An Introduction to Statistical Learning (with python application) Chapter 3 Exercise: (3), (10), (14)

3. Suppose we have a data set with five predictors, X1 = GPA, X2 = IQ, X3 = Level (1 for College and 0 for High School), X4 = Interaction between GPA and IQ, and X5 = Interaction between GPA and Level. The response is starting salary after graduation (in thousands of dollars). Suppose we use least squares to fit the model, and get

$$\hat{\beta}_0 = 50, \hat{\beta}_1 = 20, \hat{\beta}_2 = 0.07, \hat{\beta}_3 = 35, \hat{\beta}_4 = 0.01, \hat{\beta}_5 = -10.$$

a) Which answer is correct, and why?

i. For a fixed value of IQ and GPA, high school graduates earn more, on average, than college graduates.

ii. For a fixed value of IQ and GPA, college graduates earn more, on average, than high school graduates.

iii. For a fixed value of IQ and GPA, high school graduates earn more, on average, than college graduates provided that the GPA is high enough.

iv. For a fixed value of IQ and GPA, college graduates earn more, on average, than high school graduates provided that the GPA is high enough.

Answer:

Here equation for $Y = \beta0+\beta1\times GPA+\beta2\times IQ+\beta3\times Level+\beta4\times GPA\times IQ+\beta5\times GPA\times Level$

Basic salary difference between school graduate and college graduate is $35000. College graduate earn more without any condition.

All four options have fixed IQ and GPA

We can drive equation as $Ycollege- Yschool = \beta3 + \beta5\times GPA = 35 - 10GPA$

As with GPA showing negative correlation as GPA increases to significant level the school salary become more than college level.

So, iii) option is correct.

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

Answer:

From eq: $Y = \beta0+\beta1\times GPA+\beta2\times IQ+\beta3\times Level+\beta4\times GPA\times IQ+\beta5\times GPA\times Level$

$Y= 50+20*4+0.07*110+35+0.01*110*4+(-10)*4$

$= 50+80+7.7+35+4.4-40$

$= 137.1$

So, college graduate salary will be $137100

(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.

Answer:

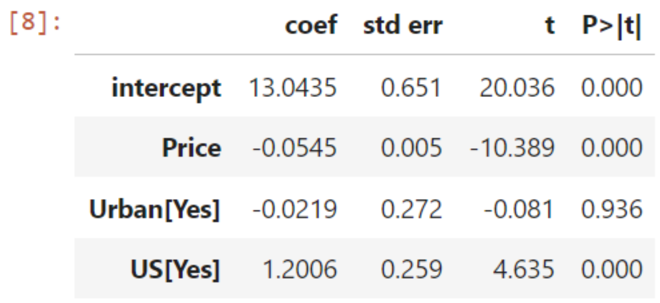
Coefficient of GPA/IQ interaction is 0.01 which is very small but still need to consider the P-value by f-statistics to prove that there is very little evidence of an interaction effect.

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

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

Python code:

