

Simple Linear Regression

Module 1

DATA 5600

Introduction to Regression and Machine Learning for Analytics

Marc Dotson

Module Overview

Introduction

- Vocabulary
- Characterize the relationship between two variables
- Correlation coefficient
- When/why use linear regression?
- Probabilistic vs deterministic model
- Linear model for describing an association
- Interpretation of model coefficients
- Fitting a linear model using least-squares regression
- Assumptions underlying a linear model

Vocabulary

Introduction

- You have (maybe) previously learned about
 - inference for a single mean (one-sample t-test),
 - for comparing two means (two-sample t-test: independent/paired), and
 - for comparing several means (ANOVA).
- What if the mean of one variable *depends* on the value of another variable?
- In the case of a linear trend (i.e., the value of one variable tends to increase or decrease linearly with an increase in another variable), we can fit a model that describes that trend.
- The tool most often used for this kind of analysis is called a **linear regression model**.

Car Gas Mileage

Introduction

Car	Weight (lbs)	MPG
1	3436	18
2	3433	16
3	3449	17
4	3086	14
5	2372	24
6	2833	22
7	2774	18
	⋮	
287	2295	32
288	2625	28
289	2720	31

- We have n pairs of points $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$.
- The **X variable** is called the **predictor**, the **explanatory variable**, or the **independent variable**.
- The **Y variable** is called the **response**, the **outcome variable**, or the **dependent variable**.
- Which variable is the independent variable and which is the dependent variable is determined based on the context of the analysis.

Introduction

- Determine the independent and dependent variable for the Car Gas Mileage example.

- $x_i =$
- $y_i =$
- $i = 1, 2, \dots, \underline{\hspace{1cm}}$
- $n =$

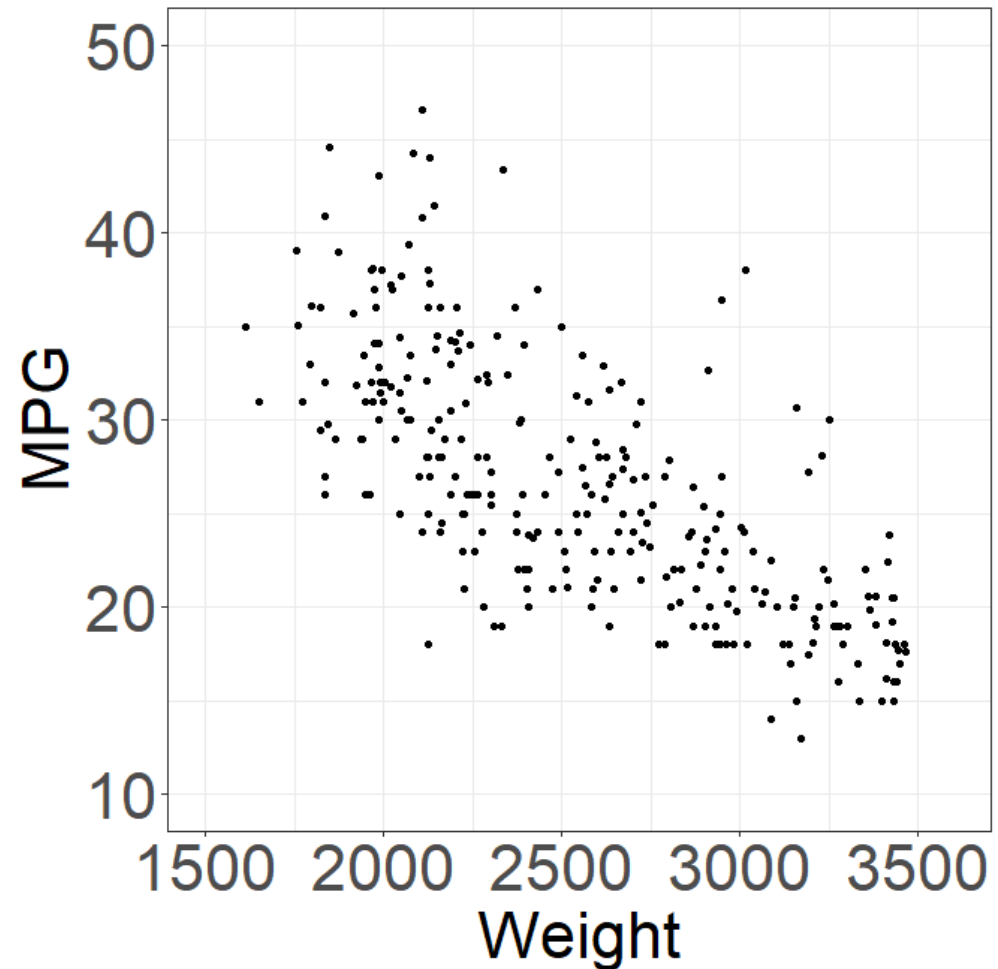
Scatterplots

Introduction

An important first step in exploring the relationship between two variables is a two-dimensional scatterplot, treating y as a function of x .

