Problem Set II Solution

Tobias Bodentien Philipp Grunenberg Alexander Haas Osama Warshagha

05-12-2025

Task 1

How could the relationship between price and demand be affected by endogeneity?

Endogeneity describes the problem, that the regressor x correlates with the true residuals. If endogeneity is present in the dataset, the estimated coefficients are biased regarding causation. There are two mechanisms creating endogeneity: Omitted variables and reverse causality.

Reverse causality: In an ideal economic model the demand is a function of the price. The higher the price, the lower the demand and vice versa. But in reality, the price can also be determined by the demand i.e. when a buyer knows that he is the only one interested in the product he can better negotiate a lower price. Then, we have a reverse causality.

Omitted variables: The relationship between price and demand (quantity of sold fish) in the fish market may be affected by endogeneity, when the price correlates well with the demand, but is not the causality. Instead, a omitted variable is the common cause for price and demand.

This omitted variable could be the supply of fish. Here, the total amount of received fish is a good control variable for the supply. In this case the quantity and price are determined by the amount of fish received at the day. When there is not much fish available, the seller just cannot sell more fish resulting in a lower quantity with higher prices. On the other hand, when there is a lot of fish available the seller can sell more, resulting in a higher quantity with lower prices. This plausability argument is supported by the following data:

```
round(summary(price_demand)$coefficients, 5)
```

```
##                   Estimate Std. Error  t value Pr(>|t|)
## (Intercept)       9615.381   1038.694  9.25718  0.00000
## daily_data$price -3709.017   1098.750 -3.37567  0.00102
```

```
round(summary(price_demand_totr)$coefficients, 5)
```

```
##                     Estimate Std. Error  t value Pr(>|t|)
## (Intercept)       3581.90280  802.71462  4.46224  0.00002
## daily_data$price -1402.96103  713.79407 -1.96550  0.05192
## daily_data$totr      0.62142    0.04802 12.93994  0.00000
```

If we include the number of received fish (totr) in the linear regression model, the effect of the price is reduced from an initial -3709 (linear regression without totr) to -1403 (linear regression with totr). Also the p-value of price in the model including totr is above 5%, indicating that it is not significant. Although -1403 is still large it is worth noting, that the average quantity is 6334.67 pounds and the average price is 0.8846. The effect of the received fish on the other hand is very significant (near 0 p-value for totr).

Implications: The quantity of sold fish is not determined by the price. Instead it is determined by the supply of fish. The buyers of fish are not sensitive to the price (meaning the price elasticity is low) but to the availability of fish.

Is weather data a suitable instrument in this context?

In her dataset Graddy classified the weather as stormy when a certain wave height and wind speed are exceeded. The wave height and wind speed are the moving averages of the last three days' wind speed and wave height before the trading day. Her argument is, that storms are an important determinant of the supply as strong winds and high waves make it difficult to catch fish. If supply is high, quantities rise and prices fall and vice versa. This is supported by the following data:

```
## diff_qty = -2370.7 diff_price = 0.322184 mean_qty = 6334.666 mean_price = 0.8845243
```

On a stormy day the average quantity of sold fish shrinks by 2370.7 pounds and the price rises by 32.22 cents. The average price on all days was 88.45 cents and average quantity was 6334.67 pounds. The correlation of price and stormy weather is also positive (0,42). Therefore the requirement of relevance is fulfilled.

```
## cor(price, stormy) = 0.4227539
```

Another requirement for instruments is exogeneity. So, to predict prices for fish, we need a variable that is independent of supply. Storms are such a variable as the supply cannot influence the weather. Therefore, the data and plausibility support that stormy weather can be used as an instrumental variable.

Re-run lin-log model with instrumental variables

First we run the lin-log model with stormy as an instrumental variable as proposed by Graddy and calculate the Hausman test of the OLS regression and the iv regression.

```
## hausman stormy p-value:  0.176175
```

The Hausman test (p = 0.176) does not reject the null hypothesis of identical estimates. This implies that the instrument based on stormy weather conditions is not strong enough to provide a statistically significant improvement over OLS. As a result, the OLS estimate appears adequate for this dataset, and the evidence for price endogeneity is weak.

The stormy variable is a binary variable indicating if wave height and windspeed are above a certain threshold. This omits information. Therefore, we now use windspeed directly as an instrumental variable. It fulfills the requirement of relevance as it is correlated with price (cor(price, windspd) = 0.42)

```
## cor(price, windspd)=  0.4150659
```

As windspeed also describes the weather the same argument as above holds regarding exogeneity.

```
## hausman windpeed p-value:  0.3155061
```

Again, the Hausman test between the OLS regression and the IV regression with windspeed indicates no evidence of price endogeneity (p-value = 0.316). Using wind speed as an instrument does not improve the model statistically.

Is there an endogeneity problem in the data? Do you see other endogeneity problems not captured by your instrument?

According to the Hausman tests using weather data as an instrument, we do not have an endogeneity problem in the data. But as seen above the total amount of received fish (totr), i.e. the supply, is an omitted variable that changes the results significantly. Therefore we conclude that we have an endogeneity problem, that is not captured by the instruments of weather.

Task 2

The data preprocessing remains the same as in PS1 Task 3, however this time we do not group by the customer id and will leave this column unmodified as dependent variable.

Expectations of using multi-level modelling

By ignoring the nested structure, OLS estimates in PS1 were likely biased and standard errors underestimated. By explicitly accounting for the hierarchy (purchases nested within customers), we expect to correct this bias and obtain more conservative (larger) standard errors. Additionally, the multilevel model separates 'within-customer' effects

from 'between-customer' differences, potentially revealing the true price sensitivity that was previously masked by aggregation bias.

```
baseline_model = lmer(quan ~ price_c + (1|cusn), data = detailed_data_prep)
print(coef(summary(baseline_model)))
```

```
##              Estimate Std. Error       df  t value     Pr(>|t|)
## (Intercept) 196.17053   13.37867 235.5046 14.66293 5.048305e-35
## price_c     -46.84754   23.95192 420.6541 -1.95590 5.113848e-02
```

```
icc(baseline_model)
```

```
## # Intraclass Correlation Coefficient
##
##     Adjusted ICC: 0.562
##   Unadjusted ICC: 0.559
```

The Baseline Model

The baseline model yields an Intraclass Correlation Coefficient (ICC) of approximately 0.56. This indicates that over half (56%) of the variance in purchase quantity is attributable to stable differences between customers rather than daily fluctuations, showing the magnitude of nestedness.

The model estimates a substantial negative price effect (-46.85), implying that a one-unit price increase reduces demand by nearly 47 units. This is a major difference to the estimated effect in PS1. Interestingly, this effect is marginally significant ($p \approx 0.051$), likely that the more conservative standard error estimation in the multilevel framework is balanced out by the removal of a bias. While not statistically significant at the strict 5% level, the coefficient's magnitude suggests considerable economic relevance and high price sensitivity.