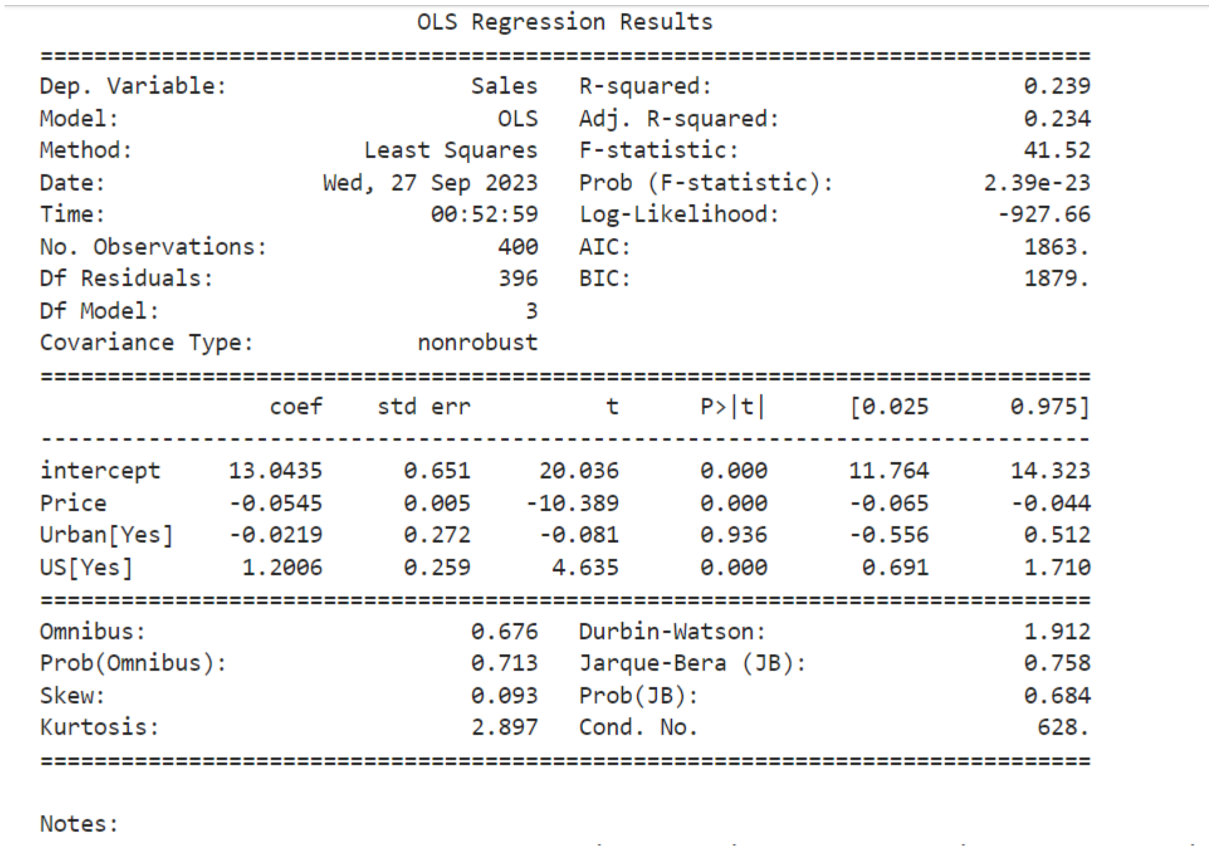
```
import numpy as np
import pandas as pd
from matplotlib.pyplot import subplots
import statsmodels.api as sm
from statsmodels.stats.outliers_influence \
    import variance_inflation_factor as VIF
from statsmodels.stats.anova import anova_lm
from ISLP import load_data
from ISLP.models import (ModelSpec as MS,
                         summarize,
                         poly)
Carseats = load_data("Carseats")
Carseats.columns
Carseats
y = Carseats['Sales']
X = MS(['Price', 'Urban', 'US']).fit_transform(Carseats)
model1 = sm.OLS(y, X)
results1 = model1.fit()
summarize(results1)
```



[8]:

	coef	std err	t	P>|t|
intercept	13.0435	0.651	20.036	0.000
Price	-0.0545	0.005	-10.389	0.000
Urban[Yes]	-0.0219	0.272	-0.081	0.936
US[Yes]	1.2006	0.259	4.635	0.000

```
print(results1.summary())
```



```
                            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:                Wed, 27 Sep 2023   Prob (F-statistic):           2.39e-23
Time:                        00:52:59   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
Price         -0.0545      0.005    -10.389      0.000      -0.065      -0.044
Urban[Yes]    -0.0219      0.272     -0.081      0.936      -0.556       0.512
US[Yes]        1.2006      0.259      4.635      0.000       0.691       1.710
==============================================================================
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.
==============================================================================

Notes:
```

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

Answer:

Coefficients are Intercept, Price, Urban and US

Here Intercept value is 13.0435 which is estimated average sales when price, Urban and US are 0.

Price is quantitative variable with negative relationship with sales having coefficient -0.0545, which means sales will decrease when price increases.

Urban is qualitative categorical variable with very less -0.0219 coefficient and p value more than 0.5 so not as effective variable to predict sales.

US is also qualitative variable with positive coefficient represent sales is more if store is in US.

