ECON 7710 Homework 3: More Regressions Restaurant Lunch Sales

Gage Thompson, Robert Merrill, Nathan Moore, Gianluca Jones, Will Noll

General instructions

The `.Rmd` source for this document will be the template for your homework submission. You must submit your completed assignment as a single html document, saved to PDF and uploaded to eLC by **11:59p on February 6** using the filename `sectiontime_teamnumber_hw3.pdf` (e.g. `935_1_hw3.pdf`).

Notes:

- Include the name of each teammate under `author` in the `yaml` .
- For questions requiring analytical solutions, you can type them in using markdown math code. Or, you can submit handwritten solutions, embedding them in the knitted document as clearly readable images.
- For questions requiring computation, some or all of the required code is included in associated chunks. Modify chunks where and how you are directed.
- For (almost) all questions about R Markdown, consult The Definitive Guide (https://bookdown.org/yihui/rmarkdown/).
- The `setup` chunk above indicates the packages required for this assignment.
- You will find a description of the variables in the referenced dataset through the Help tab in the Plot pane of RStudio.
- **Switch `eval` to `TRUE` in the global options command to execute code chunks**.

Restaurant Lunch Sales

You have been hired to advise a mom-and-pop restaurant in southern Brazil about its customer traffic. The restaurant has given you two years of data on the number of lunch sales per day. The data also include some information about weather, Brazilian holidays, the local tourist season, and common payday weeks. The full list of variables is included at the end of this assignment description. The owners have asked you to generate insights about the relationship between some of these variables and lunch sales.

The sample period spans two pre-Covid years (February 2018 to March 2020). The owners understand that customer traffic may not follow the same patterns going forward, but these data are the best available, so you have been asked to analyze them anyway.

To answer the owners' questions (below), estimate a model of lunch sales regressed on the following variables: precipitation, temperature, humidity, whether it is high tourist season, whether it is a common payday week, the number of open competitors, and day of the week.

Hint: Pay attention to which variables are categoricals. Remember to use heteroskedasticity- robust standard errors for all regressions.

1. Use your regression results to analyze the relationship between competition on lunch sales.

a. Estimate a model of lunch sales regressed on the following variables: precipitation, temperature, humidity, whether it is high tourist season, whether it is a common payday week, the number of open competitors, and day of the week. Do this also for log lunch sales. Report your estimates for the relationship with competition in a nice table.

b. Looking at the regression with sales as the dependent variable, what is the relationship between the number of open competitors and lunch sales, all else equal? Does the sign of the coefficient estimate make sense?

c. Looking at the regression with log sales as the dependent variable, what is the relationship between having one more open competitor on lunch sales?

