Problem Set I Solution

Tobias Bodentien Philipp Grunenberg Alexander Haas Osama Warshaga

21-11-2025

Task 1

The subsequent descriptive analysis is conducted on the detailed fish market data regarding whiting. First, the required packages and dataset are loaded.

```
## Load packages and dataset
library(readr)
library(dplyr)
library(ggplot2)
library(tidyr)
library(lmtest)

detailed_data <- read_tsv("../data/detailed_fish_market_data.txt")
```

Data preparation

The descriptive analysis will focus on price per pound (*pric*), the total quantity received (*totr*) and sold (*tots*) by a dealer in pounds per day in April and May 1992. Therefore, observations for which these variables are *'NA'* are removed via the 'dplyr' package. The dataset has also been filtered for Whiting, as required for Task 1, and arranged by date.

```
# delete those rows that have NA for
# "price","quan", "totr", "tots" and
# filter for whiting (no king)
detailed_whiting <- detailed_data %>%
  filter(!is.na(pric),
         !is.na(totr),
         !is.na(tots),
         type == "w") %>%
  arrange(date)

print(paste("Obs:", nrow(detailed_whiting),
            "- Vars:", ncol(detailed_whiting)))
```

```
## [1] "Obs: 478 - Vars: 17"
```

The cleaned whiting dataset now comprises 478 observations across 17 variables.

Inspecting the dataset via *'View(detailed_whiting)'* reveals that there are multiple observations for each day in the period from April to May 1992. Each observation depicts a transaction between a customer and the fish dealer. This is particularly important for understanding the next step.

Of the 478 observations, two do not fit. The total quantity that the dealers sold (*tots*) is equal for all rows (i.e. transactions) of the same dealer at a given day.

What immediately stands out is that only one dealer is observed on almost all days. For this dealer, multiple transactions are always recorded. Only on two days are two dealers observed, as indicated by multiple distinct observations for *tots*. However, these additional dealers are unrepresentative, as there is only one observation for each of them. The remaining observations all appear to come from a single, larger dealer. These two observations are therefore deleted using 'dplyr'.

```
## There seem to be two entries in the dataset, where there are two dealer per day.
# Since this is the case only for two out of all days in April and May:
# drop those two observations
detailed_whiting <- detailed_whiting %>%
  # frequency of the same tots value for different days
  group_by(date, dayw, tots) %>%
  mutate(n_same_tots = n()) %>%

  # number of distinct tots days
  group_by(date, dayw) %>%
  mutate(n_tots_values = n_distinct(tots)) %>%
  ungroup() %>%

  # delete rows for which (there are multiple different tots values
  #                                   AND
  #                         for which the tot value only appears once)
  filter(!(n_tots_values > 1 & n_same_tots == 1)) %>%

  # delete rows that are not longer needed
  select(-n_same_tots, -n_tots_values)

## two cases, where > 1 dealer is present
tots_inconsistent <- detailed_whiting |>
  group_by(date, dayw) |>
  mutate(
    n_tots = n_distinct(tots)
  ) |>
  filter(n_tots > 1) |>
  arrange(date, dayw, tots, totr)
```

With this step the data cleaning process is finished.

One part of the descriptive analysis of the Whiting data involves analysing data aggregated at a daily level. For this purpose, a new dataset, *detailed_whiting_daily*, is constructed.

```
# dataset for the daily-level
detailed_whiting_daily <- detailed_whiting %>%
  group_by(date) %>%
  summarise(
    avg_pric = mean(pric),
    totr = first(totr),
    tots = first(tots),
    dayw = first(dayw),
    n_trsact = n(),
    strate = first(tots)/first(totr)
  )
```

For each date (*date*) present in the original dataset, the average price (*avg_pric*), *totr*, *tots*, number of transactions (*n_trsact*) and sell-through rate (*strate*) are computed. For any computation involving *totr* or *tots* the first value of the day can be used. This is possible due to the previous data cleaning step, as described above. The sell-through rate measures the proportion of Whiting offered for sale on a given day that is actually sold. More on that later.

Descriptive analysis

The analysis is split into three sections: Firstly, an analysis of daily sales; secondly, an analysis of prices; and finally, an analysis based on the hours of the day.

In a code chunk that is not printed in this PDF (*'echo=FALSE'*) the different font sizes and some technical variables for the plots are defined. All plots were created using the 'ggplot2' package.

```
####
# summary of the daily datset
####
detailed_whiting_daily %>%
  select(totr, tots, n_trsact) %>%
  summary()
```

```
##       totr            tots          n_trsact
##  Min.   :  200   Min.   :  490   Min.   : 4.00
##  1st Qu.: 1990   1st Qu.: 3360   1st Qu.:18.00
##  Median : 6080   Median : 5535   Median :25.00
##  Mean   : 5881   Mean   : 5760   Mean   :25.05
##  3rd Qu.: 7927   3rd Qu.: 7495   3rd Qu.:32.50
##  Max.   :15940   Max.   :15455   Max.   :57.00
```

