Exploratory and Descriptive Methods

Contents

Configuring R

Functions from these packages will be used throughout this document:

```
library(conflicted) # check for conflicting function definitions
# library(printr) # inserts help-file output into markdown output
library(rmarkdown) # Convert R Markdown documents into a variety of formats.
library(pander) # format tables for markdown
library(ggplot2) # graphics
library(ggfortify) # help with graphics
```

```r
library(dplyr) # manipulate data
library(tibble) # `tibble`s extend `data.frame`s
library(magrittr) # `%>%` and other additional piping tools
library(haven) # import Stata files
library(knitr) # format R output for markdown
library(tidyr) # Tools to help to create tidy data
library(plotly) # interactive graphics
library(dobson) # datasets from Dobson and Barnett 2018
library(parameters) # format model output tables for markdown
library(haven) # import Stata files
library(latex2exp) # use LaTeX in R code (for figures and tables)
library(fs) # filesystem path manipulations
library(survival) # survival analysis
library(survminer) # survival analysis graphics
library(KMsurv) # datasets from Klein and Moeschberger
library(parameters) # format model output tables for
library(webshot2) # convert interactive content to static for pdf
library(forcats) # functions for categorical variables ("factors")
library(stringr) # functions for dealing with strings
library(lubridate) # functions for dealing with dates and times
library(broom) # Summarizes key information about statistical objects in tidy tibbles
library(broom.helpers) # Provides suite of functions to work with regression model 'broom::tidy()' t
```

Here are some R settings I use in this document:

```r
rm(list = ls()) # delete any data that's already loaded into R

conflicts_prefer(dplyr::filter)
ggplot2::theme_set(
  ggplot2::theme_bw() +
        # ggplot2::labs(col = "") +
    ggplot2::theme(
      legend.position = "bottom",
      text = ggplot2::element_text(size = 12, family = "serif")))

knitr::opts_chunk$set(message = FALSE)
options('digits' = 6)

panderOptions("big.mark", ",")
pander::panderOptions("table.emphasize.rownames", FALSE)
pander::panderOptions("table.split.table", Inf)
conflicts_prefer(dplyr::filter) # use the `filter()` function from dplyr() by default
legend_text_size = 9
run_graphs = TRUE
```

> **Note**
>
> This chapter is adapted from Vittinghoff et al. (2012), Chapter 2.

---

1 Introduction

Before fitting regression models, it is good practice to explore and summarize the data. Exploratory data analysis (EDA) serves several purposes:

- Understand the distribution of each variable individually.

- Detect unusual values, outliers, and missing data.
- Identify patterns and relationships among variables.
- Motivate choices of model structure (e.g., transformations).
- Provide context for interpreting model results.

In this chapter we illustrate exploratory and descriptive techniques using the Western Collaborative Group Study (WCGS) dataset.

1.0.1 Coronary heart disease (WCGS) study data

Let's use the data from the Western Collaborative Group Study (WCGS) (Rosenman et al. (1975)) to explore multiple logistic regression:

From Vittinghoff et al. (2012):

"The **Western Collaborative Group Study (WCGS)** was a large epidemiological study designed to investigate the association between the"type A" behavior pattern and coronary heart disease (CHD)".

Exercise 1.1. What is "type A" behavior?

Solution

Solution 1.1. *From Wikipedia, "Type A and Type B personality theory":*
"The hypothesis describes Type A individuals as outgoing, ambitious, rigidly organized, highly status-conscious, impatient, anxious, proactive, and concerned with time management….
The hypothesis describes Type B individuals as a contrast to those of Type A. Type B personalities, by definition, are noted to live at lower stress levels. They typically work steadily and may enjoy achievement, although they have a greater tendency to disregard physical or mental stress when they do not achieve."

Study design

from `?faraway::wcgs`*:*

3154 healthy young men aged 39-59 from the San Francisco area were assessed for their personality type. All were free from coronary heart disease at the start of the research. Eight and a half years later change in CHD status was recorded.

Details (from `faraway::wcgs`*)*

The WCGS began in 1960 with 3,524 male volunteers who were employed by 11 California companies. Subjects were 39 to 59 years old and free of heart disease as determined by electrocardiogram. After the initial screening, the study population dropped to 3,154 and the number of companies to 10 because of various exclusions. The cohort comprised both blue- and white-collar employees.

1.0.2 Baseline data collection

socio-demographic characteristics:

- age
- education
- marital status
- income

- occupation
- physical and physiological including:
- height
- weight
- blood pressure
- electrocardiogram
- corneal arcus

---

biochemical measurements:

- cholesterol and lipoprotein fractions;
- medical and family history and use of medications;

---

behavioral data:

- Type A interview,
- smoking,
- exercise
- alcohol use.

