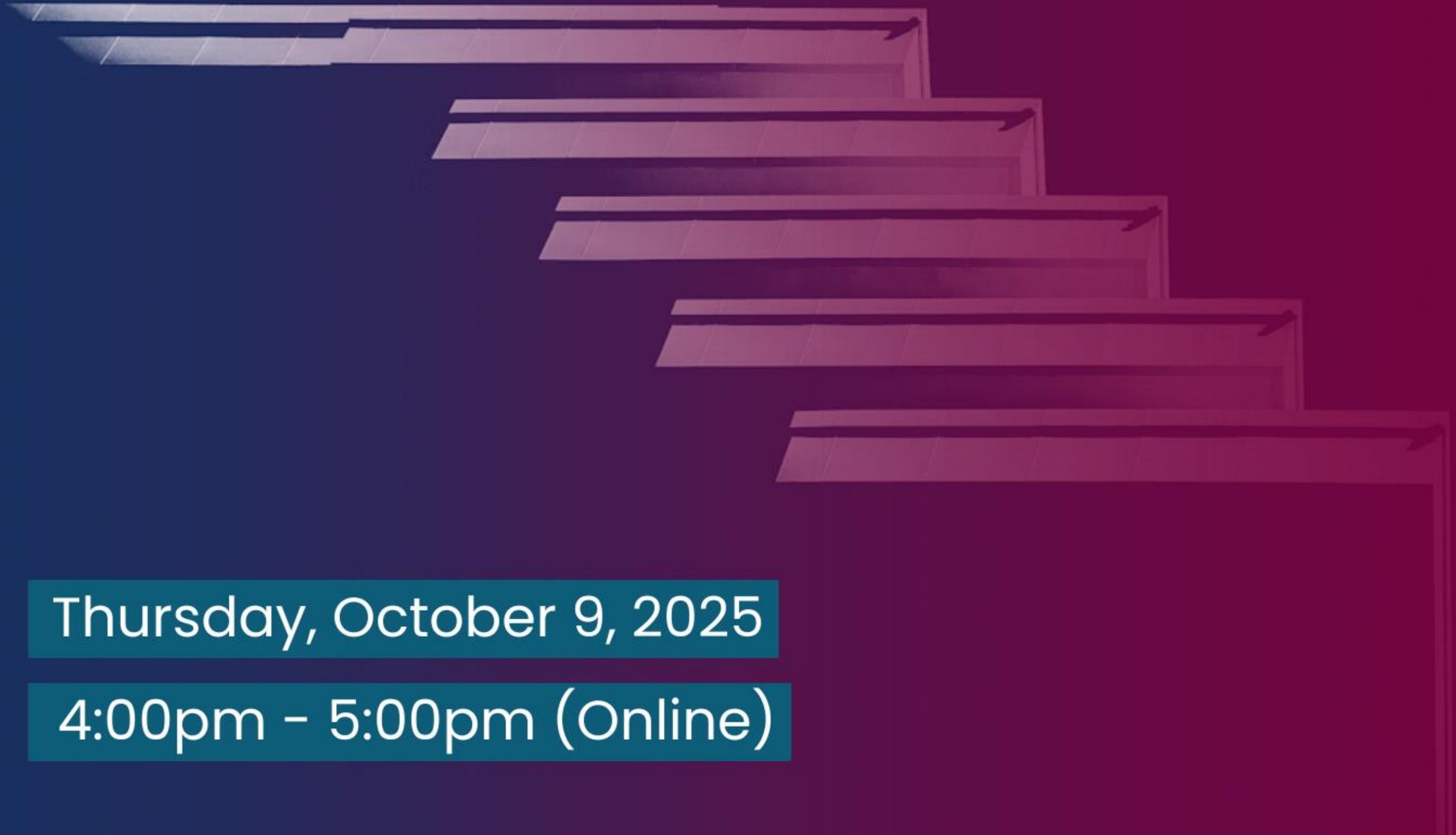
Getting Started with Linear Regression in R



Thursday, October 9, 2025

4:00pm - 5:00pm (Online)

Sherman Centre for Digital Scholarship

Getting Started with Linear Regression in R

Sahar Khademioore
PhD Candidate in Health Research Methodology

Code of Conduct

The Sherman Centre and the McMaster University Library are committed to fostering a supportive and inclusive environment for its presenters and participants.

As a participant in this session, you agree to support and help cultivate an experience that is collaborative, respectful, and inclusive, as well as free of harassment, discrimination, and oppression. We reserve the right to remove participants who exhibit harassing, malicious, or persistently disruptive behaviour.

Please refer to our code of conduct webpage for more information: scds.ca/events/code-of-conduct/

Certificate Program

The Sherman Centre offers a Certificate of Attendance that rewards synchronous participation at 7 workshops. We also offer concentrations in Data Analysis and Visualization, Digital Scholarship, and Research Data Management.

Learn more about the Certificate Program: https://scds.ca/certificate-program

Verify your participation at a session: https://u.mcmaster.ca/verification

At an unspecified point during the workshop, a code will be read aloud. This is the answer to the third question of the form.