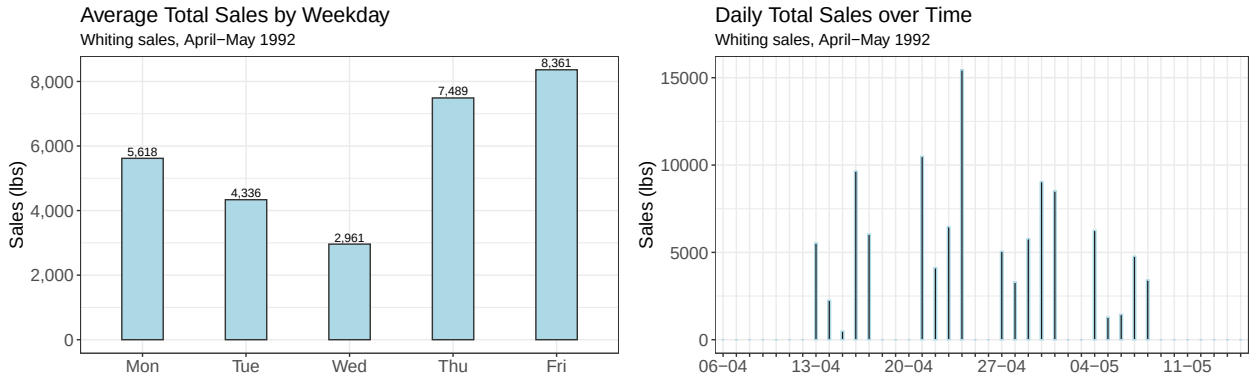
```
detailed_whiting_daily %>%
  select(totr, tots, n_trsact) %>%
  summarise(
    across(everything(), \(x) sd(x, na.rm = TRUE))
  )
```

```
## # A tibble: 1 x 3
##    totr  tots n_trsact
##   <dbl> <dbl>    <dbl>
## 1 4381. 3691.     12.3
```

Insight 1: The summary statistics for *totr*, *tots* and *n_trsact* at a daily level show that the central tendencies of *tots* and *totr* are quite similar, as expected. Furthermore, both exhibit a significant degree of variation, with standard deviations of 4,381 and 3,691, respectively. The number of daily transactions also varies considerably, ranging from very quiet to very busy days (minimum of four and a maximum of 57). Overall, these findings suggest substantial day-to-day volatility in the Whiting business.

Sales analysis

```
plot_average_sales_by_weekday
plot_daily_sales_over_time
```

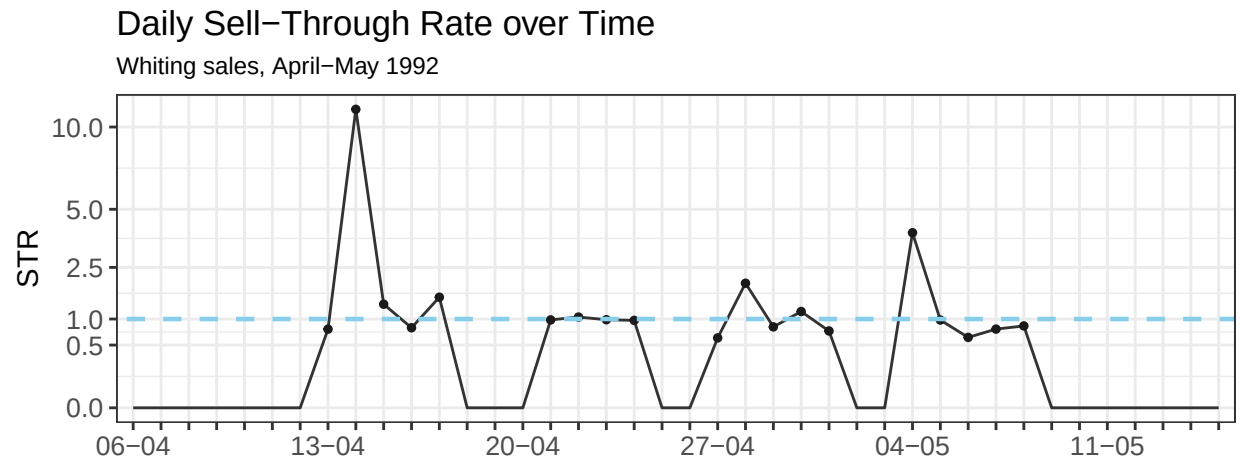


Insight 2: The plot on the left shows the average weekly total sales of whiting by day of the week. Sales are lowest on Wednesdays, increasing towards the end of the week with notably higher average volumes on Thursdays and Fridays. This suggests that demand for whiting is strongest just before the weekend.

The right-hand plot shows the total daily sales over the period April–May 1992. It is immediately clear from this plot that only 19 distinct days are observed throughout this period. As this results in a relatively small number of observations at the level of individual days, a boxplot is not used for the analysis. There is considerable fluctuation in sales from day to day, with several pronounced spikes and some much quieter days. This confirms a high level of volatility in daily whiting sales over the observed period. Another interesting observation is that the market was closed on Monday 20 April 1992. This resulted in unusually high sales the next day compared to other Tuesdays in the sample.

Please note, that the code for all plots of Task 1 can be accessed in the '*.Rmd*' file of our Problem Set 1 solution.

```
plot_daily_str_over_time
```



Insight 3: This line chart shows the daily sell-through rate per day over time. It is calculated as the ratio of *tots* to *totr* (i.e. *tots* divided by *totr*). The dashed line at one indicates full sell-through of the day's deliveries in Whiting.

On most trading days, the sell-through rate fluctuates around this value, suggesting that sales and incoming supply are roughly balanced. However, there are a few extreme spikes, especially in mid-April and early May, where the sell-through rate is far above one. These indicate days on which much more was sold than was delivered, meaning existing inventory must have been used up, as can be seen in the previous days spikes (below one). Overall, the figure indicates highly volatile, yet fairly efficient, inventory usage. This is possible because the Whiting can be frozen and sold over the next few days.

