CS 4342 Assignment #2

Ivan Martinovic

Conceptual and Theoretical Questions

1. Describe the null hypotheses to which the p-values given in the below Table correspond. Explain what conclusions you can draw based on these p-values. Your explanation should be phrased in terms of sales, TV, radio, and newspaper, rather than in terms of the coefficients of the linear model.

	Coefficient	Std. error	t-statistic	p-value
Intercept	2.939	0.3119	9.42	< 0.0001
TV	0.046	0.0014	32.81	< 0.0001
radio	0.189	0.0086	21.89	< 0.0001
newspaper	-0.001	0.0059	-0.18	0.8599

TABLE 3.4. *For the* Advertising *data, least squares coefficient estimates of the multiple linear regression of number of units sold on radio, TV, and newspaper advertising budgets.*

Intercept: Null hypothesis is that the Intercept = 0, i.e. when we don't invest any money in any kind of advertising, our sales are zero. The p-value is a measurement of the plausibility that the null hypothesis is true. Since the p-value is extremely small, i.e. there is a very small probability that the null hypothesis is true, we reject the null hypothesis, in favor of the alternative hypothesis: Intercept > 0 or Intercept < 0, i.e. the sales are non-zero when there is no advertising spending (they should be positive, because sales cannot be negative).

TV: Null hypothesis is that the TV coefficient = 0, i.e. investment in TV advertising has no effect on the total sales. Since p-value is again extremely small, we reject the null-hypothesis in favor of the alternative: TV coefficient > 0 or < 0; i.e. the sales are affected (either positively or negatively) by increased TV advertising.

radio: Null hypothesis is that the radio coefficient = 0, i.e. investment in radio advertising has no effect on the total sales. Since p-value is again extremely small, we reject the null-hypothesis in favor of the alternative: radio coefficient > 0 or < 0; i.e. the sales are affected (either positively or negatively) by increased radio advertising.

newspaper: Null hypothesis is that the newspaper coefficient = 0, i.e. investment in newspaper advertising has no effect on the total sales. In this case the p-value is very high, indicating that the plausibility of the null hypothesis is strong. Hence we do not reject the null-hypothesis.

2. Carefully explain the differences between the KNN classifier and KNN regression methods.

The KNN classifier is used in classification problems, for predicting to which class the response variable belongs to. The KNN classifier determines the K nearest neighbors for given predictors, and then assigns the class to the response variable which is found in the highest number of the K selected neighbors. For example: our response variable is whether a fruit is an orange or a lemon, and the predictors are width, height and color. Say for a certain color, width and height, the KNN classifier (with K = 5) finds that, 2 neighbors are oranges and 3 are lemons. It will therefore assign the predicted value to the lemons class.

The KNN regression method is used in estimation problems, for predicting the value of the response variable. The KNN regression method determines the K nearest neighbors for given predictors, and then assigns the value to the response variable, which is an average of the response variables of all K selected neighbors. For example: our response variable is the length of a fruit, and our predictors are width, color and class (orange or lemon). Say for a lemon, with certain width and color, the KNN regression method (K=5) finds that the values of the response variables of the 5 closest neighbors are 5.0, 5.2, 5.4, 5.6, and 5.8. It will therefore predict that the length of the lemon is: avg(5.0, 5.2, 5.4, 5.6, 5.8) = 5.4 for the given predictors.

3. Suppose we have a data set with five predictors, X_1 = GPA, X_2 = IQ, X_3 = Gender (1 for Female and 0 for Male), X_4 = Interaction between GPA and IQ, and X_5 = Interaction between GPA and Gender. The response is starting salary after graduation (in thousands of dollars). Suppose we use least squares to fit the model, and get $\widehat{\beta_0} = 50, \widehat{\beta_1} = 20, \widehat{\beta_2} = 0.07, \widehat{\beta_3} = 35, \widehat{\beta_4} = 0.01,\ \widehat{\beta_5} = -10.$

(a) Which answer is correct, and why?

i. For a fixed value of IQ and GPA, males earn more on average than females.
ii. For a fixed value of IQ and GPA, females earn more on average than males.
iii. For a fixed value of IQ and GPA, males earn more on average than females provided that the GPA is high enough.
iv. For a fixed value of IQ and GPA, females earn more on average than males provided that the GPA is high enough.

iii. is correct.
Write the equation out:

$$Salary = 50 + 20X_1 + 0.07X_2 + 35X_3 + 0.01X_4 - 10X_5$$

If two people have the same IQ and GPA, then terms $X_{1,}$ X_2 and X_4 can be replaced by a constant C, leaving only X_3 and X_5 as the variables which change the value of the salary. Now salary becomes:

$$Salary = C + 35X_3 - 10X_5$$

Assume $X_5 = Gender * GPA$. If a person is male then $X_3 = 0$ and salary is just:

$$Salary_{male} = C$$

However, if the person if female then $X_3 = 1$ and salary is:

$$Salary_{female} = C + 35 - 10 * GPA$$

Here we can see that the salary for females is higher for GPA's under 3.5; and the salary for males is higher for GPA's over 3.5.

(b) Predict the salary of a female with IQ of 110 and a GPA of 4.0.

Assuming $X_5 = IQ * GPA$:

$$Salary = 50 + 20X_1 + 0.07X_2 + 35X_3 + 0.01X_4 - 10X_{5=}$$
$$= 50 + 20GPA + 0.07IQ + 35 * Female + 0.01 * IQ * GPA - 10 * Female * GPA =$$
$$= 50 + 20 * 4.0 + 0.07 * 110 + 35 * 1 + 0.01 * 110 * 4.0 - 10 * 1 * 4.0 =$$
$$= 50 + 80 + 7.7 + 35 + 4.4 - 40 = \mathbf{137.1}$$

(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.

To conclude whether there is little evidence of an interaction effect we need to calculate the p-value for the null hypothesis. In this case we do not have the data to do so.

I will try to provide a more simplistic approach in determining whether the interaction effect exists. Let's see how much the value of the Salary is impacted as a percentage when we include or exclude $X_{4.}$

The minimum value X_4 can achieve is 0, and it occurs either when the IQ is 0 or the GPA is 0 or both. In either case adding the value $\beta_4 X_4$ will not affect the value of the salary, therefore the impact of X_4 expressed as a percentage is 0.

The maximum value X_4 can achieve is 800, and it occurs when the IQ is 200 (assuming this is the maximum one can achieve) and the GPA is 4.0. In this case $\beta_4 X_4 = 8$. The Salary then becomes:

$$Salary_{male} = 50 + 80 + 1.4 + 0 + 8 + 0 = 139.4$$

$$Salary_{female} = 50 + 80 + 1.4 + 35 + 8 - 40 = 134.4$$

In both cases removing X_4 will affect the Salary by 5.74% and 5.95% respectively. This is quite a noticeable bump (especially in someone's salary), and I would say that this is some evidence (although not so strong) of an interaction effect.

4. I collect a set of data (n = 100 observations) containing a single predictor and a quantitative response. I then fit a linear regression model to the data, as well as a separate cubic regression, i.e. $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \epsilon$

(a) Suppose that the true relationship between X and Y is linear, i.e. Y = β0 + β1 X + . Consider the training residual sum of squares (RSS) for the linear regression, and also the training RSS for the cubic regression. Would we expect one to be lower than the other, would we expect them to be the same, or is there not enough information to tell? Justify your answer.

For a training data set (regardless of the true relationship), we expect the more flexible model to have a lower RSS, since a more flexible model can always fit training data better than an inflexible model. Since the cubic regression is more flexible we would expect its RSS to be smaller.

(b) Answer (a) using test rather than training RSS.

We expect RSS for test data to be lower for the model which better reflects the true relationship. Since in this case the true relationship is linear, we expect the linear regression model to have a lower RSS.

(c) Suppose that the true relationship between X and Y is not linear, but we don't know how far it is from linear. Consider the training RSS for the linear regression, and also the training RSS for the cubic regression. Would we expect one to be lower than the other, would we expect them to be the same, or is there not enough information to tell? Justify your answer.

Just as in part (a), for a training data set (regardless of the true relationship), we expect the more flexible model to have a lower RSS, because a more flexible model can always fit training data better than an inflexible model. Since the cubic regression is more flexible we would expect its RSS to be smaller.

(d) Answer (c) using test rather than training RSS.

Just as in part (b), we expect RSS for test data to be lower for the model which better reflects the true relationship. Since now we do not know how far away a relationship is from being linear, we do not have enough information to say whether a linear or a cubic regression model would have lower RSS for test data.

5. Consider the fitted values that result from performing linear regression without an intercept. In this setting, the ith fitted value takes the form:

$$\hat{y}_i = x_i\hat{\beta}$$

where:

$$\hat{\beta} = \left(\sum_{i=1}^{n} x_i y_i\right) / \left(\sum_{i'=1}^{n} x_{i'}^2\right)$$

Show that we can write

$$\hat{y}_i = \sum_{i'=1}^{n} a_{i'} y_{i'}$$

Using algebraic manipulations:

$$\hat{y_i} = x_j\hat{\beta} = x_j * \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i'=1}^{n} x_{i'}^2} = x_j * \frac{x_1y_1 + x_2y_2 + \cdots + x_ny_n}{\sum_{i'=1}^{n} x_{i'}^2}$$

$$= \frac{x_ix_1}{\sum_{i'=1}^{n} x_{i'}^2}y_1 + \frac{x_ix_2}{\sum_{i'=1}^{n} x_{i'}^2}y_2 + \cdots + \frac{x_ix_n}{\sum_{i'=1}^{n} x_{i'}^2}y_n = \sum_{i'=1}^{n} a_{i'}y_{i'}$$

Where:

$$a_{i'} = \frac{x_ix_{i'}}{\sum_{j=1}^{n} x_j^2}$$

What is $a_{i'}$?