We will continue only with a random intercept model, instead of a random slope model, as we have a total of 478 data points and 210 distinct customer ids. Fitting a different slope for each customer would be statistically extremely unstable.

Testing H2 (from PS1 Task 3)

```
#2.2 Moderated Model 2
cash_model = lmer(quan ~ price_c+cash_dummy + (1|cusn), data=detailed_data_prep)
cash_model_moderation = lmer(quan ~ price_c*cash_dummy + (1|cusn), data=detailed_data_prep)
print("--- Model 1: Main Effects ---")
```

```
## [1] "--- Model 1: Main Effects ---"
```

```
print(coef(summary(cash_model)))
```

```
##              Estimate Std. Error       df   t value     Pr(>|t|)
## (Intercept) 155.72757   27.91705 196.8741  5.578224 7.955035e-08
## price_c     -52.01598   24.61513 418.1852 -2.113171 3.517797e-02
## cash_dummy   50.23852   31.77644 207.1202  1.580999 1.154041e-01
```

```
print("--- Model 2: Interaction ---")
```

```
## [1] "--- Model 2: Interaction ---"
```

```
print(coef(summary(cash_model_moderation)))
```

```
##                      Estimate Std. Error       df    t value     Pr(>|t|)
## (Intercept)        154.108434   27.97229 198.0996  5.5093245 1.111018e-07
## price_c             -7.436002   46.15495 367.4652 -0.1611095 8.720957e-01
## cash_dummy          51.598079   31.81981 207.8238  1.6215709 1.064114e-01
## price_c:cash_dummy -62.213993   54.54857 387.7993 -1.1405247 2.547716e-01
```

```r
cat("\nCorrelation between Price and Interaction Term:\n")
```

```
##
## Correlation between Price and Interaction Term:
```

```r
print(round(cov2cor(vcov(cash_model_moderation))[2, 4], 2))
```

```
## [1] -0.85
```

Incorporating the payment method (cash_dummy) as a control variable sharpens the model's precision regarding price sensitivity. Unlike the unconditional baseline model, this specification reveals a statistically significant negative effect of price ($b = -52.02, p < 0.05$). Holding the payment method constant, a one-unit increase in price is associated with a decrease in quantity of approximately 52 units. Again we can see substantially higher estimation Variances as in PS1 but a effect size, comparable to the baseline model. However, the payment method itself does not show a significant main effect ($p \approx 0.115$), indicating that the mere act of paying with cash does not significantly shift the baseline demand volume (intercept) compared to credit/invoice payments.

The moderated model fails to support Hypothesis 2, as the interaction between price and payment method (price_c:cash_dummy) is not statistically significant ($p \approx 0.255$). The correlation matrix shows a strong dependency between the main price effect and the interaction term ($r \approx -0.85$). This high correlation could contribute to a inflation of the standard errors (multicolinearity) as standard error for the price coefficient nearly doubles from 24.62 in the main effects model to 46.16 here. With this large standard error even the large estimated effect remains insignificant. Consequently, the data does not support Hypothesis 2.

```r
r2(cash_model)
```

```
## # R2 for Mixed Models
##
##   Conditional R2: 0.563
##      Marginal R2: 0.020
```

```r
r2(cash_model_moderation)
```

```
## # R2 for Mixed Models
##
##   Conditional R2: 0.565
##      Marginal R2: 0.022
```

```r
anova(cash_model, cash_model_moderation)
```

```
## Data: detailed_data_prep
## Models:
## cash_model: quan ~ price_c + cash_dummy + (1 | cusn)
## cash_model_moderation: quan ~ price_c * cash_dummy + (1 | cusn)
##                         npar    AIC    BIC  logLik -2*log(L)  Chisq Df Pr(>Chisq)
## cash_model                 5 6290.0 6310.8 -3140.0    6280.0
## cash_model_moderation      6 6290.7 6315.6 -3139.3    6278.7 1.3048  1     0.2533
```

Using Nakagawa's R^2, we find that ~56% of the variation in purchase quantity is driven by stable customer differences (Conditional R^2). In contrast, price and payment method explain only ~2% of the variance (Marginal R^2). While low, this is consistent with marketing data where individual heterogeneity often outweighs daily transaction factors.

The ANOVA confirms that adding the interaction term yields no statistically significant improvement. This is supported by the Information Criteria (AIC/BIC), which favor the simpler main-effects model.

Conclusion: We find a significant main effect of price, explaining about 2% of the variance. However, we reject Hypothesis 2: the data still provides no evidence that paying with cash makes customers more price-sensitive than paying with credit.

Testing H3 (from PS1 Task 3)

```
estb_model = lmer(quan ~ price_c + estb + (1|cusn), data = detailed_data_perep_estb)
estb_model_moderation = lmer(quan ~ price_c * estb + (1|cusn), data = detailed_data_perep_estb)
print("--- Model 3: Main Effects (Establishment) ---")
```

```
## [1] "--- Model 3: Main Effects (Establishment) ---"
```

```
print(coef(summary(estb_model)))
```

```
##                Estimate Std. Error       df    t value     Pr(>|t|)
## (Intercept)  278.479900   39.31438 192.8497  7.0834102 2.565251e-11
## price_c      -52.837137   25.06700 380.0506 -2.1078363 3.569948e-02
## estbs       -101.126259   42.21913 199.7287 -2.3952712 1.753164e-02
## estbsf         8.886765   63.07847 169.2815  0.1408843 8.881289e-01
```

```
print("--- Model 4: Interaction (Establishment) ---")
```

```
## [1] "--- Model 4: Interaction (Establishment) ---"
```

```
print(coef(summary(estb_model_moderation)))
```

```
##                  Estimate Std. Error       df    t value     Pr(>|t|)
## (Intercept)     279.75204   39.41145 193.9491  7.0982428 2.322878e-11
## price_c         -15.77310   62.57588 353.2813 -0.2520636 8.011385e-01
## estbs          -102.38854   42.30430 200.3881 -2.4202869 1.640021e-02
## estbsf           10.83606   63.34789 171.0348  0.1710564 8.643816e-01
## price_c:estbs   -37.17085   69.47123 360.5898 -0.5350538 5.929424e-01
## price_c:estbsf  -77.60361   90.79056 348.5321 -0.8547542 3.932744e-01
```

```
cat("\nCorrelation between Price and Interaction (Stores):\n")
```

```
##
## Correlation between Price and Interaction (Stores):
```

```
cormat <- cov2cor(vcov(estb_model_moderation))
print(round(cormat["price_c", "price_c:estbs"], 2))
```

```
## [1] -0.9
```

In the main effects model, controlling for establishment type confirms the robustness of the price effect. The coefficient for price remains stable and statistically significant ($b = -52.84, p < 0.05$), reinforcing the finding that higher prices lead to lower purchase volumes regardless of store type. Additionally, the model now reveals significant differences in baseline demand between establishment categories. 'Stores' (estbs) purchase significantly less than the reference group 'Fry Shops' (f), with a reduction of approximately 101 units ($p < 0.05$). In contrast, hybrid establishments ('sf') do not show a statistically significant difference in baseline purchase quantity compared to Fry Shops.

