Homework 1

Andrew Lei
Arizona State University

September 14, 2016

Part 1

1. `FRESH`: annual spending (m.u.) on fresh products (Continuous and Ratio);
2. `MILK`: annual spending (m.u.) on milk products (Continuous and Ratio);
3. `GROCERY`: annual spending (m.u.) on grocery products (Continuous and Ratio);
4. `FROZEN`: annual spending (m.u.) on frozen products (Continuous and Ratio)
5. `DETERGENTS_PAPER`: annual spending (m.u.) on detergents and paper products (Continuous and Ratio)
6. `DELICATESSEN`: annual spending (m.u.) on and delicatessen products (Continuous and Ratio);
7. `CHANNEL`: customers' Channel - Horeca (Hotel/Restaurant/Café) or Retail channel (Discrete and Nominal)
8. `REGION`: customers' Region - Lisbon, Oporto or Other (Discrete and Nominal)

Part 2

It isn't entirely clear whether total or average spending per region is desired. I assumed average, but for the sake of completeness, total spending is included as well. Most spending is from 'Other' as most data points fell into that region. Likewise, Oporto had the fewest, data points, and the least total spending (and, for the same reason, the highest standard errors).

For the average spending by region, the standard errors are too high for there to be a statistically significant difference in spending on most of the products. The only ones for which one exists is for fresh products (between Oporto and Other) and on delicatessen products (also between Oporto and Other). While not statistically significant, it seems that while Oporto spends less on fresh products, it instead purchases more frozen products than the other two regions.

