ØAMET2200
Business Decision Making Using Data

Lecture 9

Instructor: Fenella Carpena

November 1, 2019

Announcements

- Problem Set 3 due on November 3 at 11:59PM
- Problem set solutions are always posted on Canvas under the "Modules" section

Part 4 of this Course: Building regression models

- You want to build a model to forecast outcomes from data you have
- What's the **best model** you can build?
- Today: using regressions to make comparisons across different categories

Agenda for Today

- Building a regression model for **two (or more)** groups
- Interpreting models with **categorical variables** and **interactions**
- Chapter 25

Categorical Variables

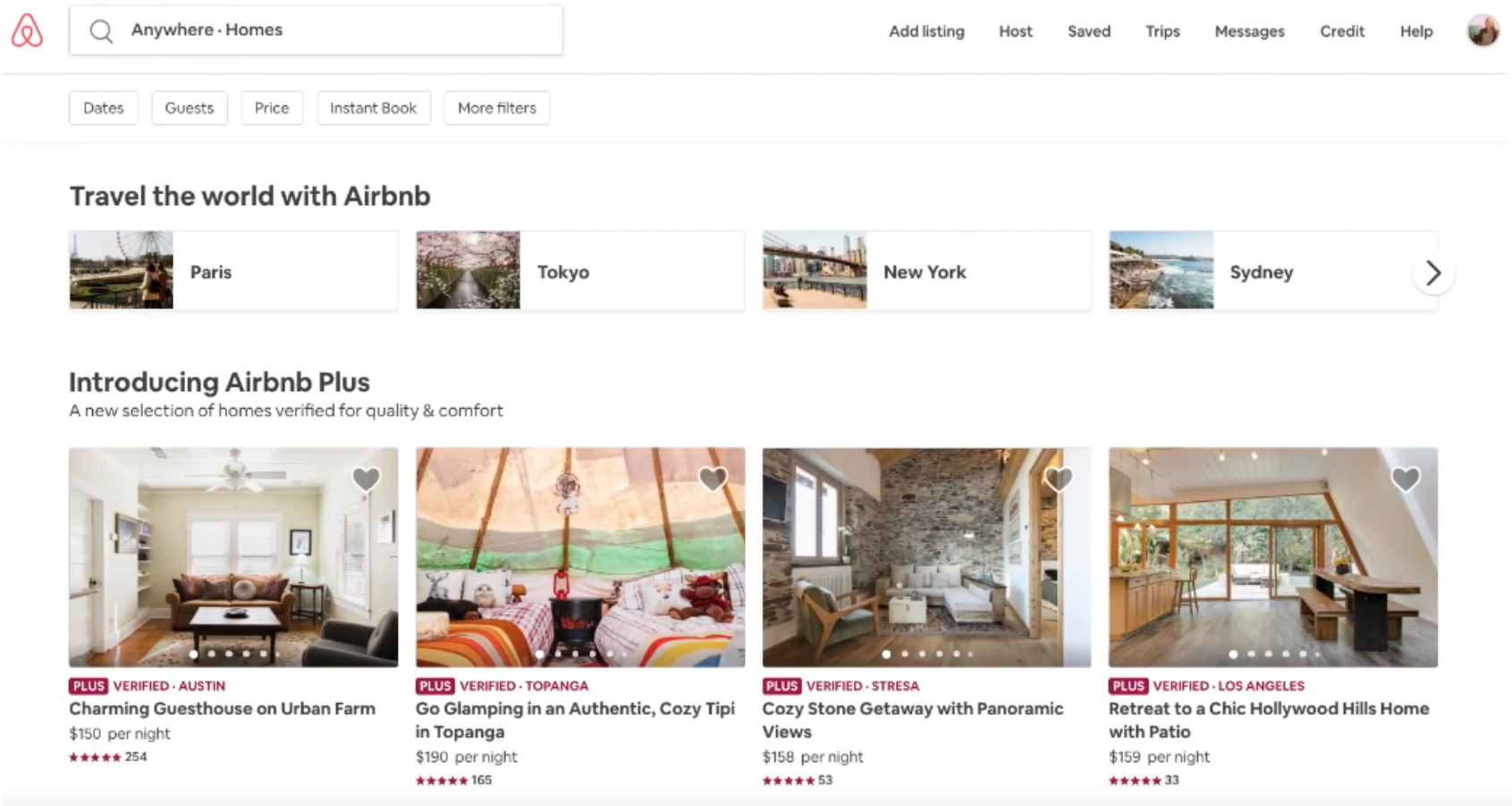
- A **categorical** variable is a variable that can take on a fixed number of possible values
  - The values assign observations to a particular group or category
- Examples of categorical variables
  - Whether a person has a car (yes/no)
  - County where a person lives
- Categorical variables can be represented in a dataset as numeric variables

