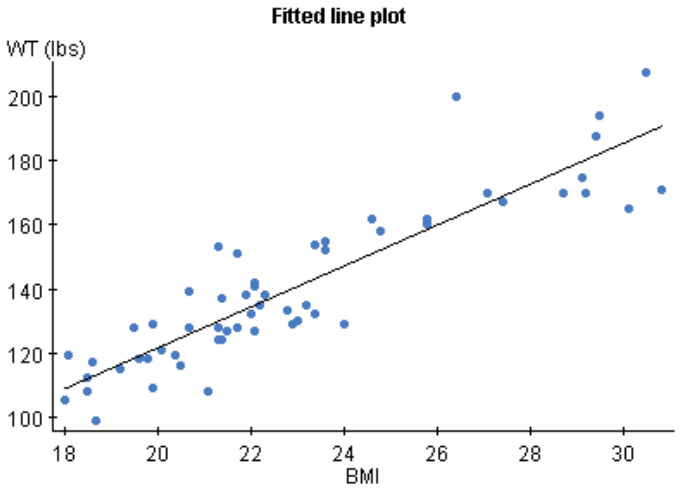
SIMPLE LINEAR REGRESSION

Why bother?

We have two variables X and Y. We are looking for a linear model of the relation between Y and Y. This is useful for:

- **Getting summary of the relation between X and Y.**
- **Prediction of Y from X, X- explanatory, Y – dependent/response variable.**
- **Setting the effect of X on Y aside (more in multiple regression).**
- **Simple Linear Regression (SLR): one explanatory variable**
- **Multiple Linear Regression (MLR): several explanatory variables.**

GOAL: Given bivariate data on (X,Y) find a line that fits the data "best"



SIMPLE LINEAR REGRESSION

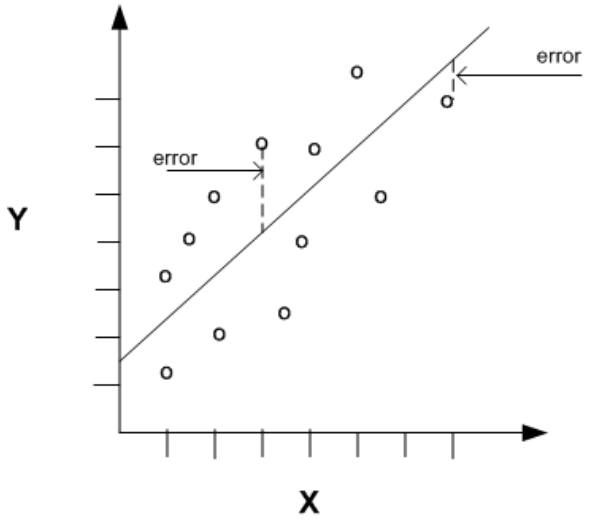
Data: $(x_1, y_1), \ldots, (x_n, y_n)$

- **The line that we are trying to fit is $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$.**
- **y_i is the dependent variable,**
- **x_i is the independent variable, and**
- **β_0 and β_1 are called the regression coefficients,**
- **ε_i *is* called the error.**

We only know the values of *x* and *y*, we must estimate β_0 and β_1.

So, we use the data to estimate β_0 and β_1.

FITTING THE LINE- THE LEAST SQUARES PRINCIPLE

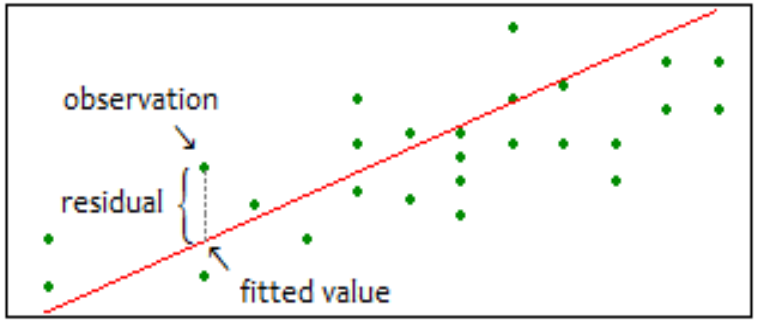


- **Equation of the fitted line:** $\hat{y} = \hat{\beta}_o + \hat{\beta}_1 x$

$\hat{\beta}_0$ and $\hat{\beta}_1$ **are called the least-squares coefficients.**

- **(Prediction) Errors =residuals =**

  predicted value-observed value



- **Objective for fitting a line: Make all prediction errors small.**
- **The least squares line is the line that minimizes the sum of squared errors (sum of squared residuals)**

FITTING THE LINE- THE LEAST SQUARES PRINCIPLE, contd.

