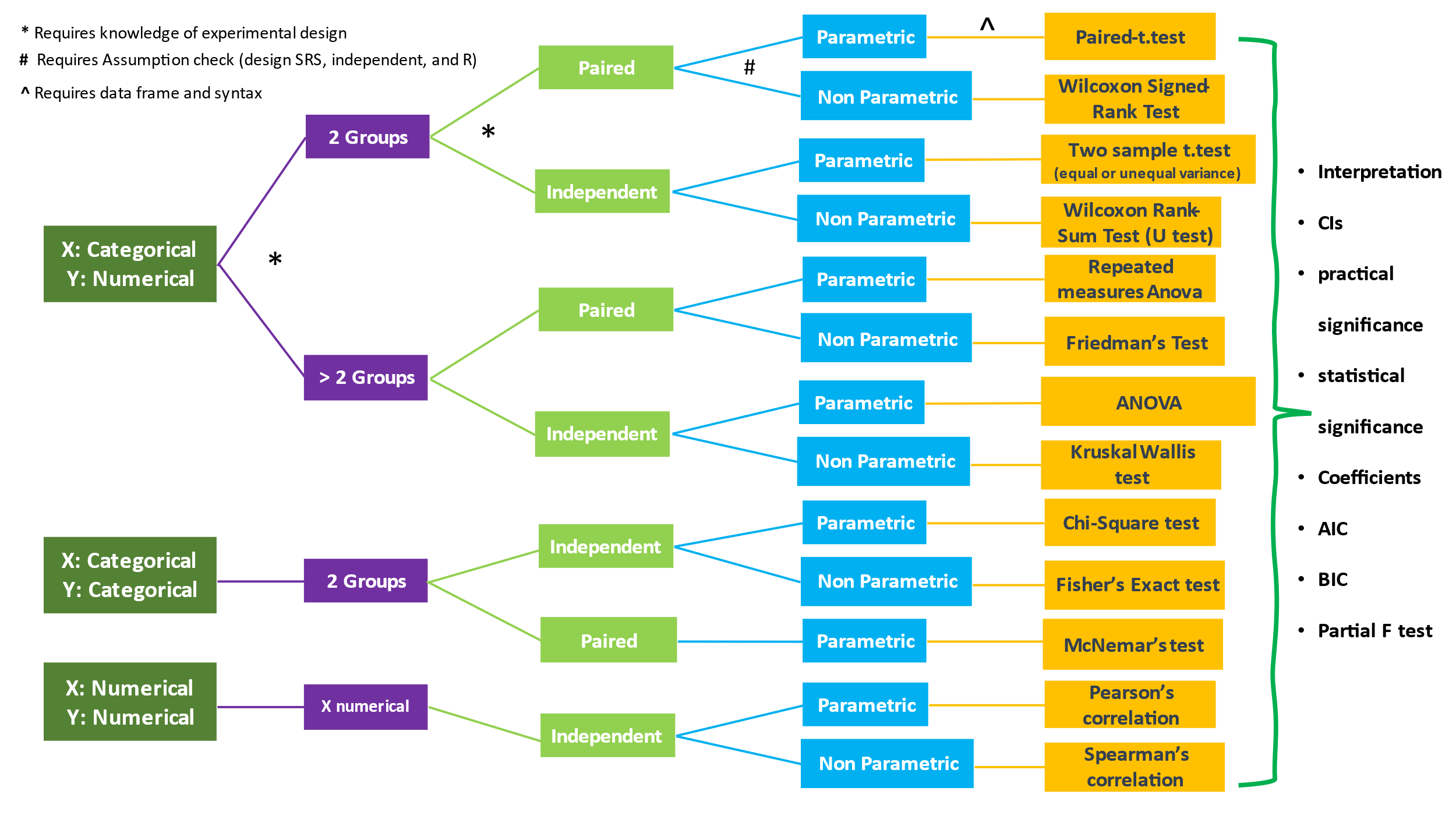
BMEN 7340 HW4

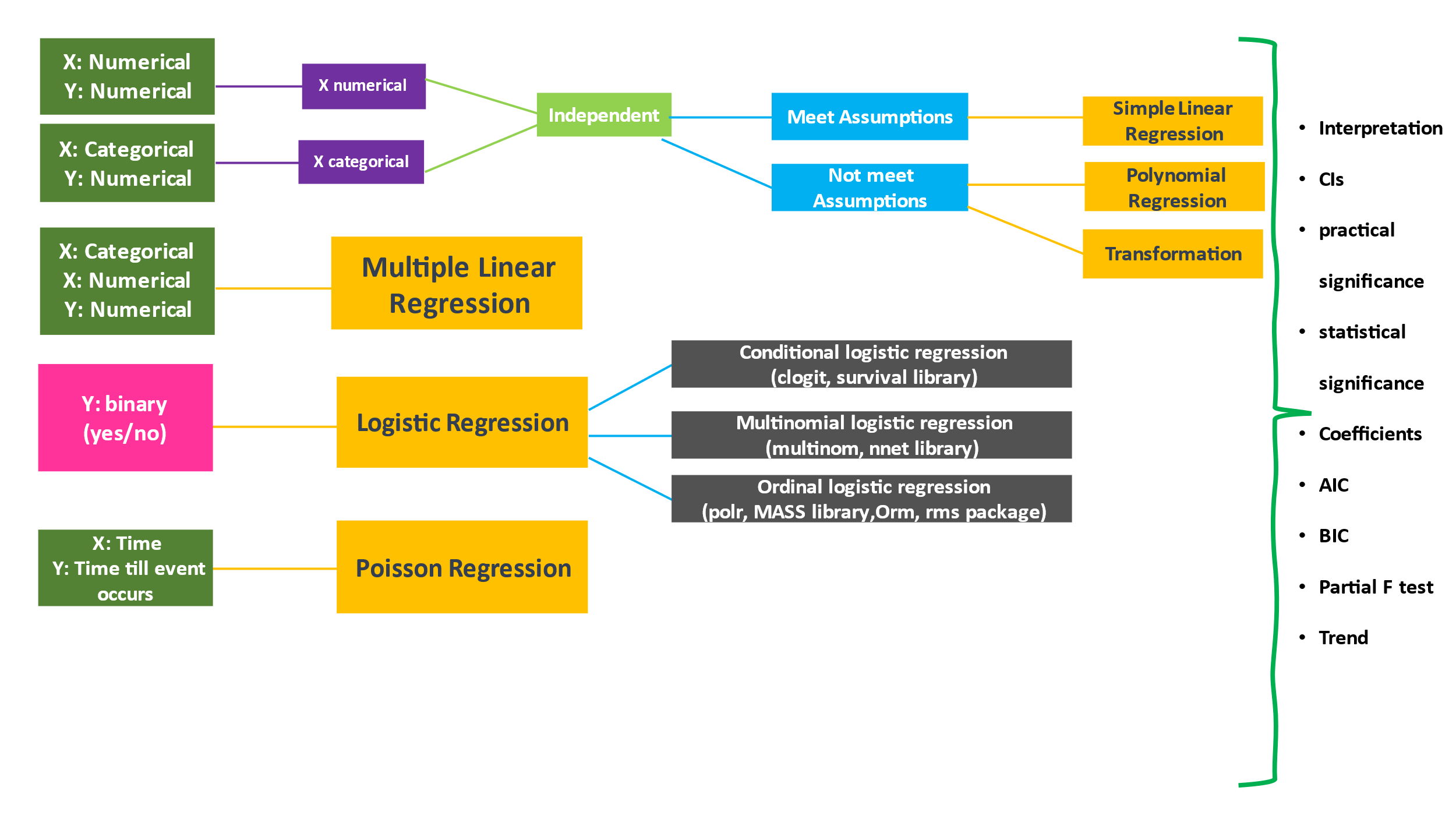
title: 'Homework #4\_FA2024\_BMEN7340'

Submission: please submit a knitted Word document

author: "First\_Last\_7340\_HW4\_FA24"

* Repeat measure ANOVA
* One-way ANOVA
* Two-way ANOVA
* Chi-square test
* Fisher’s exact test
* McNemar’s Test
* Single linear regression
* Polynomial regression
* Multiple linear regression
* Simple logistic regression
* Kaplan Meyers survival





#1:

The following data come from a study investigating drinking problems among college students. In 2010, a group of students was asked whether they had ever driven an automobile while drinking. In 2014, a different group of college students was asked the same question after congress raised the legal drinking age.

|  |  |  |
| --- | --- | --- |
|  | Year | |
| Drove while drinking | 2010 | 2014 |
| Yes | 1250 | 991 |
| No | 1387 | 1666 |

1. Evaluate the null hypothesis that the population proportions of students who drove while drinking are the same in the two calendar years (or two surveys).
2. What do you conclude about the behavior of college students?