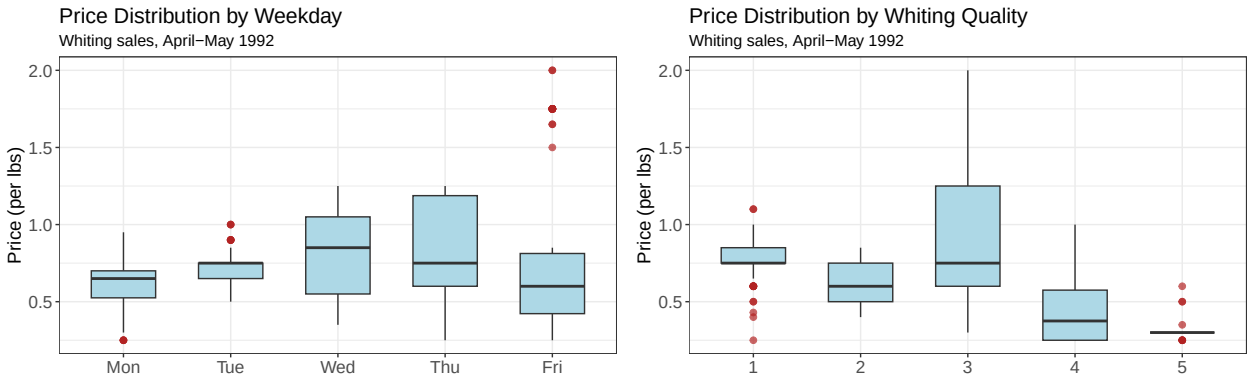
Price analysis

```
####
# correlation between tots (total daily sales) and avg_pric (average price)
####
cor(detailed_whiting_daily$avg_pric,detailed_whiting_daily$tots)
```

```
## [1] -0.4236805
```

With a correlation of around –0.42, there is a moderate negative linear relationship between average daily prices (*avg_pric*) and total daily sales (*tots*). On days with higher prices, total whiting sales tend to be lower.

```
plot_price_distr_by_weekday
plot_price_distr_by_quality
```



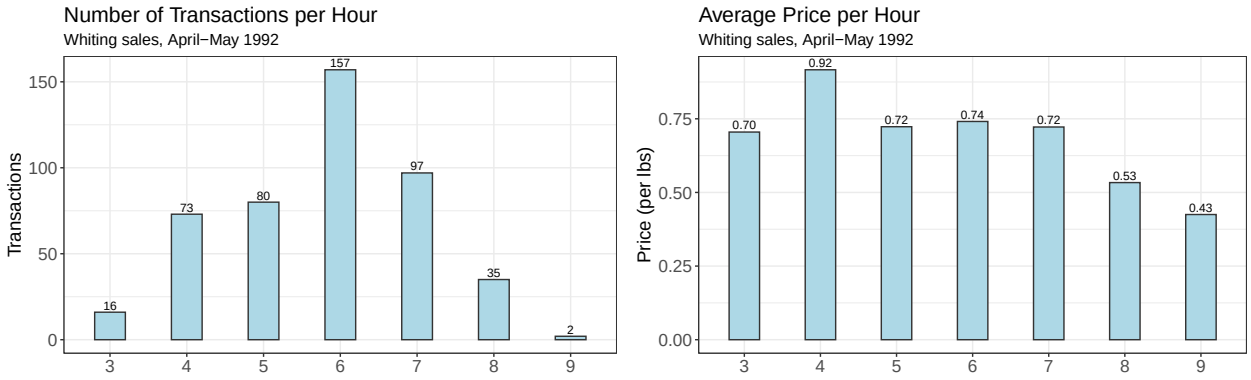
Insight 4: The boxplots of the price per pound of whiting for different weekdays (left-hand side) and different qualities of whiting (right-hand side) are both calculated at transaction level (i.e. using *detailed_whiting*).

Prices tend to be lower and less dispersed at the beginning of the week (Monday–Tuesday), becoming higher and more dispersed on Wednesday and especially Thursday. Friday's prices are usually similar to those in the middle of the week, but there are a few very high outliers. This could indicate occasional 'premium' pricing at the end of the week.

The figure on the right shows an unexpected quality effect: higher-quality grades of whiting achieve higher prices, at least in terms of the median. However, the price of grade three is similar to that of grade one and shows substantially higher variability. The maximum price for grade three is also very high compared to grade one. As expected, grades four and five fetch clearly lower prices. This could indicate that customers do not expect Whiting to be a 'premium' product, and that good quality is sufficient.

Analysis by time of day

```
plot_transactions_per_hour
plot_average_price_per_hour
```



Insight 5: The figures on the number of transactions and average price per hour confirm the typical dynamics of a fish market. The market opens at night and peak activity is observed around 6 a.m., after which the transaction volume declines rapidly. According to the bar plot on the right, prices remain relatively stable during the early morning, but then drop sharply at around 8 a.m., in line with the decline in trading activity.

Conclusion

Overall, the analyses provide a fairly consistent picture of the Whiting market. Nothing stands out as being different from what would be expected in a non-premium fish market.

Task 2: Nonlinear Regression

In Exercise 2, we examine the relationship between the quantity of sold fish and its price. First, a linear regression is conducted, and outliers are identified using three different methods: leverage, studentized residuals, and Cook's distance. In the second part of the exercise, nonlinear regressions are performed, using polynomial regression and logarithmic transformation. The results of the different regression models are then compared based on relevant criteria.

```
# For easier understanding of the code name the log values of price and quantity respectively:
library(readr)
library(dplyr)
daily_data <- read_tsv("../data/daily_fish_market_data.txt")
daily_data <- daily_data %>%
  rename(price_log=price, qty_log=qty)
# Now calculate price and quantity.
daily_data$price <- exp(daily_data$price_log)
daily_data$qty <- exp(daily_data$qty_log)
# The values of price and qty are almost identical to pricelevel and tots in most cases. Nevertheless,
daily_data$qty[24] - daily_data$tots[24]
```

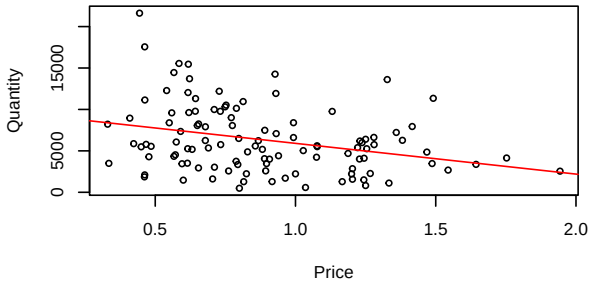
```
## [1] 119.9972
```

Task 2.1: Outliers

Before identifying the outliers we need to calculate the linear regression of quantity and price:

```
linear_reg = lm(daily_data$qty~ daily_data$price)
```

Looking at the plot we see, that some datapoints might influence the regression strongly or have a high residual. Outliers can lead to skewed statistical results or in the case of regression to overfitting.