The moderated model provides no support for Hypothesis 3, as neither of the interaction terms (price_c:estbs and price_c:estbsf) reaches statistical significance ($p > 0.39$). This suggests that the sensitivity to price changes does not vary significantly across different types of establishments. However, similar to the payment method analysis, the correlation matrix indicates an extremely high correlation between the main price effect and the interaction term for stores (price_c:estbs), with a coefficient of -0.90. This collinearity inflates the standard error for the price coefficient more than twofold (from ~25 to ~63), reducing the statistical power to detect true interaction effects. Thus, we find no statistical evidence for moderation even for the large effects, as we have high standard errors.

```
r2(estb_model)
```

```
## # R2 for Mixed Models
##
##   Conditional R2: 0.562
##      Marginal R2: 0.056
```

```r
r2(estb_model_moderation)
```

```
## # R2 for Mixed Models
##
##   Conditional R2: 0.562
##      Marginal R2: 0.057
```

```r
anova(estb_model, estb_model_moderation)
```

```
## Data: detailed_data_perep_estb
## Models:
## estb_model: quan ~ price_c + estb + (1 | cusn)
## estb_model_moderation: quan ~ price_c * estb + (1 | cusn)
##                         npar    AIC    BIC  logLik -2*log(L)  Chisq Df Pr(>Chisq)
## estb_model                 6 5665.7 5690.1 -2826.8    5653.7
## estb_model_moderation      8 5669.0 5701.4 -2826.5    5653.0 0.7365  2     0.6919
```

The Conditional R^2 remains consistent at ~56.2%, reaffirming that customer identity is the primary driver of purchase volume. However, a key difference emerges in the Marginal R^2, which rises to ~5.6%. This is more than double the explanatory power of the payment method model (~2%), indicating that the type of establishment (estb) is a much stronger predictor of daily purchase quantity than the mode of payment, even if it does not explain the price sensitivity.

The model comparison confirms that this additional explanatory power comes from the main effects (intercept differences), not the interaction. The ANOVA yields a clearly non-significant result ($\chi^2(2) = 0.74, p \approx 0.69$), and information criteria (AIC/BIC) penalize the inclusion of the interaction terms.

Conclusion: We reject Hypothesis 3. While different establishment types buy significantly different amounts on average (e.g., Stores buy less than Fry Shops), there is still no evidence that their demand curves have different slopes. Price sensitivity appears to be uniform across all establishment types.

Task 3

In Task 3, we use the **StoreData** panel dataset containing monthly county-level sales observed over 18 months, where months 1–6 form the pre-period and months 7–18 the post-period. Treatment counties are affected by a store closure, while control counties are unaffected. Our goal is (i) to develop time-series forecasting models for **total sales** (offline + online) for one treated and one control county, and (ii) to study the interaction between offline and online sales using a VAR model following Lecture Chapter 2.2. Before selecting counties, we enforce a clean panel structure by restricting the sample to counties with a constant treatment status over time and complete coverage of months 1–18, since missing months or changing treatment labels would distort time-series estimation and invalidate comparisons. From this eligible set, we then randomly draw one treated and one control county using a fixed seed to ensure reproducibility; in our run, the selected counties are **treatment: county_id = 39** and **control: county_id = 25**.

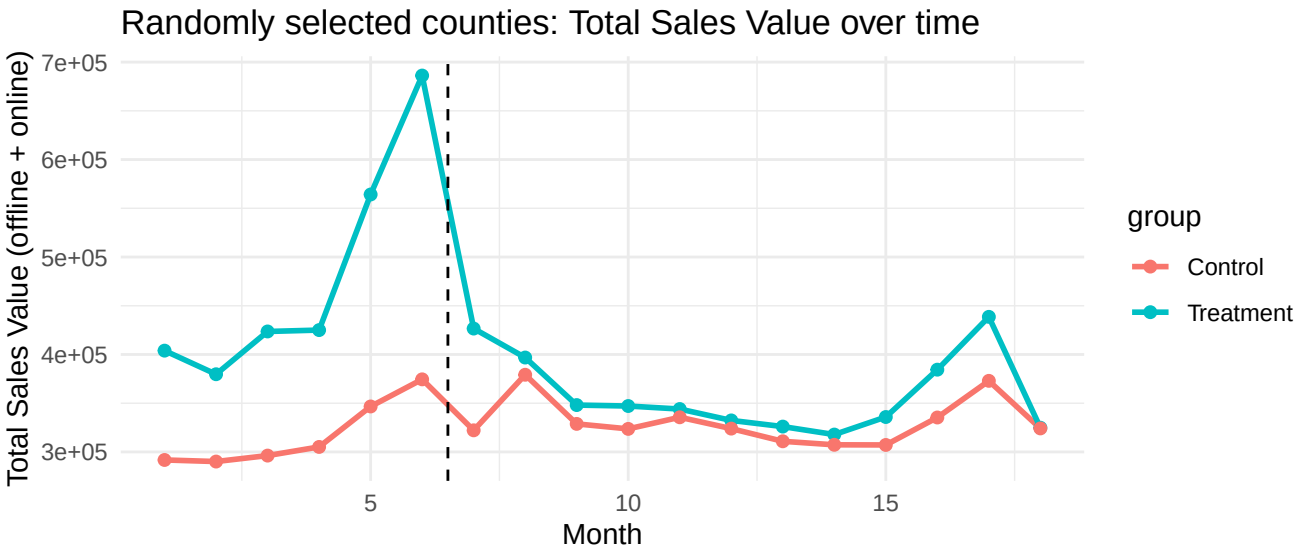

Figure 1: Total sales trajectories for the selected counties. Monthly total sales (offline + online) for the randomly selected treatment and control counties over months 1–18; the dashed vertical line indicates the start of the post-period (month 7).

The figure above plots the total sales value (offline + online) over time for both selected counties, with the dashed vertical line indicating the transition from the pre-period to the post-period (start at month 7). Visually, both counties show pronounced variation over time, especially around months 5–6 and later months, which motivates a careful time-series approach (stationarity checks, possible differencing, and lag selection) in the next step.

Time series setup (Total Sales)

We model monthly total sales value (offline + online) for each county as a univariate time series. The data are observed for 18 consecutive months; hence we treat the series as monthly data and focus on trend-stationarity / differencing, while seasonality (12-month cycle) cannot be identified reliably with such a short horizon.