Why do we care about categorical variables?

1. They allow us to compare groups while controlling for confounding variables
   - Example: you are interested in whether students who have an iPad have higher grades than those who do not
   - But is the difference in grades due to having an iPad?
2. They allow us to test whether a variable we care about has different effects on different groups
   - Example: you want to know the effect of providing extra vacation days on the performance of your employees
   - Does the effect differ depending on whether the employee has kids?

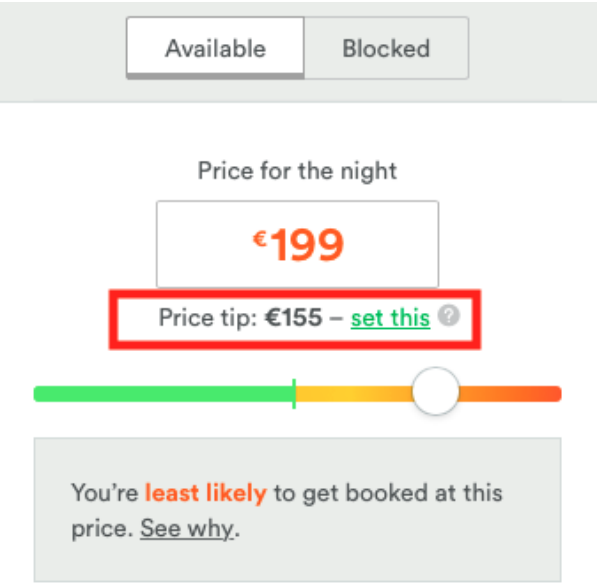
Case Study for Today: Airbnb

Airbnb is an online marketplace for sharing homes and experiences



Case Study for Today: Airbnb

- Airbnb connects hosts with spare rooms to share with travelers/guests who are looking for a place to stay
- Airbnb does not control how hosts set their price, but provides a "Price Tip" tool to help hosts set prices more effectively



- How are these price suggestions generated?

Case Study for Today: Airbnb

Airbnb develops **regression models** to generate price recommendations for hosts

Applied Data Science Track Paper KDD 2018, August 19-23, 2018, London, United Kingdom

Customized Regression Model for Airbnb Dynamic Pricing

Peng Ye*
Airbnb Inc.
San Francisco, CA
peng.ye@airbnb.com

Julian Qian*
Ant Financial
San Mateo, CA
j.qian@antfin.com

Jieying Chen
Airbnb Inc.
San Francisco, CA
jieying.chen@airbnb.com

Chen-hung Wu
Airbnb Inc.
San Francisco, CA
chen-hung.wu@airbnb.com

Yitong Zhou
Airbnb Inc.
San Francisco, CA
yitong.zhou@airbnb.com

Spencer De Mars
Impira Inc.
San Francisco, CA
spencer@impira.com

Frank Yang
Airbnb Inc.
San Francisco, CA
frank.yang@airbnb.com

Li Zhang
Airbnb Inc.
San Francisco, CA
li.zhang@airbnb.com

ABSTRACT

This paper describes the pricing strategy model deployed at Airbnb, an online marketplace for sharing home and experience. The goal of price optimization is to help hosts who share their homes on Airbnb set the optimal price for their listings. In contrast to conventional pricing problems, where pricing strategies are applied to a large quantity of identical products, there are no "identical" products on Airbnb, because each listing on our platform offers unique values and experiences to our guests. The unique nature of Airbnb listings makes it very difficult to estimate an accurate demand curve that's required to apply conventional revenue maximization pricing strategies.

