

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.

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 .

- How would you *standardize* the variable from Question 2?
- What does it mean for \bar{Y} and b_1 to be *efficient*?
- Describe a situation where $R^2 = 1$.
- What factors determine the variance (precision) of the least squares estimator?
- Why do the estimators s_Y^2 and s_ϵ^2 (see the formula sheet) use a degrees-of-freedom correction?
- For the model: $Y = \beta_0 + \beta_1 X + \epsilon$, where X is a continuous variable, what is the interpretation of β_1 ?
- 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:

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

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

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 otherwise¹. 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
```

- What is the estimated difference in the wages between men and women?
- What are the sample mean wages for men and for women?
- Test the null hypothesis that there is no wage-gender gap (that there is no difference in the wages between men and women).
- Construct a 95% confidence interval around b_1 .
- What is an interpretation of the confidence interval?
- What percentage of the variation in *wage* can be explained by *female*?
- What might be some other sources of variation in *wage*, and how are these other sources represented in the population model?
- One of the observations in the sample is $wage = 11.31$, $female = 0$. Calculate the predicted value and residual for this observation.
- 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 ?

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

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