Medical Statistics – Lab 6

R version

Welcome to lab 6 on correlation and linear regression. In today's exercises, we will be analyzing a dataset named `pockets.sav`, which you can download from the Datasets menu. This dataset contains measurements of periodontal pocket depth for a group of individuals, along with several demographic and lifestyle variables.

Below is an overview of the variables we will be working with:

Variable	Description
`pocketdepth`	Pocket depth measurement in millimeters (continuous)
`sex`	Sex (categorical: "Female" / "Male")
`age`	Age in years (continuous)
`smoking`	Smoking status (categorical: "Non-smoker" / "Smoker")
`alcohol`	Alcohol consumption categories (e.g., "None", "1–2 glasses/day", ">2 glasses/day")

```r
library(haven)   # for reading SPSS files
library(dplyr)   # for data manipulation
library(ggplot2) # for data visualization
library(car)    # for calculating type-III ANOVA tables

# Load the dataset
pockets <- read_sav("pockets.sav")

# Convert labeled variables to factors
pockets <- pockets %>%
  mutate(across(where(is.labelled), as_factor))
```

Part 1: Pearson's correlation coefficient and simple linear regression

In this section, we will investigate whether **age** is associated with **pocket depth**. We start by creating a scatterplot to visualize the relationship between `pocketdepth` and `age`:

```
# Scatterplot of pocket depth vs. age
ggplot(pockets, aes(x = age, y = pocketdepth)) +
  geom_point() +
  labs(
    x     = "Age (years)",
    y     = "Pocket Depth (mm)",
    title = "Scatterplot of Pocket Depth vs Age"
  )
```

> Question 1
>
> Based on the scatterplot, is there an indication of a linear association between `age` and `pocketdepth`? If so, is this association positive or negative?

Pearson's Correlation Coefficient

To quantify the strength and direction of the linear relationship between `age` and `pocketdepth`, we can calculate Pearson's correlation coefficient:

```
# Calculate Pearson's correlation coefficient
cor(pockets$age, pockets$pocketdepth, use="complete.obs")
```

The argument `use = "complete.obs"` tells R to exclude any missing values when calculating the correlation coefficient. In this case, there are no missing values in the `age` and `pocketdepth` variables, meaning that the argument is not strictly necessary. However, it is good practice to include it to avoid potential issues with missing data in other datasets.

> Question 2
>
> What does the correlation coefficient tell us about the relationship between `age` and `pocketdepth`? Does this align with your interpretation of the scatterplot?

We can also test whether the correlation coefficient is significantly different from zero using a hypothesis test. The null hypothesis is that the correlation coefficient is zero (i.e., no linear relationship between the variables), and the alternative hypothesis is that the correlation coefficient is not zero. In R, we can perform this test using the `cor.test()` function:

```r
# Test the significance of the correlation coefficient
cor_test <- cor.test(pockets$age, pockets$pocketdepth)
cor_test
```

> Question 3
>
> What is the p-value for the correlation coefficient test? Based on this p-value, do we have sufficient evidence to reject the null hypothesis?

Fitting a Simple Linear Regression Model

Next, we fit a simple linear regression model to quantify the relationship between `age` and `pocketdepth`. In R, this can be achieved using the `lm()` function:

```r
# Fit a simple linear regression model
model_slr <- lm(pocketdepth ~ age, data = pockets)

# Print the summary of the model
summary(model_slr)
```

The formula `pocketdepth ~ age` specifies that `pocketdepth` is the dependent variable and `age` is the independent variable. The intercept term is included by default in the linear regression model and is therefore not explicitly specified in the formula.

> Question 4
>
> Is the relationship between `age` and `pocketdepth` statistically significant (at $= 0.05$)?

> Question 5
>
> How does the p-value for `age` in the regression output compare to the p-value for the correlation coefficient test? Are they consistent with each other?

> Question 6
>
> What is the interpretation of the intercept and the coefficient for `age` in the regression output?

Question 7

Based on the fitted model, what is the expected pocket depth for a person who is 40 years old?

Question 8

How much of the variation in pocket depth is explained by age in this model?

Assumption Checking

To obtain diagnostic plots for the simple linear regression model, we can supply the fitted model to the `plot()` function. By default, R will generate four diagnostic plots. In this lab, we are going to focus on the following two plots:

- Q-Q Plot: This plot helps us assess the normality of the residuals. This plot can be obtained by setting `which = 2` in the `plot()` function.
- Residuals vs. Fitted: This plot helps us check for homoscedasticity (constant variance) and linearity assumptions. This plot can be obtained by setting `which = 1` in the `plot()` function.

Normality of Residuals

```
plot(model_slr, which = 2)  # Q-Q plot
```

In addition to the Q-Q plot, we can also create a histogram of the residuals to visually inspect their distribution:

```
# Histogram of residuals
hist(residuals(model_slr), breaks = 15, col = "lightblue",
     border = "black", main = "Histogram of Residuals",
     xlab = "Residuals")
```

Question 9

Do the histogram and Q-Q plot suggest that the residuals are reasonably normally distributed?

Homoscedasticity and linearity

```
plot(model_slr, which = 1)  # residuals vs fitted
```

> Question 10
>
> Does the residual-versus-fitted plot suggest constant variance?

