Biostatistics

- Regression -

Sophia Daskalaki

Dept. of Electrical & Computer Engineering
✉: sdask@upatras.gr ☎: 2610-997810

Recommended Reading

- *Biostatistics*, A Foundation for Analysis in the Health Sciences, W.W. Daniel and C.L. Cross (Chapter 9)
- *Probability & Statistics for Engineers and Scientists* R.E.Walpole, R.H. Myers, S.L.Myers, K.Ye (Chapter 11)
- *Engineering Biostatistics*, An introduction using MATLAB and WINBUGS, B.Vidakovic (2016) (Chapter 14)

Simple Regression Analysis

Regression and Correlation

Suppose we are interested in studying the relationship that may exist between two or more continuous variables. Two popular statistical techniques for doing so are:

- **Regression analysis**, which is used primarily for *prediction*. The goal is to develop a model (linear or non-linear) that gives the dependent variable Y as a function of the rest of variables. Since the model is developed with the help of a single sample, part of regression analysis is concerned with the validation of the model before it is actually used for prediction.
- **Correlation analysis**, which is used to *measure the strength of the linear association between two continuous variables*. It can be used even when it is understood that the variables are not related with a causal-effect relationship.

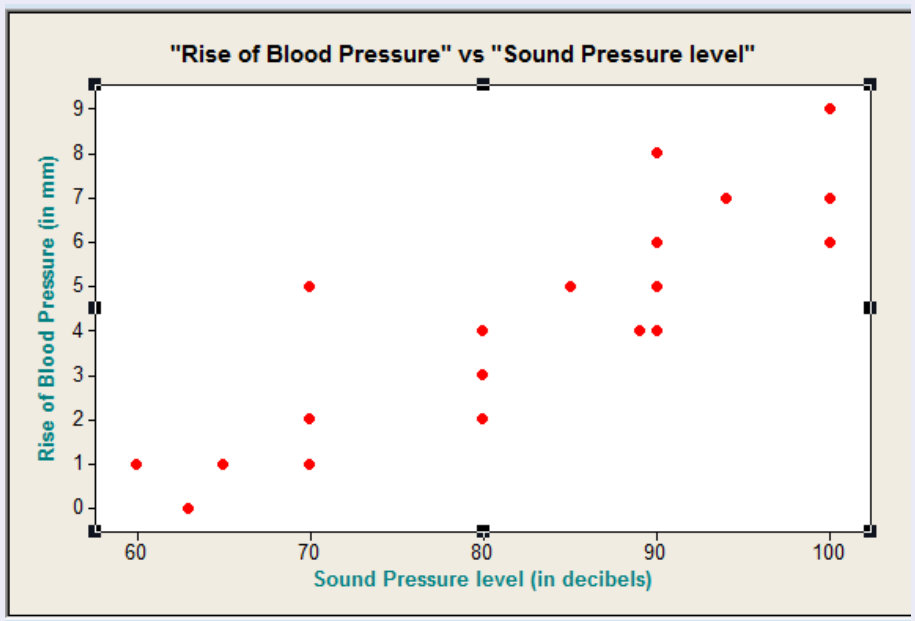
Graphical Representation

Scatterplot: a simple way to visualize the relationship between two variables.

First, get a sample which comprises of n pairs (x_i, y_i) for $i = 1, \ldots, n$. Then plot the pairs on the (x, y) 2-dimensional plane.

Example: A study investigated the relationship between noise exposure and hypertension. The sample comprised of 20 pairs (x, y), where x is the level of sound pressure (in *decibels*) and y is the blood pressure rise (in *mm*).

