# Marketing Data Science

## Modeling Techniques in Predictive Analytics with R and Python

THOMAS W. MILLER

Publisher: Paul Boger
Editor-in-Chief: Amy Neidlinger
Executive Editor: Jeanne Glasser Levine
Operations Specialist: Jodi Kemper
Cover Designer: Alan Clements
Managing Editor: Kristy Hart
Manufacturing Buyer: Dan Uhrig

Published by Pearson Education, Inc.
Old Tappan New Jersey 07675

Printed in the United States of America

First Printing May 2015

ISBN-10: 0-13-388655-7
ISBN-13: 978-0-13-388655-9

Pearson Education LTD.
Pearson Education Australia PTY, Limited.
Pearson Education Singapore, Pte. Ltd.
Pearson Education Asia, Ltd.
Pearson Education Canada, Ltd.
Pearson Educación de Mexico, S.A. de C.V.
Pearson Education—Japan
Pearson Education Malaysia, Pte. Ltd.
Library of Congress Control Number: 2015937911

# Contents

# 13 Predicting Sales

"Gentlemen, you can't fight in here! This is the War Room!"

—PETER SELLERS AS PRESIDENT MERKIN MUFFLEY
IN *Dr. Strangelove or: How I Learned
to Stop Worrying and Love the Bomb* (1964)

We are nearing the end of our journey through marketing data science. Soon it will be time for the beach.

I do not fit into the muscle-beach scene at Venice or Malibu, and I find Santa Monica too pricey. Redondo would be fine, Manhattan as well. But between those two is my favorite beach in the world, Hermosa Beach.

I first visited Hermosa Beach in my late twenties. It was during a month-long winter-term break from teaching. I rented an apartment above a bar on Pier Avenue and listened to rock music late into every night. I played volleyball with the locals and learned how to jump on the sand. I bought a hot bicycle for $25 and rode along the strand ten to fifteen miles a day.

When I returned to Hermosa many years later, I was surprised to see how little it had changed. The ocean has a way of promoting continuity. I am thankful for that.

And location? There is much to be said about the value of location.

Making predictions about sales is the job of marketing data science. For some problems we are asked to predict future sales for a company or make sales forecasts for product lines—this is time series sales forecasting, which is covered in appendix A (page 286). Here we focus on predicting sales in retail site selection.

Site selection problems involve predicting sales by location. A retailer with fifty stores in one geographic area wants to open stores in a new area. The retailer wants the new stores to be profitable, with sales revenue much higher than costs. Consumer demographics and business data can be used to guide the selection of new sites, and it is common to organize data by location and perform a cross-sectional analysis.

Site selection problems usually involve data sets with many more explanatory variables than there are stores. The challenge is to find the right combination of explanatory variables to predict sales at existing stores and then to use that combination of variables to get accurate sales forecasts for new stores.

Each store or potential store site, geocoded by longitude and latitude, represents a point on a map and may be associated with thousands of variables about population, housing, and economic conditions. We use geographical information systems to estimate population within a certain distance of the store. We get drive-time-based measures such as median income for households within five minutes of the store. We describe store sites by size and layout, location relative to nearby highways, signage, and parking. We collect data about the business environment, nearby retailers, and potential competitors. And when consumer data are available, we compute trade-area-based measures such as the percentage of families who shop at the store among the ten thousand families residing closest to the store. All of these can be used as explanatory variables in models for predicting store sales.

In site selection, we often employ cross-sectional models, ignoring the spatial data aspects of the problem. We use a store's census identifier or zip code to link sales and census data for each store or site. An example of the cross-sectional approach is shown for the thirty-three sites known as Studenmund's Restaurants, as discussed in appendix C (page 373).

*Table 13.1.* *Fitted Regression Model for Restaurant Sales*

| | Response: Sales |
|---|---|
| Competition | -9.075e+034$^{***}$ |
| Population | 3.547e-01$^{***}$ |
| Income | 1.288e+00$^{***}$ |
| Constant | 1.022e+05$^{***}$ |
| Observations | 33 |
| R$^2$ | 0.6182 |
| Adjusted R$^2$ | 0.5787 |
| Residual Std. Error | 14540$(df = 29)$ |
| F statistic | 15.65$^{***}(df = 3; 29)$ |
| *Notes:* | $^{***}$Significant at the 0.001 level. |

*Table 13.2.* *Predicting Sales for New Restaurant Sites*

| *competition* | *population* | *income* | *predicted sales* |
|---|---|---|---|
| 2 | 50,000 | 25,000 | 133,975 |
| 3 | 200,000 | 22,000 | 174,236 |
| 5 | 220,000 | 19,000 | 159,317 |

We begin as we often do, with exploratory data analysis, plotting univariate histograms, densities, and empirical cumulative distribution functions. Then we move to looking at pairwise relationships, as shown in figures 13.1 and 13.2.

