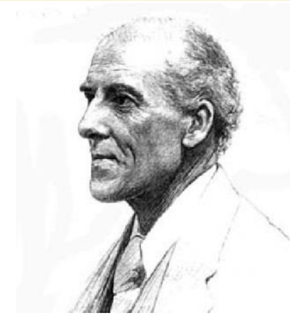
MAT20306 - Advanced Statistics

Lecture 5: Correlation & Simple linear regression

WAGENINGEN UR
For quality of life

Biometris
Quantitative Methods brought to Life

Pearson correlation coefficient



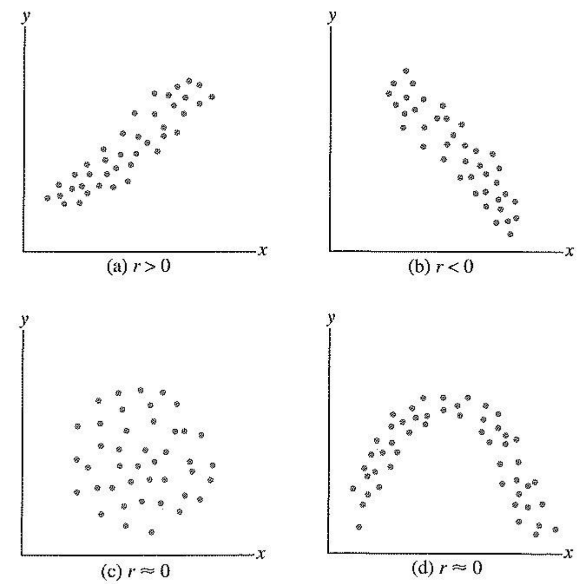
- when people talk about a correlation or correlation coefficient they usually mean Pearson's correlation coefficient
  - named after Karl Pearson (1857–1936), British statistician

- Pearson's correlation coefficient ρ_{xy} measures the strength of the linear association between two quantitative variables x and y, see figure (O&L 11.20)

- ρ_{xy} is always between –1 and +1.

- values close to 1 or –1 $\Rightarrow$ strong (linear) association, values close to 0 $\Rightarrow$ little or no (linear) association

- when correlation ρ_{xy} =1 or ρ_{xy} –1,



Pearson correlation coefficient, continued

- There is no distinction between dependent and independent variables: $\rho_{xy} = \rho_{yx}$.

- The absolute value of ρ_{xy} is not affected by linear transformations of x or y, e.g. correlation between x and y is the same as between $2\,x + 1$ and $10 + 5\,y$.
  So, it does not matter whether measurements are in e.g. grams or kilograms.

- When x and y are independent, $\rho_{xy} = \rho_{yx} = 0$, but the reverse is not necessarily true.

- The correlation ρ_{xy} is a population parameter that is estimated by the sample correlation r_{xy} :

$$r_{xy} = r_{yx} = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 \sum_{i=1}^{n} (y_i - \bar{y})^2}}, \quad -1 \le r_{xy} \le 1$$

Correlation & inference

- Test on ρ_{xy}

1. H_0: $\rho_{xy} = 0$.

2. Test statistic is:

$$t = r_{xy} \frac{\sqrt{n-2}}{\sqrt{1-r_{xy}^2}}$$

3. When H_0 is true, $t \sim t_{n-2}$

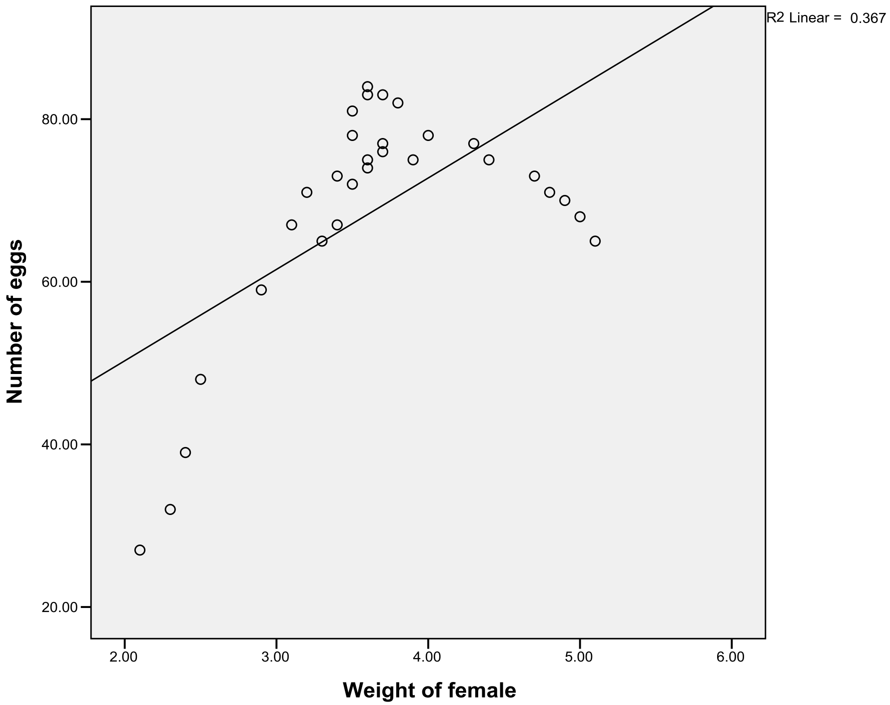
WAGENINGEN UR
For quality of life

Grasshoppers (Example 11.13 in O&L)