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

Answer:

Sales = 13.0435 - 0.0545*Price – 0.0219* Urban + 1.2006 * US

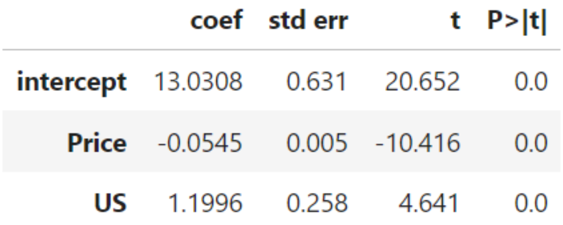
(d) For which of the predictors can you reject the null hypothesis H0 : #j = 0?

Answer: For Urban we can reject the null hypothesis as P value is more than 0.5

(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.

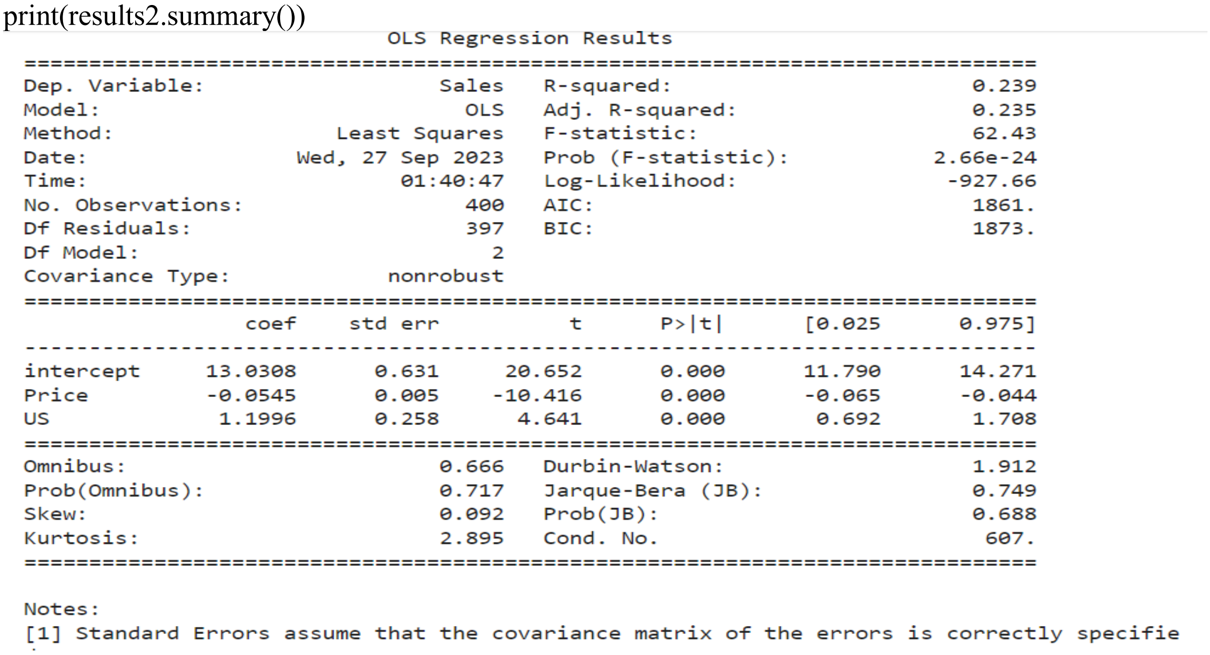
Answer:

Python code:
```
y = Carseats['Sales']
X = MS(['Price', 'US']).fit_transform(Carseats)
model2 = sm.OLS(y, X)
results2 = model2.fit()
summarize(results2)
```



	coef	std err	t	P>|t|
intercept	13.0308	0.631	20.652	0.0
Price	-0.0545	0.005	-10.416	0.0
US	1.1996	0.258	4.641	0.0

print(results2.summary())



```
                            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:                Wed, 27 Sep 2023   Prob (F-statistic):           2.66e-24
Time:                        01:40:47   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
Price         -0.0545      0.005    -10.416      0.000      -0.065      -0.044
US             1.1996      0.258      4.641      0.000       0.692       1.708
==============================================================================
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.
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specifie
```

