

# QANT 530 Homework #2

Phan Nguyen

07/12/2022

For this homework, we are going to do a multiple regression analysis on a practice dataset with 150 rows and five columns. The response variable is AverageMonthlyOrders. Below is a sample of the dataset.

Region is categorical and assumably nominal. Number of employees, inventory value and average monthly orders are continuous.

##	Region	NumberOfEmployees	InventoryValue	AverageMonthlyOrders
## 1	1	4.000000	690.2320	277.0312
## 2	1	2.067246	786.1016	254.1911
## 3	1	4.713721	764.7923	244.4685
## 4	1	5.319617	766.5935	258.6049
## 5	1	5.687089	729.1106	237.5975
## 6	1	6.831811	687.9113	245.0445

Table 1: Dataset example

## 1.Exploratory Analysis on the response variable

### Descriptive numerical summary table

##	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis
## 1	1	150	325.45	49.6	324.44	325.07	55.31	214.03	455.6	241.57	0.13	-0.4
##	se	Q0.25	Q0.75									
## 1	4.05	288.8	365.07									

Table 2: Descriptive Statistics table for AverageMonthlyOrders

AverageMonthlyOrders has a logical range of values, from 214.03 to 455.06. Its median and mean are close, meaning its distribution is not skewed toward left or right. The interquartile range is between 288.8 and 365.07, meaning 50% of the values are in between this range. Standard deviation is 49.6, meaning 95% of values are two standard deviations from the mean, between 226.25 and 424.65, according to empirical rules.

Skewness for AverageMonthlyOrders are quite close to 0, suggesting that the distribution for the variable is not quite skewed. The measures of center also suggested the same. The excess kurtosis

is around -0.4. Excess kurtosis is equal to normal kurtosis - 3. In this case, the kurtosis shows that the distribution is close to a normal distribution. AverageMonthlyOrders's distribution has heavier tails and lower bell, due to its kurtosis to be lower than 0.