DASH: Data Analysis Support Hub Workshops

Register for upcoming DASH events: u.mcmaster.ca/scds-events

October 16, 2025: "Introduction to R Programming"

October 23, 2025: "Visualizing Networks with Gephi"

October 30, 2025: "Introduction to Data Analysis with SPSS"

November 6, 2025: "Creating Interactive Data Visualizations with Power BI"

"

Book an Appointment with the DASH Team

Receive help from a member of the DASH team! DASH can assist with the following topics:

- Creating data visualizations, including charts, graphs, and scatter plots
- Figuring out which statistical tests to run (e.g., t-test, chi-square, etc.).
- Analyzing data with software including SPSS, Python, R, SAS, ArcGIS, MATLAB, and Excel
- Choosing which software package to use, including free and open-source software
- Troubleshooting problems related to file formats, data retrieval, and download
- Selecting methodology and type of data analysis to use in a thesis project

Book an appointment: https://library.mcmaster.ca/services/dash

Linear Regression using R

Objectives of the workshop:

- Review basics of linear regression
- Assess model assumptions
- How to fit a linear regression
- How to interpret our model’s output
- Evaluate the model’s fit

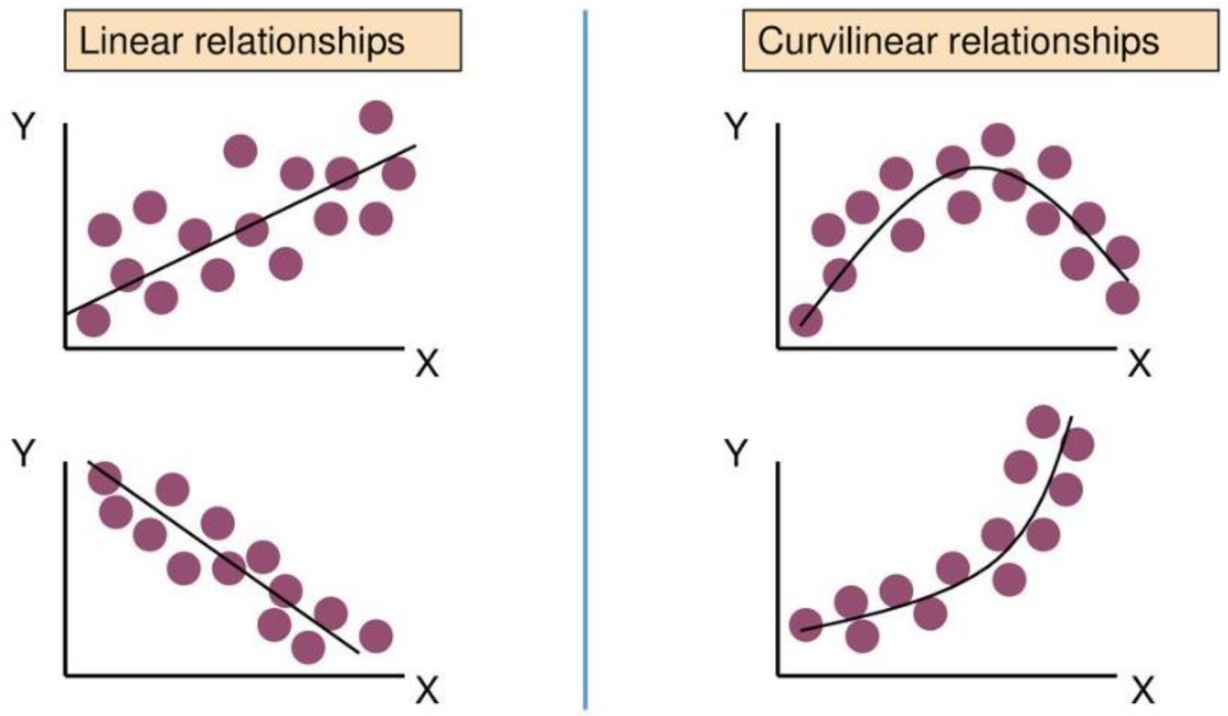
Introduction to Linear Regression

- Fundamental statistical method for modeling relationships
- Assumes linear relationship between variables
- Applications in prediction, inference, and understanding relationships

When to Use Linear Regression

Ideal Use Cases

- Predicting **continuous outcomes** (e.g., weight, BMI)
- When you suspect a linear relationship between variables



Statistics for Managers Using Microsoft Excel, 9th edition

Simple vs. Multiple Linear Regression

Simple Linear Regression

One predictor variable

Example: age and blood pressure

Multiple Linear Regression

Two or more predictor variables

Example: age, weight, smoking status and blood pressure

Purposes of Regression

Estimate association of X and Y

How big or important is the effect X on Y?

