Homework 2 - Stat 495

Example Solution

Due Monday, Sept. 25th by midnight

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

Most on separate paper.

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, such as your work for CASI 5.3. 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, ~~CASI 3.1~~, Add 1, ~~CASI 4.4 a~~, 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

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

$s \sim Bin(n, \theta)$

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:

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

SOLUTION:

a. $E(\hat{\theta}) = E\left(\frac{s+1}{n+2}\right) = \frac{1}{n+2}\left[E(s)+1\right] = \frac{1}{n+2}(\theta n+1) = \frac{n\theta+1}{n+2}$

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

$V(\hat{\theta}) = V\left(\frac{s+1}{n+2}\right) = \frac{1}{(n+2)^2}V(s) = \frac{n\theta(1-\theta)}{(n+2)^2}$

b. $SD(\hat{\theta}) = \sqrt{\frac{n\theta(1-\theta)}{(n+2)^2}}$ The plug-in principle lets us sub $\hat{\theta}$ in for θ here to get:

$SE(\hat{\theta}) = \sqrt{\frac{n\hat{\theta}(1-\hat{\theta})}{(n+2)^2}}$

We often plug in the MLE, but it can be any estimator.

$X_1, \ldots, X_n$ RS Bern(θ) $g(\theta) \sim$ Beta(α, β)

Add 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:

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

SOLUTION:

a. $f(x|\theta) = \theta^{x}(1-\theta)^{1-x} \Rightarrow f_n(\underset{\sim}{x}|\theta) = \prod \theta^{x_i}(1-\theta)^{1-x_i}$

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

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

$$g(\theta|\underset{\sim}{x}) \propto g(\theta)\cdot f_n(\underset{\sim}{x}|\theta) = \theta^{\alpha-1}(1-\theta)^{\beta-1}\theta^{\sum x_i}(1-\theta)^{n-\sum x_i}$$

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

This is a Beta($\alpha + \sum x_i$, $\beta + n - \sum x_i$).

α + # successes β + # failures

b. The Bayesian estimator for θ is the posterior mean. For a Beta(α,β), that is $\alpha/(\alpha+\beta)$. Thus, here it is!

$$\tilde{\theta} = \frac{\alpha+\sum x_i}{\alpha+\sum x_i+\beta+n-\sum x_i} = \frac{\alpha+\sum x_i}{\alpha+\beta+n}$$

Add 2

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

$E(X) = \beta$

$V(X) = \beta^2$

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

and 0, otherwise.

(a) Find the MLE for beta.

SOLUTION:

(b) Verify the MLE is unbiased.

SOLUTION:

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

SOLUTION:

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

SOLUTION:

a. $f(\underset{\sim}{x}|\beta) = \frac{1}{\beta^n} e^{-\sum X_i/\beta}$ $\quad \ell(\underset{\sim}{x}|\beta) = -n \log \beta - \frac{\sum X_i}{\beta}$

$= -n \log \beta - (\sum X_i)(\beta^{-1})$

$\ell'(\underset{\sim}{x}|\beta) = \frac{-n}{\beta} + \frac{\sum X_i}{\beta^2} = 0 \Rightarrow \frac{\sum X_i}{\beta^2} = \frac{n}{\beta} \Rightarrow \sum X_i = \beta n \Rightarrow$

The MLE is $\hat{\beta} = \overline{X}$. $\quad \hat{\beta} = \frac{\sum X_i}{n} = \overline{X}$

b. $E(\overline{X}) = \frac{1}{n} \cdot n E(X) = E(X) = \beta$. $\overline{X}$ is unbiased for β.

c. $\ell'(x|\beta) = -\beta^{-1} + x\beta^{-2} \Rightarrow \ell''(x|\beta) = +\beta^{-2} - 2x\beta^{-3}$

$I(\beta) = -E[\ell''(x|\beta)] = -E[\beta^{-2} - 2x\beta^{-3}] = \frac{-1}{\beta^2} + \frac{E(2x)}{\beta^3}$

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

d. $I_n(\theta) = nI(\theta) = n/\beta^2$. The Cramer-Rao lower bound is $1/(nI(\theta)) = \beta^2/n$.

Add 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:

In the setting with 2 simple hypotheses, the Neyman-Pearson lemma tells us the structure of the test statistic for the most powerful test @ a certain α is based on the likelihood ratio of the alternative and null densities. It allows us to get the "best" (in terms of power) test for a set α. It speaks to the power of the likelihood ratio and leads to the further development of UMP tests. If we have a test @ α^* in the form prescribed by NP, and we find another test with $\alpha < \alpha^*$, NP tells us β must be $> \beta^*$, ie. you can't improve power-wise with some other test if using the one prescribed by NP.

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.

