ØAMET2200
Business Decision Making Using Data
Lecture 5

Instructor: Fenella Carpena

September 20, 2019

Announcements

- No lecture next week
- Next lecture on October 4
- Problem Set 2 due on October 3, 11:59PM
- Textbook readings for Lectures 1-5 posted on Canvas

Part 4 of this Course: Building regression models

- You want to build a model to make a forecast or prediction from data you have
- What is the best **model you can build**?
- This lecture: Constructing and interpreting the **Ordinary Least Squares** (OLS) regression line for **linear** & **curved** patterns
- Next lectures: Statistical inference and multivariate models

Agenda for Today

- Constructing the OLS regression line
- Interpreting and using the OLS model
- The importance of residuals
- Modeling and interpreting curved patterns
- Transformations
- Chapter 19, Chapter 20.1, 20.2, 20.4

Case Study for Today: Real Estate

- Skanska Eiendomsutvikling, a residential builder in Norway, has started constructing new apartments in Oslo
- Imagine that the apartments are 50-150 square meters
- Suppose that you are a real estate analyst for this company
- You have been tasked with determining the price that the company should set for apartments of different sizes

Løren, Oslo

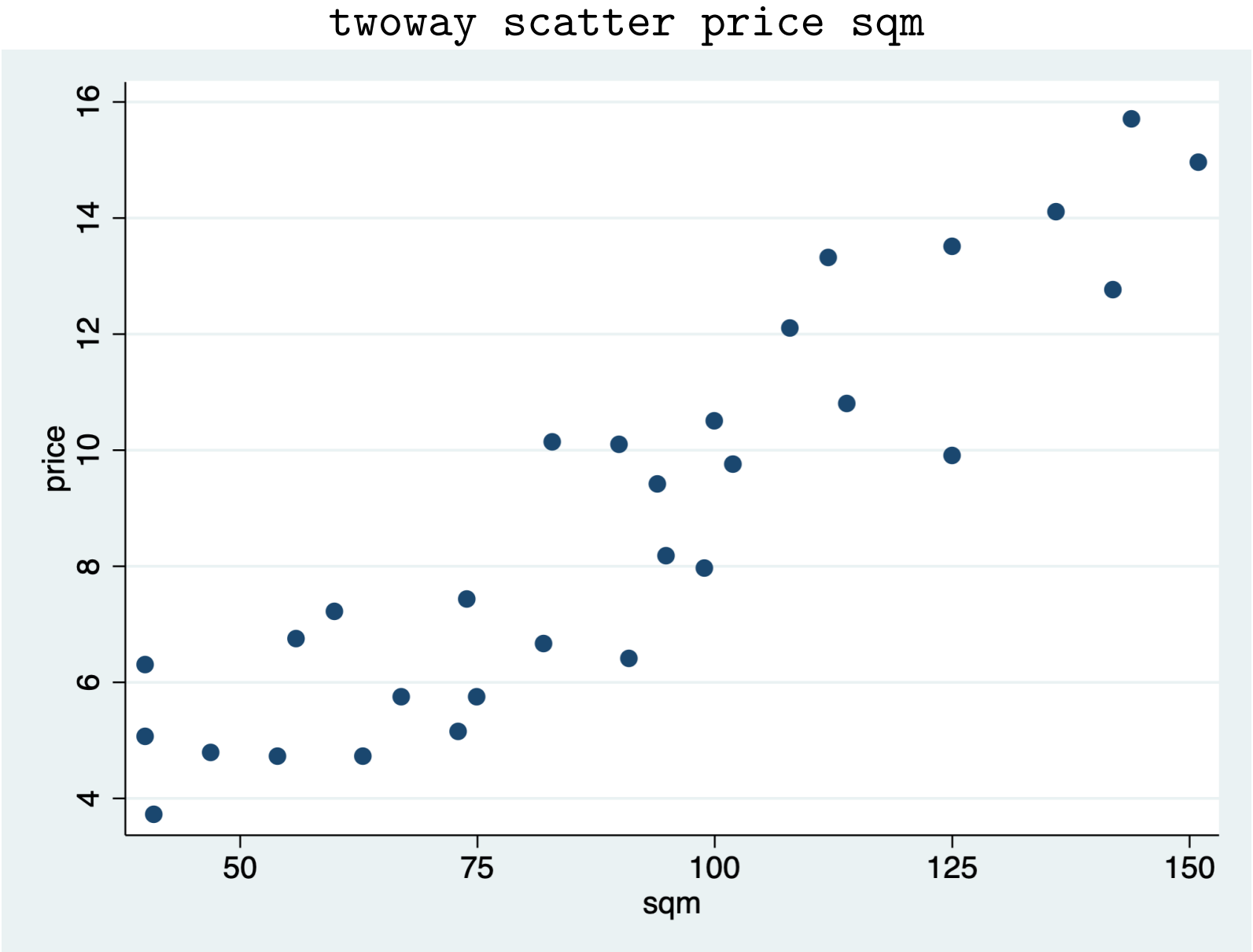
Kommer for salg!
På Løren i Oslo skal vi bygge leiligheter. Eiendommen er en del av områdeplanen for Løren.

Jeg vil holdes oppdatert om dette prosjektet

Case Study for Today: Real Estate

- From Finn.no, you gather data on two variables:
  1. Asking price in millions of kroner, variable name: `price`
  2. Size in square meters, variable name: `sqm`
- Based on this data:
  - What's the relationship between price and apartment size?
  - What's the ave. price of apartments that are 40 square meters?
  - How much more do apartments that are 50 square meters cost?
- To answer these questions, you want to build a linear model that links apartment size to price
  - Does a linear model make sense?
  - How do you build that model?