Later surveys added data on:

- anthropometry
- triglycerides
- Jenkins Activity Survey
- caffeine use

Average follow-up continued for 8.5 years with repeat examinations.

```
### load the data directly from a UCSF website:
url <- paste0(
  "https://regression.ucsf.edu/sites/g/files/",
  "tkssra6706/f/wysiwyg/home/data/wcgs.dta"
)
wcgs <- haven::read_dta(url)
```

```
wcgs <- wcgs |>
  labelled::set_variable_labels(
    age = "Age (years)",
    chol = "Cholesterol (mg/dL)",
    sbp = "Systolic BP (mmHg)",
    dbp = "Diastolic BP (mmHg)",
    bmi = "BMI (kg/m\u00B2)",
    weight = "Weight (lbs)",
    ncigs = "Cigarettes per day",
    chd69 = "CHD event by 1969",
    smoke = "Current smoker",
    arcus = "Arcus senilis",
    dibpat = "Behavioral pattern (A/B)",
    behpat = "Behavioral pattern (A1/A2/B3/B4)",
    wghtcat = "Weight category",
    agec = "Age group"
  )
```

2 Summarizing a single variable

2.1 Continuous variables

2.1.1 Measures of center

Definition 2.1 (Sample mean)**.** The **sample mean** of a variable x measured on n observations is:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Definition 2.2 (Median)**.** The **median** is the value that divides the sorted data in half: 50% of observations fall below and 50% above. The median is more robust to outliers than the mean.

2.1.2 Measures of spread

Definition 2.3 (Sample variance)**.** The **sample variance** is:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Definition 2.4 (Sample standard deviation)**.** The **sample standard deviation** is the square root of the sample variance:

$$s = \sqrt{s^2}$$

The standard deviation has the same units as the original variable, making it easier to interpret than the variance.

Definition 2.5 (Quantiles and percentiles)**.** The p**th quantile** (or $(100p)$**th percentile**) of a variable is the value below which a proportion p of the data falls.
Special cases:

- The **median** is the 50th percentile ($p = 0.5$).
- The **first quartile** (Q1) is the 25th percentile ($p = 0.25$).
- The **third quartile** (Q3) is the 75th percentile ($p = 0.75$).

Definition 2.6 (Interquartile range)**.** The **interquartile range** (IQR) is:

$$\text{IQR} = Q_3 - Q_1$$

The IQR contains the middle 50% of the data and is more robust to outliers than the standard deviation.

Table 1: WCGS: descriptive statistics for continuous variables

Characteristic	**N = 3,154**[1]
Age (years)	46.3 (5.5); 45.0 [42.0, 50.0]
Cholesterol (mg/dL)	226.4 (43.4); 223.0 [197.0, 253.0]
Unknown	12
Systolic BP (mmHg)	128.6 (15.1); 126.0 [120.0, 136.0]
Diastolic BP (mmHg)	82.0 (9.7); 80.0 [76.0, 86.0]
BMI (kg/m^2)	24.5 (2.6); 24.4 [23.0, 25.8]
Weight (lbs)	170.0 (21.1); 170.0 [155.0, 182.0]

[1]Mean (SD); Median [Q1, Q3]

2.1.3 Summary statistics in R

The `summary()` function provides a quick overview of any variable:

```
summary(wcgs$chol)
#>    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.     NAs
#>     103     197     223     226     253     645      12
```

For a formatted summary table, `gtsummary::tbl_summary()` is useful:

```
wcgs |>
  dplyr::select(age, chol, sbp, dbp, bmi, weight) |>
  gtsummary::tbl_summary(
    statistic = list(
      gtsummary::all_continuous() ~
        "{mean} ({sd}); {median} [{p25}, {p75}]"
    ),
    digits = gtsummary::all_continuous() ~ 1
  )
```

2.2 Binary and categorical variables

For binary and categorical variables, the natural descriptive statistics are **frequencies** (counts) and **proportions** (relative frequencies).

```
wcgs |>
  dplyr::select(chd69, smoke, dibpat, behpat, wghtcat) |>
  gtsummary::tbl_summary()
```

3 Graphical methods

Graphs are often more informative than summary statistics alone. Different graph types are appropriate for different variable types.

3.1 Histograms

A **histogram** displays the distribution of a continuous variable by dividing its range into bins and counting the number of observations in each.

```
wcgs |>
  ggplot2::ggplot() +
  ggplot2::aes(x = chol) +
  ggplot2::geom_histogram(bins = 30, fill = "steelblue", color = "white") +
  ggplot2::labs(y = "Count")
```

Table 2: Frequency table for categorical variables in the WCGS dataset

Characteristic	**N = 3,154**[1]
CHD event by 1969	257 (8.1%)
Current smoker	1,502 (48%)
Behavioral pattern (A/B)	
Type B	1,565 (50%)
Type A	1,589 (50%)
Behavioral pattern (A1/A2/B3/B4)	
A1	264 (8.4%)
A2	1,325 (42%)
B3	1,216 (39%)
B4	349 (11%)
Weight category	
< 140	232 (7.4%)
140-170	1,538 (49%)
170-200	1,171 (37%)
> 200	213 (6.8%)

[1] n (%)