KEYWORDS

Price Optimization, Customized Regression Model, Dynamic Pricing

ACM Reference Format:

Peng Ye, Julian Qian, Jieying Chen, Chen-hung Wu, Yitong Zhou, Spencer De Mars, Frank Yang, and Li Zhang. 2018. Customized Regression Model for Airbnb Dynamic Pricing. In *KDD '18: The 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, August 19–23, 2018, London, United Kingdom.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3219819.3219830

Case Study for Today: Airbnb

- Using OLS, we want to develop a regression model to describe listing prices
- For simplicity, we'll assume that all MRM conditions hold, but should always be checked!
- Suppose that the marketing department at Airbnb has asked us to analyze the following questions:
  1. Do superhosts charge lower prices?
  2. Is the effect of an additional bedroom on listing prices different between superhosts and non-superhosts?

Case Study for Today: Airbnb

- We have data on 8,407 Airbnb listings in Oslo
- Taken from `insideairbnb.com/get-the-data.html`, scraped from the Airbnb website

	listing_id	host_id	host_is_superhost	property_type	bedrooms	price
1	42932	187463	f	Apartment	2	1925
2	43198	4011871	t	Apartment	1	381
3	43431	189712	f	Apartment	1	472
4	69964	175633	f	Apartment	1	999
5	71725	368229	f	Apartment	1	499
6	77108	412523	f	Apartment	3	1889

Airbnb: Do superhosts charge lower prices?

- We can answer this using a two-sample t-test! (Lecture 4)
- Let's review how that works
- What is the null and alternative hypothesis of interest?

- What do we need to know to implement this test?

Airbnb: Do superhosts charge lower prices?

Summary statistics for the `price` variable by superhost status

```
.  summarize price if host_is_superhost == "t" ;
```

Variable	Obs	Mean	Std. Dev.	Min	Max
price	1114	807.342	488.7631	254	3496

```
. summarize price if host_is_superhost == "f" ;
```

Variable	Obs	Mean	Std. Dev.	Min	Max
price	7292	861.0676	478.6939	254	3796

Airbnb: Do superhosts charge lower prices?

- Standard error of the difference between the two means
- t-statistic
- Critical value (assume 5% significance level)
- Conclusion of hypothesis test

Airbnb: Do superhosts charge lower prices?

```
. ttest price, by(host_is_superhost) unequal ;

Two-sample t test with unequal variances
```

Group	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf.	Interval]
f	7292	861.0676	5.605761	478.6939	850.0787	872.0565
t	1114	807.342	14.64387	488.7631	778.6093	836.0747
combined	8406	853.9477	5.239249	480.3566	843.6774	864.2179
diff		53.7256	15.68016		22.96752	84.48367

```
    diff = mean(f) - mean(t)                                      t =   3.4263
Ho: diff = 0                     Satterthwaite's degrees of freedom =  1458.32

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.9997         Pr(|T| > |t|) = 0.0006          Pr(T > t) = 0.0003
```

The Menace of Confounding Variables

- Without an experiment or A/B test, we must be careful about **confounding variables** that could account for the significant difference in average prices between superhosts and non-superhosts
- **Number of bedrooms** is a confounding variable if
  - The two groups (superhosts and non-superhosts) differ in the number of bedrooms they offer in their listings
  - Number of bedrooms impacts price
- Do superhosts charge lower prices than non-superhosts, **after controlling for** (holding constant) the number of bedrooms?

