Econ 3040 Final Exam

Ryan T. Godwin

The exam is 3 hours long, and consists of 100 marks. **There are 15 questions**. There is a table of critical values for the F-statistic, a table of standard Normal probabilities, and a formula sheet, at the end of the exam.

Short answer - each question worth 4 marks - 40 marks total

1. A random variable X is equal to 1 with probability 0.4, and equal to 4 with probability 0.6. What is the mean and variance of X?

$$\mathrm{E}[X] = 0.4 \times 1 + 0.6 \times 4 = 2.8 \quad (2)$$
$$\mathrm{var}[X] = 0.4 \times (1 - 2.8)^2 + 0.6 \times (4 - 2.8)^2 = 2.16 \quad (2)$$

---

2. How is the least-squares estimator derived? (Where does the equation for b_0, b_1, etc. come from?) Don't try to derive the formula, just set-up the problem, or describe the process.

The b_0 and b_1 formulas result from a calculus minimization problem, where the sum-of-squared residuals ($\sum e_i^2$) are minimized by choosing b_0 and b_1.

---

3. What does it mean for least-squares to be the most "efficient" estimator?

It means that it has the smallest variance among all other linear and unbiased estimators for β. The Gauss-Markov theorem proves this result.

---

4. Why are estimators random variables?

Because they are calculated from a random sample of data.

---

5. Why does R^2 always increase when a variable is added to the model? How does $\bar{R}^2$ fix the problem?

Adding another variable adds another β. The minimization problem becomes easier. Sum of squared residuals must decrease, R^2 must increase. $\bar{R}^2$ fixes the problem by introducing a penalty for the number of variables, k.

---

6. Explain the main problem with the following population model:

$$wage = \beta_0 + \beta_1 educ + \beta_2 male + \beta_3 female + \epsilon$$

This is the dummy variable trap; there is perfect multicollinearity. For only two genders, $male + female = 1$; there is a perfect link between the two variables.

---

7. This question uses the diamond price data:

```
summary(lm(price ~ carat + I(carat^2), data=diam)
```

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    -42.51     316.37  -0.134   0.8932
carat         2786.10    1119.61   2.488   0.0134 *
I(carat^2)    6961.71     868.83   8.013  2.4e-14 ***
```

What is the predicted increase in price due to an increase in carats? Your answer should include several numbers.

Predict two different 0.1 increases.

0.1 - 0.2	487.4613
0.2 - 0.3	626.6955
0.3 - 0.4	765.9297
0.4 - 0.5	905.1639
0.5 - 0.6	1044.398
0.6 - 0.7	1183.632
0.7 - 0.8	1322.866
0.8 - 0.9	1462.101
0.9 - 1.0	1601.335
1.0 - 1.1	1740.569
1.1 - 1.2	1879.803

---

8. For the model in question 7, how would you go about determining the appropriate degree (r) of the polynomial?

You could test the significance of the highest order of the polynomial. If you fail to reject, drop the variable, and repeat. Stop once you reject the null. The highest order term still left in the equation is the "appropriate" degree of the polynomial.

---

9. What is imperfect multicollinearity?

When two variables are highly correlated. This results in uncertainty around the estimated effects of those variables; high standard errors, large confidence intervals. Interpretation and properties of other estimators are unaffected.

---

10. The following population model:

$$\log(CO_2) = \beta_0 + \beta_1 \log(GDP) + \epsilon$$

is estimated in R:

```
co2mod <- lm(log(co2) ~ log(gdp.per.cap), data = co2)
summary(co2mod)
```

```
Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      -9.94045    0.36806  -27.01   <2e-16 ***
log(gdp.per.cap)  1.20212    0.04234   28.39   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

CO_2 is per capita carbon dioxide emissions, and GDP is GDP per capita, for 134 different countries. What is the interpretation of the estimated value of 1.20212?

A 1% increase in GDP per capita is associated with a 1.2% increase in CO2 emissions.

