Unisha Aryal
Assignment 2
Sep 27, 2023

3. 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 = Interaction between GPA and IQ, and X_5 = 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?

Salary = b0 + b1x1 + b2x2 + b3x3 + b4x4 + b5x5 = 50 + 20x1 + 0.07x2 + 35x3 + 0.01x4 - 10x5

for fixed IQ and GPA at x1 and x2: - Salary (high school) = 50 + 20x1 + 0.07x2 + 35(0) + 0.01(x1.x2) -10(x1.0) = 50 + 20x1 + 0.07x2 + 0.01(x1.x2)*

Salary (college) = 50 + 20x1 + 0.07x2 + 35(1) + 0.01(x1.x2) -10(x1.1) = 50 + 20x1 + 0.07x2 + 35 + 0.01(x1.x2) - 10(x1) = Salary (high school) + 35 - 10(x1)*

From here:

Salary (college) - Salary (high school) = 35 - 10x1

Assuming the salary difference to be more than equal to zero, we get:

35 - 10x1 >= 0 → x1 <= 3.5

Assuming the salary difference to be less than equal to zero, we get:

35 - 10x1 <= 0 → x1 >= 3.5

Hence, for a fixed value of IQ and GPA, high school graduates earn more, on average, than college graduates provided that the GPA is more than equal to 3.5.

The correct answer is " iii ".

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

Ans: Salary = 50+20(4)+0.07(110)+35+0.01(110x4)-10(4) = 137.1

Hence, the predicted salary would be $137,100.

(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.

Ans: This statement is false because the magnitude of coefficient is not an indicator of statistical significance.

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

→ Check python file as well for detailed answer to these questions.

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

```
In [5]: Carseats = load_data("Carseats")
        Carseats.columns

Out[5]: Index(['Sales', 'CompPrice', 'Income', 'Advertising', 'Population', 'Price',
               'ShelveLoc', 'Age', 'Education', 'Urban', 'US'],
              dtype='object')

In [15]: import patsy
         f = 'Sales ~ Price + Urban + US'
         y, X = patsy.dmatrices(f, Carseats, return_type='dataframe')

         model = sm.OLS(y, X).fit()
         print(model.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:                        18:04:40   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.Yes]    -0.0219      0.272     -0.081      0.936      -0.556       0.512
US[T.Yes]        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.
==============================================================================

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

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

- On average the unit sales in urban location are 21.9 units less than in rural location all other predictors remaining fixed.
- *A store located in the US sells, on average, 1200 more car seats than a store situated abroad.*
- *When the price rises by $1000, and all other variables remain unchanged, the sales figures decrease by 54.5 units. In simpler terms, a $1000 increase in price leads to a reduction in car seat sales by 54.5 units.*

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

$Sales = 13.0435 + (-0.0545) \times Price + (-0.0219) \times Urban + (1.2006) \times US + \varepsilon$

with Urban=1 if the store is in an urban location and 0 if not, and US=1 if the store is in the US and 0 if not.

(d) For which of the predictors can you reject the null hypothesis $H_0: \beta_j = 0$?

We can reject the null hypothesis for "Price" and "US" variables.

(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.

```
In [16]: f = 'Sales ~ Price + US'
         y, X = patsy.dmatrices(f, Carseats, return_type='dataframe')

         model2 = sm.OLS(y, X).fit()
         print(model2.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:                        18:05:15   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.Yes]      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.
==============================================================================

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

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

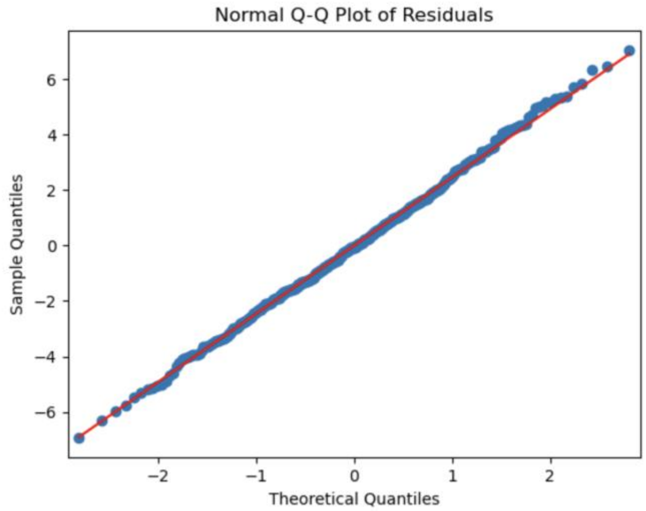
R - squared and Adjusted R squared values are same for both the models. But the smaller model has a higher f -squared value that means smaller model (e) is a better fit compared to (a).

