Simple Linear Regression

January 18, 2024

Problem

What is relationship between

- the tar content in the outlet stream in a chemical process is and the inlet temperature
- gas mileage and engine volume
- house price and square footage of living space

- inlet temperature, engine volume, square feet of living space ... are **independent variable (or regressor)**, x
- Tar content, gas mileage, house price ... are **dependent variable (or response)**, Y

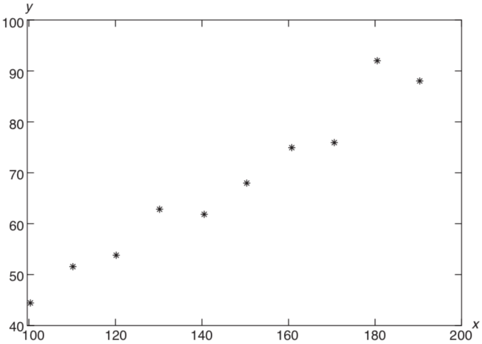
How to find out relationship between regressor and response

Data observation

i	x_i	y_i	i	x_i	y_i
1	100	45	6	150	68
2	110	52	7	160	75
3	120	54	8	170	76
4	130	63	9	180	92
5	140	62	10	190	88

y: the percent yield of a laboratory experiment
x: the temperature at which the experiment

Plotting



It seems that y is a linear function of x with some noise

Linear relationship

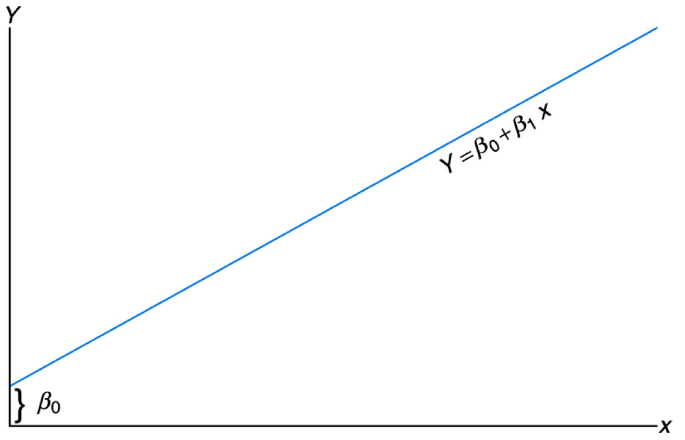


Figure: intercept β_0, slope β_1

However

- run several experiment with the same inlet temperature, tar content wil not be the same
- several automobiles with the same engine will not all have the same gas mileage.
- Houses with the same square footage are sold with different prices

Then

- Response Y is not a determisnistic function of regressor x

$$Y \neq f(x)$$

- But

$$Y = f(x) + \text{ noise}$$

Regression Analysis

- Find the best "fit" relationship between Y and x
- Qualify the strength of relationship
- Explain impact of x on Y
- Predict Y given some specific value of x

