STAT 415/615 Midterm Exam

Name: _____

Autumn 2022

Instructions

This exam contains 7 pages (including this cover page) and 13 questions, for a total of 50 points.

- This exam is closed book and closed notes.
- You do not need a calculator. Just leave equations in their unevaluated form. For example, if the solution is "2+4", then leave it as "2+4" and do not evaluate it to "6".
- If you get stuck on one question, just move on to the next.
- Good luck!

This exam concerns the data from Cesaretti et al. (2020). Modern economics says that urban centers are more efficient, producing more economic output per individual than rural areas. This is due to a variety of reasons (easier logistics, larger labor networks, etc). Cesaretti et al. (2020) wanted to test if this was also true in pre-modern cities. They collected data from $n = 93$ towns during the 1524 and 1525 tax seasons in Tudor England. They used the total tax revenue of a town as a proxy for economic output, and number of taxpayers as a proxy for town size. Thus, the two variables in the dataset were

- `taxpayers`: Number of taxpayers in the town.
- `tax`: The total tax collected by from the town, in British pounds £.

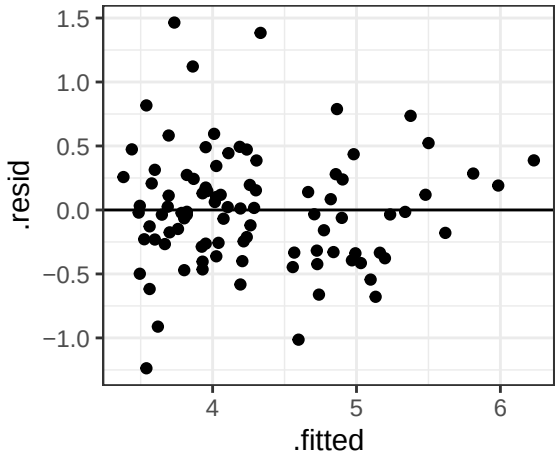
Here are some summaries for these variables:

Variable	min	Q1	med	Q3	max	mean	sd
tax	10	42	63	108	750	105.0	116.7
taxpayers	150	209	265	448	1409	355.9	228.2

The authors performed a log transformation on both `tax` and `taxpayers`. The authors then fit a regression model of log-tax (the response variable) on log-taxpayers (the explanatory variable), obtaining the following regression output (I censored some values with `NA`).

```
## # A tibble: 2 x 5
##   term          estimate std.error statistic      p.value
##   <chr>            <dbl>     <dbl>     <dbl>        <dbl>
## 1 (Intercept)      -3.00    0.535       NA    0.000000222
## 2 log_taxpayers     1.27    0.0931      13.7 NA
```

Here is a residual plot for this regression that the authors provide:



Here are some calculations which may or may not be useful (some are definitely *not* useful):

Code	value
qt(p = 0.9, df = 91)	1.29
qt(p = 0.95, df = 91)	1.66
qt(p = 0.975, df = 91)	1.99
log(1.27)	0.24
exp(1.27)	3.56
2^1.27	2.41
1.27^2	1.61

1. (5 pts) Write out the mathematical form for the simple linear regression model of log-tax on log-taxpayers. List all four assumptions in decreasing order of importance. You will have to define variables and parameters. Make sure you define variables on the original scale (not log-scale).

2. (4 pts) Do the assumptions of the linear model appear to be satisfied? List any possible issues. Also describe any possible solutions.

3. (7 pts) Provide an interpretation for the 1.27 coefficient estimate. Do not use causal language.

4. (3 pts) Fill in the value of the first `NA` from the regression table, corresponding to the `statistic` from the `(Intercept)` row.

5. (3 pts) Fill in the value of the second `NA` from the regression table, corresponding to the `p.value` from the `log_taxpayers` row. Your solution should be some R code.

6. (3 pts) What are the null and alternative hypotheses corresponding to the p-value from part 5? Use the same mathematical notation as your response from part 1.

7. (4 pts) The authors were mostly interested in if efficiency was much higher in larger towns than would be predicted based on a proportionate scaling model. Thus, they were interesting in testing if the slope coefficient of the linear model was 1. What is the t-statistic corresponding to this test? Leave it in its unevaluated form.

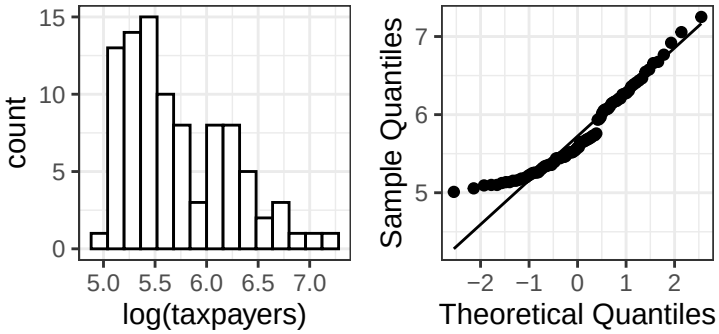
8. (4 pts) Is the p-value corresponding to the test from part 7 larger or smaller than the p-value corresponding to the test from part 5? **Explain**.

9. (4 pts) A specific theory of economics, called the "Settlement Scaling Theory", predicts that the slope of this regression should be 7/6. When testing this hypothesis, the researchers produced a p-value of 0.24.

   TRUE/FALSE and **explain**. This p-value indicates that the Tudor Tax data are consistent with the null hypothesis of the Settlement Scaling Theory.

10. (4 pts) A colleague wants a range of likely values of the slope for this regression model. Give them a 95% confidence interval for the slope parameter. No need to transform it back to the original scale.

11. (3 pts) A colleague provides the following histogram and QQ-plot of taxpayers (on the log-scale) and tells you that it's not normal, and so additional transformations are necessary for the linear model to be fully satisfied. What do you tell them?



12. (2 pts) In the below two scenarios, would you suggest calculating (i) confidence interval for the mean, (ii) prediction interval, or (iii) confidence bands for the regression line?

    a. York was not included in the data, and so researchers want a range of likely values of York's taxes given that they know York had 1160 taxpayers.

    b. What is the average tax amount for a city with 200 taxpayers?

13. (4 pts) A colleague gave you this ANOVA table

```
## Anova Table (Type II tests)
##
## Response: log_tax
##                Sum Sq Df F value Pr(>F)
## log_taxpayers   39.4  1     187 <2e-16
## Residuals       19.2 91
```

For each of the below, no need to evaluate expressions to a single numeric.

a. What is SSE?

b. What is SSR?

c. What is SSTO?

d. What is the estimated variance in the linear model?

References

Cesaretti, Rudolf, José Lobo, Luis M. A. Bettencourt, and Michael E. Smith. 2020. "Increasing Returns to Scale in the Towns of Early Tudor England." *Historical Methods: A Journal of Quantitative and Interdisciplinary History* 53 (3): 147–65. https://doi.org/10.1080/01615440.2020.1722775.