To identify outliers three methods are used.

Leverage

To measure the influence of an individual observation on the final result the weight w_i of each obersvation i and mean are used. Observations with an higher influence on x have a higher wheigt. The resulting measure is called leverage and calculates the hatvalues h for each observation.

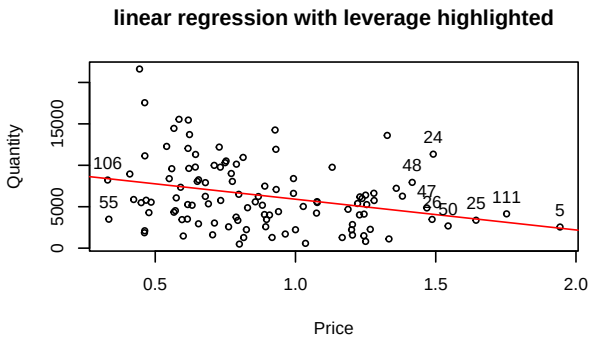
```
# Compute leverage h_i=1/n+w_i
h = hatvalues(linear_reg)
```

We now compare the highest h values with 2/n, with n being the number of observations. When all observation's hat values are h_i = 2/n, this means that they all influence the regression in the same amount.

```
## [1] "2/n = 0.018018018018018"
```

```
##           5         111          25          50          24          26          47
## 0.09973600 0.07001497 0.05567143 0.04424017 0.03880527 0.03840203 0.03660828
##         106          55          48
## 0.03387420 0.03348877 0.03189373
```

Observations 5, 111 and 25 deviate from 2/n substantially. This shows that those values have a much higher influence on the regression than other datapoints. Looking at the plot we see that those datapoints are the ones with the highest price and lowest quantity.



Studentized residuals

The idea of this method is that an outlier lies far away from the regression line and therefore has a high residual. To identify those outliers the studentized residuals are calulated:

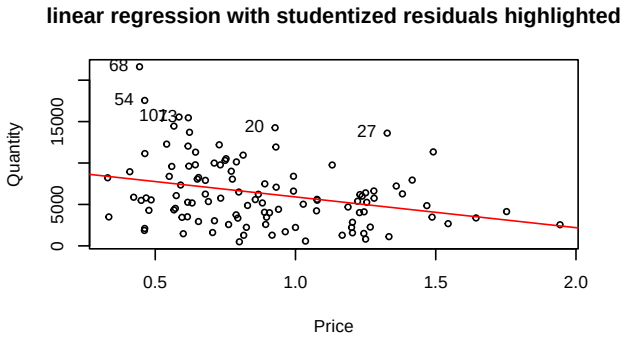
```
r_student <- rstudent(linear_reg)
```

Lets look at the highest and lowest values:

```
##       68       54       27       73      101       20       24       87
## 3.792982 2.595705 2.387418 2.158052 2.150648 2.137173 1.941637 1.832624
##       53       29
## 1.683596 1.508451
```

```
##        95        45        89        59        94
## -1.614107 -1.594851 -1.556401 -1.529101 -1.410866
```

We see that 6 values are above 1.96, which makes them stand out statistically.



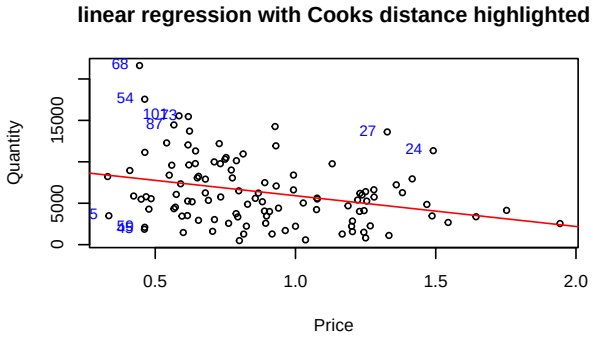
Cook's distance

The previous two methods either investigated the dependent variable (quantity) in the case of leverage or the independent variable (price) in the case of studentized residuals. Cook's distance combines both methods and considers the dependent and independent variable simultaneously.

```
# standard deviation of residuals:
sd_res = sd(resid(linear_reg))
# we want to calculate with the previously used residuals and h values
# therefore do not use the fallback of cooks.distance()
D <- cooks.distance(linear_reg, res = r_student, sd= sd_res, hat=h)
```

```
##           68           54           24           27          101           73
## 1.264998e-08 5.599887e-09 5.357574e-09 5.044084e-09 2.632765e-09 2.395923e-09
##           45           87           55           59
## 2.120240e-09 2.022045e-09 2.016477e-09 1.943305e-09
```

Very large D_i values indicate a substantial impact of observation i. In this case the values are low. The highest values are from observations 68 and 54, which also have significantly high studentized residuals. Observation 24 is among the highest values of studentized residuals, hat values and therefore also D values.



Interpretation:

As observed previously, the values with the highest h values are those that achieve the lowest quantity at the highest price. This aligns with the intuition that sales numbers decline as prices rise for general goods. For this reason, we do not assume that these values are outliers. Rather, we assume that they are influential values that occur infrequently yet represent the underlying distribution.

There are six values with statistically significant high studentized residuals. Since the studentized residuals diagnose the dependent variable, the quantity of fish sold stands out as unusually high compared to other observations. A closer look at the highest observations, 68 and 54, shows that not only the quantity sold but also the amount of fish received on those days was particularly high. Accordingly, this is also consistent with the intuition that when larger quantities are available, more is sold. Furthermore, these observations also have a comparatively low price.