#2:

In a study of intra-observer variability in the assessment of cervical smears. 3325 slides were screened for the presence or absence of abnormal squamous cells. Each slide was screened by a particular observer and then rescreened six months later by the same observer. The results are shown:

|  |  |  |
| --- | --- | --- |
|  | Second Screening | |
| First Screening | Present | Absent |
| Present | 1763 | 489 |
| Absent | 403 | 670 |

1. Do these data support the null hypothesis that there is no association between the time of screening and diagnosis (or observers’ judgment is consistent)?
2. The data could also be displayed in the following manner. Is there anything wrong with this presentation? Why?

|  |  |  |
| --- | --- | --- |
|  | Screening | |
| Abnormal cells | First | Second |
| Present | 2252 | 2166 |
| Absent | 1073 | 1159 |

#3.

Two different questionnaire formats were designed to measure alcohol consumption: one questionnaire encompassing all types of food in the diet (in the hope of minimizing bias) and another explicitly targeting alcohol use (straightforward, can be biased). Two formats were compared for independent men and women between 50-65 years of age living in a particular community. For each of the alcoholic beverages, beer, liquor, red wine, and white wine, each subject was classified as either a nondrinker (0, never or less than one drink per month) or a drinker (1, one or more drinks per month) according to each of the questionnaires. The data is in “alcohol.” Categories for the generic questionnaire are saved under the name “genques” and those for the target alcohol use are saved under “alcques.”

1. Test the null hypothesis that there is no association between drinking status on the two different types of questionnaires.
2. What do you conclude?

#4.

Thirty-five patients with ischemic heart disease participated in a series of tests designed to evaluate pain perception. In one part of the study, the patients exercised until they experienced angina or chest pain; the time until the onset of angina and the duration of the attack was recorded. The data is “ischemic.”

Time to angina in seconds is saved under the “time” variable, and the duration of angina in seconds is saved under the “duration” variable.

1. Create a two-way scatter plot for these data.
2. In the population of patients with ischemic heart disease, does there appear to be any evidence of a linear relationship between time to angina and the duration of the attack?
3. In calculating the correlation coefficient, which test will you use, Pearson or Spearmen? Why?
4. Test the null hypothesis that the population correlation is equal to 0. What do you conclude?

#5.

The dataset “lowbwt” records a sample of 100 low-birth-weight infants born in two hospitals in Boston, Massachusetts. The systolic blood pressure is saved under “sbp” variable. The values of Apgar score – an index of neonatal asphyxia or oxygen deprivation (recorded five minutes after birth) are saved under the variable name “apgar5”. The Apgar score is an ordinal random variable that takes values between 0 -10. Indicators of gender (1=male, 0=female) are saved under “sex.”

1. Use a non-parametric test to estimate the correlation between the systolic blood pressure and the Apgar score.
2. What is the Null hypothesis? What do you conclude?

#6.

Suppose that you are interested in determining whether a relationship exists between the fluoride content in a public water supply and the dental caries experience of children using this water. Data from a study examining 7257 children in 21 cities are in the file “water.” The fluoride content of the public water supply in each city, measured in parts per million, is under the “fluoride” variable, and the number of dental caries per 100 children examined is under the “caries” variable. The total dental caries experience is obtained by summing the numbers of filled teeth, teeth with untreated dental caries, teeth requiring extraction, and missing teeth.

1. Construct a scatter plot for these data.
2. What is the correlation between the number of dental caries per 100 children and the fluoride content of the water?
3. Is this correlation significantly different from 0? Why?

#7.

One of the functions of the Federation of State Medical Boards is to collect data summarizing disciplinary actions taken against nonfederal physicians by medical licensing boards. Serious actions include license revocations, suspensions, and probations. For each of the years 1991 through 1995, the number of severe actions per 1000 doctors was ranked by state from highest to lowest. The ranks are in the data “actions.” The ranks for 1991 are saved under the variable name “rank91”, those for 1992 under “rank92”, and so on.