Note: We interpret this result by saying that the fitted values from linear regression are linear combinations of the response variables.

$a_{i'}$ is the projection of our x_i on our normalized bases of our training predictor variables.

When we expand the equation for $\hat{y_i}$ we get:

$$\hat{y_i} = a_1y_1 + a_2y_2 + \cdots + a_ny_n$$

Which is a linear combination of the response variables, where coefficients are a_1, a_2 … a_n

6. Using equation (3.4) – shown below, argue that in the case of simple linear regression, the least squares line always passes through the point $(\overline{x}, \overline{y})$.

$$\widehat{\beta_1} = (\textstyle\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}))/(\sum_{i=1}^{n}(x_i - \bar{x})^2) \qquad (3.4.a)$$

$$\widehat{\beta_0} = \bar{y} - \widehat{\beta_1}\bar{x} \qquad (3.5.b)$$

In simple linear regression:

$$\hat{y} = \widehat{\beta_0} + \widehat{\beta_1}x$$

We have to show that when $\hat{y} = \bar{y}$ then $x = \bar{x}$

Suppose:

$$\hat{y} = \bar{y}$$

Substitute for $\hat{y}$:

$$\widehat{\beta_0} + \widehat{\beta_1}x = \bar{y}$$

Now substitute for $\widehat{\beta_0}$:

$$\bar{y} - \widehat{\beta_1}\bar{x} + \widehat{\beta_1}x = \bar{y}$$

With some simple algebraic manipulations we arrive at:

$$\widehat{\beta_1}(x - \bar{x}) = \bar{y} - \bar{y}$$

$$\widehat{\beta_1}(x - \bar{x}) = 0$$

$$x - \bar{x} = 0$$

$$x = \bar{x}$$

q.e.d.

<u>Applied Questions</u>

1. This question involves the use of simple linear regression on the Auto data set.

(a) Perform a simple linear regression with mpg as the response and horsepower as the predictor and answer the following questions:

	coef	std err	t	P>|t|	[0.025	0.975]
const	39.9359	0.717	55.660	0.000	38.525	41.347
horsepower	-0.1578	0.006	-24.489	0.000	-0.171	-0.145

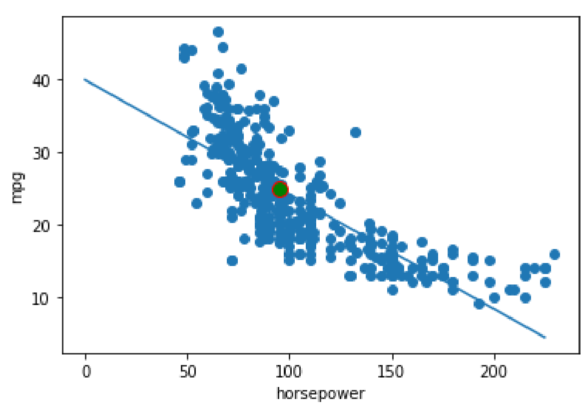
i. Is there a relationship between the predictor and the response?
Yes, there <u>seems to a relationship between mpg and horsepower</u>. The p value is very small, indicating that the plausibility of the null hypothesis is very low i.e. <u>we reject the null hypothesis that there is no relationship</u> between mpg and horsepower.

ii. How strong is the relationship between the predictor and the response?
The absolute value of the relationship is 0.1578

iii. Is the relationship between the predictor and the response positive or negative?
The relationship is negative (slope is negative). The mpg decreases as horsepower increases.

iv. What is the predicted mpg associated with a horsepower of 95?
The predicted mpg is 24.94061135

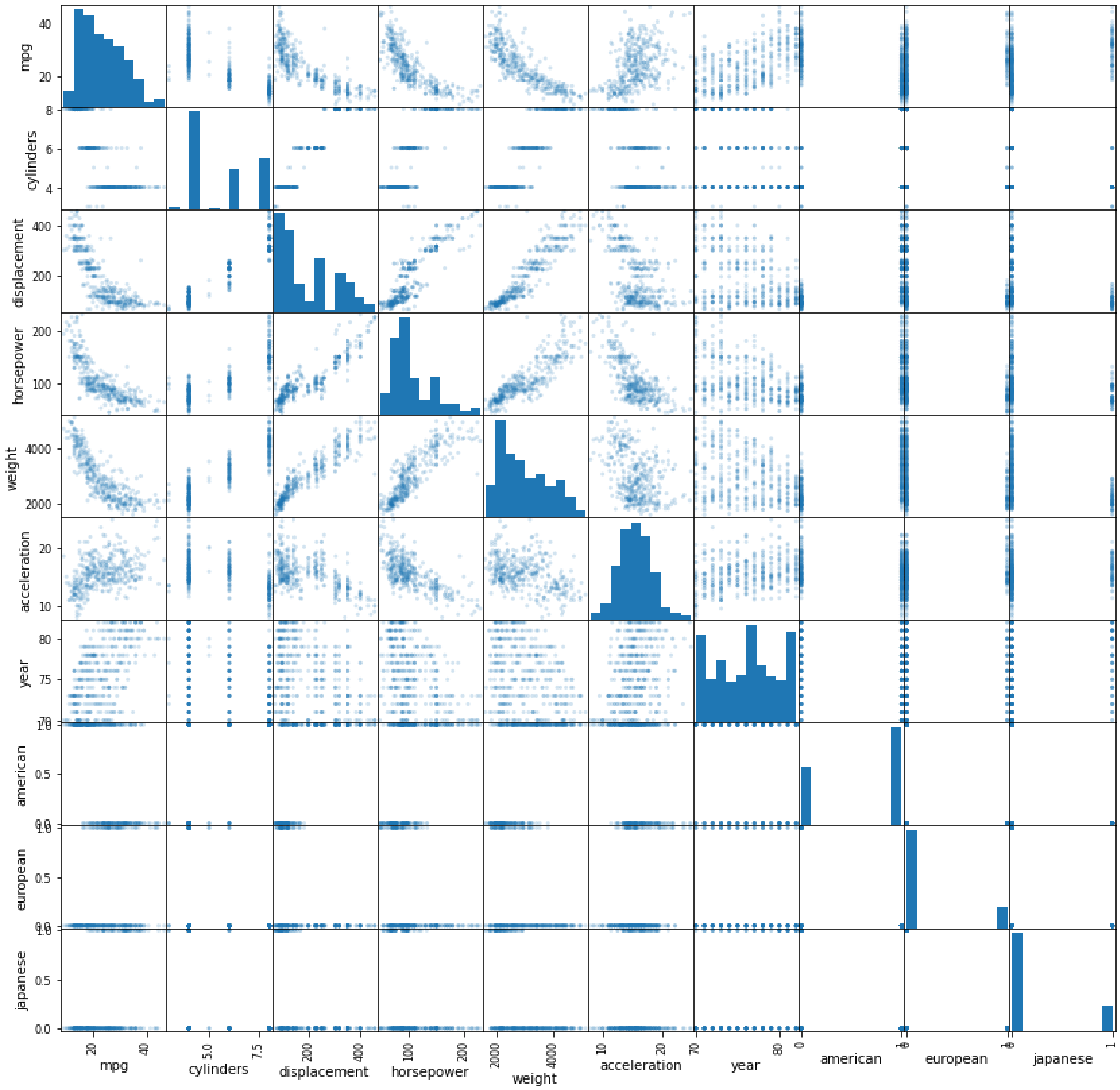
(b) Plot the response and the predictor along with the predicted line.



2. This question involves the use of multiple linear regression on the Auto data set.

I created 3 dummy variables for origin: 'american' which is 1 when origin = 1, 0 otherwise; 'european' which is 1 when origin = 2, 0 otherwise; and 'japanese' which is 1 when origin = 3, 0 otherwise.

(a) Produce a scatterplot matrix which includes all of the variables in the data set.



(b) Compute the matrix of correlations between the variables. You will need to exclude the name variable which is qualitative.

	mpg	cylinders	displacement	horsepower	weight	acceleration	year	american	european	japanese
mpg	1.000000	-0.777618	-0.805127	-0.778427	-0.832244	0.423329	0.580541	-0.565161	0.244313	0.451454
cylinders	-0.777618	1.000000	0.950823	0.842983	0.897527	-0.504683	-0.345647	0.610494	-0.352324	-0.404209
displacement	-0.805127	0.950823	1.000000	0.897257	0.932994	-0.543800	-0.369855	0.655936	-0.371633	-0.440825
horsepower	-0.778427	0.842983	0.897257	1.000000	0.864538	-0.689196	-0.416361	0.489625	-0.284948	-0.321936
weight	-0.832244	0.897527	0.932994	0.864538	1.000000	-0.416839	-0.309120	0.600978	-0.293841	-0.447929
acceleration	0.423329	-0.504683	-0.543800	-0.689196	-0.416839	1.000000	0.290316	-0.258224	0.208298	0.115020
year	0.580541	-0.345647	-0.369855	-0.416361	-0.309120	0.290316	1.000000	-0.136065	-0.037745	0.199841
american	-0.565161	0.610494	0.655936	0.489625	0.600978	-0.258224	-0.136065	1.000000	-0.591434	-0.648583
european	0.244313	-0.352324	-0.371633	-0.284948	-0.293841	0.208298	-0.037745	-0.591434	1.000000	-0.230157
japanese	0.451454	-0.404209	-0.440825	-0.321936	-0.447929	0.115020	0.199841	-0.648583	-0.230157	1.000000

(c) Perform a multiple linear regression with mpg as the response and all other variables except name as the predictors. Examine the results, and comment on the output. For instance:

i. Is there a relationship between the predictors and the response?

