Econ 425T Homework 1

Richard Grigorian (UID: 505-088-797)

1/18/23

Contents

1 Filling gaps in lecture notes (10pts)

Consider the regression model

$$Y = f(X) + \varepsilon,$$

where $\mathbb{E}(\varepsilon) = 0$.

1.1 Optimal regression function

Show that the choice

$$f_{\text{opt}}(X) = \mathbb{E}(Y|X)$$

minimizes the mean squared prediction error

$$\mathbb{E}[Y - f(X)]^2$$

where the expectations averages over variations in both X and Y. (Hint: condition on X.)

Solution

We want to show that

$$f_{\text{opt}}(X) = \mathbb{E}(Y|X) = \underset{f(X)}{\arg\min}\ \mathbb{E}[Y - f(X)]^2$$

We begin by conditioning the expression on X and expanding such that

$$\mathbb{E}[(Y - f(X))^2|X = x] = \mathbb{E}[Y^2|X = x] - 2f(x)\mathbb{E}[Y|X = x] + (f(X))^2$$

by the fact that $\mathbb{E}[f(X)|X = x] = f(X)$. Then we differentiate with respect to $f(x)$ for minimization:

$$\frac{d}{df(X)} = 0 \implies -2\mathbb{E}[Y|X = x] + 2f(X) = 0$$

Rearranging the expression gives us

$$f_{\text{opt}}(X) = \mathbb{E}(Y|X)$$

1.2 Bias-variance trade-off

Given an estimate $\hat{f}$ of f, show that the test error at a x_0 can be decomposed as

$$\mathbb{E}[y_0 - \hat{f}(x_0)]^2 = \underbrace{\text{Var}(\hat{f}(x_0)) + [\text{Bias}(\hat{f}(x_0))]^2}_{\text{MSE of } \hat{f}(x_0) \text{ for estimating } f(x_0)} + \underbrace{\text{Var}(\varepsilon)}_{\text{irreducible}},$$

where the expectation averages over the variability in y_0 and $\hat{f}$.

Solution

For simplicity, let $\hat{f} \equiv \hat{f}(x)$ and $f \equiv f(x)$. Then we derive:

$$\mathbb{E}[y_0 - \hat{f}]^2 = \mathbb{E}[(f + \varepsilon - \hat{f})^2]$$

by adding and subtracting $\mathbb{E}[\hat{f}]$ from the inside of the square. Then we can group and expand the square such that:

$$\begin{aligned}
&\mathbb{E}\Big[((f - \mathbb{E}[\hat{f}]) + \varepsilon + (\mathbb{E}[\hat{f}] - \hat{f}))^2\Big] = \\
&= \mathbb{E}\Big[(f - \mathbb{E}[\hat{f}])^2\Big] + \mathbb{E}[\varepsilon^2] + \mathbb{E}\Big[(\mathbb{E}[\hat{f}] - \hat{f})^2\Big] + \\
&+ 2\mathbb{E}\Big[(f - \mathbb{E}[\hat{f}])\varepsilon\Big] + 2\mathbb{E}\Big[\varepsilon(\mathbb{E}[\hat{f}] - \hat{f})\Big] + 2\mathbb{E}\Big[(\mathbb{E}[\hat{f}] - \hat{f})(f - \mathbb{E}[\hat{f}])\Big]
\end{aligned}$$

Note that $\mathbb{E}[\varepsilon] = 0$ and since f is deterministic, $\mathbb{E}[f] = f$:

$$\begin{aligned}
&= (f - \mathbb{E}[\hat{f}])^2 + \mathbb{E}[\varepsilon^2] + \mathbb{E}\Big[(\mathbb{E}[\hat{f}] - \hat{f})^2\Big] + \\
&+ 2(f - \mathbb{E}[\hat{f}]) \cdot 0 + 2(0 \cdot \mathbb{E}(\mathbb{E}[\hat{f}] - \hat{f})) + 2\mathbb{E}\Big[(\mathbb{E}[\hat{f}] - \hat{f})\Big](f - \mathbb{E}[\hat{f}])
\end{aligned}$$

by Linearity and Independence of Expectations. Namely, the last term cancels out since $\mathbb{E}\Big[\mathbb{E}[\hat{f}]\Big] = \mathbb{E}[\hat{f}]$ which implies $\mathbb{E}[\hat{f}] - \mathbb{E}[\hat{f}] = 0$ which cancels out the last term. This results in:

$$(f - \mathbb{E}[\hat{f}])^2 + \mathbb{E}[\varepsilon^2] + \mathbb{E}\Big[(\mathbb{E}[\hat{f}] - \hat{f})^2\Big] = (f - \mathbb{E}[\hat{f}])^2 + \text{Var}(\varepsilon) + \text{Var}(\hat{f})$$

since $\text{Var}(x) = \mathbb{E}\Big[(x - \mathbb{E}[x])^2\Big]$ and $\text{Var}(x) = \mathbb{E}[x^2] - \mathbb{E}[x]^2$ which implies $\mathbb{E}[x^2] = \text{Var}(x) + \mathbb{E}[x]^2$. Namely, $\mathbb{E}[\varepsilon^2] = \text{Var}(\varepsilon) + \mathbb{E}[\varepsilon]^2$ where the last term equals 0 such that $\mathbb{E}[\varepsilon^2] = \text{Var}(\varepsilon)$. Hence, we finally have:

$$\mathbb{E}[y_0 - \hat{f}(x_0)]^2 = \text{Var}(\hat{f}(x_0)) + [\text{Bias}(\hat{f}(x_0))]^2 + \text{Var}(\varepsilon)$$

where $\text{Bias}(\hat{f}(x_0)) \equiv \mathbb{E}[\hat{f}] - f$.

2 ISL Exercise 2.4.3 (10pts)

We now revisit the bias-variance decomposition.

2.1 Part (a)

Provide a sketch of typical (squared) bias, variance, training error, test error, and Bayes (or irreducible) error curves, on a single plot, as we go from less flexible statistical learning methods towards more flexible approaches. The x-axis should represent the amount of flexibility in the method, and the y-axis should represent the values for each curve. There should be five curves. Make sure to label each one.