Study of the reproductive success of grasshoppers.

An entomologist collected a sample of 30 female grasshoppers. She recorded the number of mature eggs produced and the body weight of each of the females (grams).

	Number	weight
1	27.00	2.10
2	32.00	2.30
3	39.00	2.40
4	48.00	2.50
5	59.00	2.90
6	67.00	3.10
7	71.00	3.20
8	65.00	3.30
9	73.00	3.40
10	67.00	3.40
11	78.00	3.50
12	72.00	3.50
13	81.00	3.50
14	74.00	3.60
15	83.00	3.60

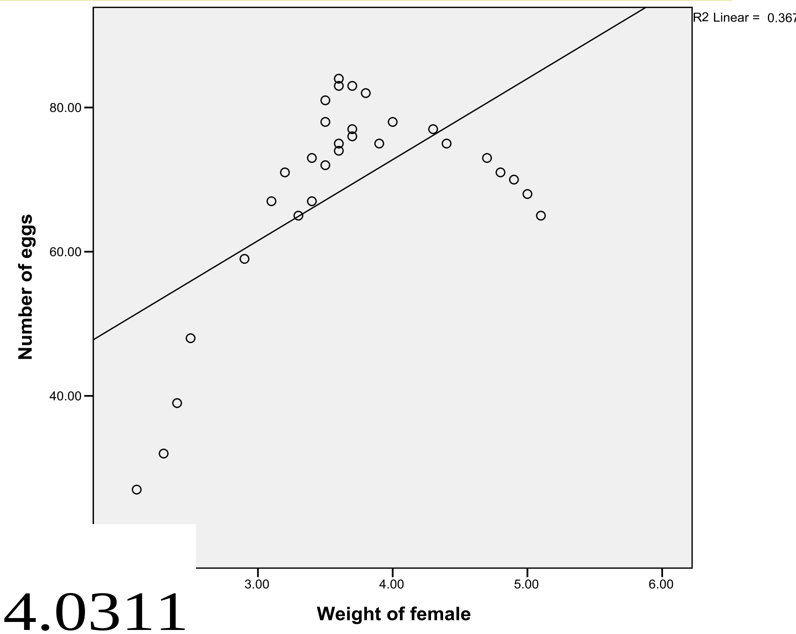


Grasshoppers (Example 11.13 in O&L)

Correlations

		Number of eggs	Weight of female
Number of eggs	Pearson Correlation	1	.606**
	Sig. (2-tailed)		.000
	N	30	30
Weight of female	Pearson Correlation	.606**	1
	Sig. (2-tailed)	.000	
	N	30	30

**. Correlation is significant at the 0.01 level (2-tailed).



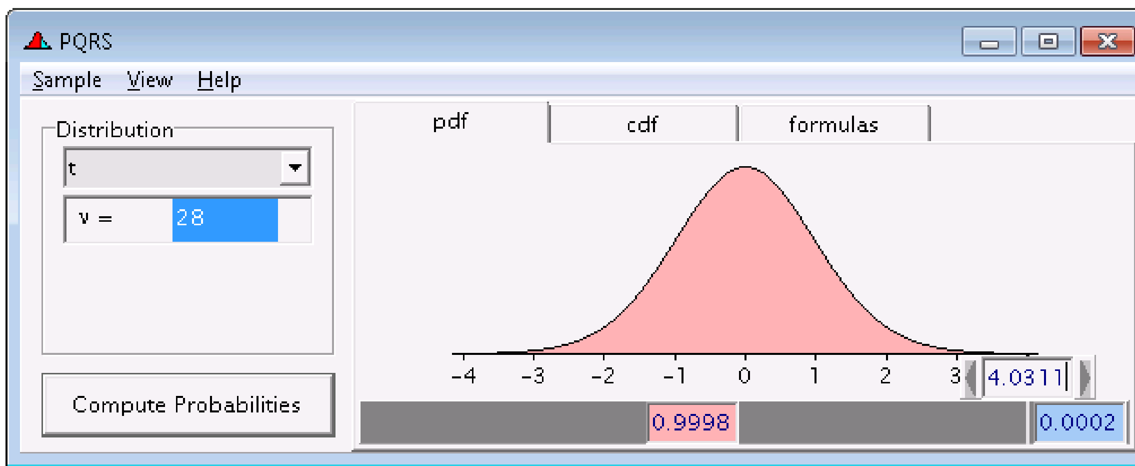
H_0: $\rho_{xy} = 0$ vs H_A: $\rho_{xy} > 0$

$$t = r_{xy}\frac{\sqrt{n-2}}{\sqrt{1-r_{xy}^2}} = 0.606\cdot\frac{\sqrt{30-2}}{\sqrt{1-0.606^2}} = 4.0311$$