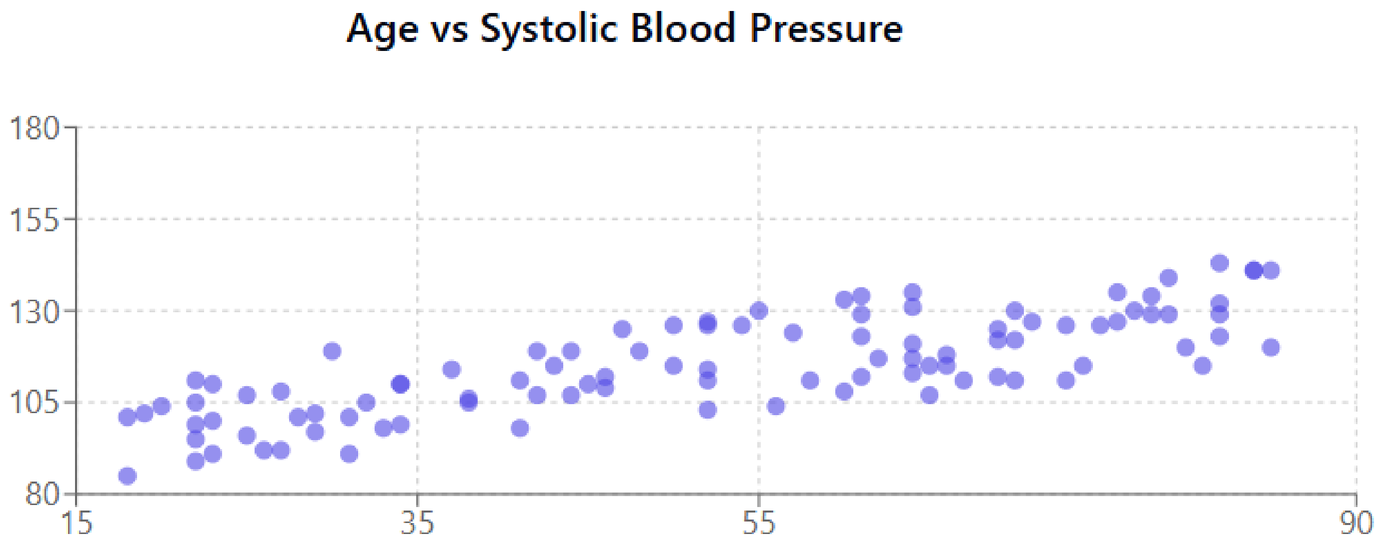
Estimate the relationship between X and Y, controlling for confounding variables

Predict Y from one or more X's

Determining what the "best model" for predicting Y from X's

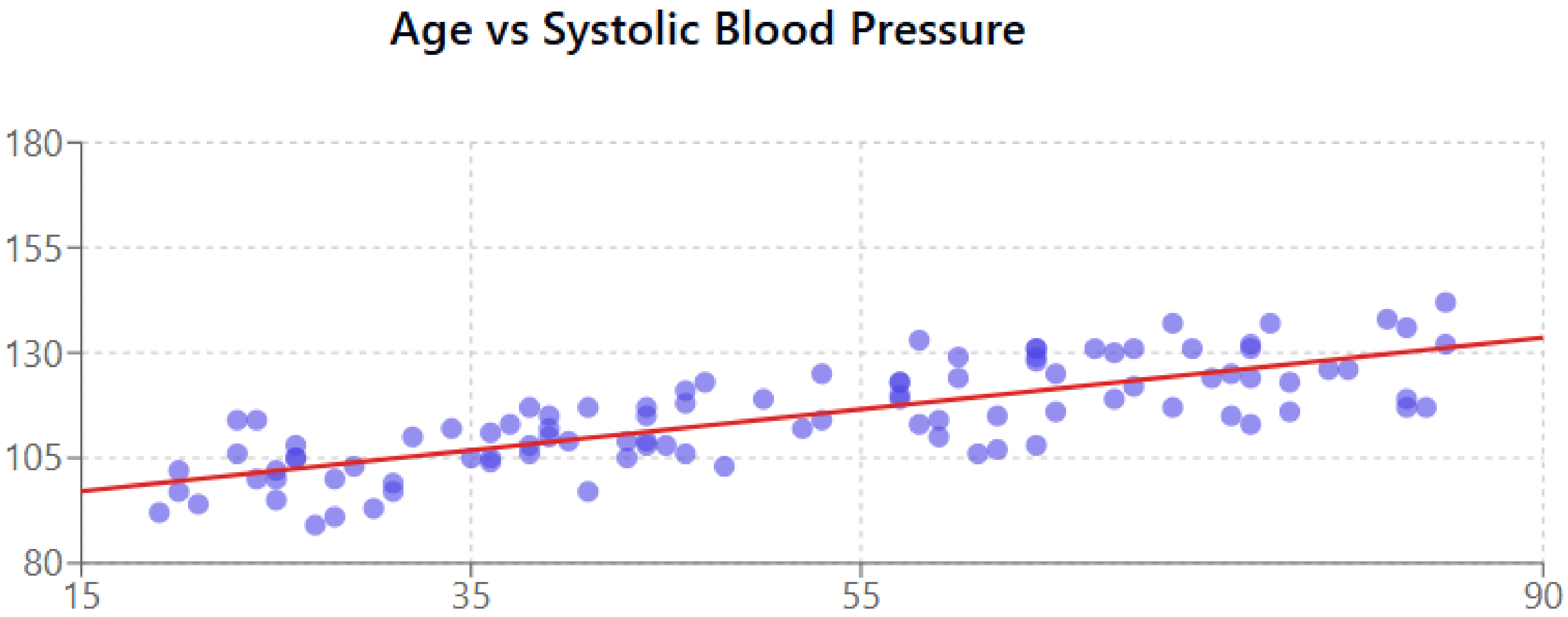
How do we assess if two continuous variables are associated?