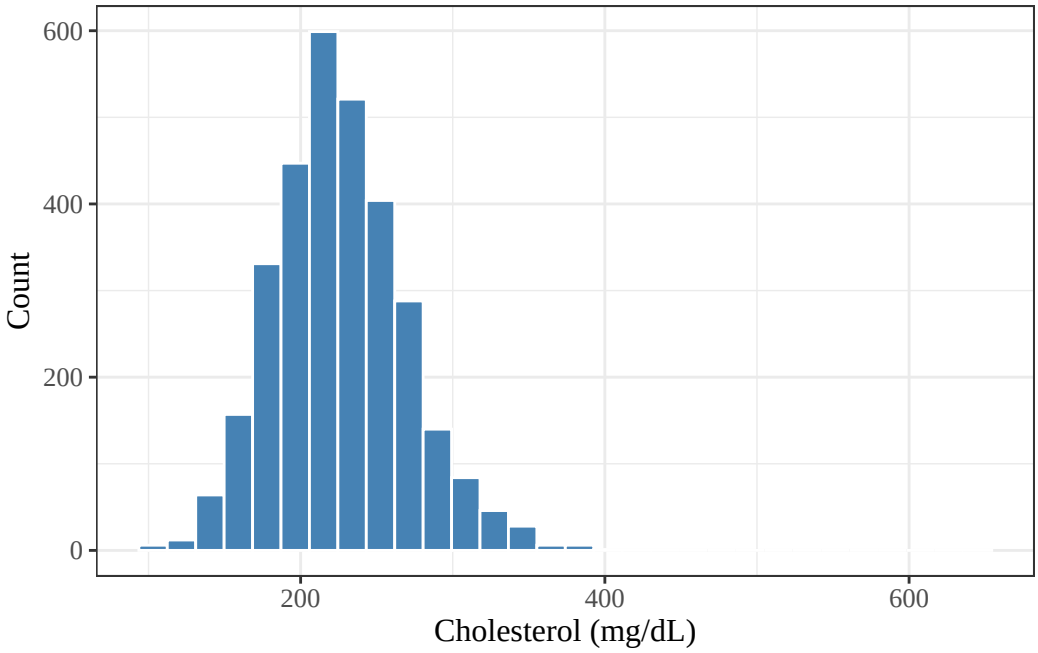

Figure 1: Histogram of total cholesterol in the WCGS dataset

The histogram in Figure 1 shows that cholesterol in the WCGS is roughly bell-shaped (approximately normal), with most values between 150 and 350 mg/dL.

3.2 Density plots

A **density plot** shows a smoothed estimate of the distribution, which can be easier to interpret than a histogram.

```
wcgs |>
  ggplot2::ggplot() +
  ggplot2::aes(x = chol) +
  ggplot2::geom_density(fill = "steelblue", alpha = 0.5) +
  ggplot2::labs(y = "Density")
```

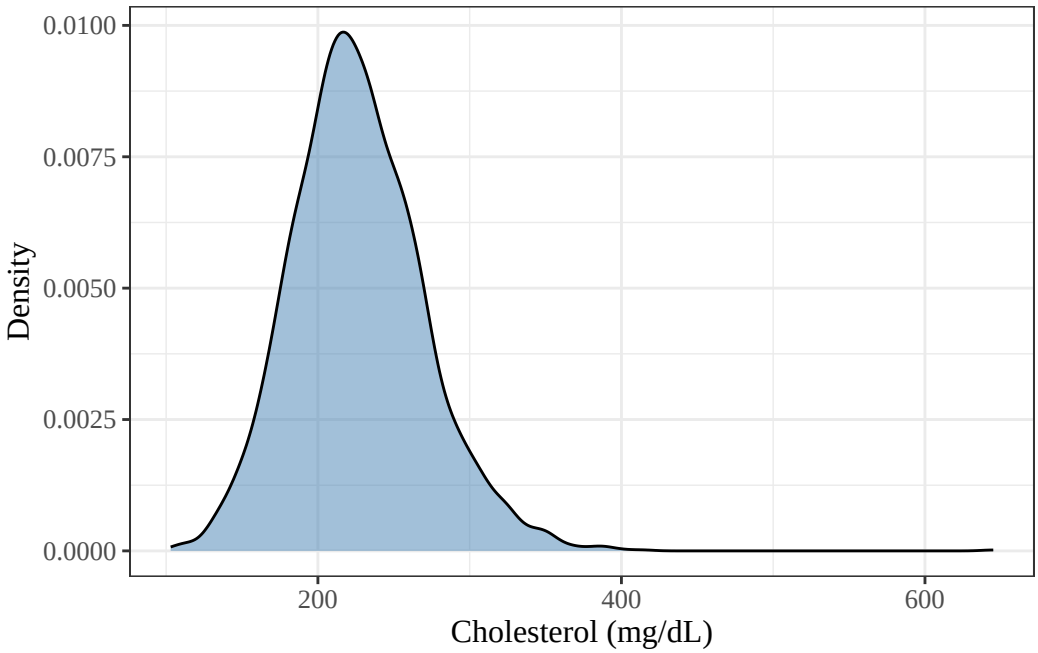

Figure 2: Density plot of total cholesterol in the WCGS dataset

3.3 Box plots

A **box plot** (or box-and-whisker plot) summarizes a distribution using the median, quartiles, and extreme values.

