ØAMET2200
Business Decision Making Using Data
Lecture 7

Instructor: Fenella Carpena

October 18, 2019

Announcements

- Nobel Prize was awarded last Monday to Esther Duflo, Abhijit Banerjee, and Michael Kremer "for their experimental approach to alleviating global poverty"
- Midsemester evaluation results
- OH today are cancelled
- Next problem set will be posted this weekend

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?
- This lecture: models with more than one explanatory variable
- Next lecture: building better models

Agenda for Today

- The Multiple Regression Model
- Interpreting Multiple Regression
- Checking Conditions
- Inference in Multiple Regression
- Steps in Fitting a Regression Model
- Chapter 23

What's the end goal for today?

- Understanding how to think about models with multiple causal factors
- You should know how to think through relationships among variables
- You should be able to interpret statistical output describing these relationships

Why is this important?

- Super important for dealing with lurking variables
- Super important for actually understanding what determines the outcome you care about
- Super important for making accurate predictions
- Having one explanatory variable is nice, but very limited

Case Study for Today: Starbucks

- Last lecture, we talked about the business decision of selecting the location of a new store
- There are multiple factors we might care about when making this decision
  - We might care about the affluence of the local population
  - We might also care about the number of competitors
- We address these questions using **Multiple Regression**
- Instead of relating y to one x (simple regression), we consider multiple x's
- By including several x's, we can obtain more accurate predictions

The Multiple Regression Model (MRM) for the Population

- The MRM is similar to the SRM
- But we now have several x's (i.e., explanatory variables)
- In the MRM, the observed response y is linearly related to k explanatory variables by the equation

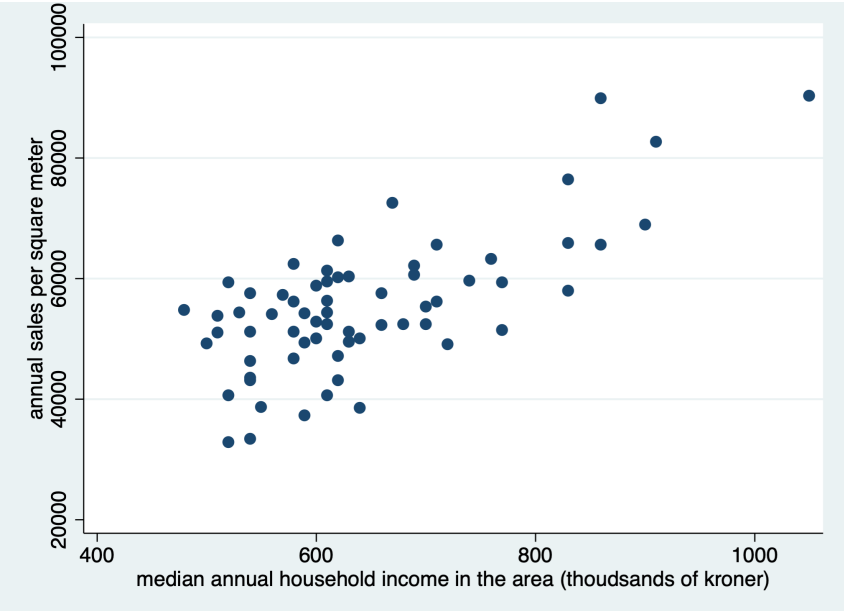
$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2)$$

