Learning to Use Regression Analysis

ECO 6416

2022-09-13

Contents

Here are all the packages needed to get started.

```
library(readxl)
library(dplyr) # for pipe operator
library(gt) # for fancier tables
library(gtsummary) # for fancier summary statistics
library(corrplot) # for fancier correlations
library(car) # for easier scatterplots
library(jtools) # for fancier regression output

sessionInfo()
```

```
## R version 4.2.1 (2022-06-23 ucrt)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
## Running under: Windows 10 x64 (build 19044)
##
## Matrix products: default
##
## locale:
## [1] LC_COLLATE=English_United States.utf8
## [2] LC_CTYPE=English_United States.utf8
## [3] LC_MONETARY=English_United States.utf8
## [4] LC_NUMERIC=C
## [5] LC_TIME=English_United States.utf8
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
## [1] jtools_2.2.0    car_3.1-0       carData_3.0-5   corrplot_0.92
## [5] gtsummary_1.6.1 gt_0.7.0        dplyr_1.0.9     readxl_1.4.1
##
```

```
## loaded via a namespace (and not attached):
##  [1] Rcpp_1.0.9          cellranger_1.1.0    pillar_1.7.0
##  [4] compiler_4.2.1      tools_4.2.1         digest_0.6.29
##  [7] evaluate_0.15       lifecycle_1.0.1     tibble_3.1.7
## [10] gtable_0.3.0        pkgconfig_2.0.3     rlang_1.0.3
## [13] cli_3.3.0           DBI_1.1.3           rstudioapi_0.14
## [16] yaml_2.3.5          xfun_0.31           fastmap_1.1.0
## [19] stringr_1.4.0       knitr_1.39          generics_0.1.3
## [22] vctrs_0.4.1         grid_4.2.1          tidyselect_1.1.2
## [25] glue_1.6.2          R6_2.5.1            fansi_1.0.3
## [28] rmarkdown_2.14      pander_0.6.5        tidyr_1.2.0
## [31] purrr_0.3.4         ggplot2_3.3.6       magrittr_2.0.3
## [34] broom.helpers_1.8.0 scales_1.2.0        ellipsis_0.3.2
## [37] htmltools_0.5.2     abind_1.4-5         assertthat_0.2.1
## [40] colorspace_2.0-3    utf8_1.2.2          stringi_1.7.8
## [43] munsell_0.5.0       crayon_1.5.1
```

1 Business Problem

You've been hired to pick the best location for the next Woody's restaurant. If you can explain gross sales as a function of location, you can use this equation to help pick the best spot.

2 Six Steps to Regression Analysis

1. Review literature and develop a theoretical model
2. Specify the model
3. Hypothesize the expected signs of the coefficients
4. Collect the data. Inspect and clean the data
5. Estimate and evaluate the equation
6. Document the results

2.1 Review the Literature and Develop the Theoretical Model

- From talking with experts at the firm, you realize that Woody's restaurants are identical regardless of location. Lot size and location type (suburban, retail, residential) will not be influential in estimating a Woody's location but may be useful if you were looking at a different restaurant chain.
- Lastly, they convince you to define your dependent variable as the number of customers served measured by the number of checks given. They say this because consumption differences and price differences between locations is not as important.
- Y = Number of checks handed out in the previous year

2.2 Specify the model

- You produce 5 potential determinants of sales:
  - Number of people near location
  - Number of competitors near location
  - Income level of individuals near location
  - Number of cars passing the location daily
  - Number of months since opened
- You decide to ignore the last two for the following reasons:

  - It would be costly to collect the traffic data
  - All other locations have been open long enough to establish a stable clientele
- As a result, you decide that these are your 3 explanatory variables:
  - N = Number of direct competitors within a 2-mile radius
  - P = Number of people living within a 3-mile radius
  - I = Average household income of the population

So in total the regression equation looks like this:

$$Y_i = \beta_0 + \beta_N N_i + \beta_P P_i + \beta_I I_i + \varepsilon_i$$

2.3 Hypothesize the Expected Signs of the Coefficients

From this linear function, you can then hypothesize the signs of the coefficients:

- $\beta_N < 0$: You would expect the number of customers to decrease as there are more competitors within the area (holding population and income constant)
- $\beta_P > 0$: You would expect the number of customers to increase as the number of people near a restaurant increases
- β_I weakly positive?: We might expect as income increases, more people will go out to eat more, however, it could also be the case that at higher-income areas, people would like to eat at a higher end restaurant

2.4 Collect the data. Inspect and Clean the Data

- You are confident in the data for 3 reasons:
  - Each manager measured the variables identically
  - All restaurants were included
  - All information is from the same year

2.4.1 Bringing in Data

```
woodys <- read_excel("../Data/Woodys.xlsx")
```

2.4.2 Inspecting Data

We can inspect the data types to make sure they make sense.

```
str(woodys)
```

```
## tibble [33 x 4] (S3: tbl_df/tbl/data.frame)
##  $ Sales      : num [1:33] 107919 118866 98579 122015 152827 ...
##  $ NCompetitor: num [1:33] 3 5 7 2 3 5 8 2 6 2 ...
##  $ Population : num [1:33] 65044 101376 124989 55249 73775 ...
##  $ Income     : num [1:33] 13240 22554 16916 20967 19576 ...
```

Check out a few rows as well

```
head(woodys)
```

```
## # A tibble: 6 x 4
##    Sales NCompetitor Population Income
##    <dbl>       <dbl>      <dbl>  <dbl>
## 1 107919           3      65044  13240
## 2 118866           5     101376  22554
## 3  98579           7     124989  16916
## 4 122015           2      55249  20967
## 5 152827           3      73775  19576
## 6  91259           5      48484  15039
```

```
tail(woodys, 2) # just to demonstrate additional args
```

```
## # A tibble: 2 x 4
##    Sales NCompetitor Population Income
##    <dbl>       <dbl>      <dbl>  <dbl>
## 1 117146           3      60457  20307
## 2 163538           2      65065  20111
```

2.4.3 Summary Statistics

```
summary(woodys)
```

```
##      Sales         NCompetitor      Population         Income     
##  Min.   : 91259   Min.   :2.000   Min.   : 37852   Min.   :13240  
##  1st Qu.:105564   1st Qu.:3.000   1st Qu.: 57386   1st Qu.:16839  
##  Median :122015   Median :4.000   Median : 95120   Median :19200  
##  Mean   :125635   Mean   :4.394   Mean   :103887   Mean   :20553  
##  3rd Qu.:140791   3rd Qu.:6.000   3rd Qu.:139900   3rd Qu.:22554  
##  Max.   :166755   Max.   :9.000   Max.   :233844   Max.   :33242  
```

```
sd(woodys$Sales)
```

```
## [1] 22404.09
```

```
sd(woodys$NCompetitor)
```

```
## [1] 1.9193
```

```
sd(woodys$Population)
```

```
## [1] 55884.51
```

```
sd(woodys$Income)
```

```
## [1] 5141.865
```

Fancier way:

```
woodys %>%
tbl_summary(statistic = list(all_continuous() ~ c("{mean} ({sd})",
"{median} ({p25}, {p75})",
"{min}, {max}"),
all_categorical() ~ "{n} / {N} ({p}%)"),
type = all_continuous() ~ "continuous2"
)
```

Characteristic	**N = 33**
Sales	
Mean (SD)	125,635 (22,404)
Median (IQR)	122,015 (105,564, 140,791)
Range	91,259, 166,755
NCompetitor	
2	5 / 33 (15%)
3	10 / 33 (30%)
4	3 / 33 (9.1%)
5	5 / 33 (15%)

Characteristic	**N = 33**
6	5 / 33 (15%)
7	3 / 33 (9.1%)
8	1 / 33 (3.0%)
9	1 / 33 (3.0%)
Population	
Mean (SD)	103,887 (55,885)
Median (IQR)	95,120 (57,386, 139,900)
Range	37,852, 233,844
Income	
Mean (SD)	20,553 (5,142)
Median (IQR)	19,200 (16,839, 22,554)
Range	13,240, 33,242

