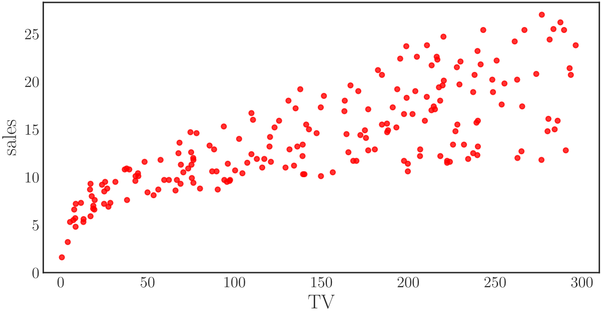
Linear Regression

229351

Example: Monthly sales data

Example: X = TV advertising budgets
Y = sales of a product



Linear Regression

- Quantitative response Y.
- Predictor variable X.

Goal: Study a linear relationship between X and Y:

$$Y \approx \beta_0 + \beta_1 X.$$

Linear Regression

- Quantitative response Y.

- Predictor variable X.

The statistical model is:

$$Y = \beta_0 + \beta_1 X + \epsilon, \qquad \epsilon \sim N(0, \sigma^2)$$

Example: $X =$ TV advertising budgets

$Y =$ sales of a product

$$sales = \beta_0 + \beta_1 \times TV + \epsilon, \qquad \epsilon \sim N(0, \sigma^2).$$

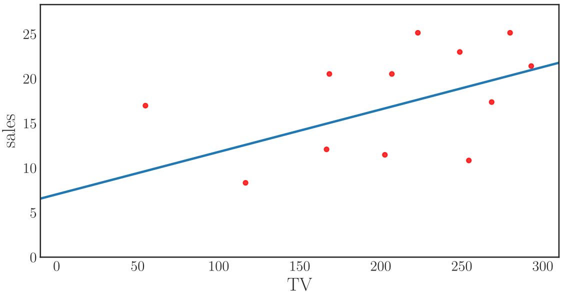
Example: $X =$ TV advertising budgets
$Y =$ sales of a product

$$sales = \beta_0 + \beta_1 \times TV + \epsilon, \qquad \epsilon \sim N(0, \sigma^2).$$

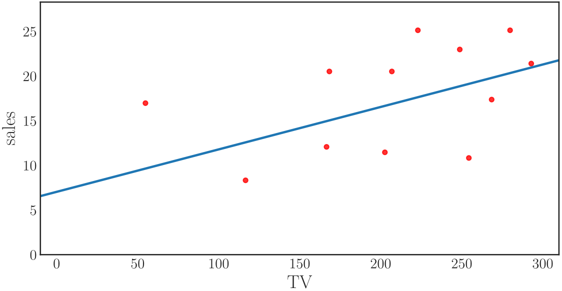
Since we do not have all possible $sales$ and TV ...

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x,$$

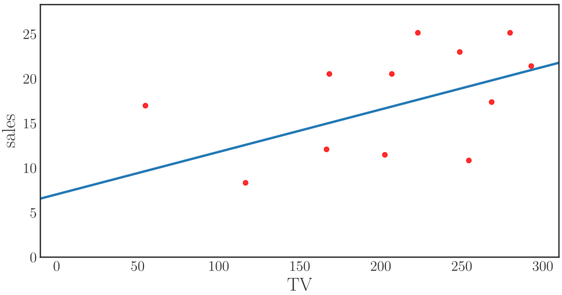
where $x =$ an observed value
$\hat{y} =$ prediction.



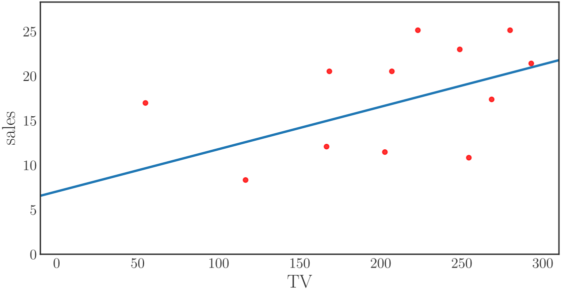
- Data: $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$



- Data: $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$
- Predictions: $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$



- Data: $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$
- Predictions: $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$
- Errors: $e_i = |y_i - \hat{y}_i|$

