Linear Algebra

General Linear Models and Least Squares

Automotive Intelligence Lab.

HANYANG UNIVERSITY

AI LAB Automotive Intelligence

Contents

General Linear Models

General Linear Model

given

$Ax = b$ data

data unknown

- **Statistical model (data-driven model)**
  - A set of equations that relates predictors to observations.
    - Predictors: independent variable.
    - Observations: dependent variable.

- **Example of the model in stock market price**
  - Independent variable: time
  - Dependent variable: stock market price

- **We will focus on General Linear Model, which is called as GLM.**
  - Regression is a type of GLM, for example.

Terminology of GLM

Difference terminology between fields of statistics and linear algebras

LinAlg	Stats	Description
$Ax = b$	$X\beta = y$	General Linear Model(GLM)
A	X	Design matrix (columns=independent variables, predictors, regressors)
x	β	Regression coefficients or beta parameters
b	y	Dependent variable, outcome measure, data

Table of terms in GLMs

$n \times N$ N n

unknown

$X =$ ↓ n data

$X\beta = Y$

$x \to \beta \to y$ 를 예측.

Setting up a GLM

Process to set up GLM

1. Define an equation that relates the predictor variables to the dependent variable.
2. Map the observed data onto the equations.
3. Transform the series of equations into a matrix equation.
4. Solve that equation.

Simple Example to Explain Process of GLM

Model: Predicts adult height based on weight and parent's height

?

$$y = \beta_0 + \beta_1 w + \beta_2 h + \epsilon$$

Equation of example model

내키 몸무게 부모님키

- $\boldsymbol{y}$: height of an individual
- $\boldsymbol{w}$: weight
- $\boldsymbol{h}$: parents' height (average of mother and father)
- $\boldsymbol{\epsilon}$: error term (also called residual)

Why we need error term ϵ (residual)?

- Weight and parents' height cannot perfectly determine an individual's height.
- Variance not attributable to weight and parents' height will be absorbed by residual.
  - Such as growing environment, sleeping time and so on.

More Explanation about Previous Simple GLM

- **What is $\boldsymbol{\beta}$?**
  - Coefficients or weights.
  - Describe how to combine weight and parent's height to predict an individual's height.
  - β_0?
    - Called an intercept or a constant.
    - Without this term, best-fit line always pass the origin.
      - It will be explained at the end of chapter.

$$\boldsymbol{y} = \boldsymbol{\beta_0} + \boldsymbol{\beta_1} \boldsymbol{w} + \boldsymbol{\beta_2} \boldsymbol{h} + \boldsymbol{\epsilon}$$

Previous GLM model

- **After defining equations, map the observed data onto the equations.**
  - Use the simple data table below.
  - For simplicity, omit ϵ.

$\boldsymbol{y}$	$\boldsymbol{w}$	$\boldsymbol{h}$
175	70	177
181	86	190
159	63	180
165	62	172

Simple data table

$$\begin{aligned} 175 &= \beta_0 + 70\beta_1 + 177\beta_2 \\ 181 &= \beta_0 + 86\beta_1 + 190\beta_2 \\ 159 &= \beta_0 + 63\beta_1 + 180\beta_2 \\ 165 &= \beta_0 + 62\beta_1 + 172\beta_2 \end{aligned}$$

$$\begin{bmatrix} 1 & 70 & 177 \\ 1 & 86 & 190 \\ 1 & 63 & 180 \\ 1 & 62 & 172 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} = \begin{bmatrix} 175 \\ 181 \\ 159 \\ 165 \end{bmatrix}$$

Transforming series of equations into a matrix equation

- **Of course, we can express this equation briefly as $\boldsymbol{X\beta} = \boldsymbol{y}$.**

Solving GLMs

Idea to Solve for the Vector of Unknown Coefficients β

Method 1

- **Simply** **left-multiply** **both sides of the equation by** **left-inverse of X.**

$$\overset{m\times n}{\boldsymbol{X}}\overset{n\times 1}{\boldsymbol{\beta}} = \overset{m\times 1}{\boldsymbol{y}}$$

$$(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$$

$$\boldsymbol{\beta} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$$

Solution to solve $\boldsymbol{\beta}$

- **Memorize** $\boldsymbol{\beta} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$
  - Also called **least squares solution**.
  - One of the most important mathematical equations in applied linear algebra.

Code Exercise of Left-Multiply to Solve Least Square

Code Exercise (11_01)

- Simply left-multiply both sides of the equation.
- Variable X: design matrix
- Variable y: data vector

```
% Clear workspace, command window, and close all figures
clc; clear; close all;

% Define a matrix X and y
X = [7; 5; 6];
y = [4; 7; 8];

% Compute the left-inverse of X
X_leftinv = ;

% Calculate beta
beta = ;
disp("beta");
disp(beta);
```

MATLAB code to solve the least square using left-multiply

Is the Solution Exact?

Method 2

$$\begin{bmatrix} | & | \end{bmatrix} \beta = \begin{bmatrix} | \end{bmatrix}$$

- **When is equation $X\beta = y$ exactly solvable?**
  - In case of y is in the column space of design matrix X.
  - Then, question would be
    - Whether data vector is guaranteed to be in the $\boxed{\text{Column Space}}$ of design matrix.
  - Answer is $\boxed{\text{"No"}}$
    - There is no such guarantee.
    - Data vector y is almost never in the column space of X.

