ECON 5140 — HW1

Problem 1

Given

Model:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1(\text{CreditScore}), \quad p = P(\text{Default} = 1 \mid \text{CreditScore}) \tag{1}$$

Estimated coefficients:

- $\hat{\beta}_0 = 5.2$
- $\hat{\beta}_1 = -0.008$

---

(1) Compute p when CreditScore = 650

$$\log\left(\frac{p}{1-p}\right) = 5.2 - 0.008(650) = 5.2 - 5.2 = 0 \tag{2}$$

Convert log-odds to probability:

$$p = \frac{1}{1+e^{-0}} = \frac{1}{2} = 0.5 \tag{3}$$

Answer: $p = 0.50$

---

(2) Compute p when CreditScore = 750

$$\log\left(\frac{p}{1-p}\right) = 5.2 - 0.008(750) = 5.2 - 6.0 = -0.8 \tag{4}$$

$$p = \frac{1}{1+e^{-(-0.8)}} = \frac{1}{1+e^{0.8}} \approx \frac{1}{1+2.226} = 0.310 \tag{5}$$

Answer: $p \approx 0.310$

---

(3) Find CreditScore when $p = 0.5$

If $p = 0.5$, then:

$$\log\left(\frac{0.5}{1-0.5}\right) = \log(1) = 0 \tag{6}$$

So:

$$0 = 5.2 - 0.008(\text{CreditScore}) \Rightarrow 0.008(\text{CreditScore}) = 5.2 \Rightarrow \text{CreditScore} = 650 \tag{7}$$

Answer: CreditScore $= 650$

---

(4) Interpret $\hat{\beta}_1 = -0.008$ using odds ratios

Compare credit scores that differ by 100 points:

$$\log(\text{odds}_2) - \log(\text{odds}_1) = -0.008(100) = -0.8 \tag{8}$$

$$\frac{\text{odds}_2}{\text{odds}_1} = e^{-0.8} \approx 0.449 \tag{9}$$

Interpretation: A 100-point increase in credit score multiplies the odds of default by **0.449** (about a **55.1% decrease** in odds).

Problem 2

Given

Model:

$$\log(E[\text{Visits} \mid X]) = \beta_0 + \beta_1(\text{Ad\_Spend}) + \beta_2(\text{Weekend}) \tag{10}$$

Where:

- Ad_Spend is in **thousands of dollars**
- Weekend = 1 if weekend, 0 otherwise

Estimated coefficients:

- $\hat{\beta}_0 = 5.5$
- $\hat{\beta}_1 = 0.12$
- $\hat{\beta}_2 = 0.30$

---

(1) Expected visits when Weekend = 0 and Ad_Spend = $5000

Convert $5000 to thousands: 5

$$\log(E[\text{Visits}]) = 5.5 + 0.12(5) + 0.30(0) = 5.5 + 0.6 = 6.1 \tag{11}$$

$$E[\text{Visits}] = e^{6.1} \approx 446 \tag{12}$$

Answer: $E[\text{Visits}] \approx 446$

---

(2) Interpret $\hat{\beta}_1 = 0.12$

A $1,000 increase in Ad_Spend increases log expected visits by 0.12, so expected visits are multiplied by:

$$e^{0.12} \approx 1.1275 \tag{13}$$

Interpretation: Each additional $1,000 in ad spend increases expected visits by about **12.75%**.

---

(3) Interpret $\hat{\beta}_2 = 0.30$

Weekend vs weekday (holding Ad_Spend fixed):

$$e^{0.30} \approx 1.3499 \tag{14}$$

Interpretation: Weekends have about **35% more expected visits** than weekdays, holding ad spend constant.

---

(4) If $E[\text{Visits}] = 500$ on a weekday, solve for Ad_Spend

Weekday means Weekend = 0:

$$\log(500) = 5.5 + 0.12(\text{Ad\_Spend}) \tag{15}$$

$$\ln(500) \approx 6.2146 \tag{16}$$

$$6.2146 = 5.5 + 0.12(\text{Ad\_Spend}) \Rightarrow 0.7146 = 0.12(\text{Ad\_Spend}) \Rightarrow \text{Ad\_Spend} = 5.955 \tag{17}$$

Convert back to dollars:

$$5.955 \times 1000 \approx \$5955 \tag{18}$$

Answer: Ad_Spend $\approx \$5955$ on a weekday to reach 500 visits.