- **Give data, we are looking for** $\hat{\beta}_0$ and $\hat{\beta}_1$ **such that the sum of squared errors is minimized:**

$$\sum_{i=1}^{n}(\hat{y}_i - y_i)^2 = (y_i - \beta_o - \beta_1 x)^2 \text{ minimized by } \widehat{\beta_o} \text{ and } \widehat{\beta_1} .$$

The estimated slope is

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

and the estimated intercept is

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

The estimated regression line is:

$$\hat{y} = \widehat{\beta_o} + \widehat{\beta_1} x$$

Note: The true values of β_0 and β_1 are unknown.

Another Representation of the Line

- **Another way to compute an estimate of β_1 is** $\hat{\beta}_1 = r\frac{S_y}{S_x}$

- **Another way to compute intercept is** $\widehat{\beta_0} = \overline{Y} - \widehat{\beta_1}\overline{X}$

- **R is the Pearson correlation coefficient between X and Y, and sx, sy are sample standard deviations of x's and y's in the sample.**

- **The slope of the regression line is proportional to the correlation coefficient.**

- **If r=0, then slope of the regression line is zero, so the line is horizontal. That means no relation between x and y.**

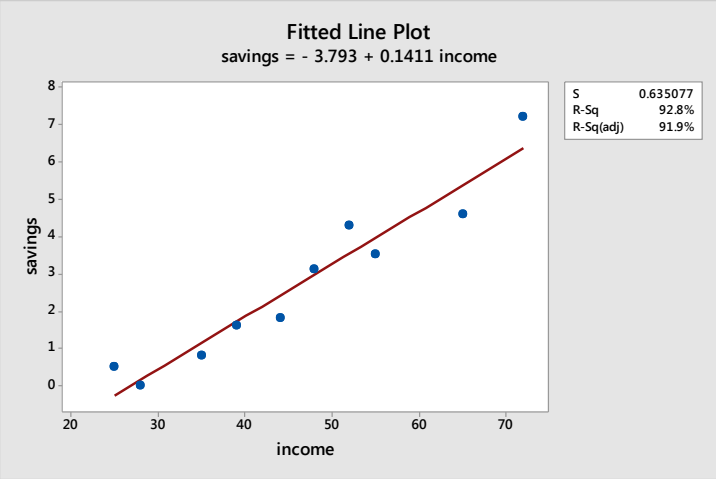
NOTE: We use regression line to predict Y from x only if X and y are significantly correlated, or equivalently if the slope of regression line is significantly different from zero.

EXAMPLE: Income and savings

In a study of income an savings, data was collected from 10 households.

Both savings and income are reported in thousands of $ in the following table :

(DATA SET: incomesav16.MTW or incomesav.xlsx)



income	savings
25	0.5
28	0.0
35	0.8
39	1.6
44	1.8
48	3.1
52	4.3
65	4.6
55	3.5
72	7.2

With the data already as two columns in MINITAB, click “Stats”, then “Regression, then “Simple Linear”. A window will come up and you need to enter information necessary to run regression: x-variable (income), an d y-variable (savings).

REGRESSION IN MINITAB

Regression Analysis: savings versus income

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	1	41.737	41.7374	103.48	0.000
income	1	41.737	41.7374	103.48	0.000
Error	8	3.227	0.4033		
Total	9	44.964			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.635077	92.82%	91.93%	86.31%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	-3.793	0.673	-5.64	0.000	
income	0.1411	0.0139	10.17	0.000	1.00

Regression Equation

savings = -3.793 + 0.1411 income

Sums of Squares

- $\sum_{i=1}^{n}(y_i - \hat{y})^2$ **is the error sum of squares and measures the overall spread of the points around the least-squares line.**

- $\sum_{i=1}^{n}(y_i - \bar{y})^2$ **is the total sum of squares and measures the overall spread of the points around the line** $y = \bar{y}$ **.**

- **The difference** $\sum_{i=1}^{n}(y_i - \bar{y})^2 - \sum_{i=1}^{n}(y_i - \hat{y})^2$ **is called the regression sum of squares and measures the reduction in the spread of points obtained by using the least-squares line rather than** $y = \bar{y}$ **.**

Total sum of squares = regression sum of squares + error sum of squares

Measure of Goodness of Fit- Correlation and Regression

- **A goodness-of-fit statistic is a quantity that measures how well a model explains a given set of data.**

- **The quantity r^2 is the square of the correlation coefficient and we call it the coefficient of determination.**

$$r^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2 - \sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

- **The proportion of variance in *y* explained by regression is the interpretation of r^2.**

The coefficient of determination r^2 expresses the reduction in variability as a proportion of the spread around $y = \bar{y}$

EXAMPLE

Income and savings. What percent of variability in savings is explained by variability in income?

