ECON3150/4150: Introductory Econometrics – Final Exam Spring 2025

Be brief and to the point.

Always motivate your answers.

Use the tables at the end of the exam where necessary.

1. [2 points] Consider the two "studies" in Figure 1 and assume the central limit theorem applies (i.e. your estimator is normally distributed, see exam appendix tables).

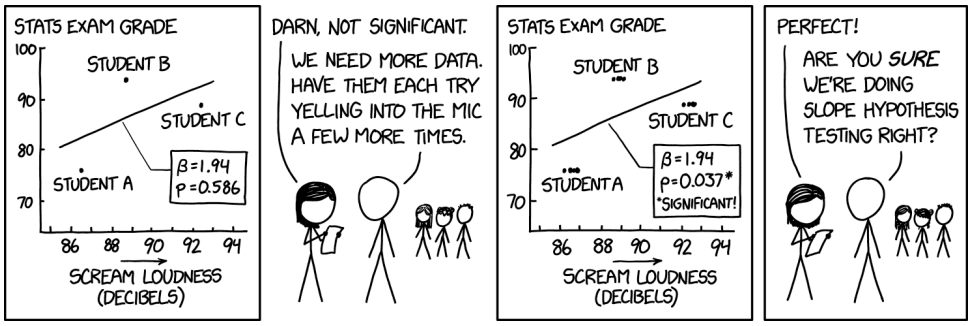

Figure 1: Statistical precision and sample size

(a) What is the standard error on the estimated slope coefficient in the first "study"?

ANSWER HINT: From the **definition of a p-value** and the CLT we know that $0.586 = 2 * \Pr(t > 1.94/SE) = 2 * (1 - \Phi(1.94/SE))$. This gives us $SE = 1.94/\Phi^{-1}(1 - 0.586/2)$. From the appendix table we see that $\Phi^{-1}(1 - 0.586/2) \approx \Phi^{-1}(0.71) = 0.55$, which gives $SE = 1.94/0.55 \approx 3.5$.

Another way of going about the same thing notes that $\Pr(|t| > |t^*|) = 0.586$, where $t^* = 1.94/SE$. This means that $\Pr(t > t^*) = 0.586/2$, or $\Pr(t < t^*) = 1 - 0.586/2 = 0.707$. From the appendix table we then know that t^* is about 0.55, which gives $SE = 1.94/0.55 \approx 3.5$.

(b) Assuming that the sample size in the first study equals 3, what sample size is implied by the p-value of the second "study" that has more data?

ANSWER HINT: In the second figure we have $0.037 = 2 * (1 - \Phi(1.94/SE)$. From the **definition of a standard error** and (a) we know that $SE = SD/\sqrt{3} \Rightarrow SD = 6.109416$. It now follows that

$$n = \left(\frac{SD}{1.94}\Phi^{-1}(1 - 0.037/2)\right)^2$$

Since $\Phi^{-1}(1 - 0.037/2) \approx \Phi^{-1}(0.98) = 2.05$, we get a sample size of $n = \left(\frac{6.109416}{1.94}2.05\right)^2 \approx 43$.

Alternatively (as in (a)) we know that $\Pr(|t| > |t^*|) = 0.037$ or $\Pr(t < t^*) = 1 - 0.037/2 = 0.9815$, from which it follows that $t^* \approx 2.085$ which gives $SE = 1.94/2.085 \approx 0.93$. Finally, $SE = SD/\sqrt{n}$ and we know from (a) that $SD = 3.5 * \sqrt{3} = 6.1$, and $n = (SD/SE)^2 = (6.1/0.93)^2 \approx 43$.

---

2. [4 points] Consider the relationships between an outcome `y`, and explanatory variables `x` (and sometimes) `d` in Figure 2. Provide for panels (a) to (d) the `R` code that estimates the minimal correctly specified regression as four sub-questions (a) to (d).

