07 Simple Regression

Tamer Çetin

Motivation

- Simple non-parametric regression and simple linear regression, and how to visualize their results.
- The regression equation, how its coefficients are uncovered (estimated) in actual data, and we emphasize how to interpret the coefficients.
- The relationship between causation and regression.

When and Why Do We Simple Regression Analysis?

- We focus on the most important pattern of association, mean-dependence.
- Mean-dependence tells whether, and to what extent, the mean of y is different when the value of x is different.
- We will use **simple regression analysis** to uncover mean-dependence between two variables.
- Regression is the most widely used method of comparison in data analysis.
- It compares average values of one variable, called the dependent variable (y) for observations that are different in the other variable, the explanatory variable (x).

When and Why Do We Simple Regression Analysis?

- We use regression to predict y given x:
  - Prediction: Hotel prices based on x features
- We also use regression to uncover the effect of one variable on other variable:
  - Estimation/causal analysis: How would sales change as a result of more or different kinds of advertising?
- Simple regression is a good starting point to answering such questions even if we want to consider more variables.
- But also a major building block for more complicated methods.

Regression: Definition

- Regression analysis is a method that uncovers the average/expected/mean value of a variable y for different values of variable x.
- Regression is a model for the conditional mean, or, in other words, the conditional expectation.
- The conditional mean shows the mean value of y for various values of x.

$$E[y|x] \;=\; f(x)$$

- This should be read as follows:
  - the expected value of y conditional on x is given by the function f.

Regression: Definition

- That function f is the model, which we call regression.

$$y^E = f(x)$$

- y^E stands for $E[y|x]$.
- It is less precise because it does not say what variable the function is conditional on (here x).
- Instead, we should infer this from the context (here the dependence is on x because it equals $f(x)$).

Regression: Definition

- Another notation for the regression is
$$y = f(x) + e$$
- In such a regression eq., the actual value of y is equal to its expected value plus a deviation from it.
- That deviation is called the **error term** of the regression.
- The main difference in notation:

$y^E = f(x)$ --- prediction --- Machine Learning

$y = f(x) + e$ --- causal analysis --- Econometrics

Regression: Definition

- Some more conventions in notation:
- $y^E = f(x)$ is the expression as a general statement of the model:
  - the relationship between the expected value of y and different values of x.
- Alternatively, $y^E = f(x = x_0)$ is the expected value of y if the variable x equals a particular value x_0.
- It shows the result of plugging a specific x value into the model.

Regression: Definition

- Regression analysis finds patterns in the data by comparing observations with different values of the x variable to see whether and how the mean value of the y variable differs across them.
- Thus, we need data with variation in x.
- Usually, the more variation in x, the better because it allows for more comparisons.

Regression: Definition

- Regression analysis may reveal that average y tends to be higher at higher values of x.
- That would be a pattern of positive mean-dependence, or positive association.
- Or, it may reveal a negative association, or negative mean-dependence, with average y being lower at higher values of x.
- However, the pattern of association may be non-monotonic, in which average y tends to be higher for higher values of x in a certain range of the x variable and lower for higher values of x in another range of the x variable.
- Regression analysis may also reveal no association between average y and x, when average y tends to be the same regardless of the value of x.
- The goal of regression analysis is to uncover this pattern and characterize it in a useful way, by visualizing it or summarizing it in one or more numbers.

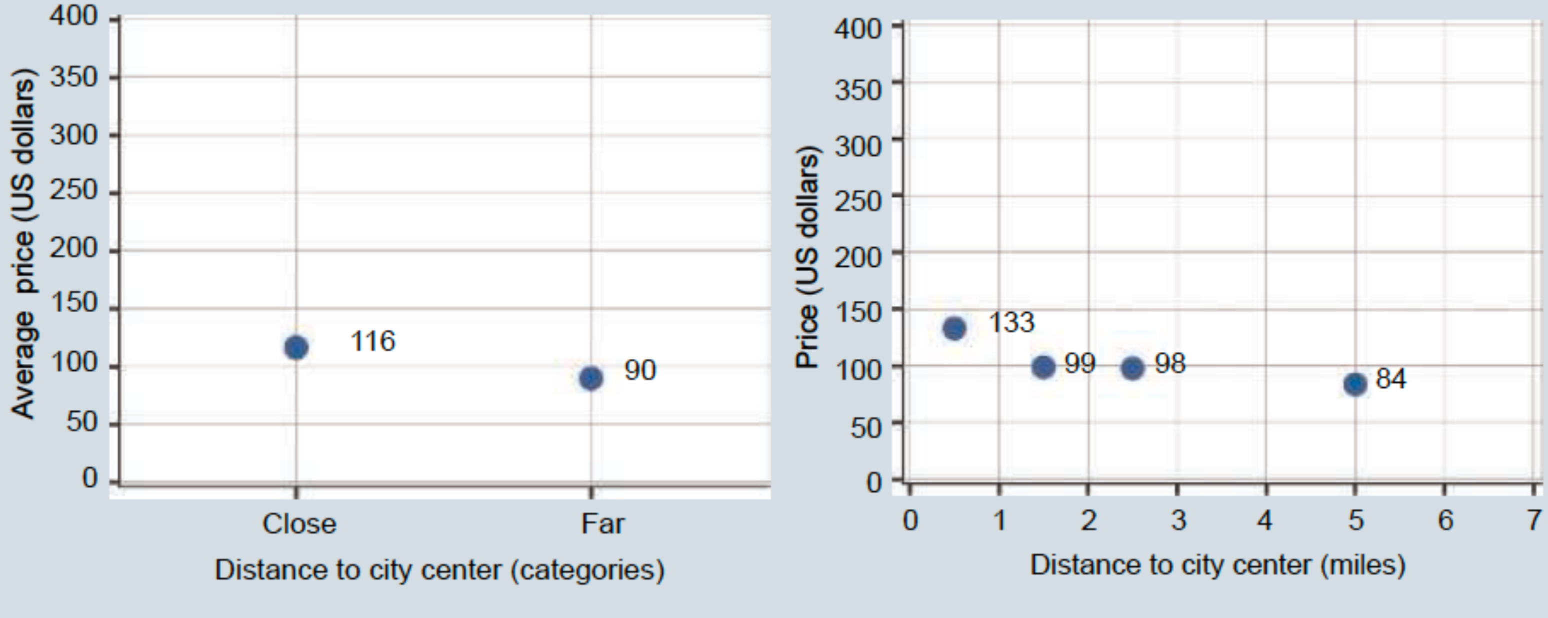
Non-parametric vs parametric

- Non-parametric regressions describe the $y^E = f(x)$ pattern without imposing a specific functional form on f.
- They let the data dictate what that function looks like, at least approximately.
- In contrast, parametric regressions impose a functional form on f.
- Parametric examples include linear functions:
    - $f(x) = a + bx$; exponential functions: $f(x) = ax^b$; quadratic functions: $f(x) = a + bx + cx^2$
- These are called parametric because the functions have parameters a, b, c.
- As we shall see, parametric regressions are restrictive, but they produce readily interpretable numbers.
- In contrast, non-parametric regressions can spot patterns that restrictive parametric functions may miss, but that comes at a price:
    - they do not produce readily interpretable numbers.