Solution

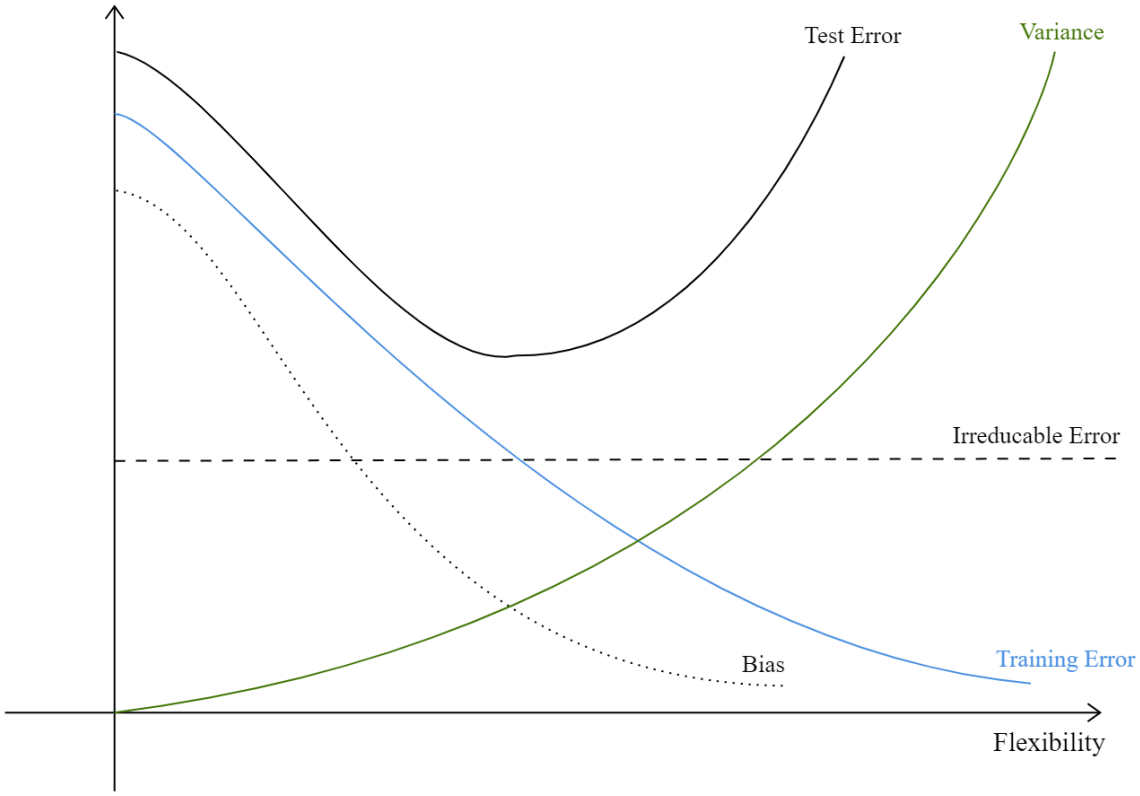


Figure 1: Bias-Variance Decomposition Sketch

Refer to figure 1.

2.2 Part (b)

Explain why each of the five curves has the shape displayed in part (a).

Solution

Squared Bias monotonically decreases due to the fact that as we increase the fledxibility of the model, we get a closer and closer fit. Thus, the bias decreaes. Variance, on the other hand, monotonically increases since the increased flexibility results in overfitting. Training error monotonically decreaes for the same reaon as bias, as we increase the flexbility we get a closer fit to the data. Testing error is concave since increased flexibility improves the fit up until the model begins to overfit to the data, then the error increases in the testing set. Lastly, irreducible error is a constant function since, as the name reveals, it is irreducible. It is ever presesent since it is inherrent to the methodology. However, the irreducible error aslo serves as bounds for the other curves. For instance, when the training error is below the irreducible error, our model has overfitt4ed the data.

Beyond all of this, we notice that all curves are weakly greater than 0.

3 ISL Exercise 2.4.4 (10pts)

You will now think of some real-life applications for statistical learning.

3.1 Part (a)

Describe three real-life applications in which classification might be useful. Describe the response, as well as the predictors. Is the goal of each application inference or prediction? Explain your answer.

Solution

Classification

1. Loan under-writing. Classification could be useful in determining whether or not a loan applicant is likely to default on their loan. For instance, the response would be a binary of *default* or *did not default* and the predictors could be any number of variable such as credit score, age, income, loan duration, etc. TThe goal of this application could be either inference or prediction, but industry professionals would most likely use this classification regression to predict loan defaults in order to determine the success of their loan portfolio.

2. Transportation using bus or car. Classification could be useful in understanding how citizens decide whether or not to take public transit or individual transportation. The response would be a binary of *took the bus* or *drove car* and the predictors could be variables such as bus trip duration, if the bus arrives on time, distance to destination, etc. The goal of this applicatino would be inference. City transportaion planners would want to understand what elements drive individuals to take public transit or their own individual means of transportation in hopes of solving city traffic problems.
3. Stock market movements. Classificatino woudl be useful in predicting whether or not the stock market prices increase or decrease on a given day. The response would be *positive movement* or *negative movement* and the predictors could be varaibles such as yesterday's price change, two previous day price change, day of the week, earnings announcements, etc. This would be a prediction problem with the hopes of being able to improve one's stock market performance.

3.2 Part (b)

Describe three real-life applications in which regression might be useful. Describe the response, as well as the predictors. Is the goal of each application inference or prediction? Explain your answer.

Solution

Regression

1. Real Exchange rate movements. The response variable would be the real exchange rate between country i and j and the predictors would be variables such as relative inflation rates, nominal bi-lateral exchange rates, distance between countries, etc. The goal of this problem would be inference. Since exchange rates, in general, are a major source of puzzles in international macro, regression models are often used to help develop inference how they respond to different macroeconomic variables.
2. The effects of minimum wage. The response variable would be unemployment in a particular city. The predictors would be variables such as the minimum wage in the particular city, indicator variables to control for industry type, etc. The goal of this regression problem would be inference. We want to understand how mininmum wage affects unemployment in different cities conditional on the type of industry to which they are applied.
3. Salary. The response variable would the salary of a given individual. The predictors would be variables such as education, years of work experience, indicator variables describing the industry they are in, etc. The goal of this regression problem could be either inference or prediction. But we say it is prediction with the idea being that we could predict the approximate salary of an individual based off their education as a way to sell higher education degrees like MBAs.

3.3 Part (c)

Describe three real-life applications in which cluster analysis might be useful.

Solution

Cluster Analysis

1. Spotify music recommendations. The idea would be that by clustering a user's msuci preferences, Spotify could recommend artists or albums based on the users who have listened to and liked similar types of music.
2. NFL analytics. The idea would be to collect data on players and conduct a cluster analysis to see which players have similar statistics to each other and thus are maybe similar in play styles. These players could then be matched up in practice to perform drills and conditioning.
3. Marketing. Using surveys, data scientists could cluster demographics for products to determine which clusters of individuals in which geography prefer to buy certain sets of goods. This would obviously be useful for firms deciding which products to promote in which markets.

4 ISL Exercise 2.4.10 (30pts)

4.1 Part (a)

Load in the Boston data set. How many rows are in this data set? How many columns? What do the rows and columns represent?

