

SDS 4130/5130: Linear Statistical Models

Topic 1. Simple Linear Regression Models

Mengxin (Maxine) Yu¹

¹Department of Statistics and Data Science
https://maxineyu.github.io/personal_web/

Washington University in St. Louis
August 25, 2025

Outline

- 1 Introduction
- 2 Description of Linear Regression Model
- 3 Least Squares Method

What is a statistical model?

- 1 An attempt to describe or explain how a give data set of interest was generated;
- 2 Due to inherent and/or systematic variability, the data changes from study to study; i.e., we cannot predict data in advanced.
- 3 Utilizes mathematical equations and probability distributions to describe the “chance” of particular data outcomes
- 4 Simplification or approximation of reality but can still be used to learn about complex systems
- 5 “All models are wrong but some are useful”- G.E. Box

Regression Models

- 1 We will focus on models that look at the relationship between a continuous or quantitative response (or dependent) variable Y and a set of explanatory variables $\mathbf{X} = (X_1, X_2, \dots, X_k)^T \in \mathbb{R}^k$ (also called predictors, or covariates). The Y and X_i 's are measurements taken from the same case, unit, or individual.
e.g., Think of Y as weight and $X_1 = \text{Height}$, $X_2 = \text{Age}$, $X_3 = \text{gender}$ (1 if female or 0 if male), etc.
- 2 Serve two major purposes:
 - **Description** of the structure or pattern in the data
 - **Prediction** of future values

Regression or Causality?

Regression is used for exploratory data analysis to

- assess the relationship between the response and some other explanatory variables,
- but regression itself is not equivalent with causality (Remember the case of “**Polio caused by Ice Cream!**” - google it). Watch out for **confounding variables!**

Motivating Example

Leaning of Tower Pisa

- Construction began in 1173 and by 1178 (2nd floor), it began to sink
- Construction resumed in 1272. To compensate for tilt, engineers built upper levels with one side taller
- Seventh floor completed in 1319 with bell tower added in 1372
- Tilt continued to grow over time and was monitored. Closed in 1990 for maintenance
- Stabilization completed in 2008 by removing ground from taller side

Leaning Tower of Pisa



The Data

- Prior to stabilization, annual measurements of its lean were taken for monitoring
- We have observations from 1975 - 1987
- Lean (Y) measured in tenths of a mm (response)
- Year (X) is the explanatory variable (predictor)
- **Goals:**
 - To **characterize** the evolution of lean over time
 - To **predict** future observations

The Data Set

Obs	year	lean
1	75	642
2	76	644
3	77	656
4	78	667
5	79	673
6	80	688
7	81	696
8	82	698
9	83	713
10	84	717
11	85	725
12	86	742
13	87	757

Studying the relationship using R

First, input the data. Two ways:

❶ Manually input in the command line:

```
pisadata=data.frame(  
  year=c(75,76,77,78,79,80,81,82,83,84,85,86,87),  
  lean=c(642,644,656,667,673,688,696,698,713,717,725,742,757))
```

❷ Import it from a *.csv file:

```
pisadata <- read.table("PisaData.csv", header=TRUE, sep=",")
```

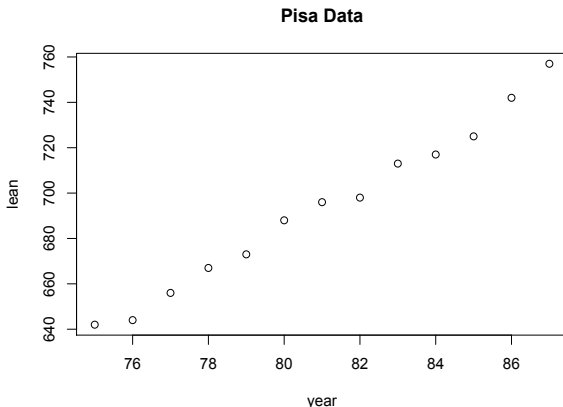
❸ **Note:** To download a data set (e.g., table.b4) from a package like MPV:

```
install.packages('MPV') # Run this line only the first time  
library(MPV)  
data(table.b4)
```

Always plot first!!!

```
plot(pisadata, main='Pisa Data') OR
```

```
plot(pisadata$year, pisadata$lean, main='Pisa Data')
```



Key Question: Do we observe a linear trend?