Long answer - each part worth 3 marks - 60 marks total

11. This question involves *heteroskedasticity.* First, a wage model is estimated using least squares:

```
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)           0.53764    0.70887   0.758 0.448521
education             0.18311    0.11333   1.616 0.106753
gendermale            0.69499    0.20315   3.421 0.000672 ***
age                  -0.06472    0.11345  -0.570 0.568616
experience            0.07754    0.11355   0.683 0.494959
education:gendermale -0.03362    0.01531  -2.196 0.028545 *
```

then, *heteroskedastic* robust standard errors are calculated (using the "sandwich" and "lmtest" packages like you did in assignment 4):

```
                      Estimate Std. Error t value  Pr(>|t|)
(Intercept)           0.537643   0.194521  2.7639 0.0059104 **
education             0.183114   0.011411 16.0471 < 2.2e-16 ***
gendermale            0.694988   0.191017  3.6384 0.0003013 ***
age                  -0.064716   0.013117 -4.9339 1.082e-06 ***
experience            0.077542   0.014099  5.4997 5.936e-08 ***
education:gendermale -0.033616   0.014731 -2.2819 0.0228902 *
```

a) What are homoskedasticity and heteroskedasticity?

Homoskedasticity is when the variance of the error term is constant for all observations: $\text{var}(\epsilon_i) = \sigma^2 \quad \forall \quad i$. Heteroskedasticity is when the variance of the error term differs between observations: $\text{var}(\epsilon_i) = \sigma_i^2$.

b) What is wrong with assuming homoskedasticity, when there is actually heteroskedasticity?

If we assume that there is homoskedasticy, when in reality the data is heteroskedastic, the estimator for the standard errors is inconsistent (it is based on the wrong formula). Confidence intervals, test statistics and their associated p-values, are all incorrect. Hypothesis testing is invalid.

c) How could you use the first estimated model to test for heteroskedasticity?

To perform White's test for heteroskedasticity, you would store the residuals from the first model. Then you would regress the squared residuals on all x variables, their squared values, and cross products. If the R^2 from this regression is high enough, then there is an explanation for the size of the squared residuals, and you would reject the null of homoskedasticity.

d) Point out the importance of using robust standard errors by using the output above.

Using robust standard errors changes everything in the ouptut table, except for the estimated βs. A very important change is that, under the robust estimator, all of the variables now appear to be statistically significant.

12. Two models are estimated to explain the effect of installing a fireplace on the selling price of a house (in dollars). The R output for the regression results are given below:

```
house.mod1 <- lm(Price ~ Fireplaces + Bathrooms, data=house)
summary(house.mod1)
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    44771       5743   7.796 1.10e-14 ***
Fireplaces     25414       3749   6.778 1.67e-11 ***
Bathrooms      79940       3167  25.241  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 77970 on 1725 degrees of freedom
Multiple R-squared:  0.3734,    Adjusted R-squared:  0.3727
F-statistic:   514 on 2 and 1725 DF,  p-value: < 2.2e-16
```

```
house.mod2 <- lm(Price ~ Fireplaces + Living.Area + Bathrooms, data=house)
summary(house.mod2)
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -118.217   5369.069  -0.022    0.982
Fireplaces   5232.053   3384.481   1.546    0.122
Living.Area    91.431      3.928  23.276  < 2e-16 ***
Bathrooms   25511.611   3620.039   7.047 2.63e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 68030 on 1724 degrees of freedom
Multiple R-squared:  0.5232,    Adjusted R-squared:  0.5224
F-statistic: 630.6 on 3 and 1724 DF,  p-value: < 2.2e-16
```

a) What is the *main* difference between the two models? (If you had to focus on *just one* difference, what would it be?)

The estimated effect of *Fireplaces* on *Price* changes from $25,414 to $5,232. This is a very large swing in the estimated marginal effect.

b) What is the problem with the first model? (Why is it worse than the second model?)

The first model has omitted variable bias. It is missing an important variable that is correlated to both *Fireplaces* and *Price*. The estimator for the marginal effect in the first model is wrong (biased and inconsistent).

c) Using the second model: how much do you *predict* a 2000 square foot house with 2 bathrooms and 1 fireplace would sell for?

$$\hat{Price} = -118.217 + 5323.053(1) + 91.431(2000) + 25511.611(2) = 239090.10$$

d) What are the F-statistics of 514 and 630.6 for?

These are F-statistics for tests of the overall significance of any of the variables in the model. For example, in the first model the null hypothesis being tested is $H_0 : \beta_{Fireplaces} = 0$ and $\beta_{Bathrooms} = 0$.