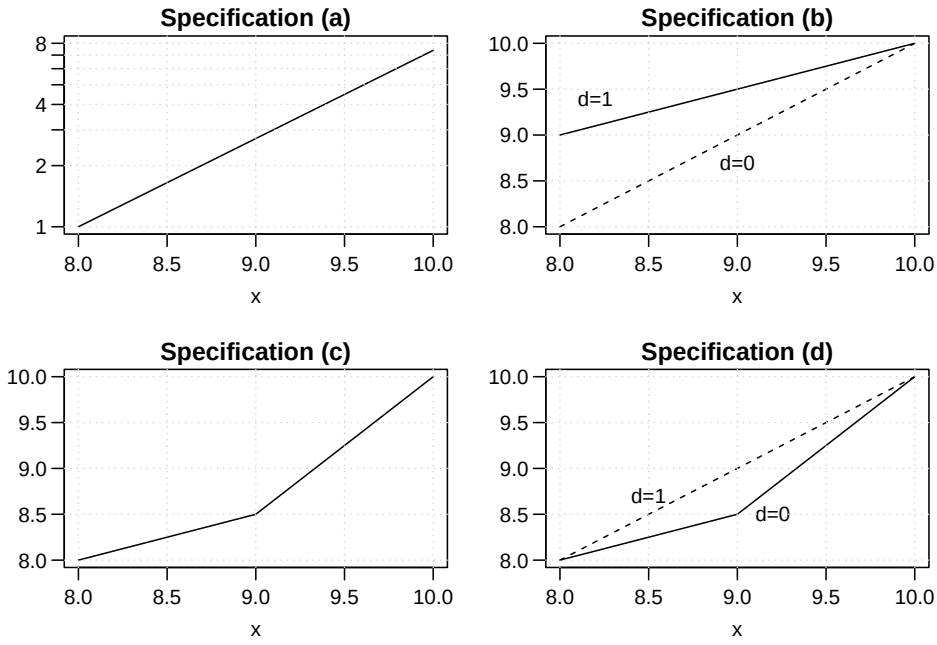

Figure 2: Regression specifications

ANSWER HINT:

```
# 2(a)
feols(log(y) ~ x, df)
# 2(b)
feols(y ~ d + x + d:x, df)
# 2(c)
feols(y ~ x + I(x-9):I(x>9), df)
# 2(d)
feols(y ~ d + d:x + (1-d):x + (1-d):I(x-9):I(x>9), df)
```

3. [8 points] You have access to data on housing prices and urban green space coverage in various neighborhoods. The relevant variables in your dataset are the sale price of homes in 1000s USD (`price`), the average neighborhood income in 1000s USD (`income_level`), the percentage of neighborhood area designated as green space (`green_space`), the number of bedrooms in the home (`bedrooms`), and the age of the home in years (`age`):

```
##                    mean    sd
## green_space       50.02  0.99
## bedrooms           3.07  0.97
## age               14.92  1.02
## income_level       4.09  0.96
## price             99.00 30.52
## log_income_level   1.38  0.27
## log_price          4.55  0.29
```

You estimated the following models:

```
##                               reg1          reg2           reg3           reg4
## Dependent Var.:         log(price)    log(price)     log(price)          price
##
## green_space         0.074 (0.020) 0.052 (0.022)  0.011 (0.022)    5.23 (2.29)
## log(income_level)                 0.200 (0.079)  0.087 (0.077)
## bedrooms                                         0.132 (0.024)
## age                                             -0.071 (0.019)
## income_level                                                      16.7 (12.0)
## income_level square                                              -1.37 (1.45)
## Constant             0.845 (1.01)   1.65 (1.05)    4.51 (1.08) -206.7 (112.1)
## ------------------- ------------- ------------- -------------- --------------
## S.E. type                     IID           IID            IID            IID
## R2                          0.063         0.092          0.232          0.090
## Observations                  200           200            200            200
```

a. Interpret the coefficient on `green_space` in `reg1`.

ANSWER HINT: A one percentage point increase in neighborhood green space is associated with 7 percent higher sales prices.

b. Construct and interpret the 90% confidence interval on `green_space` in `reg1`.

ANSWER HINT: $CI_{90\%} = 0.07 \pm 1.64 * 0.02 = (0.0372, 0.1028)$. We reject the null that the coef is zero, and say that it is statistically significant at the 10 percent level. Intervals constructed like this contain the true parameter 90 percent of the time across (infinately many) samples.

c. Interpret the coefficient of `log(income_level)` in `reg2`.

