1. The cost of attending your university has gone up once again. Although you have been told that education is investment in human capital and carries a return of roughly 10% a year, you (and your parents) are not pleased. An administrators at your school argues that you are paying more for your education because the reputation of your institution is better than that of others. To investigate this hypothesis, you collect data on a random sample of 100 national universities and liberal arts colleges from the 2000-2001 U.S. News and World Report annual ranking. Next you perform the following regression

$$\begin{aligned}\widehat{\text{cost}} &= \underset{(2,058.63)}{7,311.17} + \underset{(664.58)}{3,985.20} \cdot Reputation - \underset{(0.13)}{0.20} \cdot Size + \underset{(2,154.85)}{8,406.79} \cdot Dpriv \\ &\quad - \underset{(1,121.92)}{416.38} \cdot Dlibart - \underset{(1,007.86)}{2,376.51} \cdot Dreligion \\ R^2 &= 0.72 \qquad SER = 3,773.35\end{aligned}$$

where *Cost* is Tuition, Fees, Room and Board in dollars, *Reputation* is the index used in U.S. News and World Report (based on a survey of university presidents and chief academic officers), which ranges from 1 ("marginal") to 5 ("distinguished"), *Size* is the number of undergraduate students, and *Dpriv*, *Dlibart*, and *Dreligion* are binary variables indicating whether the institution is private, a liberal arts college, and has a religious affiliation. The numbers in parentheses are heteroskedasticity-robust standard errors.

(a) Interpret the results and determine whether or not the coefficients are significantly different from zero. Do the coefficients have the expected sign?

(b) What is the forecasted cost for a liberal arts college, which has no religious affiliation, a size of 1,500 students and a reputation level of 4.5? (All liberal arts colleges are private.)

(c) Suppose that you switch from a private university to a public university, which has a ranking that's 0.5 lower and 10,000 more students. What is the effect on your cost?

(d) What is the p-value for the null hypothesis that the coefficient on *Size* is equal to zero? Based on this, should you eliminate the variable from the regression? Why or why not?

(a) increase the cost by roughly $3985.20 (Significant at 1%, t=6)

The coefficient on school size is -0.20 which means that for each additional student the cost is reduced by 20 cents. when we have 1000 students, -0.20×1000 = -200 which mean the lowers the cost by $200 (Not Significant at 5%, t=1.54)

Private schools charge roughly $8406 more than public school (Significant at 1%, t=3.9)

a liberal art college charges roughly $416 less (Not Significant at 5%, t=0.37)

a religious affiliation is $2376 less, (Significant at 5%, t=2.36)

We can know that "Reputation, Dpriv, Dreligion", these variables have a significant impact on "costs", but "Size, Dliart show that a weak or not significant impact on "costs"

(b) We can know the reputation = 4.5, Size = 1500 students, Dpriv = 1, Dliart = 1, Dreligion = 0

cost = 7311.17 + 3985.20 × 4.5 − 0.2 × 1500 + 8406.79 − 416.38

= 7311.17 + 17933.40 − 300 + 8406.79 − 416.38 = 32934.38

So that approximately $32935.

(c) We can know that change in Dpriv = −1, change in Reputation = −0.5, change in size = 10000, So that the effect on cost = −8406.79 − 0.5 × 3985.20 − 0.2 × 10000 = −8406.79 − 1992.6 − 2000 = −12399.39 hence it's decrease of approximately 12399.39.

(d) $t = \frac{-0.20}{0.13} \approx -1.54$, we can't find a p-value greater than 0.05 and it's not significant.

(e) Eliminating the *Size* and *Dlibart* variables from your regression, the estimated regression becomes

$$\widehat{\text{cost}} = \underset{(1,772.35)}{5,450.35} + \underset{(590.49)}{3,538.84} \cdot Reputation + \underset{(875.51)}{10,935.70} \cdot Dpriv - \underset{(1,180.57)}{2,783.31} \cdot Dreligion$$

$$R^2 = 0.72, \qquad SER = 3,792.68$$

Why do you think that the effect of attending a private institution has increased now?

(f) You make one final attempt to bring the effect of Size back into the equation by forcing the assumption of homoskedasticity onto your estimation. The results are as follows:

$$\widehat{\text{cost}} = \underset{(1,985.17)}{7,311.17} + \underset{(593.65)}{3,985.20} \cdot Reputation - \underset{(0.07)}{0.20} \cdot Size + \underset{(1,423.59)}{8,406.79} \cdot Dpriv - \underset{(1,096.49)}{416.38} \cdot Dlibart - \underset{(989.23)}{2,376.51} \cdot Dreligion$$

$$R^2 = 0.72, \qquad SER = 3,682.02$$

Calculate the *t*-statistic on the *Size* coefficient and perform the hypothesis test that its coefficient is zero. Is this test reliable? Explain.