Non-parametric vs parametric

- Non-parametric regressions come in various forms.
- When x has few values and there are many observations in the data, the best and most intuitive non-parametric regression for $y^E = f(x)$ shows average y for each and every value of x.
- There is no functional form imposed on f here.
- For example, the hotels in our data have 3 stars, 3.5 stars, or 4 stars.
- If we calculate the average price of hotels with the same numbers of stars and compare these averages across 3, 3.5, and 4 stars, we are carrying out a non-parametric regression analysis.
- With many x values, things become more complicated, especially when the data has few observations for some, or all, x values.
- In such cases, there are two ways to do non-parametric regression analysis:
  - **bins** and **smoothing**.

Non-parametric vs parametric

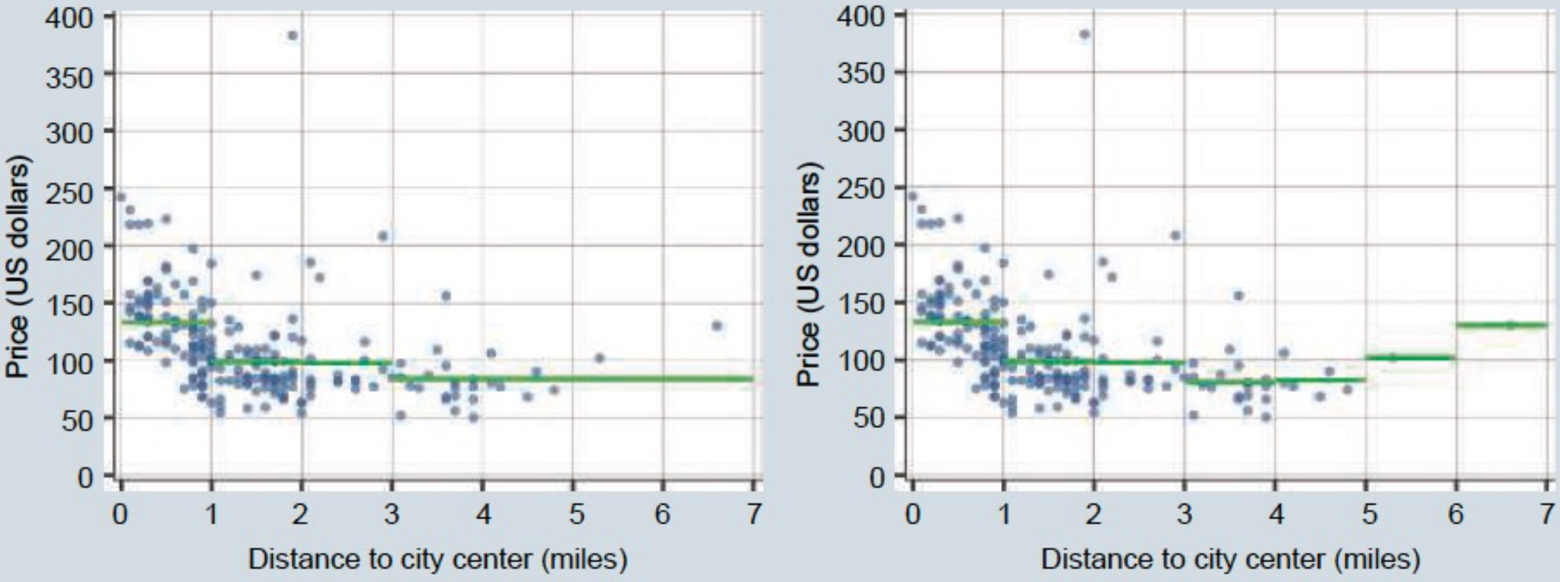


(a) Bin scatter, 2 bins

(b) Bin scatter, 4 bins

Figure 7.1 Hotel price and distance to the city center: two examples of bin scatters

Non-parametric vs parametric



(a) Non-parametric regression, 4 bins

(b) Non-parametric regression, 7 bins

Figure 7.2 Hotel price and distance to the city center: non-parametric regression and scatterplot

Non-parametric vs parametric

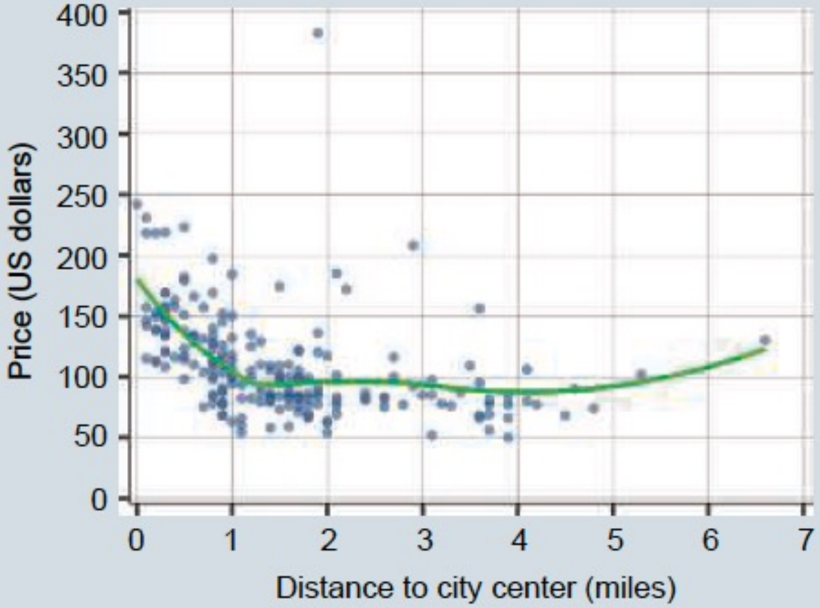


Figure 7.3 Hotel price and distance to the city center: lowess regression and scatterplot

Linear Regression: Introduction

- Linear functions have two parameters, also called coefficients: the **intercept coefficient** and the **slope coefficient**.
$$y^E = \alpha + \beta x$$
- To be more precise, linear regression imposes linearity in terms of its coefficients.
- We can have any function, including any nonlinear function, of the original variables themselves.
- It's still a linear regression as long as it consists of an intercept plus a slope coefficient, whether that slope multiplies x or any function of x (think of logarithm, square, and so on).