Solution

```
import pandas as pd
import io
import requests
import seaborn as sns
import matplotlib.pyplot as plt
import numpy as np
import statsmodels.formula.api as smf

url = "https://raw.githubusercontent.com/ucla-econ-425t/2023winter/master/slides/data/Bost
s = requests.get(url).content
Boston = pd.read_csv(io.StringIO(s.decode('utf-8')), index_col = 0)
Boston
```

	crim	zn	indus	chas	nox	rm	age	dis	rad	tax	ptratio	lstat	medv
1	0.00632	18.0	2.31	0	0.538	6.575	65.2	4.0900	1	296	15.3	4.98	24.0
2	0.02731	0.0	7.07	0	0.469	6.421	78.9	4.9671	2	242	17.8	9.14	21.6
3	0.02729	0.0	7.07	0	0.469	7.185	61.1	4.9671	2	242	17.8	4.03	34.7
4	0.03237	0.0	2.18	0	0.458	6.998	45.8	6.0622	3	222	18.7	2.94	33.4
5	0.06905	0.0	2.18	0	0.458	7.147	54.2	6.0622	3	222	18.7	5.33	36.2
...	...	...	...	...	...	...	...	...	...	...	...	...	...
502	0.06263	0.0	11.93	0	0.573	6.593	69.1	2.4786	1	273	21.0	9.67	22.4
503	0.04527	0.0	11.93	0	0.573	6.120	76.7	2.2875	1	273	21.0	9.08	20.6
504	0.06076	0.0	11.93	0	0.573	6.976	91.0	2.1675	1	273	21.0	5.64	23.9
505	0.10959	0.0	11.93	0	0.573	6.794	89.3	2.3889	1	273	21.0	6.48	22.0
506	0.04741	0.0	11.93	0	0.573	6.030	80.8	2.5050	1	273	21.0	7.88	11.9

It is immediately clear from the Python output that the datset has 506 rows and 13 columns. The rows represent the number of observations $n = 506$. In the context of this dataset, this means that we have housing data on 506 suburbs of Boston. The columns of this dataset represnt all of the variables we have access to. Namely, for this datset, they are the different characteristics of the Boston Suburbs.

4.2 Part (b)

Make some pairwise scatterplots of the predictors (columns) in this data set. Describe your findings.

Solution

```
plt.figure(figsize=(8,8))
g = sns.pairplot(
    Boston,
    x_vars = ["crim", "nox", "age", "ptratio", "lstat"],
    y_vars = ["crim", "nox", "age", "ptratio", "lstat"],
    corner = True,
)
g.fig.suptitle("Some Pairwise Scatterplots", fontsize=18)
```

```
Text(0.5, 0.98, 'Some Pairwise Scatterplots')
```

```
<Figure size 576x576 with 0 Axes>
```

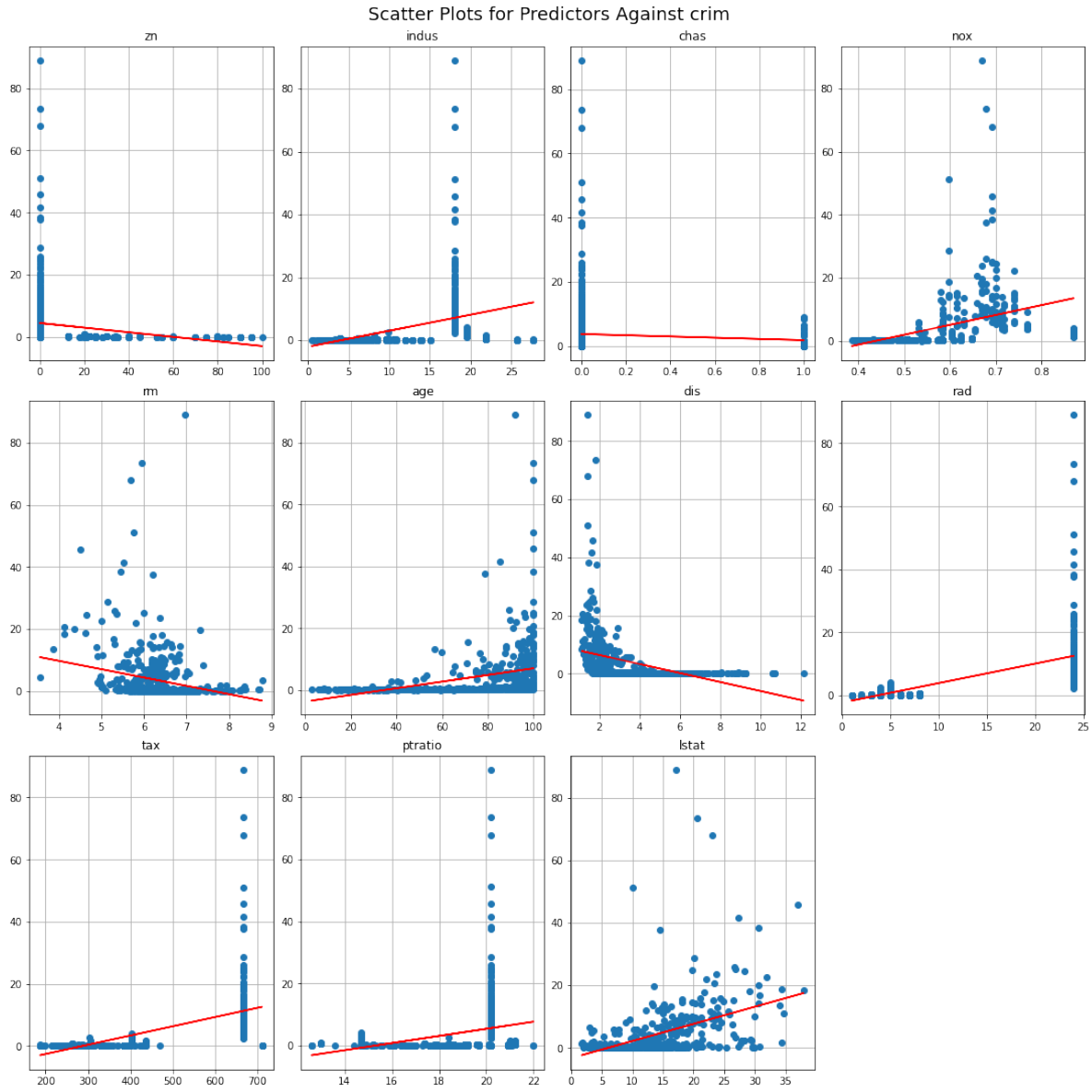
From the above scatterplot matrix, it seems that crim may be positively associated with lstat, nox may be positively associated with age, ptratio, and lstat, age may be positively associated with lstat, and ptratio may be positively associated with lstat. Some of these relationships look vaguely linear whilst others look logrithmic. Despite having only chosen a small subset of possible pairwise combinations, it does appear that some of our predictor variables are positively associated.