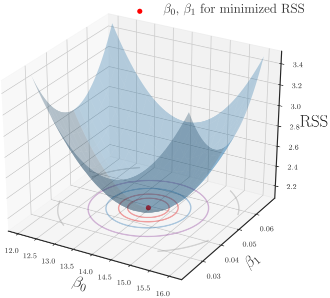


We want to minimize the *residual sum of squares*

$$\begin{aligned}\text{RSS} &= e_1^2 + e_2^2 + \ldots + e_n^2 \\ &= (y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1)^2 + \ldots + (y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n)^2.\end{aligned}$$

Residual Sum of Squares (RSS)

$$\text{RSS} = \underbrace{(y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1)^2 + (y_2 - \hat{\beta}_0 - \hat{\beta}_1 x_2)^2 + \ldots + (y_n - \hat{\beta}_0 - \hat{\beta}_1}_{\text{function of } \hat{\beta}_0, \hat{\beta}_1}$$



Least square coefficient estimate

Find $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimize

$F(\hat{\beta}_0, \hat{\beta}_1) = (y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1)^2 + (y_2 - \hat{\beta}_0 - \hat{\beta}_1 x_2)^2 + \ldots + (y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n)^2$

The solution is

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

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

Least square coefficient estimate

Find $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimize

$F(\hat{\beta}_0, \hat{\beta}_1) = (y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1)^2 + (y_2 - \hat{\beta}_0 - \hat{\beta}_1 x_2)^2 + \ldots + (y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n)^2$

The solution is

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

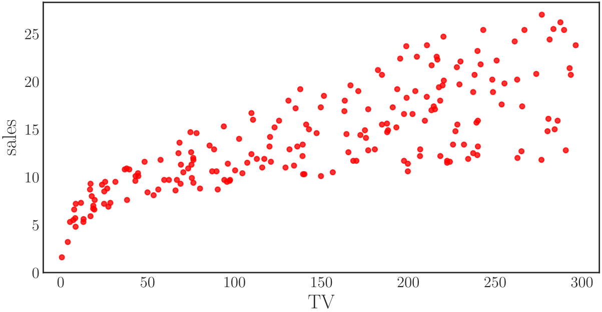
Sketch of derivation: take the partial derivatives of F with respect to $\hat{\beta}_0$ and $\hat{\beta}_1$

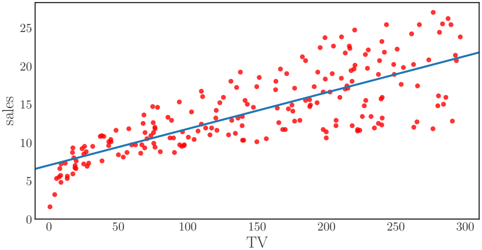
$$\frac{dF}{d\hat{\beta}_0} = -2 \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$
$$\frac{dF}{d\hat{\beta}_1} = -2 \sum_{i=1}^{n} x_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0,$$

Then solve for $\hat{\beta}_0$ and $\hat{\beta}_1$.

Example: Monthly sales data

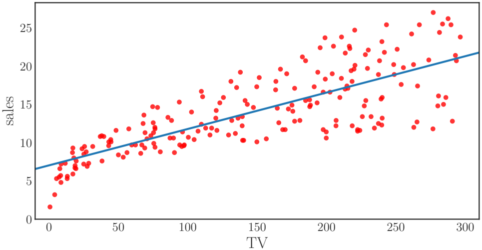
Example: $X =$ TV advertising budgets
$Y =$ sales of a product





$$\hat{\beta}_0 = 7.03, \quad \hat{\beta}_1 = 0.0475.$$

Interpreting $\hat{\beta}_0$: Without any TV advertising, the comp
7.03 units in sales **on average**.



$$\hat{\beta}_0 = 7.03, \quad \hat{\beta}_1 = 0.0475.$$

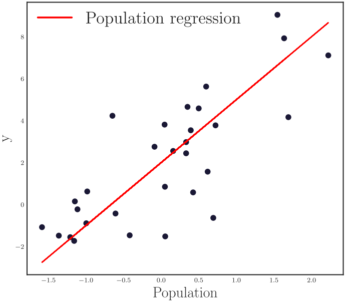
Interpreting $\hat{\beta}_1$: An additional \$100 spent on TV adver
is associated with 4.75 **more** units in sales.