Each point in the plot represents one observation in the sample.

What relationship do you observe between weight and MPG?



Exploring Linear Association

Introduction

We often want to explore how these variables correlate with each other. That is, how does the value of y depend on x ? We will talk about two ways to do this:

1. Exploratory Data Analysis (EDA): numerically summarize the data using the correlation coefficient.
2. Statistical Inference: fit a simple linear regression model.
(Note: the “simple” in simple linear regression refers to using only one independent (x) variable to explain the dependent variable (y). Later, we will talk about using multiple explanatory variables to predict the trend of y .)

1. EDA: The Correlation Coefficient

The Correlation Coefficient

Correlation

- Correlation is a measure of the strength of association between two variables.
- The correlation coefficient is a widely-used summary statistic for illustrating the degree to which variables are linearly associated.
- The correlation coefficient r is defined as

$$\text{Cor}(X, Y) = r = \frac{\sum_{i=1}^n \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)}{n - 1} = \frac{SS_{XY}}{\sqrt{SS_{XX}} \sqrt{SS_{YY}}} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}$$

The Correlation Coefficient

Correlation

- r has a range between -1 and 1 .
- r is an index and has no units.
- The closer r is to 1 , the stronger the positive linear association. If $r = 1$, this indicates perfect positive correlation (i.e., all points in the scatterplot lie on an upward-sloping straight line).
- The closer r is to -1 , the stronger the negative linear association. If $r = -1$, this indicates perfect negative correlation.
- A value of r relatively close to zero indicates weak linear association. A value of $r = 0$ indicates no linear association.
- r measures linear association only. Two variables may be highly correlated in a nonlinear way that is not captured by r .
- r is highly affected by outliers! (will talk about outliers later)