The linear regression of sales on competition, population, and income is summarized in table 13.1. Variance inflation factors computed for this problem show that multicollinearity is not an issue. Diagnostic plots in figure 13.3 indicate no special issues regarding the fitted model.

Applying the fitted model to prospective restaurant sites involves finding the associated explanatory variable values for each site and computing the predicted sales. Table 13.2 shows the results for three prospective store sites. The middle site has the highest predicted sales, so this would be the recommended site for the next new restaurant.

***Figure 13.1.*** *Scatter Plot Matrix for Restaurant Sales and Explanatory Variables*

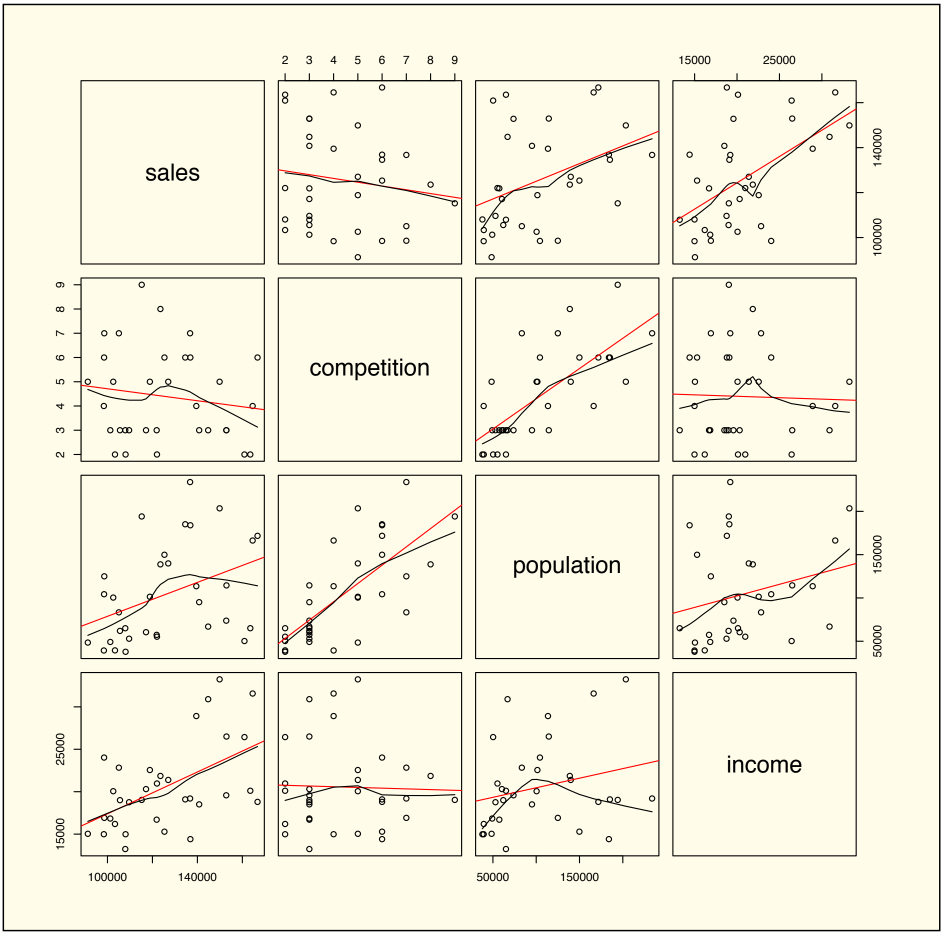


*Figure 13.2.* *Correlation Heat Map for Restaurant Sales and Explanatory Variables*

