Lecture 10 – Regression and Time-series Analysis

Dr. Jing Li
Department of Computing
The Hong Kong Polytechnic University

29 & 31 Mar 2022

Roadmap

- **Linear Regression**
  - Scenarios where linear regression will be helpful!
  - How to fit a line with least squares.
  - Residuals and homoscedasticity in fitting a line.
  - More complex data and more complex model.
  - How to measure the fit.
- **Time-series Analysis**
  - Varying time-series
  - Objectives of time-series analysis
  - Time-series vs. Regression

Roadmap

- **Linear Regression**
  - Scenarios where linear regression will be helpful!
  - How to fit a line with least squares.
  - Residuals and homoscedasticity in fitting a line.
  - More complex data and more complex model.
  - How to measure the fit.
- **Time-series Analysis**
  - Varying time-series
  - Objectives of time-series analysis
  - Time-series vs. Regression

Roadmap

- **<u>Linear Regression</u>**
  - *<u>Scenarios where linear regression will be helpful!</u>*
  - How to fit a line with least squares.
  - Residuals and homoscedasticity in fitting a line.
  - More complex data and more complex model.
  - How to measure the fit.
- Time-series Analysis
  - Varying time-series
  - Objectives of time-series analysis
  - Time-series vs. Regression

Bordeaux Wine

- Large differences in price and quality between years, although wine is produced in a similar way
- Meant to be aged, so hard to tell if wine will be good when it is on the market
- Expert tasters predict which ones will be good.
- **QUESTION**. Can analytics be used to come up with a different system for judging wine?

Bordeaux Wine

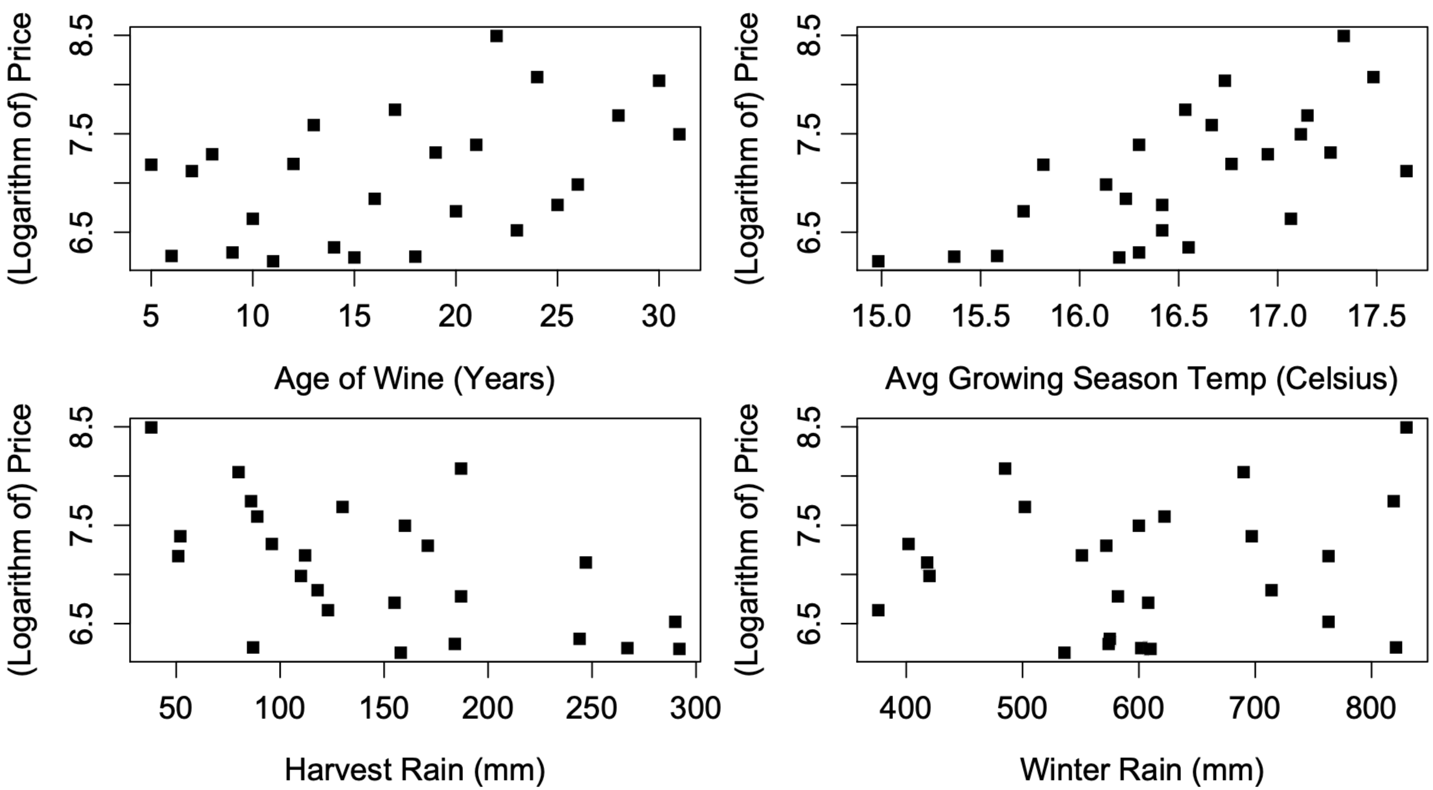
- *Orley Ashenfelter*, a Princeton economics restrictions professor, claims he can predict wine quality without tasting the wine in March 1990.
- *Robert Parker*, the world's most influential wine expert comments on *Ashenfelter*:
  - *"Ashenfelter is an absolute total sham"*
  - *"Rather like a movie critic who never goes to see the movie but tells you how good it is based on the actors and the director"*

Linear Regression vs. Bordeaux Wine

- **Two types of variables are designed:**
  - *Dependent variable* y:
    - typical price in 1990-1991 wine auctions (approximates quality)
  - *Independent variable* x:
    - Age (*older wines are more expensive*)
    - Weather
    - Average Growing Season Temperature
    - Harvest Rain
    - Winter Rain

Linear Regression vs. Bordeaux Wine

- **Relations between price (y) and diverse factors (x):**



Linear Regression

- Suppose we have collected *bivariate data* (x_i, y_i), $i = 1, \ldots, n$.
- **Goal**: to model the relationship between x and y by finding a function $y = f(x)$ that is a close fit to the data.
- **Assumptions**: x_i is *NOT* random and that y_i is a function of x_i plus some random noise.
  - x is called the *independent or predictor variable*
  - y is called the *dependent or response variable*.

Linear Regression (Example)

- **Example 1**. The cost of a first-class stamp in cents over time:

.05 (1963)	.06 (1968)	.08 (1971)	.10 (1974)	.13 (1975)	.15 (1978)	.20 (1981)	.22 (1985)
.25 (1988)	.29 (1991)	.32 (1995)	.33 (1999)	.34 (2001)	.37 (2002)	.39 (2006)	.41 (2007)
.42 (2008)	.44 (2009)	.45 (2012)	.46 (2013)	.49 (2014)			

- Using the R function *lm()* we found the *least squares fit* for a line to this data is:
  - $y = -0.06558 + 0.87574x$
  - where x is the number of years since 1960 and y is in cents.

PREDICT: What is the price for 2016 stamp?