The Correlation Coefficient

Correlation

guessthecorrelation.com

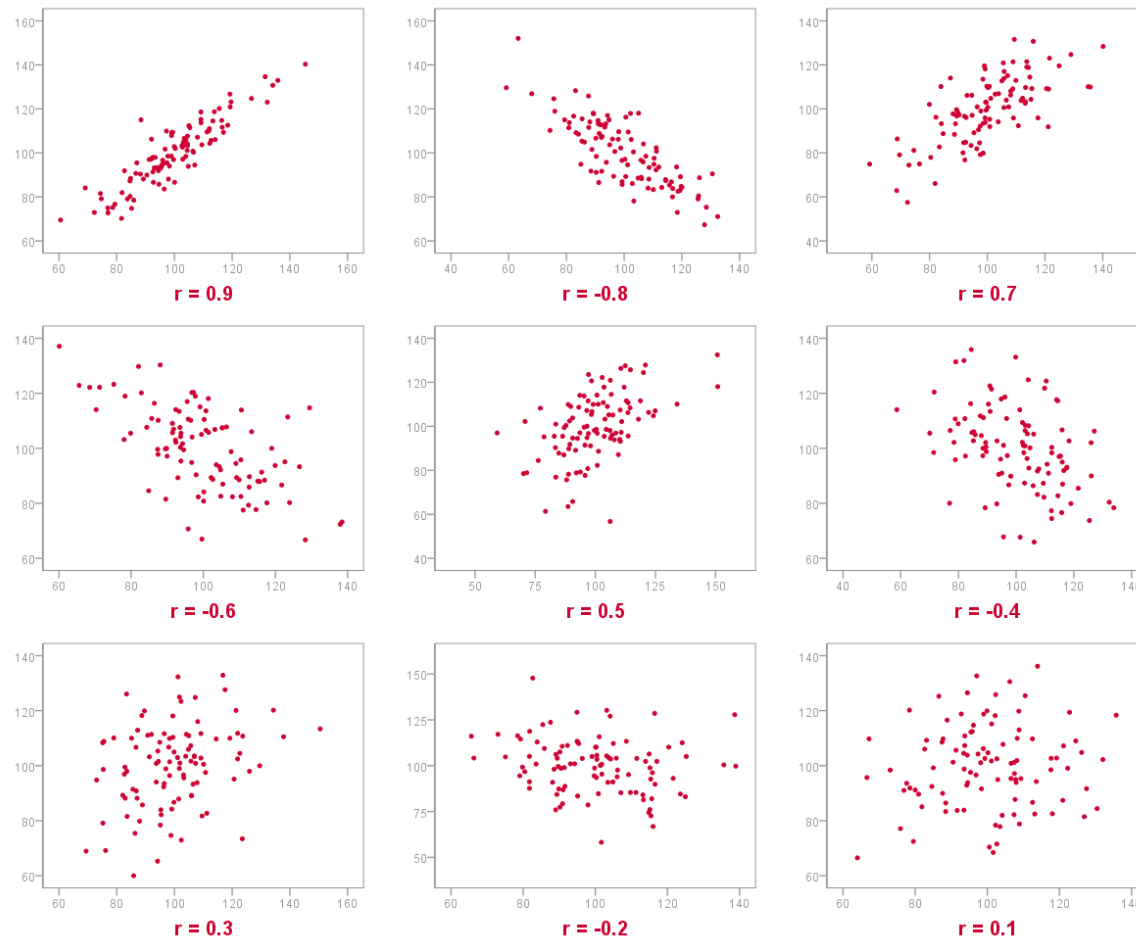


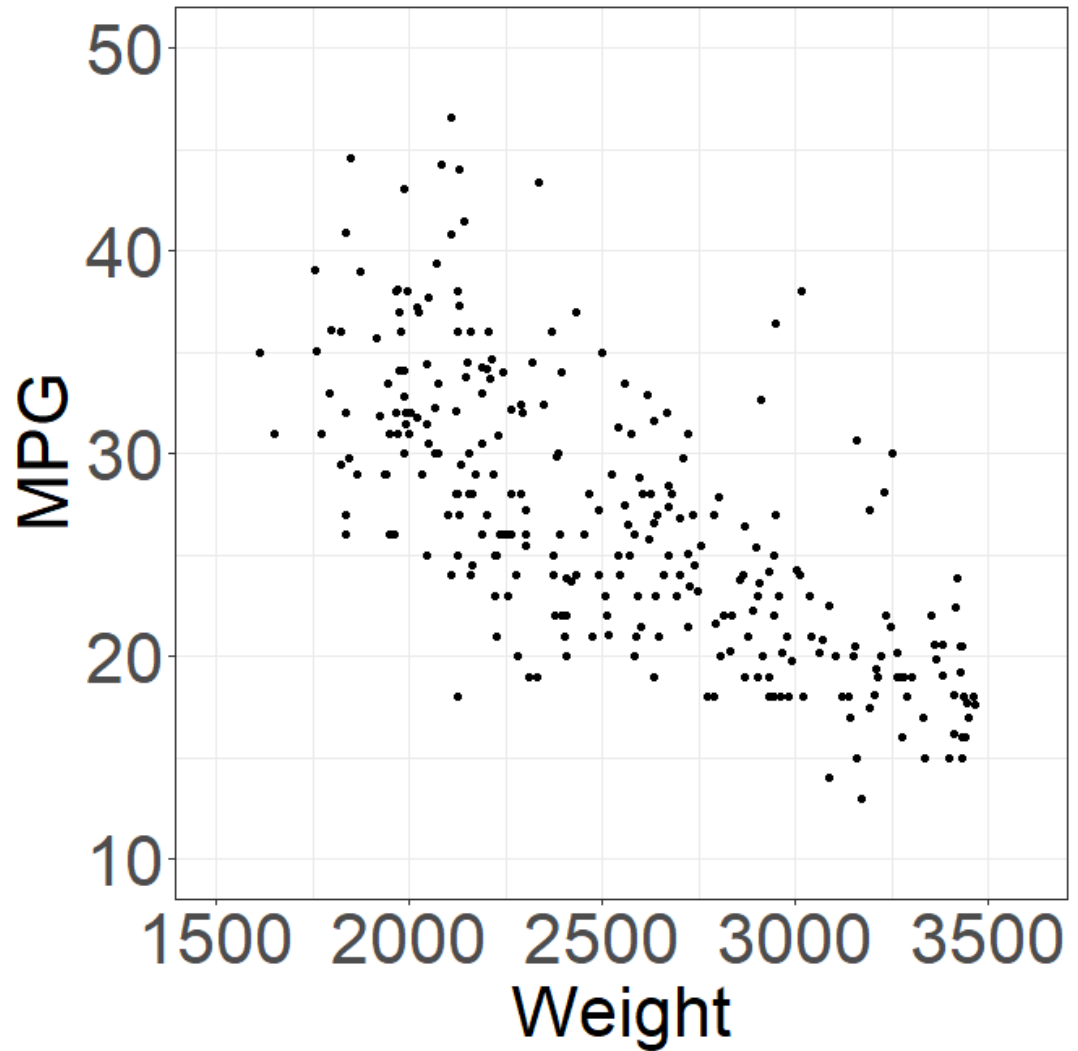
Figure taken from:

<https://www.google.com/url?sa=i&source=images&cd=&ved=2ahUKEwj2u9mf8KPkAhVGqZ4KHYNBCK8QjRx6BAGBEAQ&url=https%3A%2F%2Fwww.spsstutorials.com%2Fspss-correlation-analysis%2F&psig=AOvVaw2p-BXlwrJHWrYkE80VGQW8&ust=1567023675354711>

Scatterplots

Correlation

What do you think the correlation coefficient is for the Car Gas Mileage data?