(Simple) Linear regression model

Dependent Variable

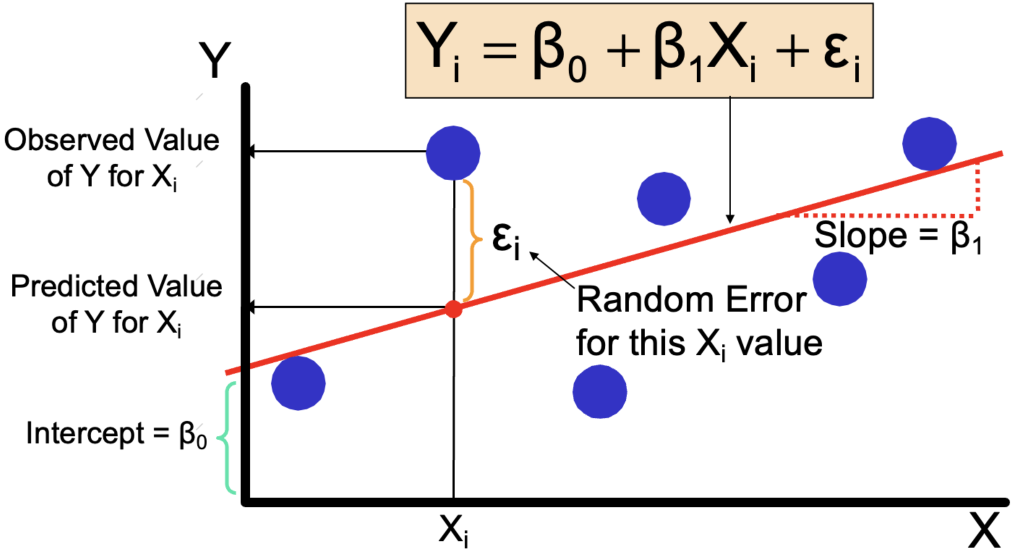
Population Y intercept

Population Slope Coefficient

Independent Variable

Random Error term

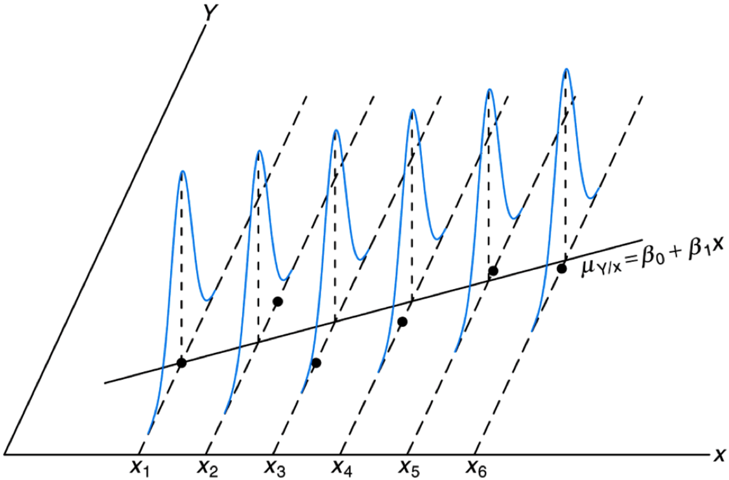
$$\boxed{Y_i = \underbrace{\beta_0 + \beta_1 X_i}_{\text{Linear component}} + \underbrace{\varepsilon_i}_{\text{Random Error component}}}$$


$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$
Y
Observed Value of Y for X_i
Predicted Value of Y for X_i
Intercept = β_0
ε_i
Random Error for this X_i value
Slope = β_1
X_i
X

Model assumption

- Error $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ are i.i.d
- Given x, response Y is normally distributed $\mathcal{N}(\beta_0 + \beta_1 x, \sigma^2)$
- True regression line $\mu_{Y|x} = \beta_0 + \beta_1 x$

The true regression line **go through the means of the response** but actually unknown



Fitted regression line

Estimated (or predicted) y value for observation i

Estimate of the regression intercept

Estimate of the regression slope

Value of x for observation i

$$\boxed{\hat{y}_i = b_0 + b_1 x_i}$$

One can use a fitted regression line to estimate predict or forecast y value given observaton x

Fitted regression line

Estimated (or predicted) y value for observation i

Estimate of the regression intercept

Estimate of the regression slope

Value of x for observation i

$$\hat{y}_i = b_0 + b_1 x_i$$

One can use a fitted regression line to estimate predict or forecast *y* value given observaton *x*

Least square and fitted model

Residual - error in fit

- Given
  - Data set $\{(x_i, y_i), i = 1, \ldots, n\}$
  - Fitted regression line

$$\hat{y}_i = b_0 + b_1 x_i$$

- Residual

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

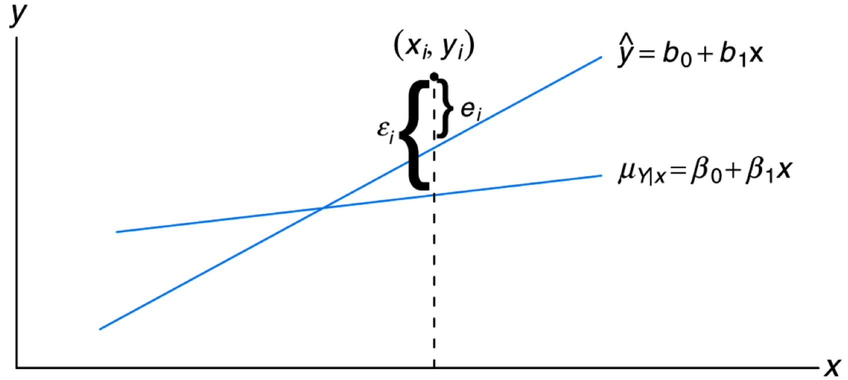
Important relationship

$$
\begin{aligned}
y_i &= b_0 + b_1 x_i + e_i \\
&= \hat{y}_i + e_i
\end{aligned}
$$

