ØAMET2200
Business Decision Making Using Data
Lecture 6

Instructor: Fenella Carpena

October 4, 2019

Announcements

- Final exam aids, see `https://student.oslomet.no/hjelpemidler-ordbok-kalkulatorreglement-hhs`
  - Calculator (as specified in regulations for use of calculator)
  - Dictionary (Native language-English/English-Native language or English-English)
  - One sheet of notes (A4-size, single-sided)
- No lecture on Oct. 11 (next week)
  - Problem set due dates and syllabus will be adjusted accordingly
  - Check Canvas
- No office hours today and next week (week 41)
- Mid-semester evaluation

Part 4 of this Course: Building regression models

- You want to build a model to make a **forecast** or a **prediction** from data you have
- What is the best **model you can build**?
- This lecture: **statistical inference** in OLS regressions
- Next lectures: **multivariate models**

Agenda for Today

- Setting up the simple regression model: theory
- Inference in regression models
- Prediction intervals
- Three potential problems affecting regression models
  1. Changing variation in the data
  2. Outliers
  3. Dependence among observations in time series data
- Chapters 21, 22

The Big Picture: Why do we focus on regression models?

- Pick a job, I'll give you a regression
- Consulting: How does competition affect price?
  `regress price competition`
- Marketing: If we increase spending on advertising by 10,000 kroner, how much would revenues increase?
  `regress revenues ad_spending`
- Finance: Does a company's sales influence its stock price?
  `regress stock_price quarterly_sales`
- Average effects of x on y are often important for decision-making in many organizations
- Regressions are powerful tools for studying these relationships

Case Study for Today: Baker Hansen

- You are in charge of deciding the location of a new Baker Hansen store
- You are considering a variety of different possible locations with different foot traffic levels
- Does foot traffic affect sales?
- One potential site has on average of 40,000 people walking by per month while another site has 32,000. How much more can we expect to sell at the busier location?

Case Study for Today: Baker Hansen

- You have (fictional) data for 80 Baker Hansen locations
- Assume that all locations charge the same price and are located in similar neighborhoods
- For each location, you have data from last month
  - `sales`: total sales in thousands of kroner
  - `foot_traffic`: # of people walking by the store in thousands
- You regress `sales` on `foot_traffic` and obtain b_1
- Want to analyze the following to determine the best location
  1. How **confident** are we in this relationship?
  2. What is the **95% CI** for the estimated slope?
  3. What is the 95% **prediction interval** for the level of sales for a given location with 50,000 people walking by?

What's new in this lecture?

- Last week, we used regressions as a **descriptive** tool
- **Key point for today:** We consider the data and the estimated OLS line to be a **sample** from a **population**
- Because our data is only a sample from a larger population, we need to distinguish between the **sample regression line** and the **population regression line**
- If we had a different sample, how different would the results be?
- We will apply statistical tools to make an **inference** about the population relationship

The Simple Regression Model

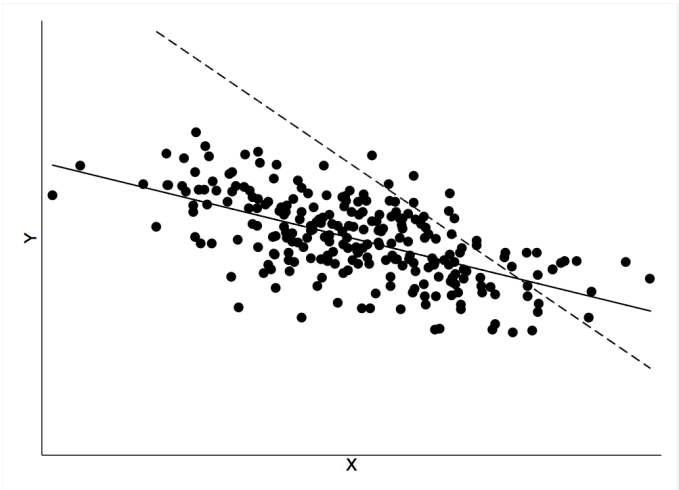
- The **Simple Regression Model (SRM)** is a model for the **population regression line** for x and y

$$y = \beta_0 + \beta_1 x + \epsilon$$

- β_0 and β_1 are **unknown** and are **constants**
- The statistical problem we have is to estimate β_0 and β_1 using **a sample of data** on x and y
- b_0 and b_1 from last lecture are our **estimates** for β_0 and β_1
- b_0 and b_1 are **random variables** and have a **sampling distribution**. Why?
- ϵ is the **population error term**

Worksheet Exercise

Consider the population regression line $y = \beta_0 + \beta_1 x + \epsilon$. The figure below shows data from a sample of 250 observations of x and y. One of the lines is the sample regression line, $b_0 + b_1 x$; the other is the population regression line $\beta_0 + \beta_1 x$. Is the sample regression line solid or dashed? Explain.



