Outline

1) Linear models – simple linear regression

2) Model formulation

Simple Linear Regression

Mitsubishi Example:

Ingrid wants to buy second-hand Mitsubishi Sigmas for her business. To estimate the cost, her team collects data on the age and price of 39 cars.

Can she use this data to predict how much she'll pay for the cars?

Mitsubishi Example

Download the dataset (**mitsub.txt**) and load it in RStudio.

A first step might be to look at some summary statistics:

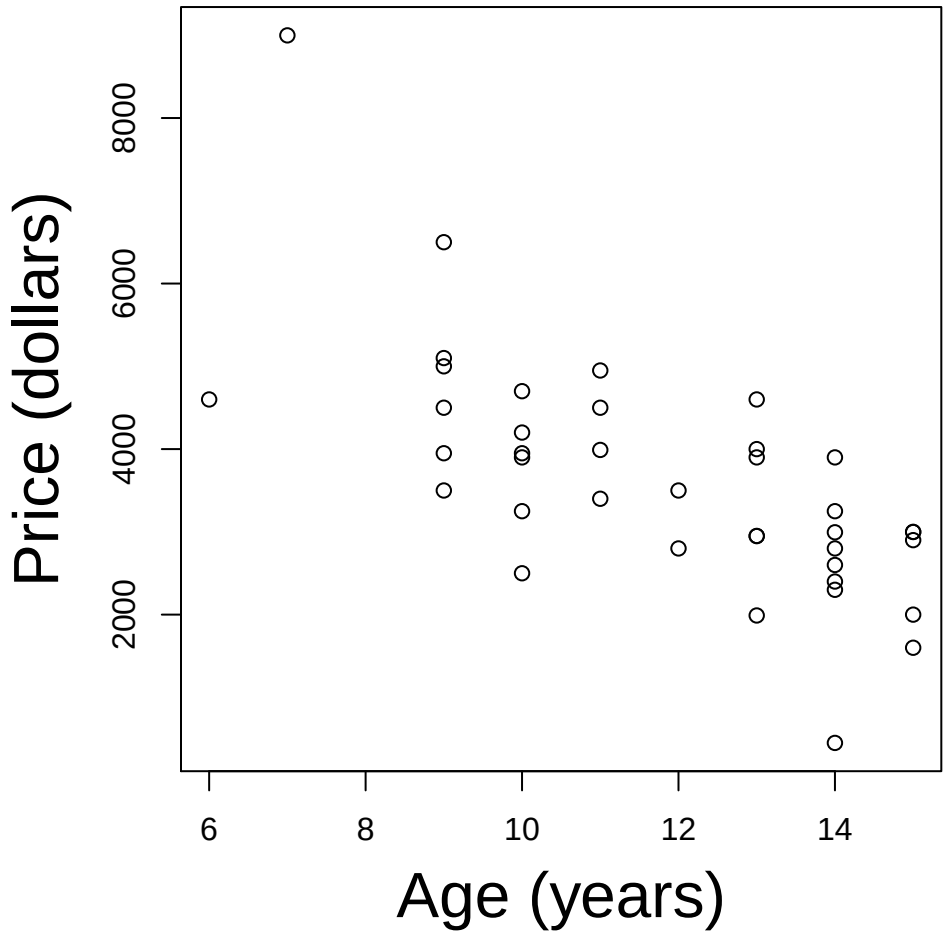
```
> summary(mitsub)
      age             price
 Min.   : 6.00   Min.   : 450
 1st Qu.:10.00   1st Qu.:2850
 Median :12.00   Median :3500
 Mean   :11.79   Mean   :3625
 3rd Qu.:14.00   3rd Qu.:4350
 Max.   :15.00   Max.   :8999
```

Some standard statistical calculations allow Ingrid to predict with 95% certainty that a randomly chosen second hand sigma will cost between \$668 and \$6625. Use the **quantile** function to find these values

Mitsubishi Example

- Ingrid needs to estimate the cost of building her fleet.
- Current prediction range is too wide to be useful.
- Improvement idea:
  - Use additional data, like car age, to improve predictions.
  - **age** and **price** are likely correlated.
- Scatterplot of **price** vs **age** (below) shows prices decrease as cars age.