Linear Regression: Introduction

- So, the regression function is linear in its coefficients, as an assumption.
- While this can accommodate various **nonlinear patterns**, it cannot capture all kinds of nonlinear patterns.
- The linearity assumption may or may not be true, and, unfortunately, that often cannot be established within the framework of linear regression.
- With y on its left-hand side and x on its right-hand side, linear regression is a line through the $x - y$ scatterplot
- It is the best fit in the sense that it is the line that is closest to all points of the scatterplot.

Linear Regression: Coefficient Interpretation

- Another, more useful way to look at linearity is to treat it as an approximation.
- Whatever the form of the $y^E = f(x)$ relationship, the $y^E = \alpha + \beta x$ regression fits a line through it.
- By fitting a line, linear regression approximates the average slope of the $y^E = f(x)$ curve.
- The average slope has an important interpretation:
  - it is the difference in average y that corresponds to different values of x, averaged across the entire range of x in the data.

Linear Regression: Coefficient Interpretation

- What makes linear regression very powerful is that its coefficients have a clear interpretation based on the idea of comparing conditional means.
- The linear regression $y^E = \alpha + \beta x$ has two coefficients.
- Less important is the **intercept**: α.
- It is the average value of y when x is zero.
- Formally: $E[y|x = 0] = \alpha + \beta \times 0 = \alpha$.

Linear Regression: Coefficient Interpretation

- The more important coefficient is the **slope**: β.
- It shows the expected difference in y corresponding to a one unit difference in x.
- y is higher, on average, by β for observations with a one-unit higher value of x.
- Or, in a longer version that is sometimes more helpful:
  - comparing two observations that differ in x by one unit, we expect y to be β higher for the observation with one unit higher x.
- Formally:

$$E[y|x = x_0 + 1] - E[y|x_0] = (\alpha + \beta \times (x_0 + 1)) - (\alpha + \beta \times x_0) = \beta\ .$$

Linear Regression: Coefficient Interpretation

- Sometimes the slope of linear regression is given more ambitious interpretations.
- One such ambitious interpretation talks about increases or decreases, e.g., saying that "β shows how much y increases, on average, when x is increased."
- This interpretation may be correct if the data is a time series where a change in y is regressed on a change in x.
- However, it is not correct in general.
- In fact, it can be very misleading for cross-sectional data where we are comparing different observations instead of comparing changes for the same observation.
- Another, even more ambitious interpretation is calling the slope the "effect" of x on y.
- Attributing a cause and effect relationship to differences uncovered by regression analysis is a conclusion that may or may not be correct.
- Typically, it is not correct in observational data

Coefficient Formula

$$\hat{\beta} = \frac{Cov[x, y]}{Var[x]} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

- One way to understand the slope coefficient formula is to view it as a normalized version of the covariance between x and y.
- Recall that the covariance measures the degree of association between the two variables, but it returns values that are hard to interpret.
- The slope measures the covariance relative to the variation in x.
- That is why the slope can be interpreted as differences in average y corresponding to differences in x.

Coefficient Formula

- The intercept is calculated after the slope is calculated.
- It is:

$$\hat{a} = \bar{y} - \hat{\beta}\bar{x}$$

- The formula of the intercept reveals that the regression line always goes through the point of average x and average y.
- To see that formally, just rearrange the formula to get $\bar{y} = \hat{a} + \hat{\beta}\bar{x}$.
- In linear regressions, the expected value of y for average x is indeed average y.

Coefficient Formula

- The derivation of the formulae is called ordinary least squares and is abbreviated as OLS.
- The idea underlying OLS is to find the values of the intercept and slope parameters that make the regression line fit the scatterplot best.

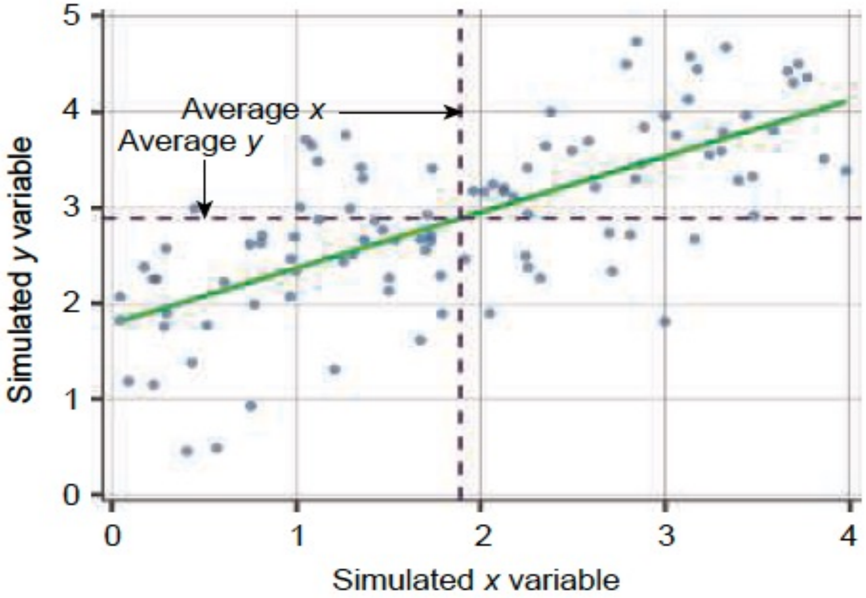


Figure 7.4 Scatterplot and best-fitting linear regression found by OLS

Coefficient Formula

- Mathematically, the OLS method finds the values of the coefficients of the linear regression that minimize the sum of squares of the difference between actual y values and their values implied by the regression, $\hat{a} + \hat{\beta}x$.
- They are called regression residuals as follows:

$$min_{\alpha,\beta} \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2$$

Predicted Dependent Variable and Regression Residual

- The predicted value of the dependent variable is our best guess for its average value if we know the value of the explanatory variable.
- The predicted value can be calculated from the regression:
  - it is the calculated value of y^E using $f(x)$ for a particular value of x.
- As the name suggests, the primary use of the predicted value is in predictive analysis.

$$\hat{y}_i = \hat{a} - \hat{\beta}\, x_i$$

Coefficient Formula

- The **residual** is another important measure that we can compute from a regression.
- It is the difference between the actual value of the dependent variable for an observation and its predicted value:

$$e_i = y_i - \hat{y}_i$$

Goodness of Fit, R-Squared

- An important property of regression is how well it fits the data, called goodness of fit.
- One result of an estimated regression are the predicted values of y for all values of x in the data.
- The fit of a regression captures how these predicted values compare to the actual values.
- The most commonly used measure of the fit of a regression is the R-squared (R^2).
- R-squared captures how close the predicted y values are to the actual values.
- ***It does so by measuring how much of the variation in $\boldsymbol{y}$ is captured by the regression, and how much is left for residual variation.***

Goodness of Fit, R-Squared

- R-squared may be defined as how much of the overall variation in y is captured by variation predicted by the regression, that is, the variation in $\hat{y}$.
- Or, we can check the residual variation.

$$R^2 = \frac{Var[\hat{y}]}{Var[y]} = 1 - \frac{Var[e]}{Var[y]}$$

- The value of R-squared is always between zero and one.
- R-squared is one if the regression fits the data perfectly.
- In this case each and every data point is equal to its predicted value from regression, and all residuals are zero.
- In a linear regression it means that all data points lie exactly on the regression line.