X (Noise level)	Y (Hypertension)
60	1
63	0
65	1
70	2
70	5
70	1
80	4
90	6
80	2
80	3
85	5
89	4
90	6
90	8
90	4
90	5
94	7
100	9
100	7
100	6

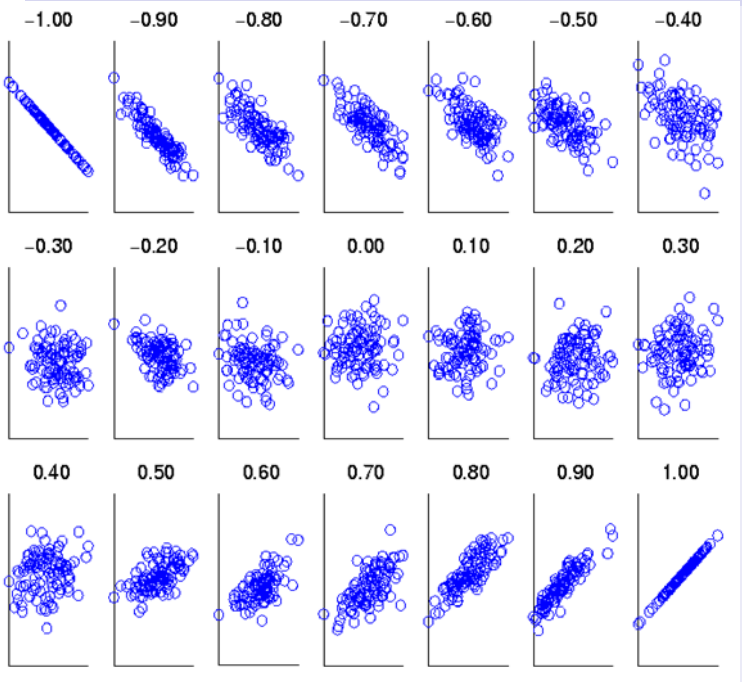


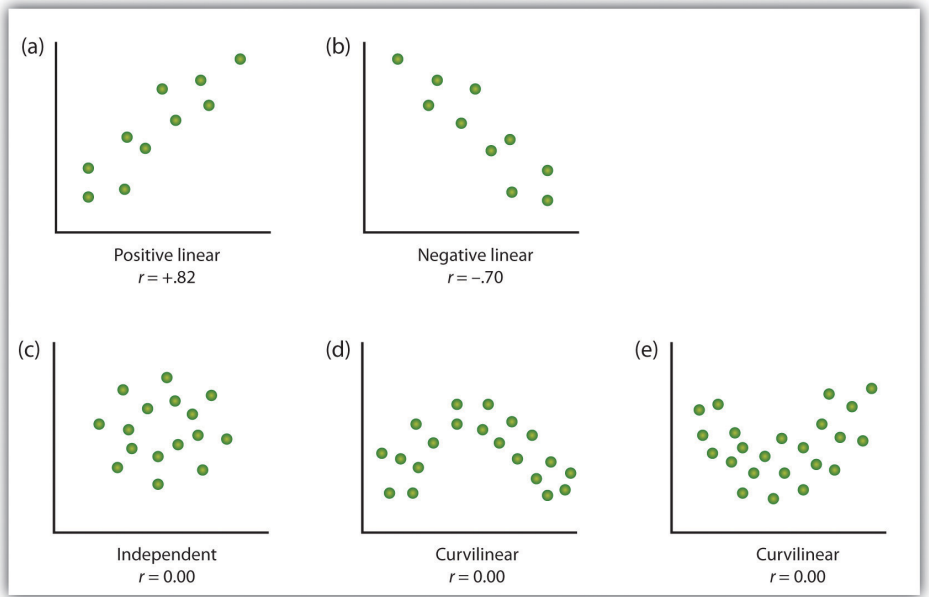
Pearson's Correlation Coefficient

Given a scatter plot for two variables it is easy to estimate the *correlation coefficient* using *Pearson's formula*.

$$r = \frac{n\sum_{i=1}^{n} x_i\ y_i - (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{\sqrt{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}\sqrt{n\sum_{i=1}^{n} y_i^2 - (\sum_{i=1}^{n} y_i)^2}}$$

This coefficient measures the *strength of the linear relation* between X and Y.





For the Simple Linear Regression it is assumed that the dependent variable Y is associated to the independent (or predictor) X with a linear relation. The goal here is to express the relation using information from a random sample. It is further assumed that the relationship between the two variables cannot be perfectly defined due to some inherent uncertainty which is embodied into the model using an error term.

