Data_Analysis_Step_by_Step.Rmd

Dustin Pluta

September 19, 2018

Objectives

1. Is supplementation with beta carotene associated with a time-averaged increase in SBC? If so, is the effect dose dependent?
2. Is time-averaged SBC associated with time-averaged SVE?

0. Hypothesize and Select Model

The variables in the data set are:

- `ptid`: patient id
- `month`: month of observation
- `bcarot`: serum beta carotne levels
- `vite`: serum vitamin E levels
- `dose`: beta carotene supplement dose
- `age`: patient age in years
- `bmi`: patient body mass index
- `chol`: patient cholesterol
- `cauc`: time-averaged serum beta carotene area under curve
- `vauc`: time-averaged serum vitamin E area under curve

1. What is the simplest model to answer the first objective?

$$SBCAUC \sim TRT$$

2. What is a reasonable target model to answer the first objective? This model should include possible confounders and precision variables as adjustment covariates. Give a brief justification for your choice of variables to include.

$$SBCAUC \sim DOSE + AGE + MALE + BMI$$

1. Import and Tidy

1. Read in the data and display the first few rows.

```
dat <- read_delim("https://raw.githubusercontent.com/dspluta/Stats-Bootcamp/master/rData/bcarotene.txt"
```

```
## Parsed with column specification:
## cols(
##   ` ` = col_integer(),
##   ptid = col_integer(),
##   month = col_integer(),
##   bcarot = col_integer(),
##   vite = col_double(),
```

```
##   dose = col_integer(),
##   age = col_integer(),
##   male = col_integer(),
##   bmi = col_double(),
##   chol = col_double(),
##   cauc = col_double(),
##   vauc = col_double()
## )
```

```
dat
```

```
## # A tibble: 699 x 12
##      `` ``  ptid month bcarot  vite  dose   age  male   bmi  chol  cauc
##    <int> <int> <int>  <int> <dbl> <int> <int> <int> <dbl> <dbl> <dbl>
##  1     1     1     0    158  8.36    30    56     0  24.0   251 1100.
##  2     2     1     1    174  7.88    30    56     0  24.0   251 1100.
##  3     3     1     2    199  7.81    30    56     0  24.0   251 1100.
##  4     4     1     3    152  7.42    30    56     0  24.0   251 1100.
##  5     5     1     4   1095  9.46    30    56     0  24.0   251 1100.
##  6     6     1     5   1193  9.39    30    56     0  24.0   251 1100.
##  7     7     1     6   1228  9.97    30    56     0  24.0   251 1100.
##  8     8     1     6   2088  9.59    30    56     0  24.0   251 1100.
##  9     9     1     7   1248  9.87    30    56     0  24.0   251 1100.
## 10    10     1     8   1207  9.9     30    56     0  24.0   251 1100.
## # ... with 689 more rows, and 1 more variable: vauc <dbl>
```

2. Verify the data is in the correct "tidy" format.

2. Exploratory Analysis

1. Identify any missing values in the data.

```
dat %>% summarise_all(.funs = ~sum(is.na(.x)))
```

```
## # A tibble: 1 x 12
##      `` ``  ptid month bcarot  vite  dose   age  male   bmi  chol  cauc  vauc
##   <int> <int> <int>  <int> <int> <int> <int> <int> <int> <int> <int> <int>
## 1     0     0     0      6     6     0     0     0     0     0     4     4
```

```
which(is.na(dat$bcarot))
```

```
## [1]  39 167 178 216 438 456
```

```
which(is.na(dat$vite))
```

```
## [1]  39 167 178 216 438 456
```

```
which(is.na(dat$cauc))
```

```
## [1] 604 605 606 607
```

```
which(is.na(dat$vauc))
```

```
## [1] 604 605 606 607
```