- **Visual Understanding**
- Scatter-plot
- a plot of paired *X* and *Y* values



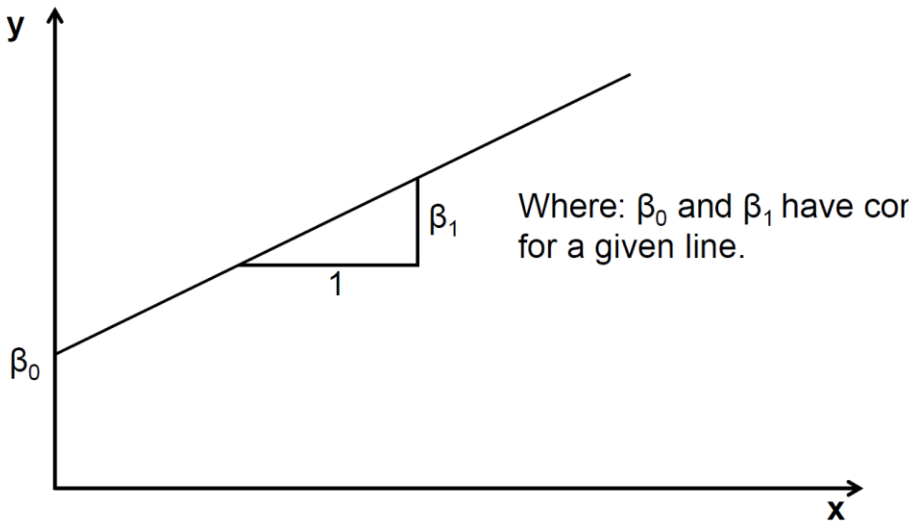
Best-fit straight line

Plot a line that fits the data point the best.



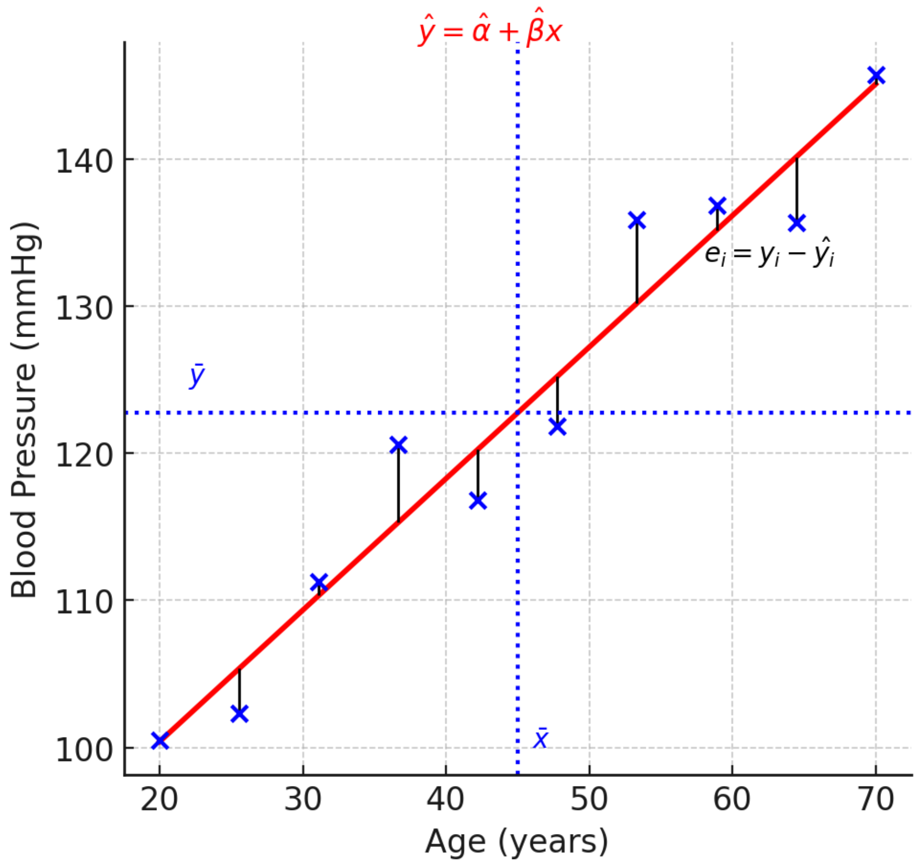
Linear regression line

- Equation: Y_i = $\boldsymbol{\beta_0 + \beta_1 X_i + e_i}$
- **Y**: Dependent variable (what we want to predict)
- **X**: Independent variable
- **β0**: Population intercept (predicted value of y when x = 0)
- **β1**: Population Slope (how much y changes for each unit change in x)
- Where: β0 and β1 have constant values for a given line.
- **e:** residual error



