Homework 2 - Stat 495

Cassandra Jin

Due Monday, Sept. 25th by midnight

Practicing Academic Integrity

If you worked with others or used resources outside of provided course material (anything besides our textbook(s), course materials in Moodle, R help menu) to complete this assignment, please acknowledge them below using a bulleted list.

I acknowledge the following individuals with whom I worked on this assignment:

Name(s) and corresponding problem(s)

-

I used the following sources to help complete this assignment:

Source(s) and corresponding problem(s)

-

For this assignment, you may do some of the problems on paper if you like, then scan in your solutions and merge with the Integrity page as a cover sheet and any work you are leaving in the .Rmd. If you decide to type in your answers, you will need to use LaTeX to show your work for some problems.

Because you may merge files into a pdf to submit in any order, be sure to assign pages in Gradescope so I can find work for each problem!

PROBLEMS TO TURN IN: CASI 2.1, Add 1, Add 2, Add 3, CASI 5.3 (modified), Add 4

CASI 2.1

A coin with probability of heads θ is independently flipped n times, after which θ is estimated by

$$\hat{\theta} = \frac{s+1}{n+2};$$

with s equal to the number of heads observed. $\hat{\theta}$ will be referred to as the estimator below.

(a) What are the bias and variance of the estimator?

SOLUTION:

$Bias(\hat{\theta}) = E(\hat{\theta}) - \theta$

$= E(\frac{s+1}{n+2}) - \theta$

$= \frac{E(s)+1}{n+2} - \theta$

We have that s follows $Bin(n, \theta)$, so $E(s) = n\theta$.

$= \frac{n\theta+1}{n+2} - \theta$

$= \frac{n\theta+1-n\theta-2\theta}{n+2}$

$= \frac{1-2\theta}{n+2}$

$Var(\hat{\theta}) = Var(\frac{s+1}{n+2})$

$= Var(\frac{s}{n+2} + \frac{1}{n+2})$

$= \frac{Var(s)}{(n+2)^2}$

Since s follows $Bin(n, \theta)$ again, $Var(s) = n\theta(1-\theta)$.

The bias and variance of the estimator are $\frac{1-2\theta}{n+2}$ and $\frac{n\theta(1-\theta)}{(n+2)^2}$, respectively.

(b) How would you apply the plug-in principle to get a practical estimate of the standard error of the estimator?

SOLUTION:

Since $Var(\hat{\theta}) = \frac{n\theta(1-\theta)}{(n+2)^2}$, $SD(\hat{\theta}) = \sqrt{\frac{n\theta(1-\theta)}{(n+2)^2}}$. Then by the plug-in principle, an estimate of the standard error of the estimator is $SE(\hat{\theta}) = \sqrt{\frac{n\hat{\theta}(1-\hat{\theta})}{(n+2)^2}}$.

Additional 1

Suppose you have a random sample of n observations drawn from a Bernoulli distribution with parameter θ. Further suppose that θ is unknown, but has a prior density of a Beta(α, β) distribution, with both α and β greater than 0.

part a: Find the posterior density for theta given the data. Be sure to fully specify/identify the posterior density in your solution.

SOLUTION:

(a) We have $X_1, \dots, X_n \overset{iid}{\sim} \text{Bern}(\theta)$ and prior density $g(\theta) \sim \text{Beta}(\alpha, \beta)$.

Then $f(x_i \mid \theta) = \theta^{x_i}(1-\theta)^{1-x_i}$

$$\prod_{i=1}^n f(x_i \mid \theta) = \prod_{i=1}^n \theta^{x_i}(1-\theta)^{1-x_i}$$

$$= \theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i}$$

and $g(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$

$$g(\theta \mid x) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1} \cdot \theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i}$$

$$\propto \theta^{\alpha-1+\sum_{i=1}^n x_i}(1-\theta)^{\beta-1+n-\sum_{i=1}^n x_i}$$

$$\propto \theta^{(\alpha+\sum_{i=1}^n x_i)-1}(1-\theta)^{(\beta+n-\sum_{i=1}^n x_i)-1}$$

$$\propto \text{Beta}\left(\alpha+\sum_{i=1}^n x_i,\ \beta+n-\sum_{i=1}^n x_i\right)$$

So $g(\theta \mid x)$ follows $\text{Beta}\left(\alpha+\sum_{i=1}^n x_i,\ \beta+n-\sum_{i=1}^n x_i\right)$.

part b: Find the Bayesian estimator for theta (i.e. the posterior mean).

SOLUTION:

(b) For $\text{Beta}(a,b)$, $E_{post}(\theta \mid X) = \frac{a}{a+b} = \frac{\alpha+\sum_{i=1}^n x_i}{\alpha+\sum_{i=1}^n x_i+\beta+n-\sum_{i=1}^n x_i}$

So the Bayesian estimator is $\frac{\alpha+\sum_{i=1}^n x_i}{\alpha+\beta+n}$.

Additional 2

Suppose $X_1 \dots X_n$ are iid from an Exponential distribution with parameter β, with pdf:

$$f(x|\beta) = \frac{1}{\beta} exp^{-x/\beta}, x > 0, \beta > 0,$$

and 0, otherwise.

