ARE 106 Winter 2019 Chalfant

Answers to First Midterm Exam

The exam consists of 20 parts spread over seven questions. Answer each part in the space provided. Each part is weighted equally, worth 5 points.

Good luck!

1. (Dougherty R.22) A random variable X has a normal distribution with mean 5 and variance 10. Another random variable is formed using $Z = X/2$.

   (a) What is the expected value of Z?

   $$E(Z) = E(X/2) = (1/2)E(X) = 5/2.$$

   (b) What is the variance of Z?

   $$V(Z) = V(X/2) = (1/2)^2 V(X) = 10/4 = 5/2.$$

   (c) What is the standard deviation of Z?

   $$s.d.(Z) = \sqrt{V(Z)} = \sqrt{5/2}.$$

   (d) Is Z normally distributed? (Yes or No will suffice here.)

   Yes; linear functions of normal random variables are themselves normally distributed.

2. You are interested in the problem of estimating $\mu_X = E(X)$, where each X in your sample of size n is an independent draw from $X \sim N(\mu_X, \sigma_X^2)$.

(a) (Dougherty, problem R.16) Show that, when you have n observations, the condition that the generalised estimator

$$\lambda_1 X_1 + \lambda_2 X_2 + \ldots + \lambda_n X_n$$

should be an unbiased estimator of μ_X is

$$\lambda_1 + \lambda_2 + \ldots + \lambda_n = 1.$$

For the estimator to be unbiased, its expected value must equal μ_X here.

$$\begin{aligned} E\left[\lambda_1 X_1 + \lambda_2 X_2 + \ldots + \lambda_n X_n\right] &= \lambda_1 E(X_1) + \ldots + \lambda_n E(X_n) \\ &= \mu_X \sum_{i=1}^{n} \lambda_i \end{aligned}$$

and this equals $E(X_i) = \mu_X$ when $\sum \lambda_i = 1$.

(b) It follows from the previous part that X_1, the first observation from your sample, provides an unbiased estimator for μ_X.

Explain why you would prefer to use the sample mean, calculated using your entire sample, to estimate μ_X.

The sample mean has a smaller variance.

A single observation has a higher probability of being far from the true value, since $V(\overline{X}) = \sigma^2/n$ instead of σ^2. The sample mean is a more efficient estimator; as n increases, the probability distribution of the sample mean is ever more concentrated in the neighborhood around the population mean.

3. For the model

$$y_i = \beta_1 + \beta_2 X_i + u_i,$$

you now assume that it is β_2 that equals zero.

(a) Derive the least-squares estimator for $\hat{\beta}_1$.

$$\min_{\beta_1} \sum_{i=1}^{n} \left(y_i - \beta_1\right)^2$$

has the single first-order condition

$$2\sum_{i=1}^{n}(y_i - \beta_1)(-1) = 0,$$

and after dividing both sides by -2, we see that

$$\sum_{i=1}^{n} y_i = n\beta_1,$$

so the solution is

$$\hat{\beta}_1 = \sum_{i=1}^{n} y_i/n = \overline{y}.$$

Be sure you keep in mind that the estimate of the intercept is only as simple as $\overline{y}$ when we impose the restriction that $\beta_2 = 0$. Otherwise, it includes the $\hat{\beta}_2$ term, as always.

(b) It is known that for your $\hat{\beta}_1$,

$$\frac{\hat{\beta}_1 - a}{b} \sim N(0,1).$$

What is a? Give a specific value or expression.

The probability distribution of the sample mean should be familiar: $\overline{Y} \sim N(\beta_1, \sigma^2/n)$. We just changed the name of $E(y_i)$ from μ to β_1.

Thus,

$$a = \beta_1$$

What is b? Give a specific value or expression.

Similarly,

$$b = \sigma/\sqrt{n}.$$

Both results simply reflect the properties of a normal distribution, and how we turn any normal random variable into a standard normal.

4. (Dougherty, problem 2.2) For the model

$$y_i = \beta_1 + \beta_2 X_i + u_i,$$

you believe that $\beta_1 = 0$ and wish to estimate only β_2.