SRM Assumptions

The SRM makes **five assumptions** about the relationship in the population between x and y

1. The conditional mean of y given x is **linear**, $\mu_{y|x} = \beta_0 + \beta_1 x$
   - Notation: $\mu_{y|x}$ is the same as $E(Y|X = x)$
   - $\mu_{y|x}$ describes our best guess for y given a particular x
   - Example: $\mu_{y|200} =$ _____
2. The expected value of the population error is **zero**, $E(\epsilon) = 0$
   - The SRM recognizes that individual responses will different from their conditional means: $y = \beta_0 + \beta_1 x + \epsilon$
   - ϵ captures deviations of responses around the conditional mean
   - Errors can be positive or negative, but is zero on average

SRM Assumptions

The last three SRM assumptions concern the population error ϵ

3. The errors are **independent** of each other

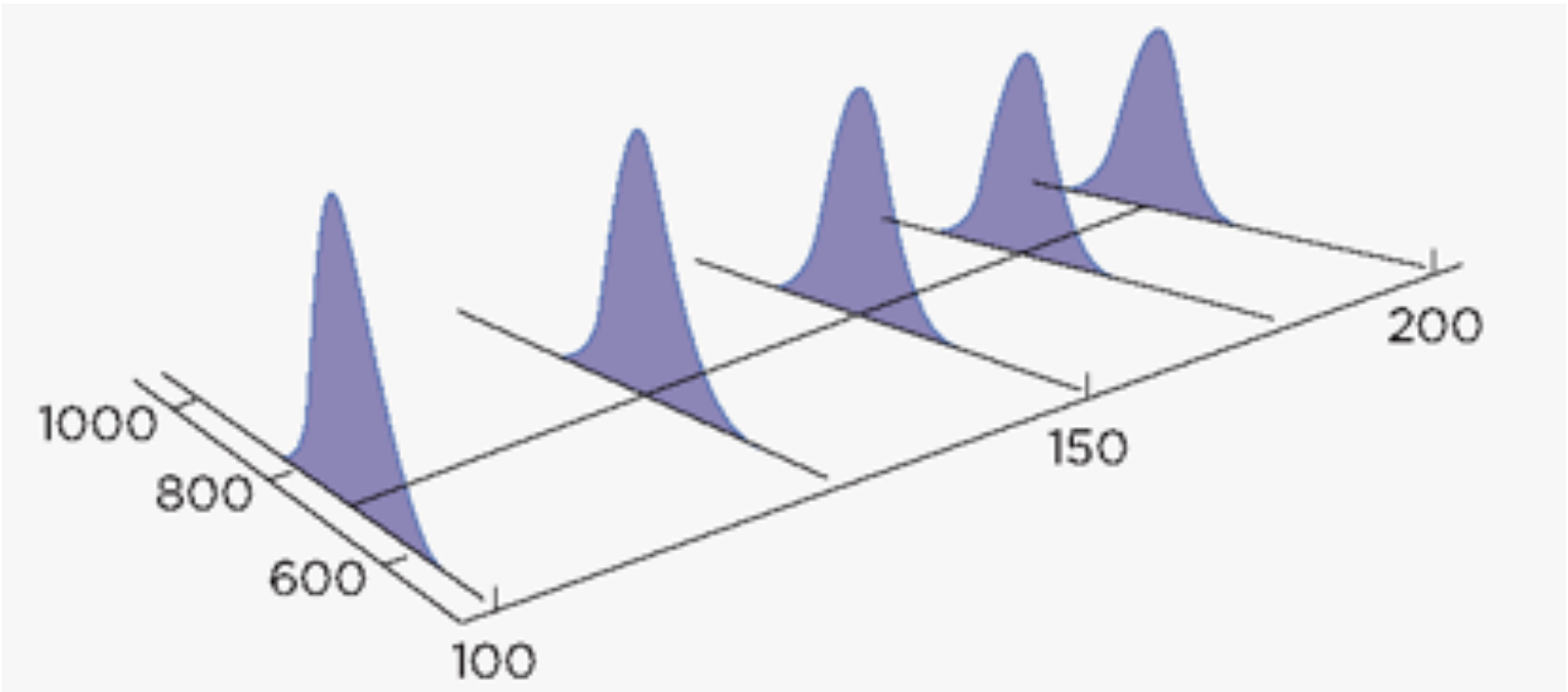
4. The errors are **normally distributed**

5. The errors all have the **same** variance, denoted $var(\epsilon) = \sigma_\epsilon^2$

Together, Assumptions 2-5 imply that $\epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2)$

Assumptions 2-5 Visually

The SRM assumes a normal distribution of y at each x around a mean predicted by the line



Estimates vs. Parameters

Data		**SRM**
b_0	Intercept	β_0
b_1	Slope	β_1
$\hat{y}$	Line	$\mu_{y\|x}$
e	Deviation	ε
s_e	SD(ε)	σ_ε

Conditions for the SRM

We **never know for sure** if the SRM is the correct model for the population relationship of x and y. We only observe a sample.

The best we can do is to **check in our data** if conditions similar to the SRM assumptions hold in the sample.

1. The relationship between x and y is linear in the sample
2. No obvious lurking variable
3. Residuals are evidently independent
4. Variances of the residuals are similar
5. Residuals have a nearly normal distribution