Why is Data Vector Not Guaranteed?

- **Imagine a survey of university students.**
  - Researchers are trying to predict average GPA based on drinking behavior.
  - Survey may contain data from 2000 students.
  - But questions are only 3.
    - How much alcohol do you consume?
    - How often do you black out?
    - What is your GPA?

- **Data will be contained in a 2000×3 table.**
  - 3 questions for 2000 students.
  - Column space of the design matrix
    - $2D$ subspace inside that $2000D$ ambient dimensionality.
    - Question of “How much alcohol do you consume?” and “How often do you black out?”
  - Data vector
    - $1D$ subspace inside that same ambient dimensionality.
    - Question of “What is your GPA?”

Meaning of Data in the Column Space of Design Matrix

- **“Data vector in the column space” means that matrix model accounts for 100% of the variance of data.**
  - This almost never happens.
  - Real world data contains **noise** and **sampling variability**.
  - Models are simplifications that don’t account for all of variability.
    - GPA is determined by myriad factors that our model ignores.

Solution to this Conundrum

- **Modify GLM equation to allow for a discrepancy between model predicted data and observed data.**
  - It can be expressed in several equivalent ways as below.

$$X\beta = y + \epsilon$$

$$X\beta - \epsilon = y$$

$$\epsilon = X\beta - y$$

three equivalent expressions

  - Interpretation of the first equation.
    - ϵ is residual, or an error term.
    - Added to the data vector.
    - So that it fits inside the column space of the design matrix.
  - Interpretation of the second equation.
    - Residual term is an adjustment to the design matrix.
    - So that it fits the data perfectly.
  - Interpretation of the third equation.
    - Residual is defined as the difference between model-predicted data and observed data.

Point of This Section

- **Observed data is almost never inside the subspace spanned by regressors.**
  - Reason why we can easily see GLM expressed as $\boldsymbol{X\beta} = \hat{\boldsymbol{y}}$, not $\boldsymbol{X\beta} = \boldsymbol{y}$.
    - $\hat{\boldsymbol{y}} = \boldsymbol{y} + \boldsymbol{\epsilon}$

- **Goal of the GLM**
  - To find linear combination of the regressors.
  - Close as possible to the observed data.

Geometric Perspective on Least Squares

- **Consider column space of design matrix $\boldsymbol{C}(\boldsymbol{X})$ is a subspace of $\mathbb{R}^{\boldsymbol{M}}$.**
  - It's typically a very low-dimensional subspace.
    - It means $N \ll M$.
    - Statistical models tend to have much more observations (M, rows) than predictors (N, columns).
  - Dependent variable is vector $y \in \mathbb{R}^M$.
  - Questions:
    - Is vector $\boldsymbol{y}$ in the column space of the design matrix? $\boxed{\text{No}}$
    - If not, what coordinate inside the column space of the design matrix is as close as possible to data vector?

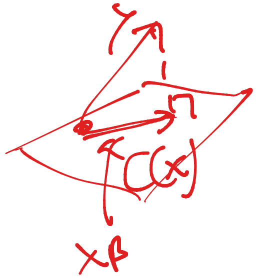
Abstracted Geometric View of GLM

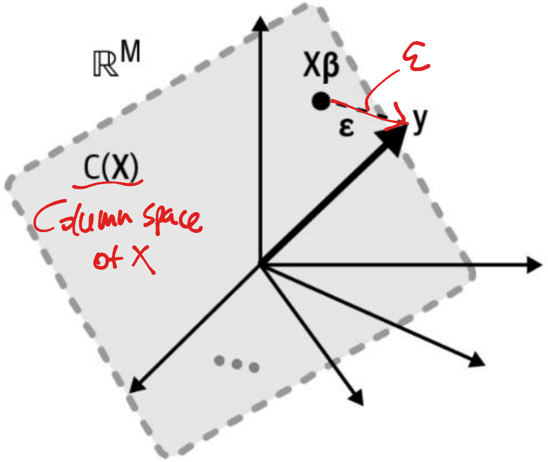
- **Our goal: find set of coefficients $\boldsymbol{\beta}$**
  - Weighted combination of columns in $\boldsymbol{X}$ minimizes distance to data vector $\boldsymbol{y}$.
  - We can call projection vector $\boldsymbol{\epsilon}$.
  - How can find vector $\boldsymbol{\epsilon}$ and coefficients $\boldsymbol{\beta}$?
    - Use orthogonal vector projection!
  - Key insight
    - Shortest distance between y and X is given by the projection vector $\boldsymbol{y} - \boldsymbol{X\beta}$ that meets $\boldsymbol{X}$ at a right angle as shown in below equation.

- **We have rederived the same left-inverse solution which we got from the algebraic approach.**

$$
\begin{aligned}
\boldsymbol{X}^T\boldsymbol{\epsilon} &= \mathbf{0} \\
\boldsymbol{X}^T\boxed{(\boldsymbol{y} - \boldsymbol{X\beta})} &= \mathbf{0} \\
\boldsymbol{X}^T\boldsymbol{y} - \boldsymbol{X}^T\boldsymbol{X\beta} &= \mathbf{0} \\
\boldsymbol{X}^T\boldsymbol{X\beta} &= \boldsymbol{X}^T\boldsymbol{y} \\
\boldsymbol{\beta} &= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}
\end{aligned}
$$

Equation expansion





Abstracted geometric view of GLM

Meaning of Least Squares

Method 3

Why is it called "least squares"?