Goodness of Fit, R-Squared

- The other polar case is an R-squared of zero.
- In this case all of the predicted $\hat{y}$ values are equal to the overall average value $\bar{y}$ in the data regardless of the value of the explanatory variable x.
- This corresponds to a slope of zero:
  - The regression line is completely flat
- When our goal is prediction, R-squared may help in choosing between different versions of regression for the same data:
  - Pick the one with higher R-squared

Correlation and Linear Regression

- Linear regression is closely related to correlation.
- The correlation coefficient measures the degree of association between two variables.
- It is a normalized version of the covariance, dividing it by the standard deviations of the two variables:

$$Corr[x,y] = \frac{Cov[y,x]}{Std[y]Std[x]}$$

- The OLS formula for the slope estimate of the linear regression $y^E = \alpha + \beta x$ is also a normalized version of the covariance, only here it is divided by the variance of the x variable:

$$\hat{\beta} = \frac{Cov[y,x]}{Var[x]}$$

- In contrast with the correlation coefficient, its values can be anything, and y and x are not interchangeable.

Correlation and Linear Regression

- Despite their differences, the covariance, the correlation coefficient, and the slope of a linear regression capture similar information:
  - the degree of association between the two variables.
- Formally, we can express the correlation coefficient in terms of the slope coefficient and vice versa.
- Recall that $Var[x] = (Std[x])^2$

$$\hat{\beta} = Corr[x, y]\,\frac{Std[y]}{Std[x]} \text{ or } Corr[x, y] = \hat{\beta}\,\frac{Std[x]}{Std[y]}$$

Regression and Causation

- Correlation does not imply causation – a warning we can find in every introduction to statistics.
- The slope of a simple regression describes the same association as correlation, so this message is equivalent to saying that the slope of a simple regression does not imply causation.
- By that we mean that, in general, the slope coefficient of a linear regression does not show the effect of the explanatory variable on the dependent variable.
- **Causality** is a concept that is not that easy to define.
- One intuitive approach postulates that x causes y if we could expect y to change if we were to change x.

Regression and Causation

- Regression analysis cannot in general uncover the effect of x on y in observational data because the variation in x is not controlled.
- In observational data we compare observations that may be different in terms of x for many reasons.
- As a result, the data does not necessarily inform us of what would happen to y if we were to change x.
- For example, using observational data to uncover the effect of advertising on sales, we may regress sales on the amount of advertising and see a strong positive coefficient.
- But that may not show an effect at all.
- For example, both advertising and sales may be above average in the holiday season.
- That in itself would lead to a positive slope coefficient without any effect of advertising.

Regression and Causation

- In contrast, the same regression analysis can uncover the effect of x on y in experimental data where variation in x is controlled.
- In well-designed experiments, an experimenter induces controlled variation in x: they manipulate the value of x in a way that rules out the influence of other effects and observe differences in the expected value of y as a result.
- For example, a firm may consciously experiment by allocating varying resources to advertising, in a random fashion, and keep track of sales.
- A regression of sales on the amount of advertising can uncover the effect of advertising here.

Regression and Causation

- This is why proper interpretation of the slope of a regression is important.
- Comparing observations that are different in x and seeing the extent to which y differs among them on average is what regression analysis does.
- The proper interpretation of the slope is necessary whether the data is observational or comes from a controlled experiment.
- A positive slope in a regression of sales on advertising means that sales tend to be higher when advertising time is higher.
- This interpretation is true both in observational and experimental data.

Regression and Causation

- Instead of saying that the correlation, or regression, does not imply causation, we should rather say that we should not infer cause and effect from comparisons, especially when the data is observational.
- That is not as catchy and short.
- But it has the advantage of being both true and useful.
- In any case, when the slope of the $y^E = \alpha + \beta x$ regression is not zero in our data ($\beta \neq 0$) and the linear regression captures the $y - x$ association reasonably well, one of three things – which are not ally exclusive – may be true:

Regression and Causation

1. x causes y.

   If this is the single one thing behind the slope, it means that we can expect y to increase by β units if we were to increase x by one unit.

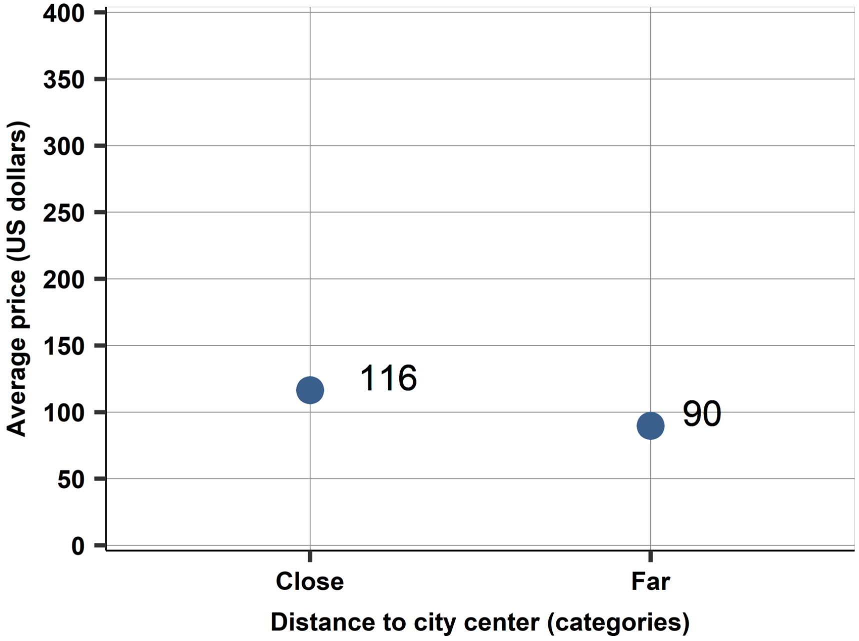
2. y causes x.