4.3 Part (c)

Are any of the predictors associated with per capita crime rate? If so, explain the relationship.

Solution

```
plt.figure(figsize=(15,15), constrained_layout=True)
plt.suptitle("Scatter Plots for Predictors Against crim", fontsize=18)
independent = ['zn', 'indus', 'chas', 'nox', 'rm', 'age', 'dis',
               'rad', 'tax', 'ptratio', 'lstat']
count = 1
for i in independent:
    plt.subplot(3, 4, count)
    x = Boston[i]
    y = Boston['crim']
    plt.plot(x, y, 'o')
    m, b = np.polyfit(x, y, 1)
    plt.plot(x, m*x+b, 'red')
    plt.title(i)
    plt.grid()
    count += 1
plt.show()
```



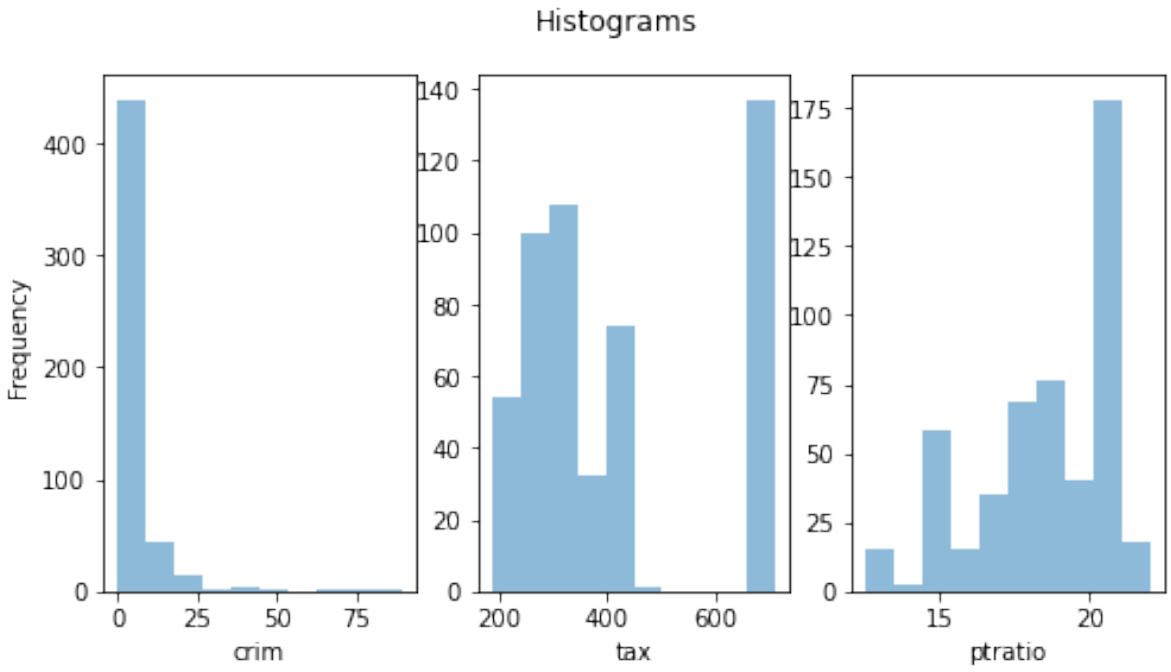
Generally speaking we see that the crime rate is decreasing in zone, increasing in industry, decreasing in proximity to the Charles River, increasing in concentration of nitric oxides, decreasing in rooms, increasing in age, decreasing in distance to employment centers, increasing in accessibility to highways, increasing in tax value, increasing in pupil-teacher ratio, and increasing in lower status percentage. It is important to note that these associations are extremely rough and are simply the result of a naive simple regression. The specific relationship as shown by the scatterplot would take a more sophisticated model to explain. Namely, some variables appear to need log transformations.

4.4 Part (d)

Do any of the census tracts of Boston appear to have particularly high crime rates? Tax rates? Pupil-teacher ratios? Comment on the range of each predictor.

Solution

```
plt.figure(figsize=(8,4))
plt.suptitle('Histograms')
plt.subplot(1,3,1)
plt.hist(Boston['crim'], alpha=.5)
plt.ylabel("Frequency")
plt.xlabel("crim")
plt.subplot(1,3,2)
plt.hist(Boston['tax'], alpha=.5)
plt.xlabel("tax")
plt.subplot(1,3,3)
plt.hist(Boston['ptratio'], alpha=.5)
plt.xlabel("ptratio")
plt.show()
```



```
print('crim: ', Boston[Boston.crim > 25].index)
```

```
crim:  Int64Index([381, 399, 401, 405, 406, 411, 414, 415, 418, 419, 428], dtype='int64')
```

From the above histograms, it is clear that most suburbs in Boston have a low crime rate; however, there is a long tail to the right. This implies that there are a handful of cities that do have a higher crime rate. Notably, there are 11 suburbs with a crime rate greater than 25. Beyond this, tax seems to have a large amount of suburbs with very high tax. The number of observations greater than 600 is amongst greater than any other bracket. And lastly, with regards to ptratio, there are many suburbs with ptratio greater than 20. Tax and Ptratio are unlike crim in that they are more evenly distributed.

4.5 Part (e)

How many of the cenus tracts in this data set bound the Charles River

Solution

```
print("Suburbs near river: ", len(Boston[Boston["chas"] ==1]))
```

```
Suburbs near river:  35
```

There are 35 suburbs near the Charles River in this dataset.

4.6 Part (f)

What is the median pupil-teacher ratio among the towns in this data set?

Solution

```
print('Median pupil-teach ratio:', Boston['ptratio'].median())
```

```
Median pupil-teach ratio: 19.05
```

The median pupil-teacher ratio among the towns in this data set is 19.05.

