Econ 3040 A01 - Midterm - Winter 2023

Ryan T. Godwin

The exam is 70 minutes long, and consists of 72 marks (**approximately 1 mark per minute**). There are 10 short answer questions, each worth 4 marks. There are two long answer questions with 8 parts total, each part worth 4 marks. Write all answers in the provided exam booklet. You may only have a calculator and writing implements at your table. You may not use any books, notes, formula sheets, computers, or phones. A table of areas under the standard Normal curve is provided at the back of the exam, as well as a formula sheet.

DO NOT OPEN THIS EXAM BOOKLET UNTIL INSTRUCTED TO DO SO.

DON'T TOUCH! (Until instructed to do so).

Short Answer

1. What is meant by "the realization of a random variable"? How does this idea relate to a sample of data?

The "realization of a random variable" is the value that the random variable takes, after the randomness has resolved. For example, before dice are rolled, the result is random. After the roll, we observe the outcome as the realization of the random variable - it is now just a number.

The sample data are realizations of random variables. While the data appear to be just numbers in a spreadsheet, it is important to remember that the values could have been different - they came from a random process.

---

2. What does it mean for an estimator to be unbiased?

An estimator is unbiased if it gives the right answer on average. That is, if the expected value of the estimator is equal to the thing it's trying to estimate, the estimator is unbiased. If $\hat{\theta}$ is an unbiased estimator for θ, then we can write:

$$\mathrm{E}[\hat{\theta}] = \theta$$

---

3. Suppose that there is a random variable Y, with $\mathrm{E}[Y] = 2$ and $\mathrm{var}[Y] = 3$. What are the mean and variance of Z, where $Z = 2Y + 1$?

$$\mathrm{E}[Z] = \mathrm{E}[2Y + 1] = 2\,\mathrm{E}[Y] + 1 = 2 \times 2 + 1 = 5$$
$$\mathrm{var}[Z] = \mathrm{var}[2Y + 1] = 4\,\mathrm{var}[Y] + 0 = 4 \times 3 = 12$$

---

4. How does the central limit theorem relate to $\bar{Y}$ and b_1?

The CLT says that the sum of random variables tends to be Normally distributed. Since the sample average and the LS estimator both involved summing the sample data, the CLT implies that they are Normally distributed.

---

5. Why are the least squares residuals sometimes called "prediction errors"?

The LS residuals are the differences between the actual Y data and the LS predicted values $\hat{Y}$. That is:

$$e = Y - \hat{Y} = \text{actual} - \text{predicted}$$

Hence, the residuals can be thought of as prediction errors.

---

6. Prove that if all data points line up (fall on the least squares estimated line), then $R^2 = 1$.

If all of the data points line up, then the LS estimated line passes through each data point exactly, and all the predictions are perfect. That is, all $\hat{Y}_i = Y_i$, and all $e_i = 0$. From the formula for R^2:

$$R^2 = \frac{ESS}{TSS} = \frac{\sum(\hat{Y}_i - \bar{\hat{Y}})^2}{\sum(Y_i - \bar{Y})^2} = \frac{\sum(Y_i - \bar{Y})^2}{\sum(Y_i - \bar{Y})^2} = 1$$

or

$$R^2 = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum(e_i^2)}{\sum(Y_i - \bar{Y})^2} = 1 - \frac{0}{\sum(Y_i - \bar{Y})^2} = 1$$

---

7. What factors determine the variance (precision) of the least squares estimator?

The variance of the LS estimator decreases (the estimator gets more precise) when:

- the sample size increases
- the variance of X increases
- the variance of ϵ decreases

---

8. Why does the formula for s_ϵ^2 have an $(n-2)$ in the denominator?

The $(n-2)$ is required so that the estimator for the sample variance is unbiased. The "-2" is a degrees of freedom correction - two things must be estimated (b_0 and b_1) before the residuals can be calculated and used to calculate s_ϵ^2.

---

9. For the model: $Y = \beta_0 + \beta_1 X + \epsilon$, where X is a continuous variable, what is the interpretation of β_1?

β_1 is the marginal effect of X on Y, or the change in Y due to a one unit change in X.

---

10. For the model: $Y = \beta_0 + \beta_1 D + \epsilon$, where D is a dummy variable, what is the interpretation of β_1?

β_1 is the difference in the population means of Y for when $D = 1$ and for $D = 0$.

Long Answer

11. This question uses a dataset with $n = 200$ and two variables: `salary` - the yearly salary of a worker in thousands of dollars, `experience` - the number of years of work experience. The population model: $salary = \beta_0 + \beta_1 experience + \epsilon$ is estimated in R:

```
summary(lm(salary ~ experience), data = mydata)
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  43.3644     2.8322  15.311   <2e-16 ***
experience    0.4999     0.1669   2.994   0.0031 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 20.26 on 198 degrees of freedom
Multiple R-squared:  0.04332,   Adjusted R-squared:  0.03849
F-statistic: 8.966 on 1 and 198 DF,  p-value: 0.003101
```

a) What is the estimated increase in salary due to an increase in experience?

The estimated increase in `salary` due to an increase in `experience` is 0.5 thousand dollars per year.