Straight Line Equation

- A straight line seems to account well for the variability in the data
- In other words, the graph suggests the following model:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i,$$

where β_1 is the slope, β_0 is the intercept, and ε_i is a “small” error accounting for the individual variability of each data point.

- Easy interpretation: β_1 is interpreted as the average lean increment per year, while β_0 is the expected initial lean at year 0.
- **Need to estimate β_0 and β_1 .**
- We are going to introduce a procedure for this:

Least Squares Method (see page 22 for details)

Least Squares Linear Regression in R

The procedure `lm()` in R is used to perform linear regression using the least-squares method:

```
> mymodel<-lm(lean~year,data=pisadata)
> summary(mymodel)
```

Call:

```
lm(formula = lean ~ year, data = pisadata)
```

Residuals:

Min	1Q	Median	3Q	Max
-5.9670	-3.0989	0.6703	2.3077	7.3956

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-61.1209	25.1298	-2.432	0.0333 *
year	9.3187	0.3099	30.069	6.5e-12 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.181 on 11 degrees of freedom

Multiple R-squared: 0.988, Adjusted R-squared: 0.9869

F-statistic: 904.1 on 1 and 11 DF, p-value: 6.503e-12

Fitted Model and Plotting

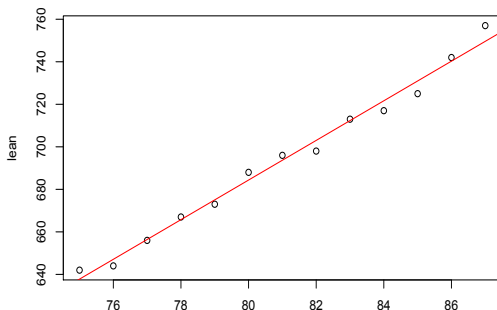
The output suggests the relationship:

$$\text{lean} = -61.1209 + 9.3187 \times \text{year} + \text{random error}$$

We can plot the best linear fit using the procedure `abline()`:

```
abline(coef(mymodel), col='red')
```

Pisa Data



Goodness of Fit Measurement: R^2

- 1 We are going to learn how to interpret and employ the regression output of `lm`. For now let me point out that one popular way to assess the goodness of fit of the model is via the so-called R^2 .
- 2 In the output, this is called "Multiple R-squared" and takes the value $R^2 = 0.988$.
- 3 The way to interpret R^2 is as the **percentage of variability in the response accounted for by the linear relationship with the independent variable**. So, in this case, about 98% is accounted by the linear relationship with year and only 2% by the error.
- 4 We shall precisely define R^2 and justify this interpretation later on.

The Simple Linear Regression Model

- ① Data: We observe the response and predictor variables (Y_i and X_i , respectively) of n cases or study units.

- ② The Simple Linear Regression Model assumes that:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \quad i=1, \dots, n,$$

- ③ ε_i is the random error of the i^{th} data point (X_i, Y_i) or i^{th} case.

- ④ Assumptions:

- Mean 0 $\longleftrightarrow \mathbb{E}(\varepsilon_i) = 0$
- Constant Variance $\sigma^2 \longleftrightarrow \text{Var}(\varepsilon_i) = \sigma^2$ (Homoskedasticity Assumption)
- Uncorrelated errors $\longleftrightarrow \text{Cov}(\varepsilon_i, \varepsilon_j) = 0$, for $i \neq j$.