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

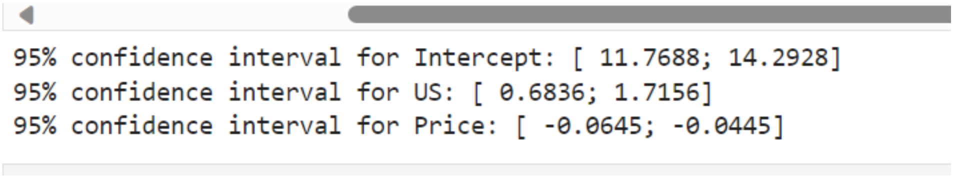
Answer:

R square value in both models is same 0.239 so model with less variable i.e. (e) is better model, However, due to low r squared value both models not fit the data very well.

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

Answer:

```
intercept_coef = 13.0308
intercept_stderr = .631
us_coef = 1.1996
us_stderr = .258
price_coef = -.0545
price_stderr = .005
print('95%% confidence interval for Intercept: [ %2.4f; %2.4f] ' %
(intercept_coef-2*intercept_stderr, intercept_coef+2*intercept_stderr))
print('95%% confidence interval for Intercept: [ %2.4f; %2.4f] ' % (us_coef-
2*us_stderr, us_coef+2*us_stderr))
print('95%% confidence interval for Intercept: [ %2.4f; %2.4f] ' % (price_coef-
2*price_stderr, price_coef+2*price_stderr))
```

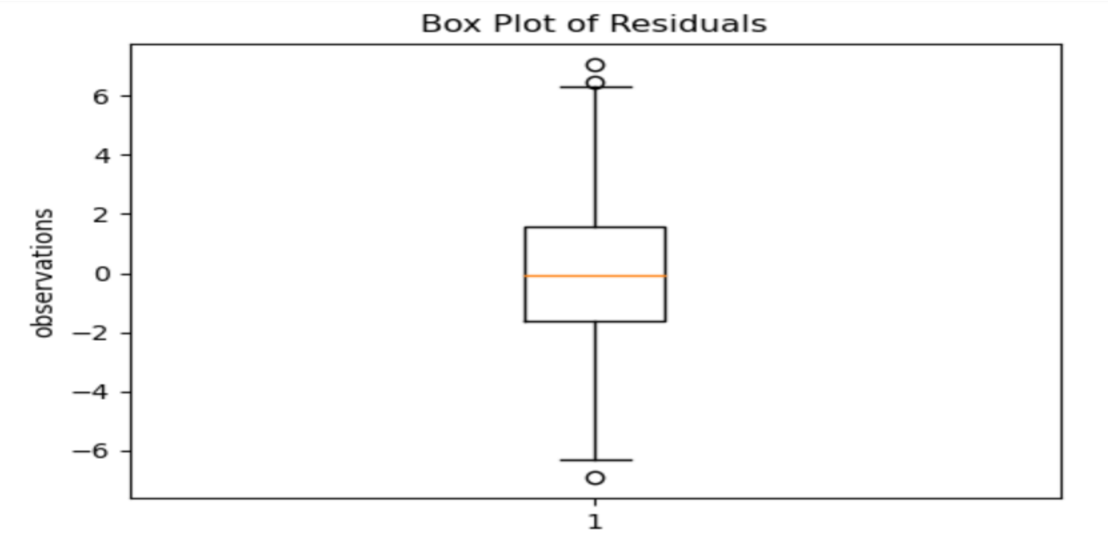


```
95% confidence interval for Intercept: [ 11.7688; 14.2928]
95% confidence interval for US: [ 0.6836; 1.7156]
95% confidence interval for Price: [ -0.0645; -0.0445]
```

