Econ 57
Gary Smith
Spring 2023

Final Examination Answers

1. Regression to the mean

2. a. Matched-pair t-test.
   b. Two-sample difference-in-means t-test
   c. Multiple regression
   d. matched-pair t-test
   e. Two-sample difference-in-means t-test.

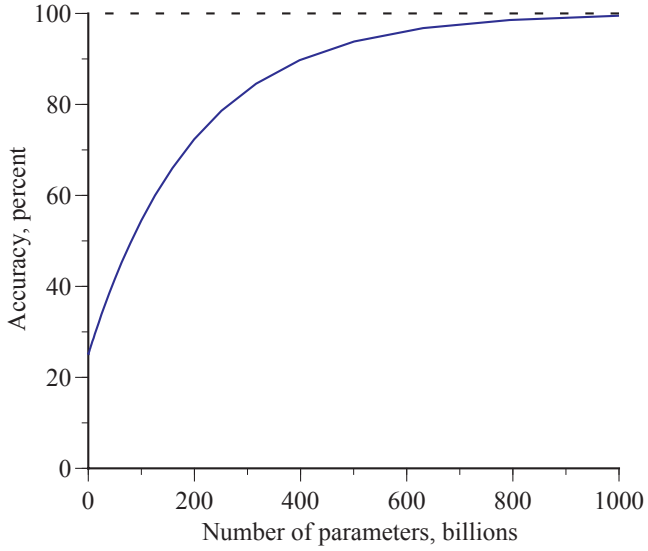
3. a. Matched-pair t-test
   b. Difference-in-proportions Z-test or chi-square test
   c. One-sample t-test
   d. simple regression
   e. Difference-in-proportions Z-test or chi-square test

4. a. HARKing
   b. p-hacking
   c. omitted-variable biased
   d. Self-selection bias.

5. The first card can be in any suit. After that the next 12 cards must be in the same suit as the first card:

$$\left(\frac{12}{51}\right)\left(\frac{11}{50}\right)\left(\frac{10}{49}\right)\cdots\left(\frac{1}{40}\right)=0.00000000000629907808979643$$

6. The use of logarithms on the horizontal axis. When the data are graphed with the number of parameters (not the log of the number of parameters) on the horizontal axis, the figure looks like this (with no tipping point to be found):



7. a. The Race variable should be a 0-1 dummy variable
   b. The 2-sided p-value for the constant should be much less than 0.05. (The correct 2-sided p-value is $3.85\ 10^{-14}$)
   c. The 2-sided p-value for Covid should not be less than 0.05. (0.0427 is the 1-sided p-value)
   d. The t-statistic for Race should be larger than 1. (The reported number if SD/b instead of b/SD.)
   e. There is no single estimate for ε

8. The answers are in bold:

Variable	Mean	Median	Standard Deviation	Minimum	Maximum
Earnings	33,830	**24,246**	27,846	2,343	130,000
Covid	**0.60**	0	0.50	0	1
Race	0.25	**0**	0.72	0	**3**

N = **37,897**

9. Regression toward the mean.

10. Think of a probability tree. The probability that the 3-person jury makes the correct decision is p(p) + p(1-p)(0.5) +(1-p)(p)0.5 = p(p + (1-p)(0.5) + (1-p)(0.5)) = p

11. Assuming that the guesses are independent, each with a 0.5 probability of being correct,
   a. $\mu = (1)(0.5) + (-2)(0.5) = -0.5$

   b. Charlie has to answer at least 6 questions to have any chance of getting 6 points. Guessing the answers to exactly 6 questions is best because every guess reduces the chances of a good score. If Charlie guesses at 6, 7, or 8 questions, Charlie has to get them all right to get an *A*. Charlie is more likely to get 6 of 6 than 7 of 7 or 8 of 8. If Charlie guesses at 9 questions, Charlie has to get 8 or 9 right to get an A and if Charlie guesses at all 10 questions, Charlie has to get 9 or 10 right to get an *A*.