Accuracy of $\hat{\beta}_0$ and $\hat{\beta}_1$

$$\text{Population model: } Y = \beta_0 + \beta_1 X + \epsilon$$

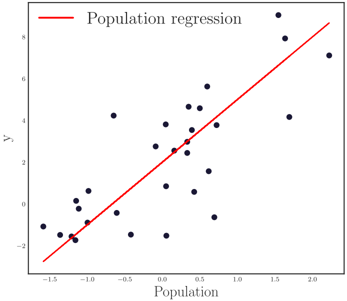
$$\text{Sample model: } Y = \hat{\beta}_0 + \hat{\beta}_1 X,$$

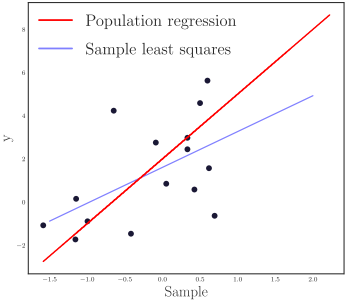
- $\hat{\beta}_0$ and $\hat{\beta}_1$ were computed from a *sample*, not a *population*.
- Can we tell anything about β_0 and β_1 from $\hat{\beta}_0$ and $\hat{\beta}_1$?



y

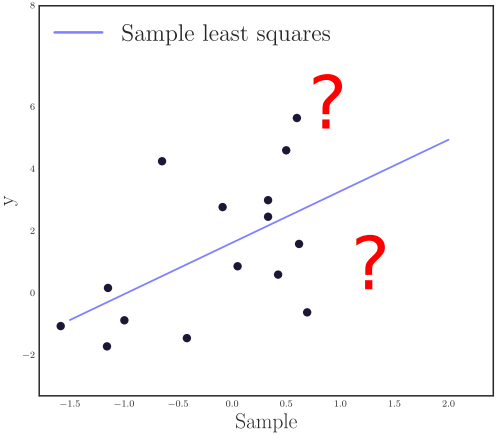
- 30 generated points from $Y = 2 + 3X + \epsilon$ where $\epsilon \sim N(0, 2)$.





- The blue line is the *least square* line of the population.The red line is the population regression line: $Y = 2 + 3X$

- The blue line is the *least square* line of the sample.



How can we locate the **population regression**

Confidence interval

We find the location of β_0's by making **confidence intervals**:

$$I_0 = [\hat{\beta}_0 - 2 \cdot \mathrm{SE}(\hat{\beta}_0), \hat{\beta}_0 + 2 \cdot \mathrm{SE}(\hat{\beta}_0)],$$

where SE is the **standard error** (next two slides)

This interval has 95% chance of containing β_0

Confidence interval

We find the location of β_1's by making **confidence intervals**:

$$I_1 = [\hat{\beta}_1 - 2 \cdot \mathsf{SE}(\hat{\beta}_1), \hat{\beta}_1 + 2 \cdot \mathsf{SE}(\hat{\beta}_1)],$$

where SE is the **standard error** (next slide)

This interval has 95% chance of containing β_1

Standard errors

$$I_i = [\hat{\beta}_i - 2 \cdot \mathsf{SE}(\hat{\beta}_i), \hat{\beta}_i + 2 \cdot \mathsf{SE}(\hat{\beta}_i)], \quad i = 0, 1$$

$$\mathsf{SE}(\hat{\beta}_0)^2 = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2} \right]$$

$$\mathsf{SE}(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}.$$

There is **95%** probability that I_i contains β_i.

Residual standard error

However, most of the time we don't know σ!

Replace σ^2 by the *residual standard error* (RSE)

$$\text{RSE} = \sqrt{\frac{\text{RSS}}{n-2}} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2}},$$

which satisfies $\mathbf{E}(\text{RSE}^2) = \sigma^2$.

Estimates of standard errors

$$I_i = [\hat{\beta}_i - 2 \cdot \widehat{\mathsf{SE}}(\hat{\beta}_i), \hat{\beta}_i + 2 \cdot \widehat{\mathsf{SE}}(\hat{\beta}_i)], \quad i = 0, 1$$

$$\widehat{\mathsf{SE}}(\hat{\beta}_0)^2 = \mathsf{RSE}^2 \left[\frac{1}{n} + \frac{\bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2} \right]$$

$$\widehat{\mathsf{SE}}(\hat{\beta}_1)^2 = \frac{\mathsf{RSE}^2}{\sum_{i=1}^n (x_i - \bar{x})^2}.$$

There is **95%** probability that I_i contains β_i.

Sales vs TV ads regression

The 95% confidence interval of β_0 is

$$I_0 = [6.135, 7.935]$$

What this means is that

- Without any advertising, the sales will fall somewhere between 6.135 and 7.935 units.

Sales vs TV ads regression

The 95% confidence interval of β_1 is

$$I_1 = [0.042, 0.053]$$

What this means is that

- For each $\$1$ additional TV advertising, there will be an increase in sale between 0.042 and 0.053 units on average.