Cook's distance is a combination of leverage and studentized residuals. Here as well, observations 68 and 54 show the highest values. For the reasons described above, these observations are not considered outliers and remain in the dataset. Observation 24 stands out across all three methods. This observation corresponds to a stormy Monday with a high amount of fish received. What is notable is the high price despite a large quantity of fish sold. However, this may be justified by the weather and the fact that it is a Monday. Since sales are generally higher on Mondays and restaurants could not purchase fish the day before, they may be more willing to pay a higher price, especially considering that the weather is expected to remain stormy for the following two days, which would likely result in a lower fish supply.

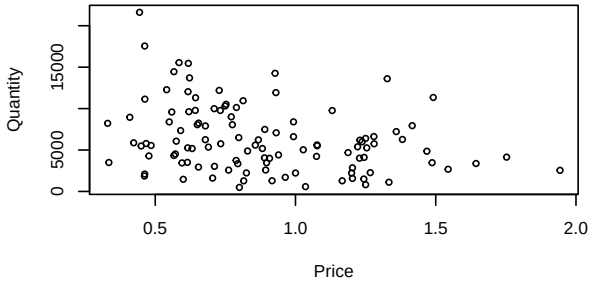
Therefore we do not identify any datapoints as outliers and will not delete any from the dataset.

Task 2.2: Models for the price-demand relationship

Until now we worked with the linear regression of quantity and price, which assumes a linear relationship between price and quantity. But often, the price-demand function is assumed to be quasi-linear. Therefore, in this task we will use variable transformation and polynomial regression to investigate whether non-linear regression fits better to the data. We decided to use a log-log model, lin-log model and squared polynomial model. The log-log model transforms the dependent and independent variable and therefore assumes the following underlying function: Quantity = alpha * Price ^ beta, beta < 0. The linear-log model only transforms the independent variable. Both functions can assume a quasi-linear shape and are therefore suitable for this analysis.

Regarding the polynomial regrassion, the following economical hypothesis can be made: Whiting is a fish that is very cheap compared to other fish. For low prices the demand can be assumed to be high and fall to a zone of indifference for mid-range prices, as the customer still wants fish. When the price is high and similar to other fish the demand falls, as people more likely consume fish of higher quality. Therefore, a s-shaped function with a zone of indifference can also be assumed with the following underlying function: quantity = alpha + beta_1 * x + beta_2 * x^2 + beta_3 * x^3, beta_3 < 0

Looking at the data a log-log and linear-log model can easily be assumed. An s-shaped curve can hardly be seen.



Although we do not see a polynomial function, the task requires to perform a polynomial regression. Therefore, firstly mean centering is conducted:

```
daily_data$price_MC = daily_data$price-mean(daily_data$price, na.rm=TRUE)
```

Then we compute the polynomials.

```
# 2.4.2 Compute polynomials
daily_data$price_MC_squared = daily_data$price_MC * daily_data$price_MC
daily_data$price_MC_cubic = daily_data$price_MC_squared * daily_data$price_MC
daily_data$price_MC_power4 = daily_data$price_MC_cubic * daily_data$price_MC
```

The linear model already explains about 9% of the variance (R-squared = 9,5% and Adjusted R-squared = 8,6%).

```
# running the linear model
summary(linear_reg)
```

```
##
## Call:
## lm(formula = daily_data$qty ~ daily_data$price)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6158.2 -2945.3  -170.3  2150.8 13651.4
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)           9615       1039   9.257 2.12e-15 ***
## daily_data$price     -3709       1099  -3.376  0.00102 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3862 on 109 degrees of freedom
## Multiple R-squared:  0.09465,    Adjusted R-squared:  0.08634
## F-statistic:  11.4 on 1 and 109 DF,  p-value: 0.001021
```

The AIC and BIC of the linear regression are:

```r
c(AIC(linear_reg), BIC(linear_reg))
```

```
## [1] 2152.457 2160.586
```

Lets look at the polynomial regression:

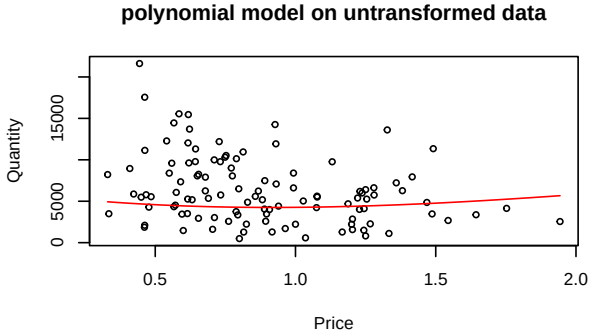
```r
# 2.4.3 Analyze and test hypothesized model
polynomial_regression_cubic = lm(daily_data$qty~ daily_data$price_MC+daily_data$price_MC_squared+daily_
summary(polynomial_regression_cubic)
```

```
##
## Call:
## lm(formula = daily_data$qty ~ daily_data$price_MC + daily_data$price_MC_squared +
##     daily_data$price_MC_cubic)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6486.2 -2843.2  -557.4  2077.1 13153.2
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    6028.6      525.7  11.468   <2e-16 ***
## daily_data$price_MC           -4208.3     1686.4  -2.495   0.0141 *
## daily_data$price_MC_squared    2837.6     4229.4   0.671   0.5037
## daily_data$price_MC_cubic      -394.6     5671.7  -0.070   0.9447
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3881 on 107 degrees of freedom
## Multiple R-squared:  0.1024, Adjusted R-squared:  0.07721
## F-statistic: 4.068 on 3 and 107 DF,  p-value: 0.008844
```

```r
c(AIC(polynomial_regression_cubic), BIC(polynomial_regression_cubic))
```

```
## [1] 2155.505 2169.053
```

As we now have more independent variables in the polynomial regression the R-squared is slightly higher than in the linear regression. Therefore, we use the adjusted R-squared to evaluate the polynomial regression, which is worse than that of the linear regression. Also AIC and BIC are worse. By visualizing the regression we see, that the assumed s-shape did not take form. The p-values also do not support the hypothesis.



There is no support for the above mentioned hypothesis. Therefore we do not analyze the extended model.