Example: Baker Hansen

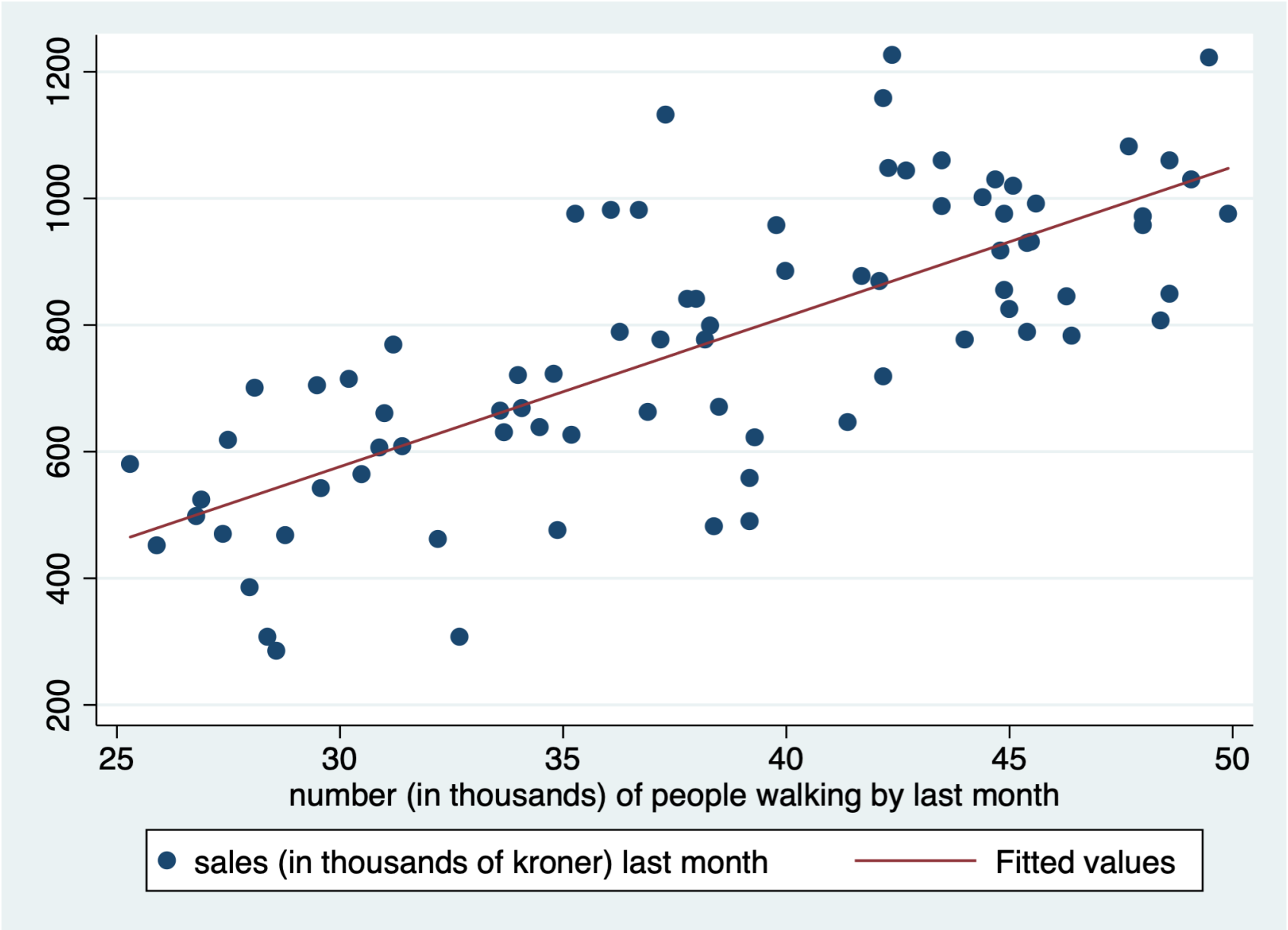
```
. regress sales foot_traffic ;
```

Source	SS	df	MS
Model	2148270.28	1	2148270.28
Residual	1767674.71	78	22662.4963
Total	3915944.99	79	49568.9239

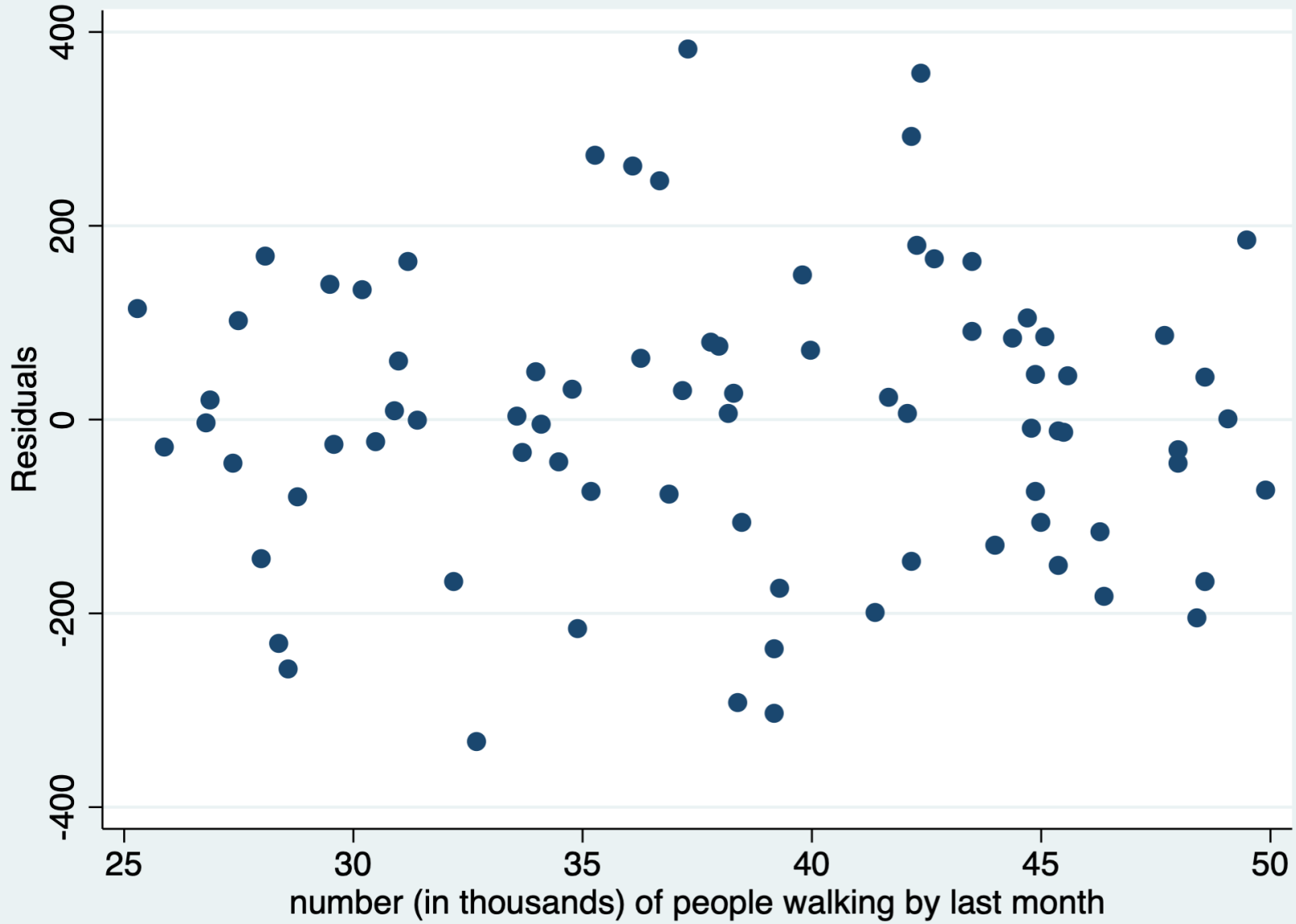
Number of obs	80
F(1, 78)	94.79
Prob > F	0.0000
R-squared	0.5486
Adj R-squared	0.5428
Root MSE	150.54