d. The owners are concerned that the estimated relationship above may be biased. What would be a plausible direction of the bias, and why?

```
# Make sure that the data csv is stored in the same directory as this file
lunch <- read.csv("lunchsales.csv")

# Some initial data cleaning
# Note the our omitted category here is Friday
lunch$log_s <- log(lunch$NumberOfLunchSales)
lunch$Monday <- as.integer(lunch$DayOfWeek =="Monday")
lunch$Tuesday <- as.integer(lunch$DayOfWeek =="Tuesday")
lunch$Wednesday <- as.integer(lunch$DayOfWeek =="Wednesday")
lunch$Thursday <- as.integer(lunch$DayOfWeek =="Thursday")
lunch$Sunday <- as.integer(lunch$DayOfWeek =="Sunday")
lunch$Saturday <- as.integer(lunch$DayOfWeek =="Saturday")
str(lunch)
```

```
## 'data.frame':    715 obs. of  19 variables:
##  $ Date                       : chr  "2018-02-14" "2018-02-15" "2018-02-16" "2018-02-17" ...
##  $ DayInData                  : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ DayOfWeek                  : chr  "Wednesday" "Thursday" "Friday" "Saturday" ...
##  $ Month                      : int  2 2 2 2 2 2 2 2 2 2 ...
##  $ NumberOfLunchSales         : int  107 123 104 138 149 95 84 80 103 113 ...
##  $ Weekend                    : int  0 0 0 1 1 0 0 0 0 0 ...
##  $ CommonPaydayWeek           : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ HighSeason                 : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ NumberOfOpenCompetitors    : int  7 7 7 7 7 7 7 7 7 7 ...
##  $ PrecipitationMm            : num  0 0 0 0 0 0 3.3 27 15 0 ...
##  $ LunchtimeTemperatureCelsius: num  28 28.6 30.6 31.5 33.4 ...
##  $ HumidityPct                : num  66.8 61.5 65.2 65.8 58.5 ...
##  $ log_s                      : num  4.67 4.81 4.64 4.93 5 ...
##  $ Monday                     : int  0 0 0 0 0 1 0 0 0 0 ...
##  $ Tuesday                    : int  0 0 0 0 0 0 1 0 0 0 ...
##  $ Wednesday                  : int  1 0 0 0 0 0 0 1 0 0 ...
##  $ Thursday                   : int  0 1 0 0 0 0 0 0 1 0 ...
##  $ Sunday                     : int  0 0 0 0 1 0 0 0 0 0 ...
##  $ Saturday                   : int  0 0 0 1 0 0 0 0 0 0 ...
```

```
# Run the two regressions
model <- lm(NumberOfLunchSales ~ PrecipitationMm + LunchtimeTemperatureCelsius + HumidityPct + H
ighSeason + CommonPaydayWeek + NumberOfOpenCompetitors + Monday + Tuesday + Wednesday + Thursday
+ Sunday + Saturday, data=lunch)
summary(model)
```

```
##
## Call:
## lm(formula = NumberOfLunchSales ~ PrecipitationMm + LunchtimeTemperatureCelsius +
##     HumidityPct + HighSeason + CommonPaydayWeek + NumberOfOpenCompetitors +
##     Monday + Tuesday + Wednesday + Thursday + Sunday + Saturday,
##     data = lunch)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -51.334 -13.696  -1.899   9.747 212.156
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 213.37076   17.42876  12.242  < 2e-16 ***
## PrecipitationMm               0.07921    0.11237   0.705  0.48110
## LunchtimeTemperatureCelsius  -1.03286    0.22211  -4.650 3.96e-06 ***
## HumidityPct                  -0.50686    0.11980  -4.231 2.63e-05 ***
## HighSeason                    1.70521    3.00607   0.567  0.57072
## CommonPaydayWeek              5.79934    2.21837   2.614  0.00913 **
## NumberOfOpenCompetitors      -4.89571    0.90496  -5.410 8.66e-08 ***
## Monday                        2.59822    3.82588   0.679  0.49729
## Tuesday                       1.95191    3.83472   0.509  0.61090
## Wednesday                     2.91182    3.79967   0.766  0.44373
## Thursday                     -2.28091    3.78830  -0.602  0.54731
## Sunday                        6.89324    3.83442   1.798  0.07265 .
## Saturday                      3.98211    3.85417   1.033  0.30187
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 27.22 on 702 degrees of freedom
## Multiple R-squared:  0.0987, Adjusted R-squared:  0.0833
## F-statistic: 6.406 on 12 and 702 DF,  p-value: 7.157e-11
```

```
log_model <- lm(log_s~ PrecipitationMm + LunchtimeTemperatureCelsius + HumidityPct + HighSeason
+ CommonPaydayWeek + NumberOfOpenCompetitors + Monday + Tuesday + Wednesday + Thursday + Sunday
+ Saturday,, data=lunch)
summary(log_model)
```

```
##
## Call:
## lm(formula = log_s ~ PrecipitationMm + LunchtimeTemperatureCelsius +
##     HumidityPct + HighSeason + CommonPaydayWeek + NumberOfOpenCompetitors +
##     Monday + Tuesday + Wednesday + Thursday + Sunday + Saturday,
##     data = lunch)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.59235 -0.12029  0.00283  0.10746  1.27304
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  5.6209386  0.1346475  41.746  < 2e-16 ***
## PrecipitationMm             -0.0003040  0.0008681  -0.350   0.7263
## LunchtimeTemperatureCelsius -0.0101077  0.0017159  -5.891 5.97e-09 ***
## HumidityPct                 -0.0047637  0.0009255  -5.147 3.44e-07 ***
## HighSeason                   0.0571822  0.0232237   2.462   0.0140 *
## CommonPaydayWeek             0.0338071  0.0171382   1.973   0.0489 *
## NumberOfOpenCompetitors     -0.0452732  0.0069914  -6.476 1.78e-10 ***
## Monday                       0.0327925  0.0295572   1.109   0.2676
## Tuesday                      0.0191011  0.0296255   0.645   0.5193
## Wednesday                    0.0333276  0.0293547   1.135   0.2566
## Thursday                    -0.0092687  0.0292669  -0.317   0.7516
## Sunday                       0.0630011  0.0296232   2.127   0.0338 *
## Saturday                     0.0327828  0.0297757   1.101   0.2713
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2103 on 702 degrees of freedom
## Multiple R-squared:  0.1546, Adjusted R-squared:  0.1401
## F-statistic: 10.69 on 12 and 702 DF,  p-value: < 2.2e-16
```

```
# Create a nice summary table
models = list(model, log_model)
modelsummary(models,
     coef_omit="^(?!NumberOfOpenCompetitors)",
     vcov="robust",
     statistic=c("std.error"), fmt=3,
     stars = c("*" = 0.05, "**" = 0.01, "***" = 0.001),
     gof_map = c("nobs", "r.squared", "mean"),
     title="Impact of Open Competitors on Lunch Sales")
```