In word

actual value = fitted value + residual

Residual vs Error



Residual e_i is observed but error term ϵ_i is unobservable

Remark

- β_0, β_1 are unknown
- true regression line $\mu_{Y|x} = \beta_0 + \beta_1 x$ is then unknown
- **Need to estimate** β_0, β_1 from observed data

Least square method

- Sum of square of residual

$$SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y_i - b_0 - b_1 x_i)^2$$

- Minimize SSE to get estimates b_0, b_1 for β_0 and β_1
- Solve the optimization problem

$$\frac{\partial SSE}{\partial b_0} = 0; \quad \frac{\partial SSE}{\partial b_1} = 0$$

Least square estimators

- $b_1 = \dfrac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$
  or equivalent
$$b_1 = \frac{n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$
- $b_0 = \bar{Y} - b_1\bar{x}$
  where $\bar{y} = \sum_{i=1}^{n} y_i/n, \ \bar{x} = \sum_{i=1}^{n} x_i/n$

Better formula

$$b_1 = \frac{S_{xY}}{S_{xx}}, \qquad b_0 = \bar{Y} - B_1\bar{x}$$

where

$$S_{xY} = \sum_{i=1}^{n}(x_i - \bar{x})(Y_i - \bar{Y}) = \sum_{i=1}^{n} x_i Y_i - n\bar{x}\bar{Y}$$

$$S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2$$

Example

Estimate regression line for raw material data

Relative humidity	46	53	29	61	36	39	47	49	52	38	55	32	57	54	44
Moisture content	12	15	7	17	10	11	11	12	14	9	16	8	18	14	12

Solution

- Independent variable x: relative humidity
- Dependent variable y: moisture content
- $$n = 1, \quad \sum x_i = 692 \quad \sum y_i = 186$$
$$\sum x_i^2 = 33212 \quad \sum y_i^2 = 2454$$
$$\sum x_i y_i = 8997, \quad \bar{x} = 46.133 \quad \bar{y} = 12.4$$

We have

$$
\begin{aligned}
S_{xx} &= \sum x_i^2 - n\bar{x}^2 = 33212 - 15 \times 46.133^2 \\
&\approx 1287.73 \\
S_{YY} &= \sum y_i^2 - n\bar{y} = 2454 - 15 \times 12.4^2 = 147.6 \\
S_{xY} &= \sum x_i y_i - n\bar{x}\bar{y} = 8997 - 15 \times 46.13 \times 12.4 \\
&= 416.2
\end{aligned}
$$

So

$$b_1 = \frac{S_{xY}}{S_{xx}} \approx 0.32$$

and

$$b_0 = \bar{y} - b_1\bar{x} \approx 12.4 - 0.32 \times 46.13 = -2.51$$

Fitted line equation

$$\hat{y} = 0.32x - 2.51$$

Comment

- b_0: the estimated average value of Y when $x = 0$
- b_1 measures the estimated change in the average value of Y as a result of a one-unit change in x
  - $b_1 = 0.323$: the average value of moisture content increases by 0.323, on average, for each additional one relative humidity

Exercise

Compressive strength x and intrinsic permeability y are related according to a simple linear regression model. Summary quantities of a sample data are $n = 14$, $\sum y_i = 572$, $\sum y_i^2 = 23,530$, $\sum x_i = 43$, $\sum x_i^2 = 157.42$ and $\sum x_i y_i = 1697.80$.

1. Calculate the least squares estimates b_0 and b_1
2. Use the fitted line to predict permeability when the compressive strength $x = 4.3$
3. Suppose that the observed value of permeability at $x = 3.7$ is $y = 46.1$. Calculate the value of the corresponding residual.

Exercise

The following data are chloride concentration x (in milligrams per roadway area in the watershed y (in percentage)

x	4.4	6.6	9.7	10.6	10.8	10.9
y	0.19	0.15	0.57	0.70	0.67	0.63

Fit the linear regression model with least square method.