Scatterplot of Price vs. Square Meters



Equation of the fitted line

- The **equation** of the **fitted line** is $\widehat{y} = b_0 + b_1 x$
  - b_0 is the **intercept**
  - b_1 is the **slope**
  - $\widehat{y}$ is the **fitted value**, an estimate or prediction of y based on the equation
- Real Estate Example: $\widehat{\text{Price}} = b_0 + b_1 \text{Square Meters}$
- **Many different lines** can be drawn through the cloud of data
- How do we **pick** b_0 and b_1?

Residuals

- We want to estimate a relationship between Price and Square Meters so that the **error** between our prediction and the actual price is **small**
- **Residuals** describe the error of our prediction vs. actual price
- **Residuals** are the vertical deviations from the datapoint to the fitted line, $e = y - \widehat{y}$

Ordinary Least Squares (OLS) Regression

- The OLS regression line picks b_0 and b_1 to minimize the **sum of squared** residuals
- The resulting line is called the **OLS regression line**
- The **slope** of OLS regression line: $b_1 = r\frac{s_y}{s_x}$
  - The slope **depends on** $corr(x, y)$
- The **intercept** of OLS regression line: $b_0 = \overline{y} - b_1\overline{x}$
  - The OLS regression line **always** goes through the point $(\overline{x}, \overline{y})$

Real Estate: Summary Statistics

```
. summarize price sqm ;
```

Variable	Obs	Mean	Std. Dev.	Min	Max
price	30	8.781667	3.438484	3.71	15.7
sqm	30	89.43333	32.37249	40	151

```
. correlate price sqm ;
(obs=30)
```

	price	sqm
price	1.0000	
sqm	0.9110	1.0000

Real Estate: Calculating the Fitted OLS Regression Line

- What is the OLS regression line for relating apartment prices to the apartment size in square meters?

Real Estate: Calculating the Fitted OLS Regression Line

```
. regress price sqm ;
```

Source	SS	df	MS
Model	284.581623	1	284.581623
Residual	58.2904057	28	2.08180021
Total	342.872029	29	11.8231734

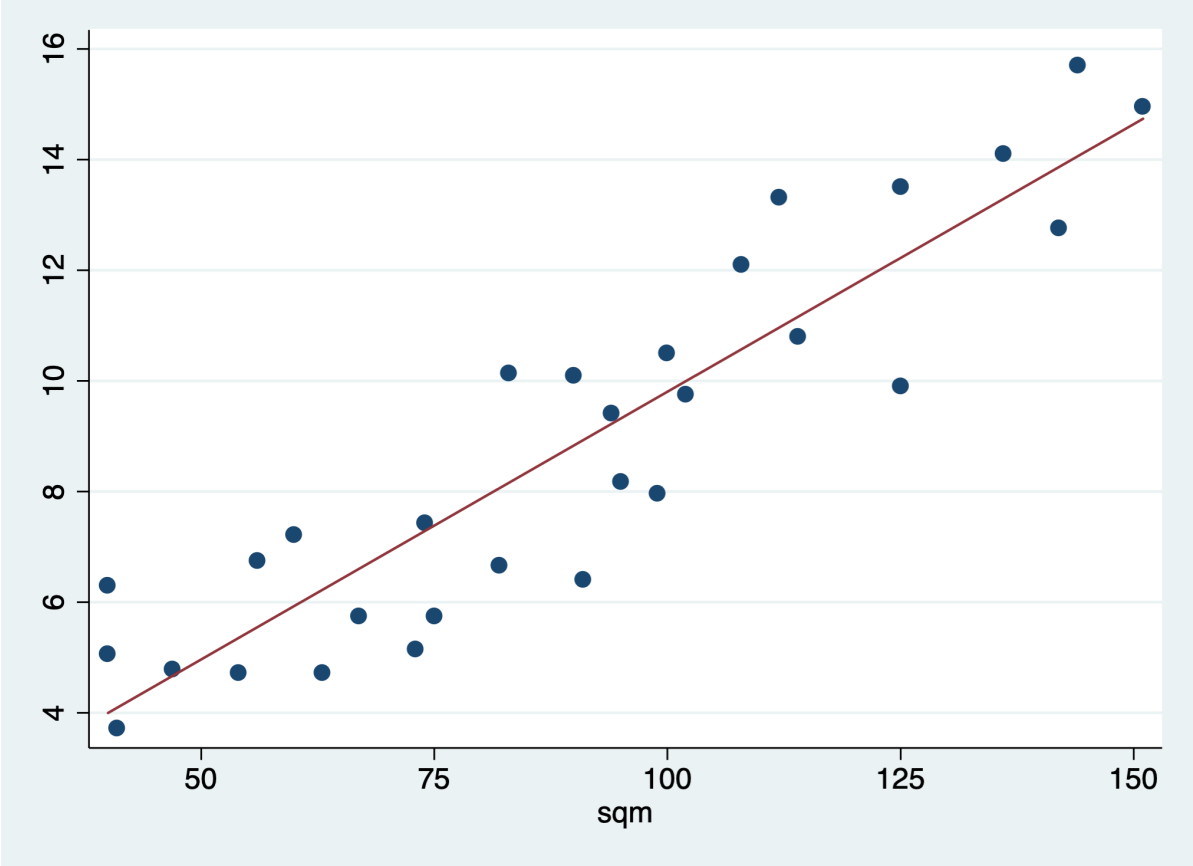
Number of obs	30
F(1, 28)	136.70
Prob > F	0.0000
R-squared	0.8300
Adj R-squared	0.8239
Root MSE	1.4428

price	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
sqm	.0967672	.0082765	11.69	0.000	.0798137	.1137208
_cons	.1274507	.7856692	0.16	0.872	-1.48192	1.736821

Interpreting the fitted line

$$\widehat{\text{Price}} = 0.127 + 0.097 \cdot \text{Square Meters}$$

- `twoway (scatter price sqm) (lfit price sqm)`



- The line gives a **prediction** for the price of an apartment of **any** size
- We can also see the residuals from the above figure