Now we investigate the log-log model: We do not need to transform the data manually, as we already have the log values of price and quantity. We do not have any complications as the minimum of price and quantity is above 0:

```
# the column price and qty or the daily dataset are already log values
min(daily_data$qty) # min > 0
```

```
## [1] 490.0003
```

```
min(daily_data$price) # min > 0
```

```
## [1] 0.330303
```

```
log_log_model = lm(daily_data$qty_log~ daily_data$price_log)
```

The results show, that 7,8% of the variance can be explained, although we cannot compare this value to the linear model as we now work with a transformed dependent variable. The p-value is lower that 5%. AIC and BIC have low values aswell. Indicating, that the log-log explains the data well.

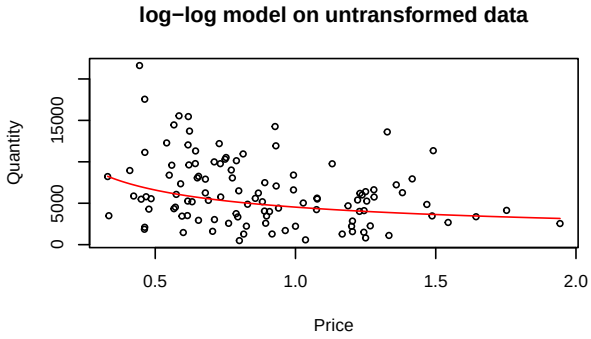
```
summary(log_log_model)
```

```
##
## Call:
## lm(formula = daily_data$qty_log ~ daily_data$price_log)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.3450 -0.3569  0.1193  0.4976  1.2528
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)            8.41867    0.07622 110.445  < 2e-16 ***
## daily_data$price_log -0.54087    0.17864  -3.028  0.00308 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7156 on 109 degrees of freedom
## Multiple R-squared:  0.07758,    Adjusted R-squared:  0.06912
## F-statistic: 9.167 on 1 and 109 DF,  p-value: 0.003075
```

```
c(AIC(log_log_model), BIC(log_log_model))
```

```
## [1] 244.6919 252.8205
```

Looking at the plot of the log-log model on the untransformed data this analysis is supported:



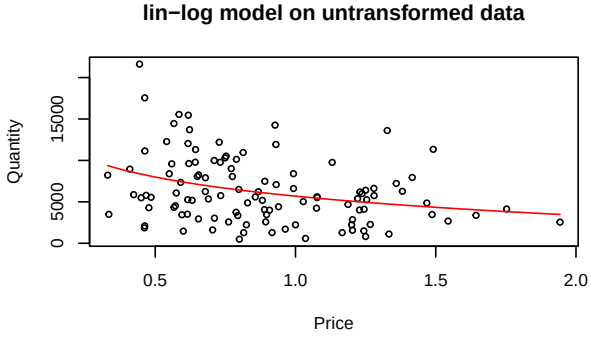
Additionally, we calculate the linear-log model, which compared to the linear regression has better results, as both the R-squared and adjusted R-squared values are higher.

```
# the column price and qty or the daily dataset are already log values
lin_log_model = lm(daily_data$qty~ daily_data$price_log)
summary(lin_log_model)
```

```
##
## Call:
## lm(formula = daily_data$qty ~ daily_data$price_log)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6403.1 -2860.4  -614.4  2154.0 13227.9
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)             5690.1      410.4   13.87  < 2e-16 ***
## daily_data$price_log  -3327.8      961.7   -3.46 0.000772 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3853 on 109 degrees of freedom
## Multiple R-squared:  0.09897,    Adjusted R-squared:  0.0907
## F-statistic: 11.97 on 1 and 109 DF,  p-value: 0.0007718
```

```
c(AIC(lin_log_model), BIC(lin_log_model))
```

```
## [1] 2151.926 2160.055
```



Conclusion

The polynomial regression did not support the hypothesis of a polynomial function. We have seen that the linear regression already explains 8.6% of the variance in the model (adjusted R-squared). When we compare the linear model with the quasi-linear models, we find that the linear-log (Lin-log) model explains 9.1% of the variance. The AIC and BIC values of the two models differ only minimally. Comparing the log-log model to the Lin-log model, it can be observed that the Lin-log model indicates a higher demand at a lower price than the log-log model. As the price increases, the quantity of fish sold in the Lin-log model decreases more than in the log-log model, eventually converging to a very similar demand level at higher prices. Since the Lin-log model allows better comparability with the linear regression and indicates higher demand at lower prices, we consider the Lin-log model to best describe the data.

Task 3: Moderated Regression

3.1 Hypotheses

In this task, we investigate whether the price sensitivity of individual customers depends on specific contextual characteristics of the transaction. Based on plausibility considerations, the following three hypotheses were developed:

Hypothesis 1 (H1): Moderation by Product Quality (`qual`)

- **Hypothesis:** The quality of the fish has an influence on the price sensitivity of customers.
- **Rationale:** We expect that higher quality (`qual`) leads to *lower* price sensitivity. A high-quality product, which might be sold to expensive restaurants with higher margins, justifies a higher price and makes customers less susceptible to price fluctuations.

Hypothesis 2 (H2): Moderation by Payment Method (`cash`)

- **Hypothesis:** The use of cash versus charge (invoice) influences price sensitivity.
- **Rationale:** Paying with cash may have a higher "emotional value" (or "pain of paying") as the amount paid is immediately visible, rather than just appearing on a bill later. We, therefore, expect that cash transactions lead to *higher* price sensitivity.

Hypothesis 3 (H3): Moderation by Establishment Type (`estb`)

- **Hypothesis:** Price sensitivity depends on the customer's type of establishment.
- **Rationale:** "Fry shops" (`f`) likely operate on lower margins for their final products and thus have a stronger incentive to watch purchase prices than "Stores" (`s`). We, therefore, expect that "fry shops" will exhibit *higher* price sensitivity.

```
#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) %>%

  # 1.1 standardization
  group_by(cusn) %>%
  mutate(Qty_Dev = quan - mean(quan, na.rm = TRUE)) %>% #target variable
  ungroup() %>%

  # 1.2 mean centering and dummy creation
  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
  )
```

This is the first cleaning step.