ANSWER HINT: This the income elasticity of prices which tells us that a 10 percent increase in income is associated with a 2 percent higher sales price, adjusting for green spaces in the neighborhood.

d. Explain why the inclusion of `bedrooms` and `age` in `reg3` reduces the estimated relationship between green space and property values compared to `reg2`, and explain how this changes your interpretation of that coefficient?

ANSWER HINT: We now need to interpret the coefficient as also adjusting for the size (`bedrooms`) and vintage (`age`) of the housing stock. We see that this matters a lot as the

coefficient drops from 0.052 to 0.011 which is explained by omitted variables bias that arises because `green_space` correlates with `bedrooms` and `age`. The drop suggests that more green neighborhoods had also larger and/or more new buildings.

e. Compute the F-statistic that tests the joint nullity of `bedrooms` and `age` in `reg3`. When would you reject this null-hypothesis with 90% confidence?

ANSWER HINT: $F = ((0.232 - 0.092)/2)/((1 - 0.232)/(200 - 5)) = 17.77344$. We reject the null if $F > 2.3026$.

f. Explain how the coefficients in `reg3` would change if housing prices were measured in USD rather than 1000s USD.

ANSWER HINT: $log(price * 1000) = log(price) + log(1000)$ and therefore only the intercept is affected.

g. What is the correlation between `green_space` and `log(income_level)` (Hint: OVB formula)?

ANSWER HINT: From `reg1` and `reg2` and the OVB formula we know that $0.074 = 0.052 + 0.2cov(green\_space, log(income\_level))/var(green\_space)$. This gives

$$cor(green\_space, log(income\_level)) = ((0.074 - 0.052)/0.2) * 0.99/0.27 = 0.4$$

h. What point estimates would you obtain if you were to estimate the following regression:

- `feols(price ~ green_space + I(income_level-100) + I((income_level-100)^2), housing_data)`

ANSWER HINT:

$$\begin{aligned} price &= \beta_0 + \beta_1 x_1 + \beta_2(x_2 - 100) + \beta_3(x_2 - 100)^2 + e \\ &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 - 100\beta_2 + \beta_3 x_2^2 - 200\beta_3 x_2 + 10000\beta_3 + e \\ &= (\beta_0 - 100\beta_2 + 10000\beta_3) + \beta_1 x_1 + (\beta_2 - 200\beta_3)x_2 + \beta_3 x_2^2 + e \end{aligned}$$

Which shows that the coefficients on `green_space` and the quadratic term `I((income_level-100)^2)` are unchanged. The coefficient on `I(income_level-100)` is $\beta_2 - 200 * -1.37 = 16.7$ wich gives $\beta_2 = -257.3$, and the intercept becomes $\beta_0 - 100 * -257.3 + 10000 * -1.37 = -206.7$ which gives $\beta_0 = -12236.7$.

---

4. [6 points] A company implemented a policy allowing employees to work remotely if their performance evaluation score exceeded 75 (not all employees who could actually did that). You have a dataset with employee performance scores and productivity metrics:
   - `productivity`: Output measure (units produced)
   - `remote_work`: Whether the employee works remotely
   - `evaluation_score`: Score on the baseline performance evaluation
   - `gender`: Gender of the employee
   - `age`: Age of the employee in years

   You wish to evaluate the effect of working remotely on productivity.

   a. Explain the method you would use and detail the regression(s) you would estimate to obtain a causal effect of remote work on productivity.

      ANSWER HINT: This is an example of a regression discontinuity design as treatment –eligibility for working remotely– is assigned on passing a threshold. Because not everybody makes use of that possiblity we are looking at a so-called fuzzy design. We know that we need to estimate this using an instrumental variable approach, and more specifically 2SLS. Our outcome equation (second stage) will look as follows

      $$productivity = \beta_0 + \beta_1 remote\_work + \beta_2 evaluation_score + e$$

      and the first-stage

      $$remote\_work = \pi_0 + \pi_1 1[evaluation_score > 75] + \pi_2 evaluation_score + e$$

      where $1[evaluation_score > 75]$ is the instrument an equals 1 if the evaluation score exceeds 75 and is zero otherwise. It is important to control for a continuous function of the assignment score. Here I chose a linear specification.

   b. How would you interpret your estimate in (a), making sure to explain the necessary assumptions underlying your interpretation.