Dep. Variable:	mpg	**R-squared:**	0.824
Model:	OLS	**Adj. R-squared:**	0.821
Method:	Least Squares	**F-statistic:**	224.5
Date:	Mon, 15 Nov 2021	**Prob (F-statistic):**	1.79e-139
Time:	00:46:43	**Log-Likelihood:**	-1020.5
No. Observations:	392	**AIC:**	2059.
Df Residuals:	383	**BIC:**	2095.
Df Model:	8		
Covariance Type:	nonrobust		

The F-statistic is large and the p-value for the F-statistic is small, meaning that we reject the null hypothesis that all coefficients are 0. We there say that there exists a relationship between some of the predictors and the response.

ii. Which predictors appear to have a statistically significant relationship to the response?

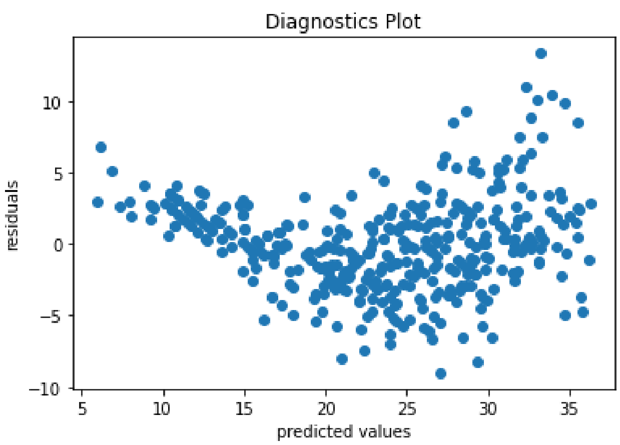
	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
const	-12.0951	3.482	-3.474	0.001	-18.941	-5.250
cylinders	-0.4897	0.321	-1.524	0.128	-1.121	0.142
displacement	0.0240	0.008	3.133	0.002	0.009	0.039
horsepower	-0.0182	0.014	-1.326	0.185	-0.045	0.009
weight	-0.0067	0.001	-10.243	0.000	-0.008	-0.005
acceleration	0.0791	0.098	0.805	0.421	-0.114	0.272
year	0.7770	0.052	15.005	0.000	0.675	0.879
american	-5.8595	1.227	-4.775	0.000	-8.272	-3.447
european	-3.2295	1.156	-2.794	0.005	-5.502	-0.957
japanese	-3.0062	1.231	-2.443	0.015	-5.426	-0.587

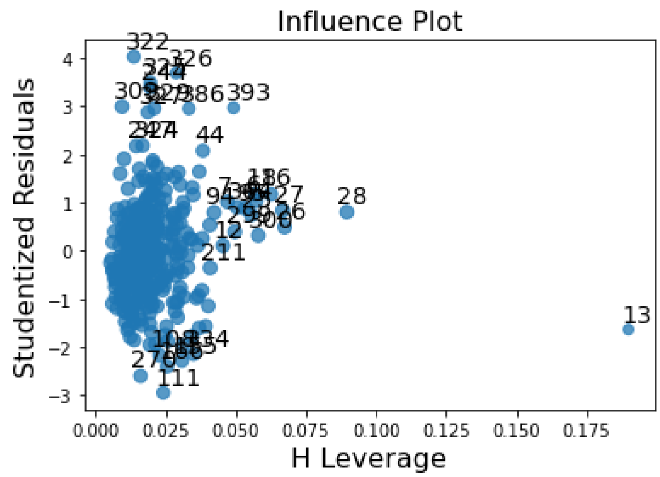
The predictors with a statistically significant (p-value < 0.01) relationship to the response are constant term, displacement, weight, year and the three dummy vairables. In other words, for these predictors we reject the null hypothesis.

iii. What does the coefficient for the year variable suggest?

Since the coefficient is positive, it suggests that newer cars have a better mpg ratio, meaning that they are more fuel efficient.

(d) Produce diagnostic plots of the linear regression fit. Comment on any problems you see with the fit. Do the residual plots suggest any unusually large outliers? Does the leverage plot identify any observations with unusually high leverage?





The residuals exhibit a clear U-shape, which provides a strong indication of non-linearity in the data. The variance in the error terms also seems to increase as the predicted value increases.

From the influence plot, we can see that quite a few points have a studentized residual over 3 in absolute value, which makes them outliers, with largest one being data point 322.

Two observations seem to stand out as having large leverage. Namely point 28 and point 13.

(e) i) Fit linear regression models with predictors and interaction terms. Do any interactions appear to be statistically significant? NOTE: there are two part e's for this question!!!

Initially I tried fitting as many predictors and interactions to obtain the following model:

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-11.8552	49.793	-0.238	0.812	-109.767	86.057
cylinders	5.3782	8.286	0.649	0.517	-10.915	21.672
displacement	-0.3237	0.182	-1.780	0.076	-0.681	0.034
horsepower	0.3830	0.328	1.169	0.243	-0.261	1.027
weight	0.0085	0.018	0.482	0.630	-0.026	0.043
acceleration	-4.8070	2.124	-2.263	0.024	-8.983	-0.631
year	1.0498	0.585	1.795	0.073	-0.100	2.200
american	-5.165e-11	1.92e-10	-0.268	0.788	-4.3e-10	3.27e-10
european	1.452e-12	8.86e-12	0.164	0.870	-1.6e-11	1.89e-11
japanese	-1.161e-12	4.32e-12	-0.269	0.788	-9.65e-12	7.33e-12
cylinders:displacement	-0.0102	0.005	-2.179	0.030	-0.019	-0.001
cylinders:horsepower	0.0320	0.024	1.345	0.179	-0.015	0.079
cylinders:weight	6.618e-05	0.001	0.080	0.936	-0.002	0.002
cylinders:acceleration	0.2946	0.168	1.751	0.081	-0.036	0.625
cylinders:year	-0.1438	0.095	-1.512	0.131	-0.331	0.043
displacement:horsepower	-0.0002	0.000	-0.667	0.505	-0.001	0.000
displacement:weight	3.476e-05	1.37e-05	2.539	0.012	7.84e-06	6.17e-05
displacement:acceleration	-0.0064	0.003	-2.008	0.045	-0.013	-0.000
displacement:year	0.0050	0.002	2.236	0.026	0.001	0.009
horsepower:weight	-3.616e-05	2.84e-05	-1.273	0.204	-9.2e-05	1.97e-05
horsepower:acceleration	-0.0057	0.004	-1.580	0.115	-0.013	0.001
horsepower:year	-0.0051	0.004	-1.301	0.194	-0.013	0.003
weight:acceleration	0.0002	0.000	1.075	0.283	-0.000	0.001
weight:year	-0.0003	0.000	-1.273	0.204	-0.001	0.000
acceleration:year	0.0540	0.025	2.118	0.035	0.004	0.104

Assuming a significance level of 0.05, we have that acceleration, cylinders*displacement, displacement*acceleration, displacement*year and acceleration*year are of statistical significance.

After thinking about the problem for a bit, to me it would make sense that displacement, acceleration, horsepower and year as well as their interactions (excluding those with year) would have the greatest statistical significance. Here are the results of fitting such a model:

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	10.4681	8.510	1.230	0.219	-6.264	27.200
displacement	0.0309	0.041	0.753	0.452	-0.050	0.112
acceleration	-0.4637	0.366	-1.267	0.206	-1.183	0.256
horsepower	-0.0375	0.075	-0.503	0.615	-0.184	0.109
weight	-0.0222	0.005	-4.361	0.000	-0.032	-0.012
year	0.7824	0.045	17.255	0.000	0.693	0.872
displacement:acceleration	-0.0042	0.002	-2.207	0.028	-0.008	-0.000
displacement:horsepower	-0.0002	0.000	-1.312	0.190	-0.001	0.000
displacement:weight	1.556e-05	5.56e-06	2.796	0.005	4.62e-06	2.65e-05
acceleration:horsepower	-0.0077	0.004	-2.184	0.030	-0.015	-0.001
acceleration:weight	0.0006	0.000	2.836	0.005	0.000	0.001
horsepower:weight	4.391e-05	1.68e-05	2.611	0.009	1.08e-05	7.7e-05

Here we see (assuming statistical significance of 0.05) that weight, year, and all interactions except displacement*horsepower.

Going even further I removed this singular interaction with no statistical significance to obtain my final model:

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	7.9109	8.292	0.954	0.341	-8.392	24.214
displacement	-0.0060	0.030	-0.201	0.841	-0.065	0.053
acceleration	-0.3762	0.360	-1.045	0.297	-1.084	0.332
horsepower	-0.0742	0.069	-1.071	0.285	-0.210	0.062
weight	-0.0168	0.003	-5.614	0.000	-0.023	-0.011
year	0.7716	0.045	17.286	0.000	0.684	0.859
displacement:acceleration	-0.0028	0.002	-1.777	0.076	-0.006	0.000
displacement:weight	1.301e-05	5.22e-06	2.492	0.013	2.75e-06	2.33e-05
acceleration:horsepower	-0.0056	0.003	-1.779	0.076	-0.012	0.001
acceleration:weight	0.0004	0.000	2.615	0.009	0.000	0.001
horsepower:weight	2.95e-05	1.27e-05	2.314	0.021	4.44e-06	5.46e-05

Here the end result is that the same predictors are of statistical significance, however interaction between displacement and acceleration and that of acceleration and horsepower lost its statistical significance.

e) ii) Fit linear regression models with only interaction terms. Do any interactions appear to be statistically significant?