(a) Derive the least-squares estimator for β_2, assuming that $\beta_1 = 0$.

Now, we solve

$$\min_{\beta_2} \sum_{i=1}^{n} (y_i - \beta_2 X_i)^2$$

The first-order condition is

$$2 \sum_{i=1}^{n} (y_i - \beta_2 X_i)(-X_i) = 0,$$

which simplifies to

$$\sum_{i=1}^{n} (y_i - \beta_2 X_i)(X_i) = 0$$

or

$$\sum_{i=1}^{n} X_i y_i = \beta_2 \sum_{i=1}^{n} X_i^2.$$

The solution is

$$\hat{\beta}_2 = \frac{\sum_{i=1}^{n} X_i y_i}{\sum_{i=1}^{n} X_i^2}.$$

(b) Show that your estimator is unbiased.

$$\hat{\beta}_2 = \frac{\sum_{i=1}^{n} X_i y_i}{\sum_{i=1}^{n} X_i^2} = \sum_{i=1}^{n} k_i y_i$$

By writing it as a linear function in this manner, we can con-

clude immediately that

$$\begin{aligned} E\left(\hat{\beta}_2\right) &= E\left(\frac{\sum_{i=1}^{n} X_i y_i}{\sum_{i=1}^{n} X_i^2}\right) \\ &= E\left(\sum_{i=1}^{n} k_i y_i\right) \\ &= \sum_{i=1}^{n} k_i E(y_i) \\ &= \sum_{i=1}^{n} k_i \cdot \beta_2 X_i \\ &= \beta_2 \underbrace{\sum_{i=1}^{n} k_i X_i}_{1} \\ &= \beta_2 \end{aligned}$$

5. (Dougherty, problem 2.5) For the model

$$y_i = \beta_1 + \beta_2 X_i + u_i,$$

you estimated β_2 using

$$\hat{\beta}_z = \frac{\sum (Z_i - \overline{Z})(y_i - \overline{y})}{\sum (Z_i - \overline{Z})(X_i - \overline{X})}.$$

(a) Demonstrate that $\hat{\beta}_z$ is a linear function of the y_i's.

As was the case for our OLS estimator, the $\overline{y}$ term adds nothing to the numerator:

$$\sum (Z_i - \overline{Z})(y_i - \overline{y}) = \sum (Z_i - \overline{Z}) y_i$$

follows from the fourth useful fact on the handout accompanying the exam. Thus,

$$\hat{\beta}_z = \frac{\sum (Z_i - \overline{Z}) y_i}{\sum (Z_i - \overline{Z})(X_i - \overline{X})} = \frac{\sum (Z_i - \overline{Z})}{\sum (Z_i - \overline{Z})(X_i - \overline{X})} \cdot y_i.$$

We can think of this as

$$\hat{\beta}_z = \sum_{i=1}^{n} b_i y_i,$$

with

$$b_i = \frac{(Z_i - \overline{Z})}{\sum (Z_i - \overline{Z})(X_i - \overline{X})} \cdot .$$

That makes the next part much easier.

(b) Demonstrate that $\hat{\beta}_z$ is unbiased.

$$E\left(\hat{\beta}_z\right) = E\left[\sum_{i=1}^{n} b_i y_i\right] = \sum_{i=1}^{n} b_i E(y_i) = \sum_{i=1}^{n} b_i \left(\beta_1 + \beta_2 X_i\right)$$

The first term is zero, since $\sum b_i \beta_1 = \sum b_i = 0$*. The second term reduces to* β_2*, since* $\sum b_i X_i = 1$*. Hence,*

$$E(\hat{\beta}_z) = \beta_2;$$

the estimator is unbiased.

(c) (Dougherty 2.12) It can be shown that the variance of your estimator is given by

$$V(\hat{\beta}_z) = \frac{\sigma^2}{\sum \left(X_i - \overline{X}\right)^2} \cdot \frac{1}{r_{xz}^2}$$

where r_{xz} is the correlation between X and Z. What are the implications for the efficiency of your estimator?

By efficiency, we mean its variance relative to the smallest possible variance, which the OLS estimator achieves.