### Descriptive chart summaries

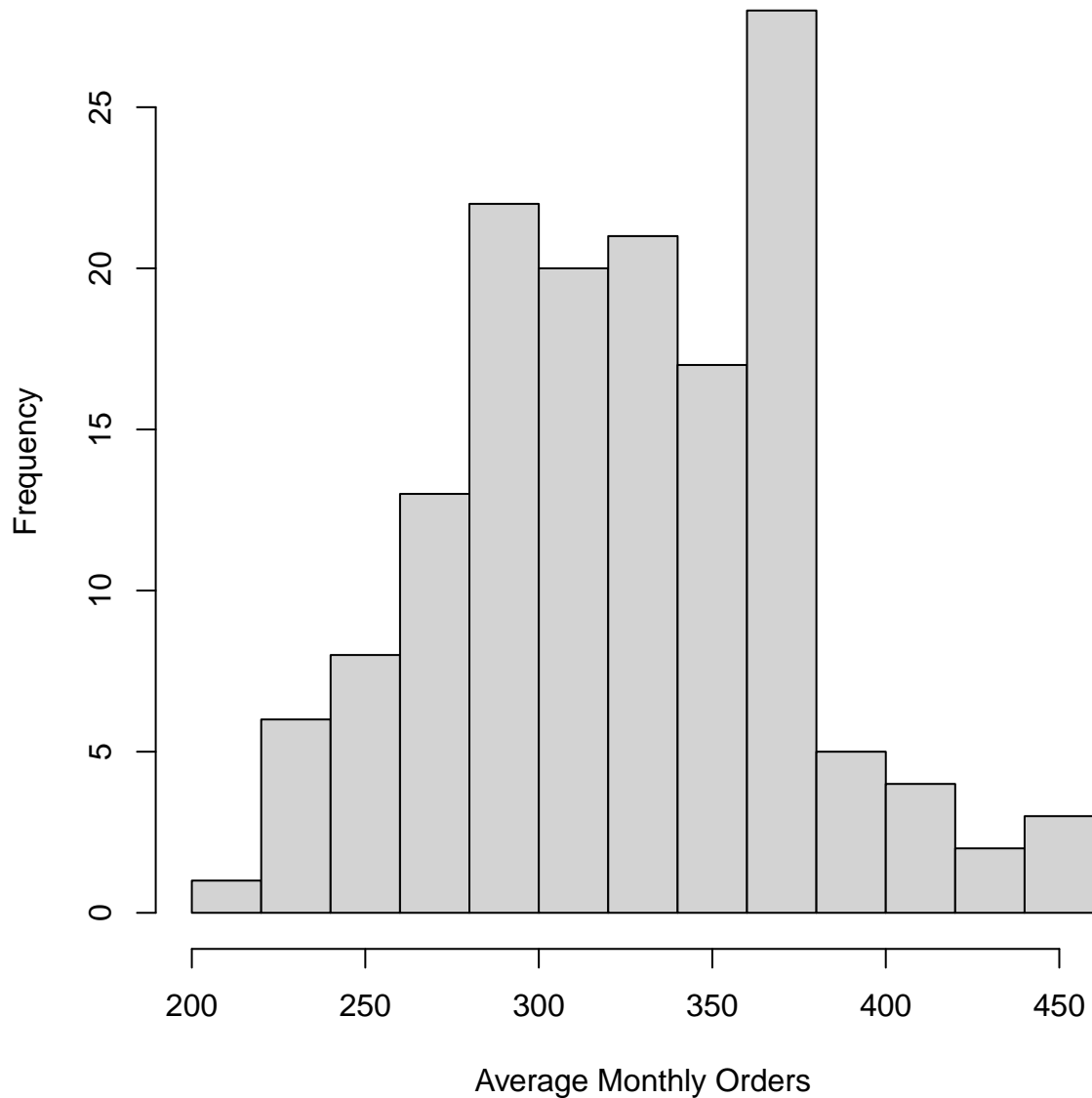


Figure 1: Histogram of AverageMonthlyOrders

This histogram shows the distribution for AverageMonthlyOrders. It does look like the distribution is bimodal, meaning there are two modes, thus two peaks in this histograms. The median and mean bin [300,350] (doesn't have a high frequency, which means this distribution is not normal. The right tail has lower frequency than the left tail.