sales	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
foot_traffic	23.67286	2.431421	9.74	0.000	18.83228	28.51345
_cons	-133.8097	94.58436	-1.41	0.161	-322.1127	54.49326

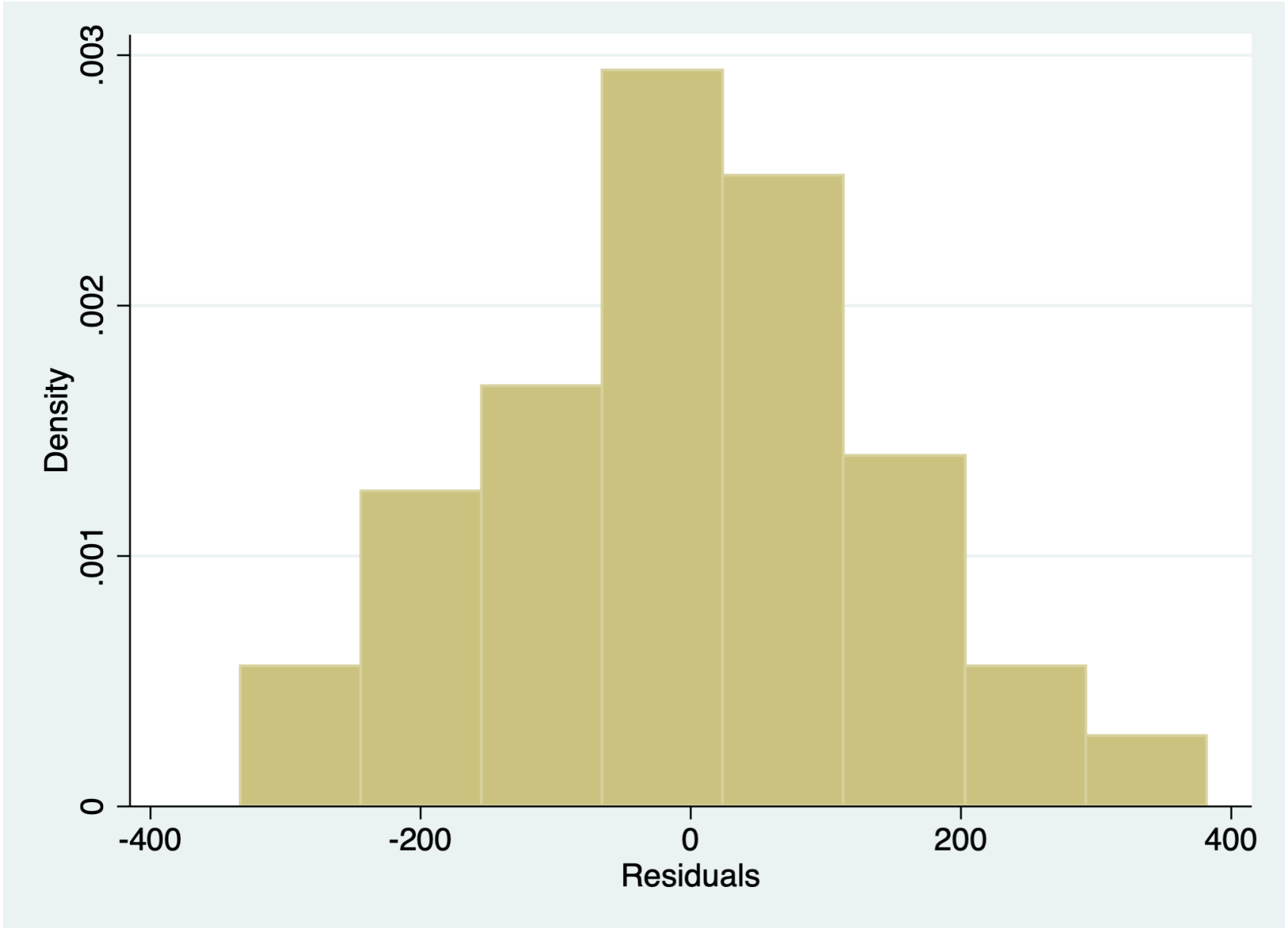
Baker Hansen: Scatter Plot and OLS Line



Baker Hansen: Residual Plot



Baker Hansen: Histogram of Residuals



Example: Baker Hansen

(a) Interpret b_0 and b_1.

