Test 3

STAT218

2024-03-16

library(tidyverse)  
library(Sleuth3)  
library(oibiostat)  
library(openintro)  
library(pander)  
library(epitools)

### Instructions

You have 48 hours from the release of this assignment to complete and submit your work. You may refer to all class materials, notes, and textbooks, but **must complete this assignment on your own**. By submitting your work, you are affirming that your work is your own and you have not consulted with anyone else in preparing your answers. Please use the word document provided (download from the class website) and write in your answers below each prompt. Please ensure that question numbering is preserved in your document. Submit your work via Gradescope.

The test comprises two parts: concepts, containing short-answer and multiple-choice questions; and applications, which requires some data analysis.

## Part I: Concepts

1. [L10] Suppose you use a simple linear regression model to predict summer high temperatures on city blocks from percent tree cover. For a block with 6% tree cover, you obtain two interval estimates for the predicted temperature. Which one is the prediction interval, and which one is the confidence interval, and why?
   1. (89.7, 92.2)
   2. (78.5, 100.6)

* ANSWER:
  1. CI
  2. PI

1. [L10] Continuing the example above, suppose your estimate for the slope parameter is -0.7. Select the correct interpretation of this parameter.
   1. Every 1% increase in tree cover is associated with an estimated 0.7-degree decrease in mean summer high temperature.
   2. Every doubling of percent tree cover is associated with an estimated 0.7-degree decrease in mean summer high temperature.
   3. Every 1% increase in tree cover is associated with an estimated 0.7-degree increase in mean summer high temperature.
   4. Every 1% increase in tree cover is associated with a 0.7-degree decrease in mean summer high temperature.

* ANSWER: (a)

1. [L8] Suppose you are interested in studying whether dairy consumption is linked to a rare disease, so you sample 40 individuals with the disease and 50 individuals without the disease, and determine whether each individual has a high or low level of dairy consumption. Which measure(s) of association are appropriate for analyzing the association between dairy consumption and disease risk?
   1. Difference in probability of disease
   2. Relative risk of disease
   3. Odds ratio
   4. All of the above

* ANSWER: (c)

1. [L6] A study asked 1,924 male and 3,666 female undergraduate college students their favorite color. A 95% confidence interval for the difference between the proportions of males and females whose favorite color is black () was calculated to be (0.02, 0.06). Select the proper interpretation of this interval.
   1. With 95% confidence, it is estimated that between 2% and 6% more men prefer the color black compared with women.
   2. With 95% confidence, the percentage of men whose favorite color is black is estimated to be between 2 and 6 percentage points higher than the corresponding percentage of women.
   3. With 95% confidence, the probability that a man’s favorite color is black is estimated to be between 2% and 6% higher than the probability that a woman’s favorite color is black.
   4. With 95% confidence, men are between 2% and 6% more likely to prefer black than women.

* ANSWER: (b)

1. [L6] It is believed that large doses of acetaminophen (the active ingredient in over the counter pain relievers like Tylenol) may cause damage to the liver. A researcher conducts a study to estimate the proportion of acetaminophen users who have liver damage, and estimates that the proportion is between 0.05 and 0.15 with 95% confidence. What was the sample size?
   1. 54
   2. 144
   3. 20
   4. 87

* ANSWER: (b)

1. [L7] Three treatments were compared to test their relative efficacy (effectiveness) in treating Type 2 Diabetes in patients aged 10-17 who were being treated with metformin. The primary outcome was lack of glycemic control (or not); lacking glycemic control means the patient still needed insulin, which is not the preferred outcome for a patient. The outcome was considered a ‘success’ if glycemic control was achieved. The chi-square test of association produced a -value of 0.01687. Interpret this result in context.

* ANSWER: The data provide sufficiently strong evidence to reject the null hypothesis of no association between achiving between glycemic control and the treatments in the experiment among 10-17 year-olds.

1. [L7] Continuing the previous problem, the residuals from the chi-square test are shown below. Identify which treatment(s) performed *better* than would be expected had there been no effect of treatment.