Linear regression with Excel

Input data $\rightarrow$ Choose Data $\rightarrow$ Data Analysis $\rightarrow$choose Regression and click Ok $\rightarrow$ select range for x and Y and click Ok

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	-2.5104577	1.31542339	-1.9084788	0.07865561	-5.3522571	0.33134181	-5.3522571	0.33134181
X Variable 1	0.32320356	0.02795527	11.5614542	3.2619E-08	0.26280988	0.38359725	0.26280988	0.38359725

Figure: Estimate parameter result in report

Practice

Estimate the regression line for pollution data

Solids Reduction, x (%)	Oxygen Demand Reduction, y (%)	Solids Reduction, x (%)	Oxygen Demand Reduction, y (%)
3	5	36	34
7	11	37	36
11	21	38	38
15	16	39	37
18	16	39	36
27	28	39	45
29	27	40	39
30	25	41	41
30	35	42	40
31	30	42	44
31	40	43	37
32	32	44	44
33	34	45	46
33	32	46	46

Properties of the Least Squares Estimators

Important remarks

- Estimate b_0, b_1 for β_0, β_1 depend on selected sample of observation
- Different experiments give different output with the same input x
- Estimates for β_0, β_1 from experiment to experiment
- Estimators are RVs B_0, B_1 while b_0, b_1 are specific realizations

Linear regression model

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

Model Assumption

Errors ϵ_i are i.i.d $\mathcal{N}(0, \sigma^2)$

Consequence

Given x_i, $Y_i \sim \mathcal{N}(\beta_0 + \beta_1 x_i, \sigma^2)$ and independent

Linear regression model

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

Model Assumption

Errors ϵ_i are i.i.d $\mathcal{N}(0, \sigma^2)$

Consequence

Given x_i, $Y_i \sim \mathcal{N}(\beta_0 + \beta_1 x_i, \sigma^2)$ and independent

Linear regression model

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

Model Assumption

Errors ϵ_i are i.i.d $\mathcal{N}(0, \sigma^2)$

Consequence

Given x_i, $Y_i \sim \mathcal{N}(\beta_0 + \beta_1 x_i, \sigma^2)$ and independent

Distribution of estimators

$$B_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x}) Y_i}{\sum_{i=1} x_i^2 - n(\bar{x})^2} \sim \mathcal{N}\left(\beta_1, \frac{\sigma^2}{\sum_{i=1} x_i^2 - n(\bar{x})^2}\right)$$

and

$$B_0 = \sum_{i=1}^{n} \frac{Y_i}{n} - B_1 \bar{x} \sim \mathcal{N}\left(\beta_0, \frac{\sigma^2 \sum_{i=1}^{n} x_i^2}{n\left(\sum_{i=1} x_i^2 - n(\bar{x})^2\right)}\right)$$

Unbiased estimator of σ^2 as mean square error

$$S^2 = \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n-2} \sim \chi^2(n-2)$$

$S = \sqrt{S^2}$ is called the **standard error**

where

- $(x_1, Y_1), \ldots, (x_n, Y_n)$ are observed data
- $\hat{Y}_i = B_0 + B_1 x_i$ is fitted value
- $n - 2$ is degree of freedom

Computational Identity for S^2

$$S^2 = \frac{S_{xx}S_{YY} - S_{xY}^2}{S_{xx}}$$

where

$$S_{xx} = \sum x_i^2 - n\bar{x}^2, \quad S_{xY} = \sum x_i Y_i - n\bar{x}\bar{Y}$$

$$S_{YY} = \sum (Y_i - \bar{Y})^2 = \sum Y_i^2 - n\bar{Y}^2$$

Inference about estimator B_1 relies on

Statistic

$$\frac{B_1 - \beta_1}{\frac{S}{\sqrt{S_{xx}}}} \sim T(n-2)$$

where $S_{xx} = \sum_{i=1}^{2}(x_i - \bar{x})^2$

$100(1-\alpha)\%$confidence interval for β_1

$$\boxed{b_1 - t_{\frac{\alpha}{2},n-2}\frac{s}{\sqrt{S_{xx}}} < \beta_1 < b_1 + t_{\frac{\alpha}{2},n-2}\frac{s}{\sqrt{S_{xx}}}}$$

Example

Relative humidity	46	53	29	61	36	39	47	49	52	38	55	32	57	54	44
Moisture content	12	15	7	17	10	11	11	12	14	9	16	8	18	14	12

Find a 95% confidence interval for β_1 in the regression line $\mu_{Y|x} = \beta_0 + \beta_1 x$