Hypothesis testing

Main question: Is there **actual** relationship between X and Y?

Hypothesis test

$$\begin{aligned} H_0 &: \beta_1 = 0 \qquad \text{(no relationship)} \\ H_1 &: \beta_1 \neq 0 \qquad \text{(some relationship)} \end{aligned}$$

Then under some rule$(\hat{\beta}_1)$, we decide to *accept* or *reject* H_0.

Hypothesis test

$$H_0 : \beta_1 = 0 \qquad \text{(no relationship)}$$
$$H_1 : \beta_1 \neq 0 \qquad \text{(some relationship)}$$

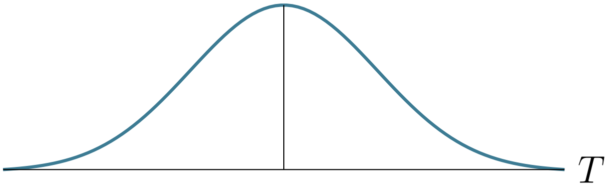
Then under some rule$(\hat{\beta}_1)$, we decide to *accept* or *reject* H_0.

How can we make a decision? Look at the *t-statistic*.

$$t = \frac{\hat{\beta}_1 - 0}{\mathrm{SE}(\hat{\beta}_1)}.$$

If $|t|$ is sufficiently large then we will reject H_0.

t-statistic

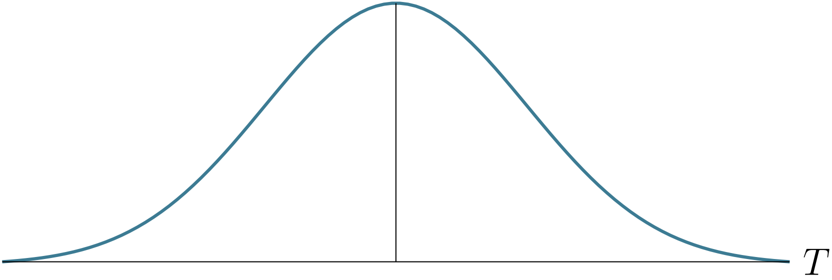


- p-value is the probability that $T > |t|$.

- If the p-value is too small, we will reject H_0.

- Typical p-value are 5% and 1% which corresponds to $|t| = 2$ and $|t| = 2.75$, respectively.

salse vs TV regression



	$\hat{\beta}_i$	SE($\hat{\beta}_i$)	t-statistic	p-value
Intercept	7.0325	0.4578	15.36	< 0.0001
TV	0.0475	0.0027	17.67	< 0.0001

Accuracy of the model

1. Residual standard error

$$\text{RSE} = \sqrt{\frac{\text{RSS}}{n-2}} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2}},$$

- In `sales` vs **TV** regression is, RSE $= 3.26$.

- Any prediction from the **true regression line** $Y = \beta_0 + \beta_1 X$ is off from the actual sales by $3,260$ units on average.

Accuracy of the model

2. R^2 statistic

$$R^2 = \frac{\text{TSS} - \text{RSS}}{\text{TSS}}$$

- where $\text{TSS} = \sum_{i=1}^n (y_i - \bar{y})^2$ is the *total sum of squares*.
  - TSS/n is the “variance” of Y.
- $\text{RSS} = \sum_{i=1}^n (y_i - \hat{y}_i)^2$
  - RSS/n is the “variance” not explained by the regression.

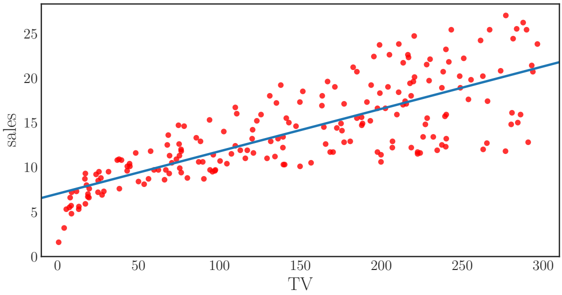
R^2 statistic

$$R^2 = \frac{\text{TSS} - \text{RSS}}{\text{TSS}}$$

R^2 statistic

$$R^2 = \frac{\text{TSS} - \text{RSS}}{\text{TSS}}$$

R^2 is the **proportion of variance of** y **explained by the regression**



$R^2 = 0.612$, so about two-thirds of the variance in Y is explained by a regression in **TV**.