(g) What can you say about causation in the above relationship? Is it possible that *Cost* affects *Reputation* rather than the other way around?

(e) Size and Dlibart are both negative coefficients. and they are excluded which mean increase in the cofficients for private school. We can see that correlation between "private↑ and ↓size" is negative and "cost↑ and ↓size" is negative, and it's can let private institution has increase.

(f) $t = \frac{-0.20}{0.07} \approx -2.86$, $|t| > 1.96$, so that it's at 95%, and we absolute value of the calculated "t" is greater than the critical value of the t-distribution like (-2 or 2), we will reject the null hypothesis that the coefficient is zero. The t-statistic is so large, if the sample size is large enough and the assumptions of the t-test. and it's show that a coefficient that is significantly different from zero. hence this is not possible to state that this is reliable.

(g) Regression analysis suggests an association and it's not causation Higher costs may lead to or require higher quality, which in turn may enhance reputation.

2. **Empirical exercise, to be solved using R (or Stata, Python, etc).** In this exercise you will investigate the relationship between housing prices and the physical characteristics of a home. On the course website you will find a data file hprice.xls (in Excel format), collected from the real estate pages of the Boston Globe in 1990 (these are homes selling in the Boston, MA area). It contains data on the selling price (*price*) of the house (in $1000), the size ($sqrft$) of the house in square feet, the number of bedrooms (*bdrms*), the size of the lot (*lotsize*) in square feet, and a dummy variable (*colonial*) which is equal to 1 if the home was colonial style. Use these data to answer the following questions. In all your regressions, please include an intercept term (i.e. β_0). **Make sure to include your output and code with your homework!**

   (a) Run a regression of selling price (*price*) on the size of the home ($sqrft$) and the number of bedrooms (*bdrms*) and report your results.

   (b) What is the estimated increase in price for a house with one more bedroom, holding square footage constant?

   (c) What is the estimated increase in price for a house with an additional bedroom that is 140 square feet in size? Compare this to your result in part b).

   (d) What percentage of the variation in price is explained by square footage and the number of bedrooms?

   (e) The first house in the sample has $sqrft = 2,438$ and $bdrms = 4$. Find the predicted selling price for this house from the OLS regression line.

   (f) The actual selling price of the first house in the sample was $300,000 (so $price = 300$). Find the residual for this house. Does it suggest that the buyer underpaid or overpaid for the house?

   (g) Suppose that instead of the regression described in part a), you ran a regression of *price* on $sqrft$ (i.e. you left out bedrooms). Would you expect this regression to suffer from an omitted variables problem? If so, in what direction would you expect the coefficient on $sqrft$ to be biased? Now run this regression and see if you were right.

   (h) Now, suppose you included *lotsize* in the original regression. Will you have a problem with perfect multicollinearity? Why or why not? Run this regression (if you can) and interpret the coefficient on *lotsize*.

1a) We can know that size of the home and the number of bedrooms results in a equation of the form:

price = −16.4743 + 0.1281 x Sqrft + 14.4144 x bdrms.

We can see that, for each additional square foot of the home size and increase by $128.10 because the selling price of the house in $1000

1b) So that increase in price for a house with one more bedroom, it's $14414.40

1c) We can calculate it

price = 14.4144 x 1 + 0.1281 x 140 = 32.3484

1d) The R-squared value of the model is 0.628, and it's approximately 62.8%

1e) sqrft = 2438, bdrms = 4

price = −16.4743 + 0.1281 x 2438 + 14.4144 x 4 = 353.4911 and it's predicted price

1f) We know that the Actual price = 300

So that the residual = 300 − 353.4911 from 1e) = −53.4911

1g) We can see that the output result for the regression:

price = 12.4032 + 0.1393 x sqrft

We can see that the "sqrft" is higher than coefficient of 0.1281 and the R-squared value is 0.618, approximately 61.8% and it's less than the 62.8%

hence the exclusion of bdrms from the model does change the

estimated effect of sqrft, and it's possibly due to variable caused by omitted price

(h) we can see that data, the regression is price = -20.2873 + 0.1218 x sqrft + 13.7342 x bdrms + 0.0021 x lotsize

sqrft: for each additional square foot of living space, increase $121.80

bdrms: additional one bedrooms, and it's approximately $13734.20

lotsize: additional for each square foot of lotsize, increase approximately $2.10

R-squard: 0.67 which mean 67% is a selling prices of the variance.

Multicollinearity: we can see that the "Cond. No." is too large = 6.27e+04 and it's maybe effect the stability of the coefficient estimates.