* ANSWER: the ‘rosi’ treatment performs better than expected.

test.out <- diabetes2 |> table() |> chisq.test()  
test.out$residuals |> pander(caption = 'Chi-square test residuals')

Chi-square test residuals

|  | failure | success |
| --- | --- | --- |
| **lifestyle** | 0.2139 | -0.196 |
| **met** | 1.373 | -1.258 |
| **rosi** | -1.584 | 1.451 |

## Part II: Applications

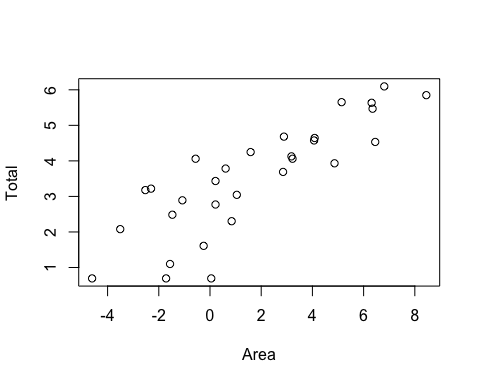
1. [L10] The dataset Sleuth3::ex1220 contains species counts for 30 islands in the Galapagos, along with information about the island geography. Estimate the relationship between island area and the number of observed species. You should include a summary of your fitted model, diagnostic plots, and a visualization of the model. Provide a 95% confidence interval for the model parameter that captures the relationship and interpret the interval in context.

* EITHER IS ACCEPTABLE:
  1. With 95% confidence, a 1-unit increase in log island area is associated with an estimated increase in mean log total number of species between 0.0416 and 0.1223.
  2. With 95% confidence, every doubling of island area is associated with an estimated 2.92 – 8.85% increase in median total number of species.

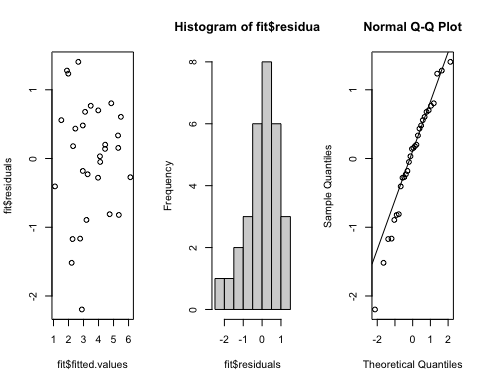
head(ex1220)

Island Total Native Area Elev DistNear DistSc AreaNear  
1 Baltra 58 23 25.09 332 0.6 0.6 1.84  
2 Bartolome 31 21 1.24 109 0.6 26.3 572.33  
3 Caldwell 3 3 0.21 114 2.8 58.7 0.78  
4 Champion 25 9 0.10 46 1.9 47.4 0.18  
5 Coamano 2 1 1.05 130 1.9 1.9 903.82  
6 Daphne Major 18 11 0.34 119 8.0 8.0 1.84

ex1220 |> select(Area, Total) |> log() |> plot()



fit <- lm(log(Total) ~ log(Area), data = ex1220)  
  
par(mfrow = c(1, 3))  
plot(fit$fitted.values, fit$residuals)  
hist(fit$residuals)  
qqnorm(fit$residuals)  
qqline(fit$residuals)



rbind(log.scale = confint(fit)[2, ],  
 back.transformed = 2^confint(fit)[2, ])

2.5 % 97.5 %  
log.scale 0.2879988 0.4815616  
back.transformed 1.2209455 1.3962541

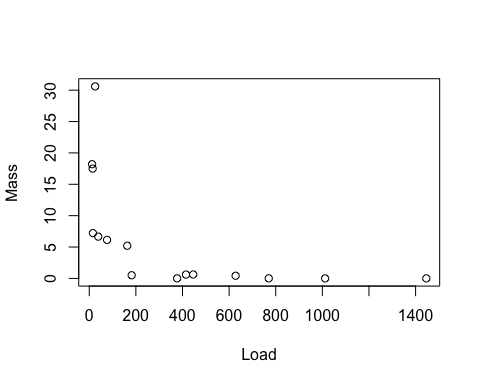
1. [L10] The dataset Sleuth3::ex0817 contains measurements from a study on the effectiveness of biological control of the weed tansy ragwort using flea beetles. Each observation is made on a plot that was planted with high density of ragwort; flea beetles were introduced at different levels on each plot and the average dry mass after harvest in the following season was measured. The objective is to determine whether dry mass decreases for plots with a higher flea beetle load.
   1. Construct a scatterplot of the data. Is the relationship linear?
   2. Construct a scatterplot of the log of the dry mass against flea beetle load. Is this relationship linear?
   3. Fit an appropriate SLR model to estimate the relationship between beetle load and dry mass of ragwort weed among the experimental plots.
   4. Show and comment on the diagnostic plots for the model you fit.
   5. Interpret the significance test on the slope parameter.
   6. Compute and interpret a 95% confidence interval for the slope parameter.