If this is the single one thing behind the slope, it means that we can expect x to increase if we were to increase y.

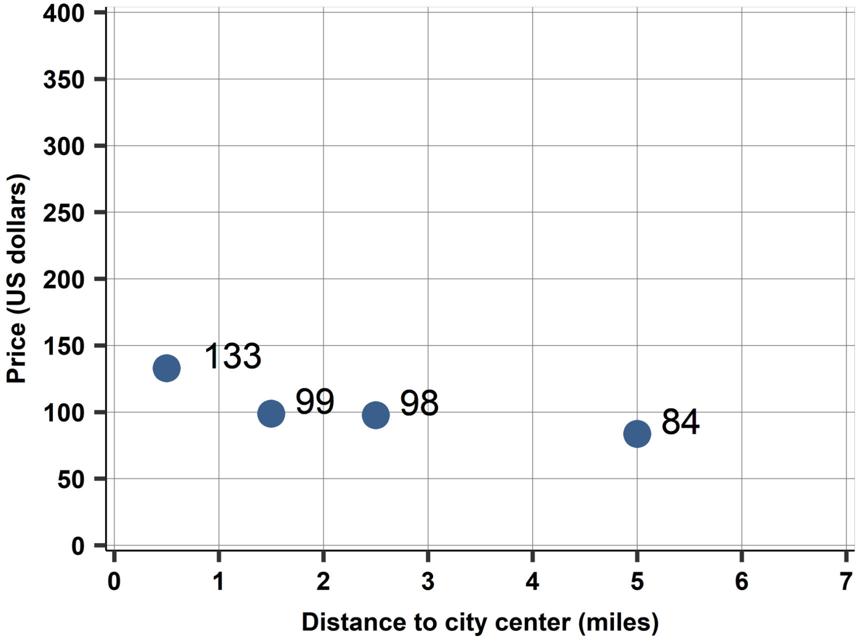
3. A third variable causes both x and y (or many such variables do).

If this is the single one thing behind the slope it means that we cannot expect y to increase if we were to increase x.