2. Inference: Simple Linear Regression

Linear Regression Aims

Regression

- Remember, we would like to explore how two (or more) variables correlate with each other. That is, how does the value of Y *depend* on X ?
- In linear regression, we are specifically interested in how the average of Y , or the expected value $E(Y)$, depends on X .
- What is a good statistical model for this data?

Linear Regression Aims

Regression

- Remember, we would like to explore how two (or more) variables correlate with each other. That is, how does the value of Y *depend* on X ?
- In linear regression, we are specifically interested in how the average of Y , or the expected value $E(Y)$, depends on X .

$E(Y|X = x_i)$ <- conditional probability like you learned in DATA 3100

- What is a good statistical model for this data?

Characteristics of Two-Variable Relationships Regression

- The Car Gas Mileage scatterplot indicates two important features with respect to weight and MPG:
 - The *trend is linear* and negative – in other words, the *average* MPG tends to decrease in direct proportion to increasing weight.
 - There is *variability* around that trend – the MPG is not exactly the same for two cars with equal or similar weights.
- These are characteristics of what we call a **probabilistic** model.



General/Theoretical Linear Regression Model

Regression

- Linear regression is a probabilistic model that provides an effective tool for analyzing the kind of relationship we see between MPG and weight.
- There are two components to a linear regression model:

$$Y = \text{Deterministic Component} + \text{Random Error}$$

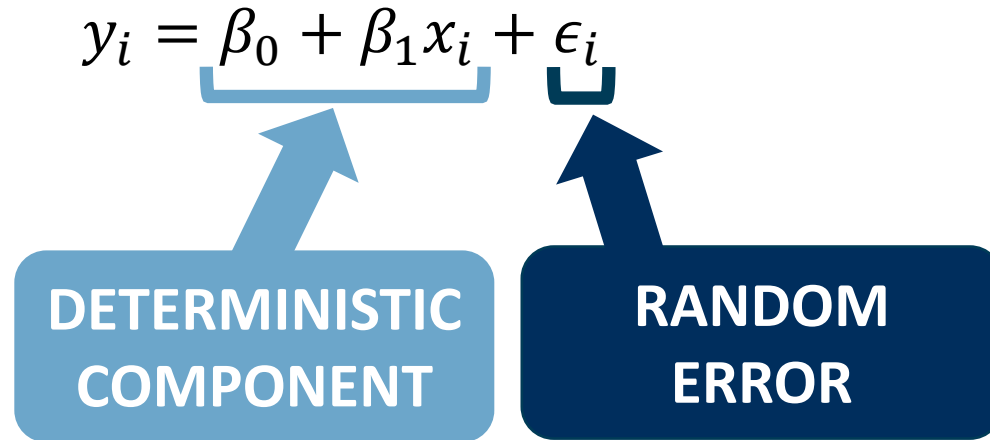
MODELS THE
TREND

MODELS THE
VARIABILITY

General/Theoretical Linear Regression Model

Regression

- A linear regression model looks more specifically like this:

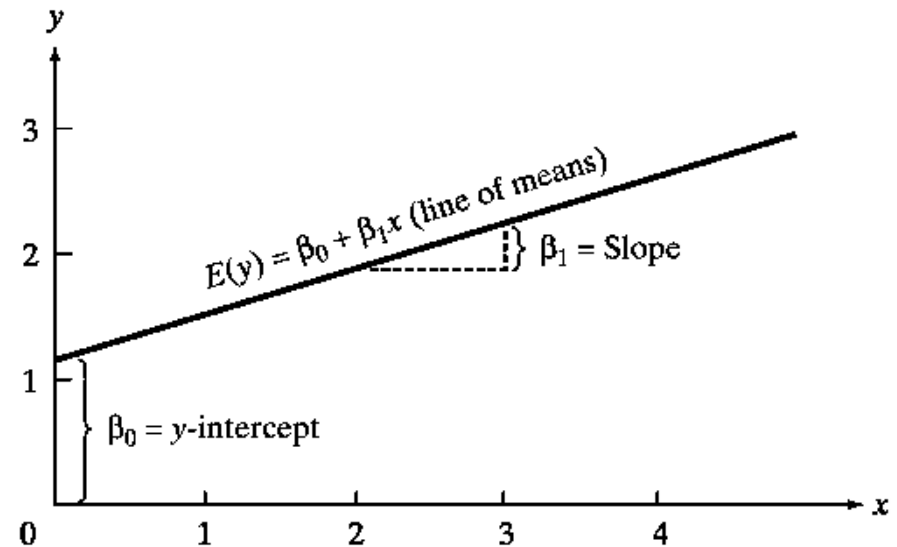
$$y_i = \underbrace{\beta_0 + \beta_1 x_i}_{\text{DETERMINISTIC COMPONENT}} + \underbrace{\epsilon_i}_{\text{RANDOM ERROR}}$$