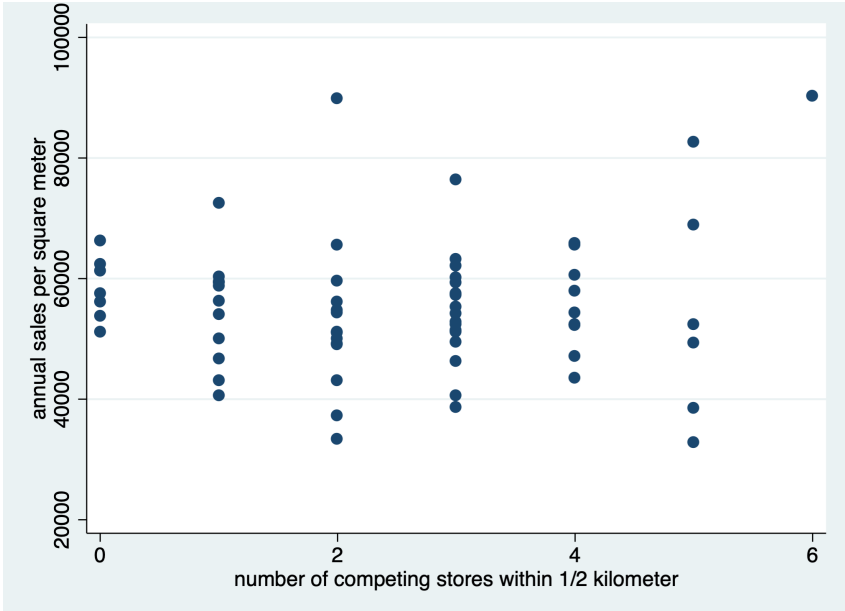
- The errors ϵ in the model are
  - independent of each other
  - have equal variance σ_ϵ^2 around the regression line
  - normally distributed
- The assumptions about the errors match those of the SRM
- SRM is a special case of MRM with $k = 1$

Starbucks: What goes into our model?

- We have (fictional) data from a sample of 65 stores
- Response variable
  - `salespersqm`: annual sales at a given Starbucks store per square meter of retail space
- Two explanatory variables
  - `medianincome`: Median annual household income in the area (measured in thousands of kroner)
  - `numcompetitors`: Number of competing stores within 1/2 kilometer

Starbucks: Scatterplots of y vs. each x





Always look at your sample data!

Starbucks: Suppose we regress sales on median income

```
. regress sales medianincome ;
```

Source	SS	df	MS
Model	4.1121e+09	1	4.1121e+09
Residual	4.0915e+09	63	64944479
Total	8.2036e+09	64	128181872

Number of obs =	65
F(1, 63) =	63.32
Prob > F =	0.0000
R-squared =	0.5013
Adj R-squared =	0.4933
Root MSE =	8058.8

salespersqm	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
medianincome	69.56166	8.741929	7.96	0.000	52.09231	87.03101
_cons	10448.78	5724.472	1.83	0.073	-990.6605	21888.23

What is the interpretation of b_1?

Starbucks: Suppose we regress sales on # competitors

```
. regress sales numcompetitors ;
```

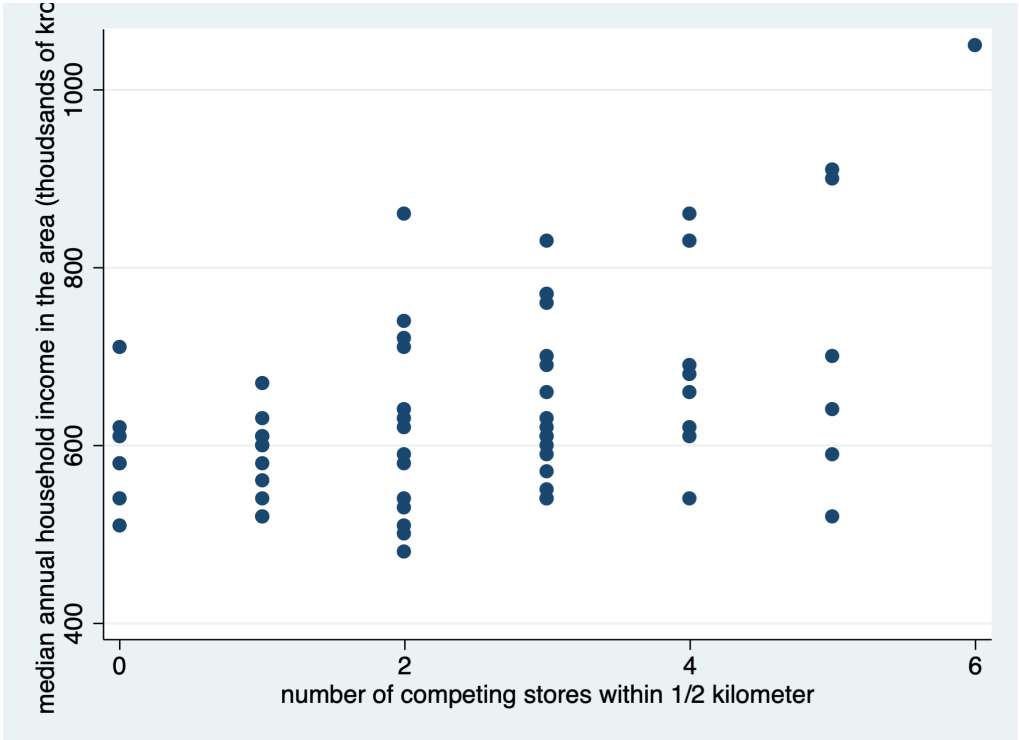
Source	SS	df	MS
Model	36404722.6	1	36404722.6
Residual	8.1672e+09	63	129638652
Total	8.2036e+09	64	128181872