      ANSWER HINT: I interpret the resulting estimate of β_2 as the average causal effect of working remotely on productivity for people whose evaluation score equals 75. It is therefore local. In the presence of heterogeneous effects the interpretation changes because then it is the average causal effect of people whose evaluation score equals 75 and who would work remotely if eligible, and who would not do so otherwise (i.e. those complying with the eligibility). The key assumption is that potential outcomes are continuous in the assignment score. (We also need a first stage, and in the heterogeneous effects case must rule out so-called defiers). We also need instrument relevance.

   c. Discuss a potential shortcoming of your model in (a) and outline how you would investigate this.

      ANSWER HINT: There are a potential number of shortcomings. I can have mispecified the functional form in my 2SLS specification above. I assumed linearity, but I can investigate this by estimating more flexible models. For example quadratif or cubic ones, or models involving splinces. Another possible violation is that of continuity. I can investigate bunching on the assignment variable by looking at a histogram of `evaluation_score`, or investigate whether people on both sides of the assignment threshold are comparable in terms of gender and age. I can do this using OLS by regressing these variables on the instrument and controls for the running variable. I can also add them to my model above to check whether my results are sensitive to their inclusion.

Critical Values for the $F_{m,\infty}$ Distribution

```
## Rows denote degrees of freedom (m), and columns significance level (%)
##          10%     5%     1%
##
##  1 :  2.7055 3.8415 6.6349
##  2 :  2.3026 2.9957 4.6052
##  3 :  2.0838 2.6049 3.7816
##  4 :  1.9449 2.3719 3.3192
##  5 :  1.8473 2.2141 3.0173
##  6 :  1.7741 2.0986 2.8020
##  7 :  1.7167 2.0096 2.6393
##  8 :  1.6702 1.9384 2.5113
##  9 :  1.6315 1.8799 2.4073
##  10 :  1.5987 1.8307 2.3209
##  11 :  1.5705 1.7886 2.2477
##  12 :  1.5458 1.7522 2.1847
##  13 :  1.5240 1.7202 2.1299
##  14 :  1.5046 1.6918 2.0815
##  15 :  1.4871 1.6664 2.0385
##  16 :  1.4714 1.6435 2.0000
##  17 :  1.4570 1.6228 1.9652
##  18 :  1.4439 1.6038 1.9336
##  19 :  1.4318 1.5865 1.9048
##  20 :  1.4206 1.5705 1.8783
##  21 :  1.4102 1.5557 1.8539
##  22 :  1.4006 1.5420 1.8313
##  23 :  1.3916 1.5292 1.8104
##  24 :  1.3832 1.5173 1.7908
##  25 :  1.3753 1.5061 1.7726
##  26 :  1.3678 1.4956 1.7554
##  27 :  1.3608 1.4857 1.7394
##  28 :  1.3541 1.4763 1.7242
##  29 :  1.3478 1.4675 1.7099
##  30 :  1.3419 1.4591 1.6964
```

The Cumulative Standard Normal Distribution Function, $\Phi(z) \equiv \Pr(Z \leq z)$

```
## Rows denote 1st decimal value of z, and columns 2nd decimal value of z

## So for example, P(Z <= 1.08) = 0.86
```