- We generally assume that the **random error ϵ_i (a.k.a. error term)** has a mean of zero ($\epsilon_i \sim N(0, \sigma^2)$), meaning that this model tells us that:

Interpreting Model Parameters

Regression

- β_0 is the model **intercept** – it represents the average of Y when X is zero
- β_1 is the model **slope** – it represents the average change in Y for every one unit increase in X
- ϵ_i are the model **residuals** – they represent the difference between an *observed* value of Y and the *average* or *fitted* value of Y based on the linear model

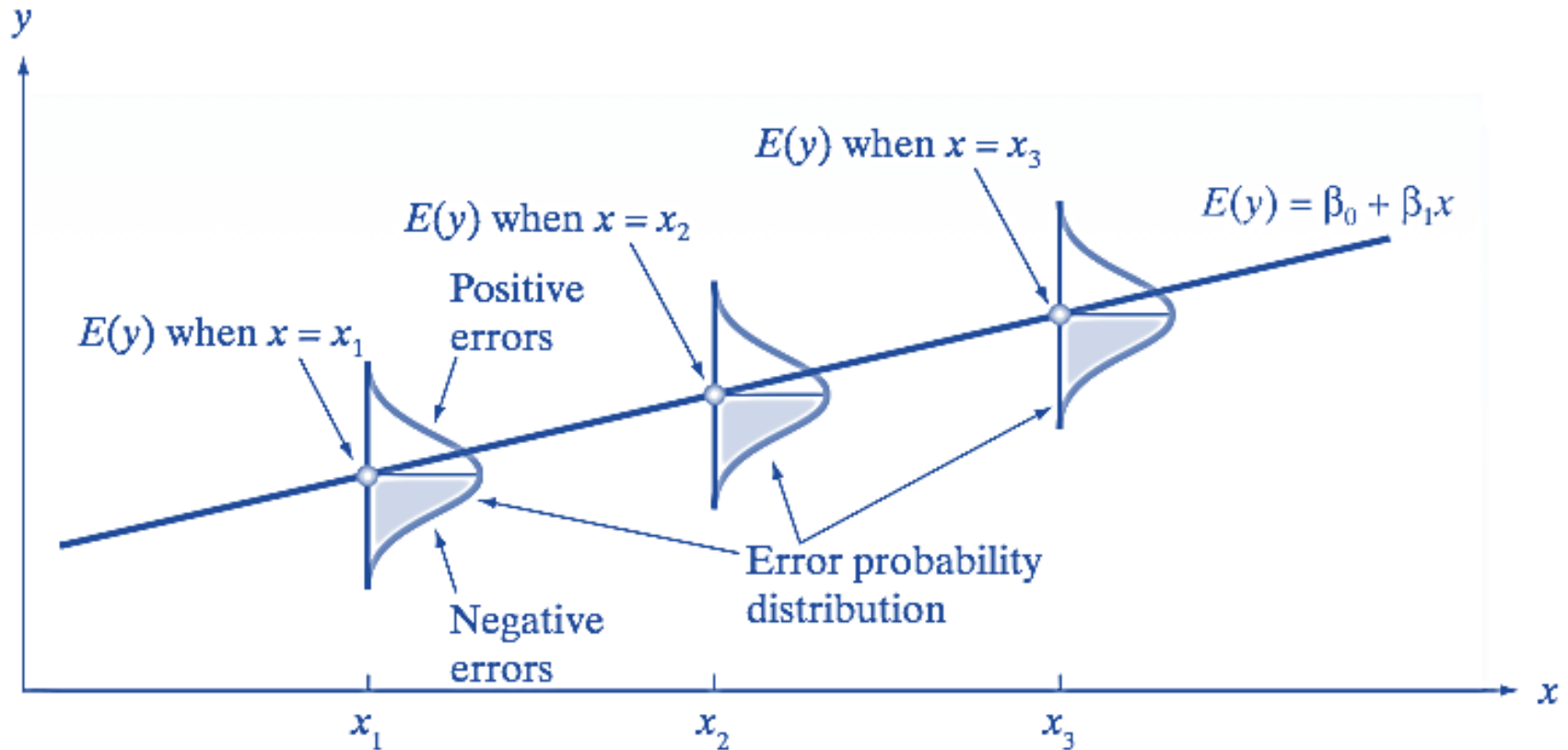


Regression Model Assumptions Regression

- **Remember:** the residuals ϵ_i in our regression model represents the difference between the regression line and the observed value of y_i .
 $\epsilon_i \sim N(0, \sigma^2)$.
- A linear regression assumes these things about the distribution of the residuals:
 - The average of the residuals is zero
 - The variance σ^2 is constant across all values of x_i (called “homoscedastic”, called “heteroscedastic” if nonconstant variance)
 - The distribution of ϵ_i is normal
 - The residuals are independent (generally true if the sample is random)

Illustrate Model Assumptions

Regression



Three Regression Parameters

Regression

- This statistical model has three parameters that we estimate:
 - the slope β_1 ,
 - the intercept β_0 ,
 - the variability around the line, also called the error variance or σ^2 .
- Note: σ^2 represents the variability of the residual values (ϵ_i).