Number of obs	65
F(1, 63)	0.28
Prob > F	0.5980
R-squared	0.0044
Adj R-squared	-0.0114
Root MSE	11386

salespersqm	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
numcompetitors	498.9266	941.5102	0.53	0.598	-1382.531	2380.385
_cons	54056.52	2738.733	19.74	0.000	48583.6	59529.45

What is the interpretation of b_1?

Starbucks: Scatterplot of median income, # competitors




```
. correlate sales medianinc numc ;
(obs=65)
```

	salesp~m	median~e	numcom~s
salespersqm	**1.0000**		
medianincome	**0.7080**	**1.0000**	
numcompeti~s	**0.0666**	**0.4743**	**1.0000**

Starbucks: Multiple Regression

```
. reg sales medianincome numcompetitors ;
```

Source	SS	df	MS
Model	4.8790e+09	2	2.4395e+09
Residual	3.3246e+09	62	53622757.4
Total	8.2036e+09	64	128181872

Number of obs	65
F(2, 62)	45.49
Prob > F	0.0000
R-squared	0.5947
Adj R-squared	0.5817
Root MSE	7322.8

salespersqm	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
medianincome	85.74509	9.022844	9.50	0.000	67.70868	103.7815
numcompetitors	-2601.102	687.8046	-3.78	0.000	-3976.004	-1226.2
_cons	6496.953	5305.549	1.22	0.225	-4108.69	17102.59

What is the sample regression line?

Interpreting Multiple Regression

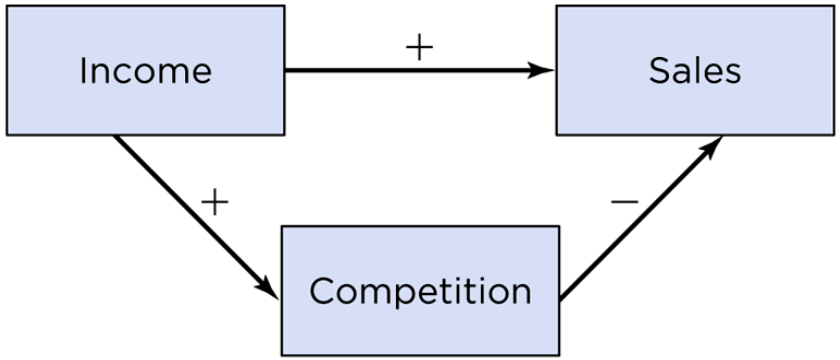
- **Partial Slope**: slope of an explanatory variable in a multiple regression; it statistically **excludes** or **controls for** the effects of other explanatory variables
- **Marginal Slope**: slope of an explanatory variable in a simple regression
- The marginal slope may combine the direct effect with other channels of influence (i.e., confounding variables)
- Partial and marginal slopes are the same only when the explanatory variables are uncorrelated

Starbucks: Interpreting partial slopes

$$\widehat{\text{SalesPerSqM}} = 6496.95 + 85.75\ \text{MedianInc} - 2601.10\ \text{NumCompetitors}$$

- $b_1 = 85.75$ means that a store in a location with 10,000 higher median income has, on average, 857.5 kroner more in sales per square meter than a store in a less affluent location **with the same number of competitors**
- $b_2 = -2601.10$ means that among stores located in areas **with the same median income**, each additional competitor reduces sales per square foot by 2,601.10 kroner on average.
- Useful to think of regression as a "matching" exercise

What's going on?