(a) Find the MLE for beta.

SOLUTION:

(a) $L(\beta) = \prod_{i=1}^{n} f(x|\beta)$

$= \left(\frac{1}{\beta} e^{-x_1/\beta}\right)\left(\frac{1}{\beta} e^{-x_2/\beta}\right) \cdots \left(\frac{1}{\beta} e^{-x_n/\beta}\right)$

$= \frac{1}{\beta^n} e^{\frac{1}{\beta}\sum_{i=1}^{n} x_i}$

$\ell(\beta) = \log(L(\beta)) = \log\left(\frac{1}{\beta^n} e^{-\frac{1}{\beta}\sum_{i=1}^{n} x_i}\right)$

$= \log(\beta^{-n}) - \frac{1}{\beta}\sum_{i=1}^{n} x_i$

$= -n \log \beta - \sum_{i=1}^{n} x_i / \beta$

$\frac{d}{d\beta}\ell(\beta) = \frac{d}{d\beta}\left(-n \log \beta - \sum_{i=1}^{n} x_i/\beta\right)$

$= -\frac{n}{\beta} + \sum_{i=1}^{n} x_i / \beta^2$

$-\frac{n}{\hat{\beta}} + \sum_{i=1}^{n} x_i / \hat{\beta}^2 = 0$

$n\hat{\beta} = \sum_{i=1}^{n} x_i$

$\hat{\beta}_{mle} = \sum_{i=1}^{n} x_i / n = \bar{x}$

(b) Verify the MLE is unbiased.

SOLUTION:

(b) $\text{Bias} = E(\hat{\beta}_{mle}) - \beta = E(\sum_{i=1}^{n} x_i/n) - \beta$

$= \frac{1}{n}\sum_{i=1}^{n} E(x_i) - \beta$

$= \frac{1}{n} \cdot n\beta - \beta$

$= \beta - \beta = 0$

So $\hat{\beta}_{mle}$ is unbiased.

(c) Find the Fisher information for a single observation.

SOLUTION:

(c) $l(x|\beta) = \log(\frac{1}{\beta} e^{-x/\beta})$

$= -\log\beta - \frac{x}{\beta}$

$l'(x|\beta) = -\frac{1}{\beta} + \frac{x}{\beta^2}$

$l''(x|\beta) = \frac{1}{\beta^2} - \frac{2x}{\beta^3}$

$-E(l''(x|\beta)) = -E(\frac{1}{\beta^2} - \frac{2x}{\beta^3})$

$= -\frac{1}{\beta^2} + \frac{2}{\beta^3} E(x)$

$= -\frac{1}{\beta^2} + \frac{2}{\beta^2} \cdot \beta$

$= \frac{1}{\beta^2} = I(\beta)$ is the Fisher information for a single observation.

(d) Find the Cramer Rao lower bound on variance for unbiased estimators of beta.

SOLUTION:

(d) We have $I_n(\beta) = nI(\beta) = \frac{n}{\beta^2}$, so the CRLB on variance for unbiased estimators of β is $\frac{1}{I_n(\beta)} = \frac{\beta^2}{n}$.

Additional 3

In a few sentences and in your own words, explain what the Neyman-Pearson lemma tells us and why it is important/useful in the context of hypothesis testing.

SOLUTION:

The Neyman-Pearson lemma guarantees the existence of the uniformly most powerful test, i.e. the likelihood-ratio test, for simple null hypothesis testing. If the lemma's conditions are satisfied, then the LRT has the greatest power $1 - \beta$ among all possible tests of a given size α when determining the H_0 against H_1. More specifically, suppose we have two tests, δ with significance level α and δ' with significance level α' where $\alpha(\delta) \leq \alpha(\delta')$, then we would find $\beta(\delta) \geq \beta(\delta')$.

CASI 5.3 (modified)

Draw a sample of 1000 bivariate normal vectors $x = (x_1, x_2)'$, with each variable having a mean of 0, a standard deviation of 1, and with a correlation between them of 0.5. Be sure your process is reproducible.

Review the chapter 4 and 5 practice problems for assistance with code.

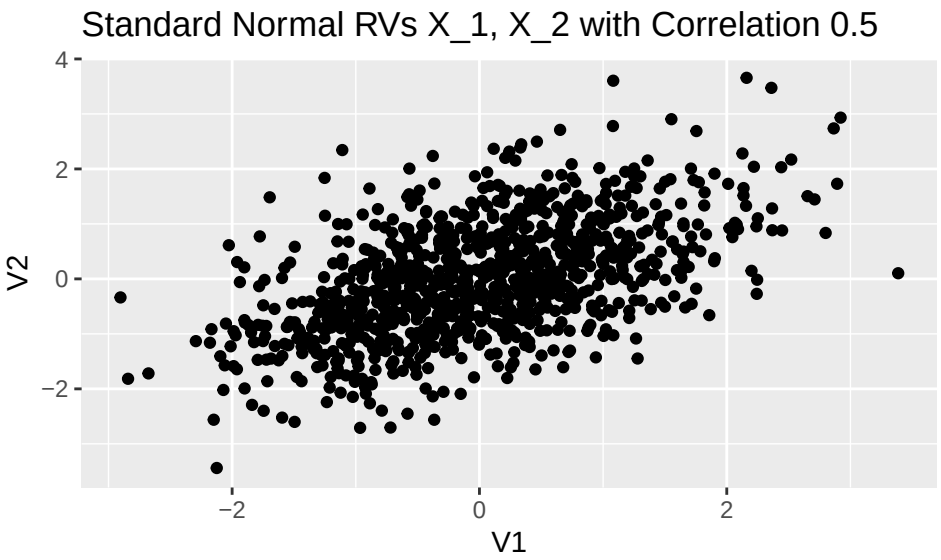
(a) Plot your sample of data points.