Linear regression line

- Equation: Y_i = **$\beta 0 + \beta 1 X_i + e_i$**
- **Y**: Dependent variable (what we want to predict)
- **X**: Independent variable
- **β1**: Population Slope (how much y changes for each unit change in x)
- **β0**: Population intercept (predicted value of y when x = 0)
- Where: β0 and β1 have constant values for a given line.
- **e:** residual error

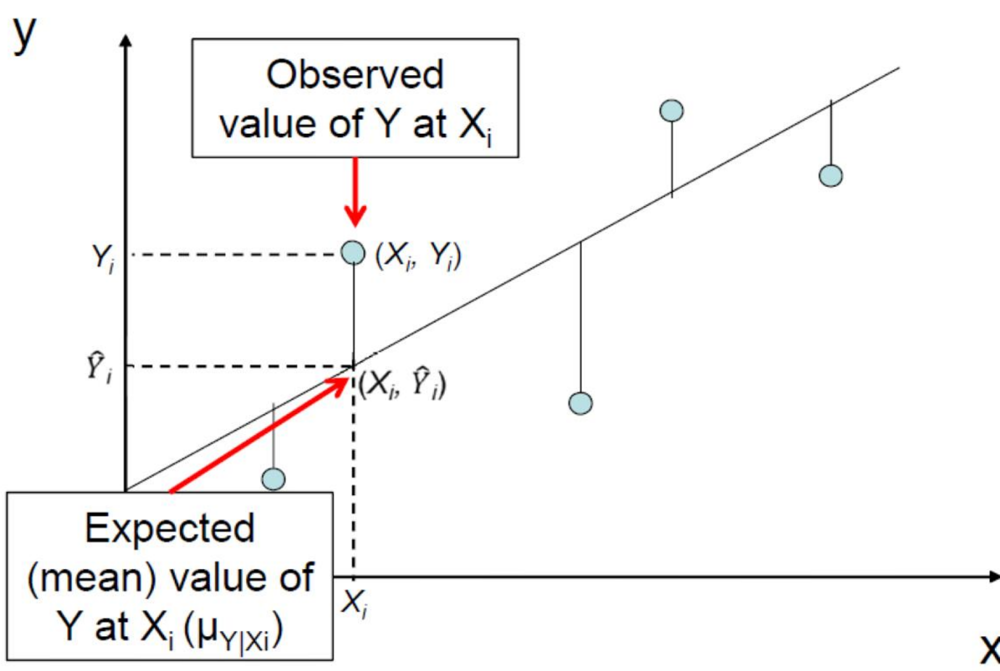


Estimating the regression line

Least-squares method:

- Determines best fitting straight line as a line that minimizes the sum of squares of the lengths of the vertical line segments from the observed data points in the scatter plot to the fitted line
- We get unbiased estimates of the slope and intercept if we minimize the sum of the squares of the vertical distances of the data points from the line.

$$SS_e = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$



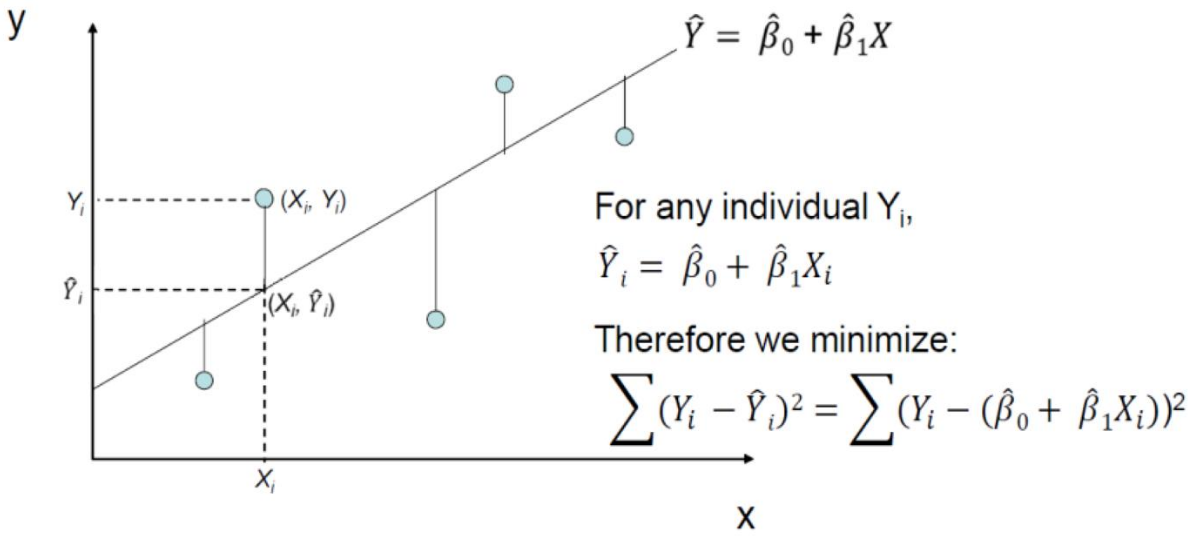
Dr. Shofiqul Islam, McMaster University

Finding best value for $\widehat{\beta_1}$ and $\widehat{\beta_0}$ to find the best-fit line

- A line that has the least error : the error between predicted values and actual values should be minimum.