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

```
In [11]: confidence_intervals = results2.conf_int(alpha=0.05)

In [12]: print(confidence_intervals)

                   0          1
intercept   11.79032  14.271265
Price       -0.06476  -0.044195
US[Yes]      0.69152   1.707766
```

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

```
In [67]: # 10 (h)
residuals = results2.resid

sm.qqplot(residuals, line='s')
plt.title('Normal Q-Q Plot of Residuals')
plt.show()
```



Answer: Considering that most of the residuals align closely with the diagonal line, it indicates that they exhibit an approximate normal distribution. While there are a few outliers represented by minor deviations from the diagonal line, these outliers do not appear to pose significant issues.

14. This problem focuses on the *collinearity* problem.

(a) Perform the following commands in Python:

```
rng = np.random.default_rng(10)
x1 = rng.uniform(0, 1, size=100)
x2 = 0.5 * x1 + rng.normal(size=100) / 10
y = 2 + 2 * x1 + 0.3 * x2 + rng.normal(size=100)
```

The last line corresponds to creating a linear model in which y is a function of x1 and x2. Write out the form of the linear model. What are the regression coefficients?

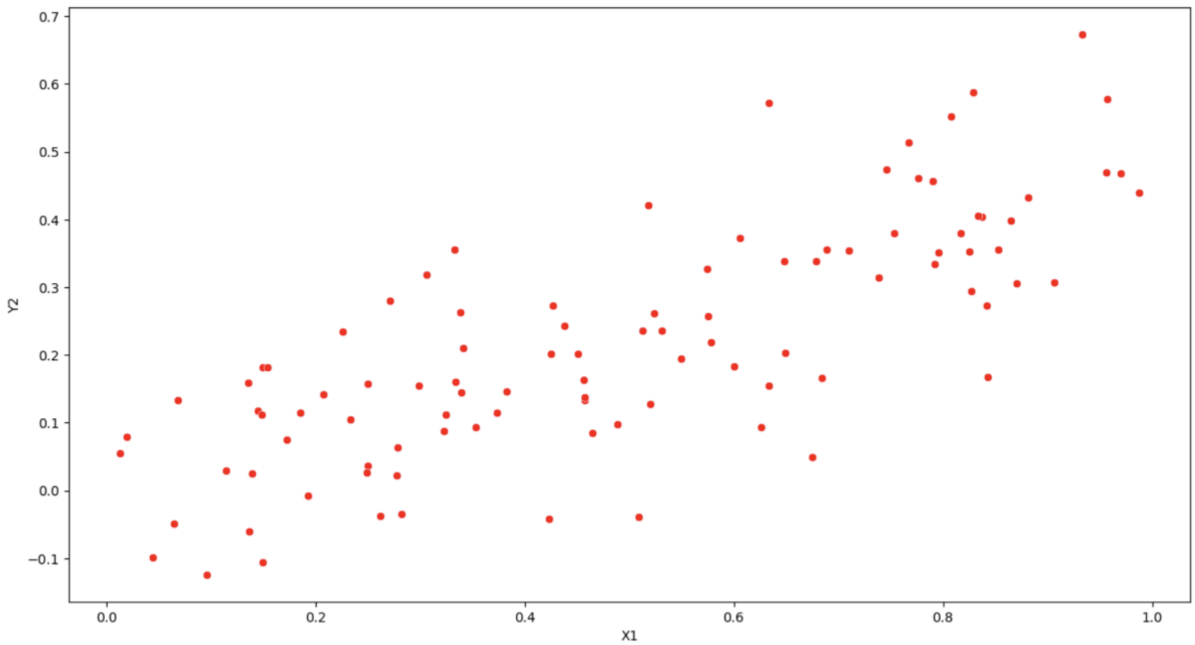
The form of the linear model is:

$y=\beta 0+\beta 1 \cdot x1+\beta 2 \cdot x2+\epsilon$

$\varepsilon \sim N(0,1)$ *random variable. The regression coefficients are respectively 2, 2 and 0.3.*

(b) What is the correlation between x1 and x2? Create a scatterplot displaying the relationship between the variables.

In [28]:
```
fig = plt.figure(figsize=(15,8))
ax = fig.add_subplot(111)
ax = sns.scatterplot(x= x1, y= x2, color='r')

ax.set_xlabel("X1")
ax.set_ylabel("Y2")

plt.show()

print("Correlation coefficient: " + str(np.corrcoef(x1, x2)[0][1]))
```




```
Correlation coefficient: 0.772324497691354
```

The correlation between x1 and x2 is 0.77.

(c) Using this data, fit a least squares regression to predict y using x1 and x2. Describe the results obtained. What are $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\beta}_2$? How do these relate to the true β_0, β_1, and β_2? Can you reject the null hypothesis $H_0 : \beta_1 = 0$? How about the null hypothesis $H_0 : \beta_2 = 0$?

```
In [31]: X = np.stack((x1, x2), axis=-1)
         X = sm.add_constant(X, prepend=True)

         model = sm.OLS(y, X)
         result = model.fit()
         print(result.summary())
```

```
                            OLS Regression Results                            
==============================================================================
Dep. Variable:                      y   R-squared:                       0.291
Model:                            OLS   Adj. R-squared:                  0.276
Method:                 Least Squares   F-statistic:                     19.89
Date:                Wed, 27 Sep 2023   Prob (F-statistic):           5.76e-08
Time:                        18:41:26   Log-Likelihood:                -130.62
No. Observations:                 100   AIC:                             267.2
Df Residuals:                      97   BIC:                             275.1
Df Model:                           2                                         
Covariance Type:            nonrobust                                         
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          1.9579      0.190     10.319      0.000       1.581       2.334
x1             1.6154      0.527      3.065      0.003       0.569       2.661
x2             0.9428      0.831      1.134      0.259      -0.707       2.592
==============================================================================
Omnibus:                        0.051   Durbin-Watson:                   1.964
Prob(Omnibus):                  0.975   Jarque-Bera (JB):                0.041
Skew:                          -0.036   Prob(JB):                        0.979
Kurtosis:                       2.931   Cond. No.                         11.9
==============================================================================

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

The coefficients $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\beta}_2$ are respectively 1.9579, 1.6154 and 0.9428. As the p-value is less than 0.05, we may reject null hypothesis for β_0 and β_1, however we may not reject H_0 for β_2 as the p-value is higher than 0.05.

(d) Now fit a least squares regression to predict y using only x1. Comment on your results. Can you reject the null hypothesis $H_0 : \beta_1 = 0$?

```
In [38]: X = sm.add_constant(x1, prepend=True)

         model2 = sm.OLS(y, X)
         result2 = model2.fit()
         print(result2.summary())
```

```
                            OLS Regression Results                            
==============================================================================
Dep. Variable:                      y   R-squared:                       0.281
Model:                            OLS   Adj. R-squared:                  0.274
Method:                 Least Squares   F-statistic:                     38.39
Date:                Wed, 27 Sep 2023   Prob (F-statistic):           1.37e-08
Time:                        19:04:23   Log-Likelihood:                -131.28
No. Observations:                 100   AIC:                             266.6
Df Residuals:                      98   BIC:                             271.8
Df Model:                           1                                         
Covariance Type:            nonrobust                                         
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          1.9371      0.189     10.242      0.000       1.562       2.312
x1             2.0771      0.335      6.196      0.000       1.412       2.742
==============================================================================
Omnibus:                        0.204   Durbin-Watson:                   1.931
Prob(Omnibus):                  0.903   Jarque-Bera (JB):                0.042
Skew:                          -0.046   Prob(JB):                        0.979
Kurtosis:                       3.038   Cond. No.                         4.65
==============================================================================

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

The coefficient for "x1" in this model is different from the one with "x1" and "x2" as predictors. In this case "x1" is highly significant as its p-value is very low, so we may reject $H_0 : \beta_1 = 0$.