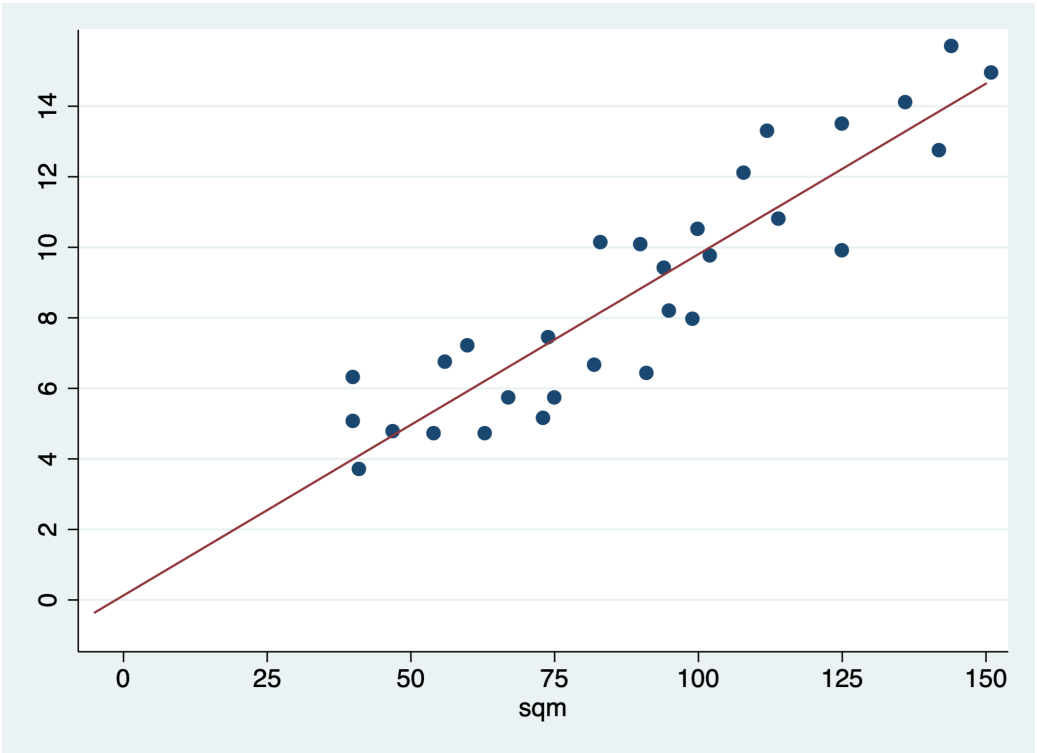
Interpreting the intercept

$$\widehat{\text{Price}} = 0.127 + 0.097 \cdot \text{Square Meters}$$

- The intercept is the portion of y that is **present for all values of** x
- The intercept estimates the **average** y (i.e., $\bar{y}$) **when** $x = 0$
- The intercept has the **same units** as y
- Real Estate Example: How do we interpret the intercept?

Extrapolation

- In some cases we should be careful in interpreting b_0 (or b_1)
- This arises with **extrapolations**, estimates based on extending the equation beyond conditions observed in the data
- In our sample, we don't have apartments with `sqm` close to 0



- $b_0 = 0.127$ is an extrapolation; must interpret with **caution**

Interpreting the slope

$$\widehat{\text{Price}} = 0.127 + 0.097 \cdot \text{Square Meters}$$

- The slope is the predicted change in y if x changes by 1 unit
  - Real Estate Example: 1 sqm increase in the apartment size is predicted to increase price by _____
- To find the effects of changes other than 1 unit, multiply b_1 with the desired change
  - Real Estate Example: 5 sqm increase in the apartment size is predicted to increase price by _____
- While tempting, it is **not correct** to describe the slope as "the change in y caused by changing x"

Exercise

Identify and interpret the slope and intercept in these regressions.

(a) A regression of the number of packages sorted per hour (y) on the number of employees working (x) in the sorting facility of a shipping company

$$\widehat{\text{Packages Sorted}} = 200 + 1,100 \text{ Number Employees}$$

(b) A regression of the cost in dollars to produce a hand-knitted sweater on the number of hours taken to knit the sweater