Case Study: Finding a good deal among hotels

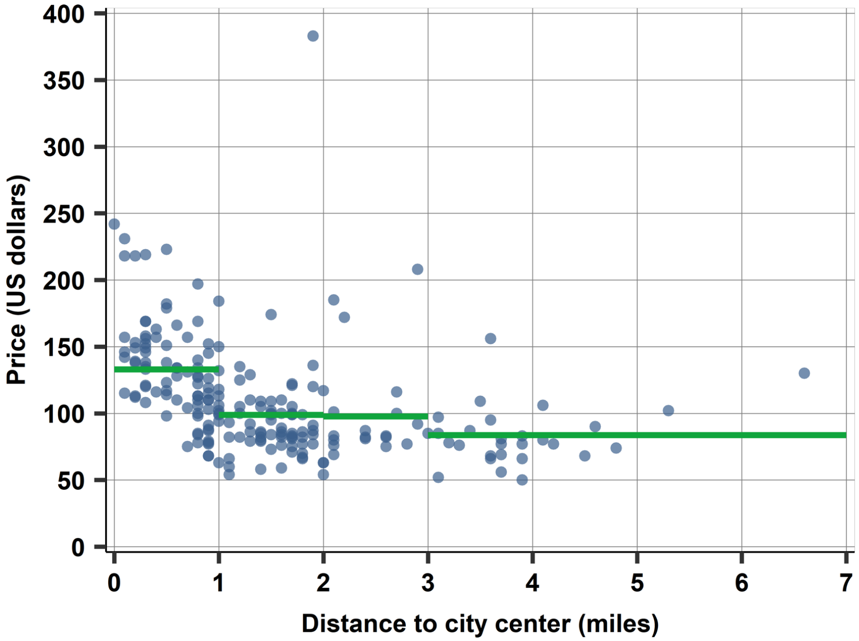


Bin scatter non-parametric regression, 2 bins

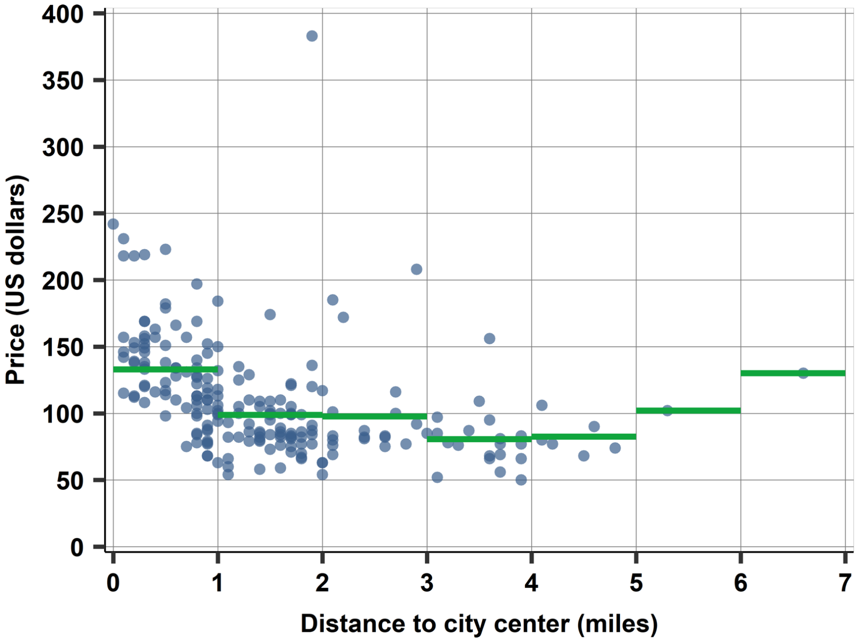


Bin scatter non-parametric regression, 4 bins

Case Study: Finding a good deal among hotels

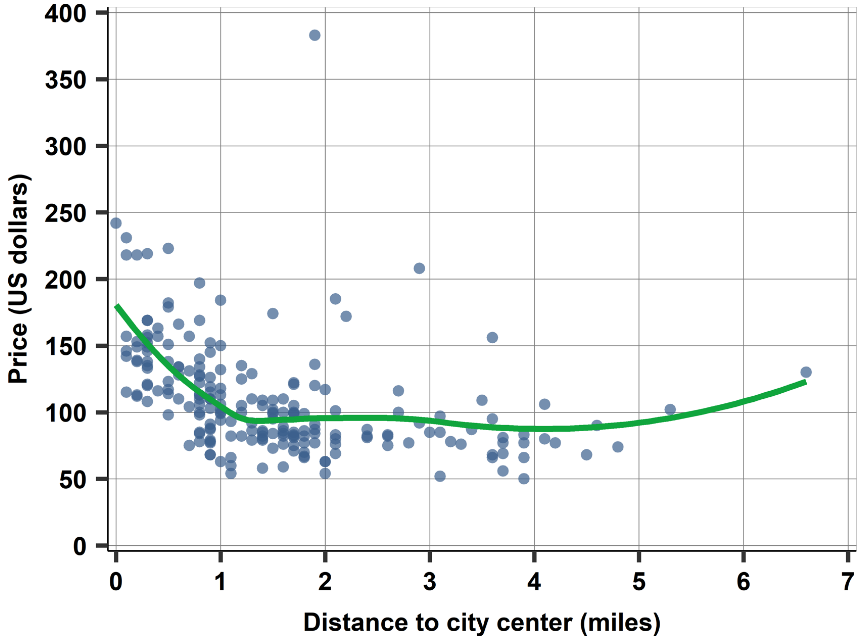


Scatter and bin scatter non-parametric regression, 4 bins



Scatter and bin scatter non-parametric regression, 7 bins

- Case Study: Finding a good deal among hotels
  - **lowess** non-parametric regression, together with the scatterplot.
  - bandwidth selected by software is 0.8 miles.
  - The smooth non-parametric regression retains some aspects of previous bin scatter – a smoother version of the corresponding
    - non-parametric regression with disjoint bins of similar width.

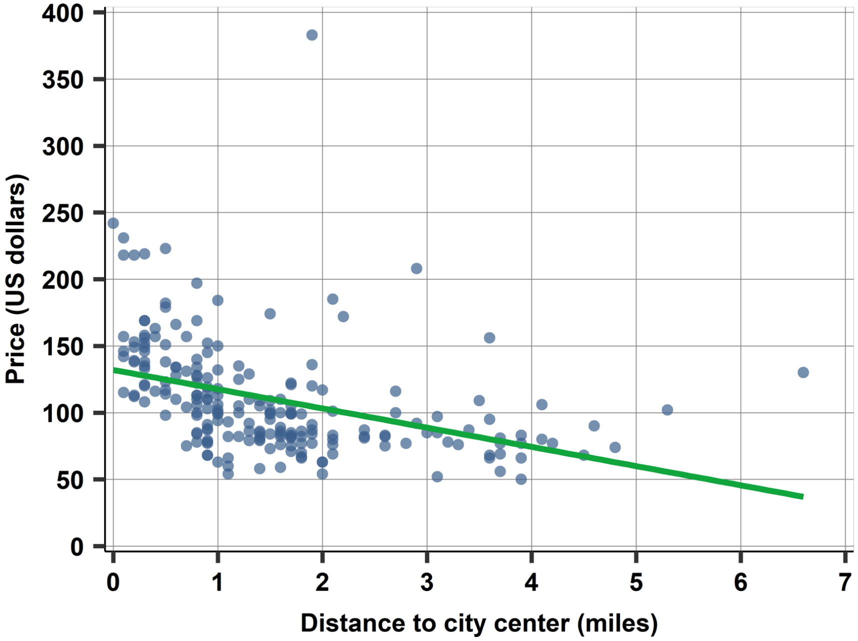


Case Study: Finding a good deal among hotels

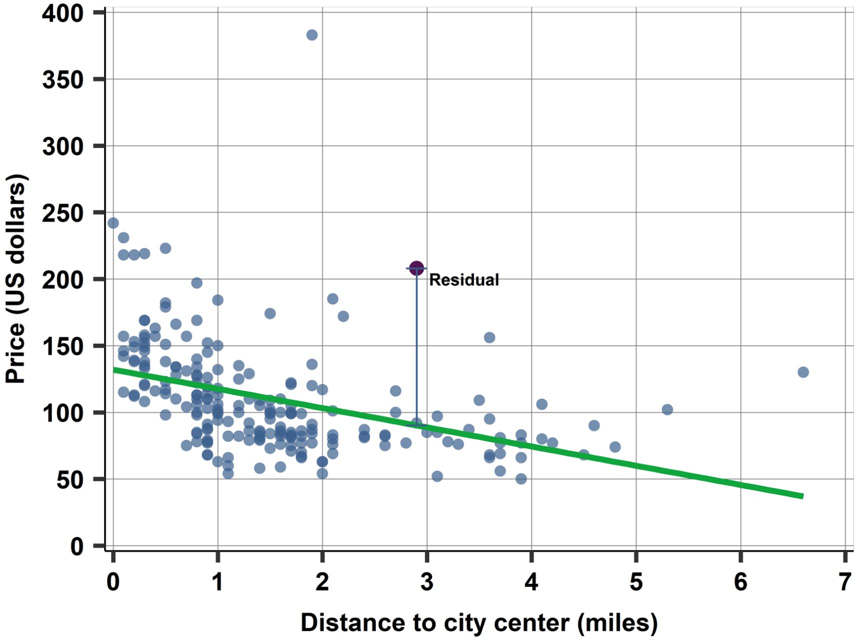
The linear regression of hotel prices (in $) on distance (in miles) produces an intercept of 133 and a slope -14.

The intercept is 133, suggesting that the average price of hotels right in the city center is $ 133.

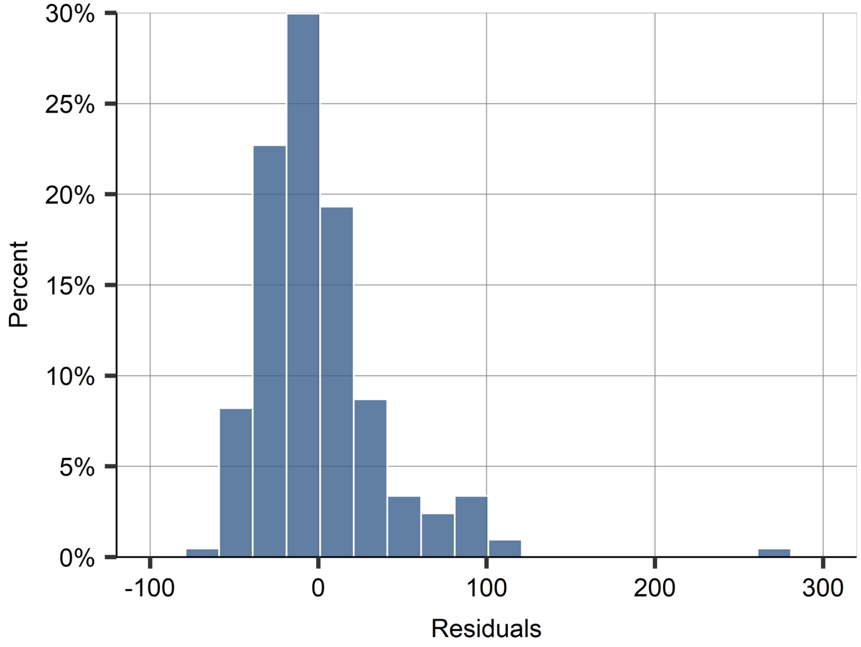
The slope of the linear regression is -14. Hotels that are 1 mile further away from the city center are, on average, $ 14 cheaper in our data.