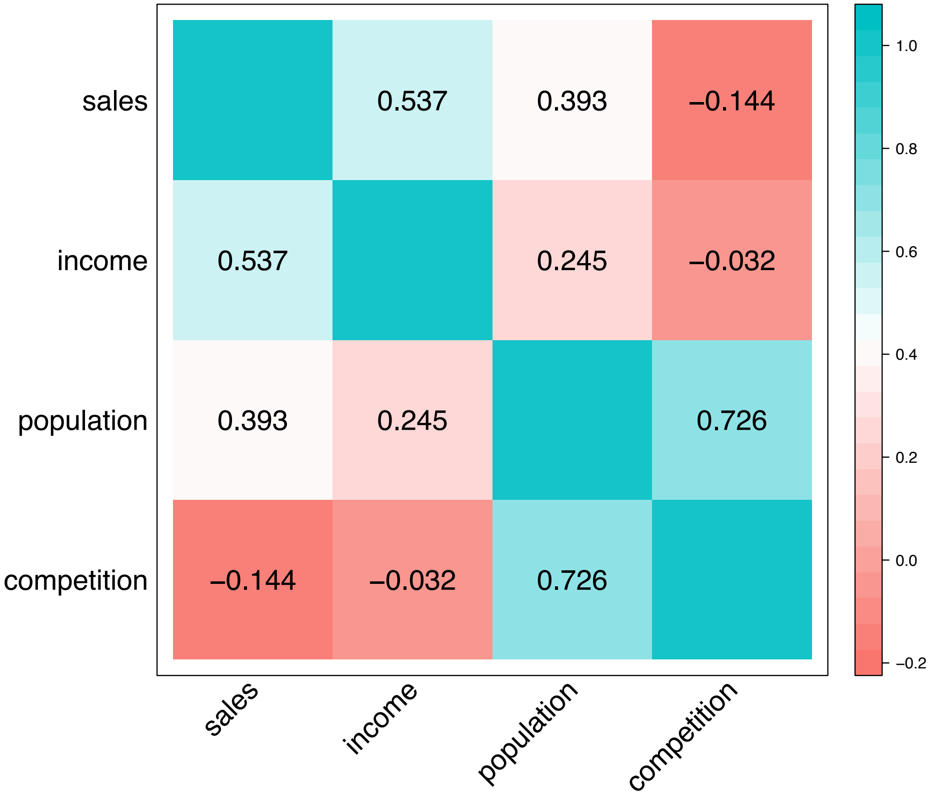
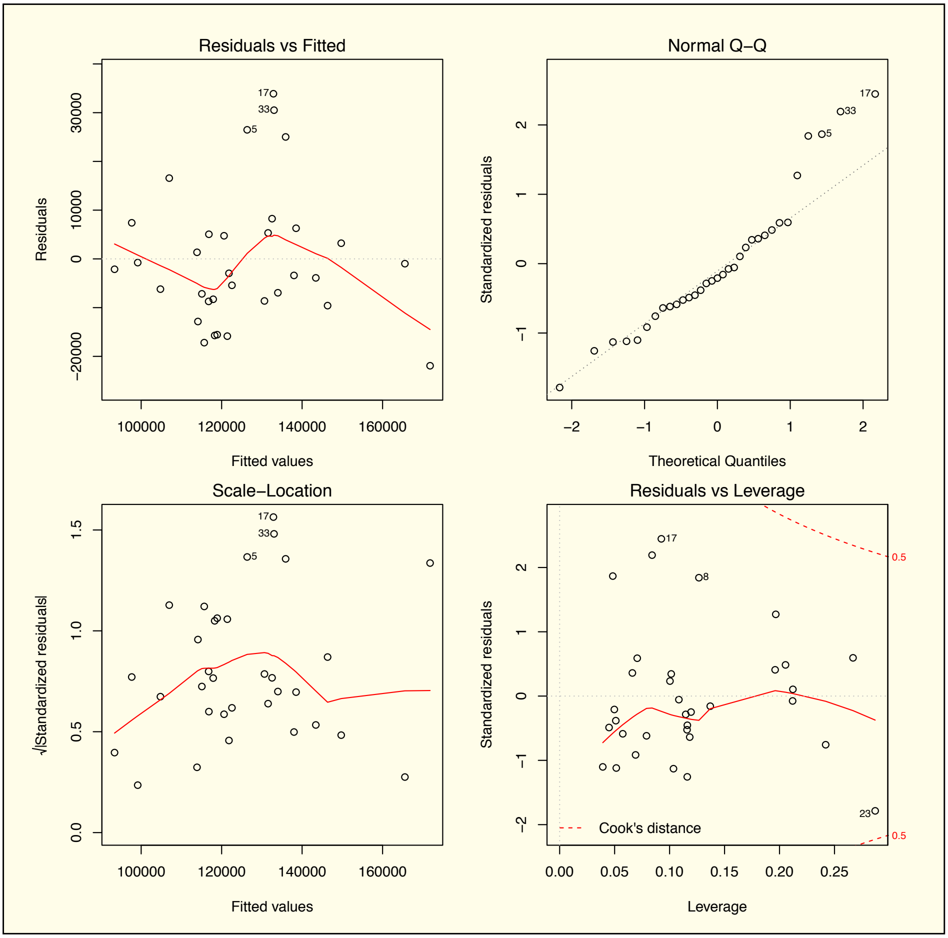