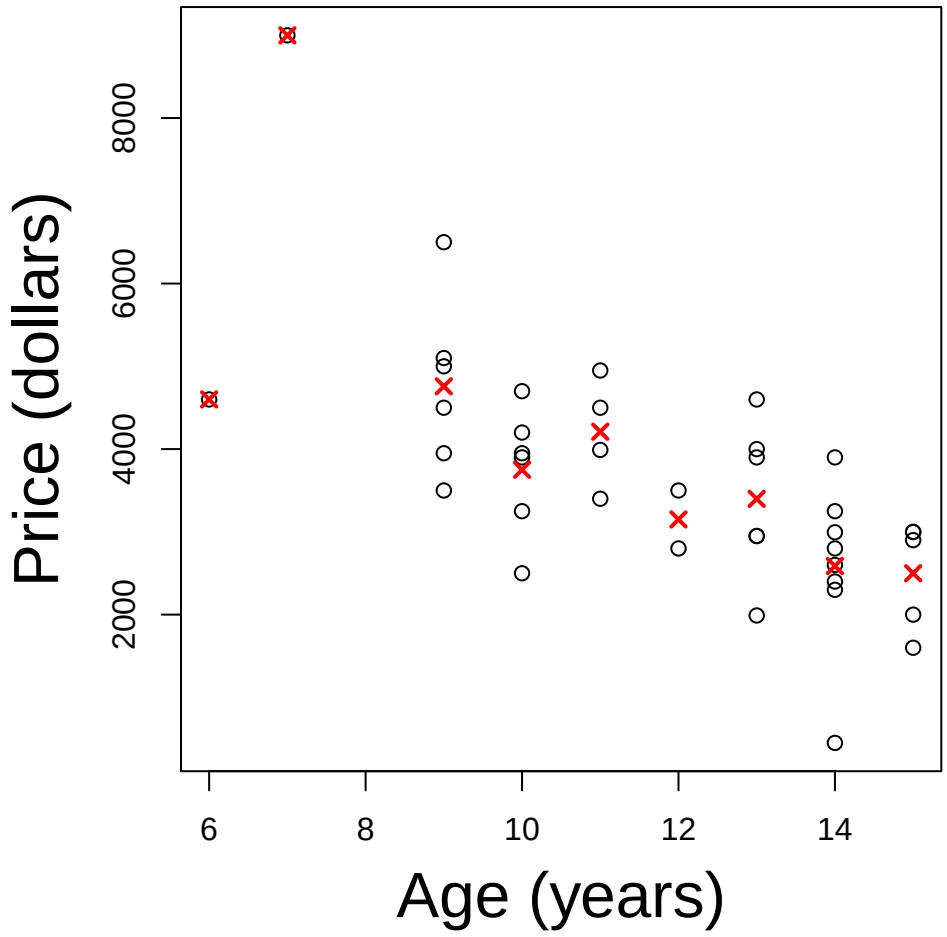
Mitsubishi Example



Mitsubishi Example

- Goal: Predict the price of a 10-year-old car.
- Average price for 10-year-old cars: $3750.
- Average price for 14-year-old cars: $2587. Find these values using R.
- Repeat for different ages to get average prices at various ages.
- The plot below shows how average price changes with age.