(e) Now fit a least squares regression to predict y using only x2. Comment on your results. Can you reject the null hypothesis $H_0 : \beta_1 = 0$?

```
In [43]: X = sm.add_constant(x2, prepend=True)

         model3 = sm.OLS(y, X)
         result3 = model3.fit()
         print(result3.summary())

                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.222
Model:                            OLS   Adj. R-squared:                  0.214
Method:                 Least Squares   F-statistic:                     27.99
Date:                Wed, 27 Sep 2023   Prob (F-statistic):           7.43e-07
Time:                        19:08:43   Log-Likelihood:                -135.24
No. Observations:                 100   AIC:                             274.5
Df Residuals:                      98   BIC:                             279.7
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          2.3239      0.154     15.124      0.000       2.019       2.629
x1             2.9103      0.550      5.291      0.000       1.819       4.002
==============================================================================
Omnibus:                        0.191   Durbin-Watson:                   1.943
Prob(Omnibus):                  0.909   Jarque-Bera (JB):                0.373
Skew:                          -0.034   Prob(JB):                        0.830
Kurtosis:                       2.709   Cond. No.                         6.11
==============================================================================

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

The coefficient for "x2" in this model is different from the one with "x1" and "x2" as predictors. The value for $\hat{\beta}_1$ is 2.9103 . In this case "x2" is highly significant as its p-value is very low, so we may again reject $H_0 : \beta_1 = 0$.

(f) Do the results obtained in (c)–(e) contradict each other? Explain your answer.

The variations in significance levels do not inherently conflict with one another; instead, they emphasize the significance of factoring in the broader context and additional predictors when interpreting the significance of individual predictors within a multiple regression framework.

(g) Suppose we obtain one additional observation, which was unfortunately mismeasured. We use the function np.concatenate() to add this additional observation to each of x1, x2 and y.

```
x1 = np.concatenate([x1, [0.1]])
x2 = np.concatenate([x2, [0.8]])
y = np.concatenate([y, [6]])
```

Re-fit the linear models from (c) to (e) using this new data. What effect does this new observation have on the each of the models? In each model, is this observation an outlier? A high-leverage point? Both? Explain your answers.

→ Check Python file for this answer.

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.

52% of adult Twitter users obtain some of their news from the platform, with a standard error estimate of 2.4%. Given this data, a normal distribution can be applied. The critical value for a 99% confidence interval is approximately 2.575829.

Calculating the confidence interval:

Lower limit: 0.52 - (2.575829)(0.024) = 0.4582

Upper limit: 0.52 + (2.575829)(0.024) = 0.5818

This results in a confidence interval of 0.4582 to 0.5818. The interpretation of the confidence interval is as follows: "We are 99% confident that the fraction of U.S. adult twitter users who get some news on twitter is between 45.82% and 58.18%. The margin of error in this context is 6.18%. This margin of error pertains to the percentage of users who receive news on Twitter and not the margin of error for the difference between the percentage obtaining news on Twitter and those who do not.

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?

(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.

a)

Null Hypothesis (H_0): There is no change in the average calorie intake for diners.

Alternative Hypothesis (H_A): There is a difference in calorie intake for diners.

H_0: $\mu = 1100$
H_A :$\mu \neq 1100$

b)

Null Hypothesis (H_0): The fraction of Wisconsin adults who consume alcohol is equal to national average of 0.7.

Alternative Hypothesis (H_A): The fraction of Wisconsin adults who consume alcohol is different from national average of 0.7.

Let, ‘p’ be the population of Wisconsin’s adult who consumed alcohol past year.

H_0: $p = 0.70$

H_A : $p \neq 0.70$