

## Problem Set #4

MACS 30100, Dr. Evans

Due Monday, Feb. 5 at 11:30am

1. **Multiple linear regression (4 points).** For this problem, you will use the 397 observations from the [Auto.csv](#) dataset.<sup>1</sup> This dataset includes 397 observations on miles per gallon (`mpg`), number of cylinders (`cylinders`), engine displacement (`displacement`), horsepower (`horsepower`), vehicle weight (`weight`), acceleration (`acceleration`), vehicle year (`year`), vehicle origin (`origin`), and vehicle name (`name`).
  - (a) Import the data using `pandas.read_csv()` function. Look for characters that seem out of place that might indicate missing values. Replace them with missing values using the `na_values=...` option.
  - (b) Produce a scatterplot matrix which includes all of the quantitative variables (`mpg`, `cylinders`, `displacement`, `horsepower`, `weight`, `acceleration`, `year`, `origin`). [Use the pandas scatterplot function in the code block below.]

```
from pandas.plotting import scatter_matrix
scatter_matrix(df_quant, alpha=0.3, figsize=(6, 6),
               diagonal='kde')
```

- (c) Compute the correlation matrix for the quantitative variables ( $8 \times 8$ ) using the `DataFrame.corr()` method.
  - (d) Estimate the following multiple linear regression model of `mpg` on all other quantitative variables, where  $u_i$  is an error term for each observation, using Python's `statsmodels.api.OLS()` function.

$$mpg_i = \beta_0 + \beta_1 cylinders_i + \beta_2 displacement_i + \beta_3 horsepower_i + \dots \\ \beta_4 weight_i + \beta_5 acceleration_i + \beta_6 year_i + \beta_7 origin_i + u_i$$

- i. Which of the coefficients is statistically significant at the 1% level?
    - ii. Which of the coefficients is NOT statistically significant at the 10% level?
    - iii. Give an interpretation in words of the estimated coefficient  $\hat{\beta}_6$  on  $year_i$  using the estimated value of  $\hat{\beta}_6$ .
  - (e) Looking at your scatterplot matrix from part (b), what are the three variables that look most likely to have a nonlinear relationship with  $mpg_i$ ?
    - i. Estimate a new multiple regression model by OLS in which you include squared terms on the three variables you identified as having a nonlinear relationship to  $mpg_i$  as well as a squared term on  $acceleration_i$ .

---

<sup>1</sup>The [Auto.csv](#) dataset comes from [James et al. \(2017, Ch. 3\)](#) and is available at <http://www-bcf.usc.edu/~gareth/ISL/data.html>.

- ii. Report your adjusted R-squared statistic. Is it better or worse than the adjusted R-squared from part (d)?
  - iii. What happened to the statistical significance of the *displacement<sub>i</sub>* variable coefficient and the coefficient on its squared term?
  - iv. What happened to the statistical significance of the cylinders variable?
- (f) Using the regression model from part (e) and the `.predict()` function, what would be the predicted miles per gallon *mpg* of a car with 6 cylinders, displacement of 200, horsepower of 100, a weight of 3,100, acceleration of 15.1, model year of 1999, and origin of 1?
2. **Classification problem: KKN by hand and in Python (3 points).** The table below provides a training data set containing six observations, three predictors, and one qualitative response variable.

Obs.	$X_1$	$X_2$	$X_3$	$Y$	Eucl. Dist. from $X_1 = X_2 = X_3 = 0$
1	0	3	0	Red	
2	2	0	0	Red	
3	0	1	3	Red	
4	0	1	2	Green	
5	-1	0	1	Green	
6	1	1	1	Red	

Suppose we wish to use this data set to make a prediction for  $Y$  when  $X_1 = X_2 = X_3 = 0$  using  $K$ -nearest neighbors.

- (a) Compute the Euclidean distance between each observation and the test point  $X_1 = X_2 = X_3 = 0$ .

$$\text{dist}(\mathbf{p}, \mathbf{q}) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + (q_3 - p_3)^2}$$

- (b) What is our KNN prediction with  $K = 1$ ? Why?
- (c) What is our KNN prediction with  $K = 3$ ? Why?
- (d) If the Bayes (optimal) decision boundary in this problem is highly non-linear, then would we expect the best value for  $K$  to be large or small? Why?
- (e) Use Python's `scikit-learn` library to estimate the KNN classifier of the test point  $X_1 = X_2 = X_3 = 1$  with  $K = 2$ .

3. **Multivariable logistic (logit) regression (3 points).**

## References

**James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani,**  
*An Introduction to Statistical Learning with Applications in R* Springer Texts in  
Statistics, Springer, 2017.