$$\widehat{\text{Cost}} = 200 + 1,100 \text{ Hours}$$

Prediction

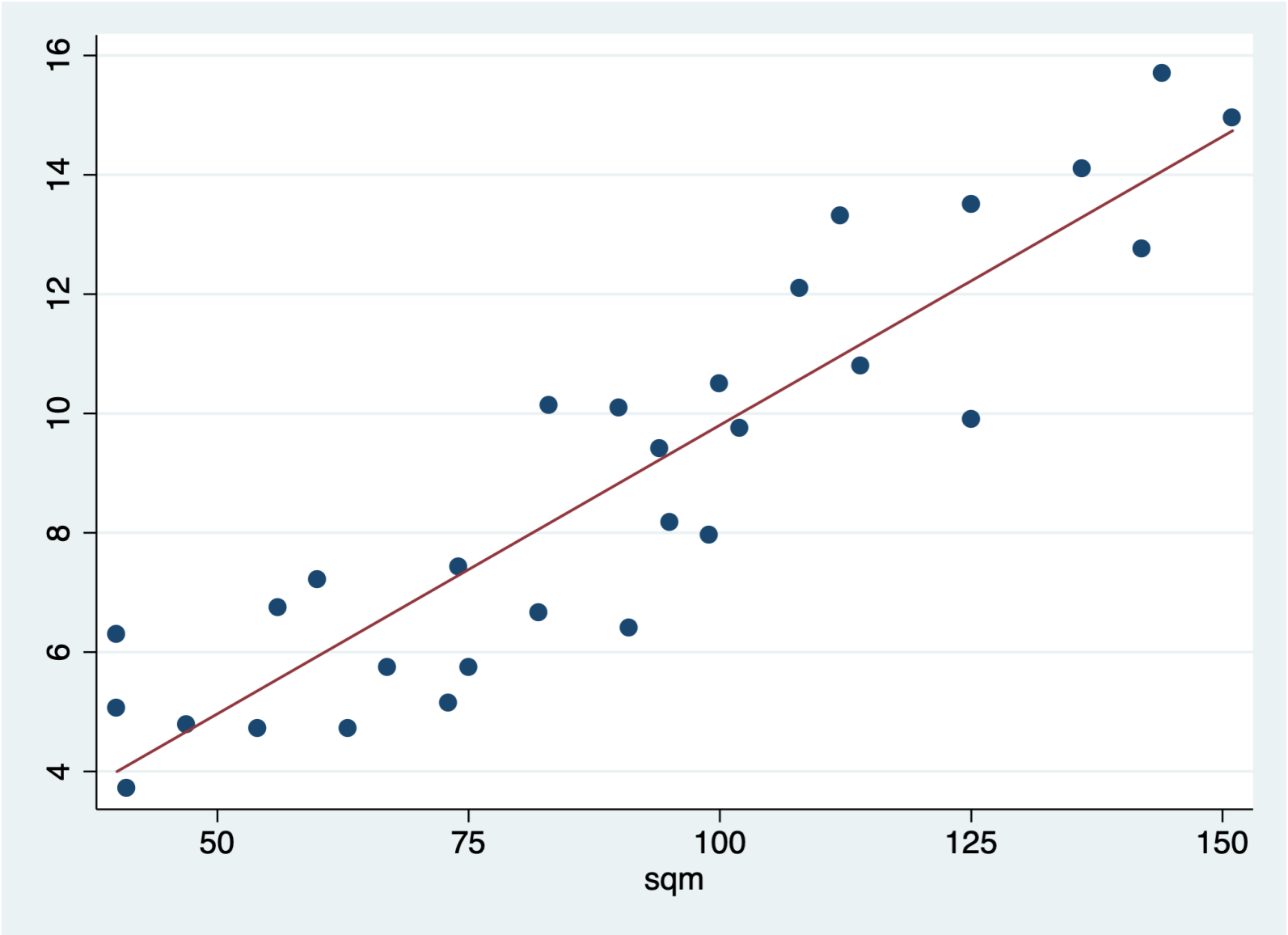
$$\widehat{\text{Price}} = 0.127 + 0.097 \cdot \text{Square Meters}$$

- What is the average price of a 75-square meter apartment?

- What is the average price of a 100-square meter apartment?

Prediction

$$\widehat{\text{Price}} = 0.127 + 0.097 \cdot \text{Square Meters}$$



Measures of Fit: How good are the predictions?

1. Standard deviation of the residuals
    - Denoted s_e
    - Also called "standard error of the regression" or the "root mean squared error"
2. R-squared
    - Denoted r^2

Standard deviation of residuals, s_e

- Measures the spread of data points around the regression line
- The formula is given by

$$s_e = \sqrt{\frac{e_1^2 + e_2^2 + \ldots + e_n^2}{n-2}}$$

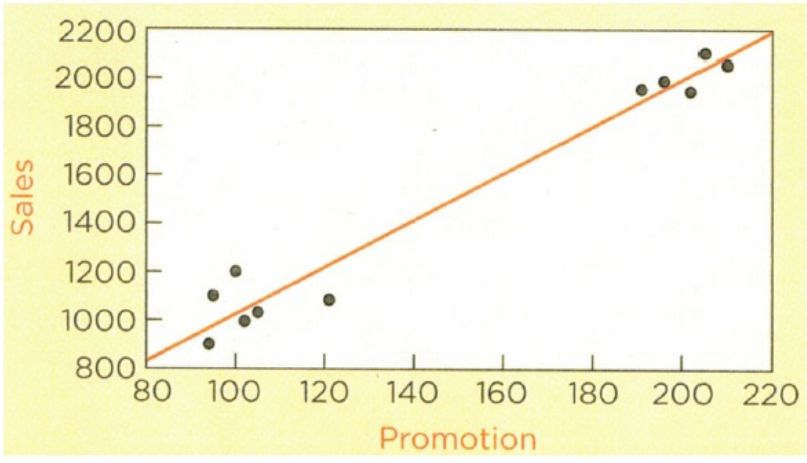
- s_e has the same units as the y variable
- If all of the data are on the fitted line, then $s_e = 0$
- Real Estate Example: $s_e = 1.4428$

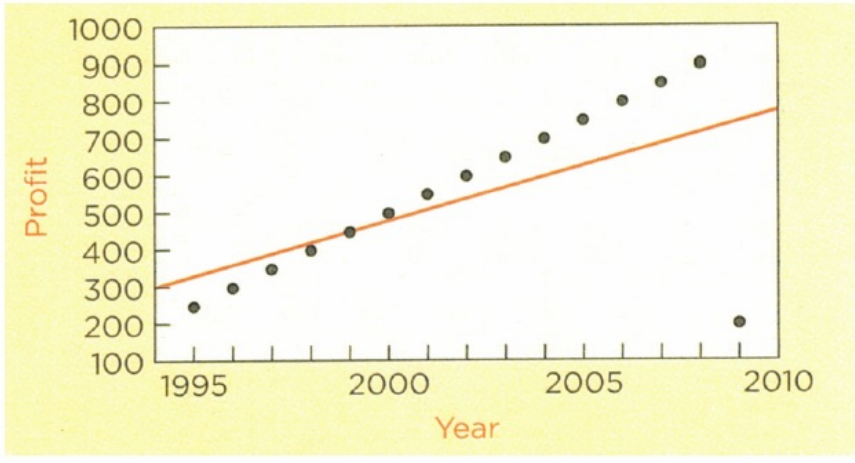
R-squared

- The residuals are what your model fails to explain
- What your model does explain is measured by the r^2
- The r^2 is equal to the square of the correlation between x and y
- It is the fraction of the variation in y that is explained by the OLS regression line
- Real Estate Example: $r^2 = 0.83$
- The OLS line explains 83% of the variation in price
- Note that a high r^2 means s_e is low

Is the linear model approporiate?