*Figure 13.3. Diagnostics from Fitted Regression Model*



Studenmund's Restaurants is a small site selection problem that shows what is possible with ordinary least squares regression for predicting sales response. Moving to larger problems, we can take what is possible with one small subset of explanatory variables and extend it to many small subsets of explanatory variables—we can develop regression ensembles.

Site selection problems often involve thousands of potential explanatory variables and small numbers of current sites. While there are formal statistical methods for dealing with high-dimensional data (Bühlmann and van de Geer 2011), these are of little use in dealing with problems on the scale we encounter with site selection. What are needed instead are trustworthy model-building heuristics, techniques for partitioning the explanatory variable space into smaller subspaces, and then working with each subspace in turn. Let us review this divide-and-conquer approach.

Partitioning of explanatory variables into meaningful groups can be a first step in dealing with large numbers of explanatory variables. Taken together, these variable groups cover the full range of the explanatory variable space. Some variables relate to demographics, others to business characteristics. Some measures are set at a one-mile radius from the site, others at two-miles, five-miles, and so on. Working with road network overlays in a geographic information system, we can also derive drive-time measures, which may be partitioned into sets at one minute, five minutes, and ten minutes from the site, for example.

Next, working with each group of explanatory variables, we can use techniques such as tree-structured regression and random forests to provide lists of the most important explanatory variables—the best predictors of sales response. With a set of thirty or so best explanatory variables, we go on to employ all possible regressions, identifying the very best subset model within each group of explanatory variables.

Repeating the subset selection and best-possible-regression search for each group of explanatory variables, we arrive at an ensemble of predictive models. The ensemble predictor can be a simple average of the predictors or a composite with weights proportional to the anticipated out-of-sample predictive accuracy of the individual models. We can use linear regression or any number of machine learning methods in hybrid component models.

Site selection problems challenge us to practice the art as well as the science of modeling. The real power of the heuristic approach we call divide-and-conquer flows from the fact that the final predictive model is an ensemble, covering the full range of the explanatory variable space. Furthermore, the entire process described here can be embedded within an internal cross-validation scheme to protect against over-fitting.

The edited volumes by Davies and Rogers (1984) and Wrigley (1988) review the historical development of cross-sectional methods of inquiry for site selection. Peterson (2004) reviews commercial supplier work in this area. Alternative approaches include nearest-neighbor methods and spatial data models based on point processes, grid, or lattice structures (Cressie 1993; Bivand, Pebesma, and Gómez-Rubio 2008).

Gravity models for retail site selection posit that shoppers are influenced by the size and location of competing stores. Shoppers are thought to be attracted to larger stores due to greater variety or lower prices. Shoppers are thought to be attracted to nearby stores because travel times to nearby stores are shorter than travel times to distant stores. Tayman and Pol (1995) and Lilien and Rangaswamy (2003) describe gravity models for retail site selection.

We can perform location-based site selection and time series sales forecasting simultaneously with longitudinal or panel data models. Frees and Miller (2004) illustrate the process using data from the Wisconsin Lottery Sales case in appendix C (page 389).

Site selection research and location-based marketing are supported by geodemographic data aggregators and providers of geographic information systems. Alteryx provides commercial solutions integrated with the R software environment.

Exhibit 13.1 shows an R program that analyzes data from Studenmund's Restaurants, drawing on regression tools from Fox (2014). The corresponding Python program is in exhibit 13.2.