Simple Linear Regression Model

$$Y = \beta_0 + \beta_1 x + \epsilon$$

where
$\beta_0 = Y-$intercept of the model
$\beta_1 =$ slope of the model
$\epsilon =$ random error in Y when $X = x$

Assumptions behind the Simple Linear Model

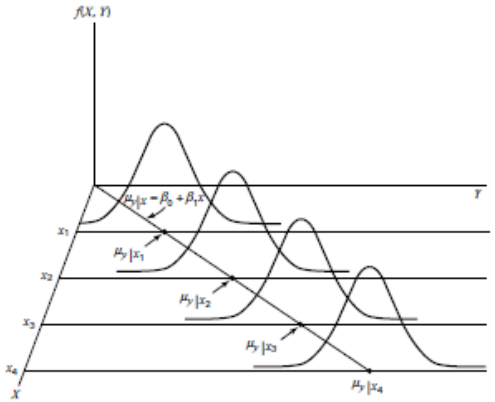


FIGURE 9.2.1 Representation of the simple linear regression model.

- Values of variable X can be considered as random or fixed at specific levels
- At each level x_i of X, it is assumed that Y follows a normal distribution with mean $\mu_{Y|X=x_i}$ and variance σ^2, the same for all levels of X
- The regression line $\widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x$ is an estimate of the line

$$\mu_{Y|X=x} = \beta_0 + \beta_1 x$$

which is the theoretical relation between X and Y

Error term in the Simple Linear Model

The *error term* is defined as:

$$\epsilon = Y - \mu_{Y|X=x}$$

and describes the variability we observe in Y when X is fixed at a certain level x.

Assumptions for the error term:

- ϵ follows the distribution of Y, i.e. it is normally distributed
- $E[\epsilon] = 0$ and $V(\epsilon) = \sigma^2$ for all levels of X
- If ϵ_i and ϵ_j are the error terms at two different levels of X, x_i and x_j, we assume that $Cov(\epsilon_i, \epsilon_j) = 0$. As a result the corresponding values of Y, Y_i and Y_j are also uncorrelated.

Meaning of Regression Parameters

The parameters β_0 and β_1 in the simple regression model are the *regression coefficients*. Their meaning results from the fact that the regression line gives the expected value of the dependent variable Y when the predictor X takes a specific value x.

- β_1 is the slope of the line and as such it gives the change in the mean of Y per unit increase in X.
- β_0 is the intercept of the line and gives the mean of Y at $X = 0$. If the value $X = 0$ is not included in the scope of the model, then we ignore the meaning of β_0.

Fitting a line

The most well known method for fitting a regression line to data is *Least Squares*. According to LS we estimate β_0 and β_1 so that the sum of squared deviations

$$Q(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 \, x_i)^2$$

is minimized. The solution of the two resulting equations are:

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

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

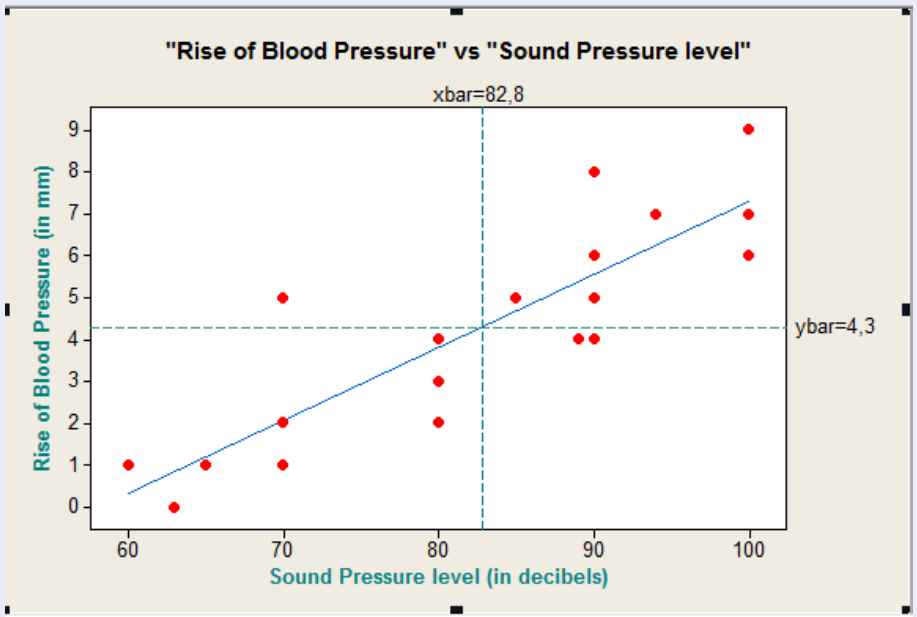
and the regression equation: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$