- r^2 and s_e do not tell you if your model is "good" or "bad"
- A high r^2 (low s_e) does not mean that x causes y
  - Switching the roles of x and y, we get the same r^2
- A high r^2 (low s_e) does not mean that a linear fit is the appropriate model to use

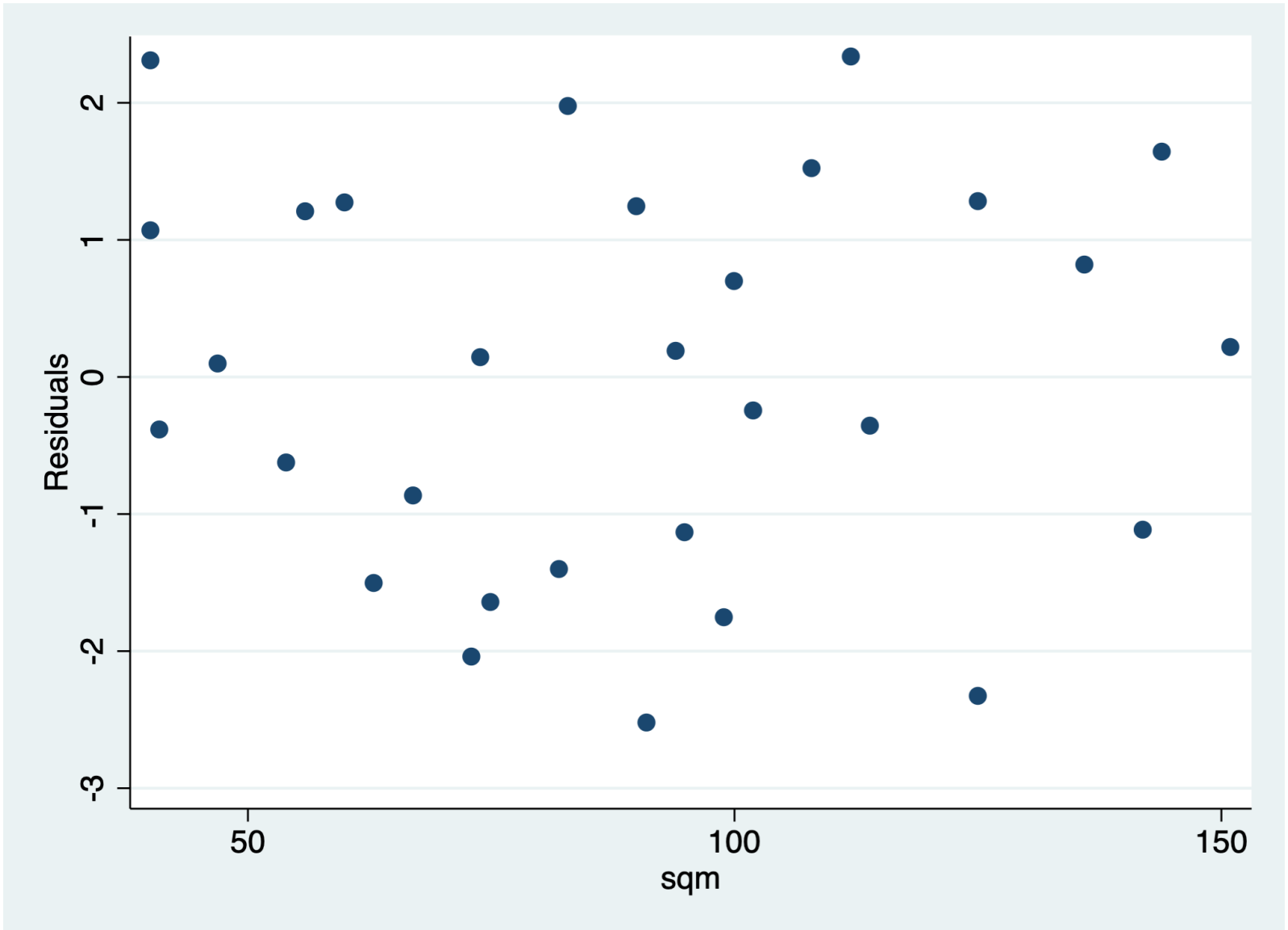




Is the linear model approporiate?

- Check patterns in the **residuals** to answer this question
- Residuals show variation that **remains** in the data after accounting for the linear relationship defined by the OLS line
- If a regression equation works well, it should capture the underlying pattern, so only **random variation** should remain in the residuals
- If the least squares model is appropriate, then the plot of the residuals vs. the x variable should **show no pattern**

Residual Plot



Exercise: True or False

Mark each statement as true or false, and provide a brief explanation for your answer.

(a) If all of the data lie along a single line with non-zero slope, then the r^2 of the regression is 1 (assume the values of the x variable are not identical).

(b) If the correlation between x and y is zero, then the slope in the OLS regression line will also be zero.

(c) The use of a linear equation to describe an association between x and y implies that the change in y when x goes from 10 to 11 is the same as when x goes from 20 to 21.

Non-Linear Patterns

- Linear models impose that **equally sized changes in** x are associated with **equal changes in** y
- Linear models are a good place to start, but **they don't work in every situation**
  - Example: decreasing marginal product of labor
- At the start of the modeling process, **ask this question**

  Should changes in x of a given size come with equal changes in y, regardless of the value of x?