Stationarity

We begin by assessing the time-series properties of total monthly sales in the treatment and control counties using the Augmented Dickey–Fuller (ADF) test. The ADF test evaluates whether a series contains a unit root. Formally, the null hypothesis is that the process has a unit root (and is therefore non-stationary), while the alternative hypothesis is that the series is stationary. In practice, a small p-value (e.g., below 0.05) provides evidence against the null, whereas a large p-value indicates that we cannot reject the presence of a unit root.

Before running the tests, we consider two representations of total sales for each county: levels and (when admissible) log-levels. Because the logarithm is only defined for strictly positive values, we first check whether each county's total sales series is positive throughout the sample. Both selected series are strictly positive, so applying the natural log transformation is valid. Using log-levels is useful in this context because it often stabilizes the variance for revenue-like data and allows changes to be interpreted approximately in percentage terms.

We then apply the ADF test to total sales in levels and log-levels for the treatment county (ID 39) and the control county (ID 25). The resulting p-values are relatively large (around **0.39–0.50** in levels and **0.38–0.48** in log-levels), implying that we cannot reject the null hypothesis of a unit root for any of the four series. In other words, both the level and log-level total sales series exhibit behavior that is statistically consistent with non-stationarity over the observed 18-month period.

Table 1: ADF test p-values

Series	ADF_p_value
Treatment 39 (level)	0.3873
Control 25 (level)	0.5011
Treatment 39 (log level)	0.3760

Series	ADF_p_value
Control 25 (log level)	0.4839

To move the data closer to stationarity, we subsequently work with the first difference of the log-transformed series, $\Delta\log(\text{Total Sales})$, for both treatment and control counties. These differenced log series can be interpreted as approximate monthly growth rates in total sales. Differencing is a standard approach to remove deterministic trends and unit-root components, thereby yielding series that fluctuate around a more stable mean. Moreover, focusing on growth rates instead of levels facilitates comparison between counties that may differ substantially in their absolute sales levels.

We visualize these growth-rate series over time and mark the beginning of the post-treatment period again with a vertical dashed line, while a horizontal dotted line at zero indicates no growth. The plot shows how growth rates fluctuate around a roughly stable mean in both the pre- and post-periods, without obvious deterministic trends, which is consistent with the objective of obtaining a more stationary series.

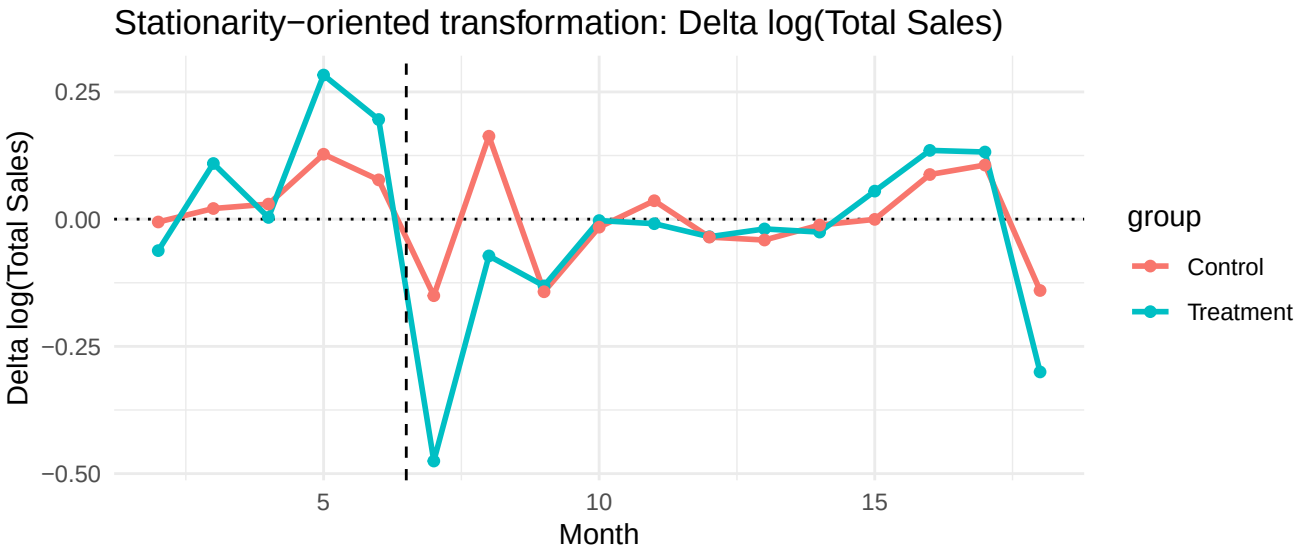

Figure 2: Growth-rate transformation of total sales (Treatment vs. Control). First differences of log total sales, $\Delta\log(\text{Total Sales})$, for the treatment and control counties over months 2–18; the dashed vertical line marks the start of the post-period (month 7) and the dotted horizontal line indicates zero growth.

Finally, we re-run the ADF test on the growth-rate series $\Delta\log(\text{Total Sales})$ for both counties. Since total sales are strictly positive throughout, the log transformation is well-defined. The resulting p-values remain relatively large (about 0.48–0.60), so we still cannot reject the unit-root null hypothesis at conventional significance levels. However, this outcome should be interpreted cautiously because the sample is extremely short (only 17 observations after differencing), which severely limits the power of unit-root tests. From a practical modeling perspective, working with $\Delta\log(\text{Total Sales})$ is still a sensible variance-stabilizing transformation that captures monthly growth rates. We therefore proceed with the differenced log series as the main outcome in the subsequent analysis, while acknowledging that formal stationarity evidence is weak in such a small sample.

Table 2: ADF tests on $\Delta\log(\text{Total Sales})$

Series	ADF_p_value
Treatment 39 (Δlog)	0.4803
Control 25 (Δlog)	0.6049

How many lags to include

To inform the choice of the dynamic structure in our subsequent time-series regressions, we examine the autocorrelation function (ACF) and the partial autocorrelation function (PACF) of the transformed outcome variable, i.e.,

the first difference of the log-transformed total sales, for both the treatment county (ID 39) and the control county (ID 25). The ACF summarizes the linear dependence of the series with its own past values at different lags, while the PACF isolates the incremental contribution of each lag after controlling for all shorter lags. Together, these diagnostics provide guidance on whether the series exhibits short-run persistence that would call for an autoregressive specification with one or more lags.

The Figure below displays the ACF and PACF of Delta log(Total Sales) for both counties. Overall, the plots do not suggest strong or long-lasting autocorrelation. At all positive lags, the sample autocorrelations are small and lie within the approximate 95% confidence bands, and the PACF does not show any pronounced spikes. Given the very short time dimension of our data (only 17 monthly observations after differencing), these diagnostics should be interpreted cautiously, but they nonetheless indicate that any remaining serial dependence is weak and, if present at all, likely to be of very low order.