- We remove any rows where either price (`pric`) or quantity (`quan`) are missing. We cannot model a price-demand relationship without a price or a quantity, so these rows are unusable for our model.
- We filter the dataset to only include "Whiting". This ensures our analysis is focused on a single product, as combining different fish types would introduce confounding variables.
- We sort the data ascending by date. While not strictly required for this regression, it's good practice to organize time-series data chronologically, which can help in identifying patterns or debugging later.
- Next we standardize the consume by customer. This is a crucial step as each customer usually buys different amounts of fish. Therefore we take the mean for every customer and model the amount consumed by the difference to the mean for this customer.
- Finally we create the main Regressor and the 3 Regressors for the hypothesis.
  - We mean center `pric` (price) . This makes the coefficients in our moderated regression models much easier to interpret. Specifically, the main effect of price will now represent the price sensitivity at the *average* level of the moderator.
  - We likewise mean-center our first moderator, `qual` (quality).
  - We convert the categorical `cash` variable into a numeric **dummy variable** for our second hypothesis. The model can interpret "1" (for cash) and "0" (for non-cash), but it cannot interpret the original 'c' and '0' letters.
  - We convert the establishment type (`estb`) variable into a **factor**. This tells R that `estb` is a categorical variable. When we include it in the `lm()` function, R will automatically create the necessary dummy variables for each establishment type, allowing us to test our third hypothesis.

```
# additional steps for establishment
detailed_data_prep %>% group_by(estb) %>% count()
```

```
## # A tibble: 9 x 2
## # Groups:   estb [9]
##   estb      n
##   <fct> <int>
## 1 d         2
## 2 f        81
## 3 fd        1
## 4 s       307
## 5 sd        2
## 6 sf       40
```

```
## 7 sh        1
## 8 sr        3
## 9 <NA>     41
```

```
detailed_data_perep_estb = detailed_data_prep %>%
  filter(estb %in% c("s", "f", "sf"))
```

This code block performs a crucial diagnostic check and a subsequent filtering action specifically to prepare for testing Hypothesis 3 (moderation by establishment type).

- Before using a categorical variable as a moderator, we must check its distribution. The output of this count reveals that while some categories like 's' (store) and 'f' (fry shop) have many observations, other categories have very few (e.g., only 1, 2, or 3).
- Attempting to run a regression or moderation analysis on a categorical level with only 1 or 2 observations is statistically unreliable; the model cannot produce a stable estimate for such a small group. It can lead to model errors or highly misleading results. By filtering down to the well-represented groups, we ensure that our analysis for Hypothesis 3 is robust and that the results are meaningful. This new dataset will be used *only* for the H3 analysis.

```
#2.1 Moderated Model 1
quality_model = lm(Qty_Dev ~ price_c + quality_c, data = detailed_data_prep)
quality_model_moderation = lm(Qty_Dev ~ price_c*quality_c, data = detailed_data_prep)

summary(quality_model)
```

```
##
## Call:
## lm(formula = Qty_Dev ~ price_c + quality_c, data = detailed_data_prep)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -458.26  -23.08   -1.96   16.28  963.39
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.6419     5.0844   0.126    0.900
## price_c     -19.7349    15.4777  -1.275    0.203
## quality_c    -0.8560     4.7298  -0.181    0.856
##
## Residual standard error: 108.7 on 454 degrees of freedom
##   (21 Beobachtungen als fehlend gelöscht)
## Multiple R-squared:  0.003572,   Adjusted R-squared:  -0.0008177
## F-statistic: 0.8137 on 2 and 454 DF,  p-value: 0.4439
```

```
summary(quality_model_moderation)
```

```
##
## Call:
## lm(formula = Qty_Dev ~ price_c * quality_c, data = detailed_data_prep)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -456.67  -23.21   -2.21   16.08  963.27
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)         0.0707     5.2992   0.013    0.989
## price_c           -15.1822    19.4567  -0.780    0.436
## quality_c          -1.9473     5.5113  -0.353    0.724
## price_c:quality_c  -8.9002    23.0116  -0.387    0.699
##
## Residual standard error: 108.8 on 453 degrees of freedom
##   (21 Beobachtungen als fehlend gelöscht)
## Multiple R-squared:  0.003901,   Adjusted R-squared:  -0.002696
## F-statistic: 0.5913 on 3 and 453 DF,  p-value: 0.6209
```

The results indicate that in both the basic and the moderated models, the estimated parameters are not statistically significant. The model's explanatory power is extremely low, with an R^2 below 1%. Consequently, we fail to reject the null hypothesis; there is no statistical evidence in this dataset that price, quality, or their interaction influences the quantity consumed.

```
#Testing if Moderator 1 is significant
anova(quality_model, quality_model_moderation)
```

```
## Analysis of Variance Table
##
## Model 1: Qty_Dev ~ price_c + quality_c
## Model 2: Qty_Dev ~ price_c * quality_c
##   Res.Df     RSS Df Sum of Sq      F Pr(>F)
## 1    454 5363293
## 2    453 5361522  1    1770.5 0.1496 0.6991
```

```
lrtest(quality_model, quality_model_moderation)
```

```
## Likelihood ratio test
##
## Model 1: Qty_Dev ~ price_c + quality_c
## Model 2: Qty_Dev ~ price_c * quality_c
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   4 -2789.6
## 2   5 -2789.5  1 0.1509     0.6977
```