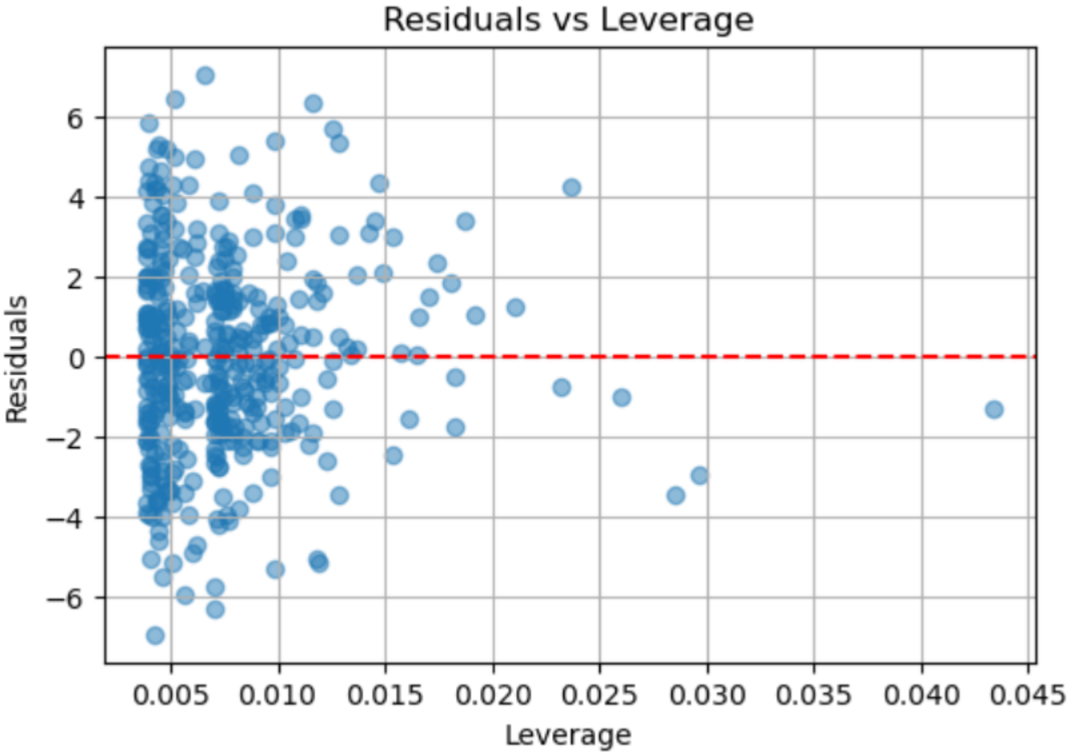
(h) Is there evidence of outliers or high leverage observations in the model from (e)?

Answer: As per below plots there seems some outliners and high leverage points.

Python code:
```
## residuals outliners
plt.figure(figsize=(6, 4))
plt.boxplot(residuals)
plt.title('Box Plot of Residuals')
plt.ylabel('Residuals')
plt.ylabel('observations')
plt.show()
```



```
## check for outliners & leverage plot
fitted_y = results2.fittedvalues
normalized_resid = results2.get_influence().resid_studentized_internal
absolute_sqrt_norm_resid = np.sqrt(np.abs(normalized_resid))
absolute_resid = np.abs(residuals)
leverage = results2.get_influence().hat_matrix_diag
plt.figure(figsize=(6, 4))
plt.scatter(leverage, residuals, alpha=0.5)
plt.axhline(y=0, color='r', linestyle='--')
plt.title('Residuals vs Leverage')
plt.xlabel('Leverage')
plt.ylabel('Residuals')
plt.grid(True)
plt.show()
```



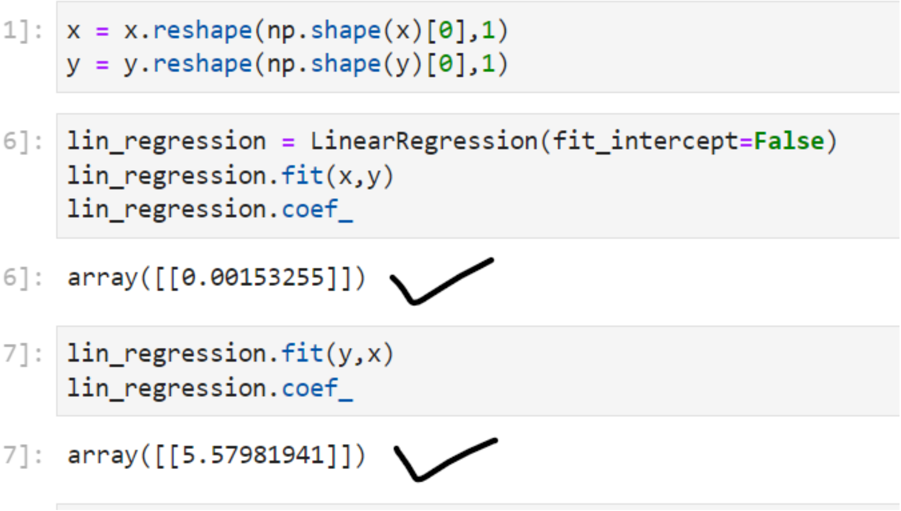
12. This problem involves simple linear regression without an intercept.
(a) Recall that the coefficient estimate β for the linear regression of Y onto X without an intercept is given by (3.38). Under what circumstance is the coefficient estimate for the regression of X onto Y the same as the coefficient estimate for the regression of Y onto X?
Answer:

When $\sum x_i^2 = \sum y_i^2$ then the coefficient estimates for the regression of X onto Y the same as the coefficient estimate for the regression of Y onto X.

(b) Generate an example in Python with n = 100 observations in which the coefficient estimate for the regression of X onto Y is different from the coefficient estimate for the regression of Y onto X.
Answer:
Python code:

```
## chapter 3 problem 12
from sklearn.linear_model import LinearRegression
```

```
x = np.arange(100)
y = np.random.normal(size=100)
## data insights
print(x)
print(y)
x = x.reshape(np.shape(x)[0],1)
y = y.reshape(np.shape(y)[0],1)
lin_regression = LinearRegression(fit_intercept=False)
lin_regression.fit(x,y)
lin_regression.coef_
lin_regression.fit(y,x)
lin_regression.coef_
```



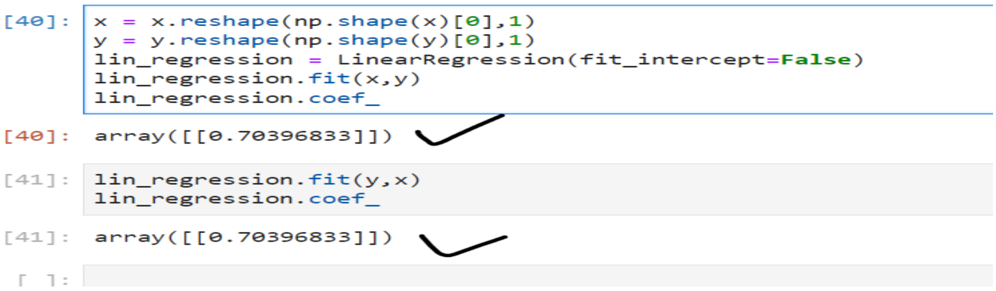
```
1]: x = x.reshape(np.shape(x)[0],1)
    y = y.reshape(np.shape(y)[0],1)

6]: lin_regression = LinearRegression(fit_intercept=False)
    lin_regression.fit(x,y)
    lin_regression.coef_

6]: array([[0.00153255]]) ✓

7]: lin_regression.fit(y,x)
    lin_regression.coef_

7]: array([[5.57981941]]) ✓
```

(c) Generate an example in Python with n = 100 observations in which the coefficient estimate for the regression of X onto Y is the same as the coefficient estimate for the regression of Y onto X.

Answer:

Python code:

```
x = np.arange(100)
y = np.random.permutation(x)
print(x)
print(y)
x = x.reshape(np.shape(x)[0],1)
y = y.reshape(np.shape(y)[0],1)
lin_regression = LinearRegression(fit_intercept=False)
lin_regression.fit(x,y)
lin_regression.coef_
lin_regression.fit(y,x)
lin_regression.coef_
```



```
[40]: x = x.reshape(np.shape(x)[0],1)
      y = y.reshape(np.shape(y)[0],1)
      lin_regression = LinearRegression(fit_intercept=False)
      lin_regression.fit(x,y)
      lin_regression.coef_

[40]: array([[0.70396833]]) ✓

[41]: lin_regression.fit(y,x)
      lin_regression.coef_

[41]: array([[0.70396833]]) ✓

[ ]:
```

5.8 Twitter users and news, Part I. A poll conducted in 2013 found that 52% of U.S. adult Twitter users get at least some news on Twitter.12. The standard error for this estimate was 2.4%, and a normal distribution may be used to model the sample proportion. Construct a 99% confidence interval for the fraction of U.S. adult Twitter users who get some news on Twitter, and interpret the confidence interval in context.