Fitting a least squares line – Example

Level of Sound Pressure vs Rise of Blood Pressure

$$\sum x_i = 1656 \sum y_i = 86, \sum x_i^2 = 140176 \sum x_i y_i = 7654, \sum y_i^2 = 494$$

$$\hat{y}_i = -10.1 + 0.174 x_i$$



Notes:

1. The line passes through the point $(\overline{x} = 82.8, \overline{y} = 4.3)$
2. It gives the mean value of Y for a given value of X

   For example: If someone is exposed to noise of 75 *db* we expect a rise in blood presure of $2.94mm$

Analysis of Variance in Regression

Variation in Y: Total Sum of Squares (SST) $= \sum_{i=1}^{n}(y_i - \bar{y})^2$

SST can be analyzed in the Sum of Squares that is due to *regression* (SSR) and the Sum of Squares that is due to *error* (SSE).

$$SST = SSR + SSE$$

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

- **Note**: $SST = \sum (y_i - \bar{y})^2 = \sum y_i^2 - \frac{1}{n}(\sum y_i)^2$ and $SSR = \sum(\hat{y}_i - \bar{y})^2 = \hat{\beta}_1 \left[\sum x_i y_i - \frac{1}{n}(\sum x_i)(\sum y_i)\right]$

ANalysis Of VAriance (ANOVA) Table

Source	**DF**	**SS**	**MS**
Regression	1	SSR	MSR=SSR/1
Error	n-2	SSE	MSE=SSE/n-2
Total	n-1	SST	

Evaluation of the model fit

- *Coefficient of Determination:*

$$R^2 = \frac{SSR}{SST}$$

It measures the proportion of variation in the dependent variable Y that is shared with or explained by the independent variable X.

- *Standard Error of the Estimate (or Standard Deviation of Residuals):*

$$s = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2}}$$

It measures the variability we observe on the actual Y_i values compared to their corresponding predicted values $\widehat{Y}_i$

Testing the significance of the model

This is a hypothesis testing procedure (*F test*) and uses the data in the sample to test whether the proposed model makes any sense.

- *Null and Alternative Hypotheses:*

$$H_0 : \beta_1 = 0 \text{ vs } H_1 : \beta_1 \neq 0$$

- *Statistical Control Function:*

$$F = \frac{MSR}{MSE} = \frac{SSR/1}{SSE/n-2}$$

- *Critical region:*

Reject H_0 at the α level of significance if $F > F_{\alpha;1,n-2}$ (significant model); otherwise if rejection of H_0 fails conclude that the model is not significant, i.e. there may be no linear relationship between the two variables

Evaluation of the model – Example

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,865[a]	,748	,734	1,318

a. Predictors: (Constant), X (Noise level)

b. Dependent Variable: Y (Hypertension)

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	92,934	1	92,934	53,502	,000[b]
	Residual	31,266	18	1,737		
	Total	124,200	19			

a. Dependent Variable: Y (Hypertension)

b. Predictors: (Constant), X (Noise level)

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95,0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	-10,132	1,995		-5,079	,000	-14,323	-5,940
	X (Noise level)	,174	,024	,865	7,314	,000	,124	,224

a. Dependent Variable: Y (Hypertension)

Statistical Inference on the Regression Model[1]

Distribution of $\hat{\beta}_1$

$$\hat{\beta}_1 \sim \mathcal{N}\left(\beta_1, \sigma\sqrt{\frac{1}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}\right) \Rightarrow \frac{\hat{\beta}_1-\beta_1}{\sigma\sqrt{\frac{1}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}} \sim \mathcal{N}(0,1)$$

$$\frac{\hat{\beta}_1-\beta_1}{s(\hat{\beta}_1)} \sim t_{n-2} \text{ where } s(\hat{\beta}_1) = s\sqrt{\frac{1}{\sum_{i=1}^{n}(x_i-\bar{x})^2}} \text{ and } s = \sqrt{\frac{SSE}{n-2}}$$