Terminology and Interpretation

$$\boxed{Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i} \quad i=1, \dots, n,$$

- β_0 is an unknown constant parameter, called the **intercept**;
this will be the average value of the response in a case with $X = 0$;
- β_1 is another unknown constant parameter, called the **slope**
This is the **average change in the value of the response y**
produced by a unit change in x

Features of the Model

- 1 $Y_i = (\text{Constant or Deterministic Part}) + (\text{Random Part})$
 - constant or deterministic part $\beta_0 + \beta_1 X_i$
 - random part ε_i
- 2 Equivalently, we can write the model only in terms of properties of Y_i 's without the ε_i 's:
 - $\mathbb{E}(Y_i) = \beta_0 + \beta_1 X_i$
 - $\text{Var}(Y_i) = \sigma^2$
→ variance is the same regardless of X_i
 - $\text{Cov}(Y_i, Y_j) = \text{Cov}(\varepsilon_i, \varepsilon_j) = 0$, for any $i \neq j$
- 3 How to choose sensible values for β_0 and β_1 given n -observation pairs $(x_1, y_1), \dots, (x_n, y_n)$?

Estimation of the parameters: Idea

- 1 Let $\tilde{\beta}_0$ and $\tilde{\beta}_1$ denotes some tentative estimates or guesses for the true values β_0 and β_1 , respectively.
- 2 Then, the **fitted** or **predicted** value for y_i , denoted by \tilde{y}_i , is given by

$$\tilde{y}_i = \tilde{\beta}_0 + \tilde{\beta}_1 x_i$$

- 3 The idea is to think of the difference between the fitted value \tilde{y}_i and the observed value y_i ,

$$\tilde{e}_i = y_i - \tilde{y}_i,$$

as some loss or error, which we would like to minimize for all i 's.

Least Squares Estimators I

- The **Sum of Squares Errors**

$$S(\tilde{\beta}_0, \tilde{\beta}_1) = \sum_{i=1}^n \tilde{e}_i^2 = \sum_{i=1}^n (y_i - \tilde{y}_i)^2 = \sum_{i=1}^n (y_i - \tilde{\beta}_0 - \tilde{\beta}_1 x_i)^2$$

can be seen as a measure of the global error of the tentative regression line $\tilde{y} = \tilde{\beta}_0 + \tilde{\beta}_1 x$:

- The least-squares method aims to minimize $S(\tilde{\beta}_0, \tilde{\beta}_1)$ over all $\tilde{\beta}_0, \tilde{\beta}_1$.
- As it turns out, the previous optimization problem is well-posed (i.e., admits a unique solution) and its solution enjoys a closed form expression. **Drawback:** It can be sensitive to outliers.

Least Squares Estimators II

- The values of $(\hat{\beta}_0, \hat{\beta}_1)$ at which $S(\tilde{\beta}_0, \tilde{\beta}_1)$ is minimized over all $\tilde{\beta}_0, \tilde{\beta}_1$ are called the **(ordinary) least squares estimators (OLSE)**:

$$\sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 \leq \sum_{i=1}^n (y_i - \tilde{\beta}_0 - \tilde{\beta}_1 x_i)^2, \quad \text{for all } \tilde{\beta}_0 \text{ and } \tilde{\beta}_1.$$

Or in shorthand notation:

$$(\hat{\beta}_0, \hat{\beta}_1) := \operatorname{argmin}_{\tilde{\beta}_0, \tilde{\beta}_1} \sum_{i=1}^n (y_i - \tilde{\beta}_0 - \tilde{\beta}_1 x_i)^2.$$

Intuitively, the so-called **least-squares line** $y = \hat{\beta}_1 x + \hat{\beta}_0$ is the line that is “closest” to all the data points $\{(x_i, y_i)\}_{i=1}^n$ when distance is measured in terms of the sum-square errors

$$S(\hat{\beta}_0, \hat{\beta}_1) = \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2.$$