Solution

- $b_1 = 0.323$
- $n = 15, \bar{x} = 46.133, \sum_{i=1}^{n} x_i^2 = 33212$
- $S_{xx} = 1287.73, S_{YY} = 147.6, S_{xY} = 416.2$
- $s^2 = \frac{S_{xx}S_{YY} - S_{xY}^2}{S_{xx}} = 1.013$
- $s = \sqrt{1.013} = 1.006$

- $1 - \alpha = 95\% \Rightarrow t_{n-2,\alpha.2} = t_{13,0.025} = 2.16$
- $ME = t_{n-2,\alpha.2}\frac{s}{\sqrt{S_{xx}}} = 0.0606$
- Lower bound $b_1 - ME = 0.263$
- Upper bound $b_1 + ME = 0.384$
- 95% CI for β_1

$$0.263 < \beta_1 < 0.384$$

Hypothesis testing on the slope β_1

$$\boxed{\text{Test } H_0 : \beta_1 = \beta_{10} \text{ versus } H_1 : \beta_1 \neq \beta_{10}}$$

Test statistic (T-test)

$$T = \frac{B_1 - \beta_{10}}{\frac{S}{\sqrt{S_{xx}}}} \sim T(n-2)$$

About conclusion for testing $H_0 : \beta_1 = 0$ vs $H_1 : \beta_1 \neq 0$

- Failure to reject H_0 suggests that there is no linear relationship between Y and x. It may mean that changing x has little impact on changes in Y
- Reject H_0: there is an implication that the linear term in x residing in the model explains a significant portion of variability in Y

Example

Test $H_0 : \beta_1 = 0$ vs $H_1 : \beta_1 \neq 0$ at level of significance $\alpha = 5\%$

Relative humidity	46	53	29	61	36	39	47	49	52	38	55	32	57	54	44
Moisture content	12	15	7	17	10	11	11	12	14	9	16	8	18	14	12

Solution

- $b_1 = 0.323$
- $s = 1.006$, $s_{xx} = 1287.73$
- $t_{obs} = \frac{b_1 - \beta_{10}}{s/\sqrt{s_{xx}}} = \frac{0.323 - 0}{\frac{1.006}{\sqrt{1287.73}}} = 11.5$
- $t_{\alpha/2, n-2} = ?$
- Conclusion: is there is a significance on impact of relative humidity on moisture content in linear relationship at $\alpha = 5\%$?

Inference about estimator B_0 relies on

Statistics

$$\frac{B_0 - \beta_0}{S\sqrt{\frac{\sum_{i=1}^{n} x_i^2}{nS_{xx}}}} \sim T(n-2)$$

$100(1-\alpha)\%$ confidence interval for β_0

$$b_0 - ME < \beta_0 < b_0 + ME$$

where

$$ME = t_{\frac{\alpha}{2},n-2}\frac{s}{\sqrt{nS_{xx}}}\sqrt{\sum_{i=1}^{n} x_i^2}$$

Hypothesis testing about the intercept β_0

To test $H_0 : \beta_0 = \beta_{00}$ against a suitable alternative H_1, we use T-test with $n-2$ degrees of freedom to establish a critical value and make decision base on the value of

$$t_{obs} = \frac{b_0 - \beta_{00}}{s\sqrt{\frac{\sum_{i=1}^{n} x_i^2}{nS_{xx}}}}$$

A Measure of Quality of Fit: Coefficient of Determination

Coefficient of Determination

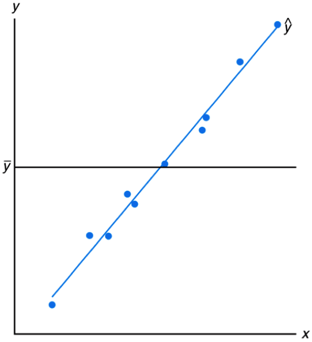
the proportion of variability explained by the fitted model

$$R^2 = 1 - \frac{SSE}{SSR}$$

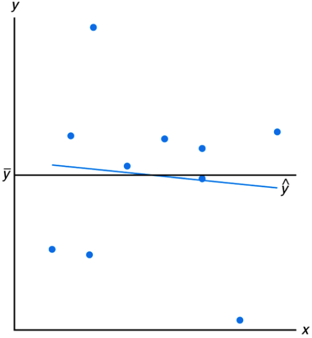
- $SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = (n-2)S^2$: sum of square error
- $SSR = \sum_{i=1}^{n}(y_i - \bar{y})^2 = S_{YY}$: sum of squares regression