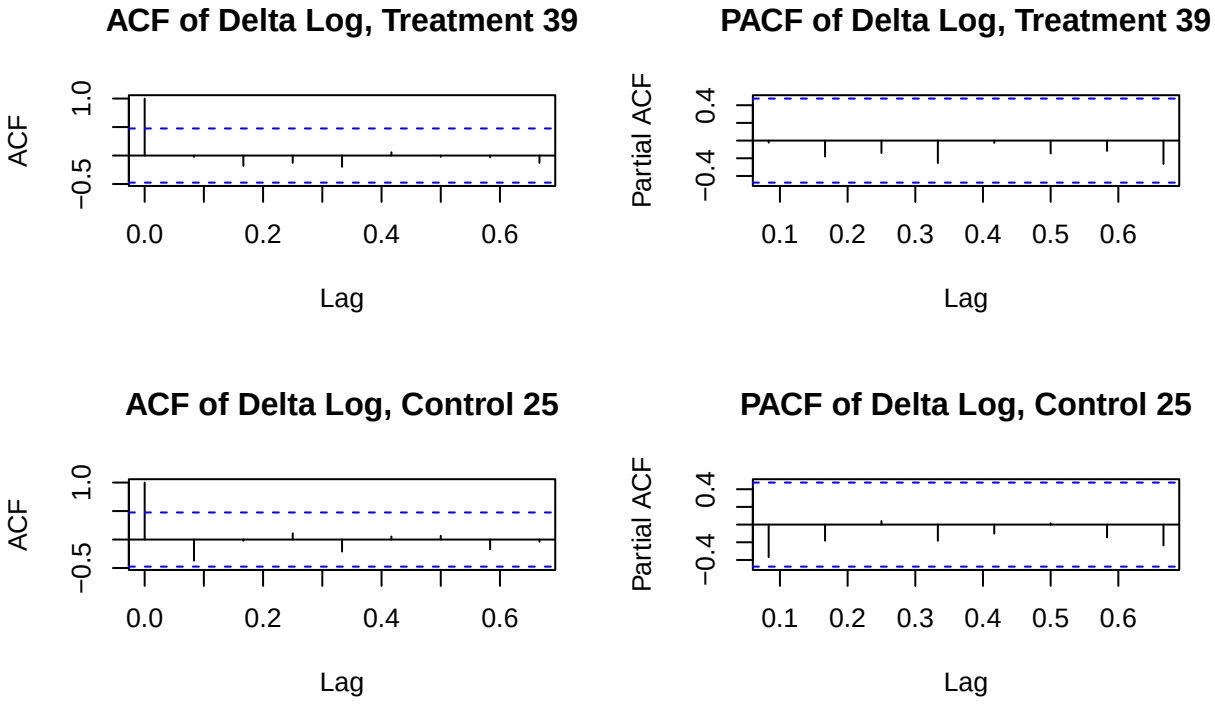

Figure 3: ACF/PACF of $\Delta \log$(Total Sales) (Treatment vs. Control). Autocorrelation (ACF) and partial autocorrelation (PACF) functions of the differenced log total sales series for the treatment county (top row) and the control county (bottom row); dashed lines indicate approximate 95% confidence bounds.

In light of these considerations and the **very short sample length** (18 months), we adopt a highly parsimonious specification and avoid heavily parameterized ARMA structures that would be poorly identified. Concretely, we model log(Total Sales) for each county using an ARIMA(0,1,0) with drift, i.e., a random walk with drift:

$$\log(y_t) = \log(y_{t-1}) + \mu + \varepsilon_t,$$

where μ denotes the average monthly growth rate in total sales (in log points) and ε_t is a white-noise innovation. This specification is equivalent to:

$$\Delta \log(y_t) = \mu + \varepsilon_t.$$

Thus, once we account for the integrated nature of log sales via first differencing (implicit in the ARIMA(0,1,0) structure), no additional autoregressive or moving-average terms are required to capture systematic short-run dynamics. This aligns with the diagnostic impression that any remaining serial dependence is weak and, in this small sample, unlikely to justify estimating additional lags. We therefore proceed with ARIMA(0,1,0) with drift for the treatment county (ID 39) and the control county (ID 25), and we interpret results cautiously given the limited time dimension.

Results (R-Output in rmd file)

The ARIMA(0,1,0) models with drift for log(Total Sales) confirm the very parsimonious dynamic structure suggested by the diagnostics. For the treatment county, the estimated drift is $\hat{\mu} = 0.0062$ (s.e. 0.0217), corresponding to an average monthly growth rate of about $e^{0.0062} - 1 \approx 0.62\%$. For the control county, the drift is $\hat{\mu} = -0.0128$ (s.e. 0.0420), i.e. approximately $e^{-0.0128} - 1 \approx -1.27\%$ per month. In both cases, the drift estimates are small relative to their standard errors, implying no statistically strong evidence of a systematic upward or downward trend in log sales over this short sample.

Consistent with the ARIMA(0,1,0) specification, no additional autoregressive or moving-average parameters are estimated. The innovation variance is $\hat{\sigma}^2 = 0.0085$ for the treatment county and $\hat{\sigma}^2 = 0.0319$ for the control county, indicating higher volatility of shocks to log sales in the control county. Overall, these results support using a random-walk-with-drift model as a defensible, low-parameter baseline given the limited time dimension.

Forecasting (Total Sales)

Based on the selected parsimonious specification, we forecast total sales (offline + online) for each county for the three months following the last observed month in the dataset (months **19**–**21**). Forecasts are generated on the log scale and then back-transformed to levels via exponentiation; the reported prediction intervals are obtained by applying the same transformation to the interval bounds and should be interpreted as an approximation.

The resulting point forecasts are very stable over the three-month horizon. For the treatment county, predicted total sales show a slight upward drift, increasing from approximately **326k** (month 19) to **330k** (month 21). For the control county, forecasts exhibit a small decline, from about **321k** (month 19) to **313k** (month 21). Overall, both counties are predicted to remain in a similar range in the immediate post-sample horizon.

Uncertainty, however, increases noticeably with the forecast horizon, as reflected in the widening 95% prediction intervals. Importantly, the prediction intervals for the control county are substantially wider than for the treatment county (e.g., by month 21 roughly **170k**–**573k** in the control county versus **242k**–**452k** in the treatment county), indicating higher volatility of shocks in the control county and therefore lower forecast precision. This difference is also clearly visible in the forecasting plots: while both panels show relatively flat point forecasts after month 18, the shaded 95% prediction band is markedly broader for the control county. In sum, while the point forecasts suggest only mild changes in expected sales over months 19–21, the wide intervals—especially for the control county—underline that inference is limited by the short sample and the inherent uncertainty of forecasting with only 18 monthly observations.