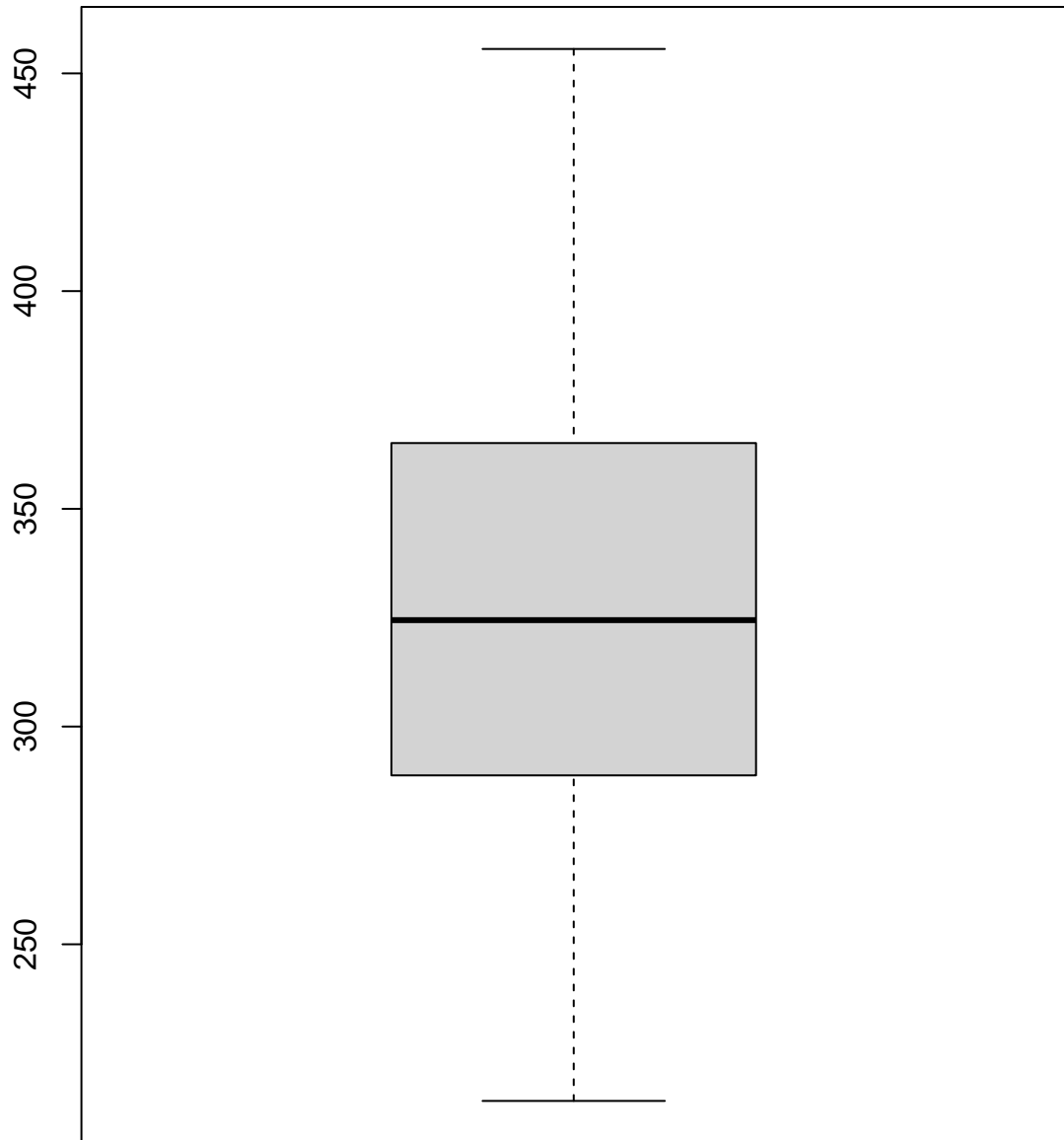


Figure 2: Boxplot of AverageMonthlyOrders

Visually, the boxplot of this variable looks normal. There are whiskers indicating the maximum and minimum values, corresponding to what we saw in the numerical descriptive statistics. The gray box represents the interquartile range, where the lower bound is the first quartile and the upper bound is the third quartile. The black line in the middle indicates the median. We can see that the median is in the middle of this gray box, meaning it has approximately the same range to the first and third quartiles. We can't tell that this variable has a mode different than the median, which is interesting.