Good fit vs Poor fit

$$0 \leq R^2 \leq 1$$



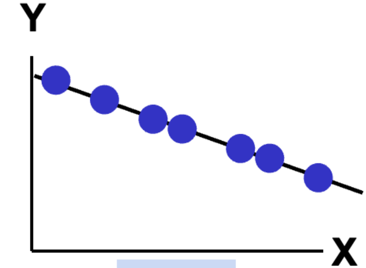
(a) $R^2 \approx 1.0$



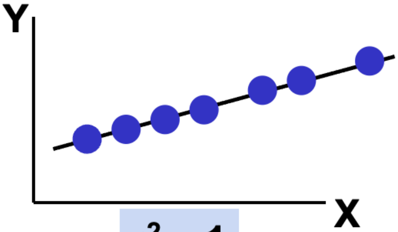
(b) $R^2 \approx 0$

R^2 as indicator

The value of R^2 is often used as an indicator of how well the regression model fits the data, with a value near 1 indicating a good fit, and one near 0 indicating a poor fit. In other words, if the regression model is able to explain most of the variation in the response data, then it is considered to fit the data well.



$r^2 = 1$

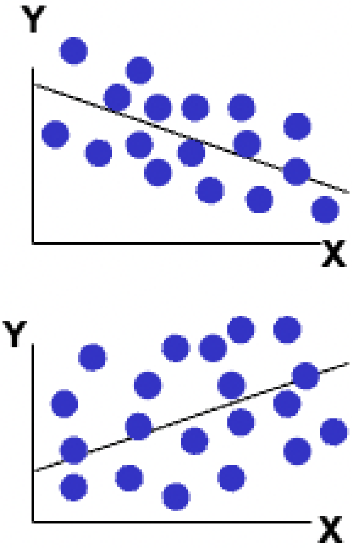


$r^2 = 1$

$r^2 = 1$

Perfect linear relationship between X and Y:

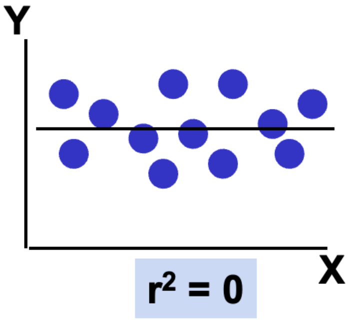
100% of the variation in Y is explained by variation in X



$0 < r^2 < 1$

Weaker linear relationships between X and Y:

Some but not all of the variation in Y is explained by variation in X



$r^2 = 0$

No linear relationship between X and Y:

The value of Y does not depend on X. (None of the variation in Y is explained by variation in X)

Compute R - square

Relative humidity	46	53	29	61	36	39	47	49	52	38	55	32	57	54	44
Moisture content	12	15	7	17	10	11	11	12	14	9	16	8	18	14	12

Solution

- Fitted regression line

$$\hat{y} = -2.51 + 0.323x$$

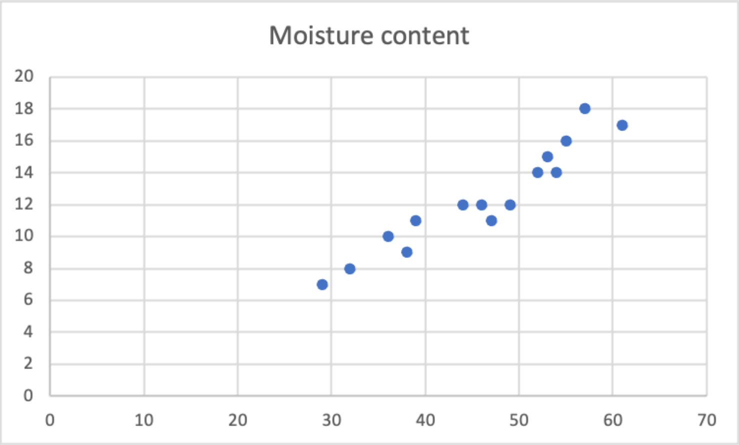
- $\bar{y} = 12.4$

- $SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = (n-2)S^2 = (15-2) \times 1.013 \approx 13.08$
- $SSR = \sum_{i=1}^{n}(y_i - \bar{y})^2 = S_{YY} = 147.6$
-

$$R^2 = 1 - \frac{SSE}{SSR} = 1 - \frac{13.08}{147.8} = 0.911$$

Comment

- The coefficient of determination suggests that the model fit to the data explains 91.1% of the variability observed in the response.
- $R^2 \approx 1$ indicates that linear model is a good fit model
- It is reasonable to use this model to estimate or predict moisture content given a value of relative humidity