One Solution: Look at Subsets of Data

- Restrict the analysis to a subset of listings with **matching** levels of the confounding variable (number of bedrooms)
- For example, we could compare superhost and non-superhost listings with exactly 2 bedrooms

```
. ttest price if bedrooms == 2, by(host_is_superhost) unequal ;

Two-sample t test with unequal variances
```

Group	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf.	Interval]
f	1836	1047.869	11.2689	482.8561	1025.768	1069.97
t	243	1079.794	32.64973	508.9589	1015.48	1144.108
combined	2079	1051.601	10.65796	485.9609	1030.699	1072.502
diff		-31.92496	34.53973		-99.8935	36.04358

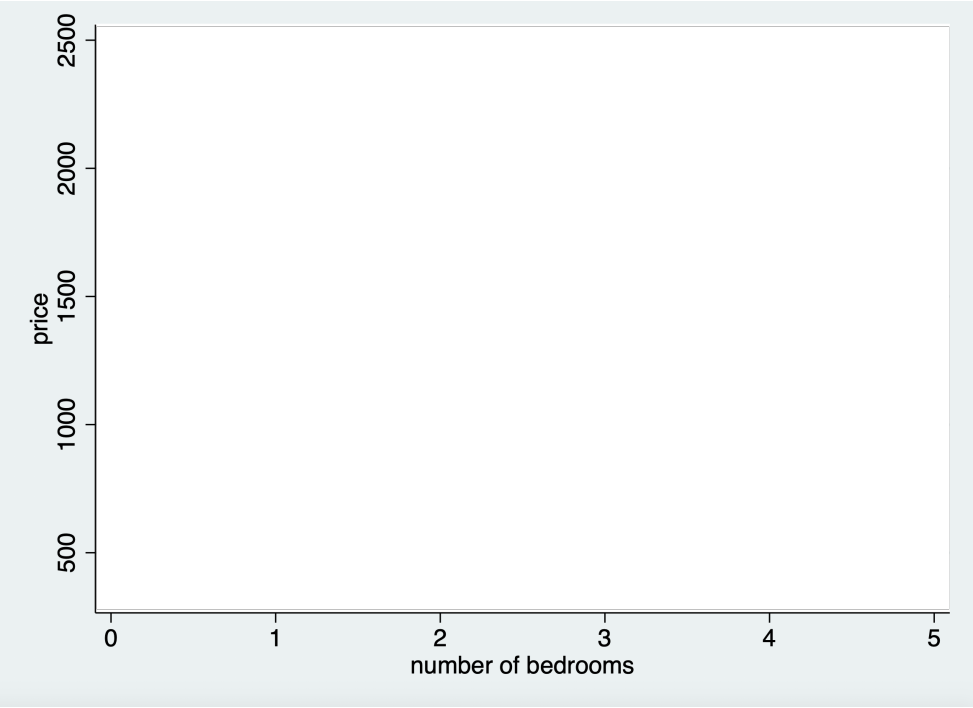
```
    diff = mean(f) - mean(t)                                      t =  -0.9243
Ho: diff = 0                     Satterthwaite's degrees of freedom =  302.525

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.1780         Pr(|T| > |t|) = 0.3561          Pr(T > t) = 0.8220
```

One Solution: Look at Subsets of Data

- By restricting the sample to listings with 2 bedrooms, we eliminate bedrooms as a confounder
- But it reduces our sample size
  - We get less precise estimates: SE increases, CI is very wide
  - We fail to reject null, but this is based on a weak "signal" so it is not satisfactory
- What about the difference between listings with 1 or 3 bedrooms?
- Regression analysis helps us overcome these issues

Two separate regressions

- Estimate separate SRMs for superhosts & non-superhosts



- This allow us to use **all of the data** to compare prices between superhosts and non-superhosts at any # of bedrooms
- For a given bedroom #, the difference in predicted price gives us an estimate of the difference in the population average price

Two separate regressions

```
. regress price bedrooms if host_is_superhost == "t" ;

      Source |       SS       df       MS              Number of obs =    1114
-------------+------------------------------           F(  1,  1112) =  554.47
       Model |  88465025.4     1  88465025.4           Prob > F      =  0.0000
    Residual |   177418791  1112  159549.273           R-squared     =  0.3327
-------------+------------------------------           Adj R-squared =  0.3321
       Total |   265883817  1113  238889.323           Root MSE      =  399.44

------------------------------------------------------------------------------
       price |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    bedrooms |   368.0675   15.63108    23.55   0.000     337.3977    398.7372
       _cons |   326.2772   23.67701    13.78   0.000     279.8205    372.7338
------------------------------------------------------------------------------

. regress price bedrooms if host_is_superhost == "f" ;

      Source |       SS       df       MS              Number of obs =    7292
-------------+------------------------------           F(  1,  7290) = 2683.45
       Model |   449522579     1   449522579           Prob > F      =  0.0000
    Residual |  1.2212e+09  7290  167516.418           R-squared     =  0.2691
-------------+------------------------------           Adj R-squared =  0.2690
       Total |  1.6707e+09  7291   229147.89           Root MSE      =  409.29

------------------------------------------------------------------------------
       price |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    bedrooms |   285.7752   5.516676    51.80   0.000     274.9609    296.5895
       _cons |   452.1175    9.23555    48.95   0.000     434.0131    470.2218
------------------------------------------------------------------------------
```

