

Linear Regression

Igor Adamiec

Important questions (based on Advertising data)

Is there a relationship between predictor and response?

- Advertising budget and sales

How strong is relationship?

- between advertising budget and sales

Which predictors contribute to the sales?

- All of them or only some?

How accurately can we estimate effect of each predictor on response?

How accurately can we predict response?

Is the relationship linear?

Is there synergy (interaction) among predictors?

Simple linear regression

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

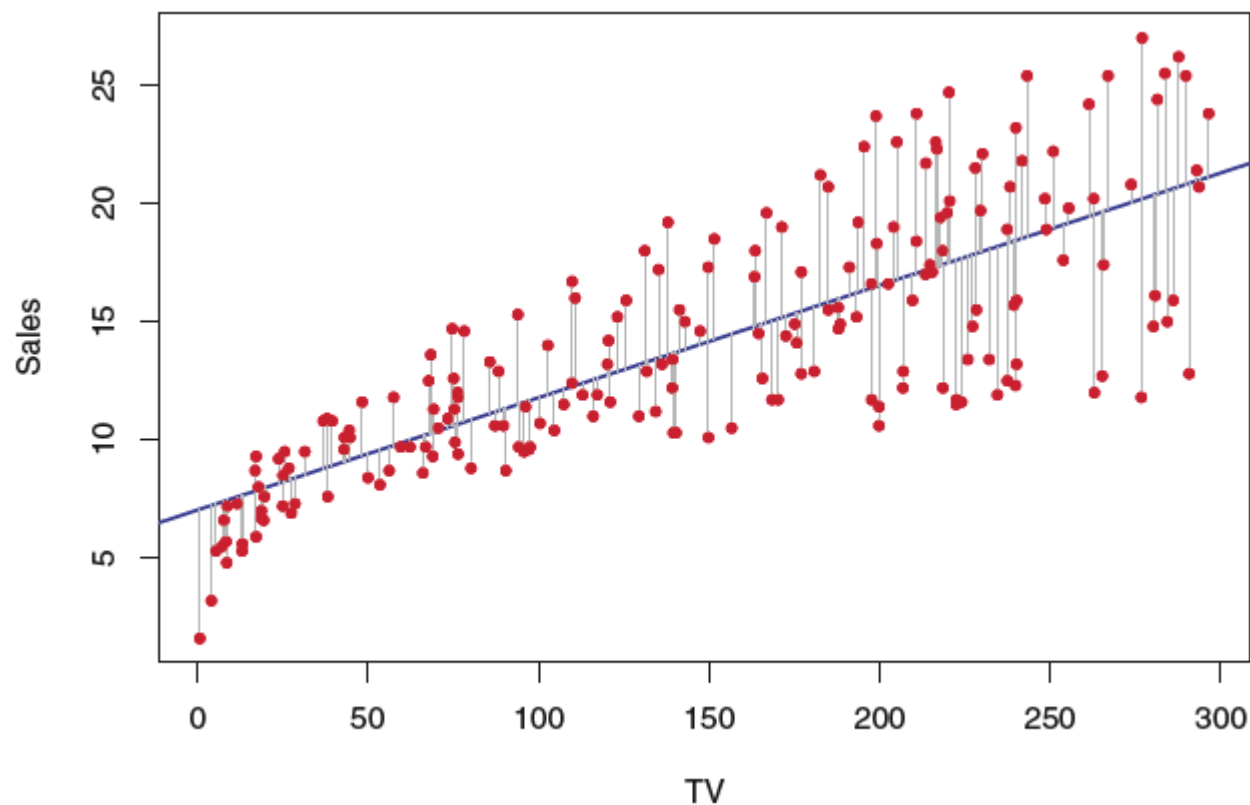
$$\text{sales} \approx \beta_0 + \beta_1 \times \text{TV}$$

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

Coefficients/parameters

intercept

slope



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

$$e_i = y_i - \hat{y}_i$$

Residual (reszta)

$$RSS = e_1^2 + e_2^2 + \dots + e_n^2$$

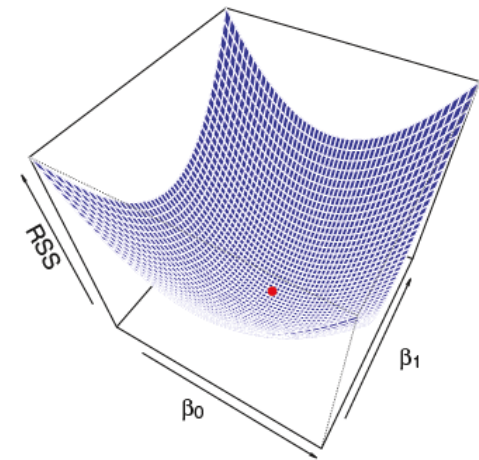
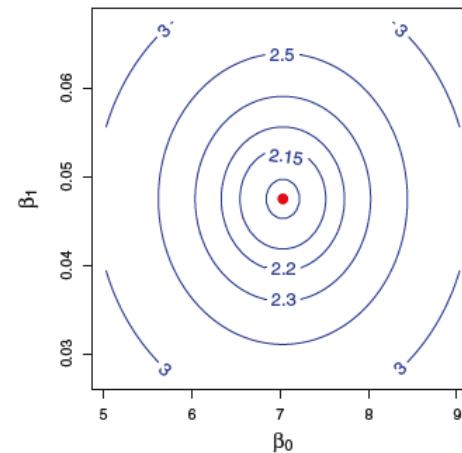
Residual sum of squares – musi być jak najmniejszy

Least squares approach

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

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

Wzory tylko dla pojedynczej regresji (trzeba się dowiedzieć czy istnieją wzory dla multi regresji)



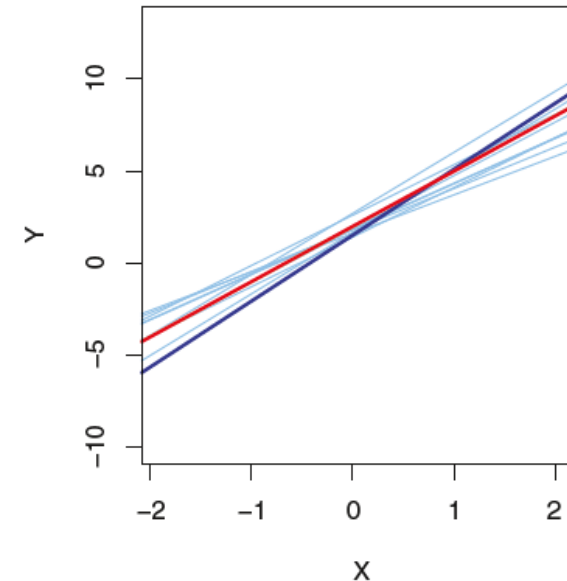
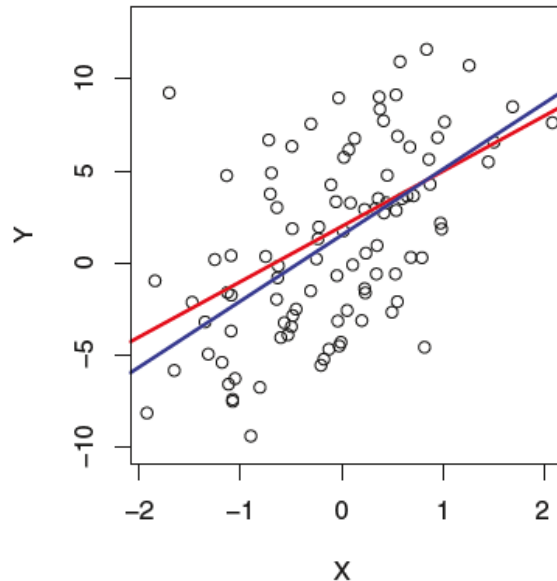
Accuracy of Coefficient estimates

Zakładamy, że relacja pomiędzy zmiennymi jest liniowa – jeżeli nie jest to prawdą, to cały model będzie bez sensu.

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

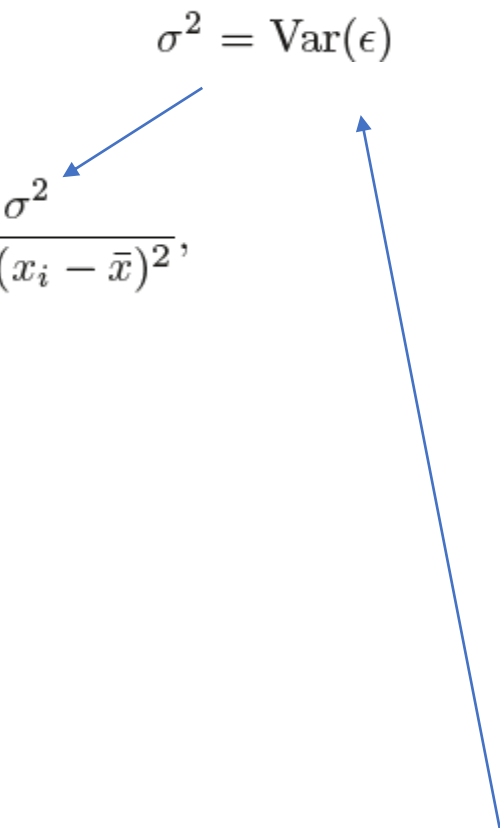
Czerwone linie, to linie regresji populacji (prawdziwa zależność, której nie znamy – $-2 + 3x$), granatowa linia, to regresja liniowa najmniejszych kwadratów. Po prawej to samo, a jasne niebieskie linie, to regresje liniowe z wycinków danych.

Różnica pomiędzy liniami jest spowodowana klasycznym problemem – populacja, a próbka.



Jeżeli zrobiliśmy regresję na wielu różnych wycinkach, to średnia parametrów powinna być zbliżona do rzeczywistych.

Jak daleko nasze estymatory są dalekie od rzeczywistości?

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


Standard error (tylko dla pojedynczej regresji; są wzory na multi?)

Im więcej obserwacji, tym mniejsza wartość błędu

$$\text{RSE} = \sqrt{\text{RSS}/(n-2)}$$

Wariancję błędu można
wyestymować przy pomocy
RSE – residual standard error

Po co jest SE?

$$\hat{\beta}_1 \pm 2 \cdot \text{SE}(\hat{\beta}_1) \longleftrightarrow [\hat{\beta}_1 - 2 \cdot \text{SE}(\hat{\beta}_1), \hat{\beta}_1 + 2 \cdot \text{SE}(\hat{\beta}_1)]$$

Ten zakres da nam 95% procentowe prawdopodobieństwo, że znajdziemy w nim prawdziwą wartość β_1 . Jest to confidence interval (przedział ufności).

[6.130, 7.935]

	Coefficient	Std. error	t-statistic	p-value
Intercept	7.0325	0.4578	15.36	< 0.0001
TV	0.0475	0.0027	17.67	< 0.0001

[0.042, 0.053]

SE służy też do testowania hipotez o tym czy istnieje relacja pomiędzy X i Y .

--

H_0 : There is no relationship between X and Y

versus the *alternative hypothesis*

H_a : There is some relationship between X and Y .

Mathematically, this corresponds to testing

$$H_0 : \beta_1 = 0$$

versus

$$H_a : \beta_1 \neq 0,$$

Testowanie hipotez

Testujemy, czy nasze β_1 jest wystarczająco różne od zera by uznać je za istotne (i odrzucić hipotezę zerową). Jeżeli SE jest stosunkowo małe, to zwiększy się prawdopodobieństwo na odrzucenie H_0 .

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

Używa się statystyki t, która mierzy ile odchyleń standardowych dzieli nasze estymowane β_1 od 0.

Na podstawie statystyki t, wyliczana jest wartość p-value – mała wartość wskazuje, że prawdopodobieństwo wystąpienia danej wartości y dla danego x jest bardzo małe, jeżeli nie ma relacji między nimi.

Na nasze: Jeżeli p-value jest małe, to znaczy, że relacja istnieje i możemy odrzucić H_0 .

	Coefficient	Std. error	t-statistic	p-value
Intercept	7.0325	0.4578	15.36	< 0.0001
TV	0.0475	0.0027	17.67	< 0.0001

Dokładność modelu - RSE

RSE – residual standard error.

Jest estymowanym odchyleniem standardowym błędu

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

Dla wartości modelu oznacza to, że rzeczywiste salesy będą się różniły o średnio 3,26 jednostek. Nawet jakbyśmy znali prawdziwą linię regresji populacji, to rzeczywiste wartości średnio by się różniły o tę wartość. Wynika to z błędu nieredukowalnego.

Średnia wartość zmiennej sales wynosi 14000 więc procentowy błąd liczymy
 $3,26/14000 = 23\%$

Quantity	Value
Residual standard error	3.26
R^2	0.612
F-statistic	312.1

RSE jest uznawane za brak dopasowania modelu do danych

Najważniejsze: im RSE bliższe 0 tym lepiej!!!

Dokładność modelu – R^2

R^2 jest proporcją wariancji, w której Y może zostać wyjaśniona przy użyciu X.
Przyjmuje wartości od 0 do 1.

Im bliższe 1, tym lepiej

$$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS}$$

$$TSS = \sum (y_i - \bar{y})^2$$

total sum of squares

R – korelacja (wartości od -1 do 1)

TSS – jak wartości rzeczywiste odbiegają od średniej,

RSS – jak wartości rzeczywiste odbiegają od naszej linii regresji

<https://www.youtube.com/watch?v=2AQKmw14mHM>