Linear Regression (Example)

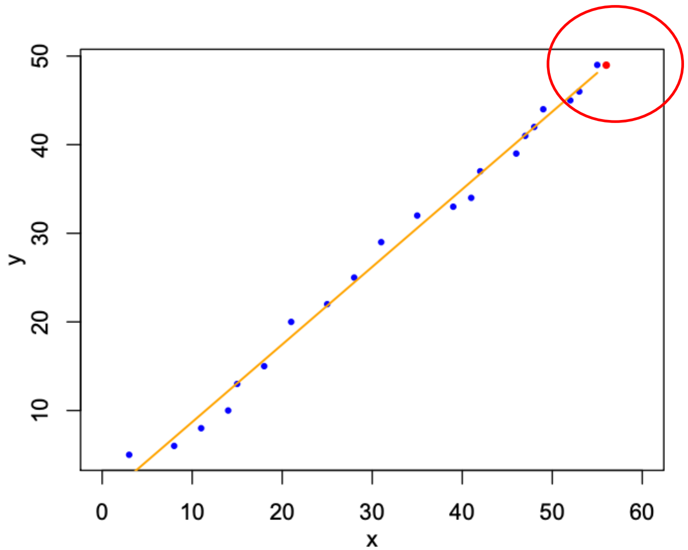
- **Example 1**. The cost of a first-class stamp in cents over time:
  - $y = -0.06558 + 0.87574x$
  - where x is *the number of years since 1960* and y is in cents.
- To predict the price for 2016 stamp, we let $x = 56$, then $y \approx 48.98$
- How to further analyze the relations of x and y?
  - Visualize the data in graphs!

QUESTION. What graph to choose? Barplots, Histograms, or scatterplots?

Linear Regression (Example)

- **Example 1**. The cost of a first-class stamp in cents over time:
  - $y = -0.06558 + 0.87574x$
  - where x is *the number of years since 1960* and y is in cents.

Red plot: the predicted price of the stamp in 2016



Stamp cost (cents) vs. time (years since 1960)

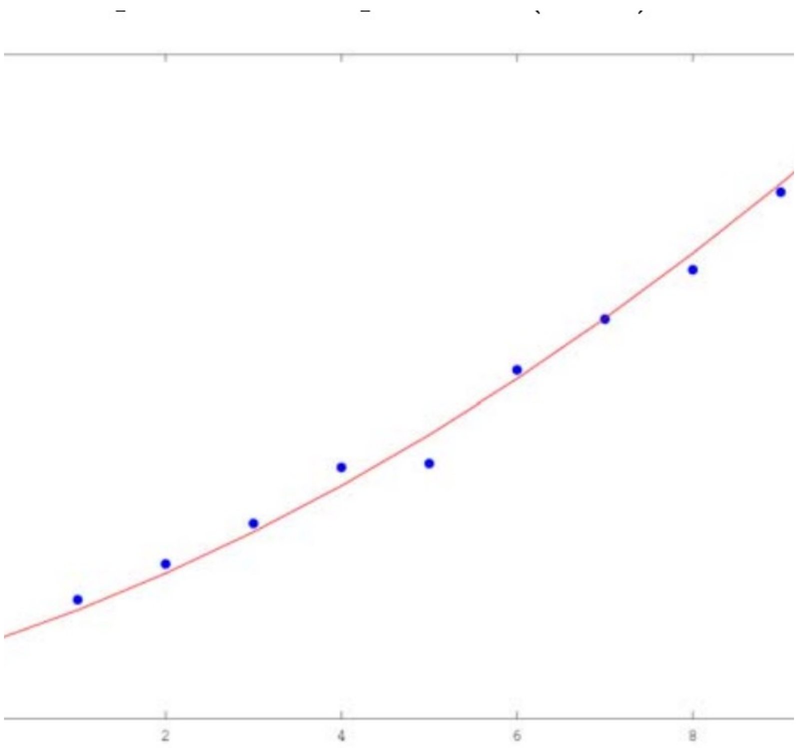
Observation.
- None of the data points actually lie on the line.
- Rather this line has the 'best fit' with respect to all the data, with a small error for each data point.

Linear Regression (More Examples)

- **Example 2**. Suppose we have n pairs of fathers and adult sons.
  - Let x_i and y_i be the heights of the i-th father and son, respectively.
  - The least squares line for this data could be used to predict the adult height of a young boy from that of his father.

Linear Regression (More Examples)

- **Example 3**. We are not limited to best fit lines (sometimes more complex model needed!).
  - For all positive d, the method of least squares may be used to find a polynomial of degree d with the *best fit* to the data.
  - Right hand side figure shows the least squares fit of a parabola (d = 2).



Fitting a parabola, $b_2x^2 + b_1x + b_0$, to data

Roadmap

- **Linear Regression**
  - Scenarios where linear regression will be helpful!
  - *How to fit a line with least squares.*
  - Residuals and homoscedasticity in fitting a line.
  - More complex data and more complex model.
  - How to measure the fit.
- **Time-series Analysis**
  - Varying time-series
  - Objectives of time-series analysis
  - Time-series vs. Regression

Fitting a Line with Least Squares

- Suppose we have several data (x_i, y_i) as above.
- **Goal**. find a line $y = \beta_1 x + \beta_0$ *best fitting* the data.

regression coefficient for the independent variable *intercept coefficient*

- **Assumption**. Each y_i is predicted by x_i up to some error ϵ_i:

Real value *Predicted value*

Error

  - $y_i = \beta_1 x_i + \beta_0 + \epsilon_i$

- **QUESTION**. How to find out the values of β_1 and β_0?

Fitting a Line with Least Squares

- **Goal**: find a line $y = \beta_1 x + \beta_0$ *best fitting* the data.
- **Assumption**: Each y_i is predicted by x_i up to some error ϵ_i:
  - $y_i = \beta_1 x_i + \beta_0 + \epsilon_i$
- **Errors**. The sum of the square errors:
  - $S(\beta_0, \beta_1) = \sum_i \epsilon_i^2 = \sum_i (y_i - \beta_1 x_i - \beta_0)^2$
  - The method of least squares finds the values $\hat{\beta}_0$ and $\hat{\beta}_1$ of β_0 and β_1 that minimize $S(\beta_0, \beta_1)$, the sum of the squared errors.
- **QUESTION**. How to find out $\hat{\beta}_0$ and $\hat{\beta}_1$?
  - **Hint**. Use the methods in *Calculus*!

Fitting a Line with Least Squares

- **Errors**. The sum of the square errors:
  - $S(\beta_0, \beta_1) = \sum_i \epsilon_i^2 = \sum_i (y_i - \beta_1 x_i - \beta_0)^2$
  - The method of least squares finds the values $\hat{\beta}_0$ and $\hat{\beta}_1$ of β_0 and β_1 that minimize $S(\beta_0, \beta_1)$, the sum of the squared errors.
- $\hat{\beta}_1 = \frac{s_{xy}}{s_{xx}}$ and $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$

n is the number of data points

  - $\bar{x} = \frac{1}{n} \sum_i x_i$ and $\bar{y} = \frac{1}{n} \sum_i y_i$

Sample Mean

  - $s_{xx} = \frac{1}{n-1} \sum_i (x_i - \bar{x})^2$ and $s_{xy} = \frac{1}{n-1} \sum_i (x_i - \bar{x})(y_i - \bar{y})$

Sample Variance

Sample Covariance of x and y

Fitting a line with Least Squares

- Use least squares to fit a line to the following three data points: (0,1), (2,1), and (3,4).
  - So, $(x_1, y_1) = (0,1)$, $(x_2, y_2) = (2,1)$, $(x_3, y_3) = (3,4)$.
- **QUESTION**.
  - What are $\bar{x}, \bar{y}, s_{xx}, s_{xy}$?
  - What are $\hat{\beta}_0$ and $\hat{\beta}_1$?