Aside: Notation Notes

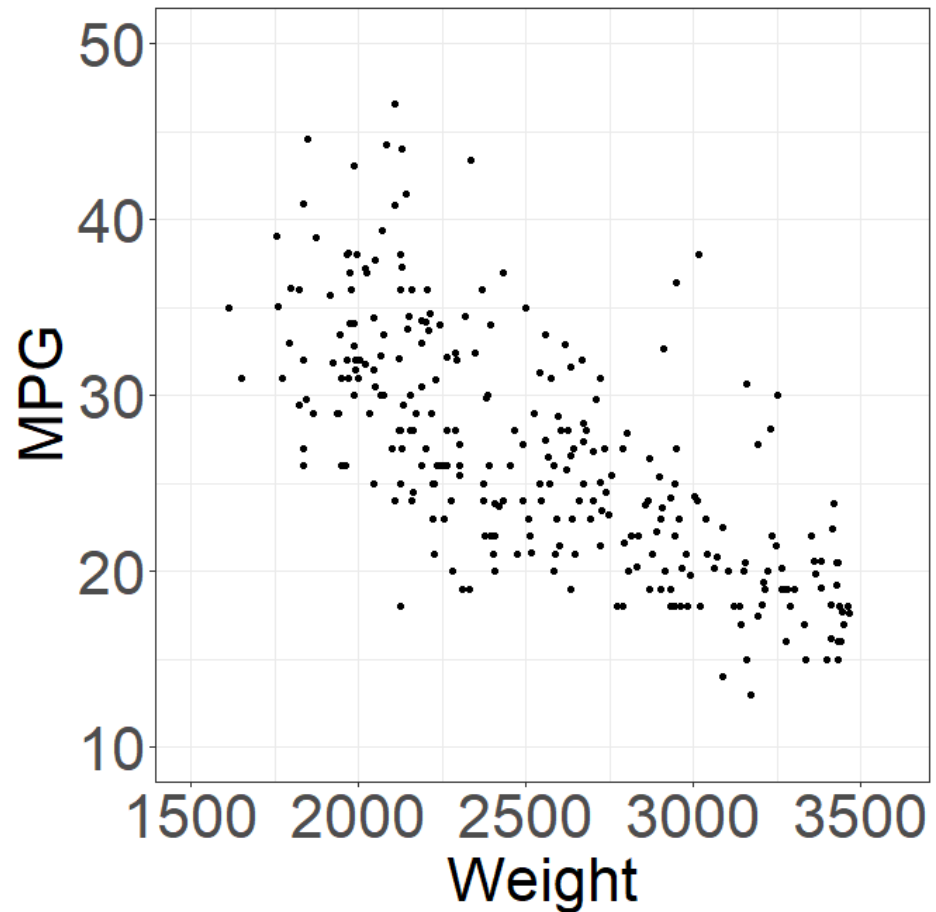
- β_1 is the true slope parameter (value is unknown)
- β_0 is the true intercept parameter (value is unknown)
- $\hat{\beta}_1$ is the estimate of β_1 computed from the linear regression model
- $\hat{\beta}_0$ is the estimate of β_0 computed from the linear regression model
- \bar{y} is the average of Y computed from the data
$$\left(\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i = \frac{y_1 + y_2 + \cdots + y_n}{n} \right)$$
- \hat{y} is the average/predicted/fitted/expected value of Y computed from the linear regression model. $(\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i)$

Fitting The Model (Estimating Parameters)

How Do We Fit a Line Using Our Data?

Estimation

Consider again the scatterplot for the Car Gas Mileage data. If we want to model average MPG as function of weight, which line represents the “best” fit for our data?



Least-Squares Regression Line Estimation

- So, which of the many lines we can draw on the scatterplot is “best”?
 - **Answer:** an effective approach is what we call the **ordinary least-squares (OLS)** model fit.
 - There are other ways: maximum likelihood estimation, Bayesian estimation, etc.
- A least-squares regression estimates the intercept and slope from the data by choosing the values that minimize the squared residuals
- In other words, *the least-squares line is the line with the smallest squared distance from the observed y values*

How are the Slope and Intercept Computed for the OLS Fit?