1. Calculate the correlations of the ranks in 1991 and those in 1992, those in 1991 and 1993, those in 1991 and 1994, and those in 1991 and 1995. What happens to the magnitude of the correlation as the years being compared get further apart?
2. Is each of these four correlations significantly different from 0?
3. Do you believe that all states are equally strict in taking disciplinary action against physicians? Why?

#8.

The dataset “lowbwt” records a sample of 100 low-birth-weight infants born in two hospitals in Boston, Massachusetts. The systolic blood pressure is saved under “sbp” variable. The values of Apgar score – an index of neonatal asphyxia or oxygen deprivation (recorded five minutes after birth) are saved under the variable name “apgar5”. The Apgar score is an ordinal random variable that takes values between 0 -10. Indicators of gender (1=male, 0=female) are saved under “sex.” The value of gestational age is under the name “gestage.”

1. Using systolic blood pressure as the response variable and the gestational age as the explanatory variable, compute the least-squares regression line. Interpret the estimated slope (b1).
2. At the 0.05 level of significance, test the null hypothesis that the true population slope is 0. What do you conclude?
3. What is the estimated mean systolic blood pressure for the population of low-birth-weight infants whose gestational age is 31 weeks?

#9.

In the 11 years before the passage of the Federal Coal Mine Health and Safety Act of 1969, the fatality rates of underground miners varied little. After the implementation of that act, however, fatality rates decreased steadily until 1979. The fatality rates for the years 1970 through 1981 are provided below; for computational purposes, calendar years have been converted to a scale beginning at 1. This information is contained in the data “miner.” Values of the response variable, fatality rate, are saved under the name “rate”, and the values of the explanatory variable, calendar year, under the name “year”.

|  |  |  |
| --- | --- | --- |
| Calendar Year | Year | Fatality Rate |
| 1970 | 1 | 2.419 |
| 1971 | 2 | 1.732 |
| 1972 | 3 | 1.361 |
| 1973 | 4 | 1.108 |
| 1974 | 5 | 0.996 |
| 1975 | 6 | 0.952 |
| 1976 | 7 | 0.904 |
| 1977 | 8 | 0.792 |
| 1978 | 9 | 0.701 |
| 1979 | 10 | 0.890 |
| 1980 | 11 | 0.799 |
| 1981 | 12 | 1.084 |

1. Construct a scatter plot of fatality rate versus year (x). What does this plot suggest about the relationship between these two variables?
2. To model the trend in fatality rates, fit the least-squares regression line.
3. Now transform the explanatory variable to ln(x). Fit the second least-squares model.
4. Transform x to 1/x. Fit the third least-squares model.
5. Which of the three models appears to fit the data best? Why?
6. Use the best model to estimate the fatality rate in year 9.

#10.

Statistics that summarize personal healthcare expenditures by the state for the years 1966 through 1982 have been examined to understand issues related to rising healthcare costs. Suppose you are interested in focusing on the relationship between expense per admission into a hospital and the average length of stay in the facility. The data “hospital” contains information for each state in the United States for the year 1982. The measures of mean expense per admission are saved under the variable name “expadm.” The corresponding average lengths of stay are saved under the variable “los.”

1. Generate numerical summary statistics for the expadm, and los. What are the means and medians of each variable?
2. Construct histogram plots for the expadm and los.
3. Using expadm as the response variable, and the los as the explanatory variable, compute the least-square regression line.
4. Interpret the estimated slope in words.
5. Construct four plots to check the model assumptions. In what ways does the residual plot help you to evaluate the fit of the model to the observed data?

Dataset: lowbwt, ischemic, water, actions, insure, hospital

#11.

The dataset “lowbwt” records a sample of 100 low-birth-weight infants born in two hospitals in Boston, Massachusetts. The systolic blood pressure is saved under “sbp” variable. The values of Apgar score – an index of neonatal asphyxia or oxygen deprivation (recorded five minutes after birth) are saved under the variable name “apgar5”. The Apgar score is an ordinal random variable that takes values between 0 -10 (although it is ordinal data, it is often treated as if it were continuous). Indicators of gender (1=male, 0=female) are saved under “sex.” The value of gestational age is under the name “gestage.”