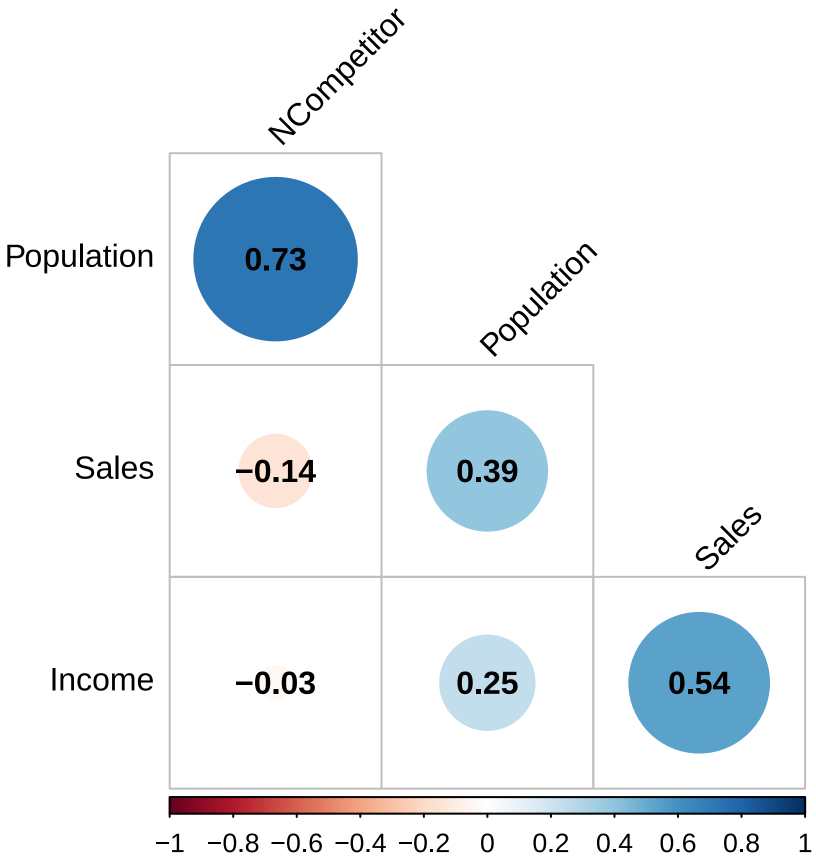
2.4.4 Correlation

```
cor(woodys)
```

```
##                  Sales NCompetitor Population      Income
## Sales        1.0000000 -0.14422464  0.3925677  0.53702201
## NCompetitor -0.1442246  1.00000000  0.7262507 -0.03153405
## Population   0.3925677  0.72625071  1.0000000  0.24519764
## Income       0.5370220 -0.03153405  0.2451976  1.00000000
```