SOLUTION:

```
set.seed(495)

# generate data from bivariate normal
bivariate_normals <- as.data.frame(rmvnorm(1000, mean = c(0, 0), sigma = matrix(c(1, 0.5, 0.5, 1), ncol

gf_point(V2 ~ V1, data = bivariate_normals) %>%
gf_labs(title = "Standard Normal RVs X_1, X_2 with Correlation 0.5")
```



(b) Following equation 5.19, what should the theoretical distribution of $x_2|x_1$ be here?

SOLUTION:

By Equation 5.19, the theoretical distribution should be $x_2|x_1 \sim N(0 + \frac{0.5}{1}(x_1 - 0), 1 - \frac{0.5^2}{1})$, so $x_2|x_1 \sim N(0.5x_1, 0.75)$.

(c) Regress x_2 on x_1 and numerically check equation 5.19.

Hint: Reading the text will help you understand what two things to check in the regression output.

SOLUTION:

```
mod <- lm(V2 ~ V1, data = bivariate_normals)
msummary(mod)
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.00323    0.02747   -0.12     0.91
## V1           0.54744    0.02800   19.55   <2e-16 ***
##
## Residual standard error: 0.869 on 998 degrees of freedom
## Multiple R-squared:  0.277,  Adjusted R-squared:  0.276
## F-statistic:  382 on 1 and 998 DF,  p-value: <2e-16
```

Theoretically, we expect the slope to be 0.5 and the R^2 value to be 0.25, and the values from the regression output, slope 0.54744 and R-squared 0.277, generally reflect these. From part (b), we have mean of $0.5x_1$, which matches the outputted slope coefficient, and since R^2 captures the variance captured by the model, it also makes sense that the R^2 value we found is approximately $1 - Var(x_2|x_1) = 1 - 0.75$.

Additional 4

Many parametric inference procedures rely on certain conditions being met in order for the procedures to be valid. In cases where the conditions are not met, nonparametric procedures can be employed. You have seen the bootstrap and permutation/randomization tests as examples of alternatives (much more exists in the field of nonparametric statistics). On Homework 1, you performed some analysis with the bootstrap, and thought through how to perform a permutation test on the gene136 data, which you later saw in the practice problems.

How are these procedures adapted to other situations?

(a) In a few sentences, describe how you would perform a permutation-based test to assess the overall significance of a multiple linear regression. If context would help, you can consider the following regression model:

```
data(penguins)
mymod <- lm(bill_length_mm ~ bill_depth_mm*species, data = penguins)
msummary(mymod)
```

```
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      23.068      3.017    7.65  2.2e-13 ***
## bill_depth_mm                     0.857      0.164    5.22  3.1e-07 ***
## speciesChinstrap                 -9.640      5.715   -1.69  0.09259 .
## speciesGentoo                    -5.839      4.535   -1.29  0.19885
## bill_depth_mm:speciesChinstrap    1.065      0.310    3.44  0.00067 ***
## bill_depth_mm:speciesGentoo       1.164      0.279    4.17  3.8e-05 ***
##
## Residual standard error: 2.44 on 336 degrees of freedom
##   (2 observations deleted due to missingness)
## Multiple R-squared:  0.802,  Adjusted R-squared:  0.799
## F-statistic:  273 on 5 and 336 DF,  p-value: <2e-16
```

Note that there aren't necessarily problems with the parametric methods in this setting, but context can be helpful for thinking through the process. You don't need to implement the procedure, but feel free if it helps your description.

SOLUTION:

To perform a permutation-based test to assess the overall significance of a multiple linear regression, I would first fit the multiple linear regression model. I would then create a vector of the residuals from the model, randomize the residual points, and reassign them to the fitted values. Following this, I would run the regression again to obtain the test statistic for the new model values. Looping through the randomization and fitting many regressions, I would have a distribution of test statistics to observe, and I would then assess how unusual the original model statistic is in the empirical null sampling distribution.

(b) Describe two ways that a bootstrap might be useful in a multiple linear regression model. Be sure that one way is tied to a specific model term, while the other is associated with the entire model.

Again, if context is useful, feel free to refer to the model fit above. There are many possible solutions here. As above, you don't need to implement the two ways, but feel free if it helps your description.

SOLUTION:

A bootstrap might be useful in a multiple linear regression model for estimating the regression coefficient for a specific variable in the model. The method could also be performed on the entire model to obtain a re-sampled and re-fitted equation.