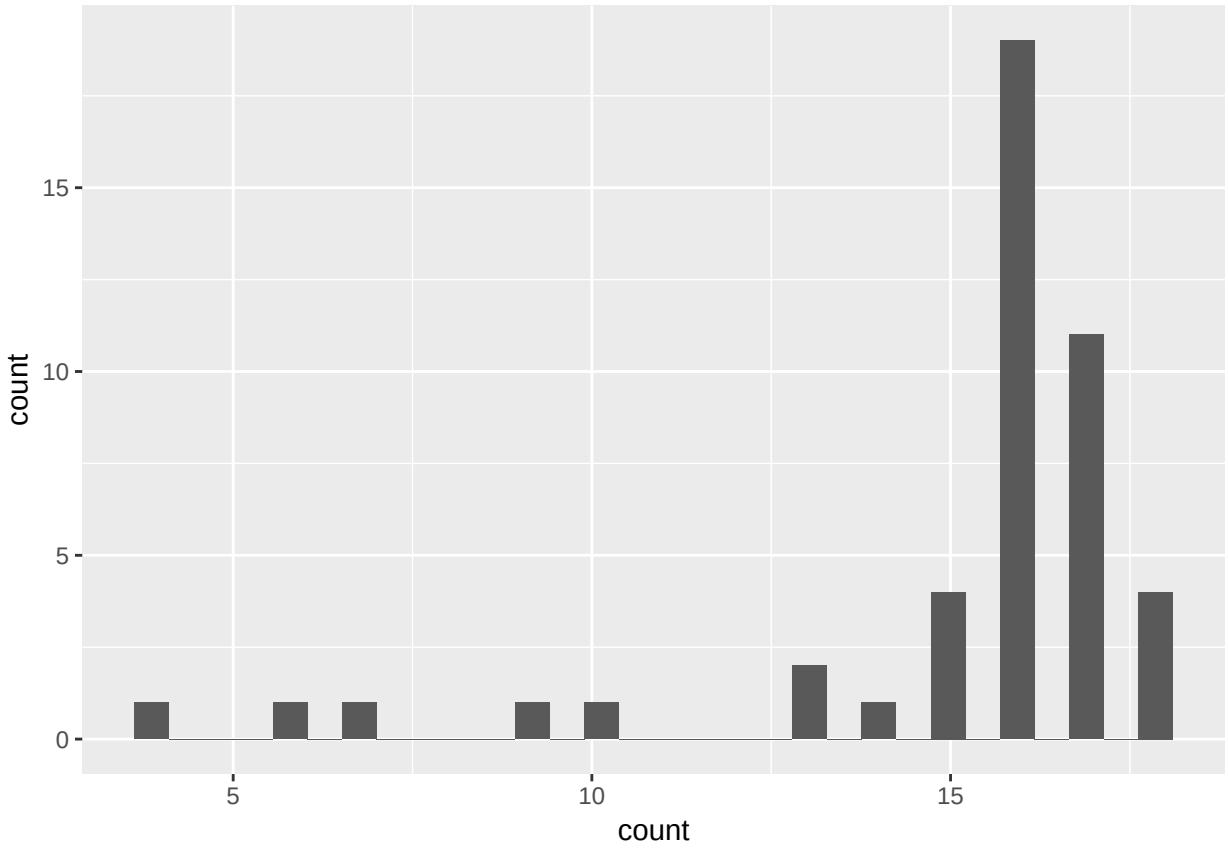
2. Plot the histogram of the number of observations per subject. What is the mean number of observations. Remove any subjects from the data set that have fewer than 3 months of observation following the baseline period (baseline is from 0 - 3 months).

```
dat %>% group_by(ptid) %>% summarize(count = n())
```

```
## # A tibble: 46 x 2
##     ptid count
##    <int> <int>
##  1     1    17
##  2     2     9
##  3     3    16
##  4     4    13
##  5     5    14
##  6     