Table 3: 3-month forecasts for Total Sales (levels), with 95% PI.

month	treat_forecast	treat_lo95	treat_hi95	ctrl_forecast	ctrl_lo95	ctrl_hi95
19	320718.8	225939.4	455257.4	326160.4	272264.3	390725.5
20	316634.4	192934.6	519644.1	328185.7	254207.3	423692.9
21	312602.0	170408.2	573446.7	330223.5	241515.6	451513.6

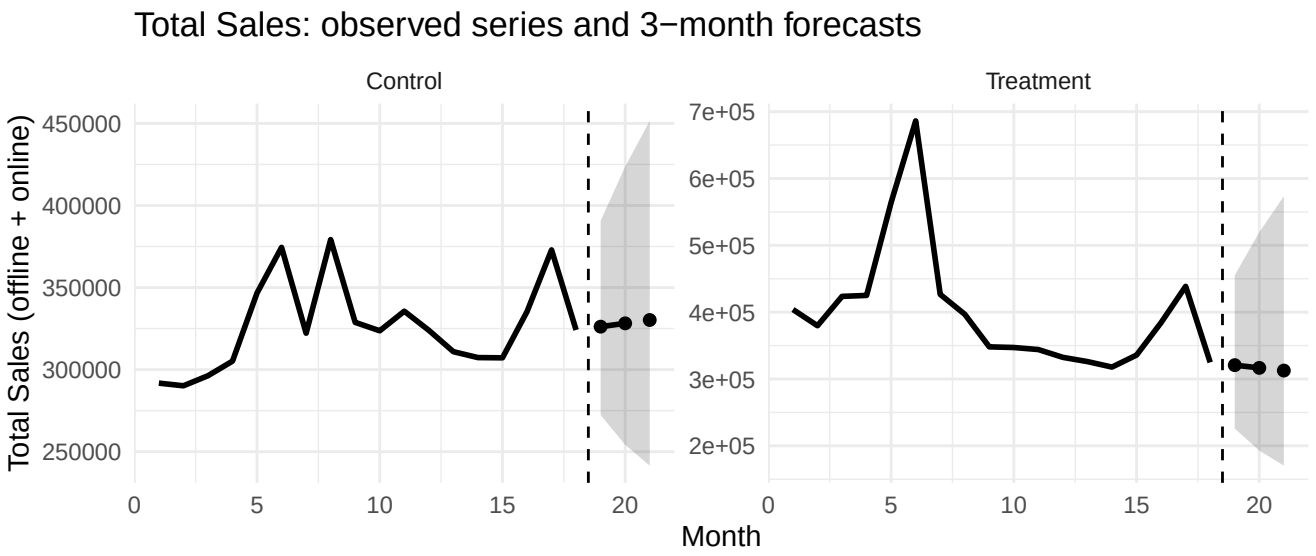

Figure 4: Total sales forecasts (Treatment vs. Control). Observed total sales (offline + online) for months 1–18 and 3-month ahead forecasts for months 19–21 from the fitted ARIMA(0,1,0) with drift model (points/line); the dashed vertical line marks the forecast start and the shaded band shows the 95% prediction interval.

Residual ACF/ Seasonality

The residual ACF plots for both counties show no pronounced spikes outside the 95% confidence bands at positive lags, suggesting that the ARIMA(0,1,0) specification captures the main time-series structure reasonably well. In particular, there is no clear seasonal pattern (e.g., no prominent spike around the 12-month lag); given the very short sample (18 months), this also supports our decision **not** to model seasonality explicitly.

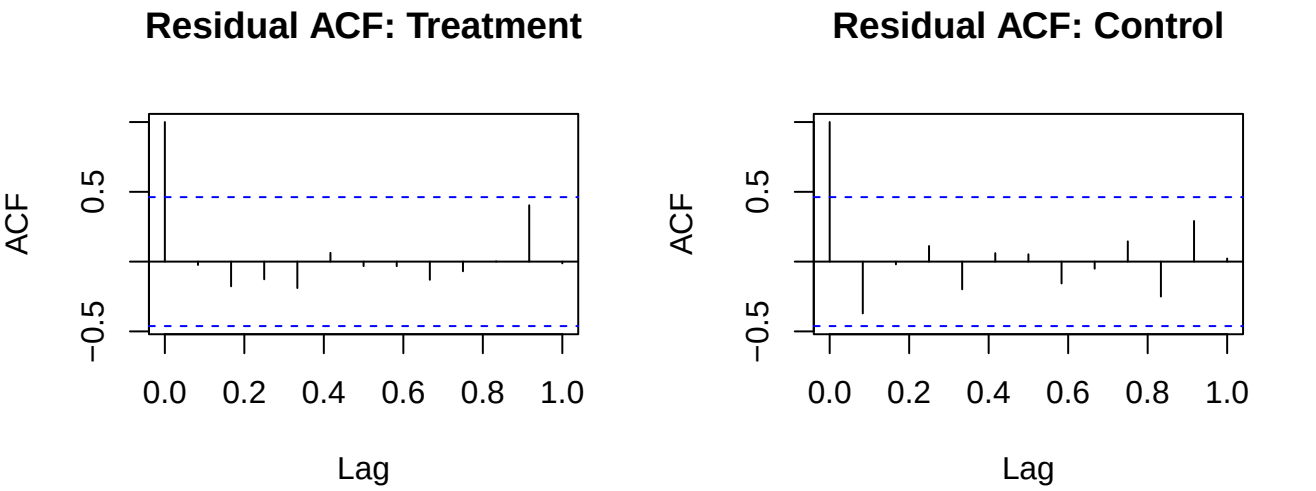

Figure 5: Residual ACF diagnostics (Treatment vs. Control). Residual ACFs of the fitted ARIMA(0,1,0) with drift models for the treatment (left) and control (right) county; dashed lines indicate approximate 95% confidence bounds.

VAR analysis (Online vs. Offline Sales)

We now turn to the second part of the task and analyze the interdependencies between offline and online sales using a Vector Autoregression (VAR) model for both the treatment and the control county. Given the very short sample (18 monthly observations), we make a few simplifying choices to keep the model identifiable and interpretable.

Transformation and simplifying assumptions.
Instead of modeling the sales levels, we work with log growth rates, i.e., the first differences of log sales, because the level series are likely non-stationary and a VAR in levels could lead to spurious dynamics in such a small sample. Working with $\Delta \log(\cdot)$ yields an approximately stationarity-oriented representation and has a natural economic interpretation as (approximate) monthly percentage changes. We further keep the specification parsimonious: we include a constant term and restrict the maximum lag order to a small number (due to limited degrees of freedom), and we do not attempt to estimate separate pre-/post-regimes.

Model setup.
For each county, we estimate a bivariate VAR on the transformed series

$$y_t \;=\; \begin{pmatrix} \Delta \log(\text{offline}_t) \\ \Delta \log(\text{online}_t) \end{pmatrix}, \qquad y_t \;=\; c + A_1 y_{t-1} + \cdots + A_p y_{t-p} + u_t,$$

where c is a constant vector and u_t is a vector of innovations. The key quantities of interest are the cross-lag coefficients, i.e., how past online growth predicts offline growth (and vice versa).