- A path diagram summarizes the relationship among explanatory variables and the response
- Income has a direct positive effect on sales an indirect negative effect on sales via the number of competitors
- The MRM estimates the direct channel of competition only!
- This is why the partial slope is negative while the marginal slope is positive

Exercise

Suppose that you have a dataset of individual's incomes and demographic information. For each individual, you know yearly income, years of education, and IQ.

You would like to know the "returns to schooling." That is, if years of education increase by one year, how much of an increase in income will it cause?

Consider two different models: (1) simple regression model, in which you regress income on years of education; (2) multiple regression model in which you regress income on years of education and IQ. In which model will you get a larger estimated return to schooling?

Path Diagrams and the Bias of Lurking Variables

- Whenever there is any variable that is omitted from the regression that is correlate with both an x variable or the y variable, our estimates will be **biased**
- Bias may be ok if we are using regression for forecasting/prediction
- However, when we are interested in causal effects, bias is a problem
- Keep this in mind when interpreting regression estimates!

Measures of explanatory power

- The R^2 measures the fraction of variation in y explained by the explanatory variables
  - Starbucks example: What is the R^2? What does this mean?
- The **adjusted** R^2 (denoted $\overline{R}^2$) adjusts for the sample size n and model size k (where k is the # of x's)
  - It is always smaller than R^2
  - $\overline{R}^2 = 1 - (1 - \overline{R}^2)\frac{n-1}{n-k-1}$
- The standard deviation of the residuals s_e is exactly the same as before
  - And you can use this to estimate prediction intervals, as with simple regression

Prediction Intervals

- As in the SRM, an approximate 95% prediction interval is given by

$$\widehat{y} \pm 2s_e$$

- We again need to be careful not to make prediction intervals outside of the range for which we have data

Starbucks: Prediction Intervals

What is predicted sales per square foot at a location with median household income of 700,000 kroner and 3 competitors? What is the 95% prediction interval around it?

MRM Conditions

- As always, conditions must be met for us to make inferences
- The conditions for inference in the MRM are the same as in the SRM
- As with the SRM, we use the residuals from the fitted model to verify that they are:
  - independent
  - have equal variance
  - follow a normal distribution
- Because there are multiple explanatory variables, we plot the residuals against the predicted values $\hat{y}$ as well as against the individual explanatory variables