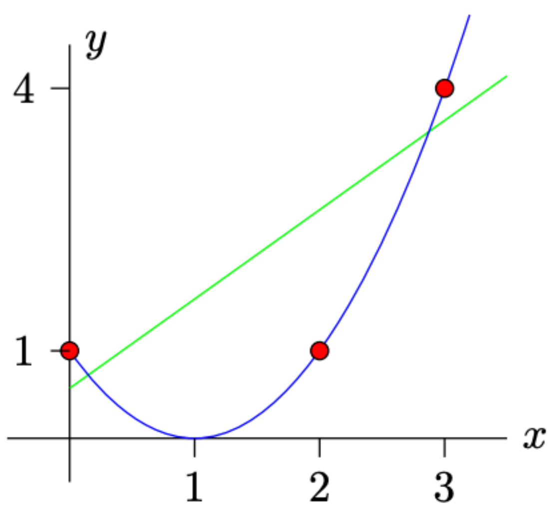
Fitting a line with Least Squares

- Use least squares to fit a line to the following three data points: (0,1), (2,1), and (3,4).
  - So, $(x_1, y_1) = (0,1)$, $(x_2, y_2) = (2,1)$, $(x_3, y_3) = (3,4)$.
- **QUESTION**.
  - $\bar{x} = \frac{5}{3}, \bar{y} = 2, s_{xx} = \frac{7}{3}, s_{xy} = 2$
  - $\hat{\beta}_0 = \frac{4}{7}$ and $\hat{\beta}_1 = \frac{6}{7}$
  - So the least squares line has equation:
    - $y = \frac{6}{7}x + \frac{4}{7}$

Fitting a line with Least Squares

- Use least squares to fit a line to the following three data points: (0,1), (2,1), and (3,4).
  - So, $(x_1, y_1) = (0,1)$, $(x_2, y_2) = (2,1)$, $(x_3, y_3) = (3,4)$.
  - The least square line: $y = \frac{6}{7}x + \frac{4}{7}$.

Which one would you prefer?

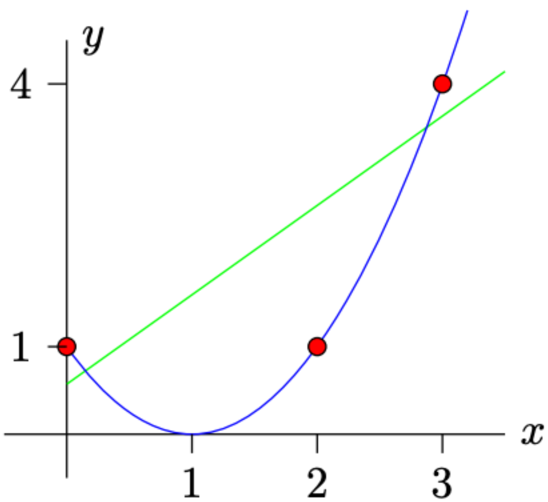


Least square fit of a line

Least square fit of a parabola

Fitting a line with Least Squares

- **Notes**.
  - The word ***"linear"*** in *linear regression* does not refer to fitting a line, though it's the most common curve to fit.
  - When we *fit a line* to *bivariate data*, it is called ***simple linear regression***.



Which one would you prefer?

Least square fit of a line

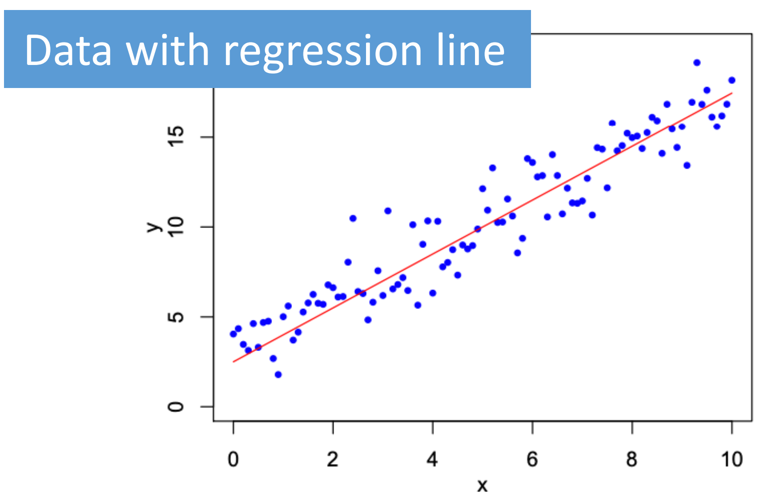
Least square fit of a parabola

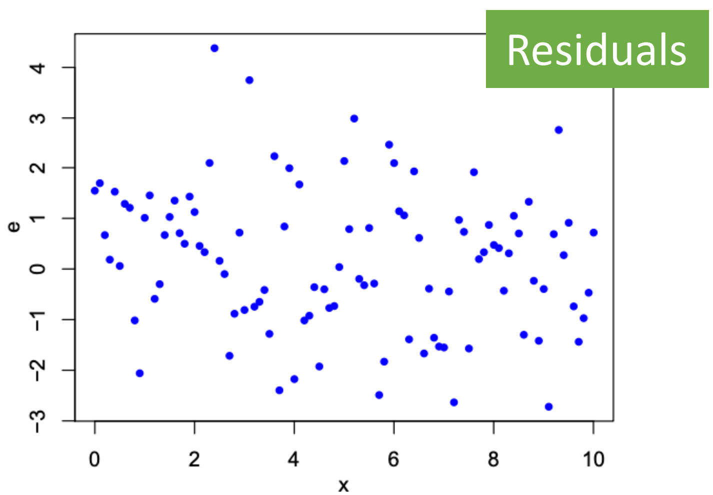
Roadmap

- **Linear Regression**
  - Scenarios where linear regression will be helpful!
  - How to fit a line with least squares.
  - *Residuals and homoscedasticity in fitting a line.*
  - More complex data and more complex model.
  - How to measure the fit.
- **Time-series Analysis**
  - Varying time-series
  - Objectives of time-series analysis
  - Time-series vs. Regression

Residuals in Fitting a Line

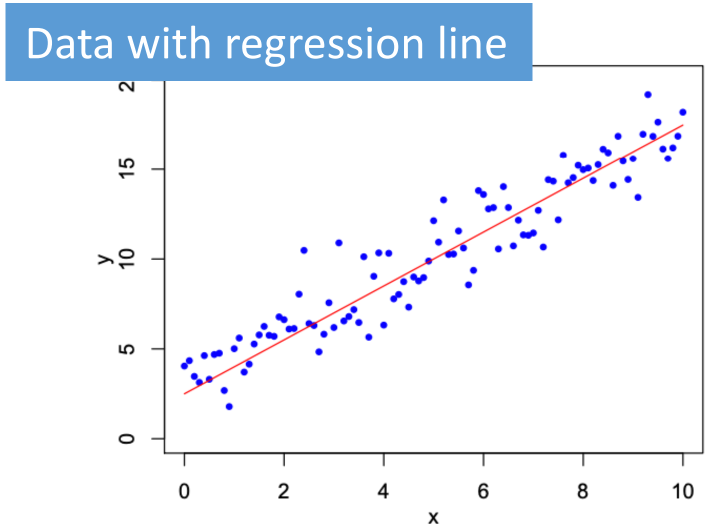
- Suppose the model is $y_i = \hat{\beta}_1 x_i + \hat{\beta}_0 + \epsilon_i$.
  - $\hat{\beta}_1 x_i + \hat{\beta}_0$ as the predicting or explaining y_i
  - The left-over term ϵ_i is called the ***residual***.
  - Residuals as *random noise* or *measurement error*
- When plot the residuals out, the data points should hover near the regression line. The residuals should look about the same across the range of x.