Problem 3

(a) Mean and Standard Deviation

The daily active users (in thousands) are:

$$Y = [45,\ 48,\ 50,\ 49,\ 52,\ 54,\ 53,\ 56,\ 58,\ 57]. \tag{19}$$

Mean

The mean is:

$$\mu = \frac{1}{10}\sum_{t=1}^{10} Y_t. \tag{20}$$

First compute the sum:

$$45 + 48 + 50 + 49 + 52 + 54 + 53 + 56 + 58 + 57 = 522. \tag{21}$$

Therefore,

$$\mu = \frac{522}{10} = 52.2. \tag{22}$$

Standard Deviation

Compute each squared deviation from the mean:

Y_T	Y_T-\MU	(Y_T-\MU)^2
45	-7.2	51.84
48	-4.2	17.64
50	-2.2	4.84
49	-3.2	10.24
52	-0.2	0.04
54	1.8	3.24
53	0.8	0.64
56	3.8	14.44
58	5.8	33.64
57	4.8	23.04

Sum of squared deviations:

$$\sum (Y_t - \mu)^2 = 159.6. \tag{23}$$

If we use the **sample standard deviation**:

$$\sigma = \sqrt{\frac{159.6}{9}} \approx \boxed{4.21}. \tag{24}$$

$$(If\ using\ the\ population\ formula, divide\ by\ 10\ instead, giving\ \sigma \approx 4.00.) \tag{25}$$

(b) Lag-1 autocorrelation

Create the two lagged series (length 9):

$$X = (Y_1, \ldots, Y_9) = [45, 48, 50, 49, 52, 54, 53, 56, 58] Z = (Y_2, \ldots, Y_{10}) = [48, 50, 49, 52, 54, 53, 56, 58, 57]$$

Compute their means:

$$\bar{X} = \frac{45 + 48 + 50 + 49 + 52 + 54 + 53 + 56 + 58}{9} = \frac{465}{9} = 51.6667 \bar{Z} = \frac{48 + 50 + 49 + 52 + 54 + 53}{9}$$

Now compute:

$$\rho(1) = \frac{\sum_{i=1}^{9} (X_i - \bar{X})(Z_i - \bar{Z})}{\sqrt{\sum_{i=1}^{9} (X_i - \bar{X})^2} \sqrt{\sum_{i=1}^{9} (Z_i - \bar{Z})^2}} \tag{28}$$

From the data:

$$-\sum (X_i - \bar{X})(Z_i - \bar{Z}) = 78.0000 - \sum (X_i - \bar{X})^2 = 133.9999 - \sum (Z_i - \bar{Z})^2 = 56.0000 \tag{29}$$

So:

$$\rho(1) = \frac{78}{\sqrt{134}\sqrt{56}} \approx \frac{78}{86.63} \approx 0.90 (More precisely, \rho(1) \approx 0.898.) \tag{30}$$

(c) What does this autocorrelation say about persistence?

A lag-1 autocorrelation of about 0.90 is strongly positive, meaning daily active users are highly persistent day-to-day. If users are high today, they tend to also be high tomorrow .

(d) Would Day 10 help forecast Day 11? Why?

Yes. Because ρ(1) is close to 1, today's value contains a lot of information about tomorrow's value. So knowing Day 10 = 57 would likely improve a forecast for Day 11, since the series tends to move smoothly rather than jumping randomly.

Problem 4

Data:

$$Y_1 = 10,\ Y_2 = 12,\ Y_3 = 11,\ Y_4 = 15,\ Y_5 = 13,\ Y_6 = 16,\ Y_7 = 14 \tag{31}$$

(a) 3-day centered moving average for Day 4

Centered MA at Day 4 uses Days 3, 4, 5:

$$MA_3(4) = \frac{Y_3 + Y_4 + Y_5}{3} = \frac{11 + 15 + 13}{3} = \frac{39}{3} = \boxed{13} \tag{32}$$

(b) 3-day trailing moving average for Day 5

A 3-day **trailing** MA for Day 5 usually means using the current day and the two previous days: Y_3, Y_4, Y_5.

(Your formula matches that.)

$$MA_3(5) = \frac{Y_3 + Y_4 + Y_5}{3} = \frac{11 + 15 + 13}{3} = \boxed{13} \tag{33}$$

(c) Weighted moving average for Day 4

$$WMA(4) = 0.25Y_3 + 0.50Y_4 + 0.25Y_5 = 0.25(11) + 0.50(15) + 0.25(13) = 2.75 + 7.50 + 3.25 = \boxed{13.5}$$

(d) Why might the weighted average produce a smoother trend estimate?