Initially I tried fitting as many interactions as possible to obtain the following model:

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	7.0312	2.232	3.150	0.002	2.642	11.420
cylinders:displacement	-0.0035	0.005	-0.704	0.482	-0.013	0.006
cylinders:horsepower	-0.0078	0.021	-0.366	0.715	-0.049	0.034
cylinders:weight	0.0001	0.001	0.160	0.873	-0.002	0.002
cylinders:acceleration	-0.4283	0.135	-3.179	0.002	-0.693	-0.163
cylinders:year	0.1100	0.031	3.492	0.001	0.048	0.172
displacement:horsepower	0.0006	0.000	2.384	0.018	9.98e-05	0.001
displacement:weight	1.165e-05	1.46e-05	0.800	0.424	-1.7e-05	4.03e-05
displacement:acceleration	0.0118	0.002	5.413	0.000	0.008	0.016
displacement:year	-0.0038	0.001	-6.143	0.000	-0.005	-0.003
horsepower:weight	-2.471e-05	2.72e-05	-0.908	0.364	-7.82e-05	2.88e-05
horsepower:acceleration	-0.0030	0.004	-0.783	0.434	-0.010	0.004
horsepower:year	-0.0002	0.001	-0.158	0.874	-0.002	0.002
weight:acceleration	-0.0008	0.000	-3.874	0.000	-0.001	-0.000
weight:year	9.885e-05	7.75e-05	1.275	0.203	-5.36e-05	0.000
acceleration:year	0.0333	0.004	9.349	0.000	0.026	0.040

Assuming again a significance level of 0.05, we have that the intercept, cylinders*acceleration, cylinders*year, displacement*horsepower, displacement*acceleration, displacement*year, weight*acceleration and acceleration*year are of statistical significance.

Going further I removed all interactions with no statistical significance to obtain the following model:

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	3.9374	2.071	1.902	0.058	-0.134	8.008
cylinders:acceleration	-0.3958	0.071	-5.541	0.000	-0.536	-0.255
cylinders:year	0.0825	0.014	5.790	0.000	0.054	0.110
displacement:horsepower	9.658e-05	4.05e-05	2.386	0.017	1.7e-05	0.000
displacement:acceleration	0.0057	0.002	2.870	0.004	0.002	0.010
displacement:year	-0.0015	0.000	-3.394	0.001	-0.002	-0.001
weight:acceleration	-0.0004	4.21e-05	-10.266	0.000	-0.001	-0.000
acceleration:year	0.0342	0.002	17.510	0.000	0.030	0.038

Here we see (assuming statistical significance of 0.05) that only the interaction between cylinders and acceleration is not of statistical significance.

For the final model I have therefore removed this interaction to obtain:

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	8.9678	1.931	4.643	0.000	5.170	12.765
cylinders:year	0.0073	0.004	1.620	0.106	-0.002	0.016
displacement:horsepower	-4.669e-05	3.23e-05	-1.445	0.149	-0.000	1.68e-05
displacement:acceleration	-0.0043	0.001	-5.000	0.000	-0.006	-0.003
displacement:year	0.0007	0.000	3.359	0.001	0.000	0.001
weight:acceleration	-0.0004	4.25e-05	-8.869	0.000	-0.000	-0.000
acceleration:year	0.0271	0.002	17.696	0.000	0.024	0.030

Here the end result is that all interactions, except those between cylinders and year and displacement and horsepower are of statistical significance.

(f) Try a few different transformations of the variables, such as log(X), $\sqrt{X}$, $X2$. Comment on your findings.

OLS Regression Results

Dep. Variable:	mpg	**R-squared:**	0.835
Model:	OLS	**Adj. R-squared:**	0.832
Method:	Least Squares	**F-statistic:**	242.4
Date:	Mon, 15 Nov 2021	**Prob (F-statistic):**	9.23e-145
Time:	01:50:00	**Log-Likelihood:**	-1008.0
No. Observations:	392	**AIC:**	2034.
Df Residuals:	383	**BIC:**	2070.
Df Model:	8		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-36.1473	6.940	-5.209	0.000	-49.792	-22.502
cylinders	-0.3383	1.535	-0.220	0.826	-3.357	2.680
displacement	0.3239	0.236	1.371	0.171	-0.141	0.788
horsepower	-0.7488	0.308	-2.433	0.015	-1.354	-0.144
weight	-0.6522	0.081	-8.066	0.000	-0.811	-0.493
acceleration	-0.7290	0.834	-0.874	0.383	-2.370	0.912
year	13.1227	0.879	14.921	0.000	11.393	14.852
american	-13.5507	2.366	-5.727	0.000	-18.203	-8.898
european	-11.4009	2.275	-5.011	0.000	-15.874	-6.928
japanese	-11.1957	2.365	-4.733	0.000	-15.846	-6.545

Omnibus:	32.960	**Durbin-Watson:**	1.303

X^2:

From the top table we see that R-squared is decently high at 0.803, meaning that our regression explains a fair amount of variance in the data.

The F-statistic is high and the probability (p-value) for that F-statistic is low which indicates that the null hypothesis (i.e. that there is no relationship) is rejected. In other words there is a relationship between at least one predictor and the response.

From the bottom table we see that (assuming a significance level of 0.05), all predictors are of statistical significance except for the constant term and the horsepower squared.

$\sqrt{X}$:

OLS Regression Results

Dep. Variable:	mpg	**R-squared:**	0.835
Model:	OLS	**Adj. R-squared:**	0.832
Method:	Least Squares	**F-statistic:**	242.4
Date:	Mon, 15 Nov 2021	**Prob (F-statistic):**	9.23e-145
Time:	01:50:00	**Log-Likelihood:**	-1008.0
No. Observations:	392	**AIC:**	2034.
Df Residuals:	383	**BIC:**	2070.
Df Model:	8		
Covariance Type:	nonrobust		

From the top table we see that the R-squared is even higher than for X^2, implying that this model explains even more of the variance in the data.

The F-statistic is high and the probability (p-value) for that F-statistic is low which indicates that the null hypothesis is rejected. In other words there is a relationship between at least one predictor and the response.

	coef	std err	t	P>|t|	[0.025	0.975]
const	-36.1473	6.940	-5.209	0.000	-49.792	-22.502
cylinders	-0.3383	1.535	-0.220	0.826	-3.357	2.680
displacement	0.3239	0.236	1.371	0.171	-0.141	0.788
horsepower	-0.7488	0.308	-2.433	0.015	-1.354	-0.144
weight	-0.6522	0.081	-8.066	0.000	-0.811	-0.493
acceleration	-0.7290	0.834	-0.874	0.383	-2.370	0.912
year	13.1227	0.879	14.921	0.000	11.393	14.852
american	-13.5507	2.366	-5.727	0.000	-18.203	-8.898
european	-11.4009	2.275	-5.011	0.000	-15.874	-6.928
japanese	-11.1957	2.365	-4.733	0.000	-15.846	-6.545

Omnibus:	32.960	Durbin-Watson:	1.303

From the bottom table we see that (assuming a significance level of 0.05), all predictors are of statistical significance, except sqrt(cylinders), sqrt(displacement) and sqrt(acceleration).

log(X):

Note: dummy variables had to be excluded (log(0) = NaN)

OLS Regression Results

Dep. Variable:	mpg	R-squared:	0.844
Model:	OLS	Adj. R-squared:	0.842
Method:	Least Squares	F-statistic:	348.1
Date:	Mon, 15 Nov 2021	Prob (F-statistic):	4.06e-152
Time:	01:51:36	Log-Likelihood:	-996.58
No. Observations:	392	AIC:	2007.
Df Residuals:	385	BIC:	2035.
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-62.4130	17.650	-3.536	0.000	-97.116	-27.710
cylinders	2.7502	1.626	1.691	0.092	-0.447	5.947
displacement	-3.4063	1.355	-2.513	0.012	-6.071	-0.741
horsepower	-6.3856	1.563	-4.085	0.000	-9.459	-3.312
weight	-11.9049	2.240	-5.316	0.000	-16.308	-7.501
acceleration	-5.3256	1.622	-3.283	0.001	-8.515	-2.137
year	54.8253	3.595	15.250	0.000	47.757	61.894

Omnibus:	48.928	Durbin-Watson:	1.371
Prob(Omnibus):	0.000	Jarque-Bera (JB):	99.706
Skew:	0.688	Prob(JB):	2.23e-22
Kurtosis:	5.053	Cond. No.	1.36e+03

From the top table we see that the R-squared is highest of all models, implying that this model explains the variance in the data the best.

The F-statistic is high and the probability (p-value) for that F-statistic is low which indicates that the null hypothesis is rejected. In other words there is a relationship between at least one predictor and the response.

From the bottom table we see that (assuming a significance level of 0.05), all predictors are of statistical significance, except for the number of cylinders.