$$P[6 \text{ of } 6] = \binom{6}{6} 0.5^6 0.5^{6-6} = 0.5^6$$

$$P[8 \text{ or } 9 \text{ of } 9] = \binom{9}{8} 0.5^8 0.5^{9-8} + \binom{9}{9} 0.5^9 0.5^{9-9} = 9(0.5^9) + 0.5^9 = 10(0.5^9)$$

$$P[9 \text{ or } 10 \text{ of } 10] = \binom{10}{9} 0.5^9 0.5^{10-9} + \binom{10}{10} 0.5^{10} 0.5^{10-10} = 10(0.5^{10}) + 0.5^{10} = 11(0.5^{10})$$

   The probability of an *F* grade with this strategy is the probability of getting 3 or fewer correct:

$$P[0, 1, 2, \text{or } 3 \text{ of } 6] = \binom{6}{0} 0.5^0 0.5^6 + \binom{6}{1} 0.5^1 0.5^5 + \binom{6}{2} 0.5^2 0.5^4 + \binom{6}{3} 0.5^3 0.5^3 = 0.656$$

12. a. No. We don't accept a null hypothesis; we fail to reject.
   b. No. For example, X might have a lower median but a larger outlier.
   c. Yes. An outlier must be far outside the box.
   d. No. Multicollinearity is when some of the explanatory variables are highly correlated.
   e. No. Oomph refers to the size of the coefficients, not the size of R^2.

13. The host's opening of a goldfish door did not affect the 1/4 probability that your initial choice has the $10,000 prize. So, if you switch, each of the two remaining doors has a (3/4)/2 = 3/8 probability of being the winning door.

This can be cast as a Bayes' Rule problem:

$$P[\text{3 wins if shows 1}] = \frac{P[\text{3 wins}]P[\text{shows 1 if 3 wins}]}{P[\text{3 wins}]P[\text{shows 1 if 3 wins}] + P[\text{2 wins}]P[\text{shows 1 if 2 wins}] + P[\text{4 wins}]P[\text{shows 1 if 4 wins}]}$$

$$= \frac{(1/4)(1/3)}{(1/4)(1/3)+(1/4)(1/2)+(1/4)(1/2)}$$

$$= \frac{1/3}{4/3}$$

$$\frac{(1/4)(1/3)}{(1/4)(1/3)+(1/4)(1/2)+(1/4)(1/2)}$$

$$= \frac{1}{4}$$

14. Simpson's Paradox. The birth rates may be higher in religions in which people are more religious.

15. a. Yes
    b. Yes
    c. No (There is about a 0.95 probability that X will be *less* than 2 standard deviations from μ
    d. No
    e. Yes

16. Each of the numerators should be 40, the expected value of the number of observations in that category.

17. Here is a contingency table with a total of 27,000 women who give birth at age 35, of whom (1/270)27,000 = 100 will suffer from Down syndrome:

	Positive Reading	Negative Reading	Total
Down	89	11	100
No Down	6,725	20,175	26,900
Total	6,814	20,186	27,000

    Of the 100 Down-syndrome babies, the blood test will give a positive reading in 0.89(100) = 89 cases and a negative reading in the remaining 11 cases. Of the 26,900 babies not suffering from Down syndrome, the blood test will give a negative reading in 0.75(26,900) = 20,175 cases and a positive reading in the remaining 6,725 cases.

    Of the 6,814 cases with positive readings, a stunning 6,725/6,814 = 0.987 are false positives. (The false negative rate is 11/ 20,186 = 0.0005; the main benefit of this blood test is that it can screen out many women from an amniotic-fluid test that risks a miscarriage. For those who get a positive blood-test result, a follow up amniotic-fluid test would be necessary.)

18. The coefficients are *ceteris paribus*. The coefficient of X_4 is is an estimate of the effect on home prices of a higher Walk Score, for a given square footage, number of bedrooms, and number of bathrooms.

19. O and E should be numbers, not fractions.

20. The 3-dimensional look doesn't help and may have confused the person who drew the figure. The year 1975 doesn't have an earnings number and the last column has no year label. The \$1.72/\$1.16 column was probably 1975 and the next two columns were probably 1976 and 1977.