- CI for β_1: With $100(1-\alpha)\%$ certainty
  $\beta_1 \in \left[\hat{\beta}_1 \pm t_{(1-\alpha/2),n-2}\, s(\hat{\beta}_1)\right]$
- Tests for β_1: $H_0 : \beta_1 = 0$ vs $H_1 : \beta_1 \neq 0$ (*test for the slope*)

  test whether $t = \frac{\hat{\beta}_1}{s(\hat{\beta}_1)} \sim t_{n-2}$ at a certain significance level α

[1] In order to build conf. intervals and hypothesis tests for the regression model we need the assumption of normality for the error term.

More Statistical Inference on the Regression Model

Distribution of $\hat{\beta}_0$

$$\hat{\beta}_0 \sim \mathcal{N}\left(\beta_0, \sigma\sqrt{\frac{\sum x_i^2}{n\sum_{i=1}^n (x_i-\bar{x})^2}}\right) \Rightarrow \frac{\hat{\beta}_0-\beta_0}{\sigma\sqrt{\frac{\sum x_i^2}{n\sum_{i=1}^n (x_i-\bar{x})^2}}} \sim \mathcal{N}(0,1)$$

$$\frac{\hat{\beta}_0-\beta_0}{s(\hat{\beta}_0)} \sim t_{n-2} \text{ where } s(\hat{\beta}_0) = s\sqrt{\frac{\sum x_i^2}{n\sum_{i=1}^n (x_i-\bar{x})^2}} \text{ and } s = \sqrt{\frac{SSE}{n-2}}$$

- CI for β_0: With $100(1-\alpha)\%$ certainty
  $\beta_0 \in \left[\hat{\beta}_0 \pm t_{(1-\alpha/2);n-2}\, s(\hat{\beta}_0)\right]$
- Tests for β_0: $H_0 : \beta_0 = \beta_0^0$ vs $H_1 : \beta_0 \neq \beta_0^0$

  test whether $t = \frac{\hat{\beta}_0-\beta_0^0}{s(\hat{\beta}_0)} \sim t_{n-2}$ at a certain significance level α

The Regression Model in Action

Major goal for building regression models is *estimation of mean values of* Y or *prediction of values of* Y when X takes a certain value x_0. The estimation and prediction can be done either with *point estimates* or with *interval estimates*.

Let x_0 be a value for X for which we are interested in estimating the mean value of Y. Then

- Point estimate: $\hat{y}_0 = b_0 + b_1 x_0$
- Confidence Interval for $E(Y_0)$:

$$\hat{y}_0 \pm t_{(1-\alpha/2);n-2}\, s \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

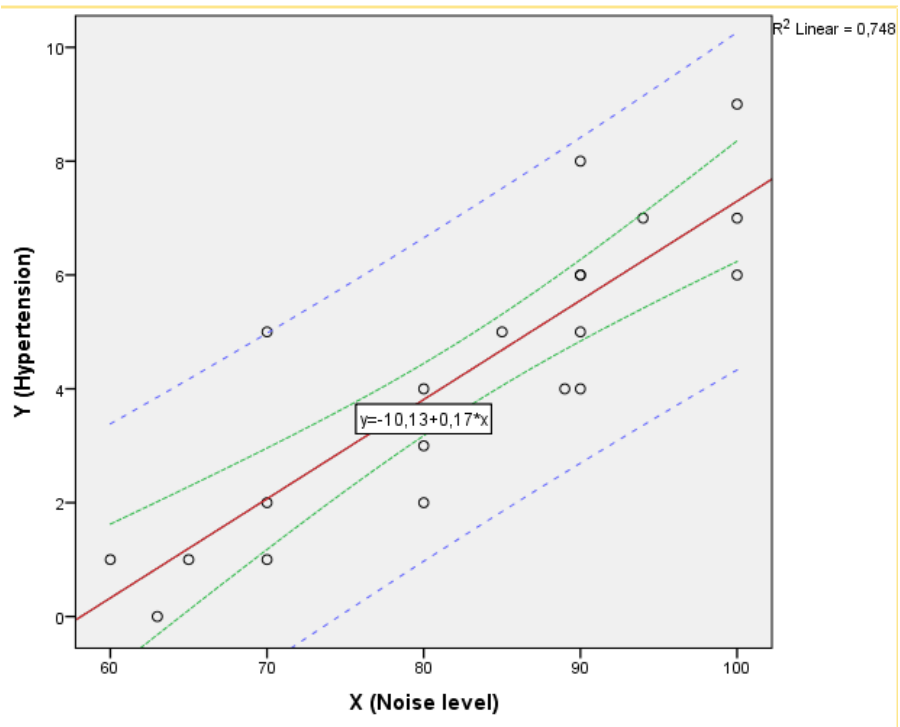
- Prediction Interval for a new observation of Y:

$$\hat{y}_0 \pm t_{(1-\alpha/2);n-2}\, s \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

Estimation and Prediction – Example

Suppose we want (1) to estimate the *mean* blood pressure rise (for the whole population) and (2) to predict the blood pressure rise (for any individual) when the noise level is measured at 75, 85, or 95 decibels.

XNoiselevel	YHypertension	PRE_1	RES_1	SEP_1	LMCI_1	UMCI_1	LICI_1	UICI_1
75	.	2,94051	.	,34842	2,20850	3,67251	,07645	5,80457
85	.	4,68345	.	,29933	4,05457	5,31232	1,84399	7,52290
				,41396	5,55669	7,29608	3,52408	9,32869



red line: Regression line

green curves: 95% confidence intervals for $\mu_Y|X = x$

Blue curves: 95% prediction intervals for Y when $X = x$

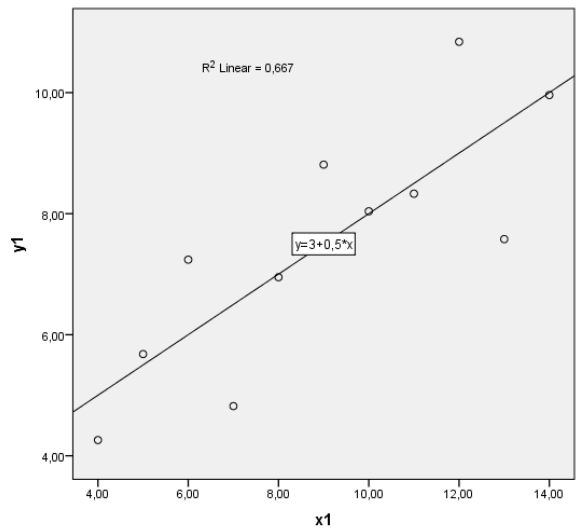
Visualization in Regression

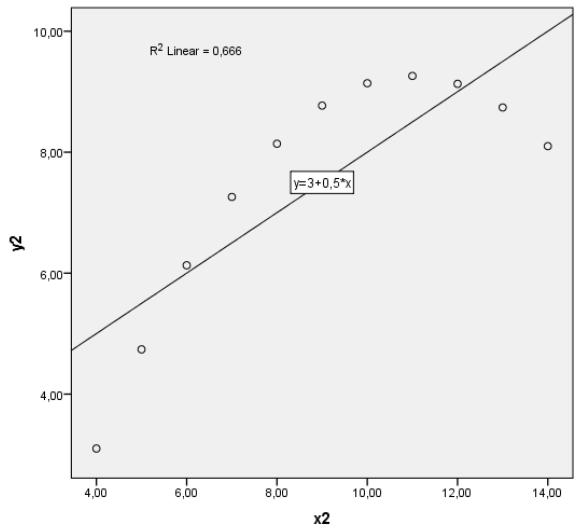
A celebrated classic example of the role of statistical graphics and residual analysis in statistical modeling was created by Anscombe in 1973[2]. He constructed four quite different data sets (x_i, y_i), $i = 1, ..., 11$ that share the same descriptive statistics and linear regression fit $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$