3. This question should be answered using the Carseats data set.

(a) Fit a multiple regression model to predict Sales using Price, Urban, and US.

OLS Regression Results

Dep. Variable:	Sales	**R-squared:**	0.239
Model:	OLS	**Adj. R-squared:**	0.234
Method:	Least Squares	**F-statistic:**	41.52
Date:	Sat, 13 Nov 2021	**Prob (F-statistic):**	2.39e-23
Time:	02:29:22	**Log-Likelihood:**	-927.66
No. Observations:	400	**AIC:**	1863.
Df Residuals:	396	**BIC:**	1879.
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	13.0435	0.651	20.036	0.000	11.764	14.323
Urban[T.1]	-0.0219	0.272	-0.081	0.936	-0.556	0.512
US[T.1]	1.2006	0.259	4.635	0.000	0.691	1.710
Price	-0.0545	0.005	-10.389	0.000	-0.065	-0.044

Omnibus:	0.676	**Durbin-Watson:**	1.912
Prob(Omnibus):	0.713	**Jarque-Bera (JB):**	0.758
Skew:	0.093	**Prob(JB):**	0.684
Kurtosis:	2.897	**Cond. No.**	628.

(b) Provide an interpretation of each coefficient in the model. Be careful—some of the variables in the model are qualitative!

The intercept coefficient is 13.0435 implying that when a car seat price is 0, the car seat is not urban or American it the sales are 13.0435.

The Urban coefficient is -0.0219, which implies that when a car seat is urban its sales are 0.0219 units lower.

The US coefficient is 1.2006, which implies that when a car seat is from the US its sales are 1.2006 units higher.

The price coefficient is -0.0545, which implies that the car seat's sales decrease by 0.0545 units for every unit increase in price.

(c) Write out the model in equation form, being careful to handle the qualitative variables properly.

$$\hat{y} = \widehat{\beta_0} + \widehat{\beta_1}X_1 + \widehat{\beta_2}X_2 + \widehat{\beta_3}X_3$$

Where $\widehat{\beta_0}$ is the intercept term, X_1 is the price and $\widehat{\beta_1}$ is its corresponding constant; X_2 is a dummy variable indicating whether a car seat is urban (1 if it is, 0 if not) price and $\widehat{\beta_2}$ is its corresponding constant; X_3 is a dummy variable indicating whether a car seat is from the US (1 if it is, 0 if not) price and $\widehat{\beta_3}$ is its corresponding constant.

(d) For which of the predictors can you reject the null hypothesis H0 : βj = 0?

The only predictor with non-zero and quite high p-value is the Urban predictor. This is the only predictor for which we should not reject the null hypothesis.

For all other predictors we may reject the null hypothesis.

(e) On the basis of your response to the previous question, fit a smaller model that only uses the predictors for which there is evidence of association with the outcome.

Here we remove the Urban predictor. Here is the outcome:

OLS Regression Results

Dep. Variable:	Sales	**R-squared:**	0.239
Model:	OLS	**Adj. R-squared:**	0.235
Method:	Least Squares	**F-statistic:**	62.43
Date:	Sat, 13 Nov 2021	**Prob (F-statistic):**	2.66e-24
Time:	02:54:12	**Log-Likelihood:**	-927.66
No. Observations:	400	**AIC:**	1861.
Df Residuals:	397	**BIC:**	1873.
Df Model:	2		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Intercept	13.0308	0.631	20.652	0.000	11.790	14.271
US[T.1]	1.1996	0.258	4.641	0.000	0.692	1.708
Price	-0.0545	0.005	-10.416	0.000	-0.065	-0.044

Omnibus:	0.666	**Durbin-Watson:**	1.912
Prob(Omnibus):	0.717	**Jarque-Bera (JB):**	0.749
Skew:	0.092	**Prob(JB):**	0.688
Kurtosis:	2.895	**Cond. No.**	607.

(f) How well do the models in (a) and (e) fit the data?

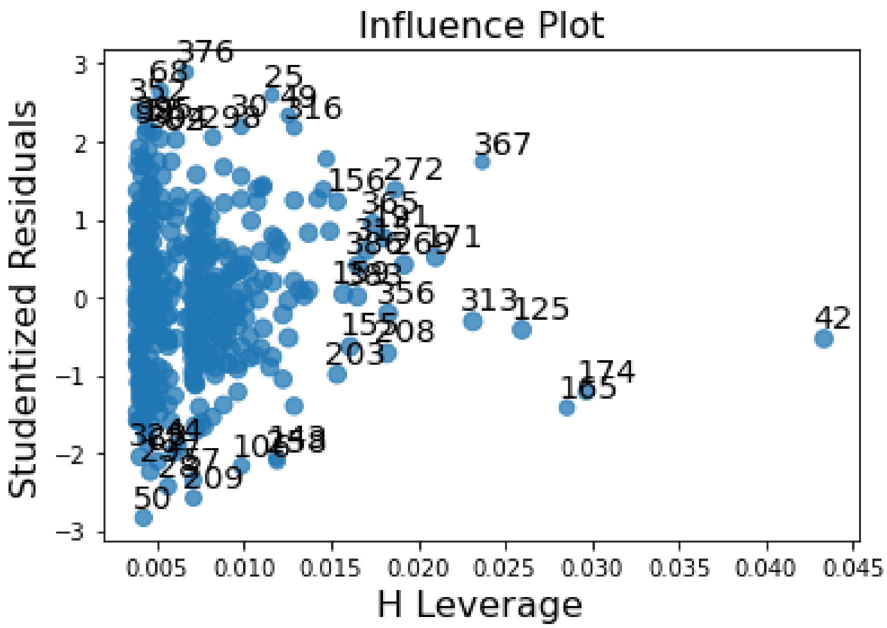
The R^2 both models is 0.239 which implies that neither model explains the variance in the data very well.

(g) Using the model from (e), obtain 95% confidence intervals for the coefficient(s).

These are given above in the last column:

Intercept: (11.790, 14.271)
US: (0.692, 1.708)
Price: (-0.065, -0.044)

(h) Is there evidence of outliers or high leverage observations in the model from (e)?



From the influence plot shown, we have no data points with studentized residuals greater than 3 in absolute value (meaning there are not outliers). However we see quite a few points with high leverage. Say our cutoff for high leverage is a leverage statistic above 0.025. This then implies that the high leverage observations are 313, 367, 125, 165, 174 and 42

4. This problem involves the Boston data set, which we saw in the previous HW. 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.

(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.

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	4.4537	0.417	10.675	0.000	3.634	5.273
zn	-0.0739	0.016	-4.594	0.000	-0.106	-0.042

p-value<0.01 -> zn is statistically significant

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-2.0637	0.667	-3.093	0.002	-3.375	-0.753
indus	0.5098	0.051	9.991	0.000	0.410	0.610

p-value<0.01 -> indus is statistically significant

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	3.7444	0.396	9.453	0.000	2.966	4.523
chas	-1.8928	1.506	-1.257	0.209	-4.852	1.066

p-value>0.01 -> chas is not statistically

significant

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-13.7199	1.699	-8.073	0.000	-17.059	-10.381
nox	31.2485	2.999	10.419	0.000	25.356	37.141

p-value<0.01 -> nox is statistically significant

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	20.4818	3.364	6.088	0.000	13.872	27.092
rm	-2.6841	0.532	-5.045	0.000	-3.729	-1.639

p-value<0.01 -> rm is statistically significant

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-3.7779	0.944	-4.002	0.000	-5.633	-1.923
age	0.1078	0.013	8.463	0.000	0.083	0.133

p-value<0.01 -> age is statistically significant

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	9.4993	0.730	13.006	0.000	8.064	10.934
dis	-1.5509	0.168	-9.213	0.000	-1.882	-1.220

p-value < 0.01 -> dis is statistically significant

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-2.2872	0.443	-5.157	0.000	-3.158	-1.416
rad	0.6179	0.034	17.998	0.000	0.550	0.685

p-value < 0.01 -> rad is statistically significant

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-8.5284	0.816	-10.454	0.000	-10.131	-6.926
tax	0.0297	0.002	16.099	0.000	0.026	0.033

p-value < 0.01 -> tax is statistically significant

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-17.6469	3.147	-5.607	0.000	-23.830	-11.464
ptratio	1.1520	0.169	6.801	0.000	0.819	1.485

p-value < 0.01 -> ptratio is statistically significant

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-3.3305	0.694	-4.801	0.000	-4.694	-1.968
lstat	0.5488	0.048	11.491	0.000	0.455	0.643

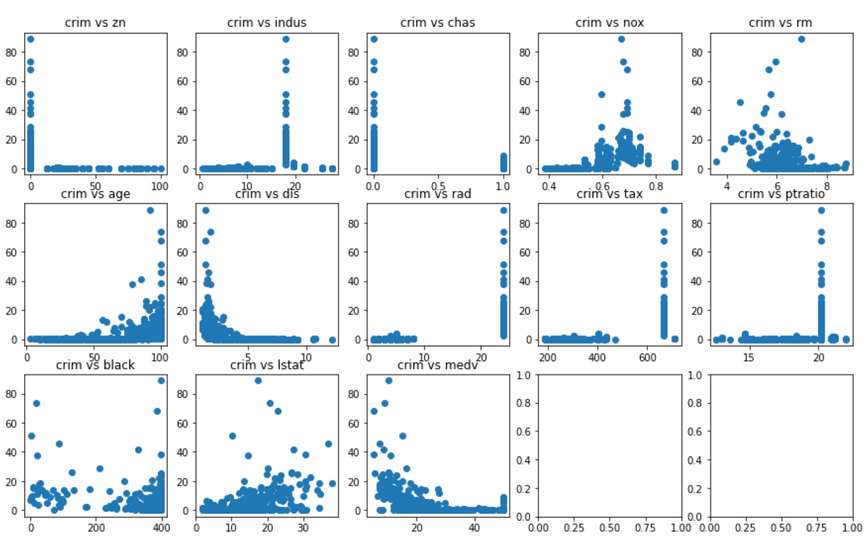
p-value <0.01 -> medv is statistically significant

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	11.7965	0.934	12.628	0.000	9.961	13.632
medv	-0.3632	0.038	-9.460	0.000	-0.439	-0.288

p-value < 0.01 -> lstat is statistically significant

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	16.5535	1.426	11.609	0.000	13.752	19.355
black	-0.0363	0.004	-9.367	0.000	-0.044	-0.029

p-value < 0.01 -> black is statistically significant



For all predictors which we deemed statistically significant, we see that as the x-value goes from zero to non-zero we see a change in the response value. For the chas predictor which we deemed statistically insignificant, when x is zero, the value for the predictor takes on a wide range of values, however when it takes the value of 1 the response goes to 0, which implies that the chas coefficient is likely zero.