Review of Residuals

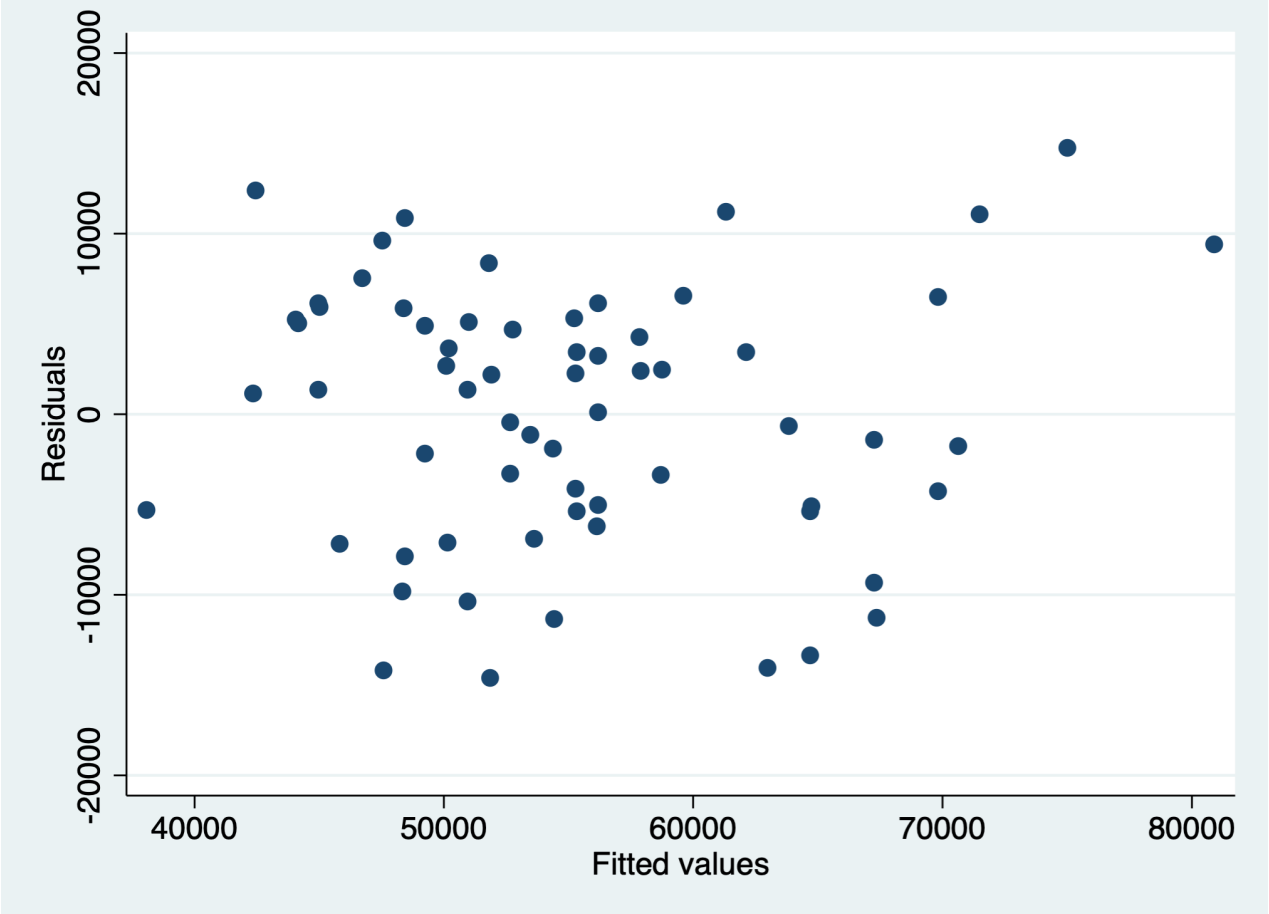
- Recall that residuals are the vertical deviations from the datapoints to the fitted line

- The sample regression in a MRM is

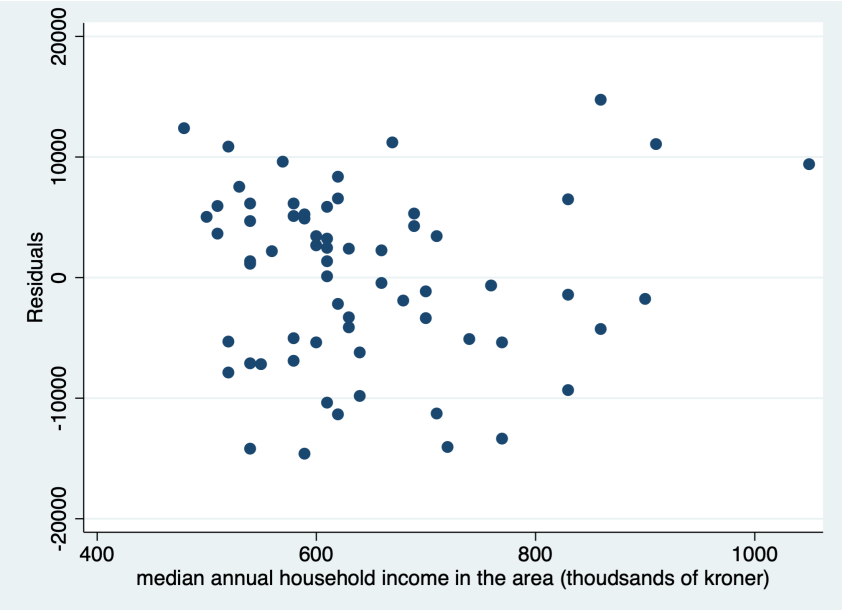
$$\widehat{y} = b_0 + b_1 x_1 + \ldots + b_k x_k$$