Impact of Open Competitors on Lunch Sales

	(1)	**(2)**
NumberOfOpenCompetitors	-4.896***	-0.045***

* p < 0.05, ** p < 0.01, *** p < 0.001

	(0.926)	(0.007)
Num.Obs.	715	715
R2	0.099	0.155

* p < 0.05, ** p < 0.01, *** p < 0.001

Answers

b. The coefficient for the number of open competitors is -4.8957 with a very significant p-value (p < 0.001). This negative coefficient indicates that an increase in the number of open competitors is associated with a decrease in lunch sales. In other words, for each additional competitor that opens, the number of lunch sales is expected to decrease by approximately 4.8957 sales, all else being equal. Yes, the negative sign makes sense. More competitors in the area provide customers with more options. This would cause customers to disperse across the various establishments and leading to fewer sales per restaurant.

c. For each additional competitor that opens, the log of lunch sales decreases by 0.0437. It is also statistically significant with a p-value of p < 0.001. Furthermore, for each additional competitor, lunch sales are expected to decrease by about 4.27%. This aligns with the expectation that more competitors in the market reduce a single restaurant's share of total sales, leading to a decrease in sales volume.

d. The concern about potential bias in the relationship between the number of open competitors and lunch sales is valid. Omitted variable bias is the most likely in this case. If key variables that affect lunch sales are not included in the model, the coefficients of the included variables could be biased. For example, if factors like the quality of competitors, local economic conditions, or specific location advantages that influence lunch sales aren't included, this could lead to either an understatement or overstatement of the impact competitors have on sales.

You can use either model (i.e., lunch sales in level or log) for the remaining question.

2. Which day of the week has the highest average lunch sales, all else equal? Which day of the week has the lowest? Are sales on these days statistically different from sales on Fridays?

Sunday has the highest average lunch sales as a unit change in the day being Sunday is a associated with a 6.893 unit change in number of lunch sales holding all else constant. Thursday has the lowest average lunch as sales as a unit change in the day being Thursday is associated with a -2.281 unit change in number of lunch sales holding all else constant. Sales on Sunday and Thursday are not statistically significant as their p-values are both greater than the alpha of .05. Thursday lowest

Answer

3. Estimate an alternative model to test whether the relationship between competitors and lunch sales is different during the peak season. What is the relationship between additional competitors and lunch sales in the high season?

```
log_model <- lm(log_s~ NumberOfOpenCompetitors*HighSeason, data=lunch)
summary(log_model)
```

```
## 
## Call:
## lm(formula = log_s ~ NumberOfOpenCompetitors * HighSeason, data = lunch)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -0.59430 -0.12243 -0.00464  0.11491  1.23918 
## 
## Coefficients:
##                                      Estimate Std. Error t value Pr(>|t|)    
## (Intercept)                         4.9541800  0.1376679  35.986  < 2e-16 ***
## NumberOfOpenCompetitors            -0.0410765  0.0128479  -3.197  0.00145 ** 
## HighSeason                          0.0917266  0.1616767   0.567  0.57066    
## NumberOfOpenCompetitors:HighSeason  0.0005181  0.0154040   0.034  0.97318    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.2172 on 711 degrees of freedom
## Multiple R-squared:  0.08638,    Adjusted R-squared:  0.08253 
## F-statistic: 22.41 on 3 and 711 DF,  p-value: 7.151e-14
```

Answer

The relationship between additional competitors and lunch sales during high season represents a one unit change in competitors during high season is associated with a .0518% change in the number of lunch sales.

Data Description

These data contain the following variables: