Linear Regression 1
DS351

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.

Goal: Study a linear relationship between X and Y:

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

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

$$\mathit{sales} \approx \beta_0 + \beta_1 \times \mathit{TV}.$$

Assumption: There are β_0 and β_1 that works for all possible *sales* and TV.

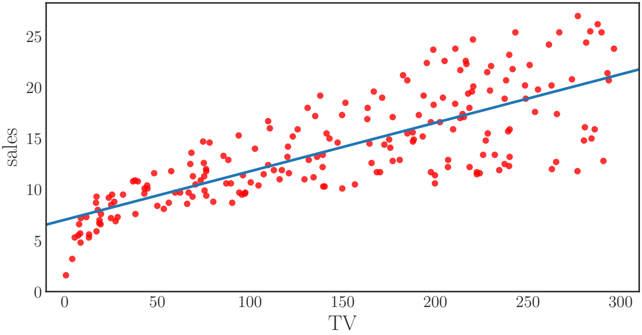
Since we do not have all possible *sales* and TV...

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

where $\hat{y}$ is a prediction of Y.

What's the difference between X and x?

$$X = \text{a random variable}$$
$$x = \text{an observed value.}$$



Let $e_i = y_i - \hat{y}_i$. We want to minimize the *residual sum of squares* (RSS)

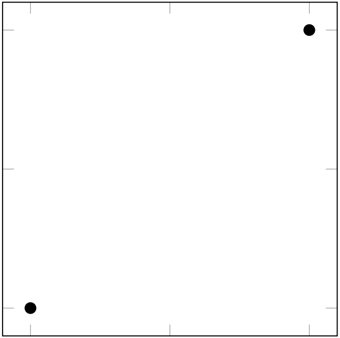
$$\text{RSS} = e_1^2 + e_2^2 + \ldots + e_n^2,$$

or equivalently

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

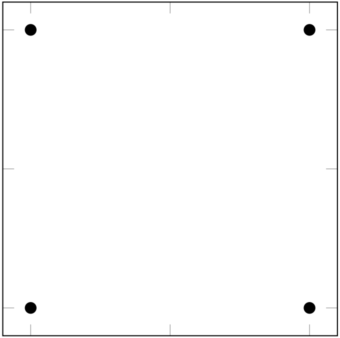
Why not minimize the *sum of absolute errors* (SAE) instead?

$$\text{SAE} = |e_1| + |e_2| + \ldots + |e_n|.$$



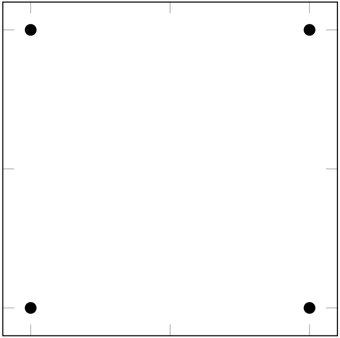
- SAE:
- SSR:

SAE vs RSS



- There are ____________ lines that minimize SAE.

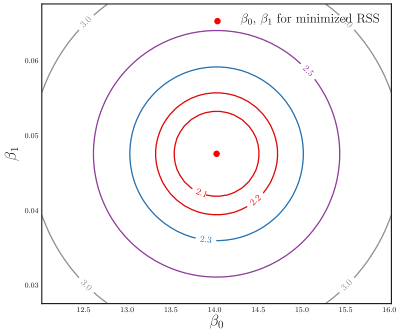
SAE vs RSS

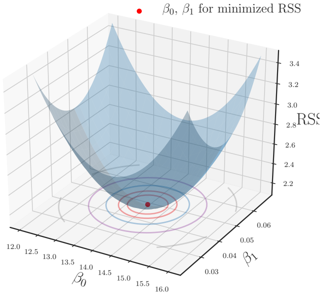


- There are ____________ lines that minimize SAE.
- There are ____________ lines that minimize RSS.

Back to RSS

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





Least square coefficient estimate

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

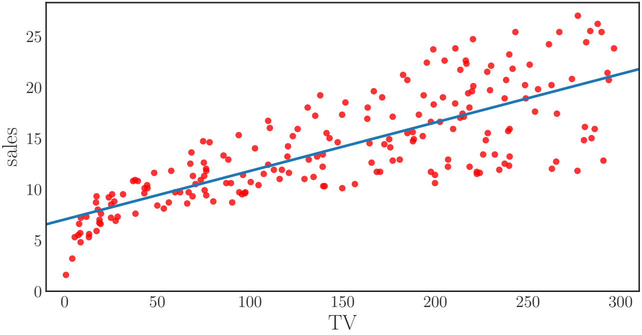
$$\mathrm{RSS} = (y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1)^2 + (y_2 - \hat{\beta}_0 - \hat{\beta}_1 x_2)^2 + \ldots + (y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n)^2.$$

The solution is

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

where

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n}$$
$$\bar{y} = \frac{y_1 + y_2 + \ldots + y_n}{n}$$



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

An additional $1,000 spent on TV advertising is associated to 47.5 more units in sales.

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

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

To be precise, this is

$$Y = \beta_0 + \beta_1 X + \epsilon,$$

where ϵ is a random variable with **zero mean** and **unknown variance** σ^2.

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

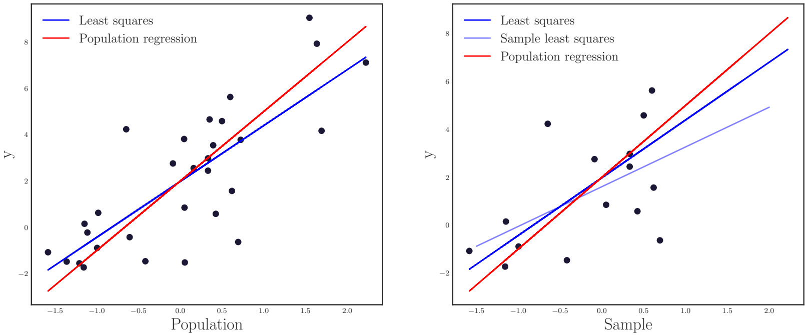
$$Y \approx \beta_0 + \beta_1 X$$

To be precise, this is

$$Y = \beta_0 + \beta_1 X + \epsilon,$$

where ϵ is a random variable with **zero mean** and **unknown variance** σ^2.

- $\hat{\beta}_0$ and $\hat{\beta}_1$ were computed from a *sample*, not a *population*.
- How close are $\hat{\beta}_0$ and $\hat{\beta}_1$ to β_0 and β_1?



- 30 generated points from $Y = 2 + 3X + \epsilon$ where $\epsilon \sim N(0, 2)$.
- The red line is the population regression line: $Y = 2 + 3X$
- The blue line is the *least square* line of the population.
- The light blue line is the *least square* line of the sample.

Confidence interval

We assess the "closeness" of $\hat{\beta}_i$'s to β_i's by making **confidence intervals**:

$$I_i = [\hat{\beta}_i - 2 \cdot \mathsf{SE}(\hat{\beta}_i), \hat{\beta}_i - 2 \cdot \mathsf{SE}(\hat{\beta}_i)],$$

Roughly speaking, $\mathsf{SE}(\hat{\beta}_i)$ tells us the distance between $\hat{\beta}_i$ and β_i **on average**.

Confidence interval

We assess the "closeness" of $\hat{\beta}_i$'s to β_i's by making **confidence intervals**:

$$I_i = [\hat{\beta}_i - 2 \cdot \mathrm{SE}(\hat{\beta}_i), \hat{\beta}_i - 2 \cdot \mathrm{SE}(\hat{\beta}_i)],$$

Roughly speaking, $\mathrm{SE}(\hat{\beta}_i)$ tells us the distance between $\hat{\beta}_i$ and β_i **on average**.

We say that $\hat{\beta}_i$ are *close* to β_i if this interval contains β_i.

Confidence interval

We assess the "closeness" of $\hat{\beta}_i$'s to β_i's by making **confidence intervals**:

$$I_i = [\hat{\beta}_i - 2 \cdot \mathrm{SE}(\hat{\beta}_i), \hat{\beta}_i - 2 \cdot \mathrm{SE}(\hat{\beta}_i)],$$

Roughly speaking, $\mathrm{SE}(\hat{\beta}_i)$ tells us the distance between $\hat{\beta}_i$ and β_i **on average**.

We say that $\hat{\beta}_i$ are *close* to β_i if this interval contains β_i.
With

$$\mathrm{SE}(\hat{\beta}_0)^2 = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right], \quad \mathrm{SE}(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2},$$

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

Residual standard error

$$\mathrm{SE}(\hat{\beta}_0)^2 = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2} \right], \quad \mathrm{SE}(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2},$$

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

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

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

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

salse vs TV regression

The 95% confidence interval of β_0 is

$$I_0 = [6.135, 7.935]$$

and the 95% confidence interval of β_1 is

$$I_1 = [0.042, 0.053]$$

What this means is that

- Without any advertising, the sales will fall somewhere between $6,130$ and $7,935$ units.
- For each $\$1,000$ additional TV advertising, there will be an increase in sale between 42 and 53 units on average.

Hypothesis test

- We want to know if there is an actual relationship between X and Y i.e. if $\beta_1 = 0$.
- However, $\hat{\beta}_1$ alone won't tell us if $\beta_1 = 0$.

Hypothesis test

- We want to know if there is an actual relationship between X and Y i.e. if $\beta_1 = 0$.
- However, $\hat{\beta}_1$ alone won't tell us if $\beta_1 = 0$.

Statistical way of making a decision: **hypothesis test**.

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

Since $\beta_1 = 0$ implies $Y = \beta_0 + \epsilon$ which means that Y does not depend on X.

Hypothesis test

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

If we decide to conclude that that H_0 is true, then we say that we *accept* H_0 and *reject* H_1.

If we decide to conclude that that H_1 is true, then we say that we *reject* H_0 and *accept* H_1.

Hypothesis test

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

If we decide to conclude that that H_0 is true, then we say that we *accept* H_0 and *reject* H_1.

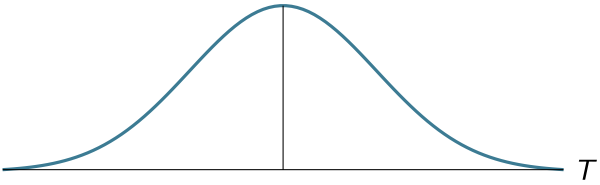
If we decide to conclude that that H_1 is true, then we say that we *reject* H_0 and *accept* H_1.

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

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

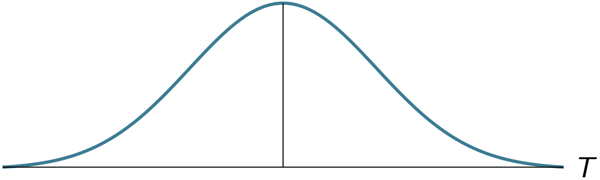
If $|t|$ is sufficiently large then we will reject H_0.

t-statistic



- p-value is the probability that $T > |t|$.
- If the p-value is too large, we will reject H_0.
- Typical p-value are 5% and 1% which corresponds to $|t| = 2$ and $|t| = 2.75$, respectively.

t-statistic



- p-value is the probability that $T > |t|$.
- If the p-value is too large, we will reject H_0.
- Typical p-value are 5% and 1% which corresponds to $|t| = 2$ and $|t| = 2.75$, respectively.

	$\hat{\beta}_i$	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

We conclude that $\beta_0 \neq 0$ and $\beta_1 \neq 0$.

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, $\text{RSE} = 3.26$.
- 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*. This is the variance of Y.
- $\text{RSS} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ is the variance that is not explained by the regression.
- Thus, R^2 is the **proportion of variance that is explained by the regression**

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*. This is the variance of Y.
- $\text{RSS} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ is the variance that is not explained by the regression.
- Thus, R^2 is the **proportion of variance that is explained by the regression**
- In `sales` vs `TV` regression, $R^2 = 0.612$, so about two-thirds of the variance in Y is explained by a regression in `TV`.