(b) Do the five SRM conditions hold?

Example: Baker Hansen

(c) Consider the following situations. Explain whether they violate the SRM of `sales` and `foot_traffic`. If so, which SRM condition is violated?

(i) Baker Hansen shops located in busier areas (more foot traffic) also have better customer service.

(ii) The variation in sales is lower in Baker Hansen shops located in busier areas.

(iii) The effect of a 1,000 increase in foot traffic on sales is larger in busier areas.

Why do we care about the SRM conditions?

- If the 5 SRM assumptions hold, then b_0 and b_1 are appropriate estimators for β_0 and β_1
- If the 5 SRM assumptions hold, the sampling distribution of b_0 and b_1 are approximately normal
- This means that when conducting hypothesis tests on b_0 and b_1, we can use the t-distribution (or for very large samples, the normal distribution)
- The point of all of this is to be able to say something about the population relationship from one sample ("statistical inference")

Inference in Regression

- Let's now consider how to make **inferences** about the population relationship between x and y using our sample
- Here are some questions we will try to answer:
  - Is the observed relationship between x and y in the sample **strong** enough to conclude that it also holds in the population?
  - How can we use the sample statistics b_0 and b_1 to determine a plausible **range** of values for β_0 and β_1 of the population regression line?
  - What **interval** predicts the value of y for a given value of x?
- We need the **standard errors** $se(b_0)$ and $se(b_1)$
- Why do b_0 and b_1 have standard errors?

SE of the slope in SRM

$$se(b_1) = \frac{s_e}{\sqrt{n-1}} \cdot \frac{1}{s_x}$$

- Smaller $se(b_1)$ means that our estimate b_1 for the population slope β_1 is **more precise**
- When is $se(b_1)$ smaller?
  - $\downarrow s_e \implies \downarrow se(b_1)$. Why?
  - $\uparrow n \implies \downarrow se(b_1)$. Why?
  - $\uparrow s_x \implies \downarrow se(b_1)$. Why?

SE of the intercept in SRM

$$se(b_0) = s_e \sqrt{\frac{1}{n} + \frac{\overline{x}^2}{(n-1)s_x^2}} \approx \frac{s_e}{\sqrt{n}} \cdot \sqrt{1 + \frac{\overline{x}^2}{s_x^2}}$$

- Less emphasis in this course
- When is $se(b_0)$ smaller?
  - $\downarrow s_e \implies \downarrow se(b_0)$
  - $\uparrow n \implies \downarrow se(b_0)$
  - $\uparrow s_x \implies \downarrow se(b_0)$
  - $\uparrow \overline{x} \implies \uparrow se(b_0)$

$se(b_0)$ and $se(b_1)$

In practice, statistics software calculate $se(b_0)$ and $se(b_1)$ automatically, so they are provided to you in the output of the regression table.

```
. regress sales foot_traffic ;
```

Source	SS	df	MS
Model	2148270.28	1	2148270.28
Residual	1767674.71	78	22662.4963
Total	3915944.99	79	49568.9239

Number of obs	80
F(1, 78)	94.79
Prob > F	0.0000
R-squared	0.5486
Adj R-squared	0.5428
Root MSE	150.54

sales	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
foot_traffic	23.67286	2.431421	9.74	0.000	18.83228	28.51345
_cons	-133.8097	94.58436	-1.41	0.161	-322.1127	54.49326

Hypothesis Tests in SRM

$$t\text{-stat} = \frac{\text{sample statistic} - \text{value from null hypothesis}}{\text{se(sample statistic)}}$$

- $H_0 : \beta_1 = 0, H_1 : \beta_1 \neq 0$

$$t\text{-stat} = \frac{b_1 - 0}{se(b_1)}$$

- $H_0 : \beta_1 \geq 4, H_1 : \beta_1 < 4$

$$t\text{-stat} = \frac{b_1 - 4}{se(b_1)}$$

- $H_0 : \beta_0 = 2, H_1 : \beta_0 \neq 2$

$$t\text{-stat} = \frac{b_0 - 2}{se(b_0)}$$

The t-distribution has $n - 2$ degrees of freedom

Example: Baker Hansen

(d) Use a t-statistic to determine if the slope is statistically significantly different from zero at the 5% level.

(e) What is the p-value of the hypothesis test in part (d)?

Confidence Intervals in SRM

sample statistic $\pm$ two-sided critical value $*$ se(sample statistic)

- The **95% CI** for β_1 is

$$b_1 \pm t_{0.025,n-2} * se(b_1)$$

- The **95% CI** for β_0 is

$$b_0 \pm t_{0.025,n-2} * se(b_0)$$

- The **90% CI** for β_0 is

$$b_0 \pm t_{0.05,n-2} * se(b_0)$$

- Example: What is $t_{0.025,34}$?

Exercise: Baker Hansen

(f) What is the 95% CI for β_1?

(g) What is the 80% CI for β_0?

