Lecture 10: Inference in regression models

Pierre Biscaye

Fall 2022

Course feedback

- 53 responses - thank you!
- Generally positive
- Some points to incorporate going forward
- Show in Jupyter (not posted on Data Hub)

Inference recap

- Can use CLT to know $t = \frac{\bar{X}-\mu}{s_x/\sqrt{n}} \sim t_{n-1}$ (or $Z \sim N(0,1)$ if we know σ_x^2)
- Use this to conduct probabilistic inference
- Confidence intervals: with a 95% confidence level, μ will be inside the calculated CI for 95% of samples.
  - CI=$[\bar{x} - c * s_x/\sqrt{n}, \bar{x} + c * s_x/\sqrt{n}]$
- Hypothesis testing: evaluate whether our sample estimate is consistent with the population parameter taking a particular value
  1. State the null and alternate hypotheses
  2. Formulate a test statistic with a known distribution (Z or t)
  3. Choose a significance level α and identify critical values
  4. State the rejection rule
  5. Implement test and interpret the outcome
- Examples in Jupyter with student feedback data
- Any MLR concerns with implementing these tests?

Critical values for t and Z

t Table

cum. prob	$t_{.50}$	$t_{.75}$	$t_{.80}$	$t_{.85}$	$t_{.90}$	$t_{.95}$	$t_{.975}$	$t_{.99}$	$t_{.995}$	$t_{.999}$	$t_{.9995}$
one-tail	**0.50**	**0.25**	**0.20**	**0.15**	**0.10**	**0.05**	**0.025**	**0.01**	**0.005**	**0.001**	**0.0005**
two-tails	**1.00**	**0.50**	**0.40**	**0.30**	**0.20**	**0.10**	**0.05**	**0.02**	**0.01**	**0.002**	**0.001**
df											
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	318.31	636.62
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	0.000	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0.000	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.000	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.000	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.000	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0.000	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.000	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.000	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0.000	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.000	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	0.000	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0.000	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	0.000	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	0.000	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	0.000	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	0.000	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	0.000	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	0.000	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	0.000	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	0.000	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	0.000	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	0.000	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	0.000	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	0.000	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	0.000	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	0.000	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	0.000	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	0.000	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	0.000	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	0.000	0.679	0.848	1.045	1.296	1.671	2.000	2.390	2.660	3.232	3.460
80	0.000	0.678	0.846	1.043	1.292	1.664	1.990	2.374	2.639	3.195	3.416
100	0.000	0.677	0.845	1.042	1.290	1.660	1.984	2.364	2.626	3.174	3.390
1000	0.000	0.675	0.842	1.037	1.282	1.646	1.962	2.330	2.581	3.098	3.300
z	0.000	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.090	3.291
	0%	50%	60%	70%	80%	90%	95%	98%	99%	99.8%	99.9%
	Confidence Level										

Mapping inference to regression models

- We now know how to form confidence intervals and test hypotheses for sample means and mean differences.
- How does this map to regression estimates?
  - We know $E[\hat{\beta}_j] = \beta_j$ if MLR1-MLR4 are true
  - We know $var(\hat{\beta}_j|X) = \frac{\sigma_u^2}{(1-R_j^2)SST_j}$
  - Suppose we wanted to test

$$H_0 : \beta_j = 0 \tag{1}$$
$$H_1 : \beta_j \neq 0 \tag{2}$$

- What would we need to know to test this?

Distribution for simple regression

- Precise distributions messy to calculate for MLR
- Math more straightforward for SLR and intuition for MLR will be the same
- Suppose $y_i = \beta_0 + \beta_1 X_i + u_i$

$$\hat{\beta_1} = \frac{\widehat{cov(X,Y)}}{\widehat{var(X)}} = \frac{\widehat{cov(\beta_0 + \beta_1 X + u, X)}}{\widehat{var(X)}} \tag{3}$$

$$\hat{\beta_1} = \frac{\widehat{cov(\beta_0, X)}}{\widehat{var(X)}} + \beta_1 \frac{\widehat{cov(X,X)}}{\widehat{var(X)}} + \frac{\widehat{cov(u,X)}}{\widehat{var(X)}} \tag{4}$$

$$\hat{\beta_1} = \beta_1 + \frac{\sum(X_i - \bar{X})u_i}{\sum(X_i - \bar{X})^2} \tag{5}$$

Why is this useful?

