Final - Spring 2025

ECON 5753 — University of Arkansas

1. The following regression uses the "College Scorecard" which describes all U.S. colleges/universities. The outcome variable is the average annual earnings ($) of students 10 years after they enroll. The explanatory variable is the median SAT Math score of the student body. I include both the variable itself and its square (quadratic in SAT math):

```
OLS estimation, Dep. Var.: mean_earnings_10yr_after
Observations: 935
Standard-errors: Heteroskedasticity-robust
                       Estimate   Std. Error  t value   Pr(>|t|)
(Intercept)           108369.50     19595.24     5.53   4.15e-08 ***
sat_math_median         -337.94        68.78    -4.91   1.05e-06 ***
I(sat_math_median^2)       0.41         0.06     6.88   1.07e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

   i. (5 pts) What is the predicted earnings for a school with an average SAT math score of 500 (round to the nearest dollar)?

   Answer: $\hat{y} = 108369.50 + -337.94 * 500 + 0.41 * 500^2 = \$41,899.50$

   ii. (5 pts) Say you take a school with an average SAT math score of 500. What is the predicted marginal change in Y for a school with a 1 unit increase in average SAT math score?

   Answer: $\frac{\partial \hat{y}}{\partial \texttt{sat\_math\_median}} = -337.94 + 2 * 0.41 * 500 = 72.06$

   iii. (5 pts) Can you reject the null that the relationship between SAT math score and average earnings is linear? Explain your reasoning.

   Answer: Since the coefficient on `sat_math_median`^{2} is statistically significantly different from zero, we can reject the null that the relationship is linear in favor of the quadratic alternative.

2. In class, we analyzed time-series data on the finishing time (in minutes) of the winner of the Boston Marathon. Below I report the coefficients from the following regression with piecewise linear trends:

$$\text{Finish Time}_t = \alpha + t\beta_1 + (t - 1950) * \mathbb{1}[t > 1950] * \beta_2 + (t - 1980) * \mathbb{1}[t > 1980] * \beta_3 + \epsilon_t$$

```
OLS estimation, Dep. Var.: minutes

Observations: 96

Standard-errors: Newey-West (L=2)

                                    Estimate  Std. Error  t value   Pr(>|t|)

(Intercept)                           772.67     223.620     3.46   8.23e-04 ***

year                                   -0.32       0.115    -2.78   6.51e-03 **

I((year - 1950) * (year > 1950))       -0.26       0.163    -1.61   0.11

I((year - 1980) * (year > 1980))        0.57       0.077     7.35   6.77e-11 ***

---

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

i. (10 pts) What is the expected finishing time in 2025? Show your work.

Answer: $\hat{y} = 772.67 - 0.32 * 2025 - 0.26 * (2025 - 1950) + 0.57 * (2025 - 1980) = 130.82$

ii. (10 pts) Notice that $\hat{\beta}_2$ is not statistically significant. What does this suggest about the time-trends in finishing times.

Answer: The estimated coefficient on $\hat{\beta}_2$ is not statistically significantly different from zero which means that we can not reject the null that the initial trend (estimated by $\hat{\beta}_1$) does not continue between 1950 and 1980.

iii. (5 pts) Looking into the future, how many minutes (or fractions of minutes) do we expect the finishing time to decrease each year?

Answer: The trend going into the future is given by $-0.32 + -0.26 + 0.57 = -0.01$ minutes per year.

3. Say you have a sample of stores where you observe the average daily revenue and the number of employees on the sales floor. You regress the average daily revenue on the number of

employees and estimate a coefficient of $\hat{\beta}_1 = 2247$ and a standard error of SE$(\hat{\beta}_1) = 980$.

i. (5 pts) Interpret this coefficient estimate in words.

   Answer: For every one additional worker a store has, the average daily revenue is expected to be $2247 higher on average.

ii. (10 pts) The company does not want to increase the number of staff if these results are not statistically significant. Perform and report a test of the null that $\beta_1 = 0$. The company is risk-adverse and want you to use a level of significance of $\alpha = 0.01$ (the z-score associated with this is 2.58).

   Answer: The t-test associated with this is given by $\frac{2247-0}{980} = 2.29$. Since this is smaller than the critical value of 2.58, we fail to reject the null at 1% significance.

4. You are an analyst at a large retail company. The retail company has a goal of "improving customer in-store experience". The company surveys a random sample of customer sentiment every week, asking customers to rate their in-store experience on a scale from 1 to 10.

   i. (10 pts) Explain why simply reporting the most recent week's average score might not be the best way to measure progress.

      Answer: Each week's average score reflects the true score μ_t and noise due to the fact only a random sample of customers is surveyed ε_t. It is hard to say whether a change from one week to the next is due to shifts in the true score or due to random sampling error. In other words, $\hat{\mu}_t = y_t$ responds too strongly to noise.

   ii. (10 pts) Suppose there is a sudden, genuine improvement in customer experience due to a new training intervention. Compare how quickly the 5-week moving average and the most recent week's score would reflect this change. Which would you recommend for reporting progress, and why?