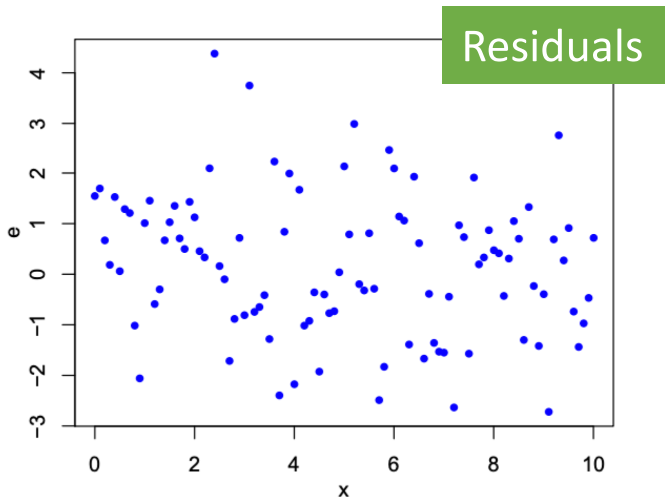




Residuals in Fitting a Line

- When plot the residuals out, the data points should hover near the regression line. The residuals should look about the same across the range of x.



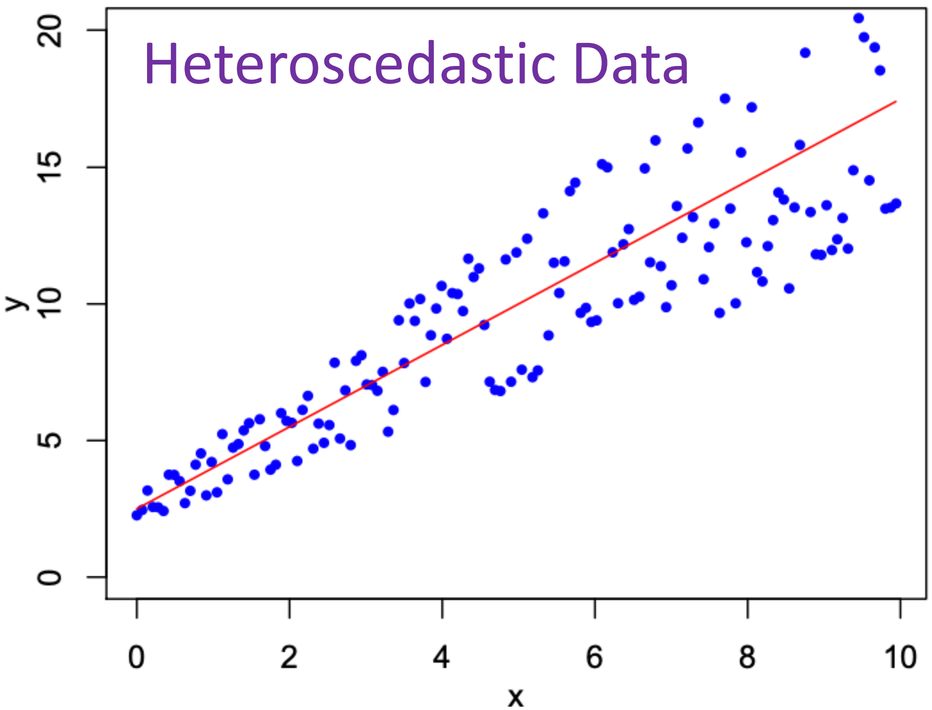


The data hovers in the band of fixed width around the regression line.

At every x the residuals have about the same vertical spread

Homoscedasticity in Fitting a Line

- **Assumption**. The residuals ϵ_i have the same variance for all i. This is called ***homoscedasticity***.
- The opposite case is called ***heteroscedasticity***.



The vertical spread of the data increases as x increases.

Roadmap

- **Linear Regression**
  - Scenarios where linear regression will be helpful!
  - How to fit a line with least squares.
  - Residuals and homoscedasticity in fitting a line.
  - *More complex data and more complex model.*
  - How to measure the fit.
- **Time-series Analysis**
  - Varying time-series
  - Objectives of time-series analysis
  - Time-series vs. Regression

Linear Regression for Multivariate

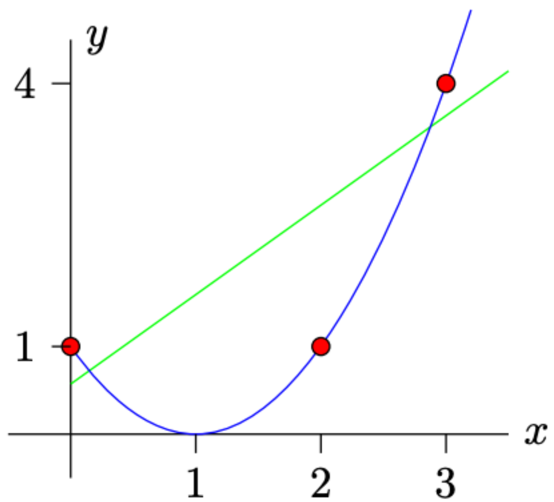
- For multivariate data: $(x_{i,1}, x_{i,2}, \dots, x_{i,m}, y_i)$
- To fit the data with a line (in high dimensional space):
  - $y = \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m + \beta_0$

Response Variable   Explanatory (or predictor) variables

- The total square error is:
  - $\sum_i (\beta_1 x_{i,1} + \beta_2 x_{i,2} + \cdots + \beta_m x_{i,m} + \beta_0 - y_i)^2$

Fitting Polynomials

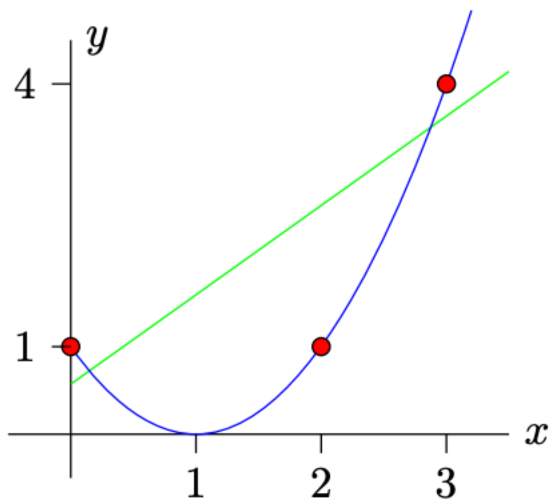
- What is the meaning of "*linear*" in *linear regression*?
  - It refers to the linear algebraic equations for the unknown parameters β_i, i.e. each β_i has exponent 1.
- Use least squares to fit a line to the following three data points: (0,1), (2,1), and (3,4).
  - So, $(x_1, y_1) = (0{,}1)$, $(x_2, y_2) = (2{,}1)$, $(x_3, y_3) = (3{,}4)$.



The parabola has the formula
$y = \beta_0 + \beta_1 x + \beta_2 x^2$

Fitting Polynomials

- The parabola has the formula $y = \beta_0 + \beta_1 x + \beta_2 x^2$
- The square error is
  - $S(\beta_0, \beta_1, \beta_2) = \sum_i \big(\overbrace{\beta_0 + \beta_1 x_i + \beta_2 x_i^2 - y_i}^{\text{The error for } (x_i, y_i)}\big)^2$



Predicted Values *Real Values*

The minimum solutions: $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$ will be used to fit polynomials.