---

b) What percentage of the variation in salary can be explained using variation in years of experience?

$R^2 = 0.04$, meaning that 4% of the variation in `salary` can be explained using variation in `experience`.

---

c) Use a 95% confidence interval to test the null hypothesis $H_0 : \beta_1 = 0$.

The 95% confidence interval is:

$$b_1 \pm 1.96 \times s.e.(b_1) = 0.4999 \pm 1.96 \times 0.1669 = [0.17, 0.83]$$

The null hypothesis of 0 is not inside of the confidence interval, so we reject the null hypothesis at the 5% significance level.

---

d) What is the p-value for the hypothesis test in part (c)?

The p-value is given in the table of R output. It is 0.0031.

---

e) One of the observations in the sample is $salary = 81.6, experience = 10$. Calculate the predicted value and residual for this observation.

For the observation with $salary = 81.6$ and $experience = 10$, the predicted salary value is:

$$\hat{Y} = 43.3644 + 0.4999 \times 10 = 48.4$$

and the residual is:

$$e = Y - \hat{Y} = 81.6 - 48.4 = 33.2$$

---

12. This question uses data on `mark` - the final percentage mark for the course (0% - 100%) for a student in ECON 3040 last semester, and `attend` - a dummy variable equal to 1 if the student was in attendance, and 0 if the student was not in attendance (attendance was only taken for one day). The population model: $mark = \beta_0 + \beta_1 attend + \epsilon$ is estimated in R:

```
summary(lm(mark ~ attend, data = attend))
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   67.075      4.989  13.445   <2e-16 ***
attend         7.576      6.340   1.195    0.239
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 19.95 on 40 degrees of freedom
Multiple R-squared:  0.03446,   Adjusted R-squared:  0.01032
F-statistic: 1.428 on 1 and 40 DF,  p-value: 0.2392
```

a) What is the sample average mark for students who *did not* attend class? What is the sample average mark for students who *did* attend class?

The sample average for those who *did not* attend is equal to $b_0 = 67.075$. The sample average for those who *did* attend is $b_0 + b_1 = 67.075 + 7.576 = 75.651$.

---

b) Test the hypothesis that attendance has no effect on marks.

The null hypothesis is $H_0 : \beta_1 = 0$. The `summary()` function automatically performs this hypothesis test. The p-value for this test is 0.239, so we fail to reject the null hypothesis at the 10% significance level.

---

c) Suppose that I used the same data, but instead estimated the model: $mark = \beta_0 + \beta_1 absent + \epsilon$, where *absent* is a dummy variable equal to 1 if the student was *missing* from class, and equal to 0 if they were in attendance (the dummy variable has been defined in the opposite way). What would be the values for b_0 and b_1 in this situation?

The new values would be $b_0^{\star} = 75.651$ and $b_1^{\star} = -7.575$.

Table 1: Area under the standard normal curve, to the right of z.

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641
0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002

Formula Sheet

expected value (mean) of Y (for discrete Y)	$\mu_Y = \sum p_i Y_i$
variance of Y (for discrete Y)	$\sigma_Y^2 = \sum p_i \left(Y_i - \mu_y\right)^2$
standard deviation of Y	$\sigma_Y = \sqrt{\sigma_Y^2}$
covariance between X and Y	$\sigma_{XY} = E\left[\left(X - \mu_X\right)\left(Y - \mu_Y\right)\right]$
correlation coefficient (between X and Y)	$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$
expected value of the sample average, $\bar{Y}$	$\mathrm{E}(\bar{Y}) = \mu_Y$
variance of the sample average, $\bar{Y}$	$\mathrm{var}[\bar{Y}] = \frac{\sigma_Y^2}{n}$
sample variance of Y (estimator for σ_Y^2)	$s_Y^2 = \frac{1}{n-1} \sum_{i=1}^n \left(Y_i - \bar{Y}\right)^2$
sample variance of e (estimator for σ_ϵ^2)	$s_\epsilon^2 = \frac{1}{n-2} \sum_{i=1}^n e_i^2$
t-statistic	$t = \frac{\text{estimate} - \text{hypothesis}}{\text{std. error}}$
95% confidence interval	estimate $\pm 1.96 \times$ std. error
LS estimator for β_1 (single regressor model)	$b_1 = \frac{\sum_{i=1}^n \left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{\sum_{i=1}^n \left(X_i - \bar{X}\right)^2}$
LS estimator for β_0 (single regressor model)	$b_0 = \bar{Y} - b_1 \bar{X}$
variance of b_1 (single regressor model)	$\mathrm{var}\left[b_1\right] = \frac{\sigma_\epsilon^2}{\sum X_i^2 - \frac{\left(\sum X_i\right)^2}{n}}$
LS predicted values (single regressor model)	$\hat{Y}_i = b_0 + b_1 X_i$
LS residuals	$e_i = Y_i - \hat{Y}_i$
R-squared	$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS}$
TSS	$\sum_{i=1}^n \left(Y_i - \bar{Y}\right)^2$
ESS	$\sum_{i=1}^n \left(\hat{Y}_i - \bar{\hat{Y}}\right)^2$
RSS	$\sum_{i=1}^n \left(e_i^2\right)$