Mitsubishi Example



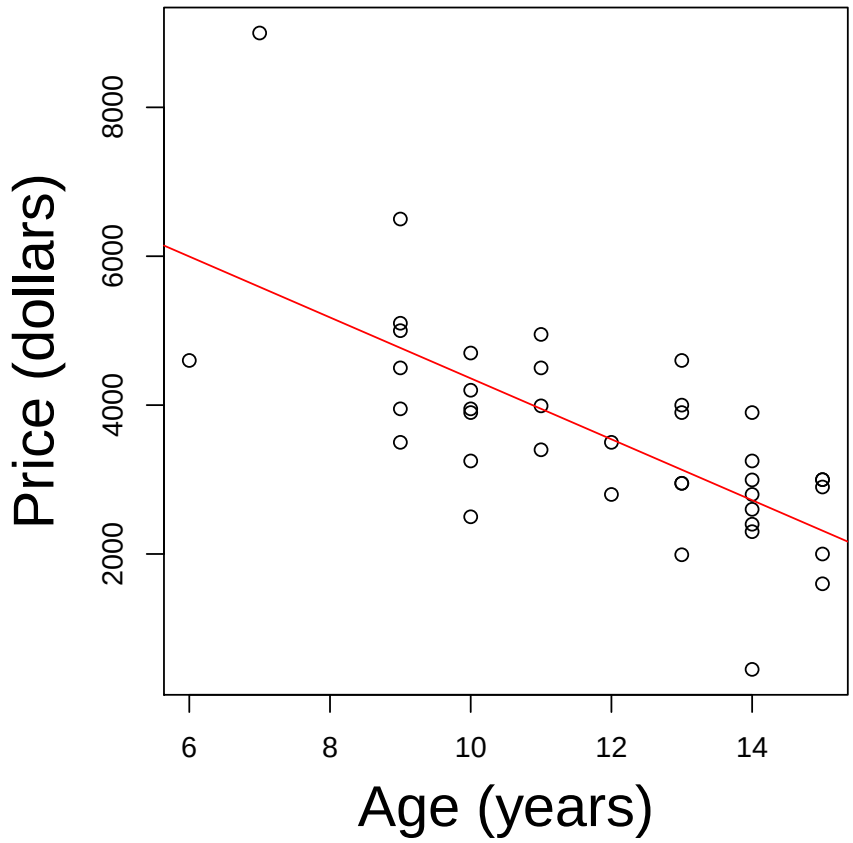
Mitsubishi Example

- A finer strip leads to a regression curve: the mean price for each age.
- We could use the prices of 10-year-old cars to create a new prediction interval.
- However, there are only 5 cars aged 10, so accuracy is limited.
- For other ages, the data is even sparser:
  - Only 2 cars are 12 years old.
  - No cars are 8 years old in the sample.
- How can we use this sparse data to predict prices for these ages?

Mitsubishi Example

- To predict prices, we model the "regression curve".
- The plot suggests the curve is roughly "linear".
- We can assume a "simple linear regression model" to predict price based on age.
- The fitted model is:

$$\hat{\text{price}} = 8452 - 409\ \textbf{age}$$


Price (dollars)
2000
4000
6000
8000
6
8
10
12
14
Age (years)

Model formulation

- In the Mitsubishi example, we say that **age** is a **predictor** for **price**. The variable **age** is called the **predictor variable** and **price** is the **response variable**.

