \epsilon_i$

$\widehat{\beta}_0 = 137.3$: The average SBP among subjects aged 80 is 137.3 mmHg.

$\widehat{\beta} = 0.260$: On average, each 1-year increase in age is associated with an increase in SBP of 0.26 mmHg.

Question 12

Perform a hypothesis test evaluating whether SBP is associated with age.

The null hypothesis is that there is no linear association between age and mean SBP: $H_0 : \beta_1 = 0$

The alternative hypothesis is that there is a linear association between age and mean SBP: $H_1 : \beta_1 \neq 0$

The p-value corresponding to this test is 0.005. Using a significance level of 0.05, the p-value of 0.005 indicates that we can reject the null and conclude that the mean SBP has a linear association with age as interpreted above $(\widehat{\beta}_1)$.

Question 13

Calculate the following quantities:

a. Model sum of squares (MSS)

Calculate the difference between the model-predicted values and the overall mean, take the square, and then the sum of all squares.

```
d$predicted_sbp <- predict(fit.age, d)

MSS <- sum( (d$predicted_sbp - mean(d$sbp))^2 )
MSS
```

```
## [1] 2797.46
```

b. Residual sum of squares (RSS)

Calculate the difference between the observed values and the model-predicted values, take the square, and then the sum of all squares. We already calculated the residuals so you don't need to do this again.

```
d$residual <- d$sbp - d$predicted_sbp
RSS <- sum(d$residual^2)
RSS
```

```
## [1] 578212.4
```

Here's another way to obtain the residual sum of squares.

```
summary(fit.age)$deviance
```

```
## [1] 578212.4
```

c. Total sum of squares (TSS)

Calculate the difference between the observed values and the overall mean, take the square, and then the sum of all squares.

```
TSS <- sum( (d$sbp - mean(d$sbp))^2 )
TSS
```

```
## [1] 581009.8
```

Or, use the formula that TSS = MSS + RSS.

```
MSS+RSS
```

```
## [1] 581009.8
```

d. R^2

The formula for R^2 is `MSS/TSS` .

```
MSS/TSS
```

```
## [1] 0.004814825
```

Another way to obtain the R^2 is to use `lm()` instead of `glm()` and extract it from the summary. It is denoted as the `Multiple R-squared` .

```
fit.age.lm <- lm(sbp~age_80,data=d)
summary(fit.age.lm)
```

```
##
## Call:
## lm(formula = sbp ~ age_80, data = d)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -45.855 -13.087  -2.491  11.339  77.679
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 137.27823    0.49229 278.856  < 2e-16 ***
## age_80        0.25535    0.09008   2.835  0.00464 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.66 on 1661 degrees of freedom
## Multiple R-squared:  0.004815,   Adjusted R-squared:  0.004216
## F-statistic: 8.036 on 1 and 1661 DF,  p-value: 0.004641
```

The R^2 is 0.004815.

e. Pearson correlation coefficient

```
cor(d$age_80,d$sbp)
```

```
## [1] 0.06938894
```

f. Take the square of the pearson correlation coefficient. What do you notice?

```
cor(d$age_80,d$sbp)^2
```

```
## [1] 0.004814825
```

The square of the pearson correlation coefficient is the R^2!