Roadmap

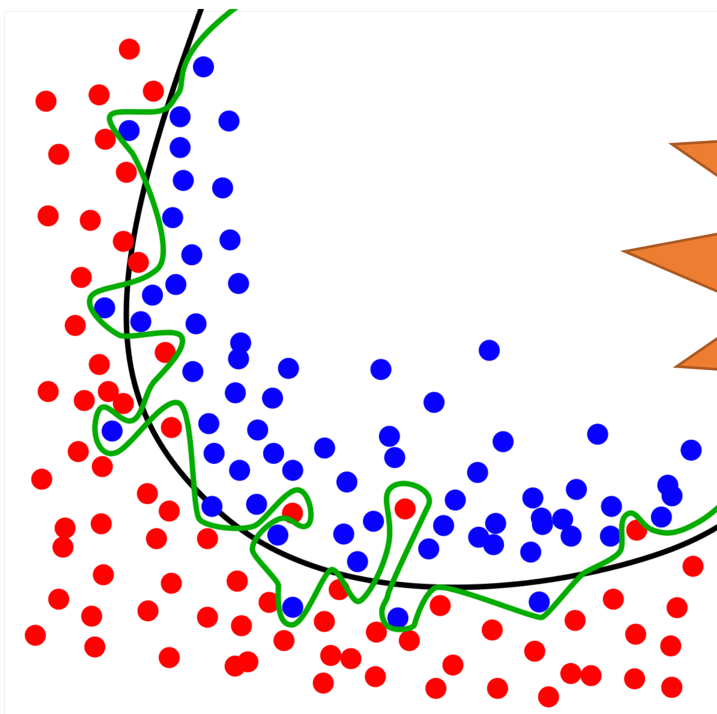
- **<u>Linear Regression</u>**
  - Scenarios where linear regression will be helpful!
  - How to fit a line with least squares.
  - Residuals and homoscedasticity in fitting a line.
  - More complex data and more complex model.
  - *<u>How to measure the fit.</u>*
- **Time-series Analysis**
  - Varying time-series
  - Objectives of time-series analysis
  - Time-series vs. Regression

Measuring the Fit

- Data and predicted values of the response variable:
  - $\boldsymbol{y} = (y_1, y_2, \dots, y_n)$
  - $\hat{\boldsymbol{y}} = (\hat{y}_1, \hat{y}_2, \dots, \hat{y}_n)$
- **Total Sum of Squares (TSS)**: $\sum_i (y_i - \bar{y})^2$
- **Residual Sum of Squares (RSS)**: $\sum_i (y_i - \hat{y}_i)^2$
- **The goodness of fit**: $R^2 = 1 - \frac{RSS}{TSS}$
  - More complex model, better fitness (and smaller R^2)!
  - Tradeoff between goodness of fit and complexity.

Overfitting in Regression

- More complex model, better fitness (and smaller R^2)!
- Tradeoff between goodness of fit and complexity.



Unable to fit new data!

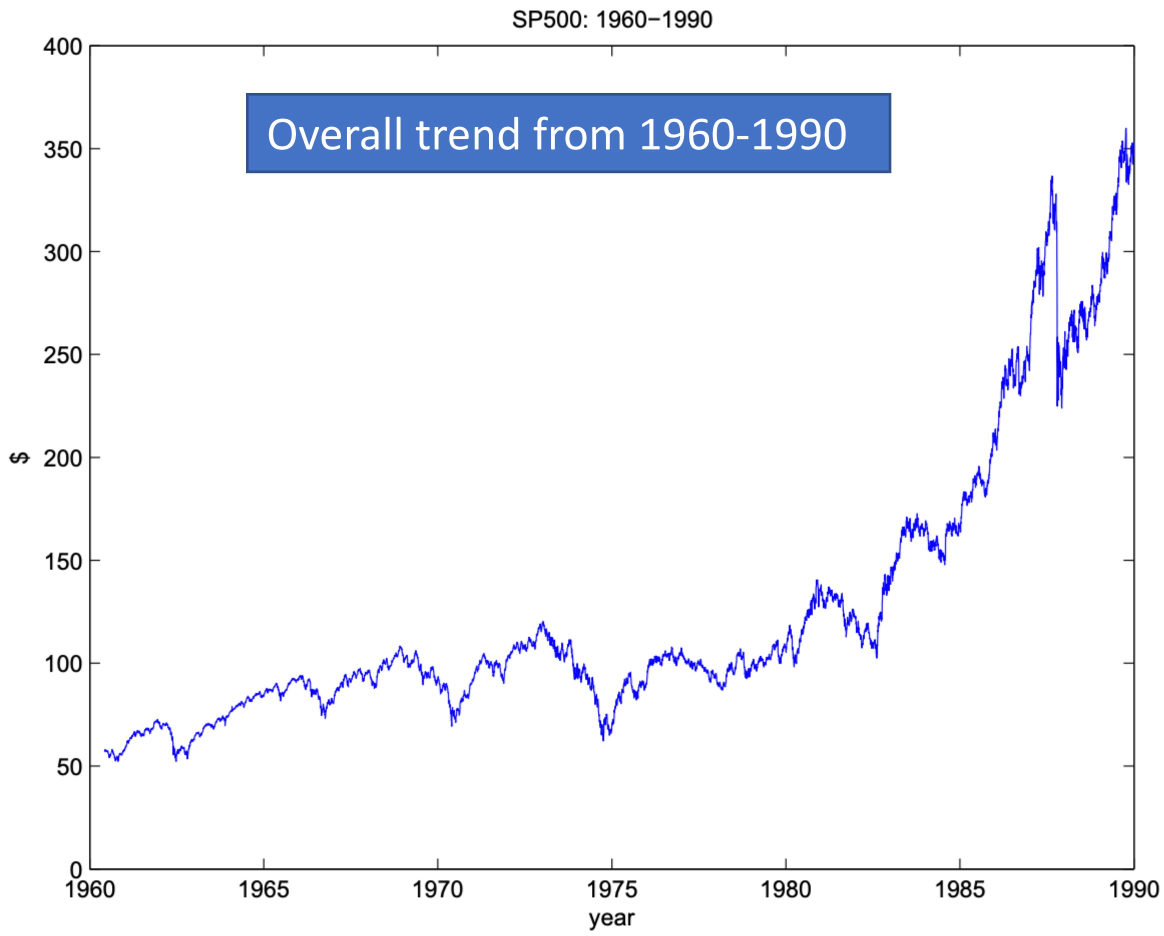
Roadmap

- **Linear Regression**
  - Scenarios where linear regression will be helpful!
  - How to fit a line with least squares.
  - Residuals and homoscedasticity in fitting a line.
  - More complex data and more complex model.
  - How to measure the fit.
- **Time-series Analysis**
  - Varying time-series
  - Objectives of time-series analysis
  - Time-series vs. Regression