6    16
##  7     7    16
##  8     8    16
##  9     9    16
## 10    10    17
## # ... with 36 more rows
```

```
ggplot(dat %>% group_by(ptid) %>% summarize(count = n())) +
  geom_histogram(aes(x = count))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

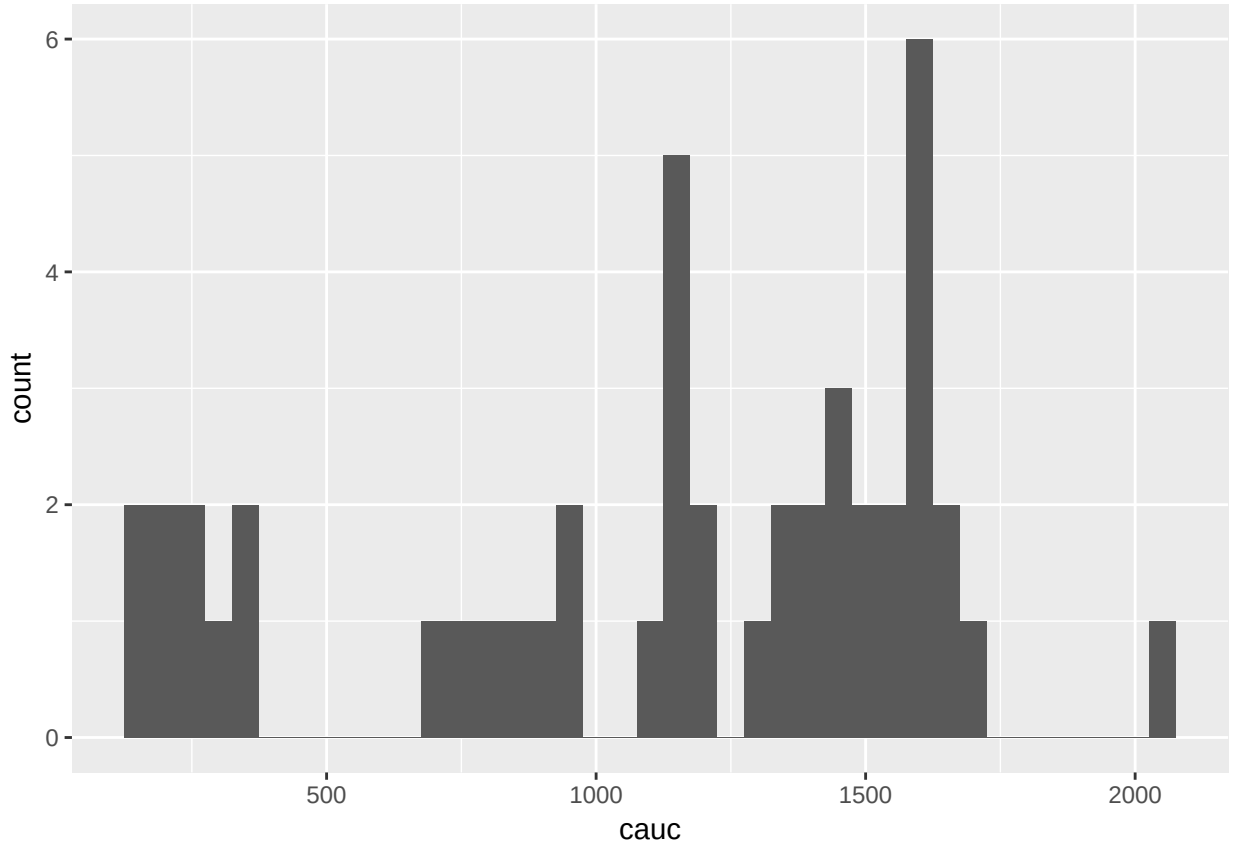


3. Since we will only analyze the time-averaged data, remove repeated rows for each subject and any columns that vary over time.

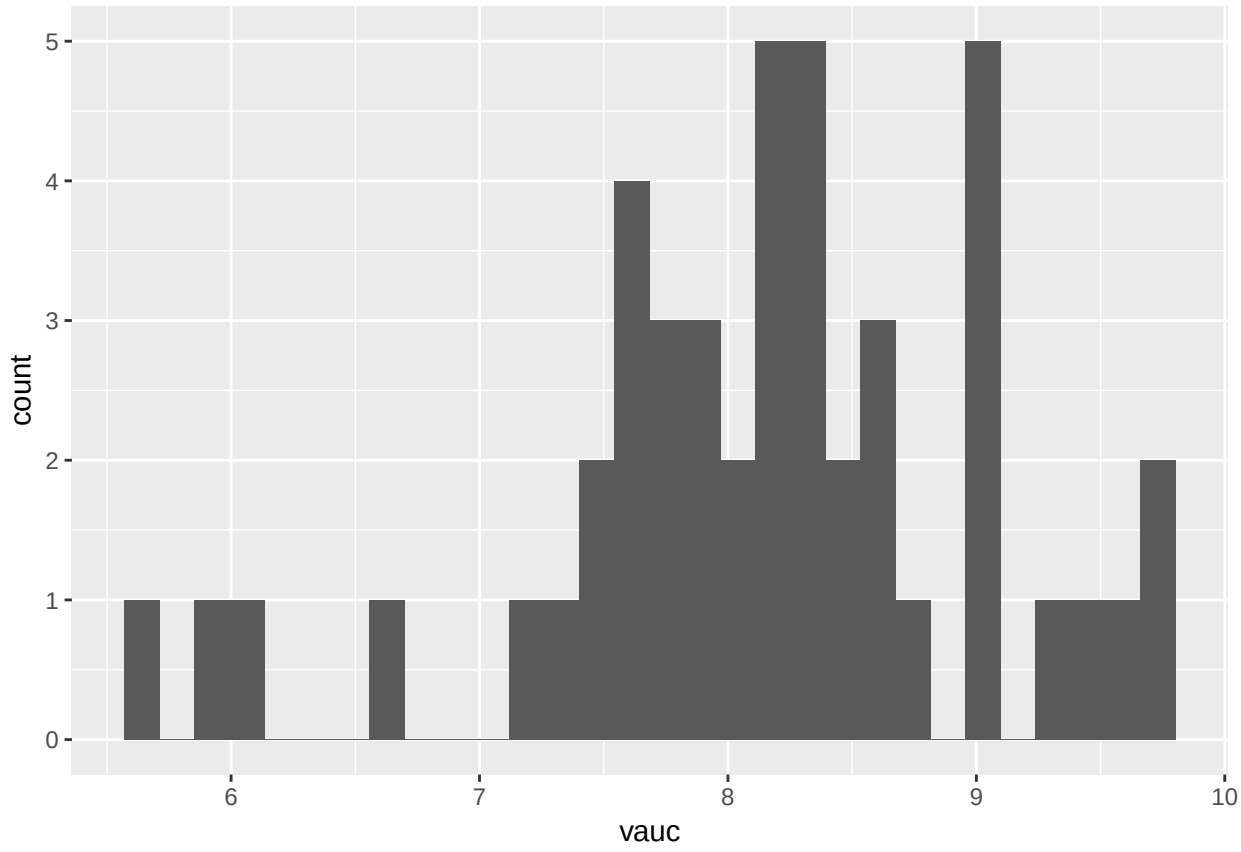
```
dat <- dat %>%
  select(-c(month, bcarot, vite)) %>%