	0	1	2	3	4	5	6	7	8	9
0.0 :	0.500	0.504	0.508	0.512	0.516	0.520	0.524	0.528	0.532	0.536
0.1 :	0.540	0.544	0.548	0.552	0.556	0.560	0.564	0.567	0.571	0.575
0.2 :	0.579	0.583	0.587	0.591	0.595	0.599	0.603	0.606	0.610	0.614
0.3 :	0.618	0.622	0.626	0.629	0.633	0.637	0.641	0.644	0.648	0.652
0.4 :	0.655	0.659	0.663	0.666	0.670	0.674	0.677	0.681	0.684	0.688
0.5 :	0.691	0.695	0.698	0.702	0.705	0.709	0.712	0.716	0.719	0.722
0.6 :	0.726	0.729	0.732	0.736	0.739	0.742	0.745	0.749	0.752	0.755
0.7 :	0.758	0.761	0.764	0.767	0.770	0.773	0.776	0.779	0.782	0.785
0.8 :	0.788	0.791	0.794	0.797	0.800	0.802	0.805	0.808	0.811	0.813
0.9 :	0.816	0.819	0.821	0.824	0.826	0.829	0.831	0.834	0.836	0.839
1.0 :	0.841	0.844	0.846	0.848	0.851	0.853	0.855	0.858	0.860	0.862
1.1 :	0.864	0.867	0.869	0.871	0.873	0.875	0.877	0.879	0.881	0.883
1.2 :	0.885	0.887	0.889	0.891	0.893	0.894	0.896	0.898	0.900	0.901
1.3 :	0.903	0.905	0.907	0.908	0.910	0.911	0.913	0.915	0.916	0.918
1.4 :	0.919	0.921	0.922	0.924	0.925	0.926	0.928	0.929	0.931	0.932
1.5 :	0.933	0.934	0.936	0.937	0.938	0.939	0.941	0.942	0.943	0.944
1.6 :	0.945	0.946	0.947	0.948	0.949	0.951	0.952	0.953	0.954	0.954
1.7 :	0.955	0.956	0.957	0.958	0.959	0.960	0.961	0.962	0.962	0.963
1.8 :	0.964	0.965	0.966	0.966	0.967	0.968	0.969	0.969	0.970	0.971
1.9 :	0.971	0.972	0.973	0.973	0.974	0.974	0.975	0.976	0.976	0.977
2.0 :	0.977	0.978	0.978	0.979	0.979	0.980	0.980	0.981	0.981	0.982
2.1 :	0.982	0.983	0.983	0.983	0.984	0.984	0.985	0.985	0.985	0.986
2.2 :	0.986	0.986	0.987	0.987	0.987	0.988	0.988	0.988	0.989	0.989
2.3 :	0.989	0.990	0.990	0.990	0.990	0.991	0.991	0.991	0.991	0.992
2.4 :	0.992	0.992	0.992	0.992	0.993	0.993	0.993	0.993	0.993	0.994
2.5 :	0.994	0.994	0.994	0.994	0.994	0.995	0.995	0.995	0.995	0.995
2.6 :	0.995	0.995	0.996	0.996	0.996	0.996	0.996	0.996	0.996	0.996
2.7 :	0.997	0.997	0.997	0.997	0.997	0.997	0.997	0.997	0.997	0.997
2.8 :	0.997	0.998	0.998	0.998	0.998	0.998	0.998	0.998	0.998	0.998
2.9 :	0.998	0.998	0.998	0.998	0.998	0.998	0.998	0.999	0.999	0.999

The Inverse of the Cumulative Standard Normal Distribution Function, $\Phi^{-1}(q)$

```
## Rows denote 1st decimal value of q, and columns 2nd decimal value of q

## So for example for P(Z <= z) = 0.86, z is (approximately) 1.08
```

	0	1	2	3	4	5	6	7	8	9
0.0 :	-Inf	-2.33	-2.05	-1.88	-1.75	-1.64	-1.55	-1.48	-1.41	-1.34
0.1 :	-1.28	-1.23	-1.17	-1.13	-1.08	-1.04	-0.99	-0.95	-0.92	-0.88
0.2 :	-0.84	-0.81	-0.77	-0.74	-0.71	-0.67	-0.64	-0.61	-0.58	-0.55
0.3 :	-0.52	-0.50	-0.47	-0.44	-0.41	-0.39	-0.36	-0.33	-0.31	-0.28
0.4 :	-0.25	-0.23	-0.20	-0.18	-0.15	-0.13	-0.10	-0.08	-0.05	-0.03
0.5 :	0.00	0.03	0.05	0.08	0.10	0.13	0.15	0.18	0.20	0.23
0.6 :	0.25	0.28	0.31	0.33	0.36	0.39	0.41	0.44	0.47	0.50
0.7 :	0.52	0.55	0.58	0.61	0.64	0.67	0.71	0.74	0.77	0.81
0.8 :	0.84	0.88	0.92	0.95	0.99	1.04	1.08	1.13	1.17	1.23
0.9 :	1.28	1.34	1.41	1.48	1.55	1.64	1.75	1.88	2.05	2.33