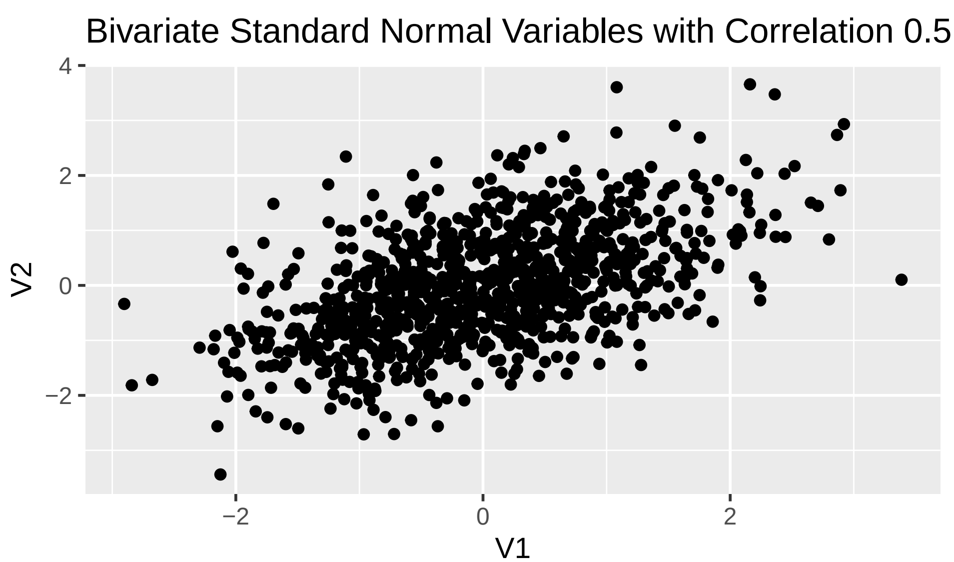
SOLUTION:

The chapter 4 and 5 problems demonstrated how to do this with the *mvtnorm* package, but other packages could be used to do it too.

```
# set desired parameters
meanvec <- c(0, 0)
sigma <- matrix(c(1, 0.5, 0.5, 1), ncol = 2)

# set seed for reproducibility and generate data
set.seed(495)
NormData <- as.data.frame(rmvnorm(1000, mean = meanvec, sigma = sigma))

#plot points
gf_point(V2 ~ V1, data = NormData) %>%
  gf_labs(title = "Bivariate Standard Normal Variables with Correlation 0.5")
```



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

SOLUTION:

Following the equations in the text, we see the conditional distribution of $x_2|x_1$ should be a Normal distribution with a mean of $0 + (0.5/1)(x_1 - 0) = (0.5)x_1$ and a variance of $1 - (0.5^2)/1 = 1 - 0.25 = 0.75$.

(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:

```
mymod <- lm(V2 ~ V1, data = NormData)
msummary(mymod)
```

```
##             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
```

The text formula says that the slope coefficient should be 0.5 and the R-squared should be $0.5^2 = 0.25$, based on the values we have been provided for the parameters here. We see that our random sample gave values very close to this, with a slope coefficient for V_1 of 0.547 and an R-squared of 0.277. These were evident in the conditional distribution in part b. The 0.5 times x_1 gave us the slope coefficient and the variance in the conditional distribution is what is NOT explained by the regression relationship, so if the un-explained variance is 0.75, the explained variance must be 0.25 (total variance = 1).

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:

When thinking about a permutation/randomization based test procedure, there are two main things to think about. First, what statistic will be recorded/studied, and what do you need to permute to generate values of that statistic possible under the null hypothesis. Here, we are thinking about the overall model significance, not focused on a particular predictor, so, a useful statistic is the overall F statistic. (Some argument could be made for R^2 as well, though I mentioned significance here to try to key you to a test statistic.)

Under the null hypothesis, all terms would have coefficients of 0. We can generate statistics from this setting simply by permuting the response variable (permuting predictors changes the relationship between the predictors). Combining these ideas, we would:

- Save the overall F statistic from the original model fit.
- Permute (shuffle) the response variable, fit the model, and record the F statistic.
- Repeat the permuting and recording step many times (say 1000 or 10000), using a seed to keep results reproducible.
- Examine where our original F statistic falls in the empirical null sampling distribution we generated to determine if the original model was useful.

(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:

There are many valid solutions here.

For a specific model term, we would probably focus on the slope coefficient. A bootstrap could be applied to get a better sense of the standard error of the slope coefficient, or used to obtain a confidence interval for the slope rather than relying on the usual theory. (You could focus on the t-test statistic as well.)

For the overall model, a bootstrap could help consider questions in regards to the overall F statistic, R^2 (or adjusted R^2), or the residual standard error (or MSE). For example, you could consider a bootstrap CI for the R^2 value to see if it's above a desired threshold.