The box spans Q1 to Q3 (the IQR). The line inside the box is the median. The whiskers extend to the most extreme observations within $1.5 \times$ IQR of the box. Points beyond the whiskers are plotted individually as potential outliers.

```
wcgs |>
  ggplot2::ggplot() +
  ggplot2::aes(y = chol) +
  ggplot2::geom_boxplot(fill = "steelblue") +
  ggplot2::theme(axis.text.x = ggplot2::element_blank())
```

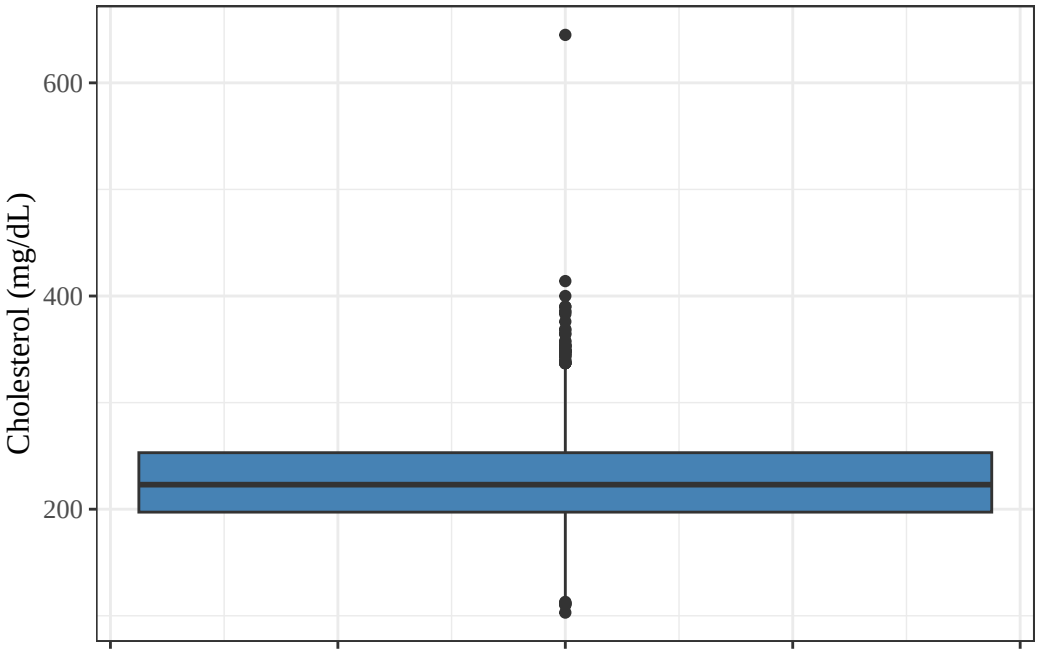

Figure 3: Box plot of total cholesterol in the WCGS dataset

3.4 Bar charts

A **bar chart** displays the frequency or proportion of each category of a categorical variable.

```
wcgs |>
  ggplot2::ggplot() +
  ggplot2::aes(x = behpat) +
  ggplot2::geom_bar(fill = "steelblue") +
  ggplot2::labs(y = "Count")
```

