Introduction to Linear Regression

BY JACK YANSONG LI
Liii Network at USTC
Email: yansong@liii.pro

Abstract

This is an abstract

Table of contents

1 Introduction/background

Linear Regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables. In matrix representation, it provides a compact and efficient way to express and solve the regression problem.

2 Case study: predicting house prices

Let's say we want to predict the price of a house (y) based on two features:

1. **Size of the house (in sq. ft.)** (x_1)
2. **Number of bedrooms** (x_2)

We collect data for N houses, where each house has:

- A known price (y)
- Known values for size (x_1) and bedrooms (x_2)

2.1 Scaler representation

Our goal is to find the best-fit linear relationship:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon, \tag{1}$$

where the symbols in Equation (1) are:

- β_0 = y-intercept (bias term)
- β_1, β_2 = coefficients for features x_1, x_2
- ϵ = error term (difference between predicted and actual price)

2.2 Matrix Representation

We can represent the regression problem in matrix form as:

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

where:

- y = **Dependent variable vector** (house prices)

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}.$$

- X = **Design matrix** (features with a column of 1's for the intercept)

$$X = \begin{bmatrix} 1 & x_{11} & x_{12} \\ 1 & x_{21} & x_{22} \\ \vdots & \vdots & \vdots \\ 1 & x_{N1} & x_{N2} \end{bmatrix}.$$

  - The first column is all `1`s (for β_0)
  - The second column is house sizes (x_1)
  - The third column is number of bedrooms (x_2)

- β = **Coefficient vector**

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix}.$$

- ϵ = **Error vector**

$$\epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_N \end{bmatrix}.$$

3 Solution to linear regression

Theorem 1. *The goal is to minimize the sum of square errors:*

$$\hat{\boldsymbol{\beta}} = \arg\min_{\beta} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2.$$

The closed-form solution is

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y}.$$

Proof. To derive the closed-form solution for the ordinary least squares (OLS) problem, we start with the minimization of the sum of squared errors:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2.$$

Here's a step-by-step derivation:

First, expand the squared norm (which is the sum of squared residuals):

$$\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\top (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$$

Expanding this gives:

$$\boldsymbol{y}^\top \boldsymbol{y} - \boldsymbol{y}^\top \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{\beta}^\top \boldsymbol{X}^\top \boldsymbol{y} + \boldsymbol{\beta}^\top \boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{\beta}.$$

Since $\boldsymbol{y}^\top \boldsymbol{X}\boldsymbol{\beta}$ is a scalar, it is equal to its transpose $\boldsymbol{\beta}^\top \boldsymbol{X}^\top \boldsymbol{y}$, so the expression simplifies to:

$$\boldsymbol{y}^\top \boldsymbol{y} - 2\,\boldsymbol{\beta}^\top \boldsymbol{X}^\top \boldsymbol{y} + \boldsymbol{\beta}^\top \boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{\beta}.$$

Take the Gradient with Respect to $\boldsymbol{\beta}$

To find the minimum, we take the derivative of the objective function with respect to $\boldsymbol{\beta}$ and set it to zero. Using matrix calculus rules:

$$\frac{\partial}{\partial \boldsymbol{\beta}} (\boldsymbol{y}^\top \boldsymbol{y} - 2\,\boldsymbol{\beta}^\top \boldsymbol{X}^\top \boldsymbol{y} + \boldsymbol{\beta}^\top \boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{\beta}) = -2\,\boldsymbol{X}^\top \boldsymbol{y} + 2\,\boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{\beta}.$$

Set the Gradient to Zero

Setting the gradient equal to zero gives the **normal equations**:

$$-2\,\boldsymbol{X}^\top \boldsymbol{y} + 2\,\boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{0}.$$

Divide both sides by 2 and rearrange:

$$\boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{X}^\top \boldsymbol{y}.$$

Solve for $\boldsymbol{\beta}$

Assuming $\boldsymbol{X}^\top \boldsymbol{X}$ is invertible (i.e., $\boldsymbol{X}$ has full column rank), multiply both sides by its inverse to obtain the closed-form solution:

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y}.$$

Summary of Key Steps:

1. Expand the squared error and simplify.

2. Compute the gradient of the objective with respect to $\boldsymbol{\beta}$.

3. Set the gradient to zero to derive the normal equations.

4. Solve the normal equations for $\boldsymbol{\beta}$ assuming $\boldsymbol{X}^\top \boldsymbol{X}$ is invertible.

Notes:

- The solution requires that $\boldsymbol{X}^\top \boldsymbol{X}$ be invertible. If $\boldsymbol{X}$ is not full rank (e.g., due to multicollinearity), a generalized inverse or regularization (e.g., ridge regression) may be used.

- This derivation assumes the errors are homoscedastic and uncorrelated; if not, weighted least squares or generalized least squares may be needed.