A weighted moving average can be smoother because it puts more weight on the middle (most relevant) observation and less weight on the neighboring days. This reduces the impact of short-term spikes or dips from the surrounding days, so the trend estimate is less “noisy.”

Problem 5

Part A: Generalized Linear Models (GLM)

A1: Exploratory Data Analysis

Key Findings from Box Plots:

- **Age**: Purchasers are older on average (35.9 vs 28.7 years)
- **Income**: Purchasers have higher income ($52.3k vs $40.2k)
- **TimeOnSite**: Purchasers spend more time on site (6.3 vs 3.9 minutes)

Correlation Matrix:

- Features show very weak correlations with each other (-0.05 to 0.00)
- This is good: no multicollinearity issues in our model

A2: Linear Probability Model (LPM)

Model Results:

VARIABLE	COEFFICIENT	P-VALUE
Constant	0.3104	0.000
Age	0.0071	0.000
Income	0.0052	0.000
TimeOnSite	0.0119	0.000

R-squared: 0.144 (14.4% of variance explained)

Problem with LPM:

- **17.4% of predictions exceed 1** (174 out of 1000)
- This is a fundamental limitation of LPM: probabilities should be bounded [0, 1]

A3: Logistic Regression

Model Results:

VARIABLE	COEFFICIENT	ODDS RATIO	P-VALUE
Constant	-5.8737	0.003	0.000
Age	0.1026	1.108	0.000
Income	0.0748	1.078	0.000
TimeOnSite	0.2692	1.309	0.000

Interpretation:

- **Age**: Each additional year increases odds of purchase by 10.8%
- **Income**: Each $1k increase raises odds by 7.8%
- **TimeOnSite**: Each additional minute increases odds by 30.9% (strongest predictor)

Advantage over LPM: All predictions are within [0, 1] ✓

A4: New Customer Predictions

AGE	INCOME	TIMEONSITE	PROBABILITY	CLASSIFICATION
25	$30k	2 min	37.15%	No Purchase
35	$50k	5 min	94.29%	Purchase
45	$70k	8 min	99.78%	Purchase
55	$90k	10 min	99.99%	Purchase

Most likely to purchase: Customer 4 (Age 55, Income $90k, TimeOnSite 10 min)

- Has the highest values for all three predictors

Part B: Time Series Analysis

B1: Time Series Visualization

Observed Patterns:

1. **Strong upward trend**: Sales grow from ~$1,000 to ~$8,000 over 2 years
2. **Yearly seasonality**: Higher sales in Nov-Dec (holiday season)
3. **Weekly variation**: Slight differences across weekdays (Wed highest, Sat lowest)
4. **Special events**: Clear spikes on Black Friday and Christmas

Mean Sales by Month:

- Lowest: January ($2,319)
- Highest: December ($5,408)

B2: Stationarity Assessment

Comparison of First 6 Months vs Last 6 Months:

PERIOD	MEAN	STD DEV
First 6 months	$1,417.52	$131.68
Last 6 months	$6,292.87	$927.96

Conclusion: The series is NOT STATIONARY

- Mean increases significantly over time (quadratic trend)
- Variance also increases (heteroskedasticity)
- Rolling mean shows clear upward movement

B3: Autocorrelation Analysis

ACF Results:

LAG	AUTOCORRELATION
Lag 1 (yesterday)	0.9983
Lag 7 (last week)	0.9982
Lag 30 (last month)	0.9955

Interpretation:

- Very high autocorrelation at all lags (due to strong trend)
- ACF decays very slowly → characteristic of non-stationary data
- Would need differencing to achieve stationarity

B4: STL Decomposition

Components:

1. **Trend**: Quadratic upward growth ($1,132 → $8,011)
2. **Seasonal**: 7-day repeating cycle (range: -183 to +410)
3. **Remainder**: Mean ≈ 7, Std ≈ 77 (contains noise + special events)

STL successfully separated:

- Long-term growth trend
- Weekly seasonal pattern
- Irregular components (including holiday spikes)

B5: Remainder Diagnostics

Statistics:

- Mean: 6.99 (close to 0 ✓)
- Std: 77.40
- Normality: Rejected ($p < 0.001$) due to outliers

Outliers Identified (|remainder| > 3×std):

- **Black Friday 2024**: Nov 24 (+833)

- **Black Friday 2025**: Nov 24 (+800)
- **Christmas 2024**: Dec 20-25 (+299 to +509)
- **Christmas 2025**: Dec 20-24 (+370 to +459)

Conclusion: STL decomposition successfully isolated special events in the remainder component.