***Exhibit 13.1.*** *Restaurant Site Selection (R)*

```
# Restaurant Site Selection (R)

# brind packages into workspace
library(car)  # regression tools
library(lattice)  # needed for correlation _heat_map function

load("correlation_heat_map.RData")  # from R utility programs

# read data for Studenmund’s Restaurants
# creating data frame restdata
restdata <- read.csv("studenmunds_restaurants.csv", header = TRUE)

# examine the data frame
print(str(restdata))
print(restdata)

# compute summary statistics
print(summary(restdata))

# exploratory data analysis... graphics for discovery
# cumulative distribution function of the sales response
pdf(file = "fig_selecting_sites_hist.pdf",
    width = 8.5, height = 8.5)
with(restdata, hist(sales/1000,
    xlab="Sales (thousands)",
    ylab="Frequency",
    main = "", las = 1))
dev.off()

pdf(file = "fig_selecting_sites_cdf.pdf",
    width = 8.5, height = 8.5)
with(restdata, plot(sort(sales/1000),
    (1:length(sales))/length(sales),
    type="s", ylim=c(0,1), las = 1,
    xlab="Restaurants Ordered by Sales (thousands)",
    ylab="Proportion of Restaurants with Lower Sales"))
dev.off()

# scatter plot matrix with simple linear regression
# models and lowess smooth fits for variable pairs
pdf(file = "fig_selecting_sites_scatter_plot_matrix.pdf",
    width = 8.5, height = 8.5)
pairs(restdata,
    panel = function(x, y) {
        points(x, y)
        abline(lm(y ˜ x), lty = "solid", col = "red")
        lines(lowess(x, y))
        }
    )
dev.off()
```

```
# correlation heat map
pdf(file = "fig_selecting_sites_correlation_heat_map.pdf",
    width = 8.5, height = 8.5)
restdata_cormat <-
    cor(restdata[,c("sales","competition","population","income")])
correlation_heat_map(cormat = restdata_cormat)
dev.off()

# specify regression model
restdata_model <- {sales ˜ competition + population + income}

# fit linear regression model
restdata_fit <- lm(restdata_model, data = restdata)

# report fitted linear model
print(summary(restdata_fit))

# examine multicollinearity across explanatory variables
# ensure that all values are low (say, less than 4)
print(vif(restdata_fit))

# default residuals plots. . . diagnostic graphics
pdf(file = "fig_selecting_sites_residuals.pdf",
    width = 8.5, height = 8.5)
par(mfrow=c(2,2),mar=c(4, 4, 2, 2))
plot(restdata_fit)
dev.off()

# define data frame of sites for new restaurants
sites <- data.frame(sales = c(NA, NA, NA),
    competition = c(2, 3, 5),
    population = c(50000, 200000, 220000),
    income = c(25000, 22000, 19000))

# obtain predicted sales for the new restaurants
# rounding to the nearest dollar
sites$predicted_sales <- round(predict(restdata_fit, newdata = sites),0)
print(sites)

# Suggestions for the student: Employ alternative methods of regression
# to predict sales response. Compare results with those obtained from
# ordinary least squares regression. Examine the out-of-sample predictive
# power of models within a cross-validation framework.
# Having predicted sales for a cross-sectional/site selection problem,
# try a time series forecasting problem, working with one of the
# cases provided for this purpose: Lydia Pinkham’s Medicine Company or
# Wisconsin Lottery Sales.
```

***Exhibit 13.2.*** *Restaurant Site Selection (Python)*

```
# Restaurant Site Selection (Python)

# prepare for Python version 3x features and functions
from __future__ import division, print_function

# import packages for analysis and modeling
import pandas as pd  # data frame operations
import numpy as np  # arrays and math functions
import statsmodels.api as sm  # statistical models (including regression)
import statsmodels.formula.api as smf  # statistical models (including regression)

# read data for Studenmund’s Restaurants
# creating data frame restdata
restdata = pd.read_csv(’studenmunds_restaurants.csv’)

# print the first five rows of the data frame
print(pd.DataFrame.head(restdata))

# specify regression model
my_model = str(’sales ~ competition + population + income’)

# fit the model to the data
my_model_fit = smf.ols(my_model, data = restdata).fit()
# summary of model fit to the training set
print(my_model_fit.summary())
# predictions from the model fit to the data for current stores
restdata[’predict_sales’] = my_model_fit.fittedvalues

# compute the proportion of response variance accounted for
print(’\nProportion of Test Set Variance Accounted for: ’,\
    round(np.power(restdata[’sales’].corr(restdata[’predict_sales’]),2),3))

# define DataFrame of sites for new restaurants
sites_data = {’sales’: [0,0,0],
             ’competition’: [2, 3, 5],
             ’population’: [50000, 200000, 220000],
             ’income’: [25000, 22000, 19000]}

sites = pd.DataFrame(sites_data)

# obtain predicted sales for the new restaurants
# rounding to the nearest dollar
sites[’sales_pred’] = my_model_fit.predict(sites)
print(’\nNew sites with predicted sales’, sites, ’\n’)
```