Roadmap

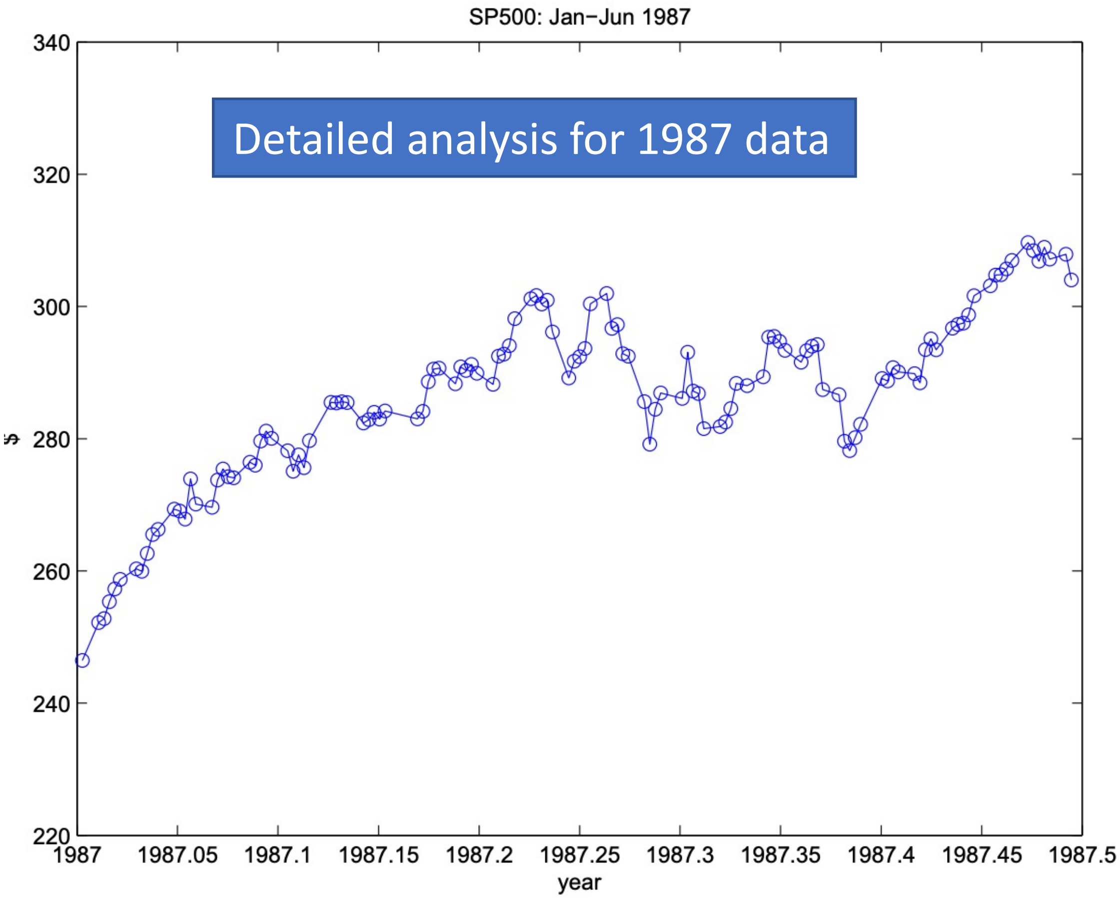
- **Linear Regression**
  - Scenarios where linear regression will be helpful!
  - How to fit a line with least squares.
  - Residuals and homoscedasticity in fitting a line.
  - More complex data and more complex model.
  - How to measure the fit.
- **<u>Time-series Analysis</u>**
  - *<u>Varying time-series</u>*
  - Objectives of time-series analysis
  - Time-series vs. Regression

Varying Time Series



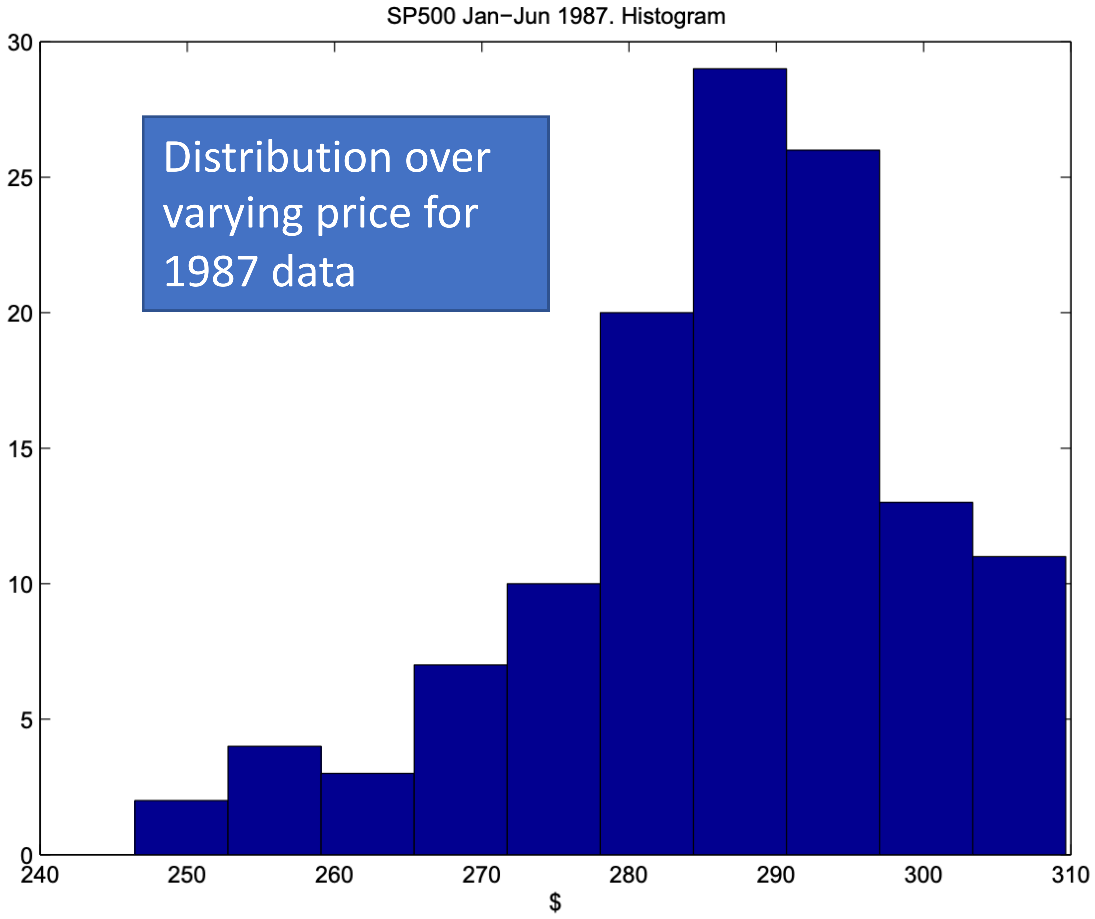
The stock performance of 500 large companies listed on stock exchanges in the United States

Varying Time Series



The stock performance of 500 large companies listed on stock exchanges in the United States

Varying Time Series



The stock performance of 500 large companies listed on stock exchanges in the United States

Roadmap

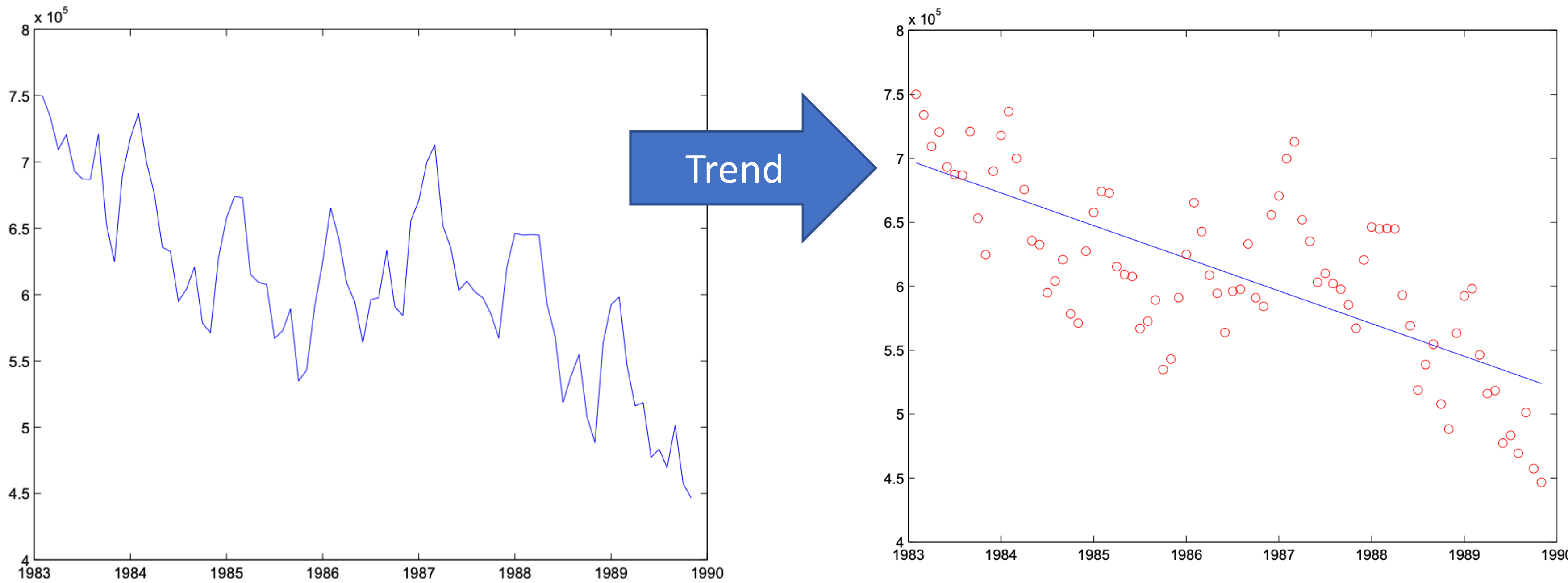
- **Linear Regression**
  - Scenarios where linear regression will be helpful!
  - How to fit a line with least squares.
  - Residuals and homoscedasticity in fitting a line.
  - More complex data and more complex model.
  - How to measure the fit.
- **Time-series Analysis**
  - Varying time-series
  - *Objectives of time-series analysis*
  - Time-series vs. Regression