- The usual generic notation for simple regression is

$$x = \text{ predictor variable}$$

$$y = \text{ response variable}$$

- The prediction equation is

$$\hat{y} = \beta_0 + \beta_1 x$$

Model formulation

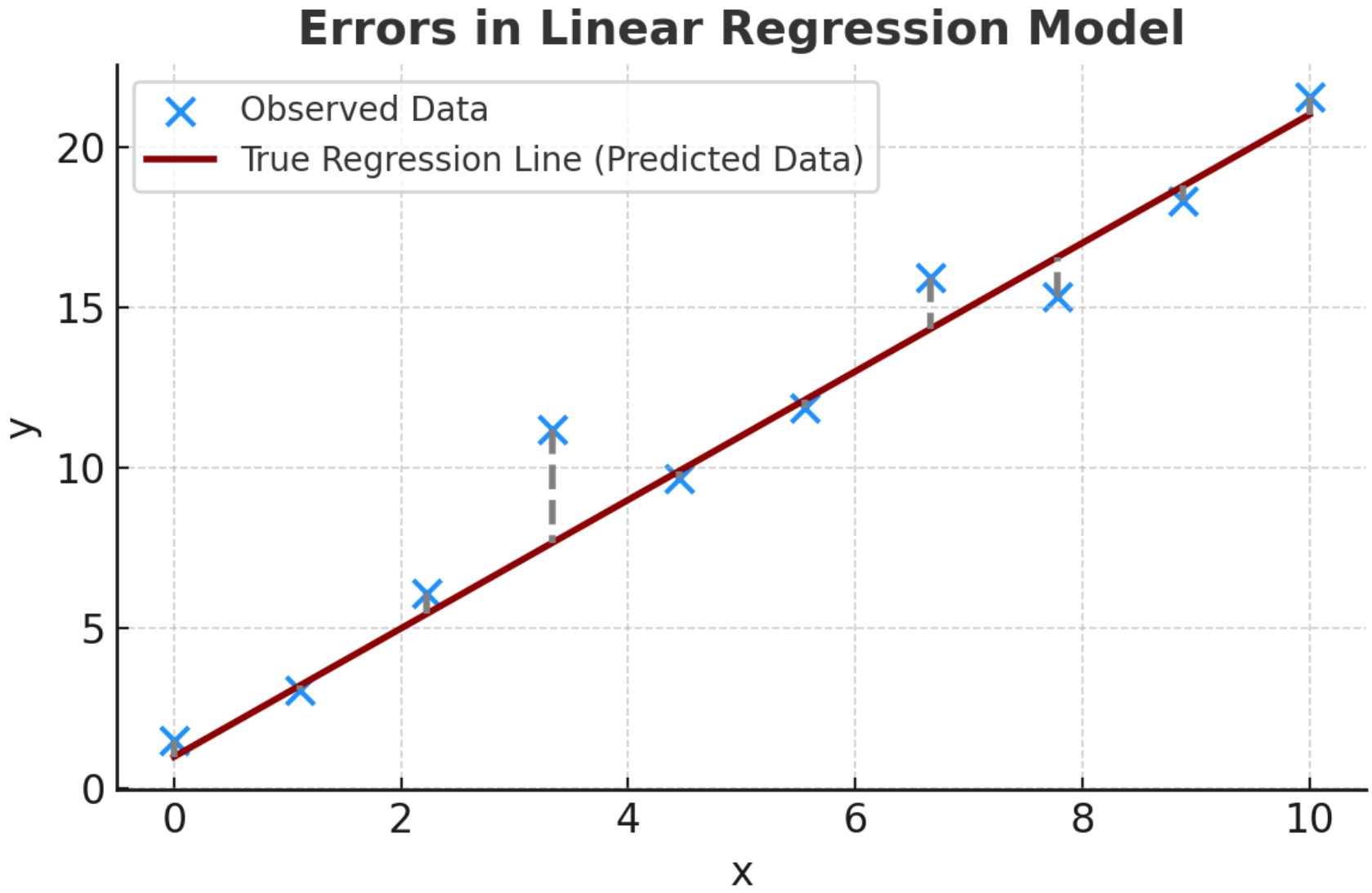
- Therefore, if y is any particular observation with corresponding x-value equal to x then we can write

$$y = \beta_0 + \beta_1 x + \epsilon$$

where the *error* ϵ is given by

$$\epsilon = y - \hat{y} = y - (\beta_0 + \beta_1 x)$$

- The **error term** (ϵ) represents the difference between what we predict ($\hat{y}$) and what we actually observe (y).


Errors in Linear Regression Model
Observed Data
True Regression Line (Predicted Data)
y
x
0
5
10
15
20
0
2
4
6
8
10

Understanding the Error Terms

- The first assumption we make is:

$$E(\epsilon) = 0$$

  This means that on average, the errors cancel out. Sometimes we predict too high, sometimes too low, but overall the average error is zero.

- Finally, we assume the errors follow a **normal distribution**:

$$\epsilon \sim N(0, \sigma^2)$$

  This means that most errors are small (close to zero), and bigger errors are less likely. The error distribution is symmetric around 0, with the spread determined by σ^2.

- The **variance of the error** tells us how spread out the errors are:

$$Var(\epsilon) = \sigma^2$$

This tells us that the errors vary by a certain amount (denoted as σ^2). It gives an idea of how far off our predictions could be from reality.

What is a Normal Distribution?

- The **normal distribution** is a bell-shaped curve where most values (in our case, errors) are close to the center (0), and fewer values are far away from the center.

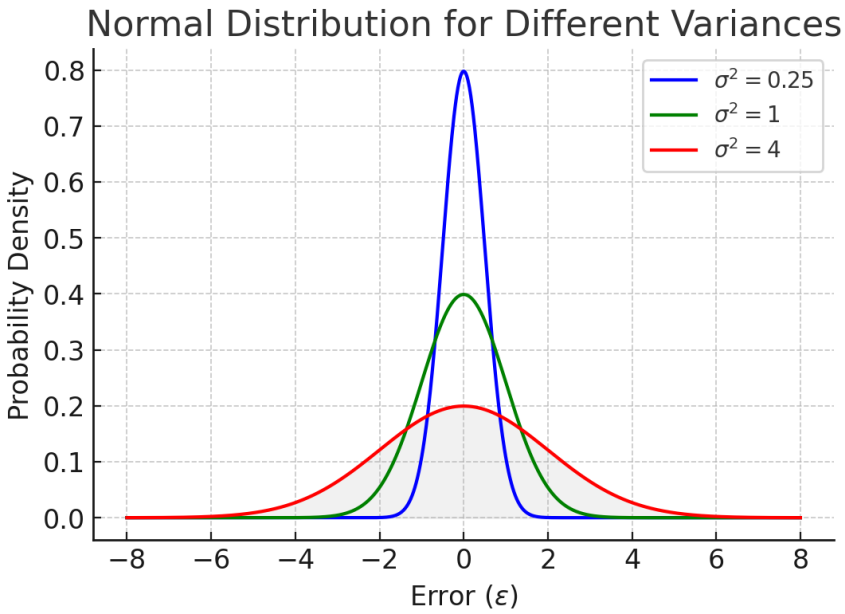
- This curve is symmetric, meaning that errors are equally likely to be positive or negative.