Moisture content
20
18
16
14
12
10
8
6
4
2
0
0
10
20
30
40
50
60
70

Excel Report

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.95465385
R Square	0.91136397
Adjusted R S	0.90454582
Standard Err	1.00317487
Observations	15

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	134.517322	134.517322	133.667224	3.26188E-08
Residual	13	13.0826776	1.00635981		
Total	14	147.6			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-2.5104577	1.31542339	-1.9084788	0.07865561	-5.352257109	0.33134181	-5.3522571	0.33134181
Relative hum	0.32320356	0.02795527	11.5614542	3.2619E-08	0.262809875	0.38359725	0.26280988	0.38359725

Response and mean of response

Response given $x = x_0$

$$Y = \beta_0 + \beta_1 x_0 + \epsilon \sim \mathcal{N}(\beta_0 + \beta_1 x_0, \sigma^2)$$

Mean of response

$$E(Y|x = x_0) = \mu_{Y|x_0} = \beta_0 + \beta_1 x_0$$

Statistic for mean of response

Estimator for mean of response

$$\hat{Y}_0 = B_0 + B_1 x_0 \sim \mathcal{N}\left(\mu_{Y|x_0}, \sigma^2 \left(1 + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)\right)$$

Statistic

$$T = \frac{\hat{Y}_0 - \mu_{Y|x_0}}{S\sqrt{\frac{1}{n} + \frac{(x_0-\bar{x})^2}{S_{xx}}}} \sim T(n-2)$$

CI for mean of response

$$\hat{y}_0 - t_{\frac{\alpha}{2},n-2}\, s\sqrt{\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}} < \mu_{Y|x_0} < \hat{y}_0 - t_{\frac{\alpha}{2},n-2}\, s\sqrt{\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}}$$

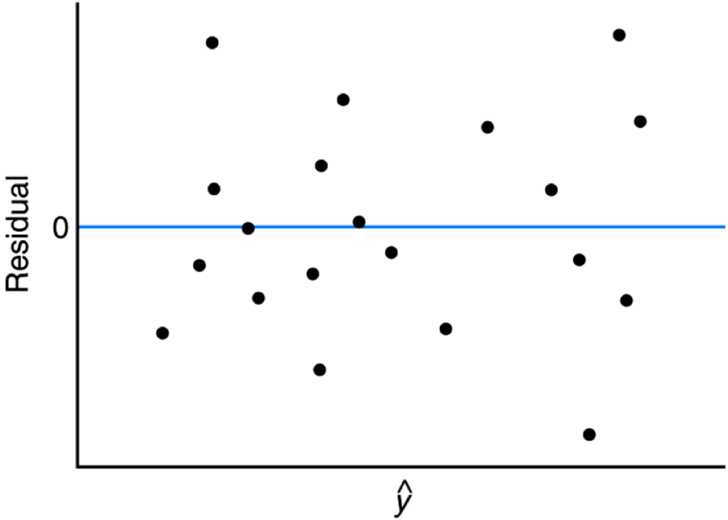
Diagnostic Plots of Residuals: Graphical Detection of Violation of Assumptions

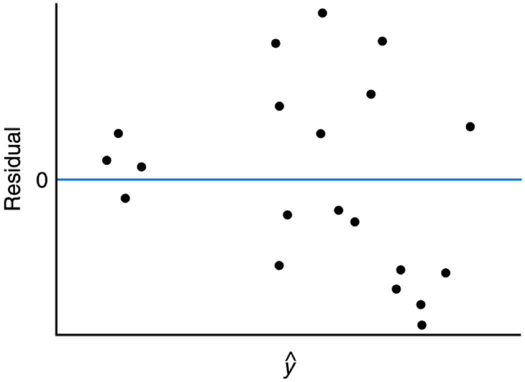
Model assumption

Errors ϵ_i are i.i.d $\mathcal{N}(0, \sigma^2)$

- Homogeneous variance
- Independence
- Normality

Ideal Residual plot





Ex: Increasing error variance with an increase in the regressor variable