4.7 Part (g)

Solution

Which census tract of Boston has lowest median value of owner- occupied homes? What are the values of the other predictors for that census tract, and how do those values compare to the overall ranges for those predictors? Comment on your findings.

```
medv_min = Boston['medv'].min()
print('Min medv Index:', Boston[Boston["medv"] == medv_min].index)
print('Observation 399:')
print(Boston.iloc[[398]])
print('Observation 406:')
print(Boston.iloc[[405]])
print(Boston.describe())
```

```
Min medv Index: Int64Index([399, 406], dtype='int64')
Observation 399:
        crim   zn  indus  chas    nox     rm    age     dis  rad  tax  \
399  38.3518  0.0   18.1     0  0.693  5.453  100.0  1.4896   24  666

     ptratio  lstat  medv
399     20.2  30.59   5.0
Observation 406:
        crim   zn  indus  chas    nox     rm    age     dis  rad  tax  \
406  67.9208  0.0   18.1     0  0.693  5.683  100.0  1.4254   24  666

     ptratio  lstat  medv
406     20.2  22.98   5.0
             crim          zn       indus        chas         nox          rm  \
count  506.000000  506.000000  506.000000  506.000000  506.000000  506.000000
mean     3.613524   11.363636   11.136779    0.069170    0.554695    6.284634
std      8.601545   23.322453    6.860353    0.253994    0.115878    0.702617
min      0.006320    0.000000    0.460000    0.000000    0.385000    3.561000
25%      0.082045    0.000000    5.190000    0.000000    0.449000    5.885500
50%      0.256510    0.000000    9.690000    0.000000    0.538000    6.208500
75%      3.677083   12.500000   18.100000    0.000000    0.624000    6.623500
max     88.976200  100.000000   27.740000    1.000000    0.871000    8.780000

              age         dis         rad         tax     ptratio       lstat  \
count  506.000000  506.000000  506.000000  506.000000  506.000000  506.000000
mean    68.574901    3.795043    9.549407  408.237154   18.455534   12.653063
```

```
std     28.148861    2.105710    8.707259  168.537116    2.164946    7.141062
min      2.900000    1.129600    1.000000  187.000000   12.600000    1.730000
25%     45.025000    2.100175    4.000000  279.000000   17.400000    6.950000
50%     77.500000    3.207450    5.000000  330.000000   19.050000   11.360000
75%     94.075000    5.188425   24.000000  666.000000   20.200000   16.955000
max    100.000000   12.126500   24.000000  711.000000   22.000000   37.970000
```

```
             medv
count  506.000000
mean    22.532806
std      9.197104
min      5.000000
25%     17.025000
50%     21.200000
75%     25.000000
max     50.000000
```

The crime rate in both of these areas is above the 75% quartile. They are both at the minimum zone. They are both at the 75% quartile indus value. They are not by the Charles River. They are both above the 75% quartile in terms of nox. They are both below the 25% quartile in the number of rooms. They are at the maximium value for age. They are below the 25% quartile for distance. They are at the maximum value for rad. They are below the 75% quartile for tax and ptratio. They are both above the 75% quartile for lstat. And of course, are both at the minimum value for medv.

Overall, most of the variables fall into what would be considered undesirable; however, the home prices are also very low so perhaps it is a trade-off.

4.8 Part (h)

In this data set, how many of the census tracts average more than seven rooms per dwelling? More than eight rooms per dwelling? Comment on the census tracts that average more than eight rooms per dwelling.

Solution

```
print('Number of suburbs > 7 rooms:', len(Boston[Boston["rm"] > 7]))
print('Number of suburbs > 8 rooms:', len(Boston[Boston["rm"] > 8]))
Boston_parth = Boston[Boston["rm"] > 8]
print('Summary for suburbs > 8 rooms:')
print(Boston_parth.describe())
```

```
Number of suburbs > 7 rooms: 64
Number of suburbs > 8 rooms: 13
Summary for suburbs > 8 rooms:
            crim         zn      indus       chas        nox         rm  \
count  13.000000  13.000000  13.000000  13.000000  13.000000  13.000000
mean    0.718795  13.615385   7.078462   0.153846   0.539238   8.348538
std     0.901640  26.298094   5.392767   0.375534   0.092352   0.251261
min     0.020090   0.000000   2.680000   0.000000   0.416100   8.034000
25%     0.331470   0.000000   3.970000   0.000000   0.504000   8.247000
50%     0.520140   0.000000   6.200000   0.000000   0.507000   8.297000
75%     0.578340  20.000000   6.200000   0.000000   0.605000   8.398000
max     3.474280  95.000000  19.580000   1.000000   0.718000   8.780000

             age        dis        rad         tax    ptratio      lstat  \
count  13.000000  13.000000  13.000000   13.000000  13.000000  13.000000
mean   71.538462   3.430192   7.461538  325.076923  16.361538   4.310000
std    24.608723   1.883955   5.332532  110.971063   2.410580   1.373566
min     8.400000   1.801000   2.000000  224.000000  13.000000   2.470000
25%    70.400000   2.288500   5.000000  264.000000  14.700000   3.320000
50%    78.300000   2.894400   7.000000  307.000000  17.400000   4.140000
75%    86.500000   3.651900   8.000000  307.000000  17.400000   5.120000
max    93.900000   8.906700  24.000000  666.000000  20.200000   7.440000

            medv
count  13.000000
mean   44.200000
std     8.092383
min    21.900000
25%    41.700000
50%    48.300000
75%    50.000000
max    50.000000
```

It is clear that for the suburbs where the average house has greater than 8 rooms that the crime rate is substantially lower than the overall average. Additionally, the median value is unsurprisingly higher and the ptratio is lower. All other variables are unremarkable to compare.

5 ISL Exercise 3.7.3 (12pts)

Suppose we have a data set with five predictors, $X_1 = GPA$, $X_2 = IQ$, $X_3 = Level$ (1 for College and 0 for High School), $X_4 = GPA \times IQ$, and $X_5 = GPA \times Level$. The response is starting

salary after graduation (in thousands of dollars). Suppose we use least squares to fit the model and get $\hat{\beta}_0 = 50$, $\hat{\beta}_1 = 20$, $\hat{\beta}_2 = 0.07$, $\hat{\beta}_3 = 35$, $\hat{\beta}_4 = 0.01$, $\hat{\beta}_5 = -10$.

5.1 Part (a)

Which answer is correct and why?

Solution

From the above information, we write the following regression equation:

$$\widehat{salary}_i = 50 + 20GPA_i + 0.07IQ_i + 35Level_i + 0.01(GPA_i \times IQ_i) - 10(GPA_i \times Level_i)$$

For a fixed value of IQ and GPA, the affect of Level is:

$$\frac{\partial salary}{\partial Level} = 35 - 10GPA_i$$

Hence, (iii) is correct. High school graduates earn more, on average, than college graduates provided that the GPA is high enough.

5.2 Part (b)

Predict the salary of a college graduate with IQ of 110 and a GPA of 4.0.

Solution

$$\widehat{salary}_i = 50 + 20 \cdot (4.0) + 0.07 \cdot (110) + 35 \cdot (1) + 0.01 \cdot (4.0 \times 110) - 10 \cdot (4.0 \times 1) = 137.1$$

Hence, the predicted average salary of a college graduate with an IQ of 110 and a GPA of 4.0 is 137.1 thousand dollars.