* ANSWERS:
  1. There is a negative association but the relationship is not linear.
  2. The relationship between load and log dry mass is roughly linear.
  3. Students should show the fit summary indicated below, or report the parameter estimates (including ).
  4. The fitted-residual plot indicates no trend or uneven spread; the residual histogram and Q-Q plot indicate some small deviations from normality, but overall residuals appear symmetric and unimodal. Overall, the plots show no obvious violations of model assumptions.
  5. Either: (i) each additional beetle per gram of dry mass is associated with an estiamted decrease in mean log dry mass of 0.006539; (ii) each additional beetle per gram of dry mass is associated with an estimated 0.652% decrease in median dry mass.
  6. Either: (i) With 95% confidence, each additional beetle per gram of dry mass is associated with an estiamted decrease in mean log dry mass between 0.00432 and 0.00878; (ii) With 95% confidence, each additional beetle per gram of dry mass is associated with an estimated decrease in median dry mass between 0.431% and 0.872%.

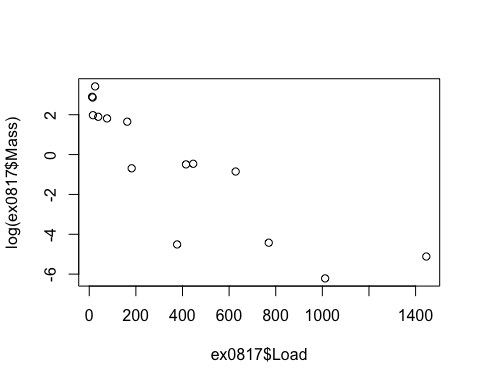
head(ex0817)

Load Mass  
1 12.2 18.20  
2 14.6 17.50  
3 15.8 7.22  
4 25.3 30.60  
5 38.6 6.66  
6 76.4 6.14

# a  
plot(ex0817)



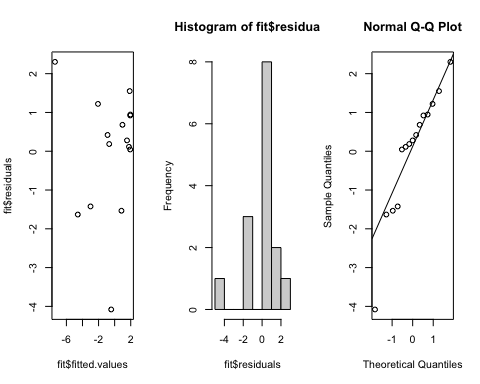
# b   
plot(ex0817$Load, log(ex0817$Mass))



# c  
fit <- lm(log(Mass) ~ Load, data = ex0817)  
summary(fit)

Call:  
lm(formula = log(Mass) ~ Load, data = ex0817)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-4.0798 -0.6889 0.2793 0.9343 2.3045   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 2.035158 0.574492 3.543 0.00361 \*\*   
Load -0.006539 0.001028 -6.361 2.49e-05 \*\*\*  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 1.65 on 13 degrees of freedom  
Multiple R-squared: 0.7568, Adjusted R-squared: 0.7381   
F-statistic: 40.46 on 1 and 13 DF, p-value: 2.494e-05

# d  
par(mfrow = c(1, 3))  
plot(fit$fitted.values, fit$residuals)  
hist(fit$residuals, breaks = 5)  
qqnorm(fit$residuals)  
qqline(fit$residuals)



# f  
confint(fit)[2,]

2.5 % 97.5 %   
-0.008760191 -0.004318129

1. [L6] Suppose that among a random sample of 144 U.S. adults 65 and older, 16 have diagnosed Alzheimer’s disease. Calculate and interpret a 95% confidence interval for the prevalence of Alzheimer’s based on the sample.

* ANSWER: With 95% confidence, the prevalence of Alzheimer’s among U.S. adults 65 and older is estimated to be between 5.87% and 16.35%.

prop.alzheimer <- 16/144  
n <- 144  
  
# solution  
prop.alzheimer.se <- sqrt(prop.alzheimer\*(1 - prop.alzheimer)/n)  
prop.alzheimer + c(-1, 1)\*2\*prop.alzheimer.se

[1] 0.05873283 0.16348939

1. [L7, L8] The dataset openintro::migraine contains observations from an experiment evaluating the effectiveness of acupuncture, relative to sham acupuncture, in treating migraines.
   1. Test for a treatment effect using an appropriate method. (*Hint*: if there’s no effect, we’d observe no association between treatment groups and pain outcomes.)
   2. Estimate the effecct of the acupuncture treatment relative to the control group. Choose an appropriate measure of association and provide a point estimate and 95% confidence interval.