Combining the separate SRMs

- Instead of separate SRMs, a better approach is to estimate a simple SRM that **combines** categorical and numeric x's
  - Allows us to conduct hypothesis test for differences across groups (controlling for other x's)
- We'll need two additional variables

1. **Dummy variable**, equal to 1 for observations that belong to some group, equal to 0 otherwise
   - Airbnb example: $\text{superhost} = \begin{cases} 1 & \text{host is a super host} \\ 0 & \text{host is NOT a superhost} \end{cases}$
   - Stata code:

Combining the separate SRMs

2. **Interactions** between dummy variable & explanatory variable
   - Airbnb example: superhost $*$ bedrooms
   - Stata code:

- How do these additional variables look like in the data?

	listing_id	host_is_superhost	bedrooms	price
1	42932	f	2	1925
2	43198	t	1	381
3	43431	f	1	472
4	69964	f	1	999
5	71725	f	1	499
6	77108	f	3	1889

- Use these variables to estimate the population regression

$$price = \beta_0 + \beta_1 superhost + \beta_2 bedrooms + \beta_3 superhost * bedrooms + \epsilon$$

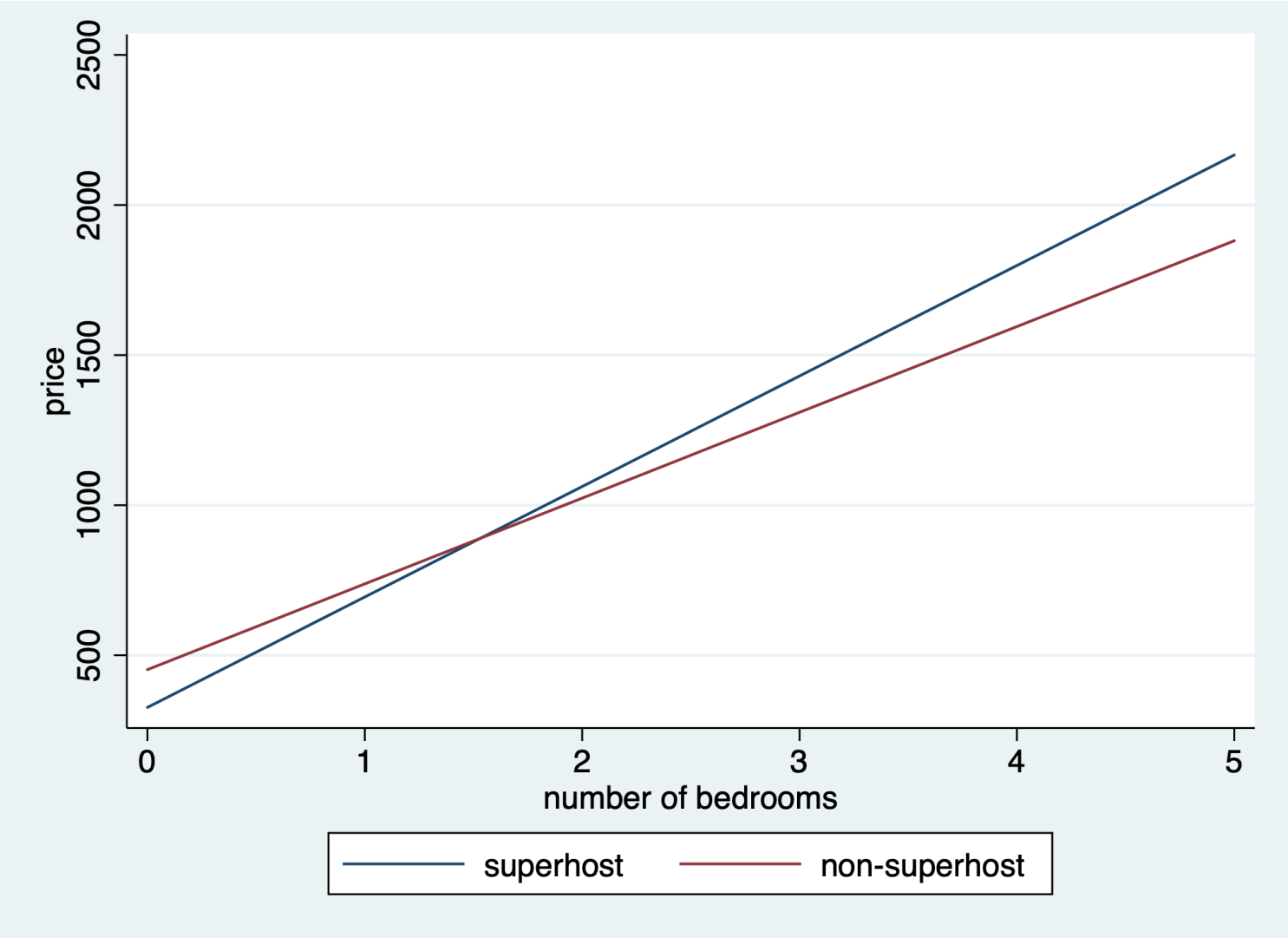
Combining the separate SRMs

```
. regress price superhost bedrooms superhostXbedrooms ;

      Source |       SS           df       MS      Number of obs   =     8407
-------------+----------------------------------   F(3, 8403)      =  1082.78
       Model |  540695868         3  180231956   Prob > F        =   0.0000
    Residual | 1.3987e+09      8403  166453.001   R-squared       =   0.2788
-------------+----------------------------------   Adj R-squared   =   0.2785
       Total | 1.9394e+09      8406  230716.207   Root MSE        =   407.99

------------------------------------------------------------------------------
       price |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   superhost |  -125.9939   25.87603    -4.87   0.000    -176.7173   -75.27051
    bedrooms |   285.6968   5.498116    51.96   0.000     274.9191    296.4744
superhostXbedrooms |   82.37067   16.88586     4.88   0.000     49.27023    115.4711