- $\hat{\beta_1} = \beta_1 + \frac{\sum_i (X_i - \bar{X}) u_i}{\sum_i (X_i - \bar{X})^2}$
- Can think of $\hat{\beta_1}$ as a random variable.
- β_1 is a constant.
- $\frac{\sum_i (X_i - \bar{X})}{\sum_i (X_i - \bar{X})^2}$ is non-random because we observe the X variables.
- The unobserved u_i is what generates the random variation.
- This implies that $\hat{\beta_1}$ will follow the distribution of u_i.

MLR6: Distribution of u_i

- SLR/MLR4: $E(u|x) = 0$
- SLR/MLR5: $Var(u|x) = \sigma_u^2$
- **SLR/MLR6**: The population error u is independent of the explanatory variables $X_1, X_2, ..., X_k$ and is *normally distributed* with mean 0 and variance σ_u^2. That is, $u \sim N(0, \sigma_u^2)$.
  - A strong assumption, but valuable.

Theorem: Sampling distribution for $\hat{\beta}$

- $\hat{\beta_1} = \beta_1 + \frac{\sum_i (X_i - \bar{X}) u_i}{\sum_i (X_i - \bar{X})^2}$
- Under SLR/MLR1-4: $E[\hat{\beta_1}] = \beta_1$
- MLR4-6: $u \sim N(0, \sigma_u^2)$.
- Theorem: under MLR1-6, $\hat{\beta_1} \sim N(\beta_1, var(\hat{\beta_1}))$.

Theorem gives us a test statistic

$$\hat{\beta_1} \sim N(\beta_1, var(\hat{\beta_1})) \tag{6}$$

$$\frac{\hat{\beta_1} - \beta_1}{sd(\hat{\beta_1})} \sim N(0,1) \tag{7}$$

- Of course, we don't know σ_u^2 so we don't know $sd(\hat{\beta_1})$:

$$\frac{\hat{\beta_1} - \beta_1}{s.e.(\hat{\beta_1})} \sim t_{n-k-1} \tag{8}$$

- Can use for CIs and hypothesis testing with MLR $\hat{\beta}$s, as long as MLR1-6 hold

Should we be concerned over MLR6?

- Lots of things in u
- Tempting to rely on a CLT-type argument - we are averaging over lots of different variables.
  - Should it be normal then? Maybe, maybe not - depends on joint distribution of those variables.
- Luckily, not too important. If n is large, under the CLT and MLR1-5

$$\frac{\hat{\beta}_j - \beta_j}{sd(\hat{\beta}_j)} \sim\approx N(0,1) \tag{9}$$

$$\frac{\hat{\beta}_j - \beta_j}{s.e.(\hat{\beta}_j)} \sim\approx t_{n-k-1} \tag{10}$$

- Then, since the t_n distribution converges to the standard normal as n is large

$$\frac{\hat{\beta}_j - \beta_j}{s.e.(\hat{\beta}_j)} \sim\approx N(0,1) \tag{11}$$

Testing hypotheses on $\hat{\beta}$: Example

$$ln(wage_i) = \beta_0 + \beta_1 Educ_i + \beta_2 Exper_i + \beta_3 Tenure_i + u_i \qquad (12)$$

- What is the interpretation of β_2?
- Suppose we test $H_0 : \beta_2 = 0$ against $H_1 : \beta_2 \neq 0$. What economic hypothesis is this associated with?

To Jupyter!

Our regression results allows us to to test this hypothesis

$$\hat{\beta_2} = 0.004$$
$$s.e.(\hat{\beta_2}) = 0.0017$$

- How do we test $H_0 : \beta_2 = 0$?
- Need to calculate our test statistic

$$t = \frac{\hat{\beta_2} - \beta_2}{s.e.(\hat{\beta_2})} = \frac{\hat{\beta_2}}{s.e.(\hat{\beta_2})} = \frac{0.004}{0.0017} = 2.39$$

- R regression output automatically displays t statistics for the hypotheses $\beta_j = 0$
- This is what we are most often interested in testing

