Econ 3040 A01 - Midterm - Fall 2023

Ryan T. Godwin

- The exam is 70 minutes long, and consists of 72 marks (**approximately 1 mark per minute**).
- **Each question is worth 4 marks** (except for the bonus question).
- Write all answers in the provided exam booklet.
- You may only have a calculator and writing implements at your table.
- No books, notes, formula sheets, computers, phones, etc.
- A table of areas under the standard Normal curve, and a **formula sheet**, are provided at the end of this booklet.

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

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

Short Answer

1. Below is a *joint* probability function, but the probabilities of the events are missing:

	$X = 0$	$X = 1$
$Y = 0$	?	?
$Y = 1$	?	?

Fill in the missing probabilities so that the correlation between X and Y is 1.

The off-diagonal elements of the table should be zero, and the diagonal elements can be any two numbers that sum to 1. For example:

	$X = 0$	$X = 1$
$Y = 0$	0.4	0
$Y = 1$	0	0.6

2. Below is the probability distribution for a random variable Y:

	Y = 1	Y = 2	Y = 3
P(Y)	0.3	0.4	0.3

Calculate the expected value, and variance, of Y.

$$\mathbb{E}(Y) = (0.3 \times 1) + (0.4 \times 2) + (0.3 \times 3) = 2$$

$$\text{var}(Y) = 0.3(1-2)^2 + 0.4(2-2)^2 + 0.3(3-2)^2 = 0.6$$

3. How would you *standardize* the variable from Question 2?

Standardizing a variable is when the variable is transformed such that its mean is 0 and variance is 1 (such as in the standard Normal distribution, and the z-test statistic). To standardize a variable, subtract the mean and divide by the standard deviation. In this case:

$$z = \frac{Y - 2}{\sqrt{0.6}}$$

z has mean 0, and variance 1.

4. What does it mean for $\bar{Y}$ and b_1 to be *efficient*?

It means that $\bar{Y}$ is the estimator for μ with the smallest variance, among all other (linear and unbiased) estimators for μ. Similarly, efficiency for b_1 means that b_1 has the smallest variance among all other (linear and unbiased) estimators for β_1.

5. Describe a situation where $R^2 = 1$.

$R^2 = 1$ when the regression "line" passes through each data point. All of the predicted values match the actual values ($\hat{y}_i = y_i \; \forall \; i$), and all of the residuals are equal to 0.

6. 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 siz n increases
- the variation in X increases
- the variation in ϵ decreases

7. Why do the estimators s_Y^2 and s_ϵ^2 (see the formula sheet) use a degrees-of-freedom correction?

The degrees of freedom correction ensures that the estimators are unbiased.

8. 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. That is, β_1 is the amount the y changes, for a change in X of 1.

9. Referring to the lectures and textbook, explain why the estimated value for β_1 changes so much between the equations:

$$Price = \beta_0 + \beta_1 Fireplaces + \epsilon$$

and:

$$Price = \beta_0 + \beta_1 Fireplaces + \beta_2 Living.Area + \epsilon$$

In class (and in the textbook and slides), we estimated β_1 to be \$66,669 for the first model, and \$8,962 for the second model. This is a big difference.

The first model suffers from *omitted variable bias*. The estimate for β_1 will be wrong (on average). The variable $Living.Area$ has been omitted from this model, causing the estimated effect of $Fireplaces$ on $Price$ to be biased (much too large).

The problem is that $Fireplaces$ and $Living.Area$ are correlated, and that $Living.Area$ is important in determining $Price$. When $Living.Area$ is omitted, the effect of a larger house on the price of the house gets mixed up with the effect that more fireplaces has on the price. When there are more fireplaces, the house is larger! This is the main reason for an increase in price.

10. (1 bonus mark) What is a sampling distribution?

A sampling distribution is the special name given to the probability function of an estimator.

Long Answer

11. This question uses a dataset with $n = 223$ and two variables: `wage` - the hourly wage of a worker in dollars, `female` - a dummy variable that takes on the value 1 if the worker is female, and 0 otherwise[1]. The population model: $wage = \beta_0 + \beta_1 female + \epsilon$ is estimated in R:

```
summary(lm(wage ~ female), data = mydata)
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  11.4348     0.2816  40.609   <2e-16 ***
female        0.9253     0.4009   2.308   0.0219 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.993 on 221 degrees of freedom
Multiple R-squared:  0.02353,   Adjusted R-squared:  0.01911
F-statistic: 5.326 on 1 and 221 DF,  p-value: 0.02193
```

a) What is the estimated difference in the wages between men and women?

On average, women make \$0.93 more per hour.

b) What are the sample mean wages for men and for women?

The sample mean wage for men is \$11.44 and the sample mean wage for women is \$12.36.

c) Test the null hypothesis that there is no wage-gender gap (that there is no difference in the wages between men and women).

This test has already been performed by R. The t-statistic is 2.308, and with a p-value of 0.0219 we reject the null hypothesis at the 5% significance level.

d) Construct a 95% confidence interval around b_1.

$$0.9253 \pm 1.96 \times 0.4009 = [0.1395\ ,\ 1.7111]$$

e) What is an interpretation of the confidence interval?

Either say: (i) A 95% confidence interval is a random interval, where 95% of such intervals contain the true parameter value; or (ii) a 95% confidence interval contains all values for the null hypothesis that will not be rejected at the 5% significance level.

f) What percentage of the variation in *wage* can be explained by $female$?

2.35% of the variation in *wage* can be explained by the dummy variable $female$.

g) What might be some other sources of variation in wage, and how are these other sources represented in the population model?

All the other variables that effect wage are contained in the error term ϵ, and could be things like education, years of work experience, age, etc.

[1] The data set is old and only allows for two genders.

h) One of the observations in the sample is $wage = 11.31, female = 0$. Calculate the predicted value and residual for this observation.

$$\hat{wage} = 11.4348 + 0.9253(0) = 11.4348$$

$$e = 11.31 - 11.4348 = -0.1248$$

i) Another researcher uses the same data, but defines the dummy variable in the opposite way ($male = 1$ if the worker is male, and $male = 0$ otherwise). They estimate the equation: $wage = \beta_0 + \beta_1 male + \epsilon$ using the same dataset. What will be the estimated values for b_0 and b_1?

The new estimates will be $b_0 = 12.36$ and $b_1 = -0.9253$.

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	$\sum_{i=1}^{n}\left(e_i^2\right)$