ps4_nocode

November 21, 2023

1 BA 222 PS4

Ethan Chang

- a) Where is King County? Use the zip codes if you are unsure.
  - It is in Washington
- b) How many observations are in the dataset? What does 1 row correspond to?
  - There are 21613 observations in the dataset
  - Each row corresponds to a house in the county
- c) What are the median statistics for price, bedrooms, bathrooms, square foot of living space, and year built?
  - Price: 450000
  - Bedrooms: 3
  - Bathrooms: 2.25
  - Square Foot of Living Space: 1910
  - Year Built: 1975
- d) Run the regression:

$$Price = a + b * Bedrooms$$

  - Write a full sentence explaining the coefficient on bedrooms.
    - The coefficient on bedrooms is about 121716 which means that every additional bedroom corresponds to an increase of house price by about 121716 dollars.
  - Is the coefficient statistically significant? What is the 95% confidence interval on the coefficient on bedrooms? Interpret the interval.
    - The coefficient is statistically significant because the p-value is less than 0.05.
    - The 95% confidence interval on the coefficient on bedrooms is $[1.17 * 10^5, 1.27 * 10^5]$ which means that we are 95% confident that the true coefficient on bedrooms is between 117000 and 127000.
  - If a house has 2 bedrooms, what does the one variable model predict the price will be?
    - Approximately $373234.61
  - Is the relationship between bedrooms and price necessarily causal?
    - The relationship between bedrooms and price is not necessarily causal. There could be other factors that affect the price of a house that are not accounted for in this model.
  - Interpert the R^2 value of this model.

    - The R^2 value of this model is 0.095 which means that 9.5% of the variation in price can be explained by the number of bedrooms.

- e) Run the regression of price on bedrooms and living square footage:

$$\text{Price} = a + b * \text{Bedrooms} + c * \text{Sqft\_living}$$

  - Write a full sentence explaining the coefficient on bedrooms. How has it changed? Why might it have changed?
    - The coefficient of bedrooms is now negative (from $1.298 * 10^5$ to $-5.707 * 10^4$), which corresponds to a decrease in the price of the house when the number of bedrooms increases. The coefficient might have changed because of the addition of a new variable or because more bedrooms could actually decrease the price of a house. In addition to this, there may be a different variable that increases the value of a house.
  - How has the R^2 changed from the first model?
    - The R^2 value has increased from 0.095 to 0.507 which means that 50.7% of the variability in housing prices can be explained by the number of bedrooms and the square footage of living space.
  - What does the model predict for the price of a 2 bedroom, 1000 square foot apartment?
    - 279,284.53
  - What does the model predict for the price of a 3 bedroom, 1000 square foot apartment?
    - 222,217.77

- f) Add dummies for zip code to your second model and run the regression:

$$\text{Price} = a + b * \text{Bedrooms} + c * \text{Sqft\_living} + d * \text{Zip}$$

  You should have 70 zip dummies. You do not need to interpret them, just include them.

  - What is the R^2 of this model? Write a full sentence.
    - The R^2 value of this model is 0.738 which means that 73.8% of the variation in price can be explained by the number of bedrooms, square footage of living space, and zip code.
  - What is the coefficient on bedrooms? How does it compare to the other models? Is it statistically significant?
    - The coefficient on bedrooms is $-4.471 * 10^4$ which is greater than the previous model. It is statistically significant because the p-value is less than 0.05.
  - Suppose we wanted to use this model to make a casual statement about the effect of bedrooms. Write a full sentence about the assumption we would have to make.
    - We would have to assume that all other variables (such as square footage and zip-code) are controlled and as good as randomly assigned to assume causality.

- g) Run one more model to evaluate the effect of bedrooms on price, picking some other variable(s) for controls. What variables did you include? Write the full estimating equation, and include a screenshot of your results. What coefficient for bedrooms do you find?

  - I included square footage of living space, bathrooms, and number of floors as controls, while still keeping bedrooms. The full estimating equation is:
    - $\text{Price} = 7.467 * 10^4 - 5.785 * 10^4 * \text{Bedrooms} + 309.3932 * \text{Sqft\_living} + 7853.5216 * \text{Bathrooms} + 200.4972 * \text{Floors}$

– The coefficient for bedrooms is $-5.785 * 10^4$ which is close to the coefficient in the first model, but smaller than the coefficient in the second model.

[1]:
```
# This is needed when I do a `.ipynb` to pdf conversion, also ignore my misspelling of regression in the file name

from IPython.display import Image

Image(filename='regressiom.png')
```

[1]:

OLS Regression Results

Dep. Variable:	price	R-squared:	0.507
Model:	OLS	Adj. R-squared:	0.507
Method:	Least Squares	F-statistic:	5554.
Date:	Fri, 17 Nov 2023	Prob (F-statistic):	0.00
Time:	21:39:18	Log-Likelihood:	-2.9996e+05
No. Observations:	21613	AIC:	5.999e+05
Df Residuals:	21608	BIC:	6.000e+05
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	7.467e+04	7679.122	9.724	0.000	5.96e+04	8.97e+04
bedrooms	-5.785e+04	2347.323	-24.644	0.000	-6.24e+04	-5.32e+04
sqft_living	309.3932	3.087	100.228	0.000	303.343	315.444
bathrooms	7853.5216	3814.223	2.059	0.040	377.363	1.53e+04
floors	200.4972	3775.505	0.053	0.958	-7199.772	7600.766

Omnibus:	14450.413	Durbin-Watson:	1.985
Prob(Omnibus):	0.000	Jarque-Bera (JB):	494760.943
Skew:	2.739	Prob(JB):	0.00
Kurtosis:	25.790	Cond. No.	1.04e+04

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.04e+04. This might indicate that there are strong multicollinearity or other numerical problems.