Find a critical value

- Choose a significance level: 95%
- $t \sim t_{n-k-1} = t_{526-3-1} = t_{522}$
- Degrees of freedom $n - k - 1$ also shown in R regression output

t Table

cum. prob	$t_{.50}$	$t_{.75}$	$t_{.80}$	$t_{.85}$	$t_{.90}$	$t_{.95}$	$t_{.975}$	$t_{.99}$	$t_{.995}$	$t_{.999}$	$t_{.9995}$
one-tail	**0.50**	**0.25**	**0.20**	**0.15**	**0.10**	**0.05**	**0.025**	**0.01**	**0.005**	**0.001**	**0.0005**
two-tails	**1.00**	**0.50**	**0.40**	**0.30**	**0.20**	**0.10**	**0.05**	**0.02**	**0.01**	**0.002**	**0.001**
df											
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	318.31	636.62
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	0.000	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0.000	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.000	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.000	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.000	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0.000	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.000	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.000	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0.000	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.000	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	0.000	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0.000	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	0.000	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	0.000	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	0.000	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	0.000	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	0.000	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	0.000	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	0.000	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	0.000	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	0.000	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	0.000	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	0.000	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	0.000	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	0.000	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	0.000	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	0.000	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	0.000	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	0.000	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	0.000	0.679	0.848	1.045	1.296	1.671	2.000	2.390	2.660	3.232	3.460
80	0.000	0.678	0.846	1.043	1.292	1.664	1.990	2.374	2.639	3.195	3.416
100	0.000	0.677	0.845	1.042	1.290	1.660	1.984	2.364	2.626	3.174	3.390
1000	0.000	0.675	0.842	1.037	1.282	1.646	1.962	2.330	2.581	3.098	3.300
z	0.000	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.090	3.291
	0%	50%	60%	70%	80%	90%	95%	98%	99%	99.8%	99.9%
						Confidence Level					

Test the null

- $t = 2.39 > 1.97$
- So we infer that...
- Note that this conclusion holds only if MLR1-4 are valid in this case; otherwise we are testing using biased $\hat{\beta}$

Statistical and *economic* significance

- Our null hypothesis is equivalent to testing: is there a *statistically significant* effect of experience on earnings?
- We interpret our test as indicating that the relationship is *statistically significant*, meaning it is statistically different from 0 at a particular significance level
- *Economic significance*: Is our effect large enough to care about?
  - An additional year of experience increases wages by 0.4%
  - Seems kind of small? An additional year of school increases wages by 9.2% (22x)
- Estimates can be statistically significant but not economically significant, or vice versa

Reading statistical significance off of R regression output

- R regression output (and other software regression packages) includes results from testing the null that $\beta_j = 0$
- t value is the test statistic
- R includes * symbols to indicate statistical significance at different levels: 0.001, 0.01, 0.05, and 0.1
- Can read *p values* in $Pr(>|t|)$ column which tell you the probability of observing a value greater than $|t|$ under the null: low probability means very unlikely
  - Note that these are two-tailed tests and probabilities
- Can compare these p values directly to your target significance level to evaluate the null hypothesis, without having to look up a critical value to compare against the t value

Example: Class size and test scores

- Do students learn *less* when they are in larger classes?
- Test scores are imperfect, but can be informative as to student learning
- We have data from 408 high schools in Michigan on the percentage of students that passed a 10th grade math test ($math10$), total student enrollment ($enroll$), and other school characteristics
- We could regress $math10_i = \beta_0 + \beta_1 enroll_i + u_i$
  - Would SLR4 hold?
  - What would you like to control for?

Statistical model

- Suppose we estimate the statistical model
- $math10_i = \beta_0 + \beta_1 enroll_i + \beta_2 totcomp_i + \beta_3 staff_i + u_i$
- *totcomp* is average total teacher compensation (salary and benefits) and *staff* is number of staff per 1,000 students
- We want to test if students learn *less* when they are in larger classes.
- Set up null hypothesis so that rejecting it provides statistical support for what we want to test.
- What is H_0?
- What is H_1?

Determine test statistic and critical value

- $H_0 : \beta_1 \geq 0$
- Recall that test statistics for one-sided hypothesis tests are calculated at equality under the null (the strictest test we could impose)