Estimation

We know:

- The errors/residuals are: $\epsilon_i = y_i - \hat{y}_i = y_i - (\beta_0 + \beta_1 x_i)$ $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$
- We square the residuals to “penalize” points above and below the line equally:

$$\epsilon_i^2 = (y_i - (\beta_0 + \beta_1 x_i))^2$$

- We add up the squared errors to get our “objective/loss function”:

$$\sum_{i=1}^n \epsilon_i^2 = \sum_{i=1}^n (y_i - (\beta_0 + \beta_1 x_i))^2 = O(\beta_0, \beta_1)$$

We can't change the data (x and y) to minimize this, but we can control the values of the parameters

- We choose β_0 and β_1 to minimize the objective function:

$$\min_{\beta_0, \beta_1} O(\beta_0, \beta_1) = \min_{\hat{\beta}_0, \hat{\beta}_1} \sum_{i=1}^n (y_i - (\beta_0 + \beta_1 x_i))^2$$

Find the β_0 that Minimizes the Estimation
Objective Function: $\sum_{i=1}^n (y_i - (\beta_0 + \beta_1 x_i))^2$

BONUS

Find the β_1 that Minimizes the Estimation
Objective Function: $\sum_{i=1}^n (y_i - (\beta_0 + \beta_1 x_i))^2$

BONUS

How are the Slope and Intercept^{Estimation} Computed for the OLS Fit?

- In summary:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{SS_{XY}}{SS_{XX}} = \frac{s_y}{s_x} \text{Corr}(X, Y)$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

- The OLS fit for the Car Gas Mileage data is given by $\hat{y}_i = 51.59 - 0.0098x_i$:
(Note that $\bar{x} = 2325$ lbs and $\bar{y} = 26.66$ mpg for these data.)
 - The estimated slope is

$$\hat{\beta}_1 = \frac{(3436 - 2535)(18 - 26.66) + (3433 - 2535)(16 - 26.66) + \dots + (2720 - 2535)(31 - 26.66)}{(3436 - 2535)^2 + (3433 - 2535)^2 + \dots + (2720 - 2535)^2} = -0.0098$$

- The estimated intercept is

$$\hat{\beta}_0 = 26.66 - (-0.0098)(2325) = 51.59$$

Estimating the Model Variance

Estimation

σ^2

- The model variance σ^2 represents the *average squared variability of the residuals around the regression line*
- To estimate σ^2 , we compute the *observed* average squared residual, based on the model fit:

$$\hat{\sigma}^2 = s^2 = \frac{\text{SSE}}{\text{degrees of freedom for error}} = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n - 2} = \frac{\sum_{i=1}^n e^2}{n - 2}$$

where SSE represents the Sum of Squared Errors (also called RSS, Residual Sum of Squares)