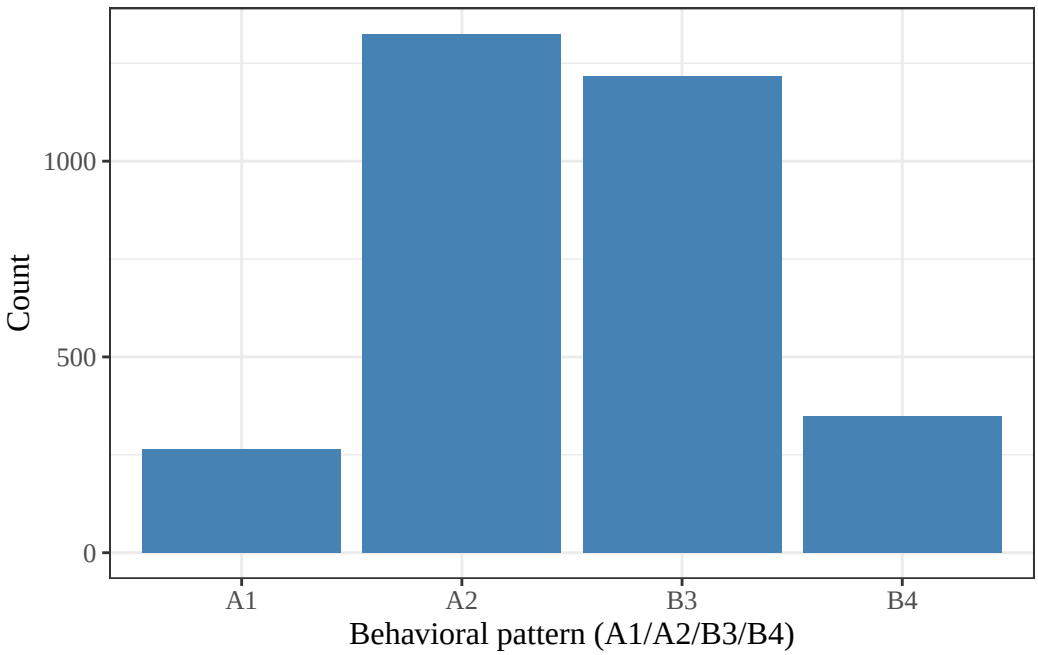

Figure 4: Bar chart of behavioral pattern in the WCGS dataset

3.5 Normal probability (Q-Q) plots

A **quantile-quantile (Q-Q) plot** compares the observed quantiles of a variable to the theoretical quantiles of a normal distribution. If the variable is approximately normally distributed, the points fall close to the diagonal reference line.

```
wcgs |>
  ggplot2::ggplot() +
  ggplot2::aes(sample = chol) +
  ggplot2::stat_qq() +
  ggplot2::stat_qq_line(color = "red") +
  ggplot2::labs(x = "Theoretical quantiles")
```

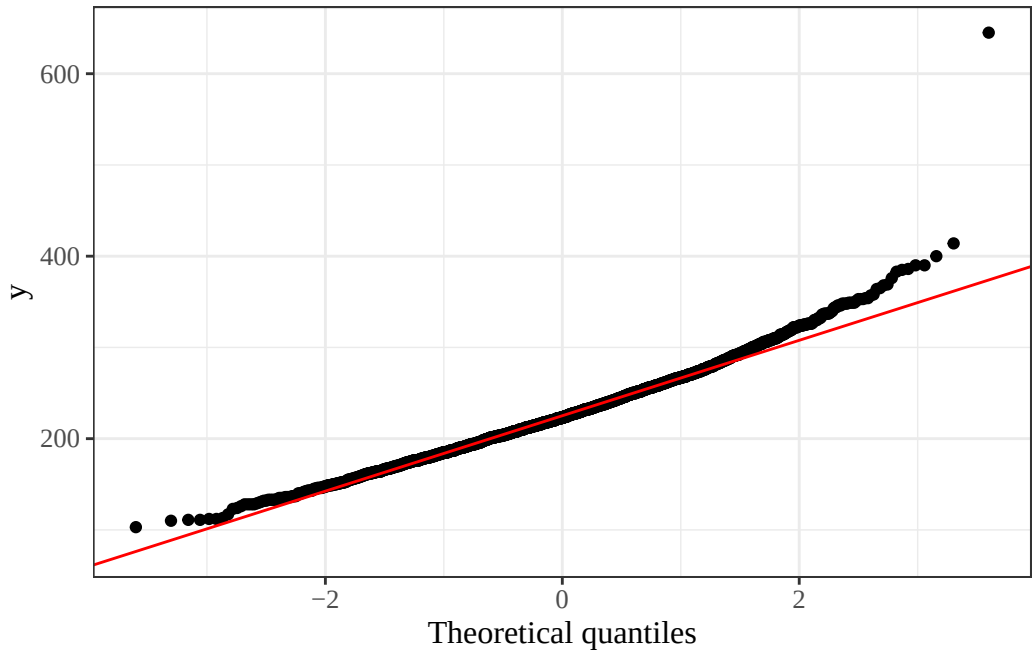

Figure 5: Normal Q-Q plot for total cholesterol in the WCGS dataset

4 Bivariate relationships

4.1 Two continuous variables: scatter plots and correlation

4.1.1 Scatter plots

A **scatter plot** displays the relationship between two continuous variables by plotting each observation as a point.

```
wcgs |>
  ggplot2::ggplot() +
  ggplot2::aes(x = sbp, y = chol) +
  ggplot2::geom_point(alpha = 0.3) +
  ggplot2::geom_smooth(method = "lm", se = TRUE, color = "red")
```