- For our model:

$$\epsilon \sim N(0, \sigma^2)$$

  This means the average error is 0, and the spread (how far errors go from 0) is determined by σ^2.

- Here is what the normal distribution looks like:



- Most errors are close to zero, but sometimes we can have larger errors, though these are less likely.

Model formulation

- Thus we can write the simple linear regression model as

$$y = \beta_0 + \beta_1 x + \epsilon, \quad \epsilon \sim N(0, \sigma^2).$$

- Suppose we have n observations, $(x_1, y_1), \ldots, (x_n, y_n)$ from this model, then

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, i = 1, \ldots, n$$

- We assume further that the observations are collected independently of each other so that the ϵ_i are independent. This means that a knowledge of one of the e_i does not tell us anything about the value of the other e_is

Mitsubishi Example:

For the Mitsubishi price/age data, using the R commands

```
> mitsub.lm <- lm(price ~ age, data = mitsub)
> summary(mitsub.lm)
```

gives that the estimates of the coefficients are

$$\hat{\beta}_0 = 8451.591 \quad \hat{\beta}_1 = -409.2175$$

and the residual variance is

$$\hat{\sigma}^2 = 1045^2$$

The R output is

```
> summary(lm(price ~ age))

Call:
lm(formula = price ~ age)

Residuals:
    Min      1Q  Median      3Q     Max
-2272.5  -504.8  -122.5   568.2  3411.9

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  8451.59     840.05  10.061 3.89e-12 ***
age          -409.22      69.79  -5.863 9.62e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1045 on 37 degrees of freedom
Multiple R-Squared: 0.4816,     Adjusted R-squared: 0.4676
F-statistic: 34.38 on 1 and 37 DF,  p-value: 9.618e-07
```

Understanding Linear Regression Results

- **Goal:** We are trying to predict the price of something based on its age. The output shows the results of this prediction.

- **Intercept (8451.59):** This means that if the age is 0, the predicted price is $8451.59. It's the starting point of our prediction line.

- **Slope (-409.22):** For every 1 year increase in age, the price decreases by $409.22. This tells us that older things are worth less.

- **R-Squared (0.4816):** About 48% of the price changes can be explained by the age. This is how well our model fits the data.

- **p-value (9.62e-07):** This number tells us if age is important for predicting the price. Since it's very small, we can say age is a very important factor.

Mitsubishi Example:

Hence the fitted line is

$$\hat{\text{price}} = 8451.59 - 409.22 \text{ age}.$$

The estimated price of new Mitsubishi Sigma cars (**age =0**) is \$8451.59

The estimated depreciation rate of Mitsubishi Sigma cars is \$409.22 per year

Mitsubishi Example:

The standard error of the intercept, $s_{\hat{\beta}_0} = 840.05$. A 95% confidence interval for the intercept, β_0 based on the t distribution with 37 df is

$$\hat{\beta}_0 \pm t_{37}(0.025) s_{\hat{\beta}_0} \Longrightarrow (6748.897, 10153.10).$$

Similarly, a 95% confidence interval for the slope β_1, is

$$\hat{\beta}_1 \pm t_{37}(0.025) s_{\hat{\beta}_1} \Longrightarrow (-550.628, -267.8120).$$

To test the null hypothesis $H_0 : \beta_0 = 0$, we would use the t statistic $\hat{\beta}_0 / s_{\hat{\beta}_0} = 10.061$. The hypothesis would be rejected at significance level $\alpha = 0.05$, there is strong evidence that the intercept is non-zero.