Lag length selection.
To choose the lag order p, we use standard information criteria (AIC, HQ, SC, FPE) while restricting the search to very small lag lengths. For the treatment county, all criteria select $p = 2$ (VAR(2)), whereas for the control county, all criteria select $p = 1$ (VAR(1)). This yields a more flexible dynamic structure for the treatment county while remaining parsimonious.

Results (coefficients, R-Output in rmd file)

Treatment county (VAR(2)).
In the offline equation, we find a statistically significant cross-lag effect from online growth: $\Delta \log(\text{online})_{t-1}$ enters with a positive coefficient (estimate ≈ 0.62, $p \approx 0.009$). This suggests that increases in online sales growth tend to be followed by higher offline sales growth one month later in the treatment county. The second lag of offline growth is negative and marginally significant (estimate ≈ -0.55, $p \approx 0.074$), indicating some mean reversion in offline growth at the two-month horizon. Overall, the offline equation is jointly significant (F-test $p \approx 0.042$), consistent with meaningful short-run dynamics.

In the online equation, the lagged offline growth rate shows a positive but only weakly significant association ($\Delta \log(\text{offline})_{t-1}$ estimate ≈ 0.89, $p \approx 0.093$), while the remaining terms are not statistically strong. Hence, the most robust dynamic linkage in the treatment county runs from online (lag 1) to offline.

Control county (VAR(1)).
For the control county, the VAR(1) results are substantially weaker. Neither cross-lag coefficient is statistically significant in either equation, and the overall explanatory power is low—especially for the online equation. The only term that comes close to significance is the own-lag in the offline equation ($\Delta \log(\text{offline})_{t-1}$ estimate ≈ -0.58, $p \approx 0.097$), which is consistent with mild mean reversion in offline growth. Overall, these estimates suggest that short-run interactions between online and offline growth are much less pronounced in the control county.

Contemporaneous comovement.
In both counties, the residual correlation between the offline and online equations is relatively high (around 0.63–0.69), indicating that both channels are affected by common shocks within the same month, even when lagged dynamics are weak.

Taken together, the coefficient estimates point to stronger dynamic interdependence in the treatment county, in particular from online to offline sales growth, whereas the control county exhibits little evidence of systematic cross-channel lag structure. In the next step, we use impulse response functions (IRFs) to summarize and compare these dynamics more directly over time.

Impulse response functions (IRFs)

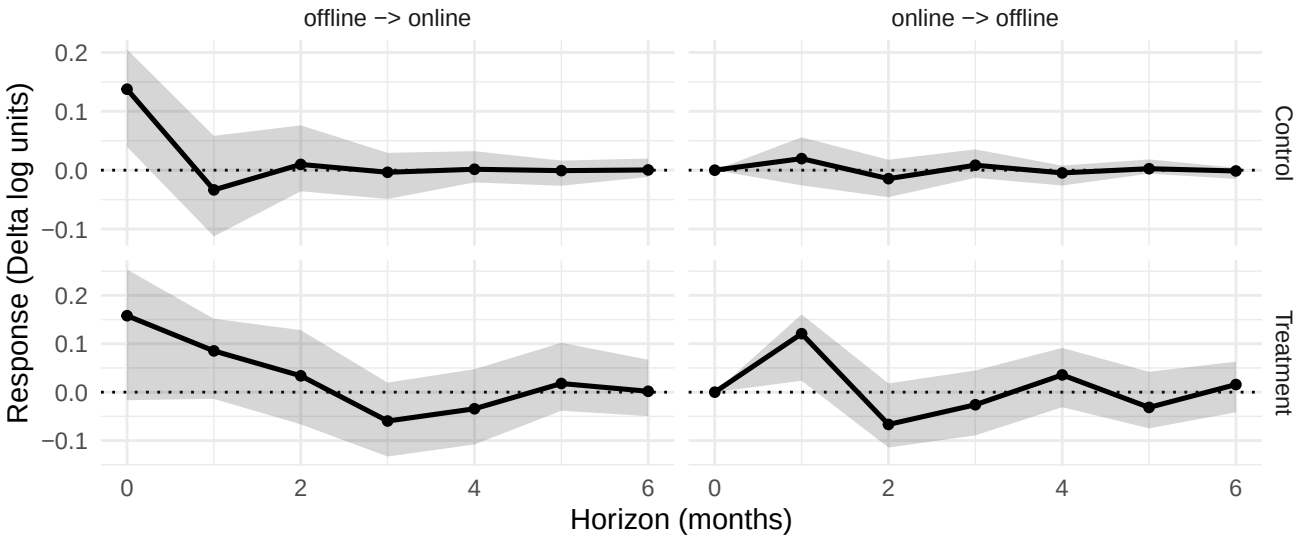

Figure 6: Impulse response functions (IRFs) from the VAR model (Treatment vs. Control). Estimated responses of $\Delta \log(\text{online})$ to a one-unit shock in $\Delta \log(\text{offline})$ (left column) and of $\Delta \log(\text{offline})$ to a one-unit shock in $\Delta \log(\text{online})$ (right column), shown separately for the control (top row) and treatment county (bottom row); the shaded areas denote bootstrap confidence bands and the dotted horizontal line marks zero response.

The IRFs broadly confirm the coefficient-based findings. In the control county, both the offline→online and online→offline responses remain close to zero across horizons, with confidence bands that typically include zero, indicating little evidence of systematic cross-channel dynamics. In contrast, the treatment county shows more pronounced short-run interactions: an offline shock is followed by a positive response of online growth at short horizons, and an online shock is associated with a noticeable response in offline growth (peaking early and then fading). Overall, the dynamic effects in the treatment county appear stronger but also imprecisely estimated, so results should be interpreted cautiously given the very short sample.

Task 4

A brief note on the approach to this subtask: since the task description requires the selection of three variables based on economic judgement and the interpretation of the logistic regression model using these variables, this is how I will proceed.

An alternative approach would be to fit a model incorporating all regressors and, if appropriate, compare it to a Lasso model. This approach would essentially treat the problem as a classic forecasting task. Although key metrics are initially easier to understand and interpret, using all available data (i.e. regressors) in this way could improve predictive performance.

Selection of three predictors

The idea is now to select three predictors that have as little conceptual overlap as possible. This will prevent redundant variables from being included in the model, which could otherwise negatively affect its interpretability (cf. multicollinearity).

1. Trend in Monetary value of sales: *sales_trend_3m*

The total sales values can be compared automatically across different products and categories. It is also intuitive to assume that regions or stores with persistently low sales are potential candidates for closure. However, it is important to note that absolute sales figures are not a suitable metric, since they can vary substantially between stores. For example, a small but profitable store may generate less revenue than a large but unprofitable one.