In [17]:
```
# Create by Kaijie Li by datahub
# run before anything
import pandas as pd
import statsmodels.api as sm
import matplotlib.pyplot as plt
# run the following if is not installed in jupyter
# !pip install pandas xlrd
```

In [18]:
```
data = pd.read_excel('hprice.xls')
print(data.head())
```

```
   price  bdrms  lotsize  sqrft  colonial
0  300.0      4     6126   2438         1
1  370.0      3     9903   2076         1
2  191.0      3     5200   1374         0
3  195.0      3     4600   1448         1
4  373.0      4     6095   2514         1
```

In [19]:
```
# Basic model with
X = sm.add_constant(data[['sqrft', 'bdrms']])
y = data['price']
model = sm.OLS(y, X)
results = model.fit()

print(results.summary())
```

```
                            OLS Regression Results
==============================================================================
==
Dep. Variable:                  price   R-squared:                       0.6
28
Model:                            OLS   Adj. R-squared:                  0.6
21
Method:                 Least Squares   F-statistic:                     82.
83
Date:                Thu, 21 Mar 2024   Prob (F-statistic):           8.69e-
22
Time:                        17:12:14   Log-Likelihood:                -555.
68
No. Observations:                 101   AIC:                             111
7.
Df Residuals:                      98   BIC:                             112
5.
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
==
                 coef    std err          t      P>|t|      [0.025      0.97
5]
------------------------------------------------------------------------------
--
const        -16.4743     28.329     -0.582      0.562     -72.693      39.7
45
sqrft          0.1281      0.013      9.993      0.000       0.103       0.1
54
bdrms         14.4144      8.635      1.669      0.098      -2.722      31.5
51
==============================================================================
==
Omnibus:                       30.321   Durbin-Watson:                   1.8
78
Prob(Omnibus):                  0.000   Jarque-Bera (JB):                62.7
84
Skew:                           1.161   Prob(JB):                     2.33e-
14
Kurtosis:                       6.086   Cond. No.                     9.89e+
03
==============================================================================
==

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is corre
ctly specified.
[2] The condition number is large, 9.89e+03. This might indicate that there
are
strong multicollinearity or other numerical problems.
```

In [20]:
```
# Second model
X_sqrft_only = sm.add_constant(data['sqrft'])
model_sqrft_only = sm.OLS(y, X_sqrft_only).fit()

print(model_sqrft_only.summary())
```

```
                            OLS Regression Results
==============================================================================
==
Dep. Variable:                  price   R-squared:                       0.6
18
Model:                            OLS   Adj. R-squared:                  0.6
14
Method:                 Least Squares   F-statistic:                     16
0.0
Date:                Thu, 21 Mar 2024   Prob (F-statistic):           2.15e-
22
Time:                        17:12:14   Log-Likelihood:                -557.
09
No. Observations:                 101   AIC:                             111
8.
Df Residuals:                      99   BIC:                             112
3.
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
==
                 coef    std err          t      P>|t|      [0.025      0.97
5]
------------------------------------------------------------------------------
--
const         12.4032     22.635      0.548      0.585     -32.510      57.3
17
sqrft          0.1393      0.011     12.649      0.000       0.117       0.1
61
==============================================================================
==
Omnibus:                       31.534   Durbin-Watson:                   1.7
61
Prob(Omnibus):                  0.000   Jarque-Bera (JB):                64.4
39
Skew:                           1.219   Prob(JB):                     1.02e-
14
Kurtosis:                       6.061   Cond. No.                     7.70e+
03
==============================================================================
==

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is corre
ctly specified.
[2] The condition number is large, 7.7e+03. This might indicate that there a
re
strong multicollinearity or other numerical problems.
```

In [21]:
```
# Third model
X = sm.add_constant(data[['sqrft', 'bdrms','lotsize']])
y = data['price']
model = sm.OLS(y, X)
results = model.fit()

print(results.summary())
```

OLS Regression Results

Dep. Variable:	price	R-squared:	0.670
Model:	OLS	Adj. R-squared:	0.660
Method:	Least Squares	F-statistic:	65.69
Date:	Thu, 21 Mar 2024	Prob (F-statistic):	2.83e-23
Time:	17:12:14	Log-Likelihood:	-549.65
No. Observations:	101	AIC:	1107.
Df Residuals:	97	BIC:	1118.
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-20.2873	26.846	-0.756	0.452	-73.570	32.995
sqrft	0.1218	0.012	9.930	0.000	0.097	0.146
bdrms	13.7342	8.179	1.679	0.096	-2.498	29.967
lotsize	0.0021	0.001	3.508	0.001	0.001	0.003

Omnibus:	23.968	Durbin-Watson:	2.128
Prob(Omnibus):	0.000	Jarque-Bera (JB):	43.998
Skew:	0.962	Prob(JB):	2.79e-10
Kurtosis:	5.599	Cond. No.	6.27e+04

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 6.27e+04. This might indicate that there are
strong multicollinearity or other numerical problems.