Normal Equations

By standard calculus (taking partial derivatives wrt $\tilde{\beta}_0$ and $\tilde{\beta}_1$, and then equating each of those to 0), it is easy to see that $\hat{\beta}_0$ and $\hat{\beta}_1$ must satisfy the following system of linear equations, called **normal equations**:

$$\begin{aligned} n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^n x_i &= \sum_{i=1}^n y_i, \\ \hat{\beta}_0 \sum_{i=1}^n x_i + \hat{\beta}_1 \sum_{i=1}^n x_i^2 &= \sum_{i=1}^n x_i y_i. \end{aligned}$$

Formulas for the Least Squares Estimators

From the normal equations we readily obtain that

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n x_i y_i - \frac{(\sum_{i=1}^n x_i)(\sum_{i=1}^n y_i)}{n}}{\sum_{i=1}^n x_i^2 - \frac{(\sum_{i=1}^n x_i)^2}{n}} =: \frac{S_{xy}}{S_{xx}},$$
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x},$$

where as usual $\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$ and $\bar{y} = \frac{\sum_{i=1}^n y_i}{n}$.

Remark: Note that the least squares line always passes through the “center” of (\bar{x}, \bar{y}) of the cloud of points $\{(x_i, y_i)\}_{i=1, \dots, n}$.

Other Formulas

Note that $S_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2$ and $S_{xy} = \sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x})$.

Therefore,

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} = r \frac{s_y}{s_x},$$

where

$$s_x^2 = \text{Sample Variance of the } x_i\text{'s} = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2,$$

$$s_y^2 = \text{Sample Variance of the } y_i\text{'s} = \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2,$$

$$r = \text{Sample Correlation} = \frac{\frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x})}{s_x s_y}.$$

Example

The Graduate Chair of Department Z administered a newly designed entrance test to 20 incoming Master's students as part of a study to determine whether a student's grade point average (GPA) at the end of the first year (Y) can be predicted from the entrance test score (X).

Based on the table in the following page:

- 1) Obtain the least squares estimates of β_0 and β_1
- 2) State the fitted regression function
- 3) Obtain an estimate of the GPA for an entrance test score of 5.0
- 4) State the expected change in grade point if the entrance test score were 0.5 units higher

X	Y	$X - \bar{X}$	$Y - \bar{Y}$	$(X - \bar{X})(Y - \bar{Y})$	$(X - \bar{X})^2$
5.5	3.1	0.5	0.6	0.30	0.25
4.8	2.3	-0.2	-0.2	0.04	0.04
4.7	3.0	-0.3	0.5	-0.15	0.09
3.9	1.9	-1.1	-0.6	0.66	1.21
4.5	2.5	-0.5	0.0	0.00	0.25
6.2	3.7	1.2	1.2	1.44	1.44
6.0	3.4	1.0	0.9	0.90	1.00
5.2	2.6	0.2	0.1	0.02	0.04
4.7	2.8	-0.3	0.3	-0.09	0.09
4.3	1.6	-0.7	-0.9	0.63	0.49
4.9	2.0	-0.1	-0.5	0.05	0.01
5.4	2.9	0.4	0.4	0.16	0.16
5.0	2.3	0.0	-0.2	0.00	0.00
6.3	3.2	1.3	0.7	0.91	1.69
4.6	1.8	-0.4	-0.7	0.28	0.16
4.3	1.4	-0.7	-1.1	0.77	0.49
5.0	2.0	0.0	-0.5	0.00	0.00
5.9	3.8	0.9	1.3	1.17	0.81
4.1	2.2	-0.9	-0.3	0.27	0.81
4.7	1.5	-0.3	-1.0	0.30	0.09
100.0	50.0	0.0	0.0	7.66	9.12

Answers

- 1) Obtain the least squares estimate of β_0 and β_1

Answer: $\hat{\beta}_1 = 0.84$, $\hat{\beta}_0 = -1.69$.

- 2) State the regression function

Answer: $\hat{y} = -1.69 + 0.84x$.

- 3) Obtain a point estimate for an entrance test score of 5.0

Answer: $\hat{y} = 2.5$.

- 4) State the expected change in grade point if the entrance test score were 0.5 units higher

Answer: $\Delta\hat{y} = 0.42$.