Medical Statistics – Lab 6

SPSS 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")

Part 1: Pearson's Correlation Coefficient and Simple Linear Regression

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

1. **Open** `pockets.sav` in SPSS.
2. Go to **Graphs** → **Legacy Dialogs** → **Scatter/Dot**.
3. Choose **Simple Scatter** and click **Define**.
4. Place `age` on the **X Axis** and `pocketdepth` on the **Y Axis**.
5. Click **OK** to generate the scatterplot.

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 will calculate Pearson's correlation coefficient:

1. Go to **Analyze** → **Correlate** → **Bivariate**.
2. Move `age` and `pocketdepth` into the **Variables** list.
3. Make sure **Pearson** is checked under **Correlation Coefficients**.
4. Click **OK**.

SPSS will output the correlation coefficient and a p-value testing whether the correlation is different from zero.

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?

You will also see a **Sig. (2-tailed)** value in the SPSS output for the Pearson correlation, which corresponds to the test that the correlation coefficient is significantly different from zero.

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 (that the correlation is zero)?

Fitting a Simple Linear Regression Model

Next, we fit a **simple linear regression** model to quantify the relationship between `age` and `pocketdepth`.

1. Go to **Analyze** → **Regression** → **Linear**.

2. Put `pocketdepth` in the **Dependent** box.

3. Put `age` in the **Independent(s)** box.

4. Click **OK** to run the analysis.

SPSS will produce output tables, including **Model Summary**, **ANOVA**, and **Coefficients**.

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

> 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?

> Question 6
>
> What is the interpretation of the intercept and the slope 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 assess assumptions (normality of residuals, homoscedasticity, etc.), we inspect **residual plots**.

1. In the **Linear Regression** dialog, click **Plots**.

2. Move **ZRESID** (standardized residuals) to the **Y:** box and **ZPRED** (standardized predicted values) to the **X:** box under **Scatter**.

3. Also check **Histogram** and **Normal probability plot**.

4. Click **Continue** → **OK**.

SPSS will generate:

- A **Histogram** of the residuals (for checking normality).
- A **Normal P-P plot** of the residuals (another way to check normality).
- A **scatterplot** of standardized residuals vs. standardized predicted values (for checking homoscedasticity and linearity).

Normality of Residuals

Inspect the **Histogram** and **Normal P-P plot** in the output.

> Question 9
>
> Do the histogram and Normal P-P plot suggest that the residuals are reasonably normally distributed?

Homoscedasticity and Linearity

Look at the standardized residuals versus standardized predicted values scatterplot.

> Question 10
>
> Does the plot suggest constant variance?

> Question 11
>
> Is there any strong curvature or systematic pattern that would indicate the model is misspecified (i.e., not truly linear)?

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`:

1. Go to **Graphs** → **Legacy Dialogs** → **Scatter/Dot**.
2. Choose **Simple Scatter** and click **Define**.
3. Place `age` on the **X Axis**, `pocketdepth` on the **Y Axis**, and `alcohol` in **Set Markers by**.
4. Click **OK** to generate the scatterplot.

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

Fitting the ANCOVA Model

In SPSS, there are multiple ways to fit linear regression models. If all your explanatory variables are continuous, you can use **Analyze** → **Regression** → **Linear**. However, if your model includes categorical predictors, it is often more convenient to use **Analyze** → **General Linear Model**, as SPSS will automatically handle the dummy coding for categorical variables in that procedure.

In this case, we have both continuous (`age`) and categorical (`alcohol`) explanatory variables, so we will use the **General Linear Model** procedure.

1. Go to **Analyze** → **General Linear Model** → **Univariate**.
2. Place `pocketdepth` in the **Dependent Variable** box.
3. Place `age` under **Covariate(s)**.
4. Place `alcohol` under **Fixed Factor(s)**.
5. Click **Options**, and select **Parameter estimates**.
6. Click **OK** to run the analysis.

> **Note**
>
> Note that SPSS uses the last category (highest numerical code) as the default reference category in the General Linear Model procedure. In this dataset, ">2 glasses/day" is the reference category.

Question 13

From the table with the estimated regression coefficients, what is the estimated difference in pocket depth between individuals who consume `"None"` and those who consume `">2 glasses/day"`, while controlling for `age`?

Question 14

From the same table, what is the estimated 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.

In SPSS, the type III ANOVA table is automatically generated when you fit a linear model using the **General Linear Model** procedure. This table is displayed under the heading `Tests of Between-Subjects Effects` in the output.

Question 15

Based on the ANOVA table, is there a significant association between `alcohol` consumption and `pocketdepth` after accounting for `age`?

Model Diagnostics

Similar to the simple regression case, you can request certain diagnostic plots in the **Options** menu of the **Univariate** dialog. However, the options are limited compared to the **linear regression** dialog. If you want more control over the diagnostic plots, you can save the residuals and predicted values to your dataset and then create the plots manually:

1. In the **Univariate** dialog, click **Save**.
2. Select **Unstandardized predicted values** and **Standardized residuals**.
3. Click **Continue** to go back to the main dialog, and Click **OK** to run the analysis.
4. After the analysis finishes, go to **Graphs** → **Legacy Dialogs** → **Scatter/Dot** (or **Histogram**) to plot the new residual and predicted-value columns, explore their relationship, or check for normality.

Exercise

Check the normality of residuals and homoscedasticity assumptions for the ANCOVA model. Do you see any notable violations?

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

1. Go to **Analyze** → **General Linear Model** → **Univariate**.
2. Place `pocketdepth` in the **Dependent Variable** box.
3. Place `age` under **Covariate(s)**.
4. Place `alcohol` under **Fixed Factor(s)**.
5. Click **Model**, and select **Build terms** under **Specify Model**.
6. Under **Build Terms**, set the type to **Main effects**.
7. In the **Factors & Covariates** box, select `alcohol` and `age` and move the two variables to the **Model** box.
8. Under **Build Terms**, set the type to **Interaction**.
9. In the **Factors & Covariates** box, select `alcohol` and `age` and move the two variables to the **Model** box.
10. Click **Continue** to go back to the main dialog.
11. Click **OK** to run the analysis.

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.