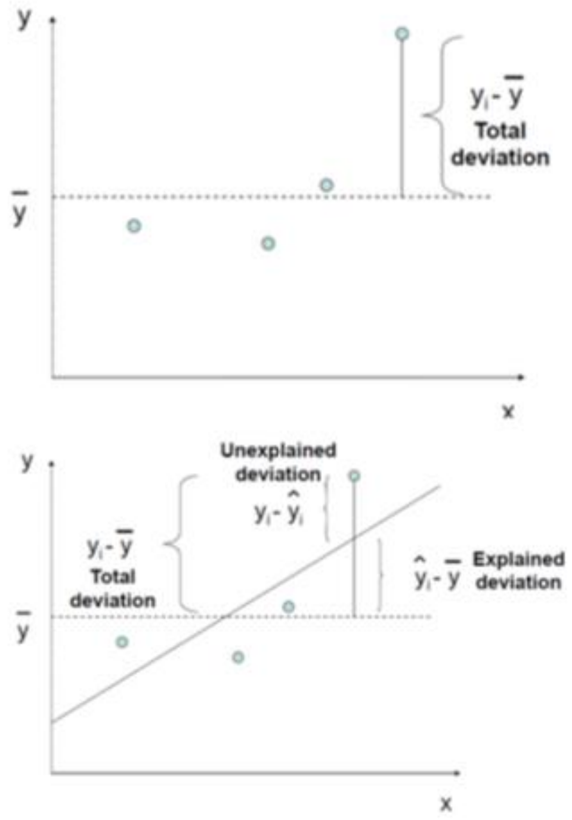
- The following formulas can be used to estimate β_1 and β_0 based on the least squares solution:

- $\widehat{\beta_1} = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2}$
- $\widehat{\beta_0} = \bar{Y} - \widehat{\beta_1}\bar{X}$



Dr. Shofiqul Islam, McMaster University

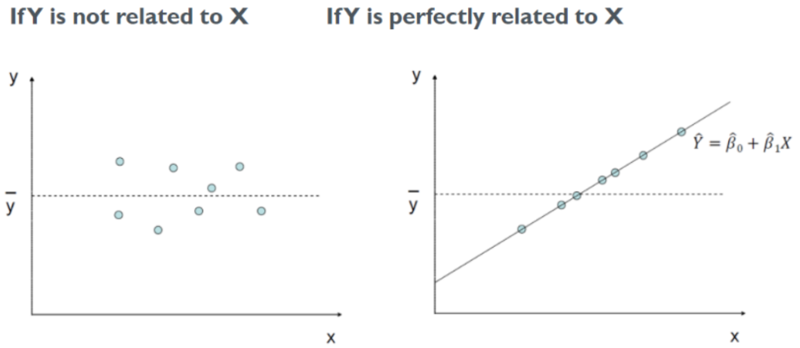
Variance decomposition (Goodness of fit)



$$Var\,Y = \frac{\sum(y_i - \bar{y})^2}{n-1}$$

$$\sum\left(y_i - \bar{y}\right)^2 = \sum\left(y_i - \hat{y}_i\right)^2 + \sum\left(\hat{y}_i - \bar{y}\right)^2$$

Total SS   Residual SS   Regression SS



$SS_{Total} \cong SS_{Error}$ $SS_{Total} \cong SS_{Reg}$

Dr. Shofiqul Islam, McMaster University

ANOVA Table and F-test

of predictors (x variables)=1

	Sum of Squares SS	Degrees of Freedom df	Mean Square MS	F Value F
Regression	SS_{reg}	p	SS_{reg}/p	MS_{reg}/MS_{err}
Residual	SS_{err}	n-p-1	SS_{err} /(n-p-1)	
Total	SS_{tot}	n-1		

# of observations - # of parameters estimated

# of observation - 1

Where:

$SS_{reg} = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$

$SS_{err} = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$

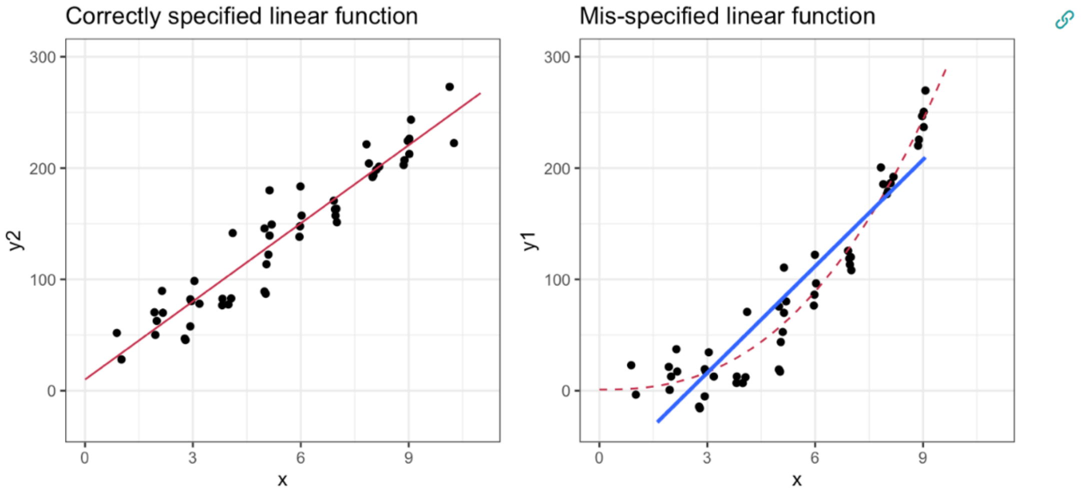
$SS_{tot} = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$