(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 $H0 : \beta j = 0$?

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	17.0332	7.235	2.354	0.019	2.818	31.248
zn	0.0449	0.019	2.394	0.017	0.008	0.082
indus	-0.0639	0.083	-0.766	0.444	-0.228	0.100
chas	-0.7491	1.180	-0.635	0.526	-3.068	1.570
nox	-10.3135	5.276	-1.955	0.051	-20.679	0.052
rm	0.4301	0.613	0.702	0.483	-0.774	1.634
age	0.0015	0.018	0.081	0.935	-0.034	0.037
dis	-0.9872	0.282	-3.503	0.001	-1.541	-0.433
rad	0.5882	0.088	6.680	0.000	0.415	0.761
tax	-0.0038	0.005	-0.733	0.464	-0.014	0.006
ptratio	-0.2711	0.186	-1.454	0.147	-0.637	0.095
black	-0.0075	0.004	-2.052	0.041	-0.015	-0.000
lstat	0.1262	0.076	1.667	0.096	-0.023	0.275
medv	-0.1989	0.061	-3.287	0.001	-0.318	-0.080

Rejecting null hypothesis for p-values under 0.01:

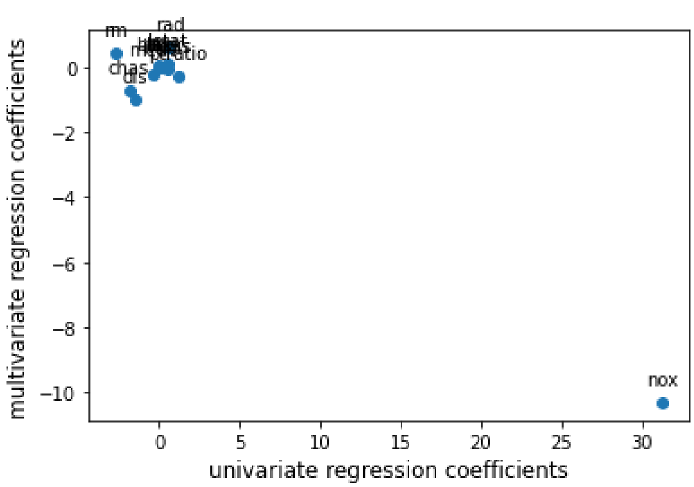
dis -> reject null

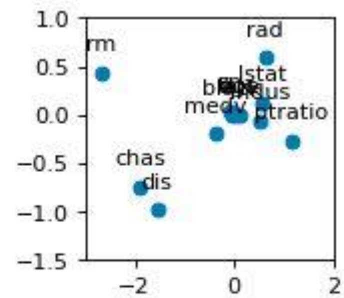
rad -> reject null

medv ->reject null

This implies that the only statistically significant predictors are dis, rad and medv.

(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 regression model is shown on the x-axis, and its coefficient estimate in the multiple linear regression model is shown on the y-axis.





We can see that the univariate regression predicts a large positive coefficient for nox, but the multivariate regression predicts a large negative coefficient. On the right I have zoomed in on the remaining coefficients. rm, ptratio and indus seem to be the only coefficients for which the univariate and multivariate regressions disagree in sign. For the remaining predictors the values are about the same for both regressions. Exception to these are chas and dis, for which the univariate regression predicts a stronger relationship than the multivariate relationship.

(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

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

	coef	std err	t	P>|t|	[0.025	0.975]
const	19.1836	11.796	1.626	0.105	-3.991	42.358
tax	-0.1533	0.096	-1.602	0.110	-0.341	0.035
tax^2	0.0004	0.000	1.488	0.137	-0.000	0.001
tax^3	-2.204e-07	1.89e-07	-1.167	0.244	-5.91e-07	1.51e-07

For tax p-value for all coefficients is greater than 0.01, implying that there is no evidence of non-linear association.

	coef	std err	t	P>|t|	[0.025	0.975]
const	112.6246	64.517	1.746	0.081	-14.132	239.382
rm	-39.1501	31.311	-1.250	0.212	-100.668	22.368
rm^2	4.5509	5.010	0.908	0.364	-5.292	14.394
rm^3	-0.1745	0.264	-0.662	0.509	-0.693	0.344

For rm p-value for all coefficients is greater than 0.01, implying that there is no evidence of non-linear association.

	coef	std err	t	P>|t|	[0.025	0.975]
const	-0.6055	2.050	-0.295	0.768	-4.633	3.422
rad	0.5127	1.044	0.491	0.623	-1.538	2.563
rad^2	-0.0752	0.149	-0.506	0.613	-0.367	0.217
rad^3	0.0032	0.005	0.703	0.482	-0.006	0.012

For rad p-value for all coefficients is greater than 0.01, implying that there is no evidence of non-linear association.

	coef	std err	t	P>|t|	[0.025	0.975]
const	53.1655	3.356	15.840	0.000	46.571	59.760
medv	-5.0948	0.434	-11.744	0.000	-5.947	-4.242
medv^2	0.1555	0.017	9.046	0.000	0.122	0.189
medv^3	-0.0015	0.000	-7.312	0.000	-0.002	-0.001

For medv p-value for all coefficients is less than 0.01, implying that there is evidence of non-linear association.

	coef	std err	t	P>|t|	[0.025	0.975]
const	1.2010	2.029	0.592	0.554	-2.785	5.187
lstat	-0.4491	0.465	-0.966	0.335	-1.362	0.464
lstat^2	0.0558	0.030	1.852	0.065	-0.003	0.115
lstat^3	-0.0009	0.001	-1.517	0.130	-0.002	0.000

For lstat p-value for all coefficients is greater than 0.01, implying that there is no evidence of non-linear association.

	coef	std err	t	P>|t|	[0.025	0.975]
const	18.2637	2.305	7.924	0.000	13.735	22.792
black	-0.0836	0.056	-1.483	0.139	-0.194	0.027
black^2	0.0002	0.000	0.716	0.474	-0.000	0.001
black^3	-2.652e-07	4.36e-07	-0.608	0.544	-1.12e-06	5.92e-07

For black p-value is less than 0.01 only for the constant coefficient, implying that there is no evidence of non-linear association.

	coef	std err	t	P>|t|	[0.025	0.975]
const	30.0476	2.446	12.285	0.000	25.242	34.853
dis	-15.5544	1.736	-8.960	0.000	-18.965	-12.144
dis^2	2.4521	0.346	7.078	0.000	1.771	3.133
dis^3	-0.1186	0.020	-5.814	0.000	-0.159	-0.079

For dis p-value for all coefficients is less than 0.01, implying that there is evidence of non-linear association.

	coef	std err	t	P>|t|	[0.025	0.975]
const	233.0866	33.643	6.928	0.000	166.988	299.185
nox	-1279.3713	170.397	-7.508	0.000	-1614.151	-944.591
nox^2	2248.5441	279.899	8.033	0.000	1698.626	2798.462
nox^3	-1245.7029	149.282	-8.345	0.000	-1538.997	-952.409

For nox p-value for all coefficients is less than 0.01, implying that there is evidence of non-linear association.

	coef	std err	t	P>|t|	[0.025	0.975]
const	-2.5488	2.769	-0.920	0.358	-7.989	2.892
age	0.2737	0.186	1.468	0.143	-0.093	0.640
age^2	-0.0072	0.004	-1.988	0.047	-0.014	-8.4e-05
age^3	5.745e-05	2.11e-05	2.724	0.007	1.6e-05	9.89e-05

For age p-value for the cubic coefficient is less than 0.01, implying that there is evidence of non-linear association.

	coef	std err	t	P>|t|	[0.025	0.975]
const	3.6626	1.574	2.327	0.020	0.570	6.755
indus	-1.9652	0.482	-4.077	0.000	-2.912	-1.018
indus^2	0.2519	0.039	6.407	0.000	0.175	0.329
indus^3	-0.0070	0.001	-7.292	0.000	-0.009	-0.005

For indus p-value for all coefficients is less than 0.01, implying that there is evidence of non-linear association.

	coef	std err	t	P>|t|	[0.025	0.975]
const	477.1840	156.795	3.043	0.002	169.129	785.239
ptratio	-82.3605	27.644	-2.979	0.003	-136.673	-28.048
ptratio^2	4.6353	1.608	2.882	0.004	1.475	7.795
ptratio^3	-0.0848	0.031	-2.743	0.006	-0.145	-0.024

For ptratio p-value for all coefficients is less than 0.01, implying that there is evidence of non-linear association.