       _cons |   452.2711   9.203846    49.14   0.000     434.2293    470.3129
------------------------------------------------------------------------------
```

Sample regression equation:

Combining the separate SRMs

$$\widehat{price} = 452 - 126 \cdot superhost + 286 \cdot bedrooms + 82 superhost * bedrooms$$



Combining the separate SRMs

$$\widehat{price} = 452 - 126 \cdot superhost + 286 \cdot bedrooms + 82 superhost * bedrooms$$

- What is the predicted price for superhosts?

- Look familiar? This is the same result we got when we estimated the SRM with superhosts only (slide 20)

Combining the separate SRMs

$$\widehat{price} = 452 - 126 \cdot superhost + 286 \cdot bedrooms + 82 superhost * bedrooms$$

- What is the predicted price for non-superhosts?

- Look familiar? This is the same result we got when we estimated the SRM with non-superhosts only (slide 20)
- **Key point:** The MRM with dummy and interaction variables combines the two separate regressions

Interpretation of coefficients on dummy and interaction

$$\widehat{price} = 452 - 126 \cdot superhost + 286 \cdot bedrooms + 82 superhost * bedrooms$$

- The estimate corresponding to `superhost` is the difference between the estimated intercepts in the SRM
- The estimate corresponding to `superhost*bedrooms` is the difference between the estimated slopes in the SRM

Airbnb Case Study

- The two questions we wanted to analyze for this case study:
  1. Do superhosts charge lower prices?
  2. Is the effect of an additional bedroom on listing prices different between superhosts and non-superhosts?
- The coefficient on `superhost` and `superhost*bedrooms` are both statistically significant at the 1% level
- We conclude that
  - With studio listings (zero bedrooms), superhosts charge lower prices on average
  - With every additional bedroom, prices increases faster for superhosts than non-superhosts
- Some caution is necessary because of lurking variables (e.g., number of bathrooms, square meter size, etc.)