Key Assumptions of Linear Regression

- Linearity: Relationship is linear in parameters
- Independence: Observations are independent
- Homoscedasticity: Constant variance of residuals
- Normality: Residual errors are normally distributed
- No perfect multicollinearity (for multiple regression)

Linear regression assumption

- **Linearity**
- Relationship between variables should be linear
- Can be checked through scatter plots
- Violations require data transformation



https://zief0002.github.io/modeling/03-03-model-assumptions.html#fig-nonlinear

Linear regression assumption

Independence (Y values are statistically independent of one another)

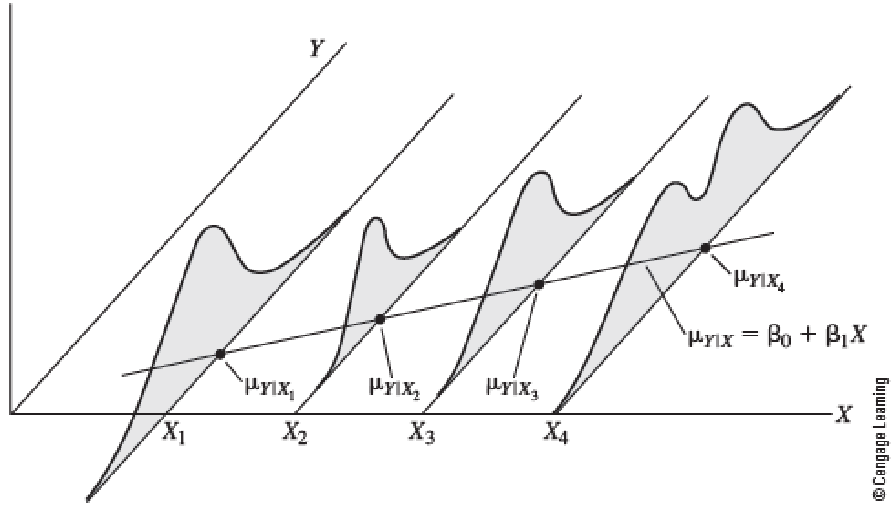
- Observations should be independent

Important for time series data (multiple observations are made on the same individual at different times)

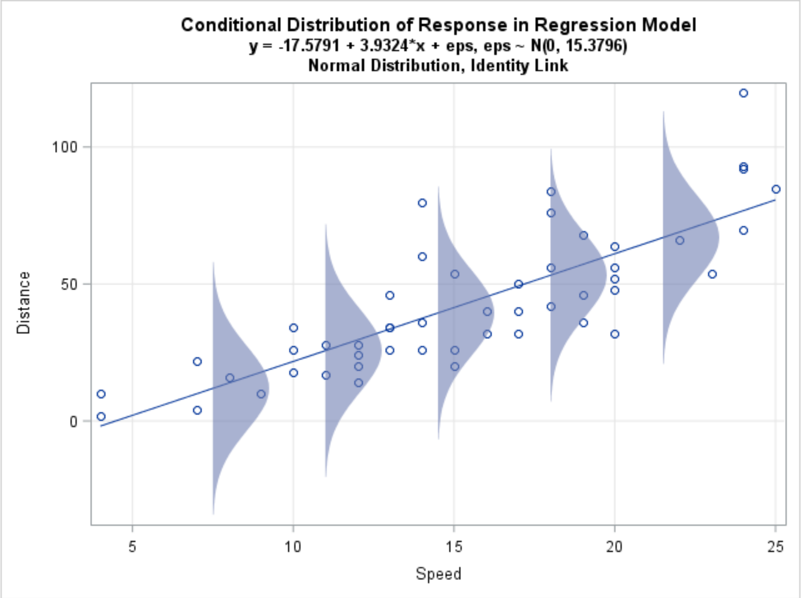
Example: blood pressure is measured on an individual over the time

Linear regression assumption

- **Homoscedasticity** (Homo- means “same,” and -scedastic means “scattered.”)
- Constant variance of residuals (The variance of Y is the same for any X.)
- Errors should be uniformly distributed
- Check using residual plots