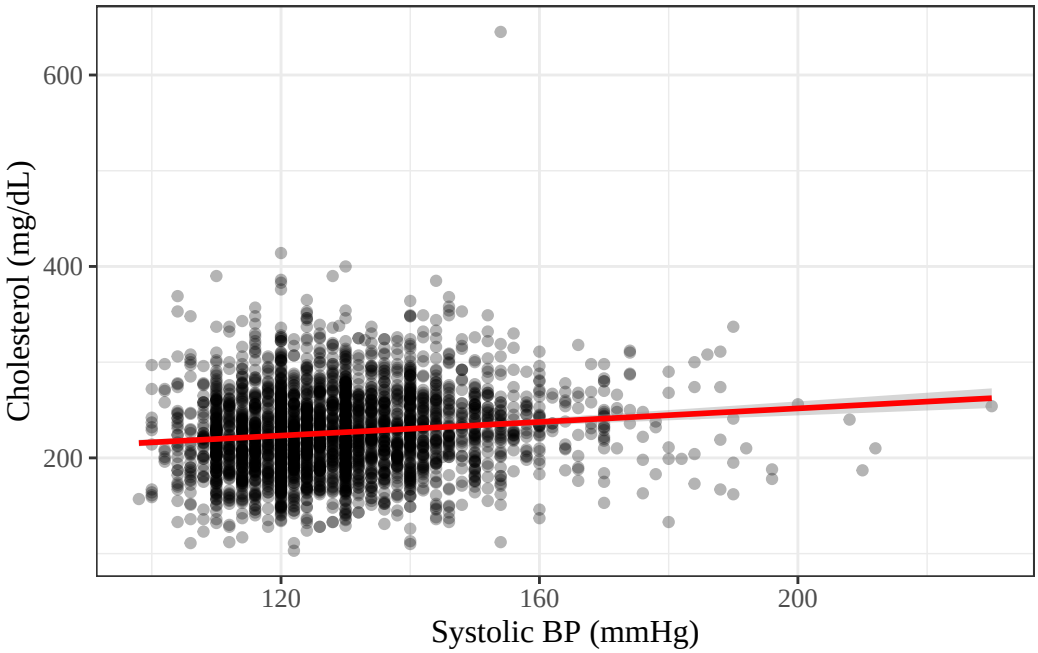

Figure 6: Cholesterol vs. systolic BP in the WCGS dataset

4.1.2 Correlation

Definition 4.1 (Pearson correlation coefficient)**.** The **Pearson correlation coefficient** measures the strength and direction of a linear relationship between two continuous variables X and Y:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

The correlation coefficient takes values in $[-1, 1]$:

- $r = 1$: perfect positive linear relationship.
- $r = -1$: perfect negative linear relationship.
- $r = 0$: no linear relationship.

```
cor(wcgs$chol, wcgs$sbp, use = "complete.obs")
#> [1] 0.123061
```

Caution

Correlation measures only **linear** relationships. Two variables can be strongly related but have a near-zero correlation if the relationship is nonlinear. Furthermore, correlation does not imply causation.

4.2 Continuous outcome by a binary or categorical variable

4.2.1 Side-by-side box plots

Side-by-side box plots are useful for comparing the distribution of a continuous variable across groups defined by a categorical variable.

```
wcgs |>
  ggplot2::ggplot() +
  ggplot2::aes(x = smoke, y = chol, fill = smoke) +
  ggplot2::geom_boxplot() +
  ggplot2::theme(legend.position = "none")
```

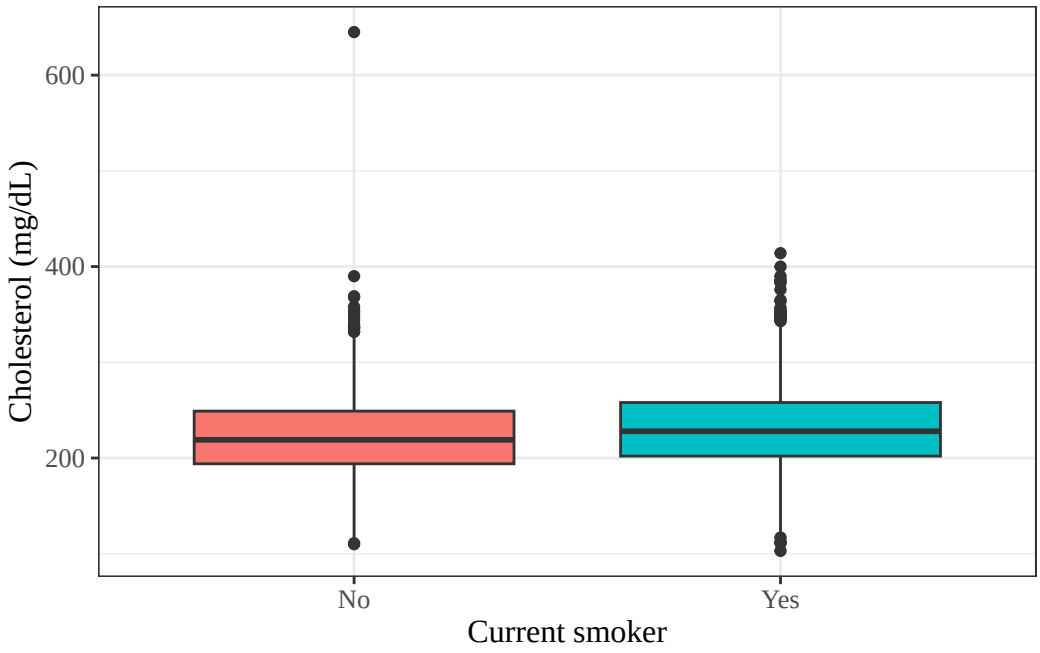

Figure 7: Box plots of cholesterol by smoking status in the WCGS dataset

---

```
wcgs |>
  ggplot2::ggplot() +
  ggplot2::aes(x = behpat, y = chol, fill = behpat) +
  ggplot2::geom_boxplot() +
  ggplot2::theme(legend.position = "none")
```

Table 3: Cholesterol summary statistics by CHD status

Characteristic	**Overall** N = 3,154[1]	**No** N = 2,897[1]	**Yes** N = 257[1]	**p-value**[2]
Cholesterol (mg/dL)	226.4 (43.4)	224.3 (42.2)	250.1 (49.4)	<0.001
Unknown	12	12	0	
Systolic BP (mmHg)	128.6 (15.1)	128.0 (14.7)	135.4 (17.5)	<0.001
BMI (kg/m^2)	24.5 (2.6)	24.5 (2.6)	25.1 (2.6)	<0.001