Model Building

- When should we include or exclude interactions in our regression?
- Use **principle of marginality**
- If the interaction is statistically significant, retain it as well as both of its components regardless of their level of significance.
- If the interaction is not statistically significant, remove it from the regression and re-estimate the equation.

Model Building: Airbnb Example

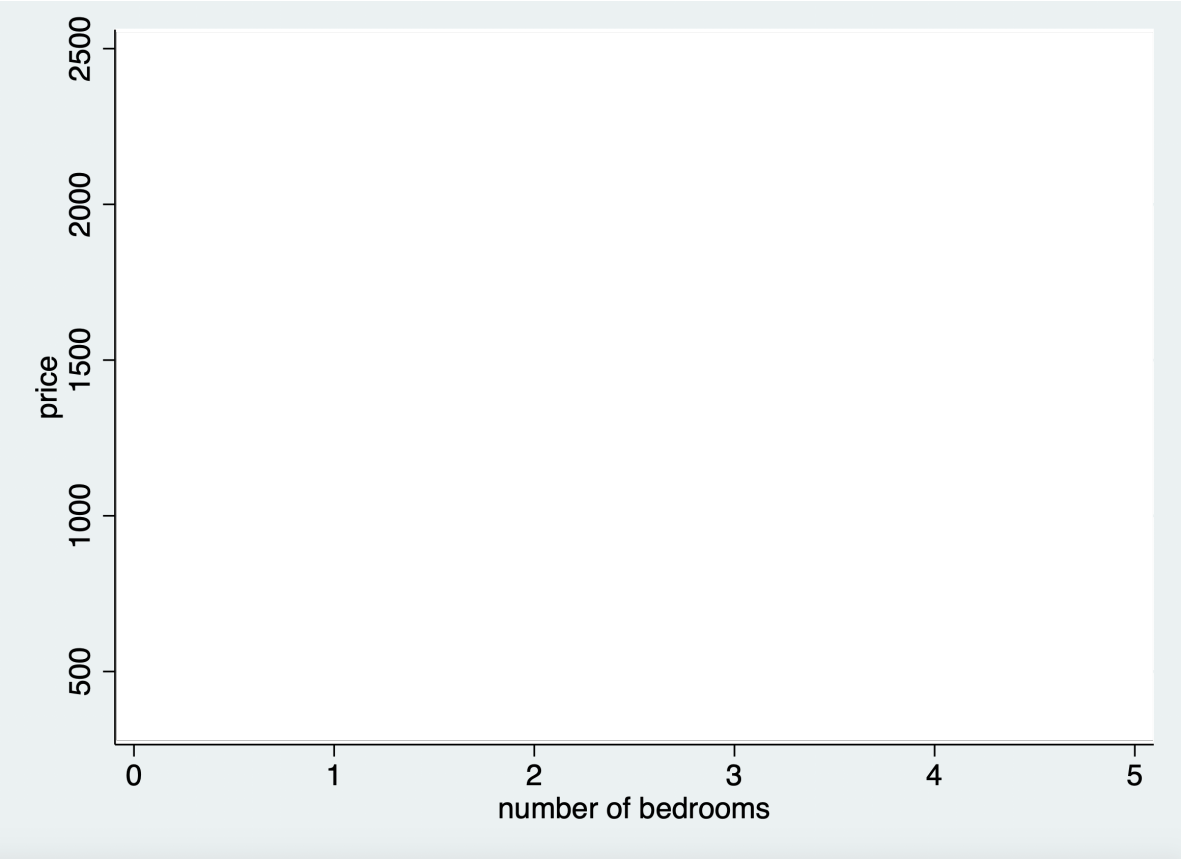
As an example, suppose we drop the interaction variable

```
. regress price superhost bedrooms ;

      Source |       SS       df       MS              Number of obs =    8407
-------------+------------------------------           F(  2,  8404) = 1607.91
       Model |  536734999     2  268367499           Prob > F      =  0.0000
    Residual | 1.4027e+09  8404  166904.502           R-squared     =  0.2768
-------------+------------------------------           Adj R-squared =  0.2766
       Total | 1.9394e+09  8406  230716.207           Root MSE      =  408.54

------------------------------------------------------------------------------
       price |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   superhost |  -17.25397   13.15772    -1.31   0.190    -43.04634    8.538396
    bedrooms |   294.4296   5.205547    56.56   0.000     284.2254    304.6337
       _cons |    439.776   8.852208    49.68   0.000     422.4235    457.1285
------------------------------------------------------------------------------
```