The red line in R's default residual-versus-fitted plot is a LOESS (locally estimated scatterplot smoothing) curve that shows the average trend in the residuals. It helps you see if there is a systematic pattern (e.g., curvature) that might indicate the linear model is misspecified. Ideally, you want that red line to be close to horizontal (i.e., around zero) with no strong curvature, suggesting that the linear fit is appropriate and there's no obvious nonlinearity or other systematic pattern left in the residuals.

> Question 11
>
> Does the residual-versus-fitted plot suggest any violation of the linearity assumption?

Part 2: ANCOVA (Analysis of Covariance)

In this section, we will fit an ANCOVA model to determine whether alcohol consumption is associated with pocket depth, controlling for age.

Exploratory Data Analysis

We start by creating a scatterplot to visualize the relationship between `age` and `pocketdepth`, using different colors to represent the levels of `alcohol`:

```
# Scatterplot of pocketdepth vs. age, colored by alcohol
ggplot(pockets, aes(x = age, y = pocketdepth, color = alcohol)) +
  geom_point() +
  labs(title = "Scatterplot of Pocket Depth vs. Age by Alcohol Consumption",
       x = "Age", y = "Pocket Depth") +
  theme_minimal() +
  theme(legend.position = "bottom")
```

> Question 12
>
> What can you infer from the scatterplot about the relationship between `age`, `pocketdepth`, and `alcohol` consumption?

Dummy Coding for `alcohol`

The variable `alcohol` has three categories (e.g., `"None"`, `"1-2 glasses/day"`, and `">2 glasses/day"`), meaning that we need to create two dummy variables to represent these categories in the model. R automatically generates these dummy variables once `alcohol` is recognized as a factor. To check if `alcohol` is a factor, we can use the `is.factor()` function:

```
# Check whether alcohol is a factor
is.factor(pockets$alcohol)
```

In this case, the output should be `TRUE`, given that we converted all categorical variables to factors when loading the dataset. Should `alcohol` not be a factor, we can either convert it to a factor using the `factor()` function or specify it as a factor in the model formula.

Fitting the ANCOVA Model

To fit the ANCOVA model, we can use the `lm()` function in R. The formula for the ANCOVA model is specified as `pocketdepth ~ age + alcohol`, where `age` is the continuous predictor and `alcohol` is the categorical predictor.

```
model_ancova <- lm(pocketdepth ~ age + alcohol, data = pockets)
summary(model_ancova)
```

> **Note**
>
> Note that R uses the first level of a factor (alphabetically or numerically) as the default reference category. In this dataset, "None" is the reference category.

Question 13

Based on the ANCOVA model output, what is the expected difference in pocket depth between individuals who consume `"None"` and those who consume `">2 glasses/day"`, while controlling for age?

Question 14

Based on the ANCOVA model output, what is the expected difference in pocket depth between individuals who consume `"1-2 glasses/day"` and those who consume `">2 glasses/day"`, while controlling for age?

To test the overall significance of the `alcohol` variable as a predictor of `pocketdepth`, we construct an analysis of variance (ANOVA) table. The ANOVA table summarizes how much each term in a linear regression model contributes to explaining the overall variation in the response variable. There are different ways to construct this table depending on how the sum of squares is partitioned among model terms. A common approach is Type III ANOVA, which evaluates each variable or interaction after all other terms have been accounted for. Each effect is tested as if it were entered last, so its sum of squares reflects the unique contribution of that variable or interaction beyond what is already explained by the remaining terms.

To obtain the type III ANOVA table, we use the `Anova()` function from the `car` package. This function provides a more detailed ANOVA output compared to the base R `anova()` function, and has an argument `type` that allows you to specify the type of sum of squares to use:

```
Anova(model_ancova, type="III")
```

> Question 15
>
> Based on the ANOVA table, is the `alcohol` variable significantly associated with `pocketdepth` after accounting for `age`?

Model Diagnostics

> Exercise
>
> Check the normality of residuals and homoscedasticity assumptions for the ANCOVA model.

Part 3: Interactions in ANCOVA

In some cases, the relationship between the outcome variable and a predictor may depend on the level of another predictor. This is known as an interaction effect. In the context of ANCOVA, we can test for interactions between the continuous predictor (`age`) and the categorical predictor (`alcohol`).

Fitting the Interaction Model

To include an interaction term in the model, we can use the `:` operator in the formula. That is, the interaction term between `age` and `alcohol` can be specified as `age:alcohol`. With this term, the interaction model can be specified as `pocketdepth ~ age + alcohol + age:alcohol`. A short form for specifying the interaction model is `pocketdepth ~ age`

`* alcohol`, which includes both the main effects of `age` and `alcohol` and their interaction terms.

```
model_interaction <- lm(pocketdepth ~ age * alcohol, data = pockets)
summary(model_interaction)
```

To test the significance of the interaction term, we again use the `Anova()` function from the `car` package:

```
Anova(model_interaction, type="III")
```

> Question 16
>
> Based on the output in the ANOVA table, is there a significant interaction between `age` and `alcohol` in predicting `pocketdepth`?

Part 4: Relationship Between Smoking and Pocket Depth

In addition to information about alcohol consumption, the dataset also contains information about smoking habits. Explore the relationship between smoking and pocket depth, and how it interacts with age. You can use the same approach as in the previous sections to fit models, test for significance, and check assumptions.