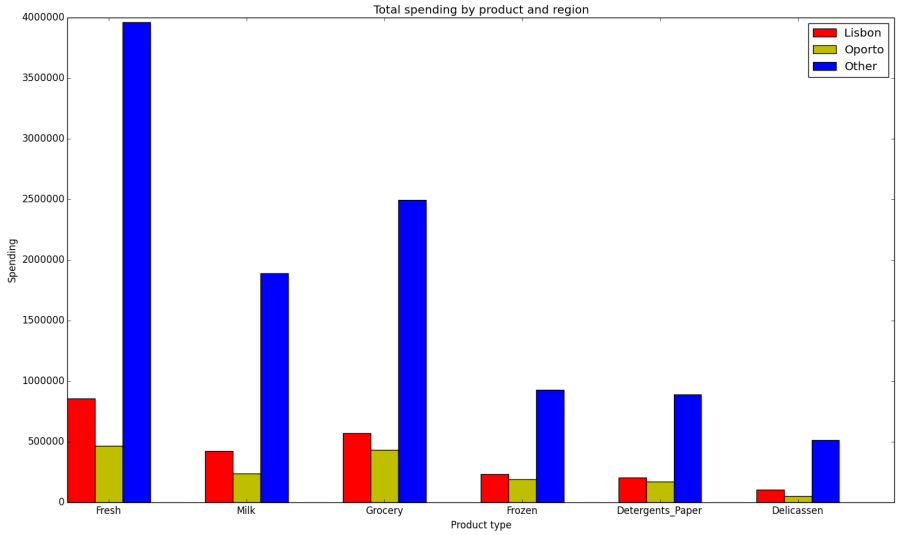

Figure 1: Total spending by region

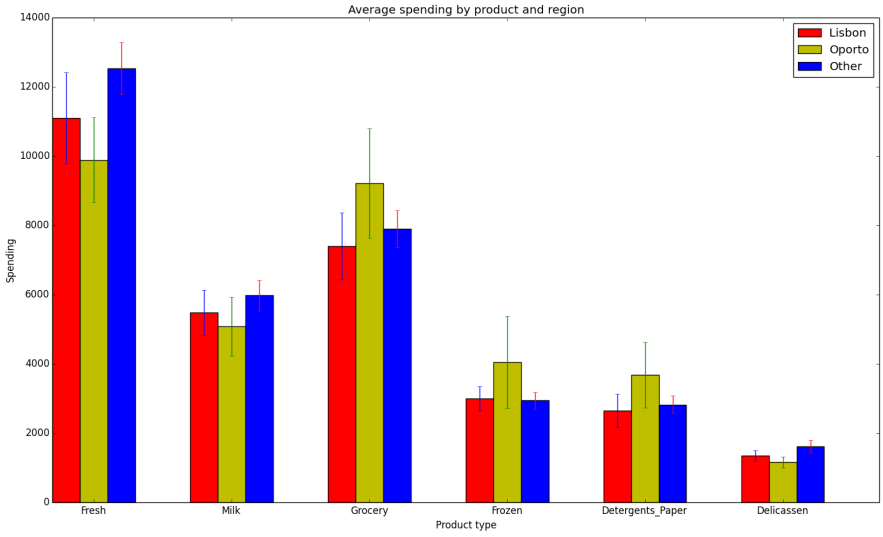

Figure 2: Average spending by region

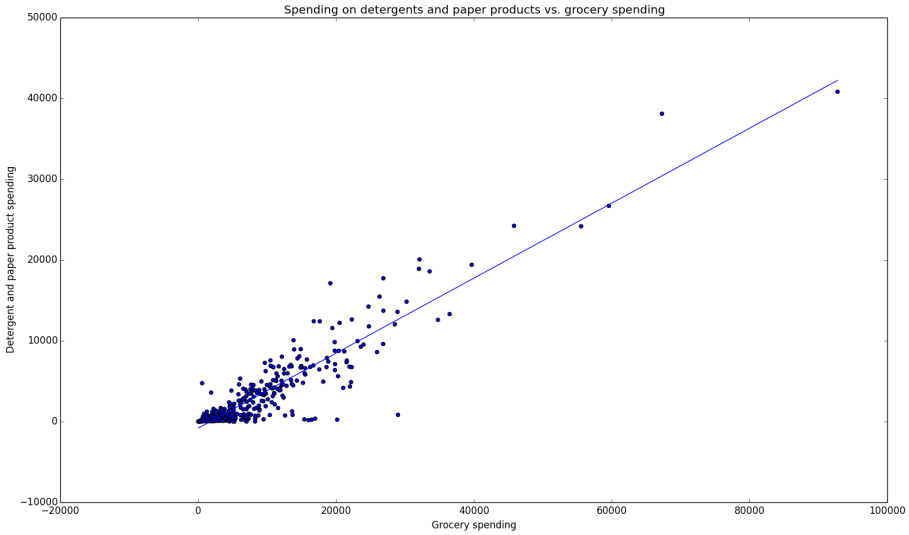

Figure 3: Relationship between spending on groceries and spending on detergents and paper, with data and best-fit line

Part 3

There is a strong correlation between spending on groceries and spending on detergents and paper. Here is the summary:

OLS Regression Results

Dep. Variable:	Detergents_Paper	R-squared:	0.855
Model:	OLS	Adj. R-squared:	0.855
Method:	Least Squares	F-statistic:	2582.
Date:	Wed, 14 Sep 2016	Prob (F-statistic):	9.56e-186
Time:	17:29:39	Log-Likelihood:	−3925.7
No. Observations:	440	AIC:	7855.
Df Residuals:	438	BIC:	7864.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[95.0% Conf.	Int.]
Intercept	−807.1336	113.050	−7.140	0.000	−1029.322	−584.945
Grocery	0.4639	0.009	50.812	0.000	0.446	0.482

Omnibus:	97.684	Durbin-Watson:	1.932
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1291.839
Skew:	−0.511	Prob(JB):	3.02e-281
Kurtosis:	11.332	Cond. No.	1.62e+04

Part 4

Below is the code I used to generate the images in parts 2 and 3. See `https://github.com/andrew-lei/cse494/blob/master/HW1/hw1.py` for the file.

```
#! /usr/bin/env python
```

```
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
import statsmodels.formula.api as sm

# Read the data
data = pd.read_csv('Wholesale customers data.csv')
totals = []
means = []
stderrs = []

# Compute average spending by category
for i in range(1, 4):
    # Get data from region i (1, 2, or 3) for spending on Fresh, Milk,↩
        &c.
    regdata = data[data.Region == i][['Fresh', 'Milk', 'Grocery', '↩
        Frozen', 'Detergents_Paper', 'Delicassen']]
    totals += [regdata.sum().as_matrix()]
    means += [regdata.mean().as_matrix()]
    # Compute standard error of mean
    stderrs += [ regdata.std().as_matrix() / np.sqrt(len(regdata)) ]


ind = np.arange(6)  # The x locations for the groups; 0, 1, 2, ..., 5
width = 0.2       # The width of the bars
fig, ax = plt.subplots()
fig.set_size_inches(18.5, 10.5) # Default size is too small

# Bars for bar graph - using total
rects = [ax.bar(ind, totals[0], width, color='r'),
    ax.bar(ind + width, totals[1], width, color='y'),
    ax.bar(ind + 2*width, totals[2], width, color='b')]

# Labels, titles, ticks, &c.
ax.set_xlabel('Product type')
ax.set_ylabel('Spending')
ax.set_title('Total spending by product and region')
ax.set_xticks(ind + 1.5 * width)
ax.set_xticklabels(['Fresh', 'Milk', 'Grocery', 'Frozen', '↩
    Detergents_Paper', 'Delicassen'])
ax.legend((rects[0][0], rects[1][0], rects[2][0]), ('Lisbon', 'Oporto'↩
    , 'Other'))

# Save and clear for next image
plt.savefig('part2a.png')
plt.cla()

# Bars for bar graph - using mean
rects = [ax.bar(ind, means[0], width, color='r', yerr=stderrs[0]),
    ax.bar(ind + width, means[1], width, color='y', yerr=stderrs[1]),
    ax.bar(ind + 2*width, means[2], width, color='b', yerr=stderrs[2])↩
        ]

# Labels, titles, ticks, &c.
ax.set_xlabel('Product type')
ax.set_ylabel('Spending')
ax.set_title('Average spending by product and region')
ax.set_xticks(ind + 1.5 * width)
ax.set_xticklabels(['Fresh', 'Milk', 'Grocery', 'Frozen', '↩
    Detergents_Paper', 'Delicassen'])
ax.legend((rects[0][0], rects[1][0], rects[2][0]), ('Lisbon', 'Oporto'↩
    , 'Other'))

# Save and clear for part 3
plt.savefig('part2b.png')
plt.cla()

# Ordinary least squares fit for Grocery (independent) and ↩
```

```
    Detergents_Paper(dependent)
fit = sm.ols(formula="Detergents_Paper ~ Grocery", data=data).fit()
ind = np.arange(min(data['Grocery']), max(data['Grocery']) )

print fit.summary()
print fit.pvalues # Print separately because very close to zero

# Plot points and best-fit
ax.scatter(data['Grocery'], data['Detergents_Paper'])
ax.plot(ind, fit.params[0] + ind*fit.params[1])

# Labels and title
ax.set_xlabel('Grocery spending')
ax.set_ylabel('Detergent and paper product spending')
ax.set_title('Spending on detergents and paper products vs. grocery ↩
    spending')

# Save
plt.savefig('part3.png')
```