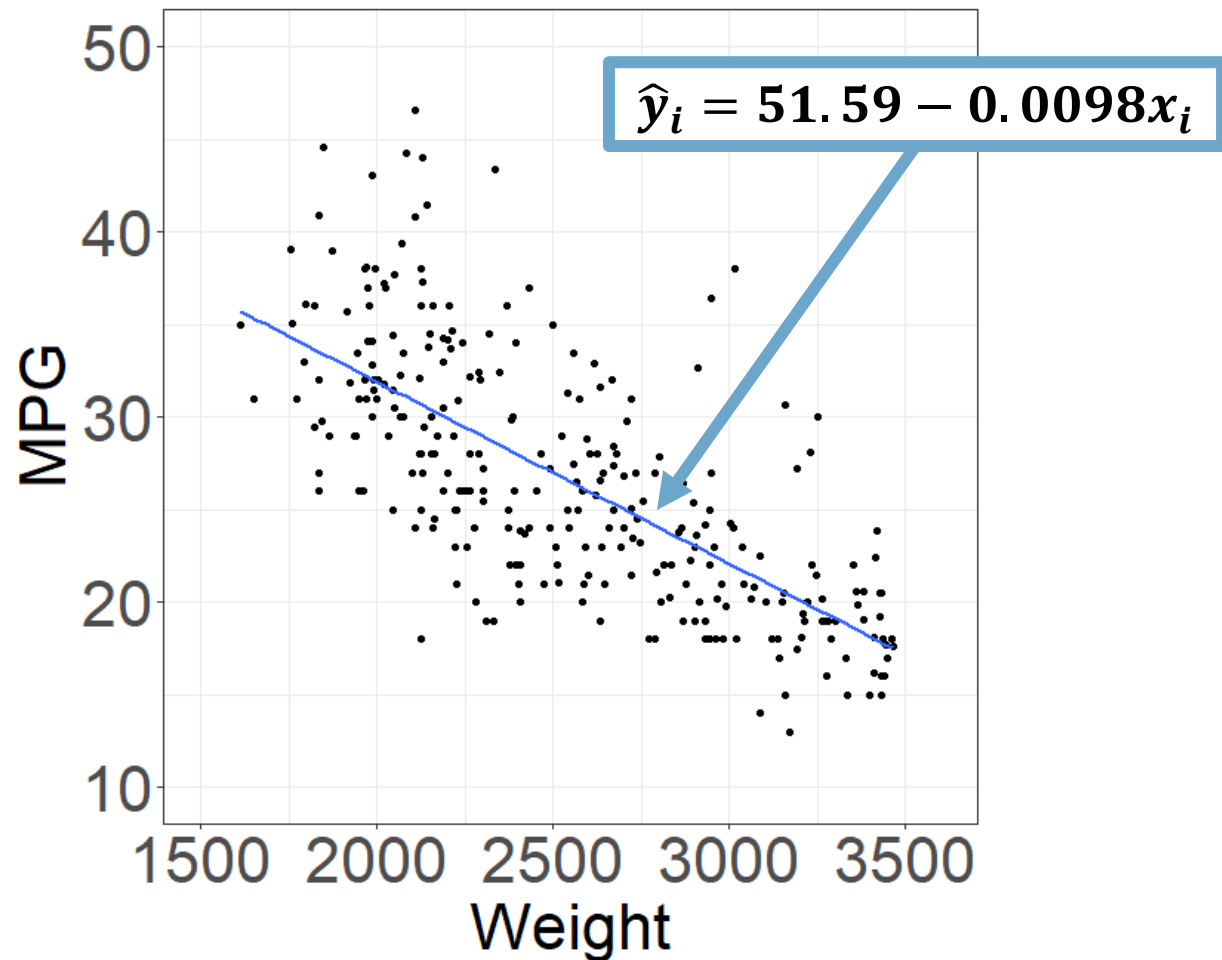
- s^2 is also often referred to as the Mean Squared Error, or MSE
- For the Car Gas Mileage data:

$$s^2 = \frac{(18 - 17.80)^2 + (16 - 17.83)^2 + \dots + (31 - 24.84)^2}{289 - 2} = 22.31$$

Using the Model

Least-Squares Line for the Car Gas Mileage Data

Model



Least-Squares Estimates from Python

Model

OLS Regression Results

```
=====
Dep. Variable:          MPG    R-squared:                0.505
Model:                  OLS    Adj. R-squared:           0.503
Method:                 Least Squares    F-statistic:            292.6
Date:                  Tue, 29 Aug 2023    Prob (F-statistic):      1.04e-45
Time:                  11:06:45    Log-Likelihood:         -857.72
No. Observations:      289    AIC:                    1719.
Df Residuals:          287    BIC:                    1727.
Df Model:               1
Covariance Type:       nonrobust
=====
```

	coef	std err	t	P> t	[0.025	0.975]
const	51.5872	1.484	34.773	0.000	48.667	54.507
Weight	-0.0098	0.001	-17.106	0.000	-0.011	-0.009

```
=====
Omnibus:                26.176    Durbin-Watson:           0.699
Prob(Omnibus):           0.000    Jarque-Bera (JB):        31.989
Skew:                   0.693    Prob(JB):                1.13e-07
Kurtosis:                3.857    Cond. No.:               1.38e+04
=====
```

Interpretation of Least-Squares Estimates for the Car Gas Mileage Data

Model

$$\hat{y}_i = 51.59 - 0.0098x_i$$

- Interpret 51.59 in context of the car gas mileage data set
- Interpret -0.0098 in context of the car gas mileage data set
- What is the residual for the first observation ($y_1 = 18$ and $\hat{y}_1 = 17.80$)?

Least-Squares Line for the Car Gas Mileage Data

Model

$$\hat{y}_i = 51.59 - 0.0098x_i$$

- How would you use the simple linear regression model to predict the MPG for a car weighing 3000 lbs.?
- How do you interpret the number you just computed?
- What is the predicted MPG for a weight of 10,000 lbs.?

Summary

Summary

Summary

Theoretical/General Model

General Form:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \text{ where } \epsilon_i \stackrel{\text{iid}}{\sim} N(0, \sigma^2)$$

or

$$y_i \stackrel{\text{ind}}{\sim} N(\beta_0 + \beta_1 x_i, \sigma^2)$$

Always add context when you know the data:

$$\text{MPG}_i = \beta_0 + \beta_1 \text{Weight}_i + \epsilon_i$$

where $\epsilon_i \stackrel{\text{iid}}{\sim} N(0, \sigma^2)$

Fitted Model

General Form:

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

Adding context:

$$\widehat{\text{MPG}}_i = 51.59 - 0.0098 \text{Weight}_i$$