

Problem Set II Solution

Tobias Bodentien

Philipp Grunenberg

Alexander Haas

Osama Warshaga

02-12-2025

#Task 1

#Task 2

```
library(lme4)
library(lmerTest)
library(readr)
library(dplyr)
library(ggplot2)
library(tidyr)
library(performance)

#data loading
detailed_data = read_tsv("../data/detailed_fish_market_data.txt")

#1 data preperation
detailed_data_prep <- detailed_data %>%
  filter(!is.na(pric),
         !is.na(quan),
         type == "w") %>%
  arrange(date) %>%
  mutate(
    price_c = as.numeric(scale(pric, center = TRUE, scale = FALSE)), #main regressor
    quality_c = as.numeric(scale(qual, center = TRUE, scale = FALSE)), # Moderator 1
    cash_dummy = if_else(cash == 'c', 1, 0), # Moderator 2
    estb = as.factor(estb), # Moderator 3
  )

# additional steps for establishment
detailed_data_perep_estb = detailed_data_prep %>%
  filter(estb %in% c("s", "f", "sf"))
```

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.

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

```
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
```

```
## Formula: quan ~ price_c + (1 | cusn)
## Data: detailed_data_prep
##
## REML criterion at convergence: 6309.2
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -2.8987 -0.3619 -0.1655  0.2003  7.0525
##
## Random effects:
## Groups Name Variance Std.Dev.
## cusn (Intercept) 24885 157.8
## Residual 19428 139.4
## Number of obs: 478, groups: cusn, 210
##
## Fixed effects:
## Estimate Std. Error df t value Pr(>|t|)
## (Intercept) 196.17 13.38 235.50 14.663 <2e-16 ***
## price_c -46.85 23.95 420.65 -1.956 0.0511 .
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
## (Intr)
## price_c -0.007
```

```
icc(baseline_model)
```

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

We fit a baseline model with just the price as explanatory variable. After that we use this model to look at the inter-class correlation to check weather we have a nested/clustered dataset. The ICC strongly indicates that the data is heavily clustered and by that justifies the employment of multi-level regression. The ICC indicates that about 56% of the variation in bought amount can be explained by the differences in the clusters.

The baseline model could not find a significant effect of price, however the p-value is just slightly above the 5% mark, meaning the effect of price can still be arguable. The model found a negative effect of price. An increase of the price of 1 unit decreases the bought quantity by 46. That is a very strong slope meaning that costumers are very price sensitive

#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)
summary(cash_model)
```

```
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: quan ~ price_c + cash_dummy + (1 | cusn)
## Data: detailed_data_prep
```

```
##
## REML criterion at convergence: 6256
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -2.8973 -0.3553 -0.1545  0.2139  7.0508
##
## Random effects:
##   Groups   Name      Variance Std.Dev.
##   cusn     (Intercept) 24205    155.6
##   Residual                19462    139.5
## Number of obs: 475, groups:  cusn, 208
##
## Fixed effects:
##              Estimate Std. Error    df t value Pr(>|t|)
## (Intercept)    155.73      27.92 196.87   5.578 7.96e-08 ***
## price_c        -52.02      24.62 418.19  -2.113  0.0352 *
## cash_dummy      50.24      31.78 207.12   1.581  0.1154
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##              (Intr) pric_c
## price_c      -0.032
## cash_dummy   -0.879  0.039
```