- Squares
  - Squared errors between predicted data and observed data
  - There is an error term for each i^{th} predicted data point.
    - Defined as $\epsilon_i = X_i\beta - y_i$.
  - Each data point is predicted using same set of coefficients.
    - Same weights for combining predictors in design matrix.
  - So, we can capture all errors in one vector: $\epsilon = X\beta - y$.

- If model is a good fit to the data,
  - Errors ϵ should be small.

- Objective of model fitting
  - Choose elements in β that minimize elements in ϵ.

Expression of Least Squares

- **Why is it called "least squares" ?**
  - If just minimizing errors,
    - It cause the model to predict values toward negative infinity.
  - Instead, minimizing squared errors
    - Corresponding to their geometric squared distance to observed data y.
    - Regardless of whether prediction error itself is positive or negative.
  - Same as minimizing the squared norm of the errors.
    - Hence named "least square".
    - Leads to the following modification:

$$||\boldsymbol{e}||^2 = ||\boldsymbol{X\beta} - \boldsymbol{y}||^2$$

Expression of least squares

View Least Squares as Optimization Problem

- **Find set of coefficients β that minimizes squared errors.**
  - Minimization can be expressed as follows:

$$\min_{\beta} \|X\beta - y\|^2$$

Minimization

  - Solution to this optimization
    - Can be found by setting derivative of objective to zero.
    - Applying a bit of differential calculus and a bit of algebra.

$$0 = \frac{d}{d\beta}\|X\beta - y\|^2 = 2X^T(X\beta - y)$$

$$0 = X^T X\beta - X^T y$$

$$X^T X\beta = X^T y$$

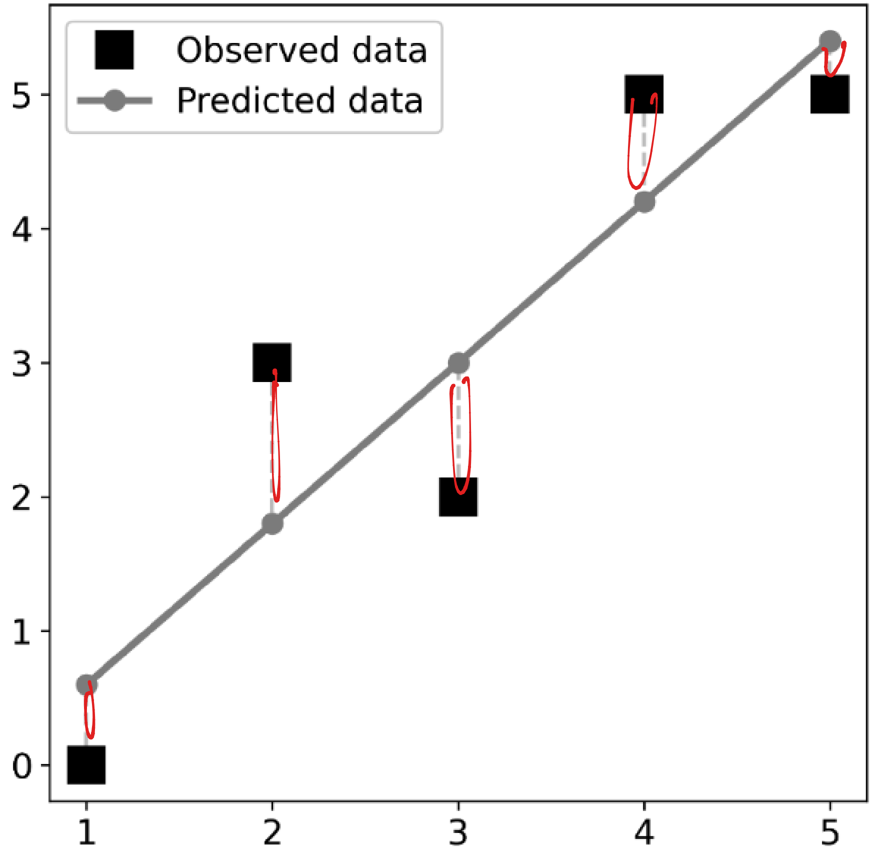
$$\beta = (X^T X)^{-1} X^T y$$

Solution of optimization

  - Rediscover same solution that reached simply by using our linear algebra intuition!
    - Although started from a different perspective to minimize the squared distance between the model-predicted values and the observed values.

Visualization Intuition for Least Squares

- **Black squares**
  - Observed data
- **Gray dots**
  - Model predicted values
- **Gray dashed lines**
  - Distances between observed data and model predicted values
- **All model predicted values lie on a line.**
- **Goal of least squares**
  - Find slope and intercept!
    - Minimize distance from predicted to observed data.



Visualization of results least squares

All Roads Lead to Least Squares

- **You’ve now seen three ways.**
  - To derive least squares solution.
- **Remarkably, all approaches lead to same conclusion.**
  - Left-multiply both sides of GLM equation by left-inverse of design matrix $\boldsymbol{X}$.
- **Different approaches have unique theoretical perspectives.**
  - Provide insight into nature and optimality of least squares.
- **But it is a beautiful thing.**
  - No matter how you begin your adventure into linear model fitting.
  - Because you end up at same conclusion.

GLM in a Simple Example

GLM in a Simple Example

Example

- Report the number of online courses they took and their general satisfaction with life.
- Experiment which is surveyed a random set of 20 students.