Case Study: Cat Food

- Rema 1000 would like to know what is the best price at which they can sell **cat food** (sold in cans)
- High price: **large profit** per can sold, but sell **fewer cans**
- Low price: **sell more** cans, but **less profit per can**
- We want to use data on prices and sales history to find the **optimal** (i.e., most profitable) price

Case Study: Cat Food

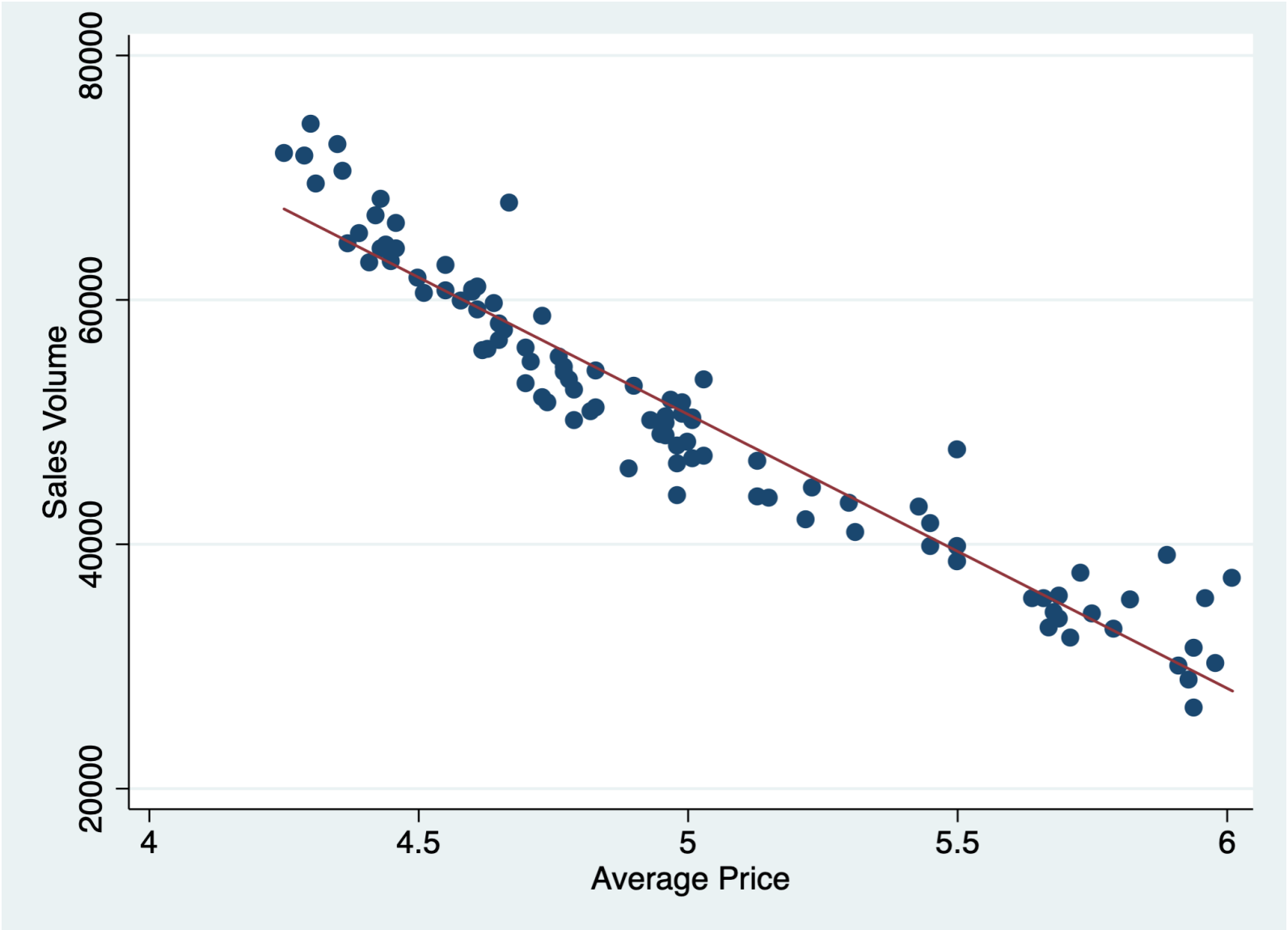
- We have **weekly data** for 99 weeks with two variables:
  1. **number of cans sold**, `SalesVolume`
  2. **average selling price**, `AvgPrice`
- Suppose that we calculated the OLS line and found

$$\widehat{\text{SalesVolume}} = 162,714 - 22,418 \cdot \text{AvgPrice}$$

$$r^2 = 0.92, \;\; s_e = 3382$$

Cat Food Data: Scatter Plot with OLS line

What is the problem here?



Cat Food Data: Residual Plot

Transformations

- A **transformation** defines a new variable by applying a function to each of the values of an existing variable.
- Why do we care? Transformations allow us to use a **linear regression** to describe a **curved** pattern
- Two common types of non-linear transformations

(1) Polynomial (e.g., square, cube)

- Example: x^2 is a polynomial transformation of x
- `generate x2 = x^2`

(2) Natural logarithm

- `generate lnx = ln(x)`

Natural Logarithm

- In a **natural logarithm**, the base is Euler's number, denoted e (which is around 2.7)
- The natural log of x is the **exponent** to which Euler's number must be raised to **produce** x
  - If $\ln(x) = y$, then $e^y = x$
- Here are some examples:
  - $\ln(1) = 0$
  - $\ln(e) = 1$
  - $\ln(10) \approx 2.3$