```
summary(cash_model_moderation)
```

```
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: quan ~ price_c * cash_dummy + (1 | cusn)
##   Data: detailed_data_prep
##
## REML criterion at convergence: 6244.8
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -2.9037 -0.3612 -0.1468  0.2135  7.0726
##
## Random effects:
##   Groups   Name      Variance Std.Dev.
##   cusn     (Intercept) 24265    155.8
##   Residual                19425    139.4
## Number of obs: 475, groups:  cusn, 208
##
## Fixed effects:
##              Estimate Std. Error    df t value Pr(>|t|)
## (Intercept)    154.108      27.972 198.100   5.509 1.11e-07 ***
## price_c         -7.436      46.155 367.465  -0.161  0.872
## cash_dummy       51.598      31.820 207.824   1.622  0.106
## price_c:cash_dummy -62.214      54.549 387.799  -1.141  0.255
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Correlation of Fixed Effects:
##           (Intr) pric_c csh_dm
## price_c      -0.060
## cash_dummy  -0.879  0.053
## prc_c:csh_d  0.051 -0.846 -0.038
```

When we control for cash the effect of price becomes with a p-value of about 3.5% significant. The effect size is even estimated to be bigger, as the model found a 52 unit decrease for a one unit increase of price. The effect of cash itself is not significant.

The moderated model could not find any significant effect. This could stem from multicollinearity that sabotages the model. We can see a strong correlation between price and price*cash of -0.86, which leads to serious estimation problems, as the real effect cannot be assigned to any of these variables. Also supported by the increased standard error of estimation for both.

```
r2(cash_model)
```

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

```
r2(cash_model_moderation)
```

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

```
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
```

```
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)
summary(estb_model)
```

```
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: quan ~ price_c + estb + (1 | cusn)
##   Data: detailed_data_perep_estb
##
## REML criterion at convergence: 5619.2
##
## Scaled residuals:
```

```
##      Min      1Q  Median      3Q      Max
## -2.8652 -0.3528 -0.1449  0.2054  6.9224
##
## Random effects:
##   Groups   Name                Variance Std.Dev.
##   cusn      (Intercept) 23242      152.5
##   Residual                20108      141.8
## Number of obs: 428, groups:  cusn, 172
##
## Fixed effects:
##              Estimate Std. Error      df t value Pr(>|t|)
## (Intercept)   278.480    39.314   192.850   7.083 2.57e-11 ***
## price_c       -52.837    25.067   380.051  -2.108  0.0357 *
## estbs        -101.126    42.219   199.729  -2.395  0.0175 *
## estbsf         8.887     63.078   169.282   0.141  0.8881
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##          (Intr) pric_c estbs
## price_c  0.017
## estbs   -0.924 -0.031
## estbsf  -0.624 -0.042  0.577
```

```
summary(estb_model_moderation)
```

```
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: quan ~ price_c * estb + (1 | cusn)
## Data: detailed_data_perep_estb
##
## REML criterion at convergence: 5597.8
##
## Scaled residuals:
##      Min      1Q  Median      3Q      Max
## -2.8535 -0.3454 -0.1449  0.1992  6.8785
##
## Random effects:
##   Groups   Name                Variance Std.Dev.
##   cusn      (Intercept) 23286      152.6
##   Residual                20177      142.0
## Number of obs: 428, groups:  cusn, 172
##
## Fixed effects:
##              Estimate Std. Error      df t value Pr(>|t|)
## (Intercept)   279.75    39.41   193.95   7.098 2.32e-11 ***
## price_c       -15.77    62.58   353.28  -0.252  0.8011
## estbs        -102.39    42.30   200.39  -2.420  0.0164 *
## estbsf         10.84    63.35   171.03   0.171  0.8644
## price_c:estbs  -37.17    69.47   360.59  -0.535  0.5929
## price_c:estbsf -77.60    90.79   348.53  -0.855  0.3933
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Correlation of Fixed Effects:
##          (Intr) pric_c estbs  estbsf pric_c:stbs
## price_c      0.053
## estbs       -0.924 -0.049
## estbsf      -0.622 -0.033  0.575
## pric_c:stbs -0.050 -0.901  0.038  0.031
## prc_c:stbsf -0.036 -0.689  0.034 -0.037  0.621
```

```
r2(estb_model)
```

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

```
r2(estb_model_moderation)
```

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

```
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
```

Task 3

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.

Conclusion

The results of the logistic regression analysis show that the three selected predictors do not have any statistically significant explanatory power when it comes to predicting whether a county experiences a store closure. A larger dataset would be interesting to work with, as it would allow for a more reliable model fit and greater statistical power. Moreover, with a larger sample, the predictive performance on unseen data could be evaluated using a train-test split.