Equivalent inferences using CI, t-stat, p-value

- For a two-sided hypothesis test, the t-stat, p-value, and CI will all reach **the same** conclusions
- The difference between the three lies in the **types** of information they contain
- The **CI** gives a list of all hypothesized values we would reject or fail to reject
- The hypothesis test using t-**stat** tells us whether we reject or fail to reject only one particular hypothesized value
- The p-**value** gives us the smallest significance level α for which we can reject the null

Prediction Intervals

- What would we predict for sales at a location with 50,000 people who walk by every month?
- Our best guess is _____
- Could sales be as low as as 500,000 kr? As high as 2,000,000 kr? How accurate is our prediction?
- To answer these questions, we need a prediction interval

Prediction Intervals

- A **prediction interval** is an interval designed to hold a fraction (usually 95%) of the values of the response y for a given value of x (denoted x_{new})
- The **95% prediction interval** is

$$\widehat{y} \pm t_{0.025,n-2} * se(\widehat{y})$$

where $se(\widehat{y}) = s_e\sqrt{1 + \frac{1}{n} + \frac{(x_{\text{new}} - \overline{x})^2}{(n-1)s_x^2}}$

- Typically, we'll rely on the **approximation**

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

which is reasonable if we have large n and are not extrapolating too much

Prediction Intervals

- The "prediction interval" differs from a "confidence interval" because we are making a statement about the location of a **new observation**, rather than a parameter of the population
- Prediction intervals are reliable only within the **range of observed data**
- They are sensitive to the **constant** variance and **normality** assumptions

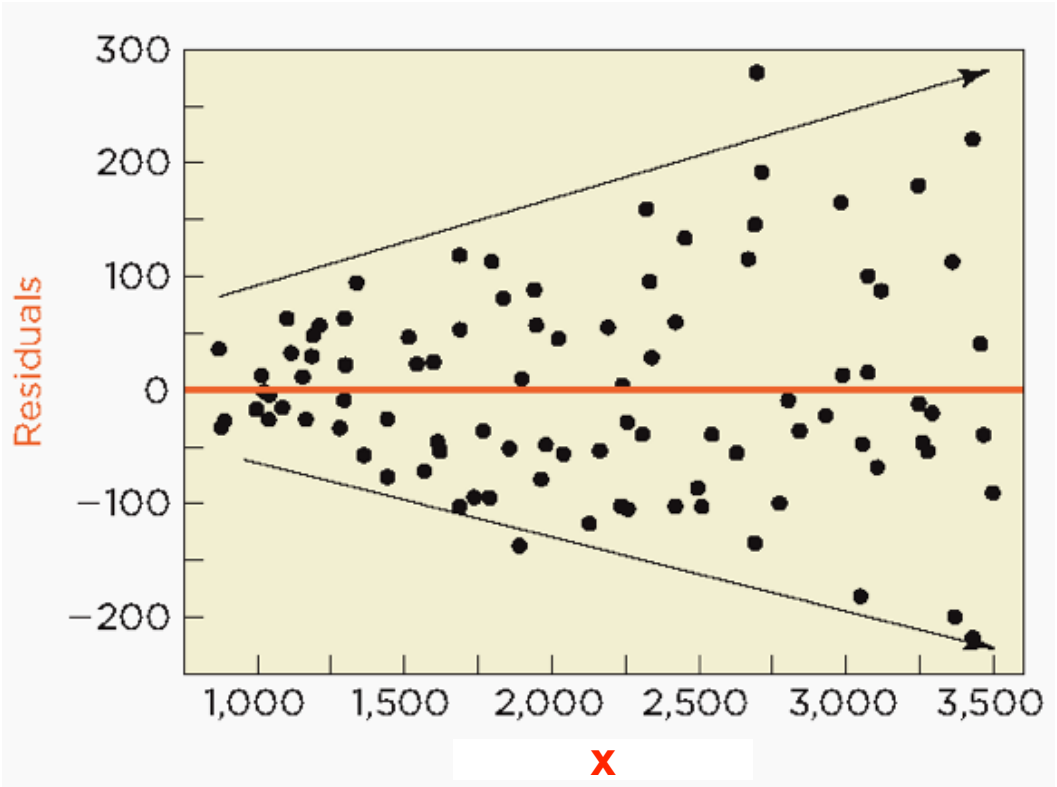
Example: Baker Hansen

(h) What is the 95% prediction interval for the monthly sales of a location with foot traffic level of 50,000?

Regression Diagnostics

- Three potential **problems** affecting regression models:
  1. Changing variation in the data
  2. Outliers
  3. Dependence among observations (time series data)
- We'll now examine:
  - The **consequences** of these problems
  - How to **detect** them
  - How to **solve** them

1. Changing variation: How to detect?



- Can detect using scatterplot of residuals vs. the x variable
- **Heteroscedastic**: errors have different variance
- **Homoscedastic**: errors have the same variance
- SRM assumes **homoskedasticity**