Sample regression equation:

Model Building: Airbnb Example

$$\widehat{price} = 440 - 17 \cdot superhost + 294 \cdot bedrooms$$



A model without an interaction term is simpler to interpret (but less flexible) since the lines fit to the two groups are **parallel**

What if we have 3 or more categories?

- In this example, the x variable has two categories: superhost and non-superhost
- Using similar techniques we can easily include categorical variables with 3+ categories
- Choosing one to be the "baseline" category, we would then include dummy variables for all other categories
- As before, the coefficients are interpreted relative to the baseline category

Airbnb example: 3+ categories

- Suppose we want to examine how prices vary by property type (3 categories: Apartment, House, Other)

listing_id	price	bedrooms	host_is_su~t	property_type
42932	1925	2	f	Apartment
43198	381	1	t	Apartment
213353	717	1	f	Other
256459	445	2	t	House

- In general, to distinguish J groups, we need to include $J - 1$ dummy variables in our regression

- For this example, use "Apartment" as the baseline category

Airbnb example: 3+ categories

```
. regress price bedrooms house other houseXbedrooms otherXbedrooms ;
```

Source	SS	df	MS
Model	541309967	5	108261993
Residual	1.3981e+09	8401	166419.53
Total	1.9394e+09	8406	230716.207

Number of obs	=	8407
F(5, 8401)	=	650.54
Prob > F	=	0.0000
R-squared	=	0.2791
Adj R-squared	=	0.2787
Root MSE	=	407.95

price	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
bedrooms	280.8157	6.976491	40.25	0.000	267.1401	294.4914
house	-47.21291	46.5031	-1.02	0.310	-138.3704	43.94463
other	39.91979	23.58616	1.69	0.091	-6.314897	86.15447
houseXbedrooms	39.4387	16.71368	2.36	0.018	6.675771	72.20162
otherXbedrooms	12.67637	13.3091	0.95	0.341	-13.41275	38.76548
_cons	444.9114	10.40237	42.77	0.000	424.5202	465.3027

Sample regression equation:

Airbnb example: 3+ categories

$$\widehat{price} = 445 + 281 \cdot bedrooms - 47 \cdot house + 40 \cdot other$$
$$+39 \cdot house * bedrooms + 13 \cdot other * bedrooms$$

- Estimated equation for property type = "Apartment"

- Estimated equation for property type = "House"

Airbnb example: 3+ categories

- -47 is the difference in the estimated intercept between "Apartments" and "House"
- 39 is the difference in the estimated slope between "Apartments" and "House"
- The interpretation of the estimates is similar to the interpretation of models with two groups

Best Practices

- Always think about the possibility of lurking variables: control for these in your cross-group comparisons (or understand how they affect estimates)
- Use interaction terms to explore different relationships between x and y across groups
- If interactions arent statistically significant, drop them from the regression
- If interactions are significant, include the individual variables even if they arent significant
  - Never include an interaction variable without its two components
- Be careful interpreting estimates on dummy variables and interactions
  - Write out the regression equations for separate groups

Big Picture Takeaways

- Categorical variables are very common in data analysis. Become comfortable with them!
  - Experiments: treatment/control
  - Allow models to become much more flexible (better prediction!)
  - Different intercepts/slopes for different groups
- Interaction terms are very useful important for testing differences in responses and/or means across groups
  - Does drinking wine lead to more/less risk of heart disease based on ethnicity?