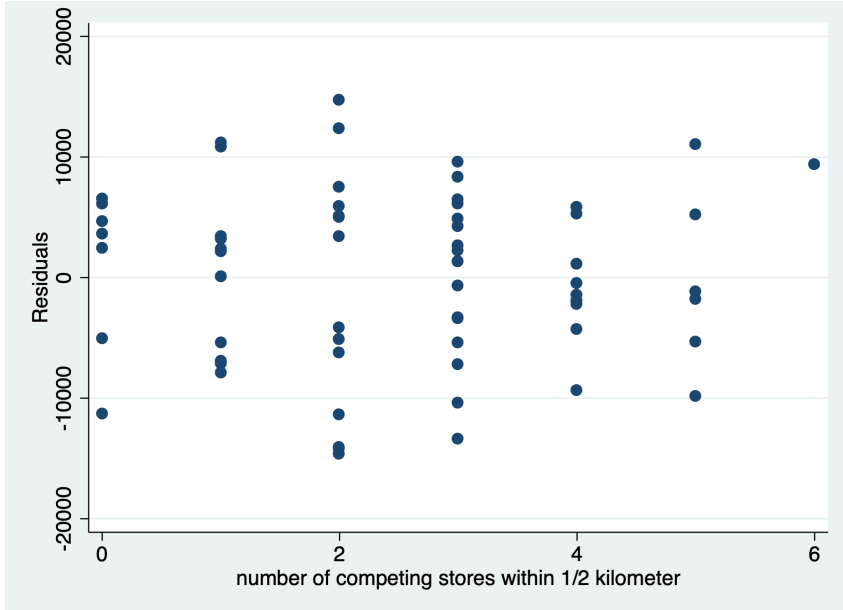
- So it is straightforward to calculate the residual once we have estimated the intercept and slopes