[1]Mean (SD)
[2]Wilcoxon rank sum test

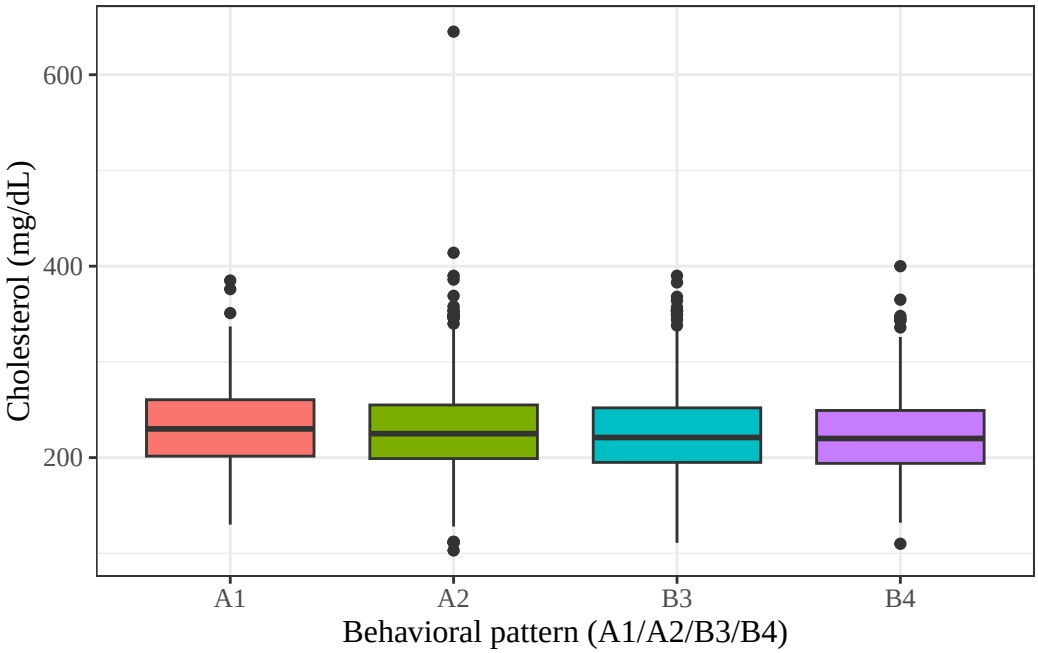

Figure 8: Box plots of cholesterol by behavioral pattern in the WCGS dataset

4.2.2 Summary statistics by group

We can compute summary statistics separately for each group:

```
wcgs |>
  dplyr::select(chol, sbp, bmi, chd69) |>
  gtsummary::tbl_summary(
    by = chd69,
    statistic = list(
      gtsummary::all_continuous() ~
        "{mean} ({sd})"
    ),
    digits = gtsummary::all_continuous() ~ 1
  ) |>
  gtsummary::add_overall() |>
  gtsummary::add_p()
```

Table 4: Cross-tabulation of smoking status and CHD event in WCGS

```
table(Smoking = wcgs$smoke, CHD = wcgs$chd69)
#>         CHD
#> Smoking   No  Yes
#>     No  1554   98
#>     Yes 1343  159
```

Table 5

```
prop.table(table(Smoking = wcgs$smoke, CHD = wcgs$chd69), margin = 1)
#>         CHD
#> Smoking       No      Yes
#>     No  0.940678 0.059322
#>     Yes 0.894141 0.105859
```

4.3 Two categorical variables: contingency tables

4.3.1 Cross-tabulation

A **contingency table** (or **cross-tabulation**) displays the joint frequency distribution of two categorical variables.

We can also compute row proportions to examine the conditional distribution of CHD given smoking status:

Using `gtsummary` for a formatted table:

```
wcgs |>
  dplyr::select(smoke, chd69) |>
  gtsummary::tbl_summary(by = chd69) |>
  gtsummary::add_overall() |>
  gtsummary::add_p()
```

5 Data transformations

When a continuous variable has a **right-skewed** distribution (long tail to the right), a **logarithmic transformation** often makes the distribution more approximately normal. Log-transformed variables also have a natural multiplicative interpretation: a one-unit increase in $\log(x)$ corresponds to multiplying x by $e \approx 2.72$.

The WCGS dataset already contains the log-transformed versions of some variables:

- `lnsbp`: log(SBP)
- `lnwght`: log(weight)

Table 6: Smoking status and CHD event in WCGS

Characteristic	**Overall** N = 3,154[1]	**No** N = 2,897[1]	**Yes** N = 257[1]	**p-value**[2]
Current smoker	1,502 (48%)	1,343 (46%)	159 (62%)	<0.001

[1]n (%)
[2]Pearson's Chi-squared test

```r
p1 <- wcgs |>
  ggplot2::ggplot() +
  ggplot2::aes(x = sbp) +
  ggplot2::geom_histogram(bins = 30, fill = "steelblue", color = "white") +
  ggplot2::labs(y = "Count", title = "Raw scale")

p2 <- wcgs |>
  ggplot2::ggplot() +
  ggplot2::aes(x = lnsbp) +
  ggplot2::geom_histogram(bins = 30, fill = "coral", color = "white") +
  ggplot2::labs(y = "Count", title = "Log scale")

gridExtra::grid.arrange(p1, p2, ncol = 2)
```