## 2. Scatterplot analysis

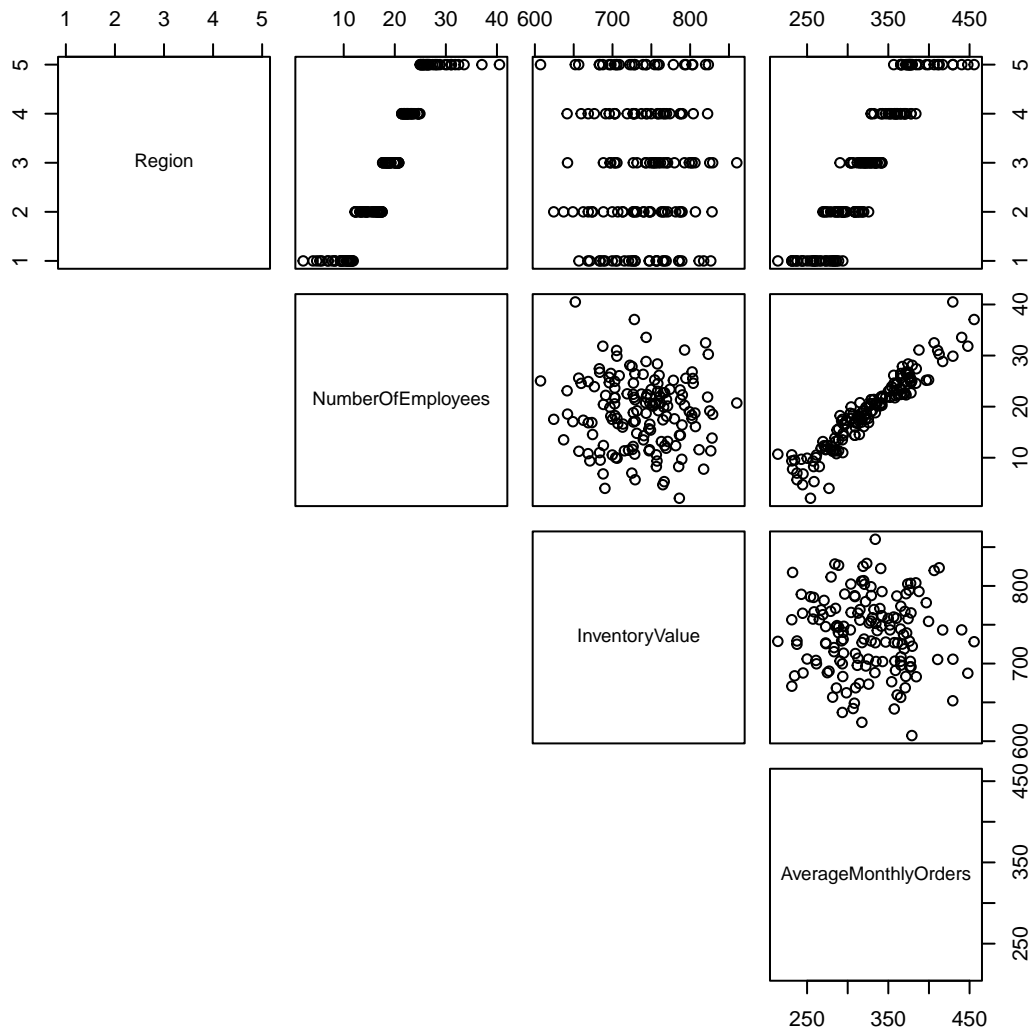


Figure 3: Scatterplot Matrix

Next, we create a scatterplot matrix between our response variable `AverageMonthlyOrders` and the explanatory variables. The scatterplot between `AverageMonthlyOrders` and `NumberOfEmployees` shows a clear linear upward trend. It looks like more employees bring more average monthly orders. There are also an upward trend between regions and average monthly orders, which is very interesting. Regions are nominal, so if we give the regions a different numbers, they are not going to change their meaning in the dataset. So it was interesting that coincidentally, they are sorted in a way that created a linear trend. However, since they are categorical, this doesn't tell us a lot about their correlation. There is no distinctive pattern between `InventoryValue` and `AverageMonthlyOrders`.

```
##               Region NumberOfEmployees InventoryValue
```

## Region	1.000000000	0.94077917	-0.003694931
## NumberOfEmployees	0.940779166	1.000000000	-0.043041229
## InventoryValue	-0.003694931	-0.04304123	1.000000000
## AverageMonthlyOrders	0.925238595	0.95017783	-0.032912135
##	AverageMonthlyOrders		
## Region	0.92523860		
## NumberOfEmployees	0.95017783		
## InventoryValue	-0.03291213		
## AverageMonthlyOrders	1.00000000		

Table 3: Correlation Matrix

We can see that there are strong correlation - near 1 - between number of employees and average monthly orders. The correlation between regions and average monthly orders is high as well, however, since Region is nominal, it doesn't mean much, Week correlation between AverageMonthlyOrders and InventoryValue.

It looks like the explanatory variables are not correlated, so there is no multicollinearity problem.

### 3. Multiple regression assumption check

First, we fit the data into a model to check all assumptions. Response variable is AverageMonthlyOrders.