1. Changing variation: Consequences & Solutions

Consequences

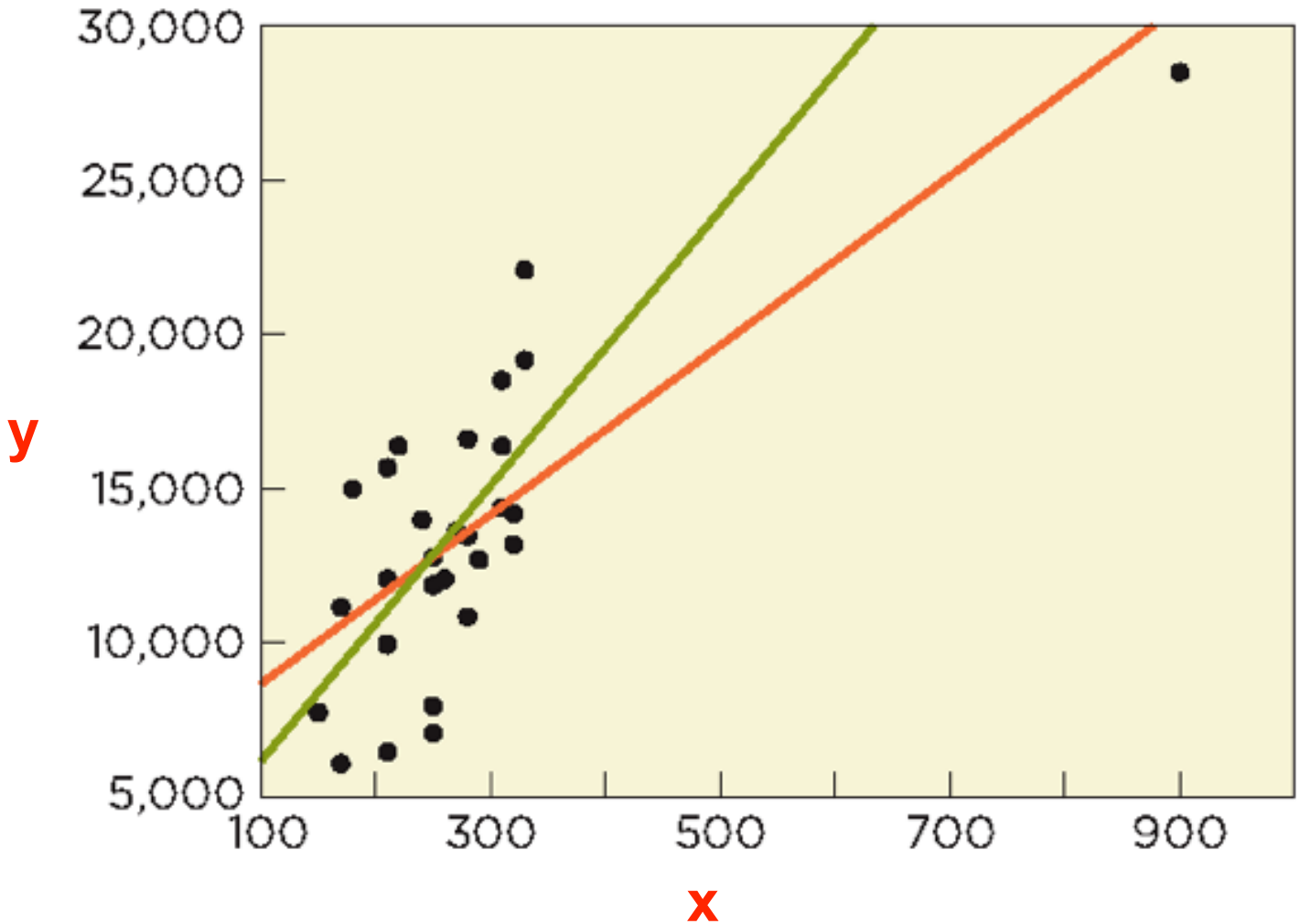
- Prediction intervals are too wide or too narrow
- Confidence intervals for the slope and intercept are not reliable
- Hypothesis tests regarding β_0 and β_1 are not reliable

Solutions

- Use standard errors that are robust to heteroskedasticity
- Stata: `regress y x, robust`

2. Outliers: How to detect?

Can detect using scatterplot of y vs. x



2. Outliers: Consequences

- To see the consequences of an outlier, fit the OLS line both with and without it
- Use the SEs obtained excluding the outlier to compare estimates
- Estimates of b_0 (or b_1) with vs. without the outlier that differ by fractions of an SE aren't so far apart
- For example
  - With outlier: $b_0 = 5887$, $se(b_0) = 1400$
  - Without outlier $b_0 = 1558$, $se(b_0) = 2877$
  - Dropping the outlier shifts b_0 by

$$\frac{5887 - 1558}{2877} \approx 1.50 \text{ standard errors}$$

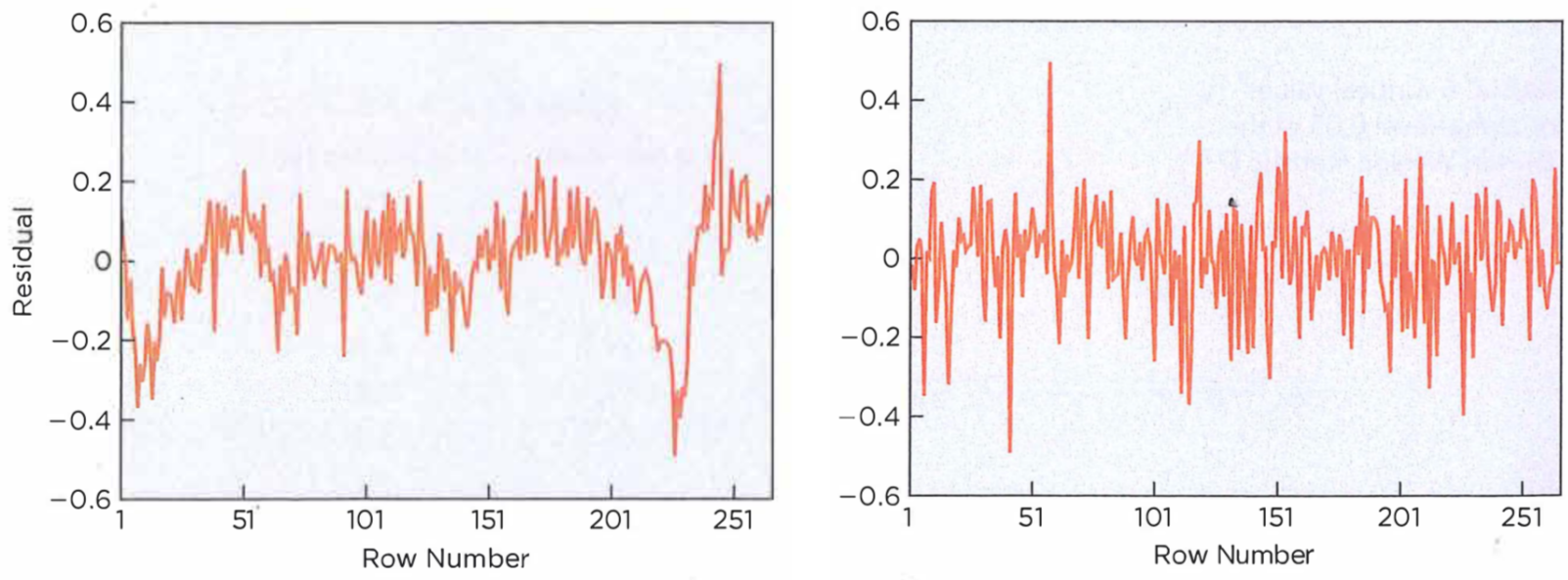
- Stata: Combine `regress` with `if` condition to exclude outliers