  select(-1) %>% unique() %>% na.omit(dat)
```

4. Construct histograms or boxplots of SBC, SVE, and any of the non-categorical covariates you will use in your models. Construct frequency tables for any categorical variables you will include in your models. Record any notable observations you make regarding the distributions of the variables.

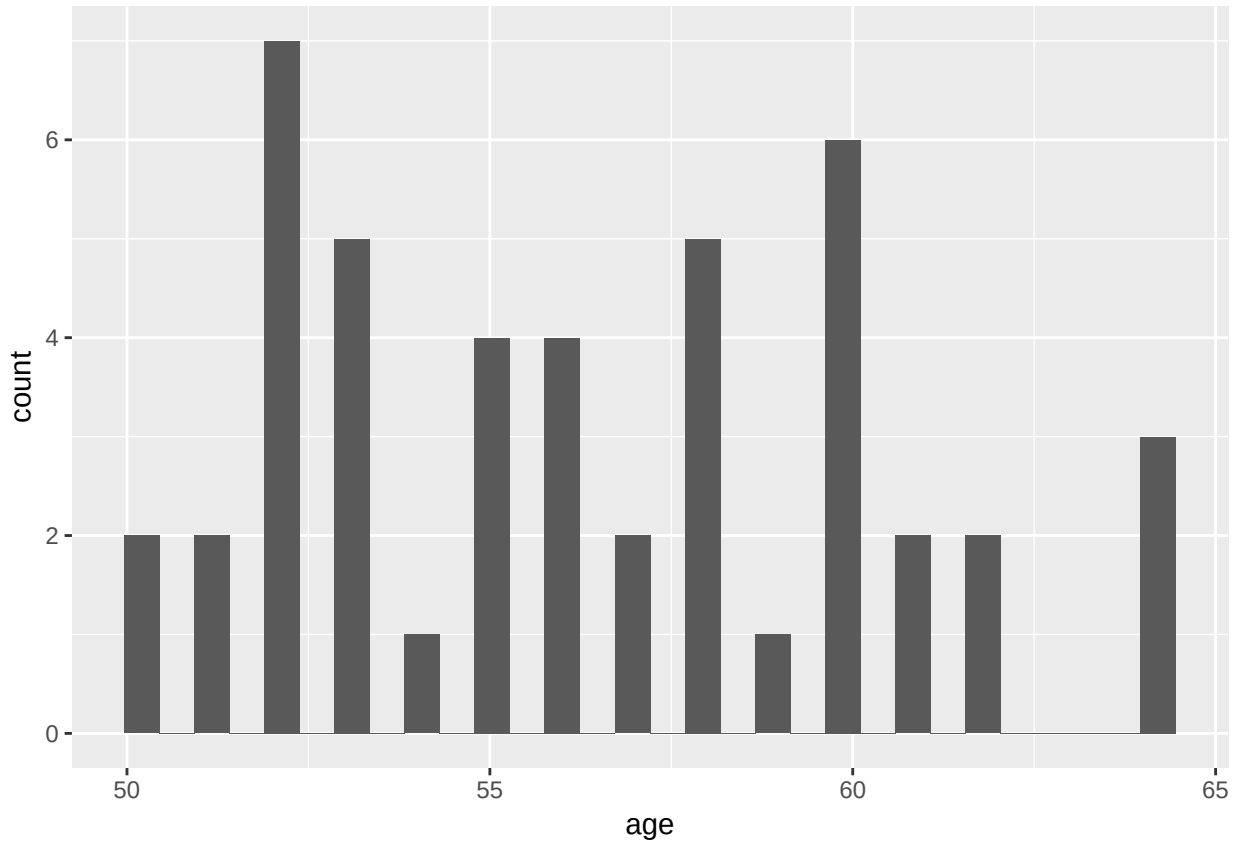
```
ggplot(dat) +
  geom_histogram(aes(x = cauc), binwidth = 50)
```



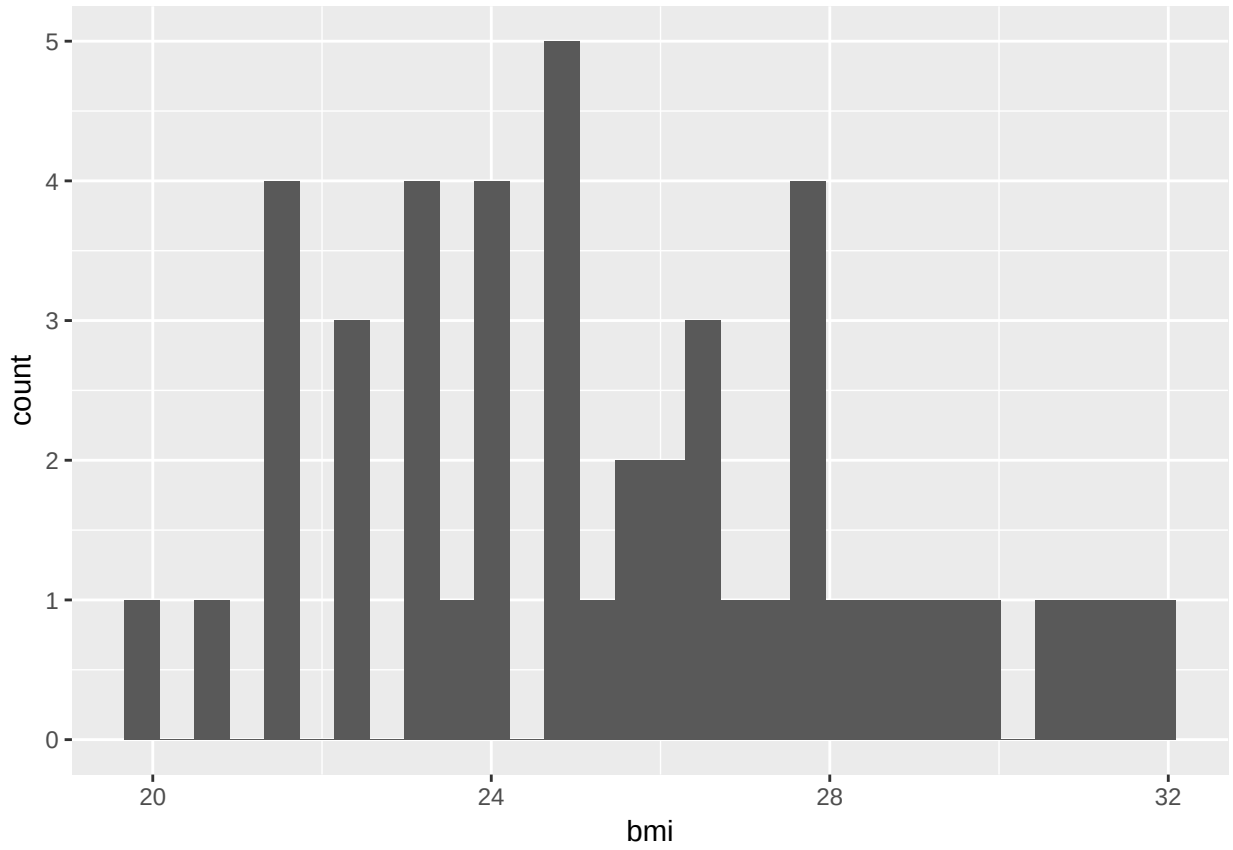