Solution. The correlation coefficient was r = 0.963.

The coefficient of determination is $r^2=(0.963)^2=0.927$.

About 92.7% of variability in savings is explained by variability in income.

Uncertainties in the Least-Squares Coefficients

- **Linear Model:** $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$

- **Assumptions for Errors in Linear Models:**

  In the simplest situation, the following assumptions are satisfied:

  1. **The errors $\varepsilon_1,\ldots,\varepsilon_n$ are random and independent. In particular, the magnitude of any error ε_i does not influence the value of the next error ε_{i+1}.**
  2. **The errors $\varepsilon_1,\ldots,\varepsilon_n$ all have mean 0.**
  3. **The errors $\varepsilon_1,\ldots,\varepsilon_n$ all have the same variance, which we denote by σ^2.**
  4. **The errors $\varepsilon_1,\ldots,\varepsilon_n$ are normally distributed.**

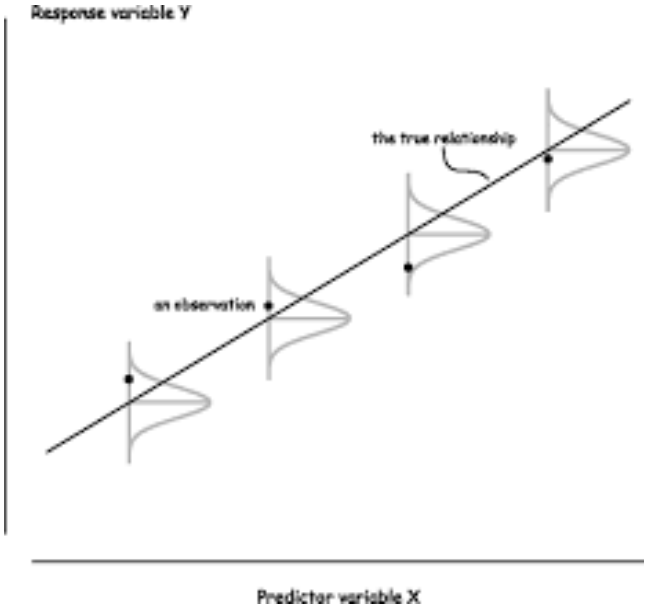
Distribution of Yi's

- **Assume linear model $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, and assumptions 1 through 4 hold.**