Table 1. shows first 4 (out of 20) rows of data matrix.

Data is easier to visualize in scatterplot as Fig 1..

- Notice that independent variable is plotted on the x-axis.
  - While dependent variable is plotted on the y-axis.
  - That is common convention in statistics.

Number of courses	Life happiness
4	25
12	54
3	21
14	80

Table 1. Data table

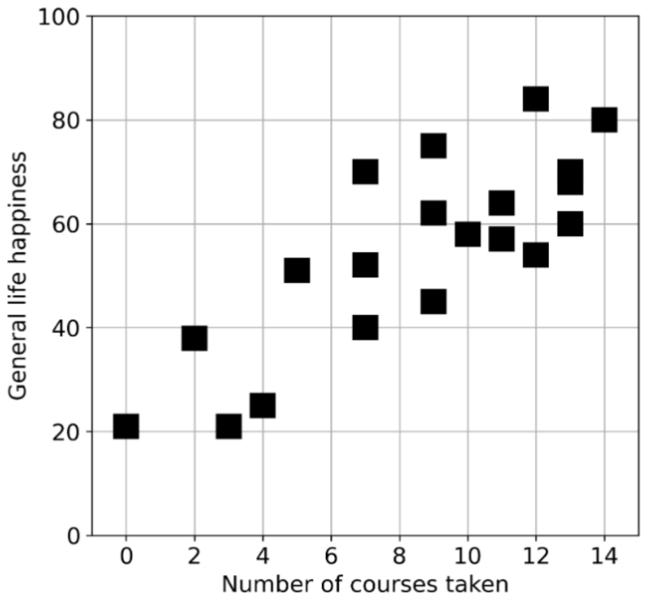


Fig 1. Fake data from fake survey

Create Design Matrix

- **Design matrix is actually only $\boxed{\text{one}}$ column vector.**
  - Because this is a simple model with only one predictor.

- **Matrix equation $X\beta = y$ looks like Eq 1. (Only first four data values).**

$$\begin{bmatrix} 4 \\ 12 \\ 3 \\ 14 \end{bmatrix} [\beta] = \begin{bmatrix} 25 \\ 54 \\ 21 \\ 80 \end{bmatrix}$$

Eq 1. Matrix equation $X\beta = y$

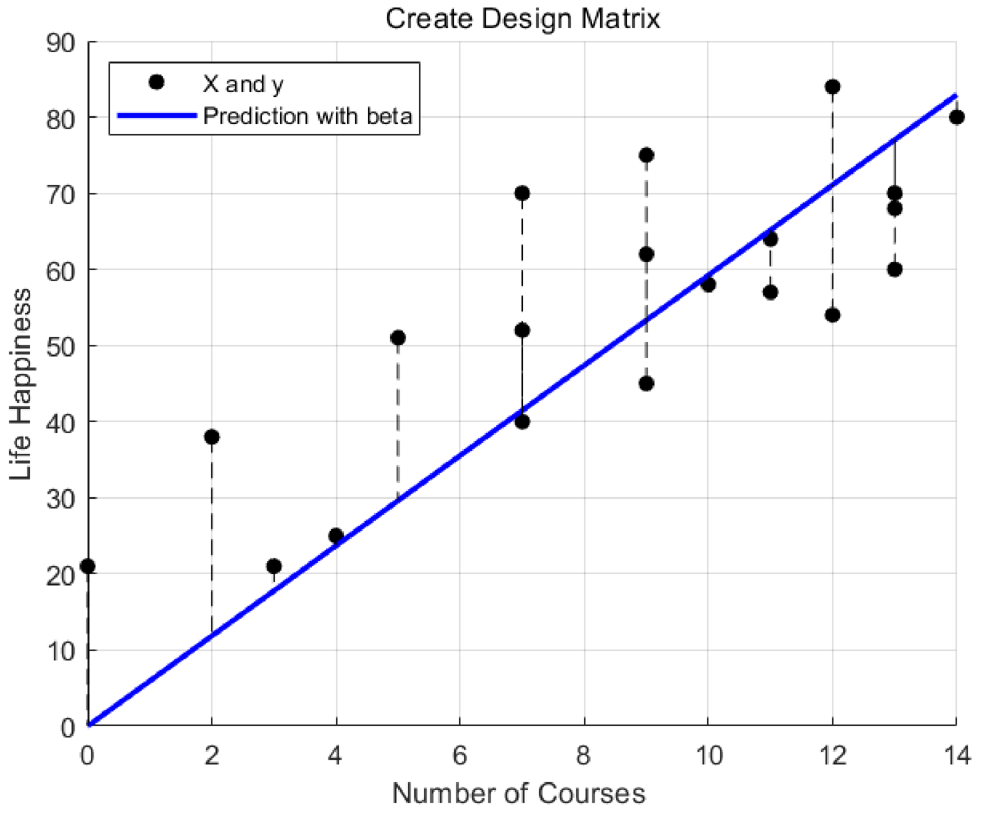
Code Exercise of Creating Design Matrix

Code Exercise (11_02)