5.3 Part (c)

True or false: Since the coefficient for the GPA/IQ interaction term is very small, there is very little evidence of an interaction effect. Justify your answer.

Solution

False!! We would need to know the p-value of the regression coefficient to decide whether or not it is statisically significant. Typically, regression results are reported with standard errors which would allow us to compute statistical significance at different confidence levels. This output does not provide that information. The β value is not reflective of statistical significance.

6 ISL Exercise 3.7.15 (20pts)

This problem involves the Boston data set, which we saw in the lab for this chapter. We will now try to predict per capita crime rate using the other variables in this data set. In other words, per capita crime rate is the response, and the other variables are the predictors.

6.1 Part (a)

For each predictor, fit a simple linear regression model to predict the response. Describe your results. In which of the models is there a statistically significant association between the predictor and the response? Create some plots to back up your assertions.

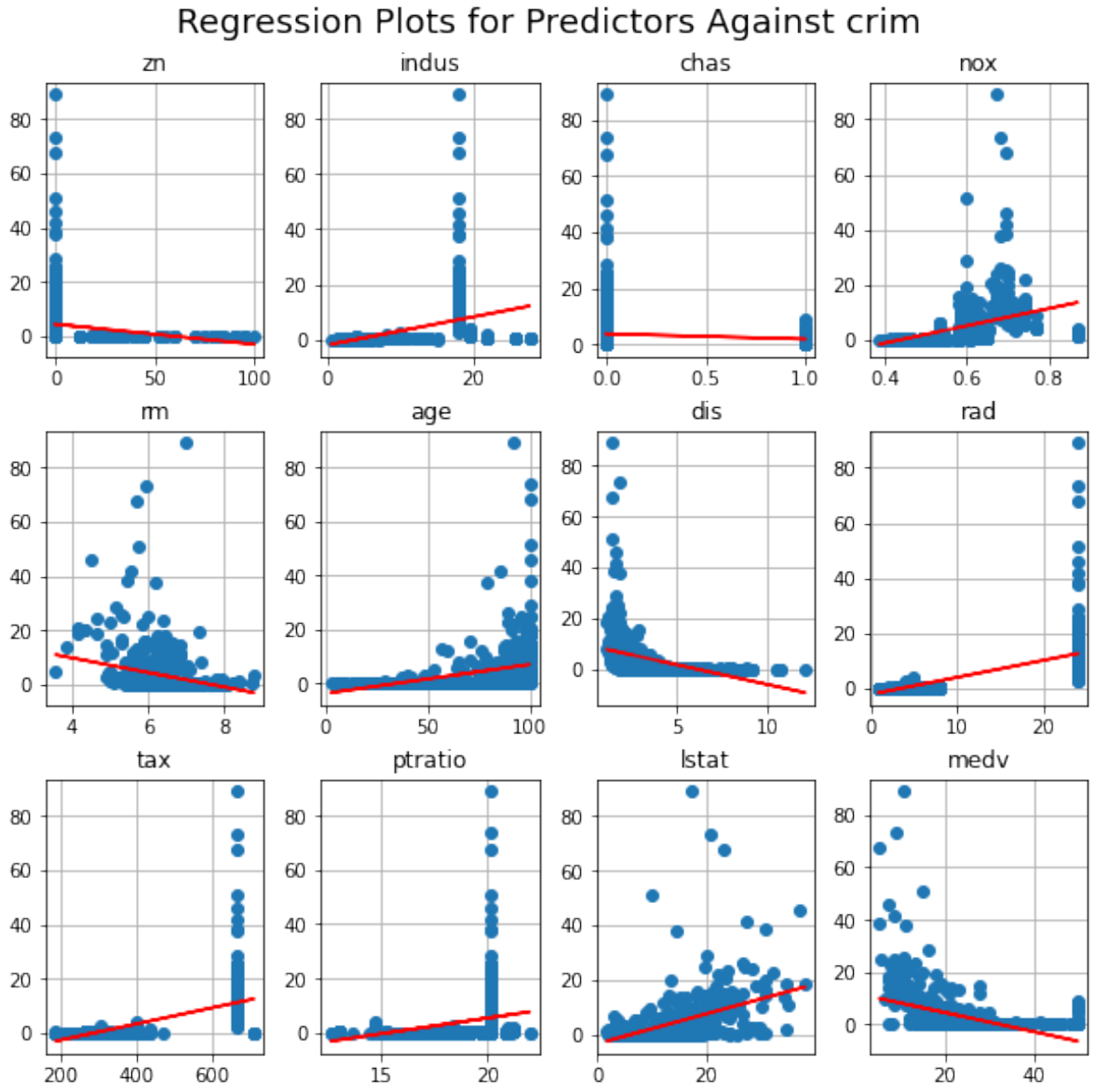
Solution

```
# Regressions
independent = ['zn', 'indus', 'chas', 'nox', 'rm', 'age', 'dis',
               'rad', 'tax', 'ptratio', 'lstat', 'medv']
for i in independent:
    reg = smf.ols('crim~' + i, Boston).fit()
    print(i,'beta:', reg.params[1], "", "|| p-value:", reg.pvalues[1])
```

```
zn beta: -0.07393497740412347  || p-value: 5.506472107679307e-06
indus beta: 0.5097763311042314  || p-value: 1.4503489330272292e-21
chas beta: -1.8927765508037604  || p-value: 0.2094345015352004
nox beta: 31.24853120112292  || p-value: 3.751739260356816e-23
rm beta: -2.6840512241139476  || p-value: 6.346702984687803e-07
age beta: 0.10778622713953306  || p-value: 2.8548693502441573e-16
dis beta: -1.5509016824100998  || p-value: 8.519948766926326e-19
rad beta: 0.6179109273272012  || p-value: 2.6938443981864414e-56
tax beta: 0.029742252822765353  || p-value: 2.357126835257048e-47
ptratio beta: 1.1519827870705859  || p-value: 2.942922447359837e-11
lstat beta: 0.548804782062398  || p-value: 2.6542772314731968e-27
```

```
medv beta: -0.36315992225760296  || p-value: 1.1739870821943694e-19
```

```
plt.figure(figsize=(8,8), constrained_layout=True)
plt.suptitle("Regression Plots for Predictors Against crim", fontsize=18)
count = 1
for i in independent:
    plt.subplot(3, 4, count)
    x = Boston[i]
    y = Boston['crim']
    plt.plot(x, y, 'o')
    m, b = np.polyfit(x, y, 1)
    plt.plot(x, m*x+b, 'red')
    plt.title(i)
    plt.grid()
    count += 1
plt.show()
```



Firstly, every predictor has a statistically significant association with the response **except** for chas. This is evident from the fact that most of the p-values are less than 0.05. Hence, the respective β is statistically significant.