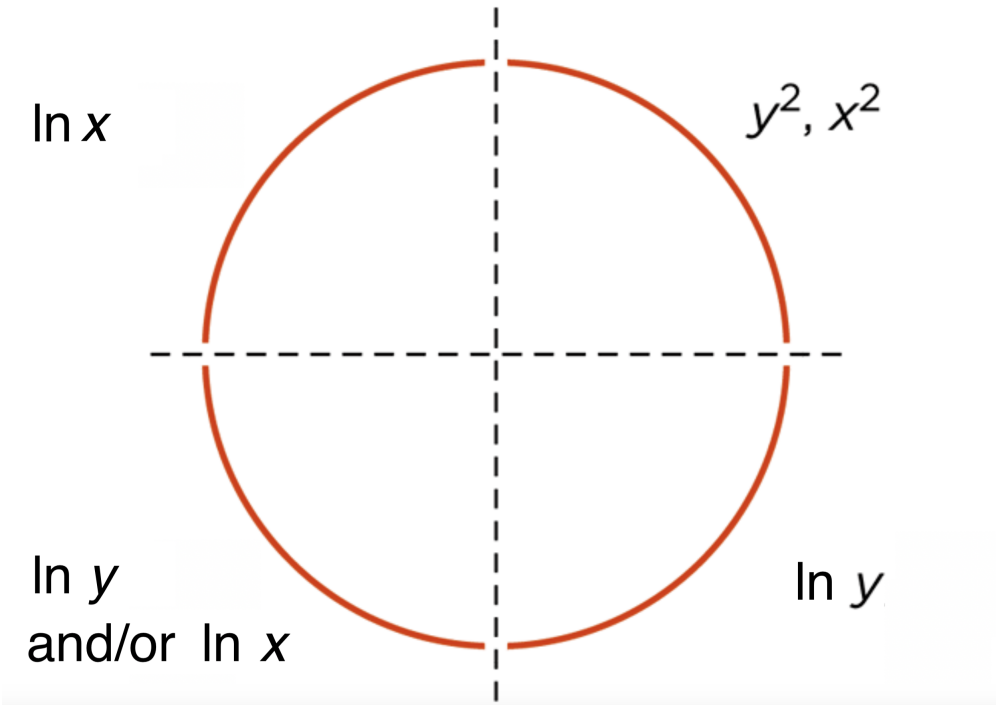
Natural Log and Percentage Changes

x	$\ln(x)$
1	0.0
10	2.3
11	2.4
100	4.6
110	4.7

- Changes in natural log approx. equal **percentage changes**
- What is the percentage change between $x = 10$ & $x = 11$?
  - What is the change between $\ln(10)$ & $\ln(11)$?
- What is the percentage change between $x = 100$ & $x = 110$?
  - What is the change between $\ln(100)$ & $\ln(110)$?

Choosing a transformation

- **Tukey's bulging rule** suggests when to use natural log, squares, and which variable to transform
- Match the **pattern in a scatterplot** to one of the shapes below to find an appropriate transformation



- Which transformation should we use in the Cat Food example?

Three logarithmic models

1. Log-Log: $\ln y = b_0 + b_1 \ln x$
   - b_1: 1% **percent** change in x is associated with b_1 **percent** change in y
   - Estimates the **elasticity**

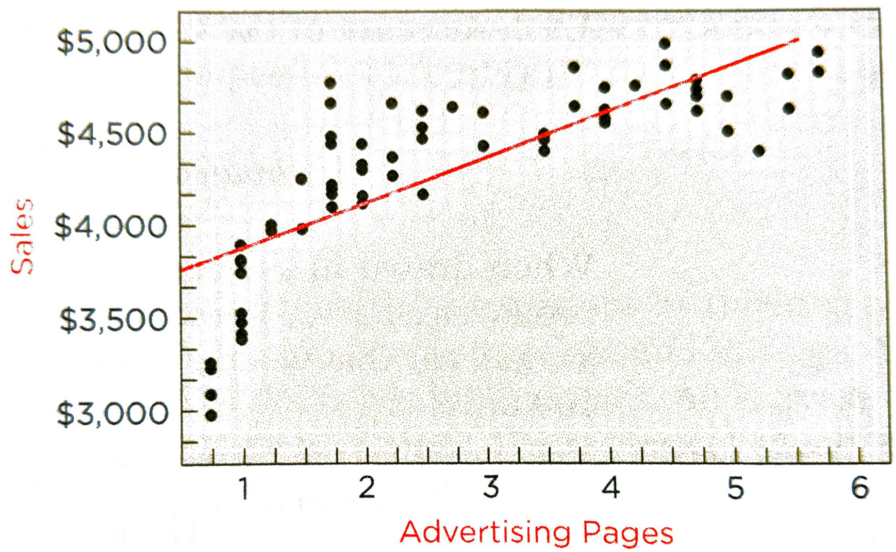
$$\text{elasticity} = \frac{\%\text{ change in } y}{\%\text{ change in } x}$$

2. Log-linear: $\ln y = b_0 + b_1 x$
   - b_1: 1 **unit** change in x is associated with $100 * b_1$ **percent** change in y

3. Linear-Log: $y = b_0 + b_1 \ln x$
   - b_1: 1 **percent** change in x is associated with $b_1/100$ **unit** change in y

Exercise: Nille

Imagine that the store Nille recently began selling grocery items (e.g., canned mackerel) in 64 stores located in Oslo. To see how advertising affects the sales of these new items, it varied the number of pages showing grocery items in the advertising circular (*kundeavis*) that it distributes to shoppers. The figure below graphs the average daily sales of grocery items versus the number of pages devoted to grocery items.



(a) Explain why these data do not satisfy the linear condition.

(b) Which change in advertising pages appear to have a larger effect on sales: increasing from 1 to 2 pages, or from 4 to 5?

(c) What transformation does Tukey's bulging rule suggest?

Cat Food: Scatter plot after log transformation

$$\widehat{\ln \text{Sales Volume}} = 14.56 - 2.41 \cdot \ln \text{Avg Price}$$

$$r^2 = 0.94, \; s_e = 0.06$$

Cat Food: Residual plot after log transformation

Comparing the linear and nonlinear equations

- Which of the two models is better?