Unless X *and* Z *are perfectly correlated, i.e., unless* Z *is an exact linear function of* X*, then* $r_{xz} < 1$ *and*

$$V(\hat{\beta}_z) > V(\hat{\beta}_2)$$

where $\hat{\beta}_2$ *denotes our OLS estimator.*

If both estimators are unbiased and the OLS estimator has a smaller variance, we prefer to use it (it is the Best Linear Unbiased Estimator, according to Dougherty's Chapter 2).

6. (Dougherty, problem 1.7) Dougherty's Chapter 1 problems included this model:

$$\mathrm{S}_i = \beta_1 + \beta_2 \mathrm{ASVABC}_i + u_i$$

He refers to S as educational attainment—years of schooling of the respondent. ASVABC is a composite measure of numerical and verbal ability for the respondent.

Use the handout accompanying this exam and the parts labeled either "Results from R" or "Results from Stata" to answer the following questions.

(a) How do you interpret the estimated coefficient on ASVABC?
It measures the effect of a change in the ASVABC score on predicted years of schooling. Specifically, a unit increase in ASVABC leads to an increase in predicted years of schooling of $\hat{\beta}_2 = 1.58$.

(b) How do you interpret the estimated intercept?
Predicted years of schooling for ASVABC=0.
That does not always make sense, but it does so here because the score is normalized to have a mean near zero.
Hence, the intercept reflects predicted years of schooling for the case where the ASVABC score is zero.
(It is sufficient to simply state that the intercept is the predicted value of the dependent variable when the explanatory variable is zero.)

(c) How do you interpret the R^2 value?
The percentage of Total SS that our model "explained."

(d) How do you interpret the t-statistic of 13.51?
It tests the hypothesis that $\beta_2 = 0$ *and this hypothesis is rejected, based on either the large* t*-value or the small* p*-value.*

(e) According to another study, β_2 equals 0.7. If you were to test the hypothesis that β_2 equals 0.7, would you reject it? Why or why not?

We would reject it.

$$t = \frac{1.58 - .7}{.117} = \frac{.88}{.117}$$

and this will be far larger than the typical t^**.*

It would also suffice to note that a confidence interval for β_2 *does not incude 0.7, so again, the hypothesis would be rejected.*

Such a confidence interval could be approximated by adding and subtracting twice the standard error to $\hat{\beta}_2$*, or you can find the interval calculated by Stata on the handout.*

The same information is produced by the `confint` *command in the results from R.*

7. Also included in the handout accompanying this exam is the set of results from running a short simulation program using R.

   (a) Give a brief overview of the results from this program. What does the program seem to be doing?

   The program repeats our Homework 3 simulations, but with $\beta_1 = 0$ *and* $\beta_2 = 2$*. Two estimators of* β_2 *are compared; the OLS estimator that we also included in our Homework 3 simulations, and the "slope-only" estimator from earlier in this exam.*

   (b) How do you interpret the results from the commands `mean(b2s)` and `mean(otherb2s)`?

   Both estimators appear to be unbiased.

   How do you interpret the results from the commands `var(b2s)` and `var(otherb2s)`?

   They provide estimates of the population variances of these two estimators.

   The results show that the slope-only estimator is more efficient.

Math Handout to Accompany First Midterm Exam: ARE 106

- Six useful facts from my notes:

$$\text{(I)} \quad \sum_{i=1}^{n}(X_i - \overline{X}) = 0$$

$$\text{(II)} \quad \sum_{i=1}^{n}(X_i - \overline{X})^2 = \sum_{i=1}^{n} X_i^2 - n\overline{X}^2$$

$$\text{(III)} \quad \sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y}) = \sum_{i=1}^{n} X_i Y_i - n\,\overline{X} \cdot \overline{Y}$$

$$\text{(IV)} \quad \sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y}) = \sum_{i=1}^{n}(X_i - \overline{X})Y_i = \sum_{i=1}^{n}(Y_i - \overline{Y})X_i$$

$$\text{(V)} \quad \sum_{i=1}^{n}(X_i - \overline{X})^2 = \sum_{i=1}^{n}(X_i - \overline{X})X_i$$

$$\text{(VI)} \quad \sum_{i=1}^{n}(X_i - \mu)^2 = \sum_{i=1}^{n}(X_i - \overline{X})^2 + n(\overline{X} - \mu)^2$$