1. Construct a model using gestational age and apgar score to predict the systolic pressure.
2. Test the null hypothesis that b1=0. (type I error is 0.05)
3. Interpret b1 and b2 with the context of this case.
4. What is the estimated mean systolic blood pressure for the population of low birth weight infants whose gestational age is 31weeks and whose five-minute apgar score is 7?
5. Explain the R2 within the context of this case.
6. Does this model pass the assumption check?

#12.

The dataset “lowbwt” records a sample of 100 low-birth-weight infants born in two hospitals in Boston, Massachusetts. The systolic blood pressure is saved under “sbp” variable. The values of Apgar score – an index of neonatal asphyxia or oxygen deprivation (recorded five minutes after birth) are saved under the variable name “apgar5”. The Apgar score is an ordinal random variable that takes values between 0 -10 (although it is ordinal data, it is often treated as if it were continuous). Indicators of gender is a dichotomous variable (1=male, 0=female) are saved under “sex.” The value of gestational age is under the name “gestage.”

1. Construct a model using gestational age, apgar score, and sex to predict the systolic pressure (without interactions). Given two infants with identical gestational age and apgar score, one male and the other female, which would tend to have the higher systolic blood pressure? By how much on average?
2. Add to the model a fourth explanatory variable, that is, the interaction between gestational age and sex. Based on the R output, what is your decision on whether you will keep the interaction in the model and why?

#13.

The Bayley Scales of Infant Development yield scores on two indices -PDI and MDI can be used to assess a child’s level of functioning at one-year-old. As part of a study investigating the development of the neurological status of children who had undergone reparative heart surgery during the first three months of life. The Bayley Scales were administered to a sample of one-year-old infants born with congenital heart disease. The children had been randomized to one of two different treatment groups, known as “circulatory arrest” and “low-flow bypass.” The groups differed in the specific way in which the reparative surgery was performed. Unlike circulatory arrest, low-flow bypass maintains continuous circulation through the brain; although it is considered preferable by some physicians, it also has its own associated risk of brain injury. The data for this study are saved in the seat set “heart.” PDI scores are saved under the variable name “pdi,” MDI score under “mdi,” and indicators of treatment group under “trtment.” (0=circulatory arrest; 1=low-flow bypass).

PDI: Psychomotor Development Index

MDI: Mental Development Index

1. Fit two simple linear regression models, one with PDI score as the dependent variable and the other with MDI score as the dependent variable, with the treatment group indicator as the independent variable in both models.
2. Who is more likely to have a higher PDI score, a child assigned to the circulatory arrest treatment group or one assigned to the low-flow bypass group? How much higher would the score be (on average)?
3. Is the treatment group difference in either PDI or MDI scores statistically significant at the 0.05 level? What do you conclude?

#14.

The relationship between expense per admission into a hospital and the average length of stay in the hospital was collected for each state. The data is “hospital”. Mean expense per admission is under variable name “expadm”, and average length of stay under “los”. The mean average salary is in “salary”.

1. Construct a scatter plot of expadm and salary. What does the graph suggest about the relationship between these two variables?
2. Fit the linear regression model between dependent variable expense per admission and average length of stay and average salary are the independent variables. Interpret the estimated regression coefficient(s) in the context of the case.
3. Does the inclusion of salary in addition to length of stay improve the model’s prediction ability? Explain.

#15.

A study was conducted to examine the roles of firearms and various other factors in the rate of homicides in the city of Detroit. Information for the years 1961 to 1973 is in the data “detroit”; the number of homicides per 100,000 population is saved under the variable name “homicide”, other variables in the data set include “police”, the number of full-time police officers per 100,000 population; the percentage of adults who are unemployed is under “unemp”, the number of handgun registrations per 100,000 population is under “register”, and the average weekly earnings for city residents is under “weekly”.

1. Fit a linear regression model using homicide rate as the dependent variable, number of handgun registrations as the primary independent variable, and using the other three variables as the rest of the independent variables. (Ignore interactions)
2. Which of the variables have a significant effect on homicide rate?
3. Find the best model.
4. Explain the coefficients (b1, b2, b3, b4 if any).
5. Explain the R2.

#16.

