Morley J Weston | Homework 2: Regression

Question 1:
This question requires us to look at an example of regression with the following characteristics:

- The dependent variable examined is the demand for parking within within 100 meters of shopping center during peak hours.
- The independent variables and their estimated effects are as follows:
  - Public bus lines within 300 meters of the shopping center. Each additional bus line in the area decreases demand for 8.796 parking spaces.
  - Area of the shopping center in square meters. Each additional square meter of shopping space implied increased demand of 0.0021 parking spaces.
  - Number of parking spaces available near the shopping center. Each additional parking space available creates demand for another 0.122 spaces.
  - Floor space of adjacent actively-used buildings. Each square meter of adjacent development adds demand for just 0.000477 parking spaces.

1.1: 此校估結果與建築技術規則規定間是否有顯著差異？什麼原因？
If the building regulations stipulate 1 parking space for every 200 square meters of floor area, that would imply that every square meter would require 0.005 parking spaces, more than double that of the actual estimation. However, this doesn't take the other factors into account, such as proximity to public transport or adjacent buildings. Overestimating the need for parking makes sense on one level; one wouldn't want a parking lot to be 100 percent full.

1.2: 依據問題 1 的結論，你會建議修改建築技術規則關於百貨公司的前述規定嗎？(如果不建議修改，是什麼原因？如果建議修改，應如何修改？)
As parking creates some level of induced demand, I wouldn't increase the amount of parking required for a commercial space. Parking has little benefit; it produced little taxable income and decreases the density of a city.

The correlation between public transport and parking space is a useful insight; perhaps some of the parking requirement could be waived if a development is close to multiple bus lines. Therefore, I would keep the parking requirement as-is, but decrease it for every bus line within a few hundred meters. This could work politically, as well, as real estate developers would be given a stake in promoting public transportation.

This change would require further study, however. Looking at this data, I wonder if the underlying reason may not actually be proximity to bus lines, but actually proximity to a city center, which would correlate with the number of bus lines in a given place. Floor space of adjacent buildings would likewise correlate with proximity to a city center, but maybe there are other reasons that this was less of a factor.

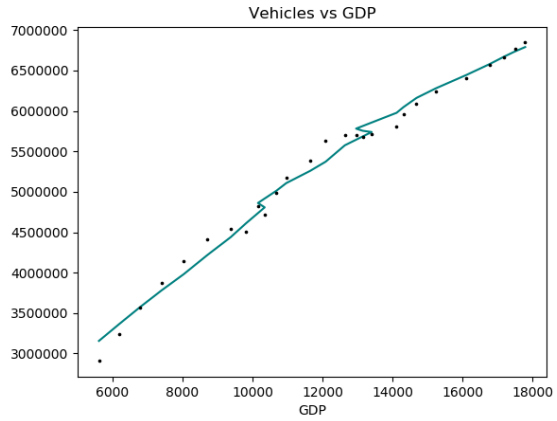
Question 2:

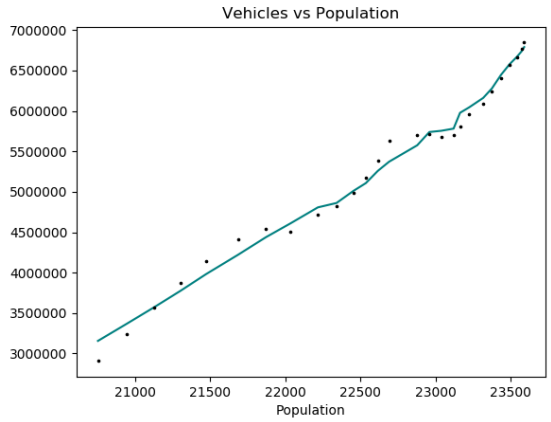
2.1: 以 OLS 法估計 Vehicle = β0 + β1 Population + β2 GDP 之迴歸模型；討論 R2、t、F、估計係數值 (β0、β1、β2)。

All calculation was done using a Python script, which can be found here:
`https://github.com/fgorkhs/miscscripts/blob/master/OLS_vehicles.py`

Despite my best efforts, I was unable to calculate the multiple regression without using outside libraries. The recurring problem is that I couldn't get the y-intercept to be the same for both cases, as indicated in the formula presented in class. I therefore used the Python library Statsmodel to complete this assignment. It provided the following output: OLS

```
Regression Results
==============================================================================
Dep. Variable:                Vehicle   R-squared:                       0.989
Model:                            OLS   Adj. R-squared:                  0.988
Method:                 Least Squares   F-statistic:                     1051.
Date:                Sun, 17 Nov 2019   Prob (F-statistic):           4.29e-24
Time:                        17:06:42   Log-Likelihood:                -352.72
No. Observations:                  27   AIC:                             711.4
Df Residuals:                      24   BIC:                             715.3
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const      -1.136e+07   2.44e+06     -4.651      0.000   -1.64e+07   -6.32e+06
GDP          144.4769     30.381      4.756      0.000      81.774     207.180
Population   660.6134    124.078      5.324      0.000     404.529     916.697
==============================================================================
Omnibus:                        0.469   Durbin-Watson:                   0.551
Prob(Omnibus):                  0.791   Jarque-Bera (JB):                0.297
Skew:                           0.244   Prob(JB):                        0.862
Kurtosis:                       2.839   Cond. No.                     2.69e+06
==============================================================================
```





R^2 in this scenario was 0.989, which means that the model explains 98.9 percent of the variance in the dependent variable, a very accurate result.