```
ggplot(dat) +
  geom_histogram(aes(x = vauc), bins = 30)
```




```
ggplot(dat) +
  geom_histogram(aes(x = age), bins = 30)
```




```
ggplot(dat) +
  geom_histogram(aes(x = bmi), bins = 30)
```




```
table(dat$dose)
```

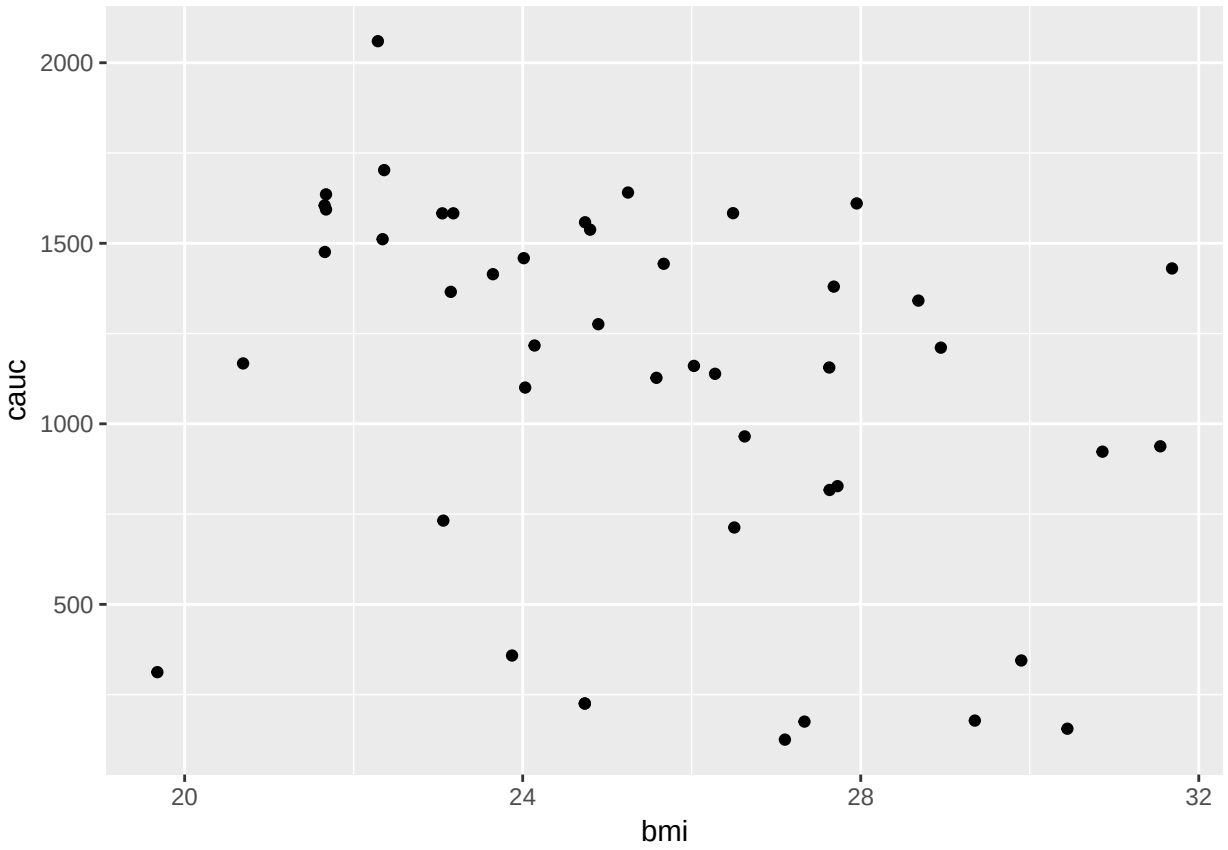
```
##
##  0 15 30 45 60
##  9 10 10  8  9
```