Let me know if you'd like further clarification on any step! □

V3> The goal is to minimize the sum of square errors:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\beta} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2.$$

The closed-form solution is

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y}.$$

how the closed-form solution derived? finish the missing details for me

V3>

4 Numerical experiment

The result of the experiment is shown in Figure 1. By Li et al. [1]

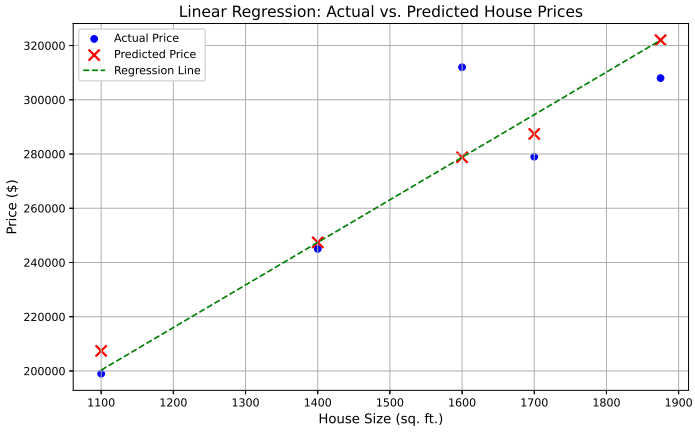


Figure 1. Result of linear regression

Bibliography

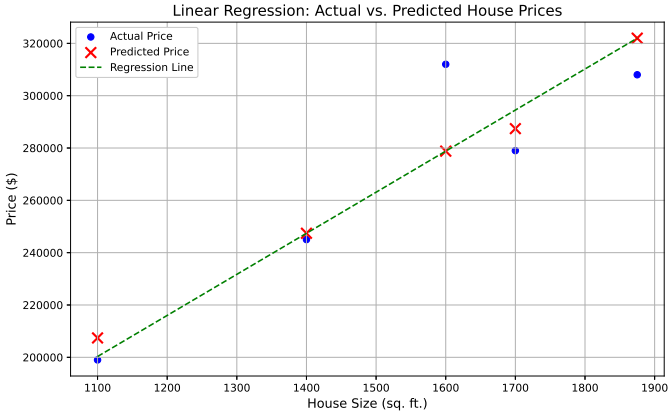
[1] Yansong Li and Shuo Han. Solving strongly convex and smooth stackelberg games without modeling the follower. In *American Control Conference, ACC 2023, San Diego, CA, USA, May 31 - June 2, 2023*, pages

2332-2337. IEEE, 2023.