The *t statistic* for each of these is small – the range by which each predicted outcome is expected to be off within a standard deviation. For example, for each 1000 people added to this hypothetical population, they are expected to buy another 660.6 vehicles, with a range of around 5.3.

As the P value is quite small – too small to show up in these results – the *f value* can be used. Using the same Python library, the f value was calculated to be 1050.87. The f value is a measure of the statistical significance of the combined regression model – it is the variance between groups divided by the variance within groups. As it is far greater than 1, it can be assumed that the combined regression is quite accurate.

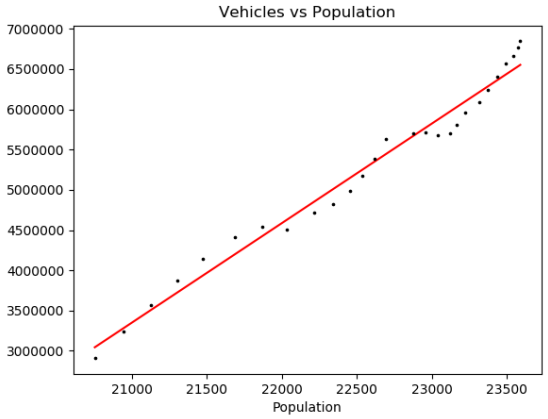
β_0 (which is sometimes written as "α" for reasons I don't entirely understand) is the y-intercept of the regression model. It is a negative number, which implies that this data set might me a good candidate for truncated regression – it is impossible to have negative people.

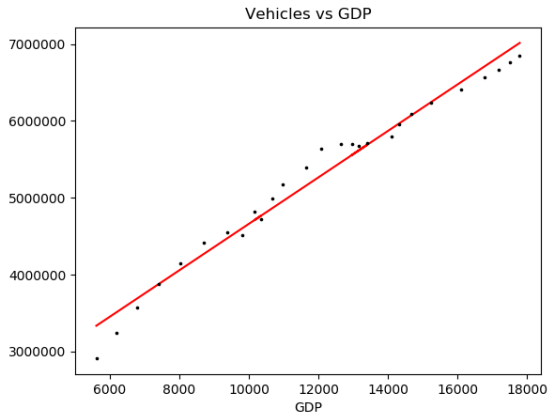
β_1 and β_2 are 144.4769 and 660.6134 respectively. As the population measured in thousands and the GDP in billions, an increase in NT$1 billion would predict 144 new vehicles bought, and a population increase of 1000 would imply 661 more vehicles on the road.

2.2 The second half of question 2 goes on to test the assumptions of the model:

2.2.1 自變數與應變數間成線性關係 (linear relationships)

A visual inspection of the data above indicates that the data scales reasonably linearly, as shown by each of the linear regressions plotted independently.





2.2.2 自變數間無嚴重的共線性關係 (no serious multicollinearity)

This can be accomplished through analyzing the variance inflation factor of the variables. Python's statsmodel output the following:

```
     VIF Factor    features
0  11003.324735       const
1     20.938680         GDP
2     20.938680  Population
```

Ignoring “const,” which is the y-intercept, GDP and population each have a VIF of 20.9, which means that the potential error is 4.6 times larger (the square root of 20.9) than if the variables had been completely random.

2.2.3 誤差項變異數同質 (homocedasticity)

While we were taught the park-glasjer test in class, I was sadly unable to accomplish it on my own, and instead used the Breusch-Pagan test for homoscedasticity, using the same Python packages:

```
Lagrange multiplier statistic:      4.648969410983504
p-value:                            0.09783384430743966
f-value:                            2.4959758660621487
f p-value:                          0.10356320092226594
```

The p-value is quite high for this, at around 9 percent, so there is a statistically significant change of heteroscedasticity. From just looking at the plots created in 2.2.1, I would not have assumed this to be the case.

2.2.4 誤差項間無自我相關 (no serious auto correlation

The Durbin-Watson test can be used to test for auto correlation, and is listed in the original regression results as producing 0.551. Weirdly, this test is on a scale of 0 to 4, where 2 means zero auto correlation; this means that this data has a strong positive auto correlation. This is apparently a common issue with time series data.

2.2.5 誤差項符合標準常態分配 (normality)

To test for normality of the residual errors for this data, the Jarque-Bera test can be used, which produced a result of 0.297. As this is below one, this means that the results are within a 5 percent margin of error.