For this reason, the three-month trend is considered here: *sales_trend_3m*, which is the relative change in average sales between the first three months and the subsequent three months of the pre-closure period. Changes at the

monthly level are less robust since short-term fluctuations are to be expected. For instance, sales may temporarily decrease for non-critical reasons, such as an abundance of public holidays or supply bottlenecks, or temporarily increase due to pre-Christmas shopping.

H1: *A decline in sales over several months is an early warning sign of structural weakness in demand, which increases the risk of closure.*

2. Percentage of sales generated online: *pct_online_sales*

If a large proportion of customers shop online, this results in lower margins for the offline business. This can cause a store to become unprofitable in the long term.

H2: *A high proportion of sales made online indicates a structural shift in demand from physical stores to digital channels. This reduces the profitability of local stores and increases their risk of closure.*

3. Percentage of discounts offered: *pct_discounts*

Using discounts to stimulate demand may indicate that a store has weak pricing power and is therefore at risk of closure. However, it is important to note that persistently high levels of discounting may also be a deliberate strategic choice. This variable is not reported at the monthly level, so it cannot be transformed into a relative change measure. It is therefore used in its original (untransformed) form.

H3: *A high proportion of discounts suggests weak local demand and limited pricing power. This indicates insufficient profitability and an increased risk of closure.*

Model analysis

To prevent information leakage, i.e. the independent variables being potentially influenced by store closures, the dataset is first filtered to include only the months prior to the closures. Next, a variable is constructed to capture the change in total sales over three months, and all other non-selected regressors are removed from the dataset. After the data transformation, since all three variables are available exclusively at the county level, duplicate observations are dropped using *'distinct()'*.

```
paste0("Proportion of closed stores: ",
       length(df_pre__treat_treated$treat)/length(store_df_pre_treat_3selected$treat))
```

```
## [1] "Proportion of closed stores: 0.5"
```

Next, a logistic regression is fitted for each of the seven possible combinations, as well as for the null model. Additionally, the in-sample classification accuracy with a threshold of 0.5 is computed for each model. A likelihood-ratio test is also performed for each model. The logistic regression models are stored in a dataset that is sorted in descending order by AIC value.

```
knitr::kable(logit_model_results_overview)
```

Formula	AIC	Accuracy	LR statistic	p-value
treat ~ sales_trend_3m	92.7278	0.5152	2.7677	0.0962
treat ~ pct_discounts	93.1854	0.6061	2.3100	0.1285
treat ~ 1	93.4954	0.5000	0.0000	1.0000
treat ~ pct_discounts + sales_trend_3m	94.0900	0.5606	3.4055	0.1822
treat ~ pct_online_sales + sales_trend_3m	94.5488	0.5455	2.9466	0.2292
treat ~ pct_online_sales + pct_discounts	95.0337	0.6061	2.4618	0.2920
treat ~ pct_online_sales	95.0889	0.5000	0.4065	0.5237
treat ~ pct_online_sales + pct_discounts + sales_trend_3m	95.9833	0.5455	3.5121	0.3192

In terms of AIC, all models are very close to each other (with a difference of only 3.3 between the best and worst model). The best model uses *sales_trend_3m* as the sole regressor, whereas the worst model incorporates all three of the aforementioned economically motivated regressors. 50% of the stores in the dataset are closed (see above). Consequently, this is the automatic baseline in-sample accuracy of the null model. None of the other

logistic regression models achieves substantially higher accuracy than this baseline (the best value is 60%). The null hypothesis of the LR test cannot be rejected at the 5% significance level for any of the models. Only the AIC-best model is statistically significant at the 10% level. Therefore, none of the additional parameters in the unrestricted models leads to a statistically significant improvement in the likelihood of observing the data.

Best-performing model according to AIC: *sales_trend_3m* only

```
summary(logit_model_results$model[[1]])
```

```
##
## Call:
## glm(formula = f, family = binomial(link = "logit"), data = store_df_pre_treat_3selected)
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -1.0339     0.6859  -1.507    0.132
## sales_trend_3m   1.8163     1.1158   1.628    0.104
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 91.495  on 65  degrees of freedom
## Residual deviance: 88.728  on 64  degrees of freedom
## AIC: 92.728
##
## Number of Fisher Scoring iterations: 4
```

With a p-value of 0.104 for sales_trend_3m, we cannot reject the null hypothesis that an increase in the three-month sales trend has no effect on the probability of store closure. It should also be noted that multicollinearity is not an issue in models with a single regressor.

For the sake of completeness, the following interpretation of this coefficient is provided; however, due to the low level of statistical significance, it should be treated with caution. A relative increase in average sales over the three-month period is associated with higher log-odds of store closure. Specifically, a 100% increase in the three-month sales trend (i.e. doubling sales) increases the log-odds of store closure by 1.8163. According to the model, therefore, a positive sales growth trend increases the risk of closure and contradicts hypothesis H1.

Worst-performing model according to AIC: all three regressors

```
summary(logit_model_results$model[[nrow(logit_model_results)]])
```

```
##
## Call:
## glm(formula = f, family = binomial(link = "logit"), data = store_df_pre_treat_3selected)
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)        0.8497     3.6784   0.231    0.817
## pct_online_sales   0.3707     1.1351   0.327    0.744
## pct_discounts     -0.2842     0.3809  -0.746    0.456
## sales_trend_3m     1.2961     1.2748   1.017    0.309
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 91.495  on 65  degrees of freedom
## Residual deviance: 87.983  on 62  degrees of freedom
## AIC: 95.983
##
## Number of Fisher Scoring iterations: 4
```

In the model incorporating all three regressors, none of the predictors are statistically significant (all p-values greater than 0.30). Due to the lack of evidence, the coefficients are not interpreted further or examined for multicollinearity. Moreover, none of the regressors in any of the regression models is significant at a level of 10%.

Conclusion

Overall, the results of the logistic regression analysis provide only weak statistical evidence that our three proposed early-warning indicators can be relied upon to predict store closures at county level. The estimates are imprecise and do not reach conventional levels of statistical significance. This suggests that these variables alone are insufficient as stand-alone warning signals in our sample. These non-significant findings have notable practical implications: managers should exercise caution when using metrics such as online share, discount intensity, and short-term sales trends to predict closures. These indicators can be affected by strategic actions such as clearance discounting, which temporarily increases sales, and they may reflect ambiguous mechanisms such as online share capturing both substitution away from stores and robust omnichannel demand. Furthermore, closure decisions may be influenced by factors not considered here, such as local fixed costs, rent levels, competition, and store network optimisation. Therefore, demand-side indicators may require the addition of cost and location/network variables.
Finally, increasing the sample size would improve statistical power and allow for a more reliable out-of-sample assessment (e.g. a train–test split). Nevertheless, the current results suggest that predicting closures probably requires a broader range of factors than the three metrics considered here.