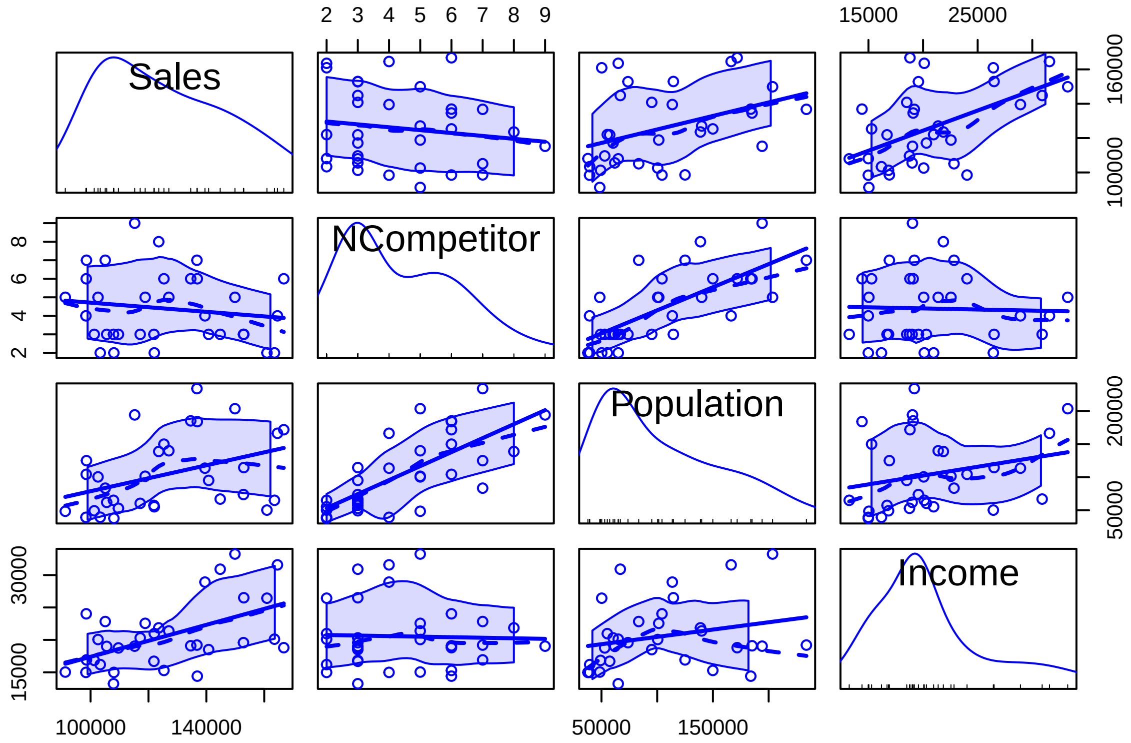
Fancier Way

```
corrplot(cor(woodys),
type = "lower",
order = "hclust",
tl.col = "black",
tl.srt = 45,
addCoef.col = "black",
diag = FALSE)
```



2.4.5 Visualize distribution and relationships of each variable

```
scatterplotMatrix(woodys)
```



2.5 Estimate and Evaluate the Equation

```
model <- lm(Sales ~ NCompetitor + Population + Income, data = woodys)

summary(model)
```

```
##
## Call:
## lm(formula = Sales ~ NCompetitor + Population + Income, data = woodys)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -21923  -8627  -2956   5328  33887
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.022e+05  1.280e+04   7.984 8.35e-09 ***
## NCompetitor -9.075e+03  2.053e+03  -4.421 0.000126 ***
## Population   3.547e-01  7.268e-02   4.880 3.54e-05 ***
## Income       1.288e+00  5.433e-01   2.371 0.024623 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14540 on 29 degrees of freedom
## Multiple R-squared:  0.6182, Adjusted R-squared:  0.5787
## F-statistic: 15.65 on 3 and 29 DF,  p-value: 3.058e-06
```

Fanicer output

```
summm(model)
```

```
## MODEL INFO:
## Observations: 33
## Dependent Variable: Sales
## Type: OLS linear regression
##
## MODEL FIT:
## F(3,29) = 15.65, p = 0.00
## R² = 0.62
## Adj. R² = 0.58
##
## Standard errors: OLS
## --------------------------------------------------------
##                          Est.       S.E.   t val.      p
## ----------------- ----------- ---------- -------- ------
## (Intercept)         102192.43   12799.83     7.98   0.00
## NCompetitor          -9074.67    2052.67    -4.42   0.00
## Population               0.35       0.07     4.88   0.00
## Income                   1.29       0.54     2.37   0.02
## --------------------------------------------------------
```

2.5.1 Prediction

Suppose we have a potential location with the following characterisitics:

- NCompetitor = 4
- Population = 90,000
- Income = 20,000

```
newPrediction <- data.frame(NCompetitor = 4,
                            Population = 90000,
                            Income = 20000)

predict(model, newdata = newPrediction)
```

```
##        1
## 123572.4
```

2.5.2 Marginal Analysis

```
model$coefficients["Population"]
```

```
## Population
##  0.3546684
```

This says with one additional person, you'll see an increase in 0.35 checks per year.

2.6 Document Results

When you knit this document, it allows you to automatically put this all together.