2. Outliers: Solutions

- Key issue: should the outlier be used in the OLS regression?
- If the outlier is representative and describes what is expected **next time** under the same conditions, it should be included
- If the outlier is the result of a mistake in the data, it should be excluded
- Bottom line: additional information may be needed to make the decision
- **Always** state in your summary if any data was excluded in the analysis and **explain** why

3. Dependent errors (time series data): How to detect?

- Random sample yields observations that are independent
- Time series data are likely to be dependent
- Can detect by plotting residuals vs. time



3. Dependent errors (time series data): How to detect?

- Can also detect using the Durbin-Watson (DW) test
- The DW statistic tests $H_0 : \rho_e = 0$ where $\rho_e = corr(\epsilon_t, \epsilon_{t-1})$
- The correlation between consecutive observations in a time series is called **autocorrelation**
- The DW statistic is calculated as

$$D = \frac{(e_2 - e_1)^2 + (e_3 - e_2)^2 + \ldots + (e_n - e_{n-1})^2}{e_1^2 + e_2^2 + \ldots + e_n^2}$$

- In practice, DW-stat provided by software

3. Dependent errors (time series data): How to detect?

Compare DW-stat to the table of critical values (at $\alpha = 0.05$)

	Reject H_0: $\rho_\varepsilon = 0$ if	
n	D is less than	D is greater than
15	1.36	2.64
20	1.41	2.59
30	1.49	2.51
40	1.54	2.46
50	1.59	2.41
75	1.65	2.35
100	1.69	2.31

3. Dependent errors (time series data): Consequences & Solutions

Consequences

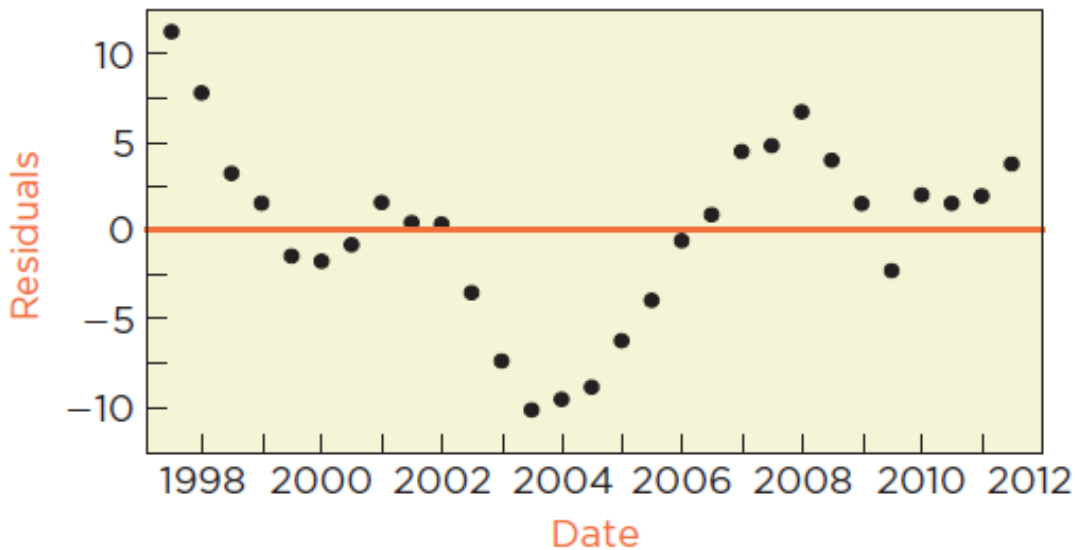
- If there is positive autocorrelation in the errors, the estimated standard errors are too small
- The estimated slope and intercept are less precise than suggested by the output

Solutions

- Best remedy is to incorporate the dependence into the regression model

3. Dependent errors (time series data): Example

- Suppose we have year time series data for 29 years
- Why does this residual plot indicate dependence?



- DW statistic = 0.25. What does the DW test conclude?

Best Practices

- Always check the conditions for the SRM
- Be aware of potential problems affecting SRM and the solutions to these problems
- Rely on software to calculate standard errors for regression estimates
- Use prediction intervals to predict ! for particular observations
- Be careful when extrapolating