

 PHC 5050 Biostatistics for Public Health

Final Project, Spring 2024

**Overview**

This is a practical application of the knowledge and skills you have acquired over the last 13 weeks of the class. For this project, you will be working with a public health dataset. Using this file, you will be asked to conduct descriptive statistics, correlation analysis and hypothesis testing of continuous and categorical variables. You will also be asked to conduct simple linear and multilinear regression.

Please include all the SPSS outputs in one document. Make sure to include the link with your initials in all SPSS outputs.

I hope this will show the practicality of what you have learned in this class and that you may be able to apply it in the future. Make sure that you answer the questions in the boxes provided for them.

**The Data**

“The National Health and Nutrition Examination Survey (NHANES) is a program of studies designed to assess the health and nutritional status of adults and children in the United States. The survey is unique in that it combines interviews and physical examinations. NHANES is a major program of the National Center for Health Statistics (NCHS). NCHS is part of the Centers for Disease Control and Prevention (CDC) and has the responsibility for producing vital and health statistics for the Nation.”

 For this project, you will analyze a subset of real-world data from NHANES to assess some of the risk factors associated with cardiovascular disease (CVD):

| **Variable Name in your dataset** | **Description** | **Coding** | **NHANES dataset****(NHANES variable(s))** |
| --- | --- | --- | --- |
| PARTICIPANT ID | Random, unique number for each participant |  | all(SEQN) |
| AGE | Age at exam (years) |  | P\_demo(RIDAGEYR) |
| GENDER | Participant’s self-identified gender\* | 1 = male2 = female | P\_demo(RIAGENDR) |
| HEIGHT | Height (m) |  | P\_bmx(BMXHT) |
| WEIGHT | Weight (kg) |  | P\_bmx(BMXWT) |
| OBESITY | Is the participant obese? (BMI > 30.0) | 1 = yes2 = no | calculated- |
| BMI | Body mass index (kg/m^2) |  | P\_bmx(BMXBMI) |
| WAIST CIRCUM | Waist circumference (cm) |  | P\_bmx(BMXWAIST) |
| TOTAL CHOL | Total cholesterol (mg/dL) |  | P\_tchol(LBXTC) |
| HDL CHOL | High-density lipoprotein cholesterol (mg/dL) |  | P\_hdl(LBDHDD) |
| LDL CHOL | Low-density lipoprotein cholesterol (mg/dL) |  | P\_trigly(LBDLDLM) |
| CHOL MEDS | Does the participant take medication for high cholesterol? | 1 = yes2 = no | P\_bpq(BPQ100D) |
| TRIGLYCERIDES | Triglycerides (mg/dL) |  | P\_trigly(LBXTR) |
| INSULIN MEDS | Does the participant take insulin? | 1 = yes2 = no | P\_diq(DIQ050) |
| INSULIN | Fasting blood insulin (microU/mL) |  | P\_ins(LBXIN) |
| GLUCOSE | Fasting blood sugar (mg/dL) |  | P\_glu(LBXGLU) |
| DIABETES | Is the participant diabetic? | 1 = yes2 = no | P\_diq(DIQ010) |
| SYS BP | Systolic blood pressure (mmHg) |  | P\_bpxo(BPXOSY1) |
| DIA BP | Diastolic blood pressure (mmHg) |  | P\_bpxo(BPXODY1) |
| BP MEDS | Does the participant take medication for hypertension? | 1 = yes2 = no | P\_bpq(BPQ050A) |
| SMOKING | Does the participant smoke cigarettes? | 1 = yes2 = no | P\_smq(SMQ040) |
| CVD | Does the participant have a history of cardiovascular disease? | 1 = yes2 = no | P\_mcq(MCQ160b, MCQ160c, MCQ160d, MCQ160e) |

\*Though gender identities other than male/female exist, the GENDER variable is binary in this dataset for ease of analysis.

**Part I: Descriptive Statistics**

1. In the table below, record the number and percentage (to 1 decimal point) of males and females with a history of cardiovascular disease (CVD) in this dataset.

| History of Cardiovascular Disease (CVD) | Frequency (N) | Percentage (%) |
| --- | --- | --- |
| Male |  |  |
| Female |  |  |

1. Attach SPSS output:
2. Next, calculate descriptive statistics for total cholesterol (mg/dL). Enter the appropriate values in the table below.

| Total Cholesterol (mg/dL) |
| --- |
| Descriptive Statistic | Value |
| N |  |
| Mean |  |
| Median |  |
| Skewness |  |
| Kurtosis |  |
| Range |  |
| Minimum |  |
| Maximum |  |
| 25th Percentile |  |
| 75th Percentile |  |
| *Interquartile range (calculate)* |  |