  1. the **linear** model with sales and average price

     $$\widehat{\text{SalesVolume}} = 162,714 - 22,418 \cdot \text{AvgPrice}, \; r^2 = 0.92, \; s_e = 3382$$

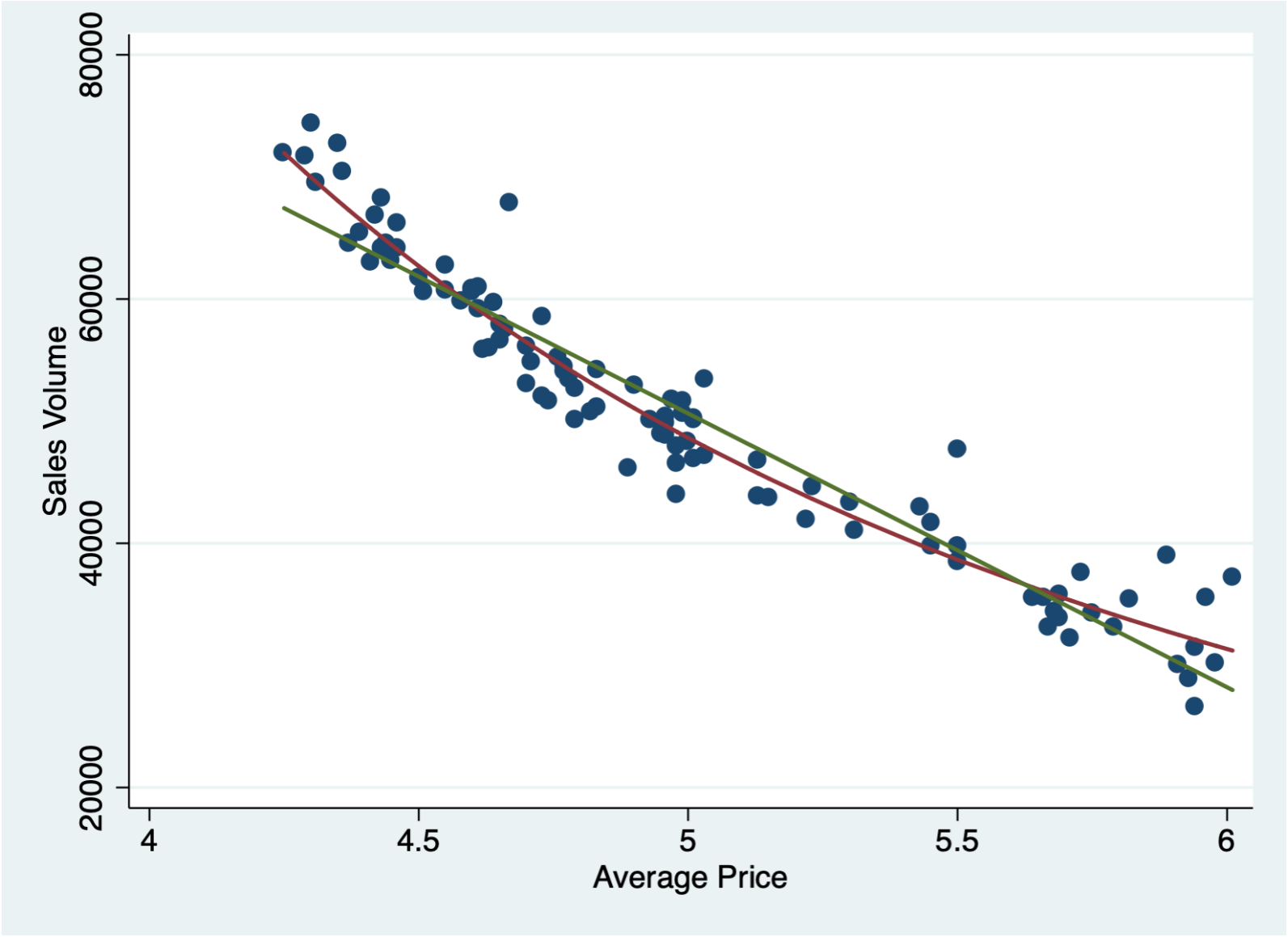
  2. the **log-log** model with log sales and log average price

     $$\widehat{\ln \text{ Sales Volume}} = 14.56 - 2.41 \cdot \ln \text{ Avg Price}, \; r^2 = 0.94, \; s_e = 0.06$$

- Here, comparing r^2 or s_e is not appropriate. Why?

- To obtain a meaningful comparison, show the fit implied by both equations in one scatter plot

Comparing the linear and nonlinear equations



Cat Food: What is the optimal price?

- From your microeconomics course (ØASØK 1000), the optimal price is based on the **elasticity**

$$\text{Optimal Price} = \text{Cost} \cdot \frac{\text{Elasticity}}{\text{Elasticity} + 1}$$

- $\text{Elasticity} = \frac{\%\text{ change in } y}{\%\text{ change in } x}$
- We get the elasticity from b_1 in a log-log regression
- We find $b_1 = -2.41$, so a 1% increase in price is associated with 2.41% decrease in quantity sold

Cat Food: What is the optimal price?

- Suppose that one can of cat food costs Rema 1000 kr. 2.70 to produce

$$\text{Optimal Price} = \text{Cost} \cdot \frac{\text{Elasticity}}{\text{Elasticity} + 1}$$

$$\text{Optimal Price} = 2.70 \cdot \frac{-2.41}{-2.41 + 1} = 4.61$$

- Rema 1000's profit is maximized when the price is kr. 4.61

Summary: Linear model vs. Non-linear transformations

- Visualize your data
  - Check if there is **curvature** in the relationship between x and y
- Use your intuition and economic knowledge
  - Use the natural log transformation when you think **percentage changes** matter
  - Examples: price and quantity demanded, measuring profits and productivity, etc.