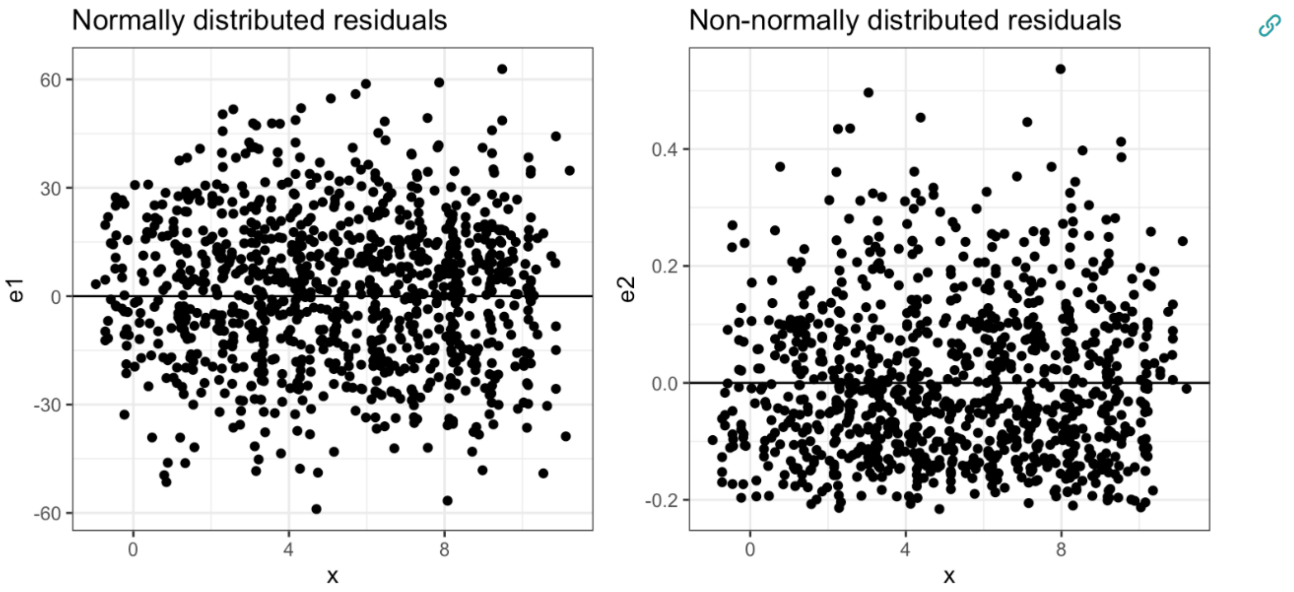
Kleinbaum, Kupper, Muller 1988



https://blogs.sas.com/content/iml/2015/09/10/plot-distrib-reg-model.html

Linear regression assumption

- **Normality**
- Residuals should be normally distributed with mean= o
- Can be checked using Q-Q plots
- Violations may require data transformation



https://zief0002.github.io/modeling/03-03-model-assumptions.html#fig-nonlinear

Residual plots

Key Residual Plots:

- *Residuals vs. Fitted Values*
- *Q-Q Plot for Normality*
- *Scale-Location Plot*
- *Residuals vs. Leverage*
- *Cook's Distance Plot*

Diagnostic Tests

Formal Tests:

- Durbin-Watson test (autocorrelation)
- Breusch-Pagan test (homoscedasticity)
- Shapiro-Wilk test (normality)
- VIF test (multicollinearity)

Key Metrics for Model Fitness

Quantitative Measures:

- R-squared (Coefficient of Determination)
- Adjusted R-squared
- F-statistic and p-value
- Root Mean Square Error (RMSE)
- Mean Absolute Error (MAE)

R-squared (R^2)

Understanding R-squared:

- Measures proportion of variance explained by the model
- Ranges from 0 to 1 (0% to 100%)
- Higher values indicate better fit
- Limitations: Can be misleadingly high with many predictors
- Formula: $R^2 = SSres / SStot$

Adjusted R-squared

Why Adjusted R-squared:

- Penalizes addition of unhelpful predictors
- Always lower than R-squared
- Better for comparing models with different numbers of predictors
- Formula: adj-$R^2 = 1 - [(1-R^2)(n-1)/(n-p-1)]$ where n = sample size, p = number of predictors

F-statistic and p-value

The F-statistic and its corresponding p-value test the overall significance of our model.

The F-statistic compares the fit of our model to a model with no predictors, and the p-value tells us if this difference is statistically significant.

Model Improvement Steps

When Fit is Poor:

- Transform variables (log, square root, etc.)
- Handle outliers appropriately
- Add interaction terms
- Consider polynomial terms

Hypothesis Testing

Statistical Significance:

- F-test for overall model significance
- t-tests for individual coefficients
- p-values interpretation
- Confidence intervals
- ANOVA table analysis

Let’s practice!

Use the link below:

https://colab.research.google.com/drive/1wD-c6NvqvFhUmKc4jiExA2r1u3fzpsZU?usp=sharing

Thank you!

- **Email:** Khades1@mcmaster.ca
- **Book an appointment** with DASH: https://library.mcmaster.ca/services/dash
- **Contact DASH:** Data Analysis Support Hub: libdash@mcmaster.ca