- Follow the previous slide.
- The matrix equation looks like a form of $\boldsymbol{X\beta} = \boldsymbol{y}$.

```matlab
% Clear workspace, command window, and close all figures
clc; clear; close all;

% Define a matrix number of courses and life happiness
X = [4,12,3,14,13,12,9,11,7,13,11,9,2,5,7,10,0,9,7,13]';    % number of course
y = [25,54,21,80,68,84,62,57,40,60,64,45,38,51,52,58,21,75,70,70]';  % life happiness

% Compute the left-inverse of X
X_leftinv = ;

% Calculate beta
beta = ;
disp("beta");
disp(beta);

% Calculate y_pred with beta
y_pred = ;

% Plot
figure;
hold on;
grid on;
scatter(X, y, 'k', 'filled');  % X and y
plot(X, y_pred, 'b', 'LineWidth', 2);  % Plot predicted line with beta
for i = 1:length(X)
    plot([X(i) X(i)], [y(i) y_pred(i)], 'k--');  % Plot residuals as dashed lines
end
title('Create Design Matrix');
xlabel('Number of Courses ');
ylabel('Life Happiness');
legend('X and y', 'Prediction with beta', 'Location', 'northwest');
hold off;
```

MATLAB code of creating design matrix



Result of code

Meaning of Least Squares Formula's Result

- **Following least squares formula tells $\beta = 5.92$.**

- **What does this number mean?**
  - It means $\boxed{}$ in formula.
    - For each additional course that someone takes, their self-reported life happiness increases by 5.92 points.

- **Let's see how that result looks in plot as Fig 1..**

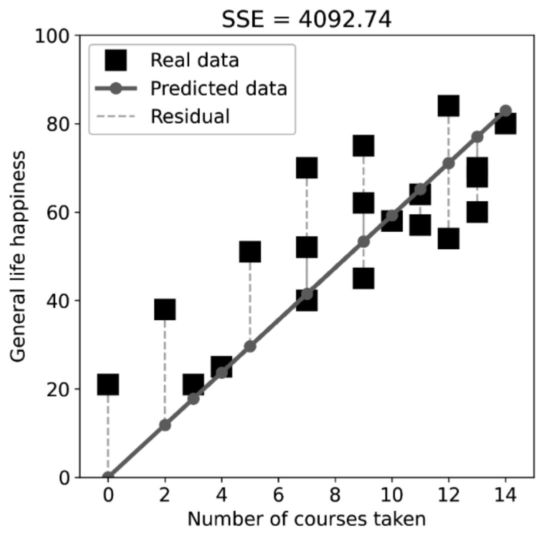


Fig 1. Observed and predicted data (SSE=sum of squared errors)

Feeling of Unease While Looking at Fig 1.

- **If you experience feeling of unease while looking at Fig 1.,**
  - Then, that's good signal!
    - It means you are thinking critically and noticed that **model doesn't do great job at minimizing errors**.
  - You can easily imagine pushing left side of best-fit line up to get better fit.

- **What's the problem here in term of mathematics?**

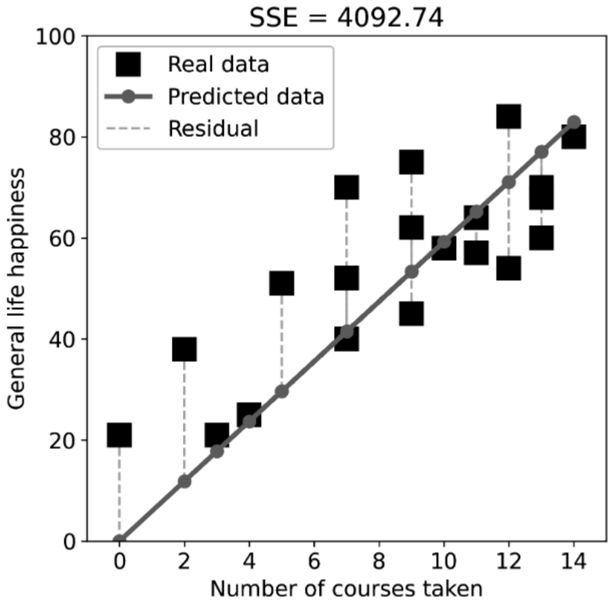


Fig 1. Observed and predicted data

Problem in Fig 1.

- **Design matrix contains no**
  - Equation of the best-fit line is $y = mx$.
    - Which means $x = 0,\ y = 0$.
    - That constraint doesn't make sense for this problem.
      - Because it means anyone who doesn't take courses is completely devoid of life satisfaction.

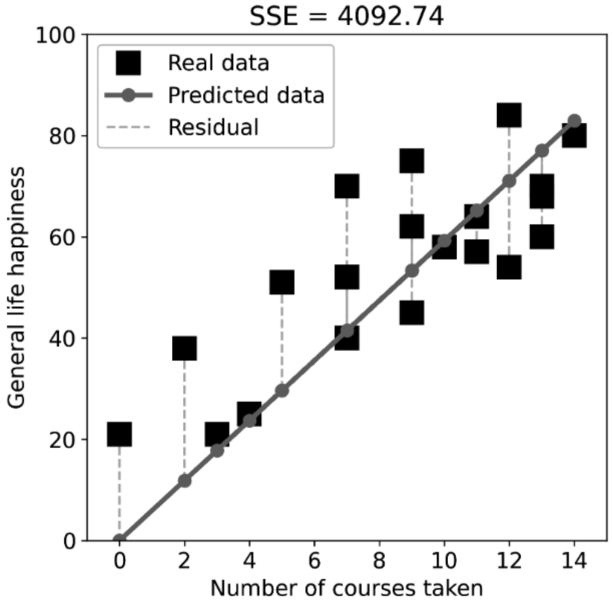


Fig 1. Observed and predicted data

Add Intercept Term

- **In form of $y = mx + b$.**
  - b is intercept term.
    - Allows the best-fit line to cross the y-axis at any value.