* ANSWERS:
  1. The data provide sufficient evidence to reject the null hypothesis that acupuncture has no pain relief effect relative to sham acupuncture (Fisher’s exact test, *p = 0.01212*).
  2. With 95% confidence, the probability of pain relief from acupuncture is estimated to be between 1.24 and 23.04 times greater than the probability of pain relief from sham acupuncture, with a point estiamte of 5.349.

head(migraine)

# A tibble: 6 × 2  
 group pain\_free  
 <fct> <fct>   
1 treatment yes   
2 treatment yes   
3 treatment yes   
4 treatment yes   
5 treatment yes   
6 treatment yes

# a-b  
table(migraine) |> riskratio()

$data  
 pain\_free  
group no yes Total  
 control 44 2 46  
 treatment 33 10 43  
 Total 77 12 89  
  
$measure  
 risk ratio with 95% C.I.  
group estimate lower upper  
 control 1.000000 NA NA  
 treatment 5.348837 1.242012 23.03526  
  
$p.value  
 two-sided  
group midp.exact fisher.exact chi.square  
 control NA NA NA  
 treatment 0.01071837 0.01211965 0.009057886  
  
$correction  
[1] FALSE  
  
attr(,"method")  
[1] "Unconditional MLE & normal approximation (Wald) CI"

1. [L8] An observational study of 227,571 Medicare beneficiaries who initiated treatment with one of two diabetes medications recorded whether each patient reported cardiovascular problems. Determine which medication is linked to a higher probability of cardiovascular problems, and provide a point and interval estimate for an appropriate measure of comparison.

* ANSWER: The probability of cardiovascular problems is estimated to be 0.0337 among patients on Pioglitazone, and 0.0383 among patients on Rosiglitazone; it appears that Rosiglitazone is associated with a higher rate of cardiovascular problems. With 95% confidence, the risk of cardiovascular problems is estimated to be between 8.82% and 19.31% higher for patients on Rosiglitazone compared with Pioglitazone.

diabetes <- avandia |> table()  
  
diabetes |> prop.table(margin = 1)

cardiovascular\_problems  
treatment no yes  
 Pioglitazone 0.96633287 0.03366713  
 Rosiglitazone 0.96163804 0.03836196

riskratio(diabetes)$measure

risk ratio with 95% C.I.  
treatment estimate lower upper  
 Pioglitazone 1.000000 NA NA  
 Rosiglitazone 1.139449 1.088246 1.19306

1. [L7] The dataset openintro::census contains a random subsample of 500 respondents in the 2000 U.S. Census. Construct a contingency table of marital status by sex, and test for an association. Interpret the results of your test.

* ANSWER: The data do not provide sufficient evidence to reject the null hypothesis of no association between sex and marital status ( = 8.1315 on 5 degrees of freedom, *p = 0.1491*).

head(census)

# A tibble: 6 × 8  
 census\_year state\_fips\_code total\_family\_income age sex race\_general   
 <int> <fct> <int> <int> <fct> <fct>   
1 2000 Florida 14550 44 Male Two major races  
2 2000 Florida 22800 20 Female White   
3 2000 Florida 0 20 Male Black   
4 2000 Florida 23000 6 Female White   
5 2000 Florida 48000 55 Male White   
6 2000 Florida 74000 43 Female White   
# ℹ 2 more variables: marital\_status <fct>, total\_personal\_income <int>

census |> select(marital\_status, sex) |> table()

sex  
marital\_status Female Male  
 Divorced 21 17  
 Married/spouse absent 5 9  
 Married/spouse present 92 100  
 Never married/single 93 129  
 Separated 1 2  
 Widowed 20 11

census |> select(marital\_status, sex) |> table() |> chisq.test()

Pearson's Chi-squared test  
  
data: table(select(census, marital\_status, sex))  
X-squared = 8.1315, df = 5, p-value = 0.1491

1. [L6] A random sample of 1629 U.S. adults obtained from NHANES data was used to investigate whether asthma is more common in women than in men. Estimate the difference in asthma prevalence between women and men; provide both a point estimate and an interval estimate.