```
table(dat$male)
```

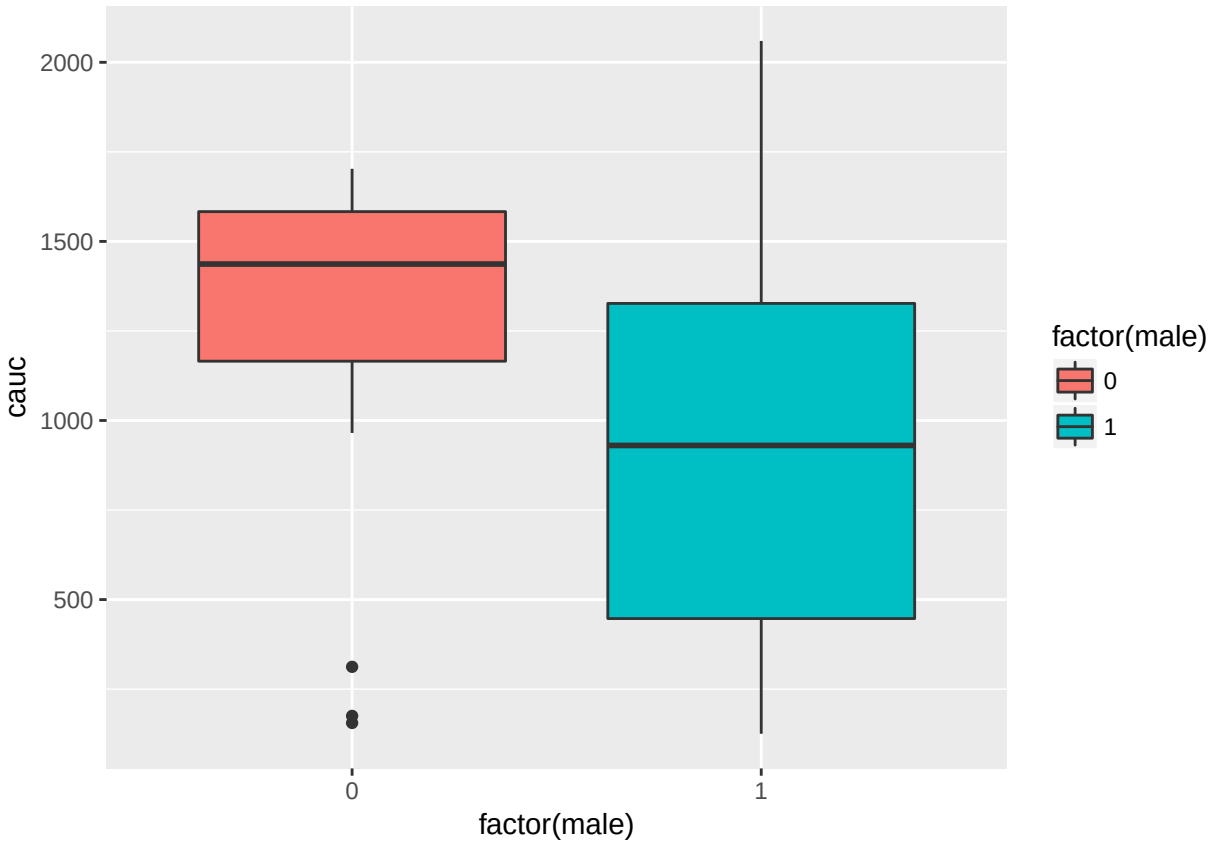
```
##
##  0  1
## 24 22
```

5. Construct scatterplots and stratified boxplots for SBC against each of the covariates you will use.

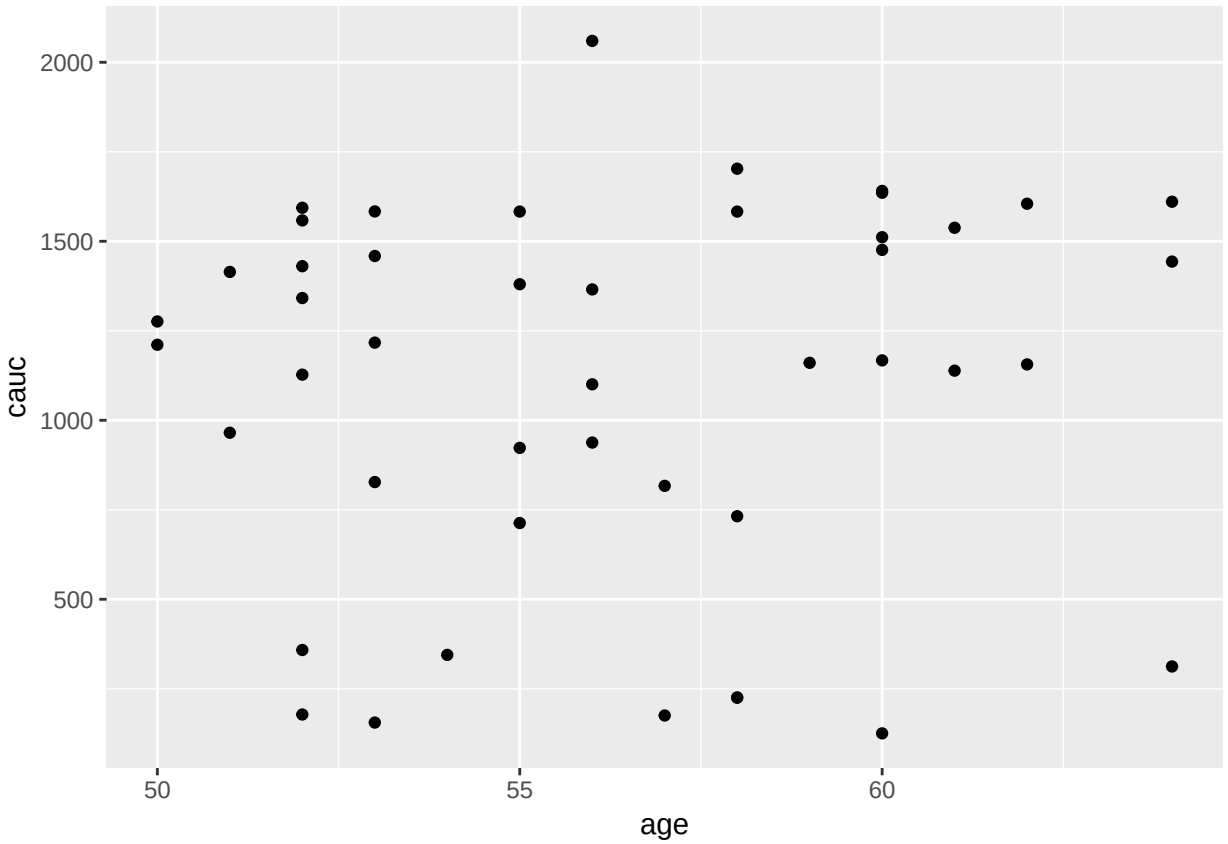
```
ggplot(dat) +
  geom_point(aes(x = bmi, y = cauc))
```



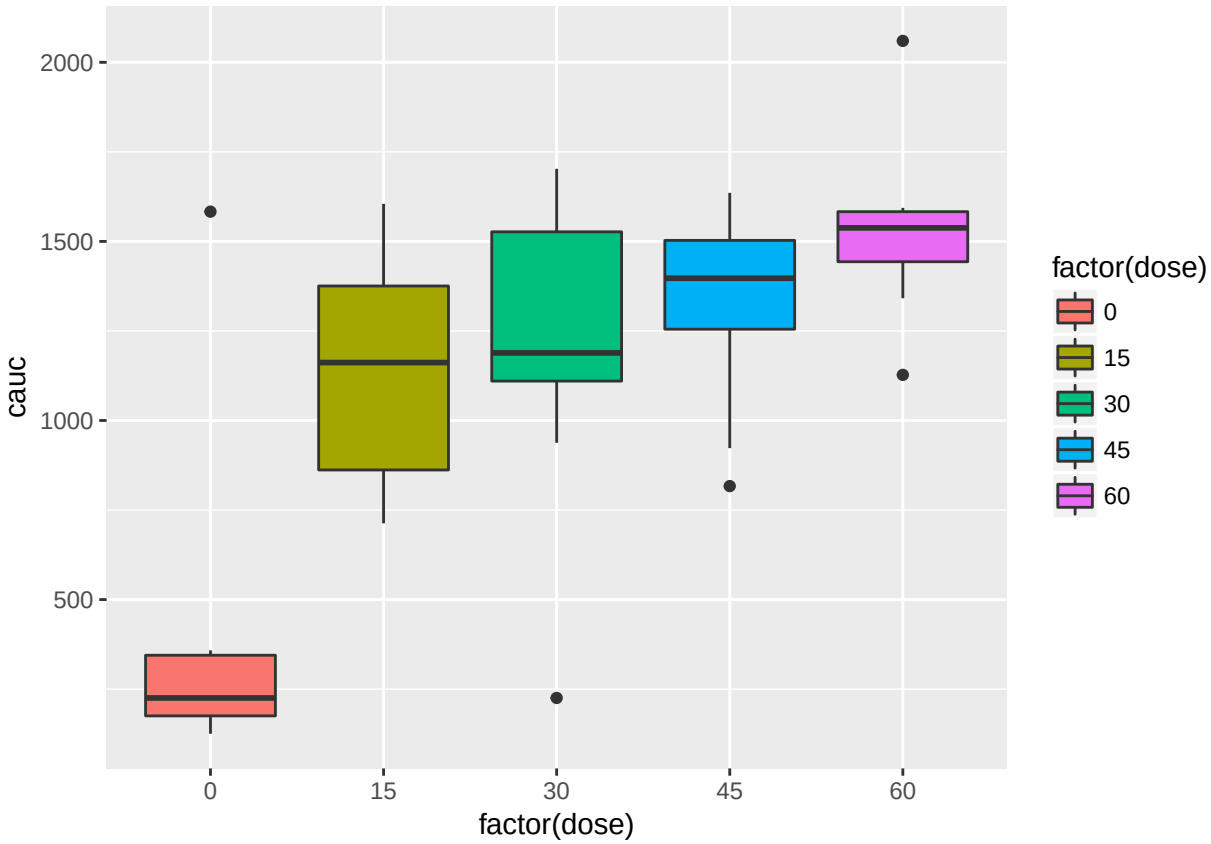