$$t = \frac{\hat{\beta_1}}{s.e.(\hat{\beta_1})} \sim t_{404} \tag{13}$$

- What is our critical value for a 95% confidence level?
- Graph on board

t Table

cum. prob	$t_{.50}$	$t_{.75}$	$t_{.80}$	$t_{.85}$	$t_{.90}$	$t_{.95}$	$t_{.975}$	$t_{.99}$	$t_{.995}$	$t_{.999}$	$t_{.9995}$
one-tail	**0.50**	**0.25**	**0.20**	**0.15**	**0.10**	**0.05**	**0.025**	**0.01**	**0.005**	**0.001**	**0.0005**
two-tails	**1.00**	**0.50**	**0.40**	**0.30**	**0.20**	**0.10**	**0.05**	**0.02**	**0.01**	**0.002**	**0.001**
df											
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	318.31	636.62
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	0.000	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0.000	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.000	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.000	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.000	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0.000	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.000	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.000	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0.000	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.000	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	0.000	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0.000	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	0.000	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	0.000	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	0.000	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	0.000	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	0.000	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	0.000	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	0.000	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	0.000	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	0.000	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	0.000	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	0.000	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	0.000	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	0.000	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	0.000	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	0.000	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	0.000	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	0.000	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	0.000	0.679	0.848	1.045	1.296	1.671	2.000	2.390	2.660	3.232	3.460
80	0.000	0.678	0.846	1.043	1.292	1.664	1.990	2.374	2.639	3.195	3.416
100	0.000	0.677	0.845	1.042	1.290	1.660	1.984	2.364	2.626	3.174	3.390
1000	0.000	0.675	0.842	1.037	1.282	1.646	1.962	2.330	2.581	3.098	3.300
z	0.000	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.090	3.291
	0%	50%	60%	70%	80%	90%	95%	98%	99%	99.8%	99.9%

Confidence Level

Implement our test

To Jupyter!

Interpret our test

$$t = \frac{\hat{\beta_1}}{s.e.(\hat{\beta_1})} = \frac{-0.0002}{0.0002} = -0.918 > -1.65 \qquad (14)$$

- Plot t on the board graph
- Interpretation?

In a MLR framework, can conduct other hypothesis tests

- $math10_i = \beta_0 + \beta_1 enroll_i + \beta_2 totcomp_i + \beta_3 staff_i + u_i$
- We can use the same regression output to test other hypotheses
- E.g., does teacher compensation matter for test performance?
- What is H_0 and what is H_1?
- Implement test with R output

To Jupyter!

Are these patterns causal? Key question

- Our statistical model said that there was a positive statistically significant relationship between teacher compensation and student test scores.
- But we could find no statistical evidence of a relationship between class size and test scores.
- What must be true for these estimates to be causal?

Remember MLR1-MLR4

- Suppose MLR2-4 hold.
- We estimated $math10_i = \beta_0 + \beta_1 enroll_i + \beta_2 totcomp_i + \beta_3 staff_i + u_i$
- Consider MLR 1: what if in the population $math10_i = \beta_0 + \beta_1 log(enroll_i) + \beta_2 log(totcomp_i) + \beta_3 log(staff_i) + u_i$

To Jupyter!

Testing the hypothesis $\beta_1 \geq 0$ in logs

$$t = \frac{\hat{\beta_1}}{s.e.(\hat{\beta_1})} = \frac{-1.268}{0.69} = -1.83 < -1.65 \tag{15}$$

- What do we conclude?
- What about the effects of total compensation?

Wrapping up on MLR1

- Being able to reject the null hypothesis *is not* evidence in favor of the log specification.
    - It only says that if in the population, we think a 1% increase in enrollment is associated with a decrease in test scores, we have evidence for it.
    - If the number of students is what matters and not percent increases, we do not have evidence for it.
    - In some cases, theory may guide appropriate choice.
    - One piece of evidence: the $\bar{R}^2$ increases going to the level-log specification: level-log model fits the data better.
    - More on this later.
- The effect of total compensation mattered in either specification
    - May consider this result to be more robust if we are unsure of the true functional form.
- But what about MLR4 - is this correlation causal?
    - Are there variables in u that could be driving this relationship?

Testing MLR4

- Suppose MLR1-3 hold.
- We estimated
  $math10_i = \beta_0 + \beta_1 enroll_i + \beta_2 totcomp_i + \beta_3 staff_i + u_i$
- What if students' household income levels are associated with these X variables?

To Jupyter!

Interpreting results

- Estimating $math10_i = \beta_0 + \beta_1 enroll_i + \beta_2 totcomp_i + \beta_3 staff_i + \beta_4 lnchprg_i + u_i$
- *totcomp* still statistically significant, but less so and smaller magnitude
  - Result seems pretty robust, but may still be explained by something else in u.
  - Think hard before concluding that policymakers should just start paying teachers more to increase test scores.
- What changed with this new regression specification?
  - Magnitude of β_3 is now larger, and have $t = 1.7834$
  - What can we conclude about $H_0 : \beta_3 = 0$ in this case?
  - What does this imply about relationship between *lnchprg* and *staff*?
- In both cases, still not quite convinced we can treat relationships as causal. Future classes will talk about how we can get to this point.