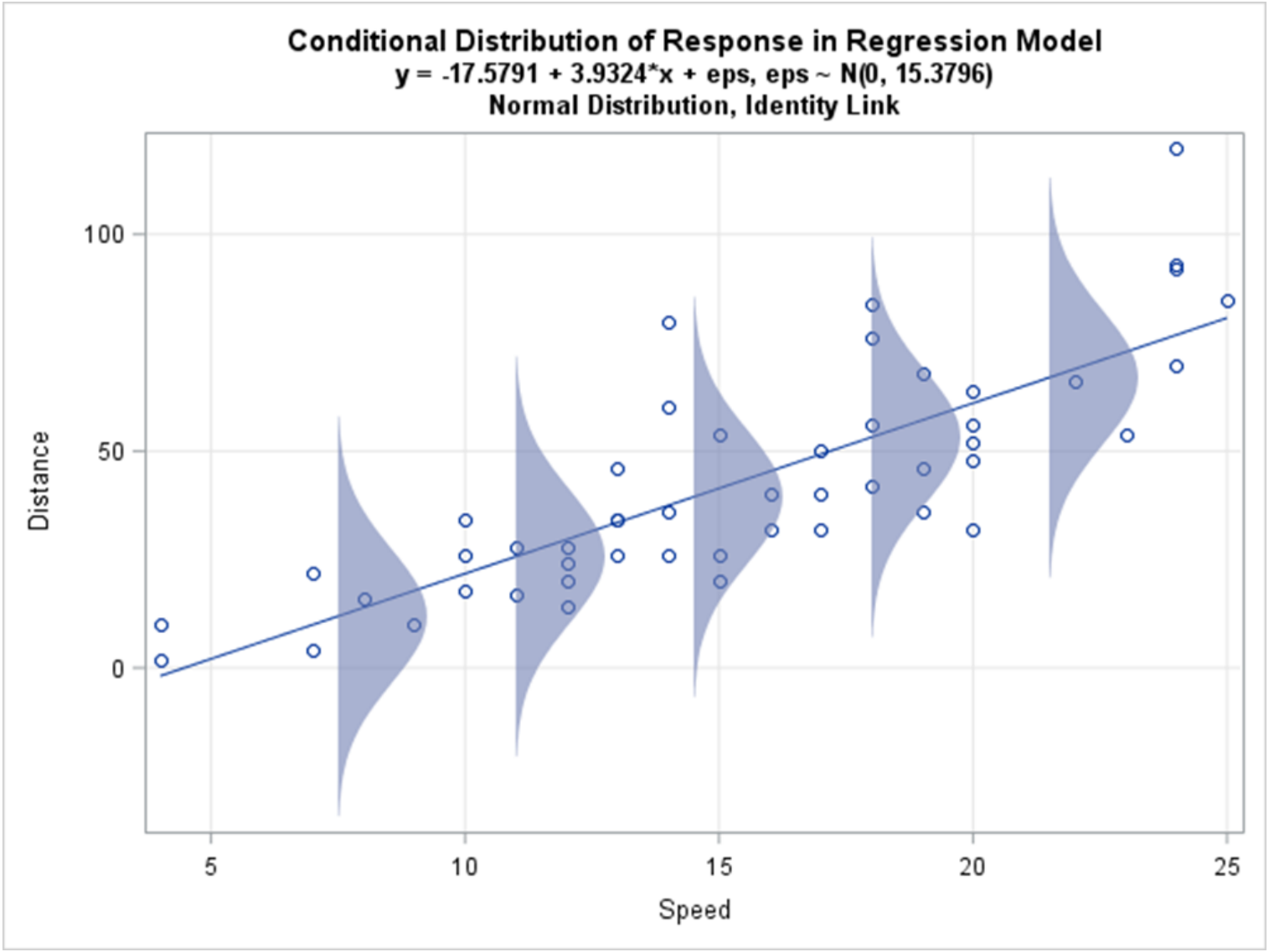
- **Then, the observations $y_1,\ldots, y_n$ are independent random variables that follow the normal distribution. The mean and variance of y_i are given by**

$$\mu_{y_i} = \beta_0 + \beta_1 x_i \qquad \sigma^2_{y_i} = \sigma^2$$

- **The slope represents the change in the mean of *y* associated with an increase in one unit in the value of *x*.**



Distribution of Y's for different x's



Distributions of the estimators for the slope and intercept

Under assumptions 1 – 4:

- **The quantities** $\hat{\beta}_0$ and $\hat{\beta}_1$ **are normally distributed random variables.**
- **The means of** $\hat{\beta}_0$ and $\hat{\beta}_1$ **are the true values β_0 and β_1, respectively.**
- **The standard deviations of** $\hat{\beta}_0$ and $\hat{\beta}_1$ **are estimated with**

$$s_{\hat{\beta}_0} = s\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}} \quad \text{and} \quad s_{\hat{\beta}_1} = \frac{s}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

where $s = \sqrt{\dfrac{(1-r)^2\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-2}}$ **is an estimate of the error standard deviation σ.**

All standard deviations above are reported in MINITAB within the "Coefficients" table

INFERENCE FOR REGRESSION: t-test

- The main purpose of regression is prediction of y from x.
- For prediction to be meaningful, we need y to depend significantly on x.
- In terms of the regression equation: $Y = \beta_0 + \beta_1 x + \varepsilon$, we need $\beta_1 \neq 0$.

Goal: Test hypothesis: Ho: $\beta_1 = 0$ (y does not depend on x) vs. Ha: $\beta_1 \neq 0$

Test statistic is based on the point estimate of β_1, $\widehat{\beta_1}$.

Test statistic

$$t = \frac{\widehat{\beta_1}}{SE_{\widehat{\beta_1}}} \text{ where } SE_{\widehat{\beta_1}} = \frac{S_Y}{S_X}\sqrt{\frac{1-r^2}{n-2}}.$$

Under Ho, the test statistic has t distribution with df=n-2.