Neither the ANOVA (F-test for nested models) nor the Likelihood Ratio Test (LRT) indicates a significant difference between the models ($p > 0.05$). This confirms that adding the interaction term does not improve model fit, meaning we find no support for the hypothesis that quality moderates price sensitivity (H1).

```
#2.2 Moderated Model 2
cash_model = lm(Qty_Dev ~ price_c+cash_dummy, data=detailed_data_prep)
cash_model_moderation = lm(Qty_Dev ~ price_c*cash_dummy, data=detailed_data_prep)
summary(cash_model)
```

```
##
## Call:
## lm(formula = Qty_Dev ~ price_c + cash_dummy, data = detailed_data_prep)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -455.49  -23.55   -0.94   14.87  965.08
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.9983     9.1851   0.109    0.913
## price_c     -18.2086    15.5898  -1.168    0.243
## cash_dummy   -1.5193    10.9672  -0.139    0.890
##
## Residual standard error: 108.2 on 472 degrees of freedom
##   (3 Beobachtungen als fehlend gelöscht)
## Multiple R-squared:  0.002882,   Adjusted R-squared:  -0.001343
## F-statistic: 0.6821 on 2 and 472 DF,  p-value: 0.5061
```

```
summary(cash_model_moderation)
```

```
##
## Call:
## lm(formula = Qty_Dev ~ price_c * cash_dummy, data = detailed_data_prep)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -456.53  -23.23   -0.50   16.07  965.35
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)         -0.3933     9.3002  -0.042    0.966
## price_c              7.1742    30.7550   0.233    0.816
## cash_dummy          -0.3788    11.0327  -0.034    0.973
## price_c:cash_dummy -34.1625    35.6796  -0.957    0.339
##
## Residual standard error: 108.2 on 471 degrees of freedom
##   (3 Beobachtungen als fehlend gelöscht)
## Multiple R-squared:  0.004819,   Adjusted R-squared:  -0.00152
## F-statistic: 0.7602 on 3 and 471 DF,  p-value: 0.5168
```

Similar to the previous analysis, the parameters in both the main effects and interaction models are statistically insignificant. The adjusted R^2 remains negligible ($< 1\%$). We cannot reject the null hypothesis that the payment method (cash vs. credit) has no impact on consumption behavior or price sensitivity.

```
#Testing if Moderator 2 is significant
anova(cash_model, cash_model_moderation)
```

```
## Analysis of Variance Table
##
## Model 1: Qty_Dev ~ price_c + cash_dummy
## Model 2: Qty_Dev ~ price_c * cash_dummy
##   Res.Df     RSS Df Sum of Sq      F Pr(>F)
## 1    472 5526664
## 2    471 5515928  1     10736 0.9168 0.3388
```

```
lrtest(cash_model, cash_model_moderation)
```

```
## Likelihood ratio test
##
## Model 1: Qty_Dev ~ price_c + cash_dummy
## Model 2: Qty_Dev ~ price_c * cash_dummy
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   4 -2897.4
## 2   5 -2897.0  1 0.9237     0.3365
```

Both the ANOVA and LRT yield non-significant p-values. This suggests that the inclusion of the payment method interaction does not provide a better fit than the simple additive model. Thus, H2 is not supported.

```
#2.3 Moderated Model 3
estb_model = lm(Qty_Dev ~ price_c + estb, data = detailed_data_perep_estb)
estb_model_moderation = lm(Qty_Dev ~ price_c * estb, data = detailed_data_perep_estb)
summary(estb_model)
```

```
##
## Call:
## lm(formula = Qty_Dev ~ price_c + estb, data = detailed_data_perep_estb)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -457.01  -24.83   -0.38   17.31  966.03
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -1.465     12.577  -0.116    0.907
## price_c      -21.989     16.571  -1.327    0.185
## estbs          1.905     14.158   0.135    0.893
## estbsf         3.234     21.925   0.148    0.883
##
## Residual standard error: 112.7 on 424 degrees of freedom
## Multiple R-squared:  0.004136,   Adjusted R-squared:  -0.00291
## F-statistic: 0.5869 on 3 and 424 DF,  p-value: 0.6238
```

```
summary(estb_model_moderation)
```

```
##
## Call:
## lm(formula = Qty_Dev ~ price_c * estb, data = detailed_data_perep_estb)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -457.49  -24.14   -1.02   15.34  963.48
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)      0.9751    12.8901   0.076    0.940
## price_c         13.8828    43.7572   0.317    0.751
## estbs           -0.4756    14.4158  -0.033    0.974
```

```
## estbsf           2.5946    22.3083   0.116    0.907
## price_c:estbs  -38.7734    47.9334  -0.809    0.419
## price_c:estbsf -58.2433    62.6070  -0.930    0.353
##
## Residual standard error: 112.9 on 422 degrees of freedom
## Multiple R-squared:  0.006358,   Adjusted R-squared:  -0.005415
## F-statistic:  0.54 on 5 and 422 DF,  p-value: 0.746
```

The analysis of establishment types yields comparable results. None of the coefficients for price, establishment type, or their interaction terms achieve statistical significance. The low R^2 values suggest that these variables do not effectively predict deviations in purchase quantity.

```
#Testing if Moderator 3 is significant
anova(estb_model, estb_model_moderation)
```

```
## Analysis of Variance Table
##
## Model 1: Qty_Dev ~ price_c + estb
## Model 2: Qty_Dev ~ price_c * estb
##   Res.Df     RSS Df Sum of Sq      F Pr(>F)
## 1    424 5388709
## 2    422 5376685  2     12023 0.4718 0.6242
```

```
lrtest(estb_model, estb_model_moderation)
```

```
## Likelihood ratio test
##
## Model 1: Qty_Dev ~ price_c + estb
## Model 2: Qty_Dev ~ price_c * estb
##   #Df  LogLik Df Chisq Pr(>Chisq)
## 1   5 -2627.6
## 2   7 -2627.1  2 0.956       0.62
```

The model comparison tests (ANOVA and LRT) evaluate the collective contribution of the interaction terms associated with the `estb` factor. The results are non-significant, indicating that the relationship between price and demand does not vary significantly across the different establishment types. Therefore, H3 is not supported.