In terms of association, the slope and direction can be seen from the above plot. The plots look the most convincing for nox, rm, age, lstat, and medv. In these plots, the best fit line seems to cross through the centroid of the data and looks to fit the data well. In other instances, a transformation may be necessary.

6.2 Part (b)

Fit a multiple regression model to predict the response using all of the predictors. Describe your results. For which predictors can we reject the null hypothesis $H_0 : \beta_j = 0$?

Solution

```
mreg = smf.ols('crim ~ zn + indus + chas + nox + \
    + rm + age + dis + rad + tax + ptratio + \
        + lstat + medv', Boston).fit()
print(mreg.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                   crim   R-squared:                       0.449
Model:                            OLS   Adj. R-squared:                  0.436
Method:                 Least Squares   F-statistic:                     33.52
Date:                Fri, 13 Jan 2023   Prob (F-statistic):           2.03e-56
Time:                        15:48:32   Log-Likelihood:                -1655.4
No. Observations:                 506   AIC:                             3337.
Df Residuals:                     493   BIC:                             3392.
Df Model:                          12
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     13.7784      7.082      1.946      0.052      -0.136      27.693
zn             0.0457      0.019      2.433      0.015       0.009       0.083
indus         -0.0584      0.084     -0.698      0.486      -0.223       0.106
chas          -0.8254      1.183     -0.697      0.486      -3.150       1.500
nox           -9.9576      5.290     -1.882      0.060     -20.351       0.436
rm             0.6289      0.607      1.036      0.301      -0.564       1.822
age           -0.0008      0.018     -0.047      0.962      -0.036       0.034
dis           -1.0122      0.282     -3.584      0.000      -1.567      -0.457
rad            0.6125      0.088      6.997      0.000       0.440       0.784
tax           -0.0038      0.005     -0.730      0.466      -0.014       0.006
ptratio       -0.3041      0.186     -1.632      0.103      -0.670       0.062
lstat          0.1388      0.076      1.833      0.067      -0.010       0.288
medv          -0.2201      0.060     -3.678      0.000      -0.338      -0.103
==============================================================================
Omnibus:                      663.436   Durbin-Watson:                   1.516
```

```
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            80856.852
Skew:                           6.579   Prob(JB):                        0.00
Kurtosis:                      63.514   Cond. No.                     1.24e+04
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.24e+04. This might indicate that there are
strong multicollinearity or other numerical problems.
```

The multiple regression model has an R^2 of 0.449. Beyond this, the F-stat is high enough such that we can conclude that the coefficients are not all 0. However, despite this, the multiple regression reveals that most of our variables are now jointly insignificant. For instance, indus, chas, nox, rm, age, tax, ptratio, and medv now all have p-values > 0.05. Hence, these coefficients are not significant.

In particular, the only remaining significant predictors (such that we can reject $H_0 : \beta_j = 0$) are dis and rad. Intuitively, this seems strange since these do not seem to economically make the most sense as predictors of crime rates.

6.3 Part (c)

How do your results from (a) compare to your results from (b)? Create a plot displaying the univariate regression coefficients from (a) on the x-axis, and the multiple regression coefficients from (b) on the y-axis. That is, each predictor is displayed as a single point in the plot. Its coefficient in a simple linear regres- sion model is shown on the x-axis, and its coefficient estimate in the multiple linear regression model is shown on the y-axis.

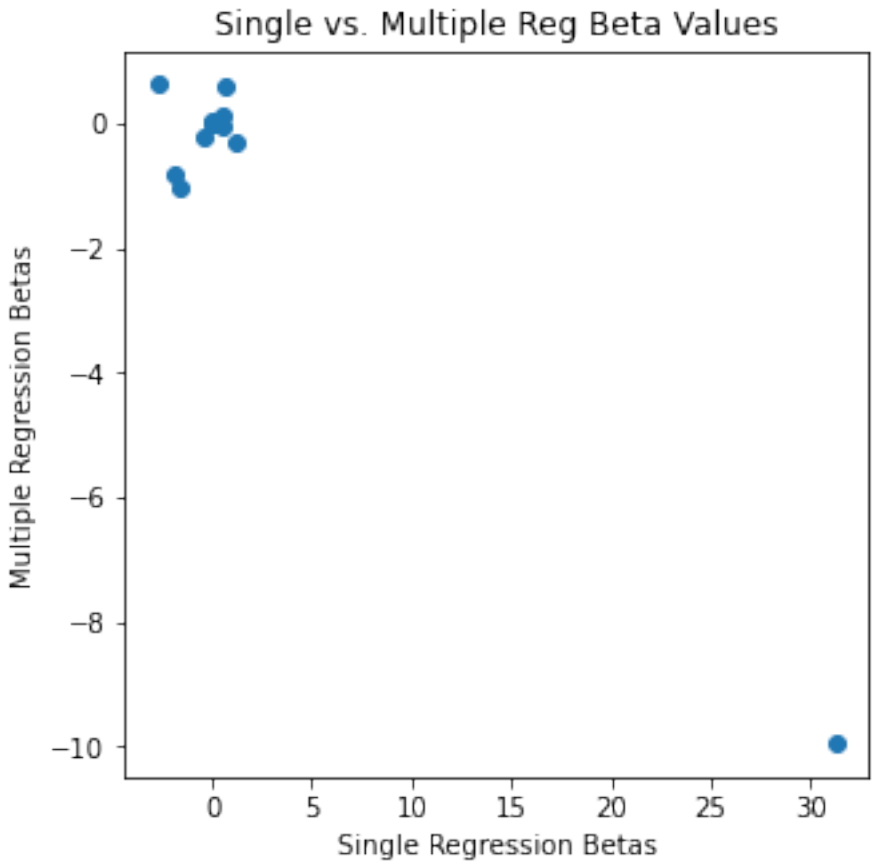
Solution

```
# Putting single regressions coeffs into list
slr_coeffs = []
for i in independent:
    reg = smf.ols('crim~' + i, Boston).fit()
    beta = reg.params[1]
    slr_coeffs.append(beta)

# Putting multiple reg coeffs into list
mlr = mreg.params[1:]
mlr_list = mlr.tolist()

# Scatter plot
```

```
plt.figure(figsize=(5,5))
plt.scatter(slr_coeffs, mlr_list)
plt.title('Single vs. Multiple Reg Beta Values')
plt.ylabel('Multiple Regression Betas')
plt.xlabel('Single Regression Betas')
plt.show()
```



From the above scatterplot it is clear that the majority of predictors did not have their beta values change by a large value (in absolute terms). However, **nox** went from **-10 in single regression** to **31 in multiple regression**. So not only did it change signs, it also change by a large value in absolute terms.