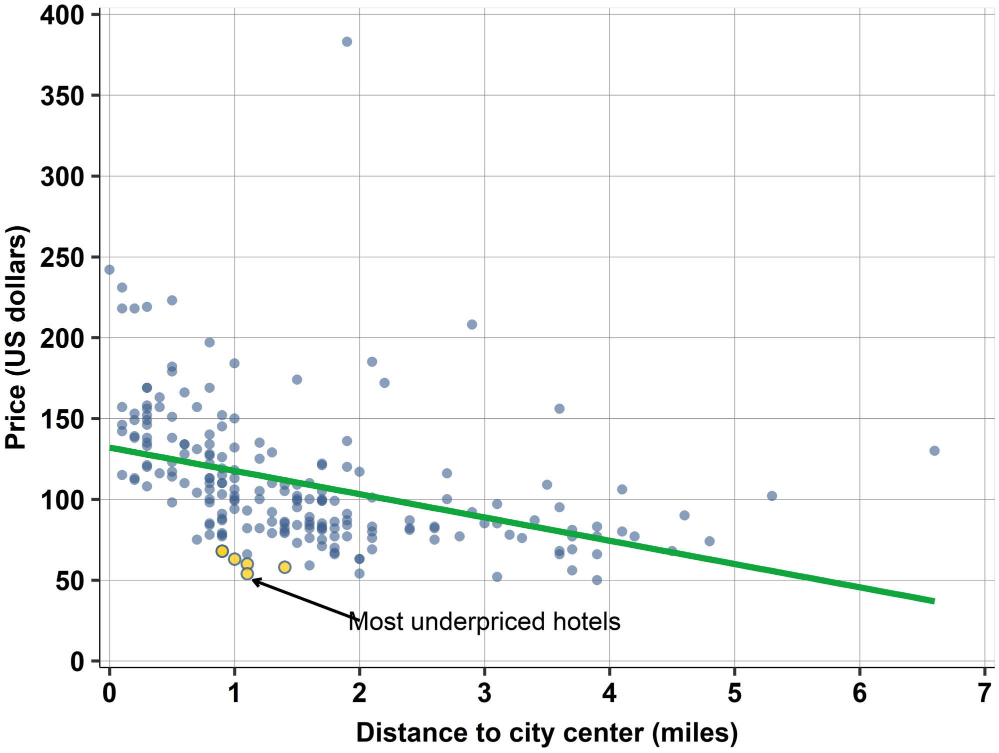
- Case Study: Finding a good deal among hotels
  - Residual is vertical distance
  - Positive residual shown here - price is above what predicted by regression line



- Case Study: Finding a good deal among hotels
  - Can look at residuals from linear regressions
  - Centered around zero
  - Both positive and negative



- Case Study: Finding a good deal among hotels
  - If linear regression is accepted model for prices
  - Draw a scatterplot with regression line
  - With the model you can capture the over and underpriced hotels



Case Study: Finding a good deal among hotels

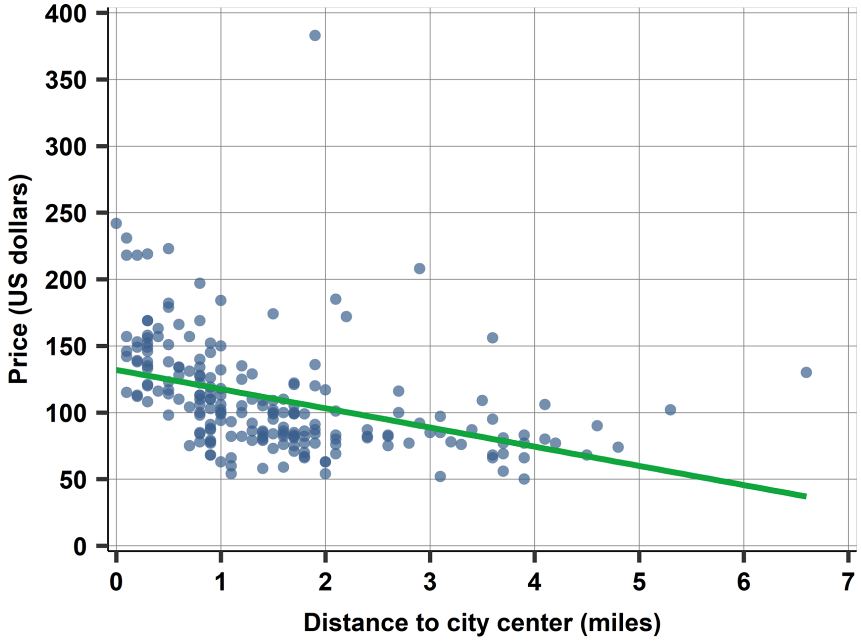
A list of the hotels with the five lowest value of the residual.

No.	Hotel_id	Distance	Price	Predicted price	Residual
1	22080	1.1	54	116.17	-62.17
2	21912	1.1	60	116.17	-56.17
3	22152	1	63	117.61	-54.61
4	22408	1.4	58	111.85	-53.85
5	22090	0.9	68	119.05	-51.05