13. When estimating the model:

$$wage = \beta_0 + \beta_1 education + \beta_2 gender + \beta_3 age + \beta_4 experience + \epsilon$$

the results indicate that `age` and `experience` are *insignificant*:

```
summary(lm(wage ~ education + gender + age + experience, data=cps))
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   -1.9574     6.8350  -0.286    0.775
education      1.3073     1.1201   1.167    0.244
genderfemale  -2.3442     0.3889  -6.028 3.12e-09 ***
age           -0.3675     1.1195  -0.328    0.743
experience     0.4811     1.1205   0.429    0.668
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.458 on 529 degrees of freedom
Multiple R-squared:  0.2533,    Adjusted R-squared:  0.2477
F-statistic: 44.86 on 4 and 529 DF,  p-value: < 2.2e-16
```

so, the variables `age` and `experience` are dropped from the model, and we get:

```
summary(lm(wage ~ education + gender, data=cps))
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.21783    1.03632   0.210    0.834
education     0.75128    0.07682   9.779  < 2e-16 ***
genderfemale -2.12406        ?     -5.273 1.96e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.639 on 531 degrees of freedom
Multiple R-squared:  0.1884,    Adjusted R-squared:  0.1853
F-statistic: 61.62 on 2 and 531 DF,  p-value: < 2.2e-16
```

a) What are the benefits to "dropping" variables from a model?

Simpler models are always better; they are easier to understand, estimate, and communicate. Statistically speaker, smaller models are more efficient (the estimators have smaller variance). This is because the entire dataset can focus on estimating fewer βs.

b) Why shouldn't we use t-tests to determine if these two variables can be dropped?

Even though the individual t-statistics on these two variables indicate that they are insignificant, we shouldn't use the t-test to decide whether to drop both of the variables. If they are correlated, then so are the bs, and so are the t-statistics. We need to take into account this correlation if we want to test a *joint* hypothesis.

c) Test the null hypothesis:

$$H_0 : \beta_3 = 0 \text{ and } \beta_4 = 0$$

What do you conclude?

As this is a joint hypothesis, we need to use an F test. The F-statistic is:

$$
\begin{aligned}
F &= \frac{\left(R_U^2 - R_R^2\right)/q}{\left(1 - R_U^2\right)/\left(n - k_U - 1\right)} \\
&= \frac{(0.2533 - 0.1884)/2}{(1 - 0.2533)/(529)} \\
&= 22.99
\end{aligned}
$$

Looking at Table 2, the appropriate critical value is 3.00. Since thhe F-stat of 22.99 is greater than this critical value, we reject the null at the 5% significance level. Even thought these variables appear to be insignificant according to the t-tests, they are *jointly* significant (we can't drop them both).

d) In the second table, what is the value for the missing (`?`) `Std. Error`?

The null hypothesis in the table is $H_0 : b_i = 0$, so the t-statistic is:

$$
t = \frac{b_i - 0}{\text{se}(b_i)}
$$

so the missing standard error is:

$$
\text{se}(b_i) = \frac{b_i}{\text{se}(t_i)} = \frac{-2.124}{-5.273} = 0.403
$$

14. This question is about differences-in-differences (DiD). A minimum wage increase happened in City B (this is the "treatment" group). There was no minimum wage increase in City A (the "no-treatment" group), but City A and City B are otherwise very similar. The number of employees in 100 retail stores (where workers are paid minimum wage) is observed in each city, both before and after the minimum wage increase.

Table 1: Average number of workers in 100 retail stores in City A (where there was no minimum wage increase) and City B (where there was a minimum wage increase). The number of workers is measured both before the minimum wage increase (at `time = 0`) and after the minimum wage increase (at `time = 1`).

	`time` = 0	`time` = 1
City A `treatment.group` = 0	35.2	25.7
City B `treatment.group` = 1	32.1	27.1

The variables in the data are:

Variable	Description
`employed`	the number of workers employed in a retail store
`time`	= 1 if after the minimum wage increase
	= 0 if before the minimum wage increase
`treatment.group`	= 1 if in City B (where the minimum wage increase happened)
	= 0 if in City A (no minimum wage increase)

a) The average number of employees in the retail stores in City B fell by 5 after the minimum wage increase. What is the problem with claiming that the minimum wage increase *caused* this decline in employment?

The causal effect is the difference between reality, and counterfactual. The reality is that employment fell from 32.1 to 27.1. But what would it have been if there had been no minimum wage increase? It is unlikely that it would have stayed flat at 32.1 over time. Other things could have caused the decrease, for example the economy could have been heading into a recession. Maybe it would have dropped to 27.1 anyway?

b) What is the DiD estimator for the effect of the minimum wage increase on employment?

The difference in employment for City A is $25.7 - 35.2 = -9.5$. DiD assumes that this is what would have happened for City B if there was no wage increase. What actually happened was -5. Employment is actually 4.5 higher than what it should have been. This is the DiD estimate. DiD estimate = (difference in treatment group) - (difference in control group).

c) What assumption needs to be made for the DiD estimator in part (b) to work?