Code Section

Questions 1&2:

```
import numpy as np
import matplotlib
import matplotlib.pyplot as plt
from matplotlib.pyplot import figure


import statsmodels.api as sm
from statsmodels.formula.api import ols
from statsmodels.stats.outliers_influence import OLSInfluence
```

```
import statsmodels.formula.api as smf

!pip install pandas
import pandas as pd

from google.colab import data_table
data_table.disable_dataframe_formatter()

data = pd.read_csv('Auto.csv', na_values='?').dropna()


#Question1
'''
hp = pd.to_numeric(data[data['horsepower'] != '?']['horsepower'])
mpg = pd.to_numeric(data[data['horsepower'] != '?']['mpg'])

plt.scatter(hp, mpg)
hp = sm.add_constant(hp)
model = sm.OLS(mpg, hp)
results = model.fit()
results.summary().tables[1]
print(results.params)

print("The predicted value of mpg for a 95 horsepower car is: ")
print(results.predict([1, 95]))

plt.plot([0, 225], [results.params[0],  results.params[0] + 225 * results.
params[1]])
plt.plot(95, results.predict([1, 95]), marker="o", markersize = 10, marker
edgecolor="red", markerfacecolor="green")

plt.xlabel('horsepower')
plt.ylabel('mpg')

'''
#Question2
data = data.loc[:, data.columns!='name']
data['american'] = 0
data['european'] = 0
data['japanese'] = 0
data.loc[(data.origin == 1), 'american']= 1
data.loc[(data.origin == 2), 'european']= 1
data.loc[(data.origin == 3), 'japanese']= 1
data =  data.loc[:, data.columns!='origin']
'''
```

```
#a)
pd.plotting.scatter_matrix(data, alpha=0.2, figsize=(14,14))
#b)
corrM = data.corr()
corrM
#c
predictors = data.loc[:, data.columns!='mpg']
predictors['american'] = 0
predictors['european'] = 0
predictors['japanese'] = 0
predictors.loc[(predictors.origin == 1), 'american']= 1
predictors.loc[(predictors.origin == 2), 'european']= 1
predictors.loc[(predictors.origin == 3), 'japanese']= 1
predictors =  predictors.loc[:, predictors.columns!='origin']



response = data['mpg']
predictors = sm.add_constant(predictors)
model = sm.OLS(response, predictors.astype(float))
results = model.fit()
print(results.params)
results.summary()
#d)

predictions = results.predict(predictors)
residuals = response - predictions;

#residual plot
plt.title('Diagnostics Plot')
plt.xlabel('predicted values')
plt.ylabel('residuals')
plt.scatter(predictions, residuals)

figure(figsize=(16, 16), dpi=80)
sm.graphics.influence_plot(results, size=6)
#e1) plot interactions
model_all_interactions = smf.ols(formula='mpg ~ cylinders + displacement +
 horsepower + weight + acceleration + year + american + european + japanes
e + cylinders:displacement + cylinders:horsepower + cylinders:weight + cyl
inders:acceleration + cylinders:year + displacement:horsepower + displacem
ent:weight + displacement:acceleration + displacement:year + horsepower:we
ight + horsepower:acceleration + horsepower:year + weight:acceleration + w
eight:year + acceleration:year', data=data).fit()
model_all_interactions.summary().tables[1]
```

```
model_some_interactions_1 = smf.ols(formula='mpg ~ displacement + accelera
tion + horsepower + weight + year + displacement:acceleration + displaceme
nt:horsepower + displacement:weight + acceleration:horsepower + accelerati
on:weight + horsepower:weight', data=data).fit()
model_some_interactions_1.summary().tables[1]

model_some_interactions_2 = smf.ols(formula='mpg ~ displacement + accelera
tion + horsepower + weight + year + displacement:acceleration + displaceme
nt:weight + acceleration:horsepower + acceleration:weight + horsepower:wei
ght', data=data).fit()
model_some_interactions_2.summary().tables[1]
'''
#e2) plot interactions

model_all_interactions = smf.ols(formula='mpg ~ cylinders:displacement + c
ylinders:horsepower + cylinders:weight + cylinders:acceleration + cylinder
s:year + displacement:horsepower + displacement:weight + displacement:acce
leration + displacement:year + horsepower:weight + horsepower:acceleration
 + horsepower:year + weight:acceleration + weight:year + acceleration:year
', data=data).fit()
model_all_interactions.summary().tables[1]

model_some_interactions_1 = smf.ols(formula='mpg ~ cylinders:acceleration
+ cylinders:year + displacement:horsepower + displacement:acceleration + d
isplacement:year + weight:acceleration + acceleration:year', data=data).fi
t()
model_some_interactions_1.summary().tables[1]

model_some_interactions_2 = smf.ols(formula='mpg ~ cylinders:year + displa
cement:horsepower + displacement:acceleration + displacement:year + weight
:acceleration + acceleration:year', data=data).fit()
model_some_interactions_2.summary().tables[1]
'''
#f)
predictors =  data.loc[:, data.columns!='mpg']
response = data['mpg']
predictorsSquared = sm.add_constant(np.power(predictors.astype(float), 2))
predictorsSquareRoot = sm.add_constant(np.power(predictors.astype(float),
0.5))

resultsSquared = sm.OLS(response, predictorsSquared).fit()
resultsSquareRoot = sm.OLS(response, predictorsSquareRoot).fit()

predictors =  predictors.loc[:, predictors.columns!='american']
predictors =  predictors.loc[:, predictors.columns!='european']
predictors =  predictors.loc[:, predictors.columns!='japanese']
```

```
predictorsLog = sm.add_constant(np.log(predictors.astype(float)))
resultsLog = sm.OLS(response, predictorsLog).fit()

resultsSquared.summary()
resultsSquareRoot.summary()
resultsLog.summary()
'''
plt.show()
```

Question 3

```
import numpy as np
import matplotlib
import matplotlib.pyplot as plt
from matplotlib.pyplot import figure


import statsmodels.api as sm
from statsmodels.formula.api import ols
from statsmodels.stats.outliers_influence import OLSInfluence

import statsmodels.formula.api as smf

!pip install pandas
import pandas as pd

from google.colab import data_table
data_table.disable_dataframe_formatter()

data = pd.read_csv('Carseats.csv', na_values='?').dropna()

#Question3
#a)

data.loc[(data.Urban == 'Yes'),'Urban']= 1
data.loc[(data.Urban == 'No'),'Urban']= 0

data.loc[(data.US == 'Yes'),'US']= 1
data.loc[(data.US == 'No'),'US']= 0

model = smf.ols(formula='Sales ~ Price + Urban + US', data=data).fit()
```

```
#e)
model_noUrban = smf.ols(formula='Sales ~ Price + US', data=data).fit()

#g)

#h)
sm.graphics.influence_plot(model_noUrban, size=6)


plt.show()
```