Starbucks: Residual vs. $\widehat{y}$

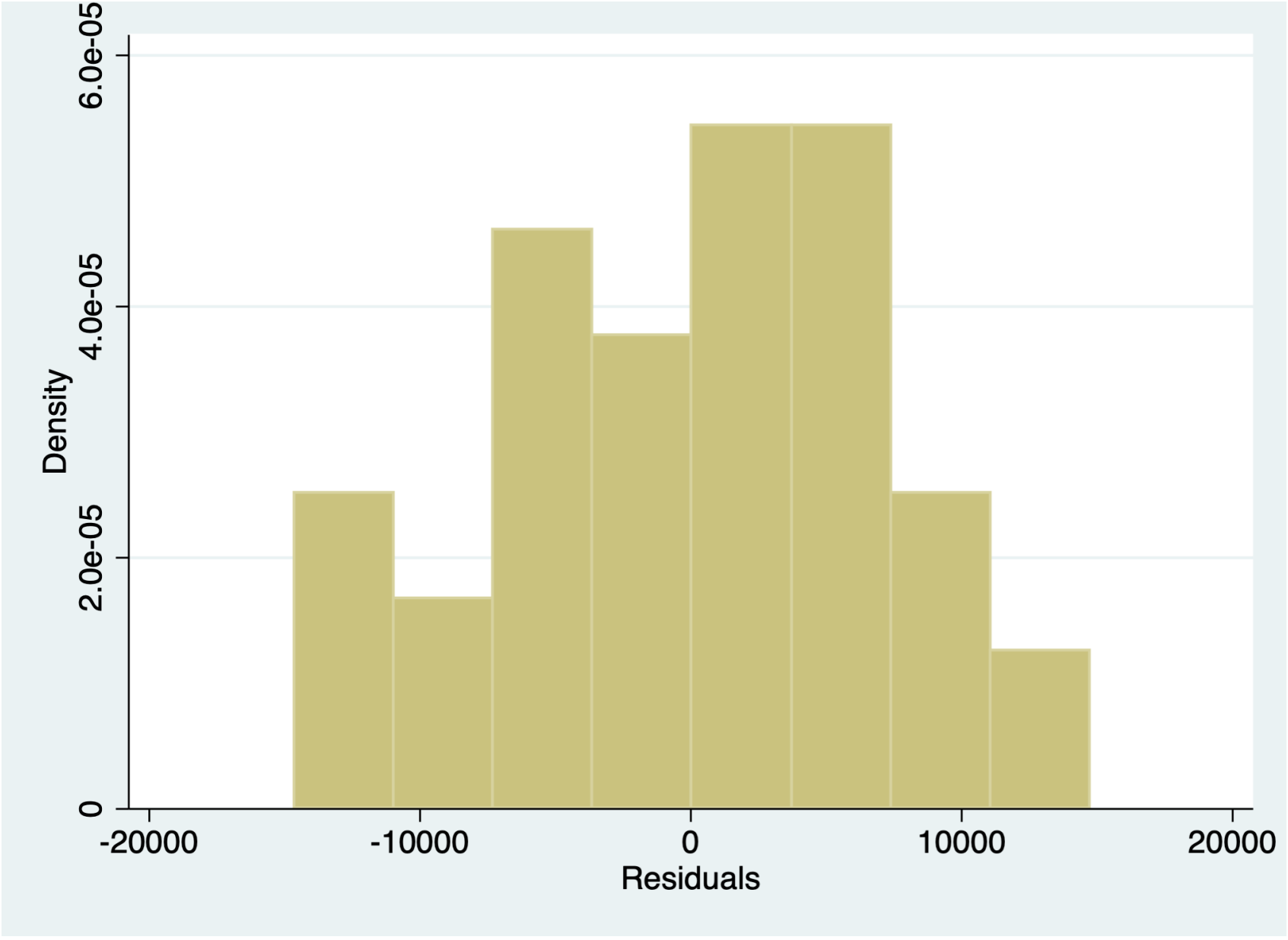


Starbucks: Residual vs. x variables





Starbucks: Histogram of Residuals



Inference in Multiple Regression: F-statistic

- The F-statistic is a measure of the **overall** explanatory power of a MRM

$$F\text{-stat} = \frac{R^2}{1 - R^2} \cdot \frac{n - k - 1}{k}$$

- Used to test the null hypothesis that **all** slopes are equal to zero
- Software automatically reports the F-stat and its p-value
- Is your model weak given the number of explanatory variables?

Starbucks: Multiple Regression

```
. reg sales medianincome numcompetitors ;
```

Source	SS	df	MS
Model	4.8790e+09	2	2.4395e+09
Residual	3.3246e+09	62	53622757.4
Total	8.2036e+09	64	128181872

Number of obs	65
F(2, 62)	45.49
Prob > F	0.0000
R-squared	0.5947
Adj R-squared	0.5817
Root MSE	7322.8

salespersqm	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
medianincome	85.74509	9.022844	9.50	0.000	67.70868	103.7815
numcompetitors	-2601.102	687.8046	-3.78	0.000	-3976.004	-1226.2
_cons	6496.953	5305.549	1.22	0.225	-4108.69	17102.59

- Does the model have explanatory power?
- Is median income a statistically significant determinant of sales per square meter?

Putting this into practice

- How to "read" multiple regression output?
- Check the overall F-statistic before looking at the t-statistics: is there some explanatory power?
- What makes a variable important?
  - **Statistical** significance
  - **Economic** significance
- Keep in mind interpretation of the partial slopes estimated
  - Holding other factors fixed
- Always be careful in making statements about causality
  - Statistical significance or high R^2 doesn't mean causality
  - Controls don't mean there are no other confounding/lurking variables

Steps in Fitting a Multiple Regression

1. What problem are you trying to solve? Do the data help you? Does your data include all the relevant x variables so that there are no confounding factors?

2. Check the scatterplots of y vs. each x. Identify outliers and check for curved patterns

3. If the scatterplots of y on each x appear straight, fit the multiple regression

4. Obtain residuals e and fitted values $\widehat{y}$

5. Make scatterplots of y vs. $\widehat{y}$, e vs. $\widehat{y}$. Check for similar variances condition. If heteroskedastic, avoid prediction intervals

Steps in Fitting a Multiple Regression

6. Check residuals for dependence. If time series data, inspect residuals vs. time and use DW test.

7. Plot residuals vs. each x. Should find no patterns.

8. Check whether residuals are nearly normal. If not, be wary of prediction intervals.

9. Use F-stat to test the null hypothesis that the model has no explanatory power

10. If F-stat is significant, test and interpret individual slopes. If not, proceed with caution of tests of individual coefficients.