      Answer: The 5-week moving average would take a few weeks for this sudden change to show up in $\hat{\mu}_t$ That is one primary cost of using a longer time-span in a moving average. The 1-week moving "average" responds immediately, but also is a very noisy measure. There is no correct answer at which is better, but in this case the most recent week will better pick up on the genuine sudden improvement.

   iii. (5 pts) Suppose the company wants to test whether the new training intervention has had a statistically significant impact on customer sentiment. Describe a time-series regression you could run to test if there is a statistically significant jump in responses after the intervention. Describe the hypothesis test you would use after running this regression.

      Answer: To test for a significant jump in average score after the intervention, the time-series regression should add an indicator for the post-period ($t >$ the intervention date) along with other useful time-series terms to predict systematic patterns. Then, a test that the coefficient on the post indicator can test for a statistically significant jump after the intervention date.

5. On the following page, I present time-series data on the log of the daily number of airline passengers (millions). Additionally, I present output from `feols` from a time-series regression. Use both the figure and the regression output to answer the following questions:

   i. (5 pts) Forecast the daily number of airline passengers for the January 1st in 1965 (note the outcome in the regression is logged, so you need to exponentiate). Show your work. What are you assuming when forecasting multiple years into the future?

      Answer: $\log \hat{y} = -230.75 + 0.12 * 1965 = 5.05$

      $\hat{y} = e^{5.05} = 156.02$ million passengers.

      This forecast is relying on the trend that we estimated in the time-series to continue into the years after 1960.

   ii. (5 pts) Interpret the coefficient on the `July` indicator. Form a 95% confidence interval for this coefficient and comment on its statistical significance.

      Answer: Days in July on average have about 36% more passengers compared to January. This difference is statistically significant at the 1% significance level. With 95% confidence, the true difference falls between $(0.36 - 1.96 * 0.023, 0.36 + 1.96 * 0.023) = (0.315, 0.405)$.

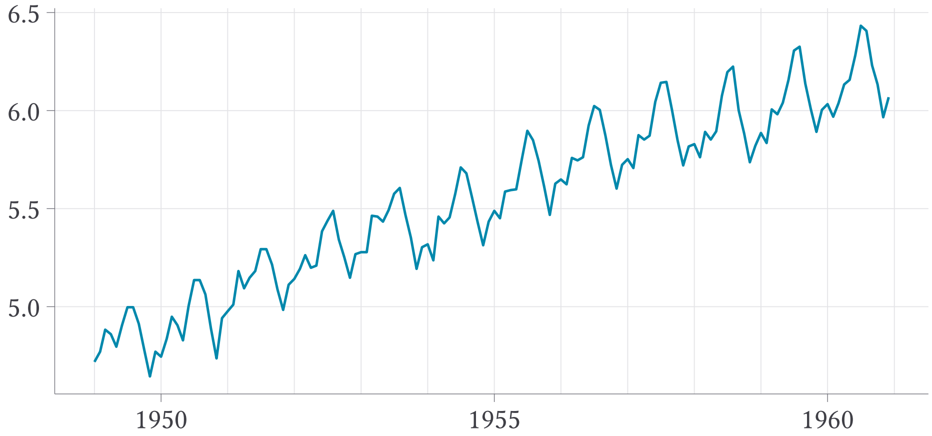
   iii. (5 pts) Describe how you might change this regression model to test if the number of passengers in June is statistically significantly larger than May.

      Answer: I would change the omitted category to be May and then look at the coefficient on June.

   iv. (5 pts) You might be concerned that your time-series model is missing some predictable patterns. Describe in words what you could do to visually assess that.

      Answer: I would plot the time-series of $y_t - \hat{y}_t$ where $\hat{y}_t$ is the predicted value from my current time-series regression. I would then visually inspect the residual time-series figure for any systematic patterns.

Figure 1 — log **Daily number of passengers (millions)**



```
OLS estimation, Dep. Var.: log(n_passengers)
Observations: 144
Standard-errors: Newey-West (L=2)
                  Estimate  Std. Error   t value   Pr(>|t|)
(Intercept)        -230.75       4.690    -49.20  < 2.2e-16 ***
year                  0.12       0.002     50.36  < 2.2e-16 ***
month::February      -0.01       0.020     -0.60   5.43e-01
month::March          0.13       0.023      5.46   1.99e-07 ***
month::April          0.10       0.023      4.59   9.53e-06 ***
month::May            0.11       0.024      4.70   5.91e-06 ***
month::June           0.25       0.024     10.43  < 2.2e-16 ***
month::July           0.36       0.023     15.76  < 2.2e-16 ***
month::August         0.36       0.023     16.20  < 2.2e-16 ***
month::September      0.23       0.020     11.15  < 2.2e-16 ***
month::October        0.10       0.020      4.92   2.31e-06 ***
month::November      -0.03       0.020     -1.76   7.93e-02 .
month::December       0.09       0.016      5.69   6.73e-08 ***
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```