Question 4

```
import numpy as np
import matplotlib
import matplotlib.pyplot as plt
from matplotlib.pyplot import figure


import statsmodels.api as sm
from statsmodels.formula.api import ols
from statsmodels.stats.outliers_influence import OLSInfluence

import statsmodels.formula.api as smf

!pip install pandas
import pandas as pd

from google.colab import data_table
data_table.disable_dataframe_formatter()

data = pd.read_csv('Boston.csv', na_values='?').dropna()

crim = data['crim']
predictors =  data.loc[:, data.columns!='crim']

zn = data['zn']
indus = data['indus']
chas = data['chas']
nox = data['nox']
rm = data['rm']
age = data['age']
dis = data['dis']
```

```
rad = data['rad']
tax = data['tax']
ptratio = data['ptratio']
black = data['black']
lstat = data['lstat']
medv = data['medv']

#a)
model_zn = smf.ols(formula='crim ~ zn', data=data).fit()
model_indus = smf.ols(formula='crim ~ indus', data=data).fit()
model_chas = smf.ols(formula='crim ~ chas', data=data).fit()
model_nox = smf.ols(formula='crim ~ nox', data=data).fit()
model_rm = smf.ols(formula='crim ~ rm', data=data).fit()
model_age = smf.ols(formula='crim ~ age', data=data).fit()
model_dis = smf.ols(formula='crim ~ dis', data=data).fit()
model_rad = smf.ols(formula='crim ~ rad', data=data).fit()
model_tax = smf.ols(formula='crim ~ tax', data=data).fit()
model_ptratio = smf.ols(formula='crim ~ ptratio', data=data).fit()
model_black = smf.ols(formula='crim ~ black', data=data).fit()
model_lstat = smf.ols(formula='crim ~ lstat', data=data).fit()
model_medv = smf.ols(formula='crim ~ medv', data=data).fit()

model_zn.summary().tables[1]
model_indus.summary().tables[1]
model_chas.summary().tables[1]
model_nox.summary().tables[1]
model_rm.summary().tables[1]
model_age.summary().tables[1]
model_dis.summary().tables[1]
model_rad.summary().tables[1]
model_tax.summary().tables[1]
model_ptratio.summary().tables[1]
model_black.summary().tables[1]
model_lstat.summary().tables[1]
model_medv.summary().tables[1]

fig, axs = plt.subplots(3, 5)
axs[0, 0].set_title('crim vs zn')
axs[0, 0].scatter(zn, crim)

axs[0, 1].set_title('crim vs indus')
axs[0, 1].scatter(indus, crim)

axs[0, 2].set_title('crim vs chas')
axs[0, 2].scatter(chas, crim)
```

```
axs[0, 3].set_title('crim vs nox')
axs[0, 3].scatter(nox, crim)

axs[0, 4].set_title('crim vs rm')
axs[0, 4].scatter(rm, crim)

axs[1, 0].set_title('crim vs age')
axs[1, 0].scatter(age, crim)

axs[1, 1].set_title('crim vs dis')
axs[1, 1].scatter(dis, crim)

axs[1, 2].set_title('crim vs rad')
axs[1, 2].scatter(rad, crim)

axs[1, 3].set_title('crim vs tax')
axs[1, 3].scatter(tax, crim)

axs[1, 4].set_title('crim vs ptratio')
axs[1, 4].scatter(ptratio, crim)

axs[2, 0].set_title('crim vs black')
axs[2, 0].scatter(black, crim)

axs[2, 1].set_title('crim vs lstat')
axs[2, 1].scatter(lstat, crim)

axs[2, 2].set_title('crim vs medv')
axs[2, 2].scatter(medv, crim)

fig.set_figheight(9)
fig.set_figwidth(15)
'''
#b)
model = smf.ols(formula='crim ~ zn + indus + chas + nox + rm + age + dis +
 rad + tax + ptratio + black + lstat + medv', data=data).fit()
model.summary().tables[1]
#c)

X = [model_zn.params[1], model_indus.params[1], model_chas.params[1], mode
l_nox.params[1], model_rm.params[1], model_age.params[1], model_dis.params
[1], model_rad.params[1], model_tax.params[1], model_ptratio.params[1], mo
del_black.params[1], model_lstat.params[1], model_medv.params[1]]
Y = [model.params[1], model.params[2], model.params[3], model.params[4], m
odel.params[5], model.params[6], model.params[7], model.params[8], model.p
```

```
arams[9], model.params[10], model.params[11], model.params[12], model.para
ms[13]]
#figure(figsize=(2, 2), dpi=80)

#plt.xlim([-3, 2])
#plt.ylim([-1.5, 1])
plt.xlabel('univariate regression coefficients', fontsize=12)
plt.ylabel('multivariate regression coefficients', fontsize=12)
plt.scatter(X,Y)

plt.annotate('zn', (X[0],Y[0]),  textcoords="offset points",  xytext=(0,10
), ha='center')
plt.annotate('indus', (X[1],Y[1]),  textcoords="offset points",  xytext=(0
,10), ha='center')
plt.annotate('chas', (X[2],Y[2]),  textcoords="offset points",  xytext=(0,
10), ha='center')
plt.annotate('nox', (X[3],Y[3]),  textcoords="offset points",  xytext=(0,1
0), ha='center')
plt.annotate('rm', (X[4],Y[4]),  textcoords="offset points",  xytext=(0,10
), ha='center')
plt.annotate('age', (X[5],Y[5]),  textcoords="offset points",  xytext=(0,1
0), ha='center')
plt.annotate('dis', (X[6],Y[6]),  textcoords="offset points",  xytext=(0,1
0), ha='center')
plt.annotate('rad', (X[7],Y[7]),  textcoords="offset points",  xytext=(0,1
0), ha='center')
plt.annotate('tax', (X[8],Y[8]),  textcoords="offset points",  xytext=(0,1
0), ha='center')
plt.annotate('ptratio', (X[9],Y[9]),  textcoords="offset points",  xytext=
(0,10), ha='center')
plt.annotate('black', (X[10],Y[10]),  textcoords="offset points",  xytext=
(0,10), ha='center')
plt.annotate('lstat', (X[11],Y[11]),  textcoords="offset points",  xytext=
(0,10), ha='center')
plt.annotate('medv', (X[12],Y[12]),  textcoords="offset points",  xytext=(
0,10), ha='center')


#d)
zn_cubic = pd.DataFrame(columns = ['zn', 'zn^2', 'zn^3'])
zn_cubic['zn'] = zn
zn_cubic['zn^2'] = zn*zn
zn_cubic['zn^3'] = zn*zn*zn
zn_cubic = sm.add_constant(zn_cubic)
model_zn_cube = sm.OLS(crim, zn_cubic).fit()
```

```
indus_cubic = pd.DataFrame(columns = ['indus', 'indus^2', 'indus^3'])
indus_cubic['indus'] = indus
indus_cubic['indus^2'] = indus*indus
indus_cubic['indus^3'] = indus*indus*indus
indus_cubic = sm.add_constant(indus_cubic)
model_indus_cube = sm.OLS(crim, indus_cubic).fit()

chas_cubic = pd.DataFrame(columns = ['chas', 'chas^2', 'chas^3'])
chas_cubic['chas'] = chas
chas_cubic['chas^2'] = chas*chas
chas_cubic['chas^3'] = chas*chas*chas
chas_cubic = sm.add_constant(chas_cubic)
model_chas_cube = sm.OLS(crim, chas_cubic).fit()
model_chas_cube.summary().tables[1]

nox_cubic = pd.DataFrame(columns = ['nox', 'nox^2', 'nox^3'])
nox_cubic['nox'] = nox
nox_cubic['nox^2'] = nox*nox
nox_cubic['nox^3'] = nox*nox*nox
nox_cubic = sm.add_constant(nox_cubic)
model_nox_cube = sm.OLS(crim, nox_cubic).fit()

rm_cubic = pd.DataFrame(columns = ['rm', 'rm^2', 'rm^3'])
rm_cubic['rm'] = rm
rm_cubic['rm^2'] = rm*rm
rm_cubic['rm^3'] = rm*rm*rm
rm_cubic = sm.add_constant(rm_cubic)
model_rm_cube = sm.OLS(crim, rm_cubic).fit()

age_cubic = pd.DataFrame(columns = ['age', 'age^2', 'age^3'])
age_cubic['age'] = age
age_cubic['age^2'] = age*age
age_cubic['age^3'] = age*age*age
age_cubic = sm.add_constant(age_cubic)
model_age_cube = sm.OLS(crim, age_cubic).fit()

dis_cubic = pd.DataFrame(columns = ['dis', 'dis^2', 'dis^3'])
dis_cubic['dis'] = dis
dis_cubic['dis^2'] = dis*dis
dis_cubic['dis^3'] = dis*dis*dis
dis_cubic = sm.add_constant(dis_cubic)
model_dis_cube = sm.OLS(crim, dis_cubic).fit()

rad_cubic = pd.DataFrame(columns = ['rad', 'rad^2', 'rad^3'])
rad_cubic['rad'] = rad
rad_cubic['rad^2'] = rad*rad
```

```
rad_cubic['rad^3'] = rad*rad*rad
rad_cubic = sm.add_constant(rad_cubic)
model_rad_cube = sm.OLS(crim, rad_cubic).fit()

tax_cubic = pd.DataFrame(columns = ['tax', 'tax^2', 'tax^3'])
tax_cubic['tax'] = tax
tax_cubic['tax^2'] = tax*tax
tax_cubic['tax^3'] = tax*tax*tax
tax_cubic = sm.add_constant(tax_cubic)
model_tax_cube = sm.OLS(crim, tax_cubic).fit()

ptratio_cubic = pd.DataFrame(columns = ['ptratio', 'ptratio^2', 'ptratio^3
'])
ptratio_cubic['ptratio'] = ptratio
ptratio_cubic['ptratio^2'] = ptratio*ptratio
ptratio_cubic['ptratio^3'] = ptratio*ptratio*ptratio
ptratio_cubic = sm.add_constant(ptratio_cubic)
model_ptratio_cube = sm.OLS(crim, ptratio_cubic).fit()

black_cubic = pd.DataFrame(columns = ['black', 'black^2', 'black^3'])
black_cubic['black'] = black
black_cubic['black^2'] = black*black
black_cubic['black^3'] = black*black*black
black_cubic = sm.add_constant(black_cubic)
model_black_cube = sm.OLS(crim, black_cubic).fit()

lstat_cubic = pd.DataFrame(columns = ['lstat', 'lstat^2', 'lstat^3'])
lstat_cubic['lstat'] = lstat
lstat_cubic['lstat^2'] = lstat*lstat
lstat_cubic['lstat^3'] = lstat*lstat*lstat
lstat_cubic = sm.add_constant(lstat_cubic)
model_lstat_cube = sm.OLS(crim, lstat_cubic).fit()

medv_cubic = pd.DataFrame(columns = ['medv', 'medv^2', 'medv^3'])
medv_cubic['medv'] = medv
medv_cubic['medv^2'] = medv*medv
medv_cubic['medv^3'] = medv*medv*medv
medv_cubic = sm.add_constant(medv_cubic)
model_medv_cube = sm.OLS(crim, medv_cubic).fit()

model_zn_cube.summary().tables[1]
model_indus_cube.summary().tables[1]
model_nox_cube.summary().tables[1]
model_rm_cube.summary().tables[1]
model_age_cube.summary().tables[1]
model_dis_cube.summary().tables[1]
```

```
model_rad_cube.summary().tables[1]
model_tax_cube.summary().tables[1]
model_ptratio_cube.summary().tables[1]
model_black_cube.summary().tables[1]
model_lstat_cube.summary().tables[1]
model_medv_cube.summary().tables[1]
'''

plt.show()
```