Answer:

CI formula = $\text{point estimate} \pm z^{\star} \times SE$

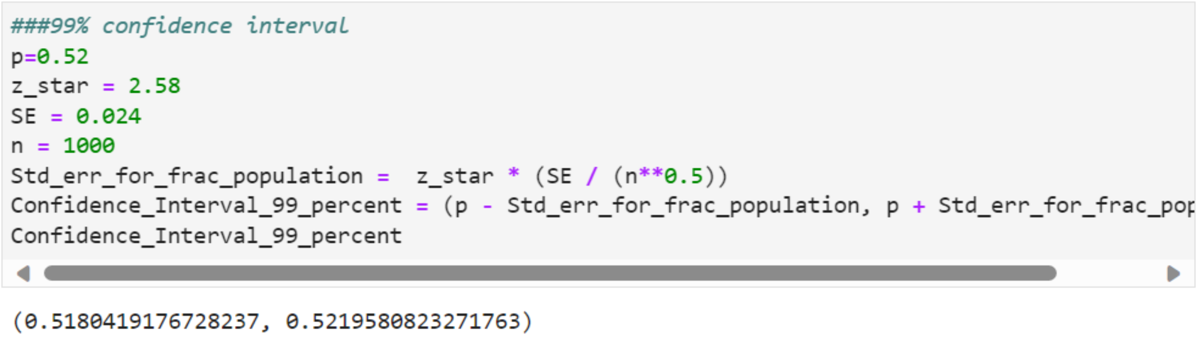
Here, p = 0.52

Z* for 99% CI = 2.58

SE = 0.024

As sample size need to consider for fraction of U.S. adult twitter population so taking n = 1000

As per formula we are getting 99% CI = 0.518% & 0.522%



```
###99% confidence interval
p=0.52
z_star = 2.58
SE = 0.024
n = 1000
Std_err_for_frac_population =  z_star * (SE / (n**0.5))
Confidence_Interval_99_percent = (p - Std_err_for_frac_population, p + Std_err_for_frac_pop
Confidence_Interval_99_percent
```

```
(0.5180419176728237, 0.5219580823271763)
```

From this we can say or 99 percent confident that 0.518% to 0.522% of US adult twitter users get some news on twitter.

5.16 Identify hypotheses, Part II. Write the null and alternative hypotheses in words and using symbols for each of the following situations.

(a) Since 2008, chain restaurants in California have been required to display calorie counts of each menu item. Prior to menus displaying calorie counts, the average calorie intake of diners at a restaurant was 1100 calories. After calorie counts started to be displayed on menus, a nutritionist collected data on the number of calories consumed at this restaurant from a random sample of diners. Do these data provide convincing evidence of a difference in the average calorie intake of a diners at this restaurant?

Answer:

Null Hypotheses(H_0) = Average Calories count of diners at chain restaurants after calorie count displayed on menus is equal to 1000 (μ=1100).

Alternative Hypotheses(H_a) = Average Calories count of diners at chain restaurants after calorie count displayed on menus is not equal to 1000 ($\mu \neq 1100$).

Null Hypotheses = The calorie count display on menus has no evidence on change of average calorie intake of diners at chain restaurants.

Alternative Hypotheses = The calorie count display on menus has significant evidence on change of average calorie intake of diners at chain restaurants.

(b) The state of Wisconsin would like to understand the fraction of its adult residents that consumed alcohol in the last year, specifically if the rate is different from the national rate of 70%. To help them answer this question, they conduct a random sample of 852 residents and ask them about their alcohol consumption.

Answer:

Null Hypotheses(H0) = In last year Fraction of Wisconsin adult residents who consumed alcohol is equal to 70% (p=70).

Alternative Hypotheses(Ha) = In last year Fraction of Wisconsin adult residents who consumed alcohol is not equal to 70% (p ≠ 70).

Null Hypotheses = Alcohol consumption rate among Wisconsin adult resident is same as national rate of 70%

Alternative Hypotheses = Alcohol consumption rate among Wisconsin adult resident is different as national rate of 70%