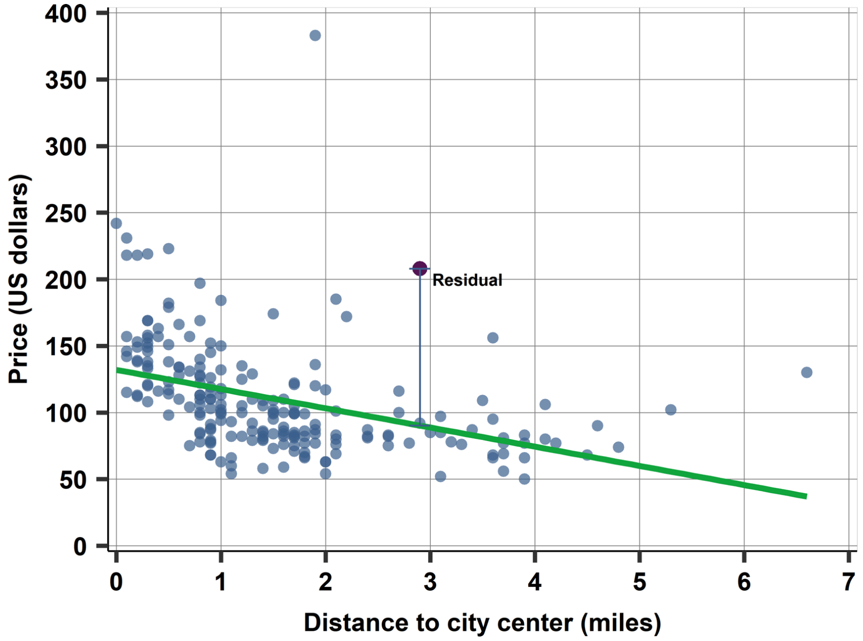
- Bear in mind, we can (and will) do better - this is not the best model for price prediction.
  - Non-linear pattern
  - Functional form
  - Taking into account differences beyond distance

- Case Study: Finding a good deal among hotels

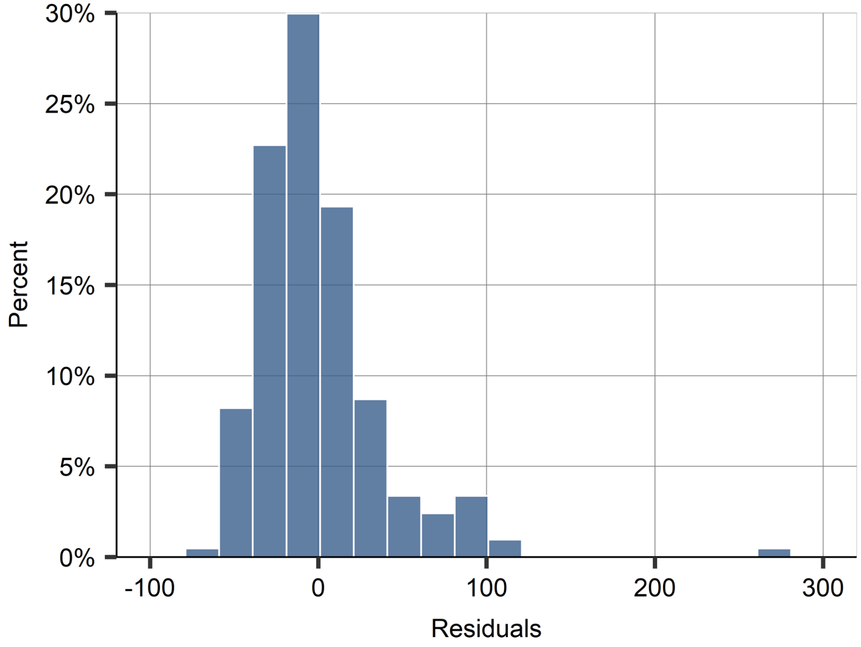
  - The linear regression of hotel prices (in $) on distance (in miles) produces an intercept of 133 and a slope -14.
  - The intercept is 133, suggesting that the average price of hotels right in the city center is $ 133.
  - The slope of the linear regression is
    - -14. Hotels that are 1 mile further away from the city center are, on average, $ 14 cheaper in our data.



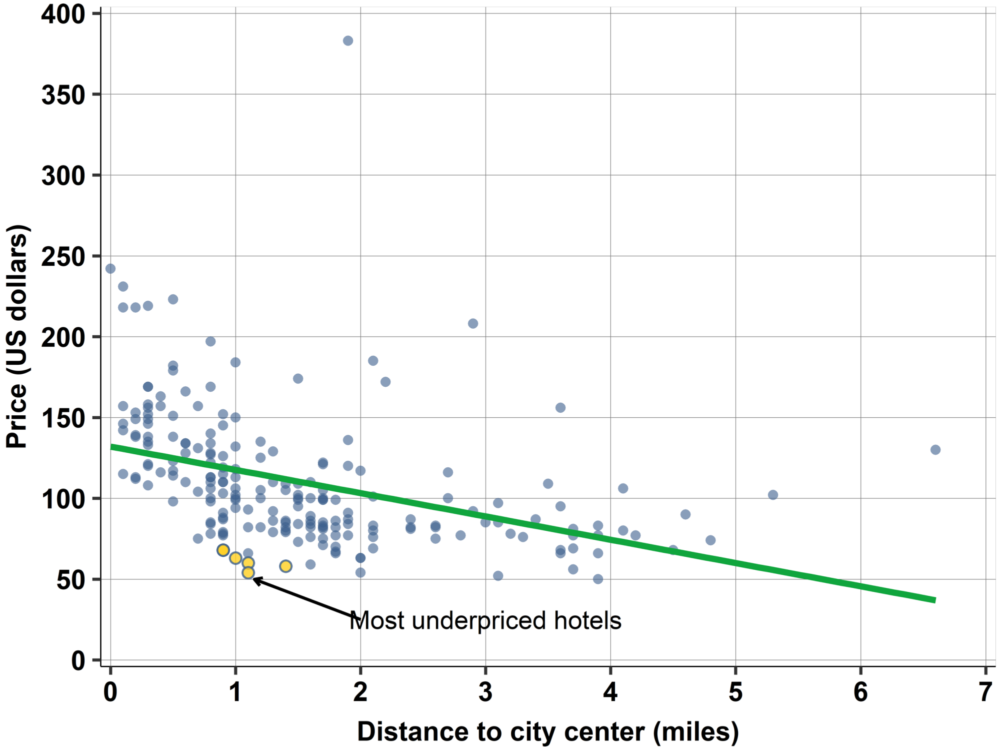
- Case Study: Finding a good deal among hotels
  - Residual is vertical distance
  - Positive residual shown here - price is above what predicted by regression line



- Case Study: Finding a good deal among hotels
  - Can look at residuals from linear regressions
  - Centered around zero
  - Both positive and negative



- Case Study: Finding a good deal among hotels
    - If linear regression is accepted model for prices
    - Draw a scatterplot with regression line
    - With the model you can capture the over and underpriced hotels



Case Study: Finding a good deal among hotels

A list of the hotels with the five lowest value of the residual.

No.	Hotel_id	Distance	Price	Predicted price	Residual
1	22080	1.1	54	116.17	-62.17
2	21912	1.1	60	116.17	-56.17
3	22152	1	63	117.61	-54.61
4	22408	1.4	58	111.85	-53.85
5	22090	0.9	68	119.05	-51.05

- Bear in mind, we can (and will) do better - this is not the best model for price prediction.
  - Non-linear pattern
  - Functional form
  - Taking into account differences beyond distance