Why we analyze time-series?

- **Compact description of data**:
  - **Level**: The average value in the series.
  - **Trend**: The increasing or decreasing value in the series.
  - **Seasonality**: The repeating short-term cycle in the series.
  - **Noise**: The random variation in the series.
- **Interpretation**: e.g., *seasonal adjustment.*
- **Forecasting**: e.g., *predict unemployment.*

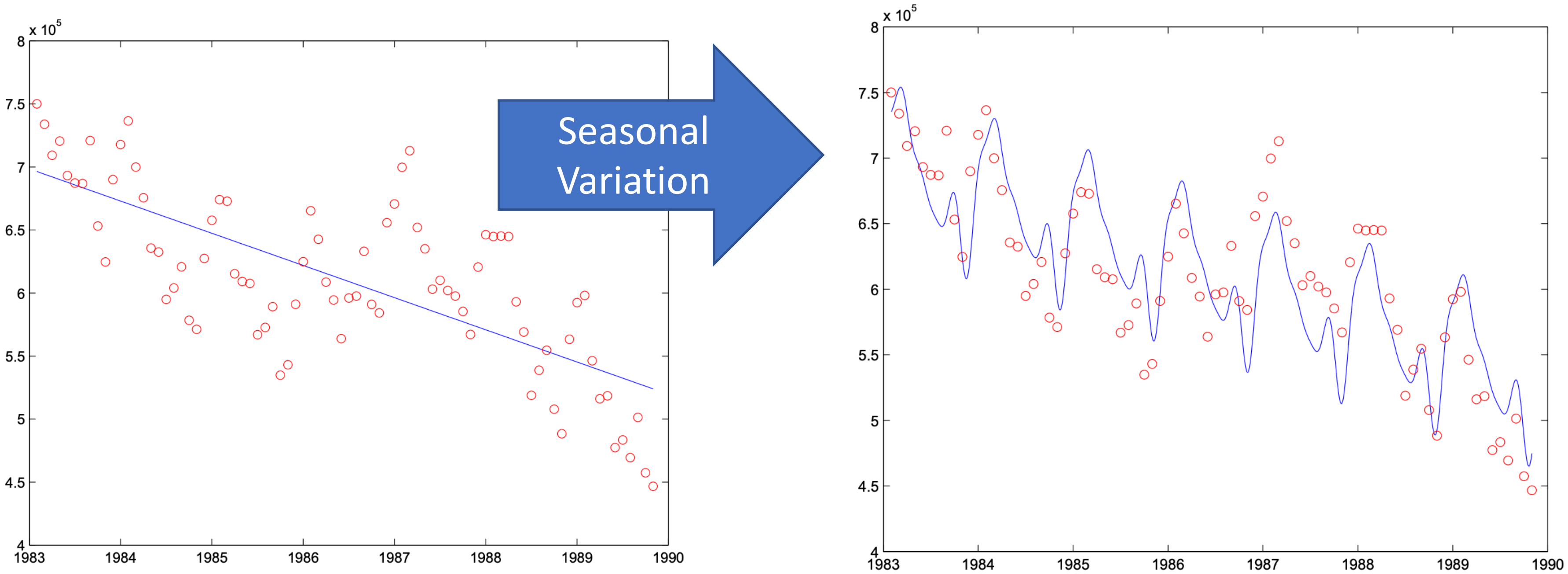
Example: Unemployment Data

- Monthly number of unemployed people over years in Australia. (*Hipel and McLeod, 1994*)



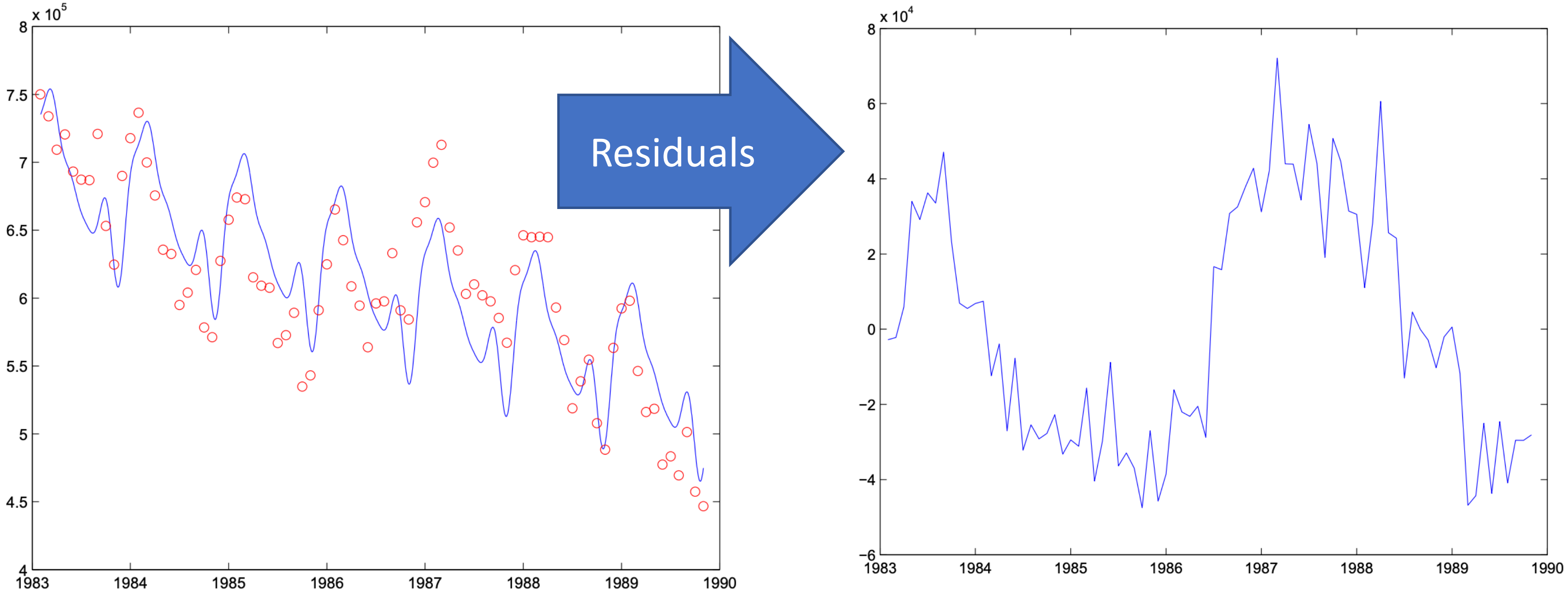
Example: Unemployment Data

- Monthly number of unemployed people over years in Australia. (*Hipel and McLeod, 1994*)



Example: Unemployment Data

- Monthly number of unemployed people over years in Australia. (*Hipel and McLeod, 1994*)



Why we analyze time-series?