- **Statistical interpretation of intercept**
  - Expected numerical value of observations when predictors are set to zero.

- **Adding intercept term to design matrix as below Eq 1.**
  - Only showing first four rows.

- **Code doesn't change with one exception of creating design matrix.**

$$\begin{bmatrix} 1 & 4 \\ 1 & 12 \\ 1 & 3 \\ 1 & 14 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} = \begin{bmatrix} 25 \\ 54 \\ 21 \\ 80 \end{bmatrix}$$

Eq 1. Adding intercept term at design matrix

Code Exercise to Add Intercept Term

Code Exercise (11_03)

- Code is same with **Code Exercise (11_02)** with one exception.
- The difference is design matrix.
- Add intercept term in design matrix following the previous slide.

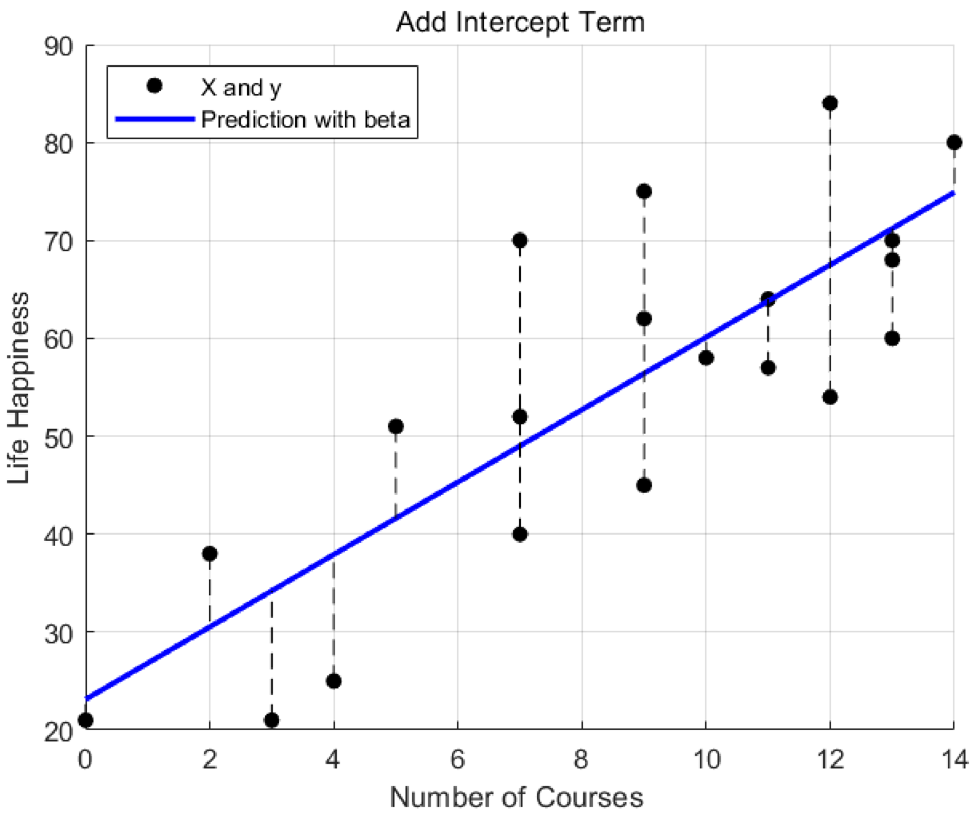
```
% Clear workspace, command window, and close all figures
clc; clear; close all;

% Define a matrix number of courses and life happiness
number_of_course = [4,12,3,14,13,12,9,11,7,13,11,9,2,5,7,10,0,9,7,13]';
life_happiness =
[25,54,21,80,68,84,62,57,40,60,64,45,38,51,52,58,21,75,70,70]';

% Define a new design matrix X that contains the intercept term and
% dependent variable matrix y
X = ; % Use number_of_course
y = ;

% Compute the left-inverse of X
X_leftinv = ;

% Calculate the beta
beta = ;   % [beta0 beta1]
beta = flip(beta);      % [beta1 beta0]
y_pred = polyval(beta, number_of_course);  % Predict y values using
beta
```

```
% Plot
figure;
hold on;
grid on;
scatter(number_of_course, y, 'k', 'filled');  % X and y
plot(number_of_course, y_pred, 'b', 'LineWidth', 2);  % Plot predicted
line with beta
for i = 1:length(X)
    plot([number_of_course(i) number_of_course(i)], [y(i) y_pred(i)],
'k--');  % Plot residuals as dashed lines
end
title('Add Intercept Term');
xlabel('Number of Courses ');
ylabel('Life Happiness');
legend('X and y', 'Prediction with beta', 'Location', 'northwest');
hold off;
```

MATLAB code to add Intercept term

Visual result of Code Exercise

Code Exercise (11_03)



Visual result of code exercise (11_03)

Result of Including Intercept Term

- **Now, $\boldsymbol{\beta}$ is two-element vector $[\mathbf{23.1, 3.7}]$.**
  - Expected level of happiness for someone who has taken zero courses is 23.1.
  - For each additional course someone takes, their happiness increase by 3.7 points.