A Code

```
>>> import matplotlib.pyplot as plt
    import numpy as np

    # Sample data
    X = np.array([
        [1400, 3],
        [1600, 3],
        [1700, 4],
        [1875, 3],
        [1100, 2]
    ])
    y = np.array([245000, 312000, 279000, 308000, 199000])

    # Add intercept term and calculate coefficients
    X_with_intercept = np.c_[np.ones(X.shape[0]), X]
    beta_hat = np.linalg.inv(X_with_intercept.T @ X_with_intercept) @
    X_with_intercept.T @ y
    y_pred = X_with_intercept @ beta_hat

    # Create plot without any transparency
    plt.figure(figsize=(10, 6))

    # Explicitly set solid colors (no alpha channel)
    plt.scatter(X[:, 0], y, color='blue', label='Actual Price', alpha=1.0)
    plt.scatter(X[:, 0], y_pred, color='red', marker='x', s=100,
                label='Predicted Price', alpha=1.0, linewidth=2)

    # Regression line with solid color
    mean_bedrooms = np.mean(X[:, 1])
    x_range = np.linspace(min(X[:, 0]), max(X[:, 0]), 100)
    plt.plot(x_range,
             beta_hat[0] + beta_hat[1]*x_range + beta_hat[2]*mean_bedrooms,
             color='green', linestyle='--', label='Regression Line', alpha=1.0)

    # Formatting
    plt.xlabel('House Size (sq. ft.)', fontsize=12)
    plt.ylabel('Price ($)', fontsize=12)
    plt.title('Linear Regression: Actual vs. Predicted House Prices', fontsize=14)
    plt.legend(framealpha=1.0)  # Solid legend background
    plt.grid(True, alpha=1.0)   # Solid grid lines

    # Print coefficients
    print("\nRegression coefficients:")
    print(f"Intercept (β0): {beta_hat[0]:.2f}")
    print(f"Size coefficient (β1): {beta_hat[1]:.2f}")
    print(f"Bedrooms coefficient (β2): {beta_hat[2]:.2f}")

Regression coefficients:
Intercept (β0): 48910.46
Size coefficient (β1): 156.99
Bedrooms coefficient (β2): -7097.42
```

```
>>> # Output the figure
    ps_out(plt.gcf())
```




```
>>>
```

```
   Elvish plugin v2025012216 by LiiiLabs
```

```
Elvish] ls
Elvish] pwd
```

```
   /Users/jackyansongli
```

```
Elvish] cat .zshrc
```

```
   # Enable Powerlevel10k instant prompt. Should stay close to the top of
   ~/.zshrc.
   # Initialization code that may require console input (password prompts, [y/n]
   # confirmations, etc.) must go above this block; everything else may go below.
   if [[ -r "${XDG_CACHE_HOME:-$HOME/.cache}/p10k-instant-prompt-${(%):-%n}.zsh"
   ]]; then
     source "${XDG_CACHE_HOME:-$HOME/.cache}/p10k-instant-prompt-${(%):-%n}.zsh"
   fi

   # If you come from bash you might have to change your $PATH.
   # export PATH=$HOME/bin:/usr/local/bin:$PATH
```

```
# Path to your oh-my-zsh installation.
export ZSH="$HOME/.oh-my-zsh"

# Set name of the theme to load --- if set to "random", it will
# load a random theme each time oh-my-zsh is loaded, in which case,
# to know which specific one was loaded, run: echo $RANDOM_THEME
# See https://github.com/ohmyzsh/ohmyzsh/wiki/Themes
ZSH_THEME="powerlevel10k/powerlevel10k"

# Set list of themes to pick from when loading at random
# Setting this variable when ZSH_THEME=random will cause zsh to load
# a theme from this variable instead of looking in $ZSH/themes/
# If set to an empty array, this variable will have no effect.
# ZSH_THEME_RANDOM_CANDIDATES=( "robbyrussell" "agnoster" )

# Uncomment the following line to use case-sensitive completion.
# CASE_SENSITIVE="true"

# Uncomment the following line to use hyphen-insensitive completion.
# Case-sensitive completion must be off. _ and - will be interchangeable.
# HYPHEN_INSENSITIVE="true"

# Uncomment one of the following lines to change the auto-update behavior
# zstyle ':omz:update' mode disabled  # disable automatic updates
# zstyle ':omz:update' mode auto      # update automatically without asking
# zstyle ':omz:update' mode reminder  # just remind me to update when
it's␣time

# Uncomment the following line to change how often to auto-update (in days).
# zstyle ':omz:update' frequency 13

# Uncomment the following line if pasting URLs and other text is messed up.
# DISABLE_MAGIC_FUNCTIONS="true"

# Uncomment the following line to disable colors in ls.
# DISABLE_LS_COLORS="true"

# Uncomment the following line to disable auto-setting terminal title.
# DISABLE_AUTO_TITLE="true"

# Uncomment the following line to enable command auto-correction.
# ENABLE_CORRECTION="true"

# Uncomment the following line to display red dots whilst waiting for
completion.
# You can also set it to another string to have that shown instead of the
default red dots.
# e.g. COMPLETION_WAITING_DOTS="%F{yellow}waiting...%f"
# Caution: this setting can cause issues with multiline prompts in zsh < 5.7.1
(see #5765)
# COMPLETION_WAITING_DOTS="true"

# Uncomment the following line if you want to disable marking untracked files
# under VCS as dirty. This makes repository status check for large
repositories
```

```
# much, much faster.
# DISABLE_UNTRACKED_FILES_DIRTY="true"

# Uncomment the following line if you want to change the command execution
time
# stamp shown in the history command output.
# You can set one of the optional three formats:
# "mm/dd/yyyy"|"dd.mm.yyyy"|"yyyy-mm-dd"
# or set a custom format using the strftime function format specifications,
# see 'man␣strftime' for details.
# HIST_STAMPS="mm/dd/yyyy"

# Would you like to use another custom folder than $ZSH/custom?
# ZSH_CUSTOM=/path/to/new-custom-folder

# Which plugins would you like to load?
# Standard plugins can be found in $ZSH/plugins/
# Custom plugins may be added to $ZSH_CUSTOM/plugins/
# Example format: plugins=(rails git textmate ruby lighthouse)
# Add wisely, as too many plugins slow down shell startup.
plugins=(git)

source $ZSH/oh-my-zsh.sh

# User configuration

# export MANPATH="/usr/local/man:$MANPATH"

# You may need to manually set your language environment
# export LANG=en_US.UTF-8

# Preferred editor for local and remote sessions
# if [[ -n $SSH_CONNECTION ]]; then
#   export EDITOR='vim'
# else
#   export EDITOR='mvim'
# fi

# Compilation flags
# export ARCHFLAGS="-arch␣x86_64"

# Set personal aliases, overriding those provided by oh-my-zsh libs,
# plugins, and themes. Aliases can be placed here, though oh-my-zsh
# users are encouraged to define aliases within the ZSH_CUSTOM folder.
# For a full list of active aliases, run `alias`.
#
# Example aliases
# alias zshconfig="mate␣~/.zshrc"
# alias ohmyzsh="mate␣~/.oh-my-zsh"

# To customize prompt, run `p10k configure` or edit ~/.p10k.zsh.
[[ ! -f ~/.p10k.zsh ]] || source ~/.p10k.zsh
source /opt/homebrew/share/zsh-autosuggestions/zsh-autosuggestions.zsh
source /opt/homebrew/share/zsh-syntax-highlighting/zsh-syntax-highlighting.zsh
source ~/.myzshrc
```

```
[ -f ~/.fzf.zsh ] && source ~/.fzf.zsh

# >>> juliaup initialize >>>

# !! Contents within this block are managed by juliaup !!

path=('/Users/jackyansongli/.juliaup/bin' $path)
export PATH

# <<< juliaup initialize <<<
```

```
Elvish]
```