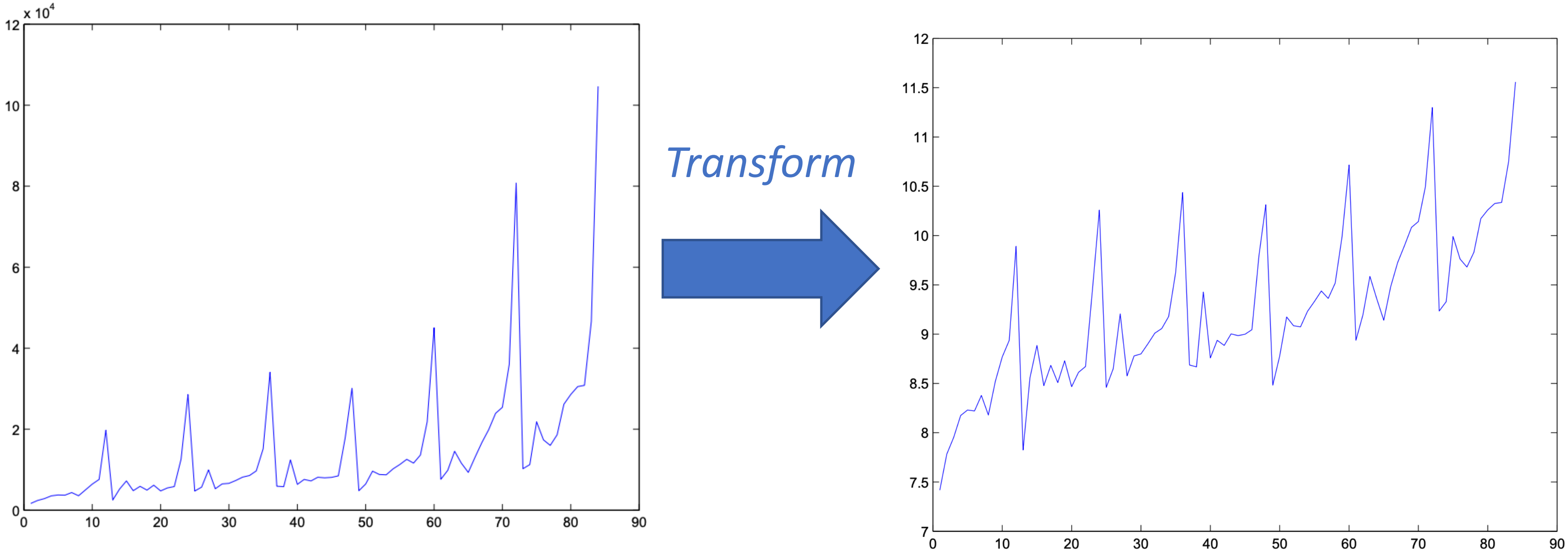
- **Compact description of data**
- **Interpretation**: e.g., *seasonal adjustment.*
- **Forecasting**: e.g., *predict unemployment.*
- **Control**: e.g., analyze impact of monetary policy on unemployment.
- **Hypothesis Testing**: e.g., *global warming*.
- **Simulation**: e.g., *estimate probability of catastrophic events*.

Roadmap

- **Linear Regression**
  - Scenarios where linear regression will be helpful!
  - How to fit a line with least squares.
  - Residuals and homoscedasticity in fitting a line.
  - More complex data and more complex model.
  - How to measure the fit.
- **Time-series Analysis**
  - Varying time-series
  - Objectives of time-series analysis
  - *Time-series vs. Regression*

Time-series Analysis vs. Regression

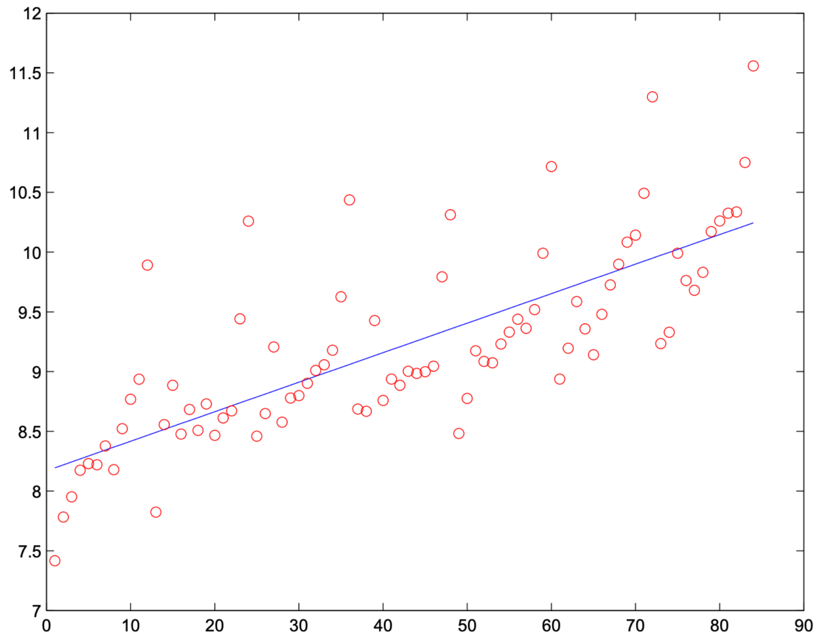
- Monthly sales for a souvenir shop at a beach resort town in Queensland.
  - (*Makridakis, Wheelwright and Hyndman, 1998*)



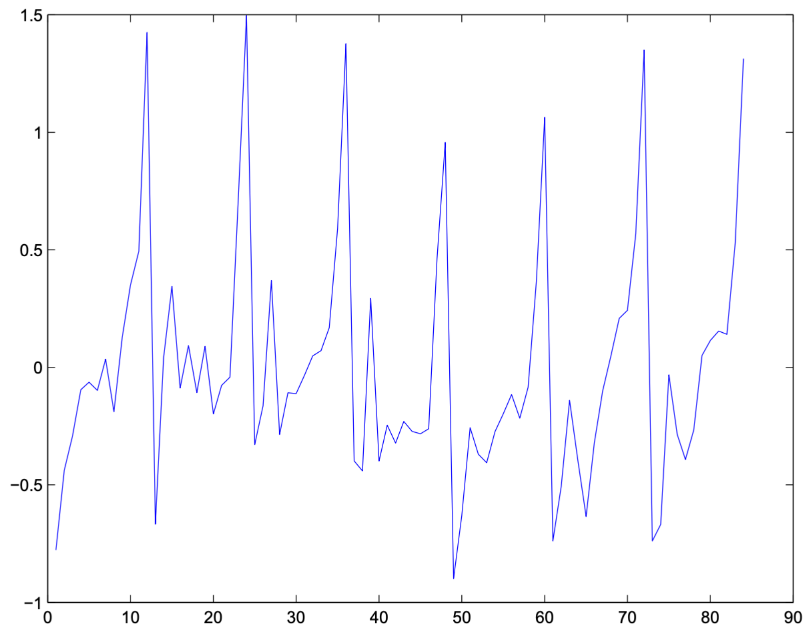
Time-series Analysis vs. Regression

- Monthly sales for a souvenir shop at a beach resort town in Queensland.
  - (*Makridakis, Wheelwright and Hyndman, 1998*)

Trend: fitted line



Residuals



A slide to take away

- What is linear regression?
- How to use linear regression to fit data?
- How to evaluate the regression results?
- What are time-series?
- Why we do time-series analysis?
- What are the useful tools to analyze time series?