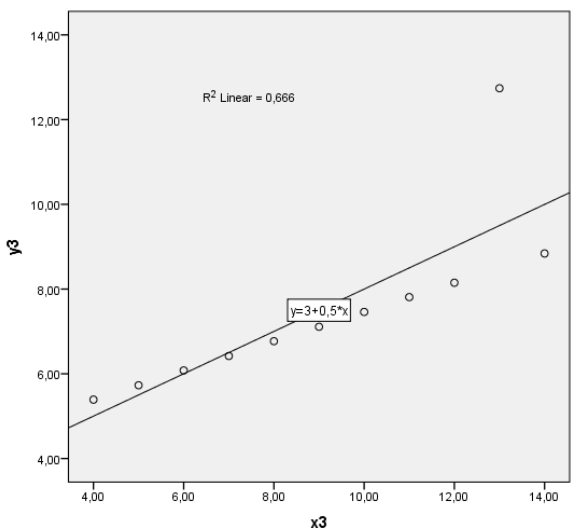
Anscombe's Data										
Observation	x1	y1		x2	y2		x3	y3		x4	y4
1	10	8.04		10	9.14		10	7.46		8	6.58
2	8	6.95		8	8.14		8	6.77		8	5.76
3	13	7.58		13	8.74		13	12.74		8	7.71
4	9	8.81		9	8.77		9	7.11		8	8.84
5	11	8.33		11	9.26		11	7.81		8	8.47
6	14	9.96		14	8.1		14	8.84		8	7.04
7	6	7.24		6	6.13		6	6.08		8	5.25
8	4	4.26		4	3.1		4	5.39		19	12.5
9	12	10.84		12	9.13		12	8.15		8	5.56
10	7	4.82		7	7.26		7	6.42		8	7.91
11	5	5.68		5	4.74		5	5.73		8	6.89
				Summary Statistics							
N	11	11		11	11		11	11		11	11
mean	9.00	7.50		9.00	7.50091		9.00	7.50		9.00	7.50
SD	3.16	1.94		3.16	1.94		3.16	1.94		3.16	1.94
r	0.82			0.82			0.82			0.82	

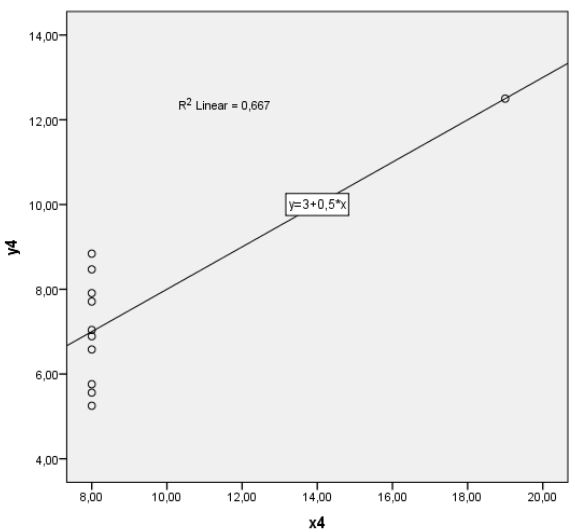
[2] Anscombe,F. Graphs in Statistical Analysis. *American Statistician*, 27, 17-21

Fitting Lines in Anscombe Data









ANOVA for Anscombe Data

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	27,510	1	27,510	17,990	,002[b]
	Residual	13,763	9	1,529		
	Total	41,273	10			

a. Dependent Variable: y1

b. Predictors: (Constant), x1

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	27,500	1	27,500	17,966	,002[b]
	Residual	13,776	9	1,531		
	Total	41,276	10			

a. Dependent Variable: y2

b. Predictors: (Constant), x2

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	27,470	1	27,470	17,972	,002[b]
	Residual	13,756	9	1,528		
	Total	41,226	10			

a. Dependent Variable: y3

b. Predictors: (Constant), x3

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	27,490	1	27,490	18,003	,002[b]
	Residual	13,742	9	1,527		
	Total	41,232	10			

a. Dependent Variable: y4

b. Predictors: (Constant), x4

- All four datasets give the same regression line and the same ANOVA table. So how can we evaluate the models?

Residual Plots for Anscombe Data