The dataset “lowbwt” records a sample of 100 low-birth-weight infants born in two hospitals in Boston, Massachusetts. The systolic blood pressure is saved under “sbp” variable. The values of Apgar score – an index of neonatal asphyxia or oxygen deprivation (recorded five minutes after birth) are saved under the variable name “apgar5”. The Apgar score is an ordinal random variable that takes values between 0 -10 (although it is ordinal data, it is often treated as if it were a continuous variable “grmhem” is a dichotomous random variable indicating whether an infant experienced a germinal matrix hemorrhage. The value 1 indicates that a hemorrhage occurred and 0 that is did not. The indicators of toxemia are stored in the variable name “tox”, where 1 represents a diagnosis of toxemia during pregnancy for the child’s mother and 0 is no such diagnosis). Indicators of gender (1=male, 0=female) are saved under “sex.” The value of gestational age is under the name “gestage.”

1. Using germinal matrix hemorrhage as the dependent variable, fit a logistic regression model, using the five-minute apgar score as the primary independent variable and the toxemia status as the second independent variable.
2. If a child has a five-minute apgar score of 3, and the child’s mother was diagnosed with toxemia during pregnancy, what is the predicted probability of experiencing a geminal matrix hemorrhage?
3. At the 0.05 level of significance, test the null hypothesis that the b1=0. What do you conclude?

#17.

To investigate the association between smoking and the development of aortic stenosis, a narrowing or stricture of the aorta that impedes blood flow to the body, for males and females, the data in “stenosis.” Smoking status is saved under the variable “smoke” (0=non-smoker, 1=smoker), and the presence of aortic stenosis is under the variable “diseases (0= no disease, 1=disease).

1. Using the presence of aortic stenosis as the response (dependent variable), fit a logistic regression model with smoking status as the primary explanatory variable and gender as the second explanatory variable to estimate the probability of the presence of aortic stenosis. What are the estimated odds of aortic stenosis for smokers versus nonsmokers, adjusting for gender? Hint: odds = eb1
2. Do you believe that the relationship between the presence of aortic stenosis and smoking status differs for males and females? Explain.

#18

In a group of patients undergoing dialysis for chronic renal failure for a period of at least two years, it was determined which of the individuals had experienced at least one episode of peritonitis, an inflammation of the membrane lining of the abdominal cavity, and which had not. The results are contained in a data set called “dialysis.” The variable “perito” is a dichotomous random variable taking the value 1 if an individual experienced an infection and 0 otherwise. Potential explanatory are age, gender, and racial background. The variable age is continuous; sex and race are dichotomous and take the value 1 for female and non-white patients, respectively. Male and white individuals are represented by 0.

1. Fit three separate logistic regression models investigating the effects of age, gender, and racial group on the probability that an individual experiences peritontitis. Interpret the estimated coefficients of each explanatory variables.
2. Fit a logistic model to predict the probability that a white male patient (age = 50) will experience peritonitis?
3. Do you see any problems with the way in which the response variable is categorized?

#19.

In the 1980s, a study was conducted to examine the effects of the drug ganciclovir on AIDs patients suffering from disseminated cytomegalovirus infection. Two groups of patients were followed. 18 were treated with the drug, and 11 were not. The results of this study are in the dataset “cyto”. Survival times in months after diagnosis are saved under the variable “time”, and indicators of censoring status – where 0 designates that observation was censored and 1 that a death occurred – under the name “censor”. Values of treatment group, where 1 indicates that a patient took the drug, and 2 that the patient did not, are saved under “group”.

1. Construct survival curves for the two treatment groups based on the S(t).
2. Does it appear that the individuals in one group survive longer than those in the other group?
3. Use the log-rank test to evaluate the null hypothesis that the distributions of survival times are identical in the two groups. What do you conclude?

#20.

In a study of bladder cancer, tumors were removed from the bladders of 86 patients. Subsequently, the individuals were assigned to be treated either with a placebo or with the drug thiopeta. Data is in “bladder”. Time to the first recurrence of tumor in months is saved under the variable “time”. Treatment status is saved under the name “group”. 1=placebo. Indicators of censoring status: 0 = tumor did not recur, 1 = tumor did recur; in variable “censor”, 0 = censored.

1. Construct survival curves.
2. Does it appear that the individuals in one group have a longer time to fist recurrence of tumor than those in the other group? Why?
3. Test the null hypothesis that the distributions of recurrence times are identical in the two treatment groups. What do you conclude?

Dataset: *stenosis, detroit, dialysis, cyto, bladder*