- OLS estimators:

$$\hat{\beta}_1 = \overline{Y} - \hat{\beta}_2\overline{X}$$

and

$$\hat{\beta}_2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2}.$$

- Probability distributions for OLS estimators:

$$\hat{\beta}_1 \sim \left[\beta_1, \sigma^2\left(\frac{1}{n} + \frac{\overline{X}^2}{\sum\left(X_i - \overline{X}\right)^2}\right)\right]$$

$$\hat{\beta}_2 \sim \left[\beta_2, \frac{\sigma^2}{\sum\left(X_i - \overline{X}\right)^2}\right]$$

Results to Accompany First Midterm Exam

Results from R

```
library(foreign)
mydata = read.dta("/home/are106/EAWE21.dta")
attach(mydata)
mymodel = lm(S~ASVABC)
summary(mymodel)
```

```
##
## Call:
## lm(formula = S ~ ASVABC)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.7126 -1.6813  0.1044  1.7147  6.2663
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  14.4368     0.1097  131.56   <2e-16 ***
## ASVABC        1.5809     0.1170   13.51   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.349 on 498 degrees of freedom
## Multiple R-squared:  0.2682, Adjusted R-squared:  0.2668
## F-statistic: 182.6 on 1 and 498 DF,  p-value: < 2.2e-16
```

```
anova(mymodel)
```

```
## Analysis of Variance Table
##
## Response: S
##            Df Sum Sq Mean Sq F value    Pr(>F)
## ASVABC      1   1007 1007.00  182.56 < 2.2e-16 ***
## Residuals 498   2747    5.52
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
confint(mymodel)
```

```
##                 2.5 %    97.5 %
## (Intercept) 14.221174 14.652369
## ASVABC       1.351015  1.810787
```

Results from Stata

```
library(foreign)
library(RStata)
options("RStata.StataPath"='/usr/local/stata15/stata')
options("RStata.StataVersion"=15)
```

Now some Stata commands:

```
stata_commands="
use /home/are106/EAWE21
reg S ASVABC
"
stata(stata_commands)
```

```
## .
## . use /home/are106/EAWE21
## . reg S ASVABC
##
##       Source |       SS           df       MS      Number of obs   =       500
## -------------+----------------------------------   F(1, 498)       =    182.56
##        Model |  1006.99534         1  1006.99534   Prob > F        =    0.0000
##     Residual |  2747.02666       498   5.5161178   R-squared       =    0.2682
## -------------+----------------------------------   Adj R-squared   =    0.2668
##        Total |    3754.022       499  7.52309018   Root MSE        =    2.3486
##
## ------------------------------------------------------------------------------
##            S |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
## -------------+----------------------------------------------------------------
##       ASVABC |   1.580901   .1170059    13.51   0.000     1.351015    1.810787
##        _cons |   14.43677   .1097335   131.56   0.000     14.22117    14.65237
## ------------------------------------------------------------------------------
```

Question 6

```
n=20
nreps=10000
X = c(1:20)
ys = matrix(-9999,n,nreps)

b1s = matrix(-9999,nreps,1)
b2s = matrix(-9999,nreps,1)
otherb2s = matrix(-9999,nreps,1)
for (i in 1:nreps) {
  ys[,i] = 2*X + rnorm(n)
  mymodel = lm(ys[,i]~X)
  b1s[i]=mymodel$coefficients[1]
  b2s[i]=mymodel$coefficients[2]

  otherb2s[i] = sum(X*ys[,i])/sum(X^2)
}
mean(b1s)
```

```
## [1] 0.005544138
```

```
mean(b2s)
```

```
## [1] 1.999384
```

```
mean(otherb2s)
```

```
## [1] 1.99979
```

```r
var(b2s)
```

```
##              [,1]
## [1,] 0.001503329
```

```r
var(otherb2s)
```

```
##               [,1]
## [1,] 0.0003550251
```