- **You will agree that Fig 2. looks much better than Fig 1..**
  - And SSE is around half of what it was when we excluded intercept.

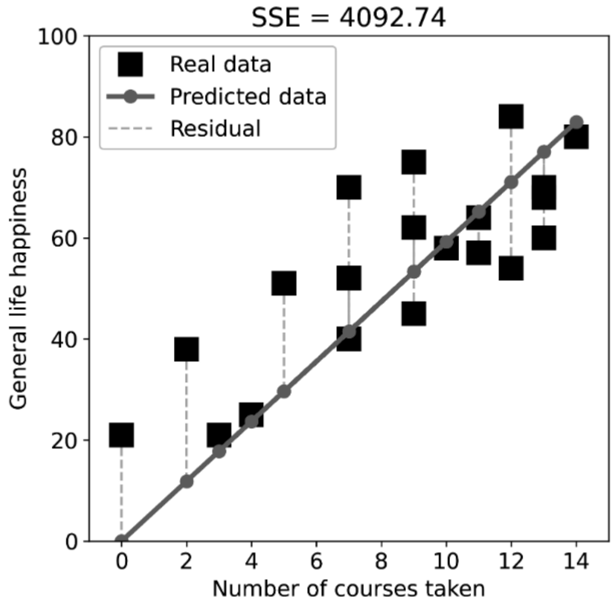


Fig 1. Observed and predicted data

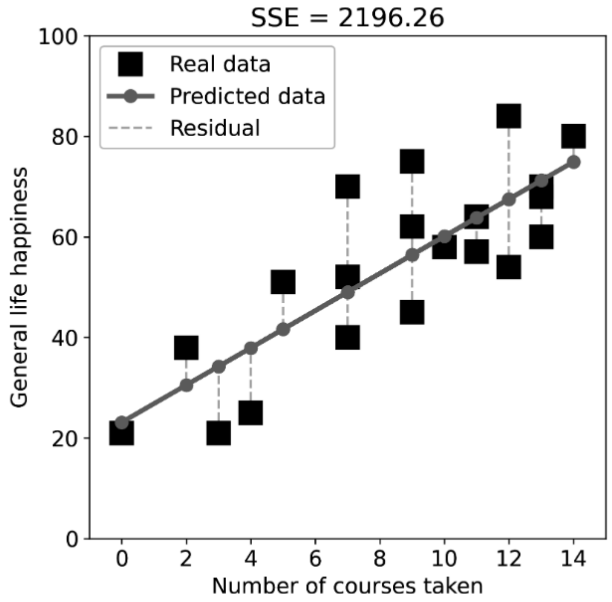


Fig 2. Observed and predicted data, with an intercept term

Least Squares via QR

Problem of Left-Inverse

- **Left-inverse method is theoretically reasonable but risks numerical instability.**
  - Because of computing matrix inverse which can be numerically unstable.

  - Matrix $\boldsymbol{X}^T\boldsymbol{X}$ itself can introduce difficulties.
    - Multiplying matrix by its transpose has implications.
      - Properties such as norm and condition number which you will learn more later.
    - Matrices with high condition number can be numerically unstable.
      - Thus, design matrix with high condition number will become even less numerically stable when squared.

Stable Way to Solve Least Squares Problem

QR decomposition

- Observe following sequence of equations as Eq 1..
- Eq 1. is slightly simplified.
  - From actual low-level numerical implementations.

$$
\begin{aligned}
\boldsymbol{X\beta} &= \boldsymbol{y} \\
\boldsymbol{QR\beta} &= \boldsymbol{y} \\
\boldsymbol{R\beta} &= \boldsymbol{Q}^{T}\boldsymbol{y} \\
\boldsymbol{\beta} &= \boldsymbol{R}^{-1}\boldsymbol{Q}^{T}\boldsymbol{y}
\end{aligned}
$$

Eq 1. Sequence of equation

Example of How to Increase Numerical Stability

- **$\boldsymbol{R}$ is same shape as $\boldsymbol{X}$.**
  - Tall (and therefore noninvertible)
  - Although only first N rows are nonzero.
    - Rows $N+1$ through M do not contribute to the solution.
    - In matrix multiplication, rows of zeros produce results of zeros.
  - Those rows can be removed.
    - From $\boldsymbol{R}$ and from $\boldsymbol{Q}^T\boldsymbol{y}$.
- **Row swaps**
  - Implemented via $\boxed{}$ matrices.
  - Might be used to increase numerical stability.

$$\boldsymbol{X\beta} = \boldsymbol{y}$$
$$\boldsymbol{QR\beta} = \boldsymbol{y}$$
$$\boldsymbol{R\beta} = \boldsymbol{Q}^T\boldsymbol{y}$$
$$\boldsymbol{\beta} = \boldsymbol{R}^{-1}\boldsymbol{Q}^T\boldsymbol{y}$$

Eq 1. Sequence of equation

Best Part of Eq 1.