```
ggplot(dat) +
  geom_boxplot(aes(x = factor(male), y = cauc, fill = factor(male)))
```




```
ggplot(dat) +
  geom_point(aes(x = age, y = cauc))
```




```
ggplot(dat) +
  geom_boxplot(aes(x = factor(dose), y = cauc, fill = factor(dose)))
```




```
dat
```

```
## # A tibble: 46 x 8
##     ptid  dose   age  male   bmi  chol  cauc  vauc
##    <int> <int> <int> <int> <dbl> <dbl> <dbl> <dbl>
##  1     1    30    56     0  24.0   251  1100.  9.08
##  2     2    15    60     0  20.7   229  1167.  9.32
##  3     3    30    56     1  31.5   182   938.  8.64
##  4     4    60    52     0  28.7   216  1341.  9.04
##  5     5     0    57     0  27.3   217   175.  8.76
##  6     6     0    54     1  29.9   190   345.  7.71
##  7     7    60    53     1  24.0   236. 1459.  8.13
##  8     8    15    62     1  27.6   219  1156.  7.89
##  9     9    45    53     0  26.5   225  1583.  8.23
## 10    10    45    51     1  23.6   254. 1414.  8.50
## # ... with 36 more rows
```

6. Compute the correlation matrix for any continuous variables you will use in the analysis. Do any variables exhibit high correlation?

```
cor(dat[, c(2, 5, 6, 7, 8)])
```

```
##             dose        bmi       chol       cauc       vauc
## dose  1.0000000 -0.1749027  0.0896257  0.6667336  0.1298459
## bmi  -0.1749027  1.0000000 -0.1469465 -0.3395622 -0.1960592
## chol  0.0896257 -0.1469465  1.0000000  0.0451586  0.5245305
## cauc  0.6667336 -0.3395622  0.0451586  1.0000000  0.1519869
## vauc  0.1298459 -0.1960592  0.5245305  0.1519869  1.0000000
```

7. Did you notice any significant violations of model assumptions?
8. Are there obvious outliers or potentially erroneous observations?
9. Are there any transformations we should consider applying to the data?

3. Model Fitting

1. Fit the simple model for objective (1) and print the summary.

```
fit_simple <- lm(cauc ~ I(dose != 0), data = dat)
summary(fit_simple)
```

```
##
## Call:
## lm(formula = cauc ~ I(dose != 0), data = dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1061.81  -201.25   -32.77   244.01  1198.85
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)         384.2      125.1   3.072  0.00364 **
## I(dose != 0)TRUE    903.0      139.4   6.475 6.74e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 375.2 on 44 degrees of freedom
## Multiple R-squared:  0.488,  Adjusted R-squared:  0.4763
## F-statistic: 41.93 on 1 and 44 DF,  p-value: 6.744e-08
```

2. Fit the target model for objective (1) and print the summary.