For a two-sided Ha, p-value = 2P(t > t*), where t* is the computed value of the test statistic, and t denoted a t-distribution with n-2 df. Test statistic and p-value are reported in MINITAB.

REGRESSION IN MINITAB

Regression Analysis: savings versus income

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	1	41.737	41.7374	103.48	0.000
income	1	41.737	41.7374	103.48	0.000
Error	8	3.227	0.4033		
Total	9	44.964			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.635077	92.82%	91.93%	86.31%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	-3.793	0.673	-5.64	0.000	
income	0.1411	0.0139	10.17	0.000	1.00

The p-value reported in MINITAB is the p-value for the t-test of significance of regression (testing if slope is zero or not).

Regression Equation

savings = -3.793 + 0.1411 income

PREDICTION of an individual value of Y given a value of X

Given a value of x, say x* Predict individual value of y: Two cases:

1. If there is NO significant linear relationship between X and Y: best prediction of Y is average y:

$$\overline{y}$$

2. If there is a sign. relationship between X and Y, then for the best prediction of Y given X= x* use the regression line:

$$\hat{y} = \hat{\beta}_o + \hat{\beta}_1 x^*$$

And the (1-α)100% prediction interval for the individual future value

$$\hat{y} \pm t_{\alpha/2}(SE_{\hat{y}})$$

Standard prediction error

We compute the prediction interval in MINITAB

PREDICTION, contd.

Example. Income and savings. Predict savings for a family with income of $50k and find a 95% prediction interval for savings of a family with income of $50k.

Solution: savings = -3.793 + 0.1411 income

- Variable Setting
  income 50

- Fit SE Fit 95% CI 95% PI
  3.26211 0.207283 (2.78411, 3.74010) (1.72158, 4.80263)

For a family with income of $50k, we predict savings of $3,262. A 95% prediction interval for the savings of a family with income of 50k is (1.72158, 4.80263).

REGRESSION DIAGNOSTICS: RESIDUAL ANALYSIS I

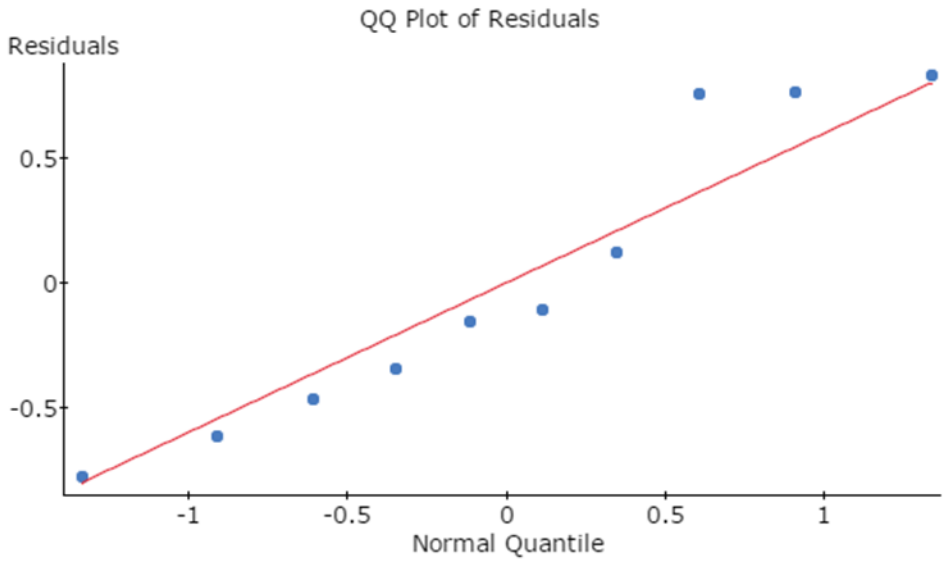
Regression model: $Y = \alpha + \beta x + \varepsilon$, $\varepsilon \sim N(0, \sigma)$

For the inference to work, we need the residuals to be approximately normal. Standard method is probability plot.

The model works well, if the normal probability plot is an approximately straight line.

Example. Income and savings.

The plot is approximately a straight line, so the model works well.



REGRESSION DIAGNOSTICS: RESIDUAL ANALYSIS II

For the inference to work, we need the residuals to HAVE THE SAME STANDARD DEVIATION. Standard method is plot of residuals versus fitted values : use a statistical package like MINITAB

The model works well, if the plot has no patterns.

Example. Income and savings.

The plot does not show any pattern, so the model works well.