- **Unnecessary to invert $\boldsymbol{R}$.**
  - Matrix is
  - Therefore, solution can be obtained via back substitution.
    - As solving simultaneous equations via Gauss-Jordan method.
      - Augment coefficients matrix by constants.
      - Reduce row to obtain RREF.
      - Extract solution from final column of augmented matrix.

$$\boldsymbol{X\beta} = \boldsymbol{y}$$
$$\boldsymbol{QR\beta} = \boldsymbol{y}$$
$$\boldsymbol{R\beta} = \boldsymbol{Q}^{\boldsymbol{T}}\boldsymbol{y}$$
$$\boldsymbol{\beta} = \boldsymbol{R}^{-1}\boldsymbol{Q}^{\boldsymbol{T}}\boldsymbol{y}$$

Eq 1. Sequence of equation

Conclusion of Least Squares Via QR Decomposition

- **QR decomposition solves least squares problem.**
  - **Without squaring $X^T X$.**
  - **Without explicitly inverting a matrix.**

- **Main risk of numerical instability comes from computing Q.**
  - This is fairly numerically stable.
    - When implemented via **Householder reflections**.

Summary

Summary

- **GLM is a statistical framework.**
  - To understand our rich and beautiful universe.
  - Works by setting up simultaneous equations.
    - Like that you learned about in previous lecture.
- **Different terms between linear algebra and statistics**
  - Once you learn terminological mappings, statistics becomes easier.
    - Because you already know math.
- **Least squares method of solving equations via left-inverse**
  - Foundation of many statistical analysis
  - You will often see least squares solution “hidden” inside seemingly complicated formulas.
- **Least squares formula**
  - Derived via algebra, geometry or calculus.
  - Multiple ways of understanding and interpreting least squares

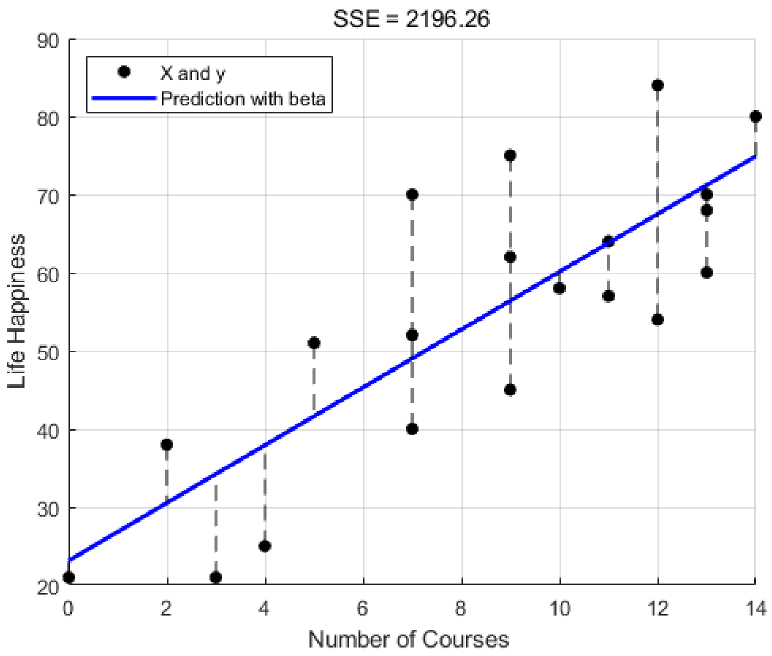
Summary

- **Multiplying observed data vector by left-inverse**
  - Right way to think about least squares
- **In practice, other methods are more numerically stable.**
  - Such as LU and QR decomposition

Code Exercises

SSE Calculation

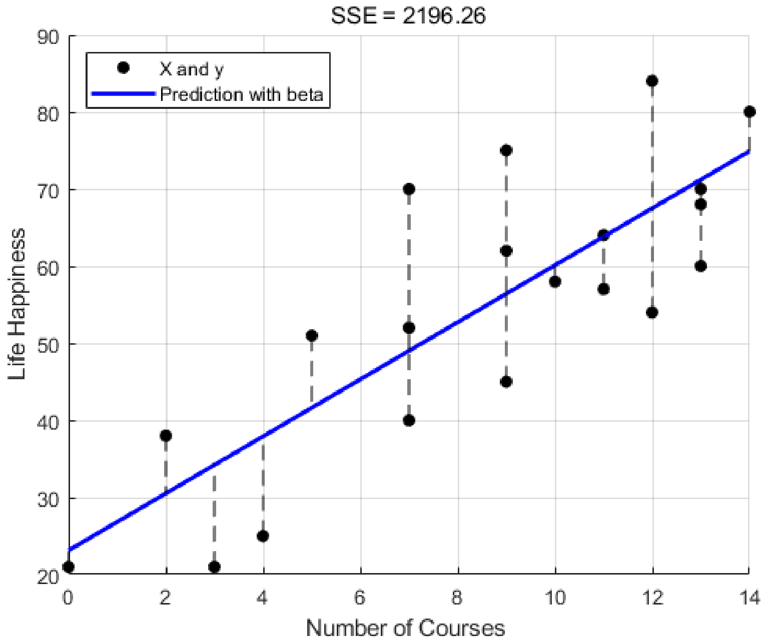
- Code Exercise (11_03) introduced the best fit line $y = mx + b$.
- Write code that calculate SSE(Sum of Squares Error) between real data and predicted data.



Result of the code

SSE Calculation using QR Decomposition

- You can also calculate beta using QR Decomposition.
- Write code that calculate SSE(Sum of Squares Error) between real data and predicted data using QR Decomposition.
- Hint: Calculate Economy-sized QR.
- Hint: Use '\' for get inverse of matrix R.



Result of the code

THANK YOU
FOR YOUR ATTENTION