The assumption that: what happened in City A would also have happened in City B (in the absence of a wage increase) is called the *parallel trends* assumption.

d) The model:

$$employed = \beta_0 + \beta_1 treatment.group + \beta_2 time + \beta_3(treatment.group \times time) + \epsilon$$

is estimated using the data above. What is the estimated value of β_3?

The estimate for β_3 is also the DiD estimator, so it will be equal to 4.5. β_3 is the extra difference in employment over time, for the treatment group.

e) **Bonus question.** What are the estimated values of β_0, β_1, and β_2?

$b_0 = 35.2, b_1 = -3.1, b_2 = -9.5$

15. This question involves *instrumental variables.* Consider the simple model:

$$y = \beta_0 + \beta_1 x + \epsilon$$

a) Suppose that there is a missing variable m that is correlated with both the dependent variable y, and a regressor x. In this case, what happens to the least-squares estimator b_1? (What are the properties of b_1?)

The LS estimator is biased and inconsistent.

b) What properties must an instrument z have, in order to be "valid"? (In order for it to work in instrumental variables estimation?)

z must be correlated to the "problem" endogenous x variable, and must be uncorrelated with the missing variable (uncorrelated with the error term).

Now, consider the *wage*, *education*, and *distance from college* data. First a model is estimated by LS:

```
college <- read.csv("https://rtgodwin.com/data/collegedist.csv")
ls <- lm(wage ~ education + urban + gender + ethnicity + unemp, data=college)
summary(ls)
```

```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       8.000192   0.156928  50.980   <2e-16 ***
education         0.005369   0.010362   0.518   0.6044
urbanyes          0.070117   0.044727   1.568   0.1170
gendermale        0.085242   0.037069   2.300   0.0215 *
ethnicityhispanic 0.012048   0.062385   0.193   0.8469
ethnicityother    0.556056   0.052167  10.659   <2e-16 ***
unemp             0.133101   0.006711  19.834   <2e-16 ***
```

and then by instrumental variables (IV) estimation, using *distance from college* as the instrument:

```
iv <- ivreg(wage ~ education + urban + gender + ethnicity + unemp |
                    distance + urban + gender + ethnicity + unemp,
            data=college)
summary(iv)
```

```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       -0.65702    1.83641  -0.358   0.7205
education          0.64710    0.13594   4.760 1.99e-06 ***
urbanyes           0.04614    0.06039   0.764   0.4449
gendermale         0.07075    0.04997   1.416   0.1569
ethnicityhispanic -0.12405    0.08871  -1.398   0.1621
ethnicityother     0.22724    0.09863   2.304   0.0213 *
unemp              0.13916    0.00912  15.259  < 2e-16 ***
```

c) Describe the major important difference between the two estimated models.

There is a massive swing in the estimated returns to education.

d) How does the two-stage least squares (2SLS) procedure work? Explain the steps using the above example.

First, education is regressed on the instrument and the other x variables, getting the predicted values $\widehat{education}$ from this regression. Second, the model is estimated using LS, but replacing $education$ with $\widehat{education}$.

END

Table 2: Critical values for the F-test statistic.

q	5% critical value
1	3.84
2	3.00
3	2.60
4	2.37
5	2.21

Table 3: 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

Econ 3040 Final Exam 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, $\bar{Y}$	$\sigma_{\bar{Y}}^2 = \frac{\sigma_Y^2}{n}$
sample variance of Y (estimator for σ^2)	$s_Y^2 = \frac{1}{n-1} \sum_{i=1}^n \left(Y_i - \bar{Y}\right)^2$
sample variance of y in a regression model	$s_y^2 = \frac{1}{n-k-1} \sum_{i=1}^n e_i^2$
t-statistic (assuming large n)	$t = \frac{\text{estimate } - \text{ hypothesis}}{\text{std. error}}$
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	$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS}$
adjusted-R-squared	$\bar{R}^2 = 1 - \frac{RSS/(n-k-1)}{TSS/(n-1)}$
F-statistic	$F = \frac{\left(R_U^2 - R_R^2\right)/q}{\left(1 - R_U^2\right)/(n - k_U - 1)}$
IV estimator	$\hat{\beta}_{IV} = \frac{\sum[(y-\bar{y})(z-\bar{z})]}{\sum[(x-\bar{x})(z-\bar{z})]}$