* ANSWER: With 95% confidence, the prevalence of asthma among women is estimated to be between 0.04 percentage points lower and 4.35 percentage points higher than the prevalence of asthma among men, with a point estimate of 2.15 percentage points in difference (women - men).

asthma <- matrix(data = c(49, 781, 30, 769),   
 nrow = 2,   
 byrow = T,   
 dimnames = list(sex = c('female', 'male'),  
 asthma = c('yes', 'no'))  
 )  
  
asthma |> prop.table(margin = 1)

asthma  
sex yes no  
 female 0.05903614 0.9409639  
 male 0.03754693 0.9624531

asthma |> prop.test()

2-sample test for equality of proportions with continuity correction  
  
data: asthma  
X-squared = 3.6217, df = 1, p-value = 0.05703  
alternative hypothesis: two.sided  
95 percent confidence interval:  
 -0.0004958005 0.0434742223  
sample estimates:  
 prop 1 prop 2   
0.05903614 0.03754693

prop.test(asthma)$estimate |> diff()

prop 2   
-0.02148921

1. [L10] The dataset openintro::babies\_crawl contains observations of average crawling ages for babies born in each month of the year, along with the average outdoor temperature at six months of age. Fit a simple linear regression model to the data and respond to the following prompts.
   1. How much variation in average crawling age is explained by average outdoor temperature at 6 months old?
   2. Estimate the association between average temperature at 6 months old and average crawling age.
   3. Predict the average crawling age for babies when the average outdoor temperature at 6 months is 70 degrees.

* ANSWERS:
  1. 48.96% of variability in average crawling age is explained by average outdoor temperature at 6 months old.
  2. With each one-degree increase in average outdoor temperature at six months, average crawling age is estimated to decrease by 0.07774 months (approximately 2.33 days).
  3. For an average outdoor temperature of 70 degrees at six months of age, the average crawling age is predicted to be 30.236 months.

head(babies\_crawl)

# A tibble: 6 × 5  
 birth\_month avg\_crawling\_age sd n temperature  
 <fct> <dbl> <dbl> <int> <int>  
1 January 29.8 7.08 32 66  
2 February 30.5 6.96 36 73  
3 March 29.7 8.33 23 72  
4 April 31.8 6.21 26 63  
5 May 28.6 8.07 27 52  
6 June 31.4 8.1 29 39

# a-b  
fit <- lm(avg\_crawling\_age ~ temperature, data = babies\_crawl)  
summary(fit)

Call:  
lm(formula = avg\_crawling\_age ~ temperature, data = babies\_crawl)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-3.0556 -0.5712 0.5221 0.8029 1.4334   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 35.67806 1.31753 27.080 1.09e-10 \*\*\*  
temperature -0.07774 0.02510 -3.097 0.0113 \*   
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 1.319 on 10 degrees of freedom  
Multiple R-squared: 0.4896, Adjusted R-squared: 0.4386   
F-statistic: 9.592 on 1 and 10 DF, p-value: 0.01131

# c  
predict(fit, newdata = data.frame(temperature = 70))

1   
30.23632

1. [L8] Suppose a retrospective study is done in which data are collected on men ages 50-54 who died over a 1-month period. Of 185 men who died from CVD, 15 had a diet with high salt intake before they died, while of the 196 men who died from other causes, 6 had a diet with high salt intake. Test for an association between salt intake and CVD mortality, and use an appropriate measure of association to estimate the relationship; provide both point and interval estimates.

* ANSWER: The data provide sufficient evidence to reject the null hypothesis of no association between salt intake and cardiovascular disease mortality among men aged 50-54 ( = 4.6545 on 1 degree of freedom, *p = 0.031*). With 95% confidence, the odds of dying from cardiovascular disease are estimated to be between 1.078 and 7.977 times greater among men aged 50-54 with high dietary salt intake compared with men of the same age with low dietary salt intake.

salt <- matrix(data = c(15, 170, 6, 190),   
 nrow = 2, byrow = T,  
 dimnames = list(cause = c('cvd', 'other'),  
 salt = c('high', 'low')))  
  
oddsratio(salt, rev = 'both')

$data  
 salt  
cause low high Total  
 other 190 6 196  
 cvd 170 15 185  
 Total 360 21 381  
  
$measure  
 odds ratio with 95% C.I.  
cause estimate lower upper  
 other 1.000000 NA NA  
 cvd 2.745102 1.078439 7.976611  
  
$p.value  
 two-sided  
cause midp.exact fisher.exact chi.square  
 other NA NA NA  
 cvd 0.03365918 0.04177721 0.03097138  
  
$correction  
[1] FALSE  
  
attr(,"method")  
[1] "median-unbiased estimate & mid-p exact CI"