1. Attach SPSS output:
2. What do you notice about the mean and median? What does this tell you about the distribution of total cholesterol in this dataset?
3. Create a histogram of the distribution of total cholesterol. Include a normal curve. SPSS output:
4. Next, calculate descriptive statistics for diastolic blood pressure (mmHg). Enter the appropriate values in the table below:

| Diastolic Blood Pressure (mmHg) |
| --- |
| Descriptive Statistic | Value |
| N |  |
| Mean |  |
| Median |  |
| Skewness |  |
| Kurtosis |  |
| Range |  |
| Minimum |  |
| Maximum |  |
| 25th Percentile |  |
| 75th Percentile |  |
| *Interquartile range (calculate)* |  |

1. Attach SPSS output:
2. What do you notice about the mean and median? What does this tell you about the distribution of diastolic blood pressure in this dataset?
3. Create a histogram of the distribution of diastolic blood pressure. Include a normal curve. SPSS output:

**Part II: Correlation analysis of continuous variables**

1. Test for normality of the total cholesterol and diastolic blood pressure variables.
	1. What are your null and alternative hypotheses?
	2. What is your conclusion (include p-value)?
	3. Attach SPSS output:
	4. If the distributions of total cholesterol and/or diastolic blood pressure are not normal, can we proceed with correlation analysis? Why or why not? (hint: look at sample size).
2. Test for correlation between total cholesterol and diastolic blood pressure.
3. What are your null and alternative hypotheses?
4. Create a scatterplot of Total Cholesterol (mg/dL) x Diastolic Blood Pressure. Include a total linear fit line. Attach SPSS output:

1. Perform the correlation analysis and fill in the table below.

|  | Pearson Correlation Coefficient (r) | p-value | 95% Confidence Interval |
| --- | --- | --- | --- |
| Lower | Upper |
| Total Cholesterol (mg/dL) x Diastolic Blood Pressure (mmHg) |  |  |  |  |

1. Attach SPSS output:
2. What is your conclusion (include r and p-value)? Interpret the 95% confidence interval. Interpret the R2 value.

**Part III: Hypothesis Testing (Continuous Data)**

1. We are interested in whether mean total cholesterol differs between males and females. Conduct a t-test and fill in the table below.

|  | Male | Female |
| --- | --- | --- |
| N |  |  |
| Mean total cholesterol (mg/dL) |  |  |
| Mean difference in total cholesterol between males and females |  |
| Levene’s Test for Equality of Variances p-value |  |
| Can we assume equality of variances? (yes/no) |  |
| t-value |  |
| p-value |  |

* 1. What are the two-sided null and alternative hypotheses?
	2. What is your conclusion (include p-value)?
	3. Attach SPSS output:
1. We are interested in whether mean diastolic blood pressure differs between males and females. Conduct a t-test and fill in the table below.

|  | Male | Female |
| --- | --- | --- |
| N |  |  |
| Mean diastolic blood pressure (mmHg) |  |  |
| Mean difference in total cholesterol between males and females |  |
| Levene’s Test for Equality of Variances p-value |  |
| Can we assume equality of variances if α=0.05? (yes/no) |  |
| t-value |  |
| p-value |  |

* 1. What are the two-sided null and alternative hypotheses?
	2. What is your conclusion (include p-value)?
	3. Attach SPSS output:

**Part IV: Hypothesis Testing (Categorical Data)**

1. Using a cross-tabulation table of the variable smoker (in layer box), CVD (in row), and gender (in column box), fill in the table below.

| Proportion of smokers with history of CVD |  |
| --- | --- |
| Proportion of smokers with history of CVD who are femaleProportion of smokers with history of CVD who are male |  |
| Proportion of females who smokeProportion of males who smoke |  |
| Proportion of female smokers with history of CVDProportion of male smokers with history of CVD |  |

Tip: Focus on determining the numerator, then the denominator.

* 1. Attach SPSS output:
	2. What do you notice about the patterns of CVD among male and female smokers?
1. We would like to assess whether the proportion of people with CVD is different between non-smokers and smokers. Develop a contingency table and calculate the Chi-square statistic comparing both variables. CDC (in column box) and the variable smoker (in row box).

|  |  | History of CVD |  |
| --- | --- | --- | --- |
|  |  | Yes | No | Total |
| Current Smoker | Yes |  |  |  |
| No |  |  |  |
|  | Total |  |  |  |

| 𝝌2 value |  |
| --- | --- |
| p-value |  |

* 1. What are your null and alternative hypotheses?
	2. What is your conclusion (include p-value)?
	3. Attach SPSS output:

**Part V: Linear Regression**

In Part II, we investigated the correlation between total cholesterol and diastolic blood pressure. Next, we will conduct a univariate analysis to test the hypothesis that total cholesterol is related to diastolic blood pressure.

Run a simple linear regression model with total cholesterol as the independent variable (X) and diastolic blood pressure as the dependent variable (Y).