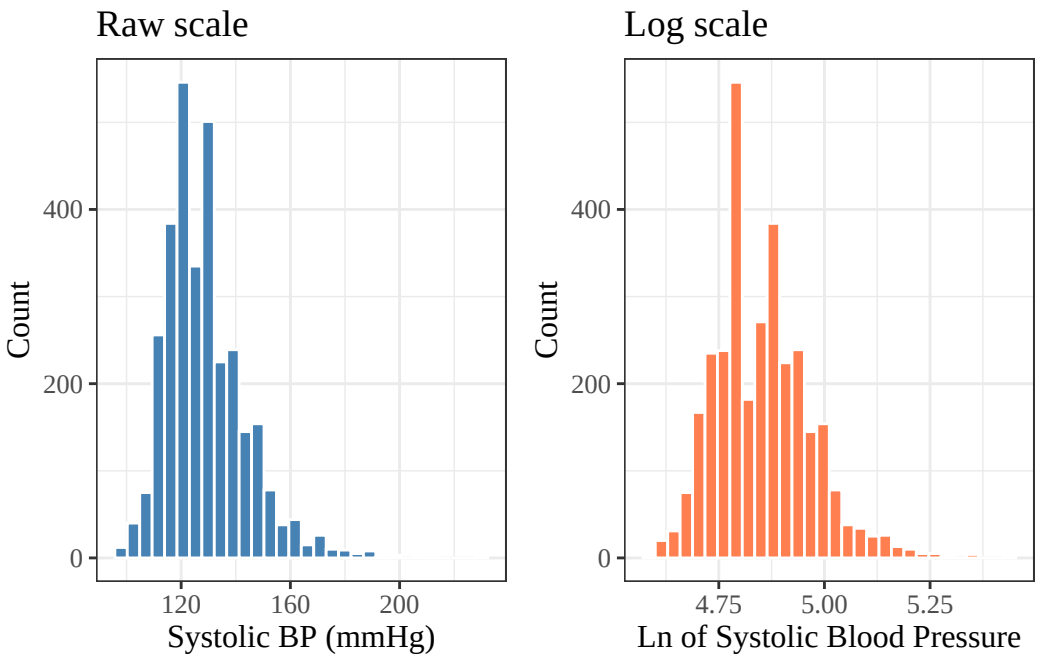

Figure 9: SBP distribution: raw and log scale (WCGS)

The log-transformed SBP is slightly more symmetric than the raw SBP. Whether to use the transformation in a regression model depends on assumptions of that model and the scientific question.

6 Putting it all together: an EDA workflow

A typical exploratory data analysis workflow might proceed as follows:

1. **Understand the dataset**: number of observations, variables, and their types.
2. **Examine each variable individually**:
   - Continuous: histogram, box plot, mean, SD, median, IQR.
   - Categorical: bar chart, frequency table.
   - Note missing values, unusual values, and outliers.
3. **Examine relationships between variables**:
   - Two continuous variables: scatter plot, correlation.

Table 7: Summary of selected WCGS variables

Characteristic	**N = 3,154**[1]
Age (years)	45.0 (42.0, 50.0)
Cholesterol (mg/dL)	223 (197, 253)
Unknown	12
Systolic BP (mmHg)	126 (120, 136)
Diastolic BP (mmHg)	80 (76, 86)
BMI (kg/m²)	24.39 (22.96, 25.84)
Weight (lbs)	170 (155, 182)
Cigarettes per day	0 (0, 20)
CHD event by 1969	257 (8.1%)
Current smoker	1,502 (48%)
Behavioral pattern (A/B)	
Type B	1,565 (50%)
Type A	1,589 (50%)
Behavioral pattern (A1/A2/B3/B4)	
A1	264 (8.4%)
A2	1,325 (42%)
B3	1,216 (39%)
B4	349 (11%)

[1]Median (Q1, Q3); n (%)

- Continuous by category: side-by-side box plots, group means.
- Two categorical variables: contingency table.

4. **Consider transformations** for skewed continuous variables.
5. **Summarize findings** to guide model-building decisions.

```
wcgs |>
  dplyr::select(
    age, chol, sbp, dbp, bmi, weight, ncigs,
    chd69, smoke, dibpat, behpat
  ) |>
  gtsummary::tbl_summary()
```

References

Rosenman, Ray H, Richard J Brand, C David Jenkins, Meyer Friedman, Reuben Straus, and Moses Wurm. 1975. "Coronary Heart Disease in the Western Collaborative Group Study: Final Follow-up Experience of 8 1/2 Years." *JAMA* 233 (8): 872–77. https://doi.org/10.1001/jama.1975.03260080034016.

Vittinghoff, Eric, David V Glidden, Stephen C Shiboski, and Charles E McCulloch. 2012. *Regression Methods in Biostatistics: Linear, Logistic, Survival, and Repeated Measures Models*. 2nd ed. Springer. https://doi.org/10.1007/978-1-4614-1353-0.