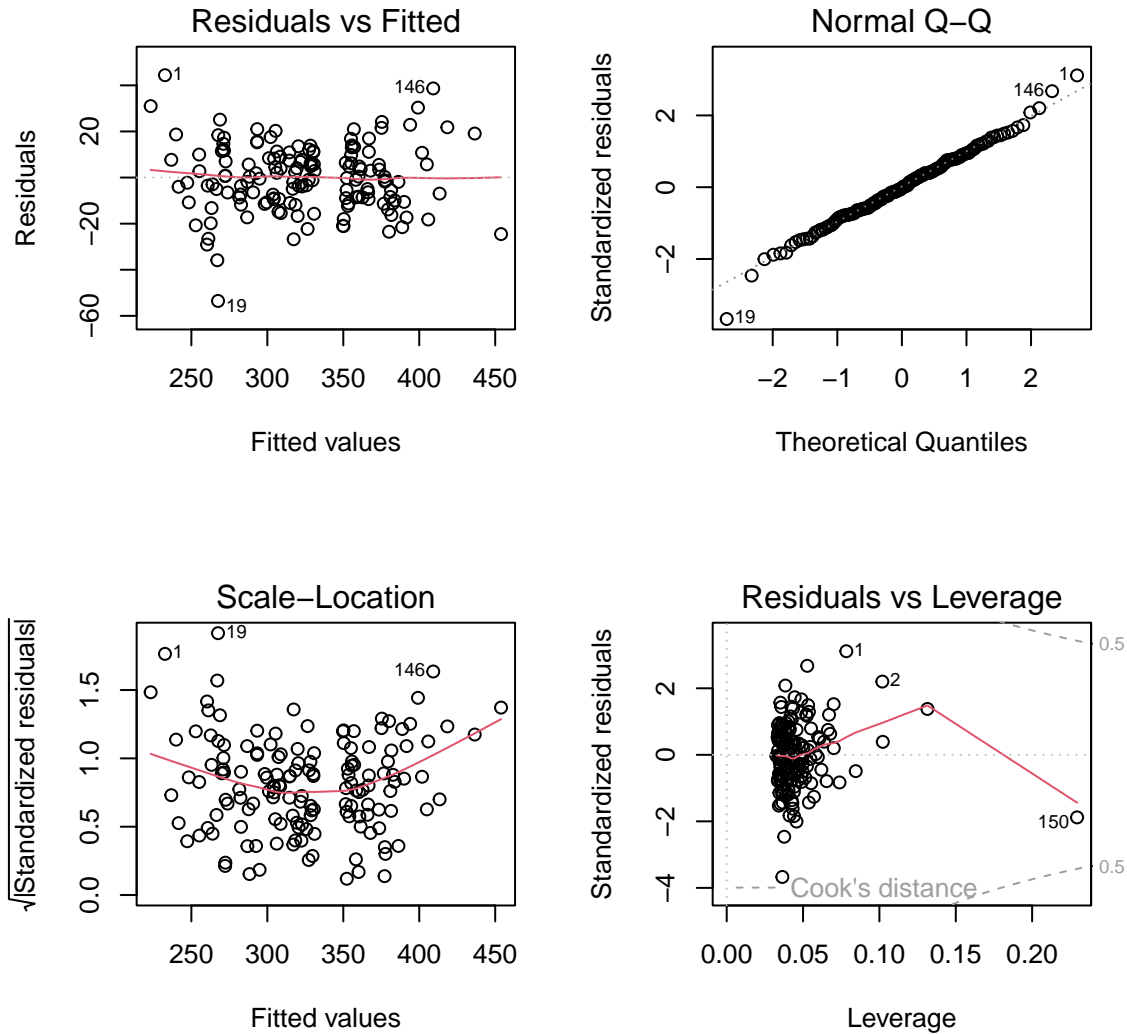


Figure 4: Simple Regression Plots for models

We can check the assumptions for the multiple regression model using these plots. The residual plot on the left hand side, shows how the predicted versus observed values. Ideally, the red line will stay horizontally at 0. In this case, the red line is approximately horizontal and at line 0. The points are scattered and don't follow any pattern. We can conclude that the **linearity and equal variances assumptions are met**.

Independence of variance can be checked using the Scale-Location plot, on the bottom left hand. The red line supposed to be horizontally at zero for the assumption to be met. We can say that **the independence of variance assumption is met**.

The normal Q-Q plots on the top right shows visually a straight line, meaning that **the normality assumption is met**.

Earlier in the analysis, we agreed that there is no multicollinearity problem, meaning the independent variables are not correlated with other.

#### 4. Multiple Linear Regression model

The assumptions are warranted for us to run a multiple regression model.

```
##
## Call:
## lm(formula = AverageMonthlyOrders ~ NumberOfEmployees + InventoryValue +
##     factor(Region), data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -53.526  -9.338  -0.352   9.929  44.400
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    2.077e+02  2.020e+01  10.281 < 2e-16 ***
## NumberOfEmployees  5.180e+00  5.403e-01   9.588 < 2e-16 ***
## InventoryValue    6.076e-03  2.555e-02   0.238  0.81239
## factor(Region)2    6.194e+00  5.022e+00   1.233  0.21944
## factor(Region)3    1.083e+01  6.794e+00   1.595  0.11300
## factor(Region)4    2.701e+01  8.349e+00   3.235  0.00151 **
## factor(Region)5    3.240e+01  1.112e+01   2.914  0.00415 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.85 on 143 degrees of freedom
## Multiple R-squared:  0.914, Adjusted R-squared:  0.9104
## F-statistic: 253.2 on 6 and 143 DF,  p-value: < 2.2e-16
```

Table 3: Regression statistics

The slope of the regression is -0.14, meaning that for every one unit change in UtilityCosts, MaintenanceCosts changes averagely -0.14 units. Both the intercept and the slope are statistically significant with  $\alpha = 0.05$ , due to p-value is close to 0. R squared is 0.08, meaning that the variation in UtilityCosts only explains 8% of variations in MaintenanceCosts.