* + 1. What is the F-value and its p-value? What does the F test tell us about the full model (slope and intercept) versus the reduced model (intercept only)?
	1. Attach SPSS output (ANOVA table):
	2. Fill out the table below:

| 𝜷0 (**beta-coefficient for the intercept)** |  |
| --- | --- |
| 𝜷1 (**beta coefficient for cholesterol)** |  |
| p-value, 𝜷1 |  |

* 1. What is your conclusion about the relationship between total cholesterol and diastolic blood pressure? Interpret the slope coefficient.

BONUS+1: There is limited clinical utility in describing this relationship in terms of just a one mg/dL change in total cholesterol. We want an interpretation that is more clinically relevant. What is the change in diastolic blood pressure that corresponds to a 50 mg/dL increase in total cholesterol? Round up to the nearest integer.

* 1. Attach SPSS output (Coefficients table):
	2. Write out the full simple linear regression equation for this model.
	3. Based on this model, what is the anticipated diastolic blood pressure for an individual with a total cholesterol of 200 mg/dL? Round up to the nearest integer.

**Part VI: Multiple Linear Regression**

In Part V, we determined that there is a statistically significant relationship between total cholesterol and diastolic blood pressure. However, we are unaware of the extent to which other variables may confound this relationship. In this section, we will create a multiple linear regression model to adjust for these potential confounders, including age, gender, and treatment for hypertension (high blood pressure). The first step is to evaluate the univariate association of each of these variables with diastolic blood pressure.

**AGE**

Run a simple linear regression model with age as the independent variable (X) and diastolic blood pressure as the dependent variable (Y). Fill in the table below.

| r |  |
| --- | --- |
| R2 |  |
| F |  |
| p-value, F |  |
| 𝜷0 (**beta-coefficient for the intercept)** |  |
| 𝜷1 (**beta coefficient for age)** |  |
| p-value, 𝜷1 |  |

* 1. What are your conclusions? Write in APA style. Interpret the correlation coefficient, F statistic, and 𝜷1.
	2. Attach SPSS output:

**GENDER**

Run a simple linear regression model with gender as the independent variable (X) and diastolic blood pressure as the dependent variable (Y). Fill in the table below.

 Gender (1 = male, 2 = female)

| r |  |
| --- | --- |
| R2 |  |
| F |  |
| p-value, F |  |
| 𝜷0 (**beta-coefficient for the intercept)** |  |
| 𝜷1 (**beta coefficient for age)** |  |
| p-value, 𝜷1 |  |

* 1. What are your conclusions? Write in APA style. Interpret the correlation coefficient, F statistic, and 𝜷1.
	2. Attach SPSS output:

**TREATMENT FOR HYPERTENSION**

Run a simple linear regression model with BPmeds as the independent variable (X) and diastolic blood pressure as the dependent variable (Y). Fill in the table below.

BPmeds, Is the participant taking blood pressure medication? (1 = yes, 2 = no)

| r |  |
| --- | --- |
| R2 |  |
| F |  |
| p-value, F |  |
| 𝜷0 (**beta-coefficient for the intercept)** |  |
| 𝜷1 (**beta coefficient for age)** |  |
| p-value, 𝜷1 |  |

1. What are your conclusions? Write in APA style. Interpret the correlation coefficient, F statistic, and 𝜷1.
2. Attach SPSS output:

**TOTAL CHOLESTEROL, AGE, GENDER, and TREATMENT FOR HYPERTENSION**

Run a multiple linear regression model with total cholesterol, age, gender, and BPmeds as the independent variables (Xs) and diastolic blood pressure as the dependent variable (Y). Fill in the table below.

| r |  |
| --- | --- |
| R2 |  |
| F |  |
| p-value, F |  |
| 𝜷0 (**beta-coefficient for the intercept)** |  |
| Total cholesterol, 𝜷1 (**beta coefficient for cholesterol)** |  |
| p-value, 𝜷1 |  |
| Age, 𝜷1 (**beta coefficient for age)** |  |
| p-value, 𝜷1 |  |
| Gender, 𝜷1 (**beta coefficient for gender)** |  |
| p-value, 𝜷1 |  |
| BPmeds, 𝜷1 (**beta coefficient for BPmeds)** |  |
| p-value, 𝜷1 |  |

1. Which independent variable became statistically ***in***significant (𝛂 = 0.05) upon inclusion in this multivariate model?
2. Attach SPSS output:

**2) Re-run the multiple linear regression model without this statistically insignificant variable.** Write the complete regression equation below.

1. Use your regression equation to estimate the diastolic blood pressure of a 70 year-old with a total cholesterol of 150 mg/dL who is not on treatment for hypertension.
2. Use your regression equation to estimate the diastolic blood pressure of a 48 year-old with a total cholesterol of 300 mg/dL who is on treatment for hypertension.

 **Congrats! You have successfully completed this assignment!!**