```
fit_target <- lm(cauc ~ dose + age + bmi + male, data = dat)
summary(fit_target)
```

```
##
## Call:
## lm(formula = cauc ~ dose + age + bmi + male, data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -787.64 -201.98  -25.24  155.71  743.51
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1314.447   1110.244   1.184   0.2433
## dose          15.300      2.617   5.847 7.13e-07 ***
## age            2.381     14.459   0.165   0.8700
## bmi          -26.138     20.305  -1.287   0.2052
## male        -252.015    112.685  -2.236   0.0308 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 363.8 on 41 degrees of freedom
## Multiple R-squared:  0.5515, Adjusted R-squared:  0.5077
## F-statistic:  12.6 on 4 and 41 DF,  p-value: 8.942e-07
```

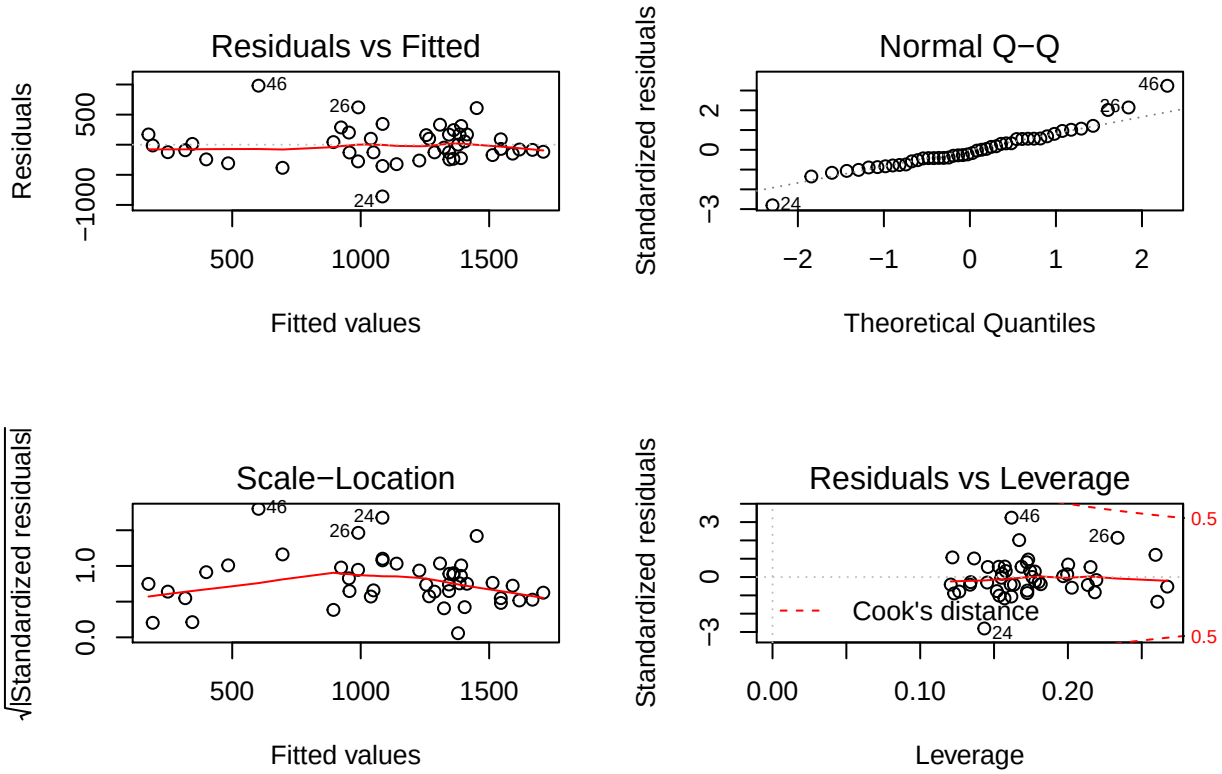
```r
fit_target <- lm(cauc ~ factor(dose) + age + bmi + male, data = dat)
summary(fit_target)
```

```
##
## Call:
## lm(formula = cauc ~ factor(dose) + age + bmi + male, data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -859.24 -166.04  -59.02  169.22  980.69
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)    1285.9305  1024.9417   1.255   0.2173
## factor(dose)15  720.5888   152.1494   4.736 3.02e-05 ***
## factor(dose)30  767.3577   153.4339   5.001 1.32e-05 ***
## factor(dose)45  903.3327   161.4669   5.595 2.05e-06 ***
## factor(dose)60 1064.1785   159.4445   6.674 6.81e-08 ***
## age              -0.5658    13.3747  -0.042   0.9665
## bmi             -28.1483    18.6298  -1.511   0.1391
## male           -239.6364   102.8388  -2.330   0.0252 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 330.6 on 38 degrees of freedom
## Multiple R-squared:  0.6567, Adjusted R-squared:  0.5934
## F-statistic: 10.38 on 7 and 38 DF,  p-value: 3.316e-07
```

4. Diagnostics

1. Examine the diagnostic plots for each of the models you fit above. Comment on any issues or potential problems you see.

```r
par(mfrow = c(2, 2))
plot(fit_target)
```



2. Compare the adjusted R^2 for the simple and target models for each of the two objectives. Do the more complicated models seem to yield significantly better fits than the simple models?

```
summary(fit_simple)$adj.r.squared
```

```
## [1] 0.4763164
```

```
summary(fit_target)$adj.r.squared
```

```
## [1] 0.5934332
```

3. Do any of the models seem to be invalid due to violation of assumptions? If so, discard these models for the remainder of the analysis.

5. Test and Interpret

1. Perform hypothesis tests and construct confidence intervals to answer the questions of interest for objective (1).
2. Produce a plot to help convey the main message of your results for objective (1).
3. Give practical interpretations of your coefficient estimates related to the questions of interest. Use plain English, and explain the results in a way that would be meaningful to a clinician. Avoid causal language.

6. Post-hoc Analysis

1. Based on the results of the analysis, can you propse a better model? Fit a few other candidate models and use goodness-of-fit measures (e.g. adjusted-R^2) to compare to the simple and target models.

2. Were there any outliers or highly influential observations? Do the results change substantially if we remove these observations?

7. Report

1. Compile the above work into a concise and insightful report.