6.4 Part (d)

Is there evidence of non-linear association between any of the predictors and the response? To answer this question, for each predictor X, fit a model of the form

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \varepsilon$$

Solution

Below I will conduct a Ramsey RESET test for higher order terms. Namely, this test does exactly what the prompt asks for. It adds in higher order terms (of degree 2 or 3 depending on the specification) and then it conducts a Wald Test to see whether or not these higher order terms are statistically jointly statistically significant. Hence, I will conduct a RESET test with power of 3. This will allow us to see, in the context of single regression, what predictors display evidence of non-linear association. Namely, the null hypothesis is that the higher order terms have coefficients equal to 0. Hence, if we have low p-values and reject the null then that implies that the higher order terms are statisitcally significant.

```
import statsmodels.stats.diagnostic as dg
import warnings
warnings.filterwarnings("ignore")
print("Ramsey RESET Test:")
for i in independent:
    hp_reg = smf.ols('crim~' + i, Boston).fit()
    RESETtest =  dg.linear_reset(hp_reg, power=3,  test_type='fitted', use_f = True)
    print(i, "RAMSEY RESET p-value:", RESETtest.pvalue)
```

```
Ramsey RESET Test:
zn RAMSEY RESET p-value: 0.008511995129361908
indus RAMSEY RESET p-value: 8.408753824091315e-14
chas RAMSEY RESET p-value: 6.874894185508546e-19
nox RAMSEY RESET p-value: 7.122383468919643e-18
rm RAMSEY RESET p-value: 0.005229426705161654
age RAMSEY RESET p-value: 4.125056401314371e-07
dis RAMSEY RESET p-value: 3.07183726991073e-19
rad RAMSEY RESET p-value: 0.026078319325960476
tax RAMSEY RESET p-value: 1.144238029894827e-05
ptratio RAMSEY RESET p-value: 0.0002541646638093135
lstat RAMSEY RESET p-value: 0.036983220241183876
medv RAMSEY RESET p-value: 2.504778348783454e-42
```

From the above output, it is clear that RESET test for all of the predictors results in a case where each variable displays some non-linear association with the variable crim. Note that this was done in the context of single regression. These results would likely change if done in different combinations of multiple regression. But, in this case, it seems that every model could improve its fit by including a higher order term of the form X^2 or $X^2 + X^3$.

7 Bonus question (20pts)

For multiple linear regression, show that R^2 is equal to the correlation between the response vector $\mathbf{y} = (y_1, ..., y_n)^T$ and the fitted values $\hat{\mathbf{y}} = (\hat{y}_1, ..., \hat{y}_n)^T$. That is

$$R^2 = 1 - \frac{\text{RSS}}{\text{TSS}} = [\text{Cor}(\mathbf{y}, \hat{\mathbf{y}})]^2.$$

Solution

By definition, let

$$\text{RSS} = \hat{\varepsilon}'\varepsilon = (\mathbf{y} - \hat{\mathbf{y}})'(\mathbf{y} - \hat{\mathbf{y}}), \quad \text{TSS} = (\mathbf{y} - \bar{\mathbf{y}})'(\mathbf{y} - \bar{\mathbf{y}}), \quad \text{ESS} = (\hat{\mathbf{y}} - \bar{\mathbf{y}})'(\hat{\mathbf{y}} - \bar{\mathbf{y}})$$

Now, we expand our correlation term such that

$$[\text{Cor}(\mathbf{y}, \hat{\mathbf{y}})]^2 = \left(\frac{\text{Cov}(\mathbf{y}, \hat{\mathbf{y}})}{\sqrt{\text{Var}(\mathbf{y})\,\text{Var}(\hat{\mathbf{y}})}} \right)^2 = \frac{\text{Cov}(\mathbf{y}, \hat{\mathbf{y}})\,\text{Cov}(\mathbf{y}, \hat{\mathbf{y}})}{\text{Var}(\mathbf{y})\,\text{Var}(\hat{\mathbf{y}})}$$

Substituting that $\mathbf{y} = \hat{\mathbf{y}} + \varepsilon$ gives

$$\frac{\text{Cov}(\hat{\mathbf{y}} + \varepsilon, \hat{\mathbf{y}})\,\text{Cov}(\hat{\mathbf{y}} + \varepsilon, \hat{\mathbf{y}})}{\text{Var}(\mathbf{y})\,\text{Var}(\hat{\mathbf{y}})} = \frac{(\text{Cov}(\hat{\mathbf{y}}, \hat{\mathbf{y}}) + \text{Cov}(\hat{\mathbf{y}}, \varepsilon))(\text{Cov}(\hat{\mathbf{y}}, \hat{\mathbf{y}}) + \text{Cov}(\hat{\mathbf{y}}, \varepsilon))}{\text{Var}(\mathbf{y})\,\text{Var}(\hat{\mathbf{y}})}$$

by covariance expansion properties. Additionally, by OLS assumptions $\text{Cov}(\hat{\mathbf{y}}, \varepsilon) = 0$. Thus, our expression can be rewritten as

$$\frac{\text{Var}(\hat{\mathbf{y}})\,\text{Var}(\hat{\mathbf{y}})}{\text{Var}(\mathbf{y})\,\text{Var}(\hat{\mathbf{y}})} = \frac{\text{Var}(\hat{\mathbf{y}})}{\text{Var}(\mathbf{y})}$$

where the covariance of a term with itself is just the variance of the term. We can expand the variance terms into quadratic terms using vector transposes such that

$$\frac{\text{Var}(\hat{\mathbf{y}})}{\text{Var}(\mathbf{y})} = \frac{\frac{1}{n}(\hat{\mathbf{y}} - \bar{\mathbf{y}})'(\hat{\mathbf{y}} - \bar{\mathbf{y}})}{\frac{1}{n}(\mathbf{y} - \bar{\mathbf{y}})'(\mathbf{y} - \bar{\mathbf{y}})} = \frac{\text{ESS}}{\text{TSS}}$$

By the Partitioning the Sum of Squares Theorem for least squares, we say $\text{TSS} = \text{ESS} + \text{RSS}$ which implies

$$\frac{\text{ESS}}{\text{TSS}} = \frac{\text{TSS} - \text{RSS}}{\text{TSS}} = 1 - \frac{\text{RSS}}{\text{TSS}} = R^2$$

Thus,

$$[\text{Cor}(\mathbf{y}, \hat{\mathbf{y}})]^2 = 1 - \frac{\text{RSS}}{\text{TSS}} = R^2.$$