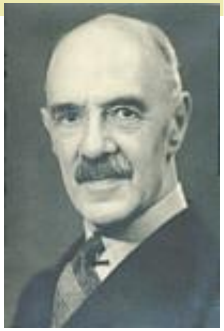
Under H_0: $t_{n-2}=t_{28}$ distribution

RSP=0.000 <0.05, so reject H_0

We have shown there is a positive correlation between weight and number of eggs



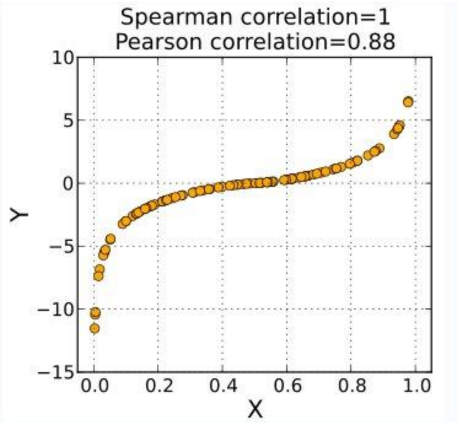
Spearman rank correlation



- r_{xy} is highly sensitive to outlying observations (outliers)
- an alternative is Spearman's rank correlation r_S (not mentioned in O&L), named after Charles Spearman (1863 – 1945), English psychologist

- observations are replaced by rank numbers
  ranking x and y separately, with mid ranks in case of ties
- Spearman's r_S is the ordinary correlation, but derived from these rank numbers

- r_s measures the strength of a monotonic relationship between two quantitative variables x and y.
  The relationship need not be linear, see figure from Wikipedia.

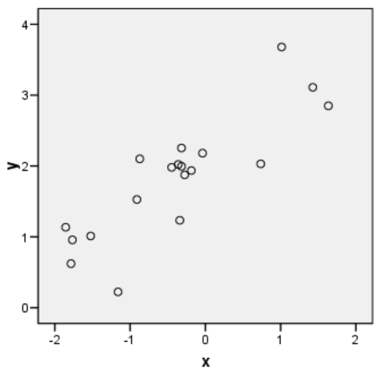


- when data are approximately normally distributed (without outliers), r_S and r_{xy} tend to be similar.
- but r_s is not estimating a population parameter, in contrast to r_{xy} ,

WAGENINGEN UR
For quality of life

An example of Spearman’s rank correlation

x	y
-.91	1.53
-.04	2.18
-.28	1.88
-.36	2.02
-1.86	1.14
-1.77	.96
-.32	2.25
1.63	2.85
-.19	1.94
-.32	2.00
-.34	1.23
-1.16	.22
1.43	3.11
-.87	2.10
-.45	1.98
-1.79	.62
-1.52	1.01
1.01	3.68
.74	2.03



Correlations

		x	y
x	Pearson Correlation	1	.852**
	Sig. (2-tailed)		.000
	N	19	19
y	Pearson Correlation	.852**	1
	Sig. (2-tailed)	.000	
	N	19	19

**. Correlation is significant at the 0.01 level

Note that here Pearson correlation and Spearman rank correlation are similar.

x	y	Rx	Ry
-.91	1.53	6	7
-.04	2.18	15	15
-.28	1.88	13	8
-.36	2.02	9	12
-1.86	1.14	1	5
-1.77	.96	3	3
-.32	2.25	11	16
1.63	2.85	19	17
-.19	1.94	14	9
-.32	2.00	12	11
-.34	1.23	10	6
-1.16	.22	5	1
1.43	3.11	18	18
-.87	2.10	7	14
-.45	1.98	8	10
-1.79	.62	2	2
-1.52	1.01	4	4
1.01	3.68	17	19
.74	2.03	16	13

Spearman’ correlation of 0.821 can be Obtained by calculating Pearson’s correlation on rank numbers

Correlations

		Rank of x	Rank of y
Rank of x	Pearson Correlation	1	.821**
	Sig. (2-tailed)		.000
	N	19	19
Rank of y	Pearson Correlation	.821**	1
	Sig. (2-tailed)	.000	
	N	19	19

**. Correlation is significant at the 0.01 level (2-tailed).

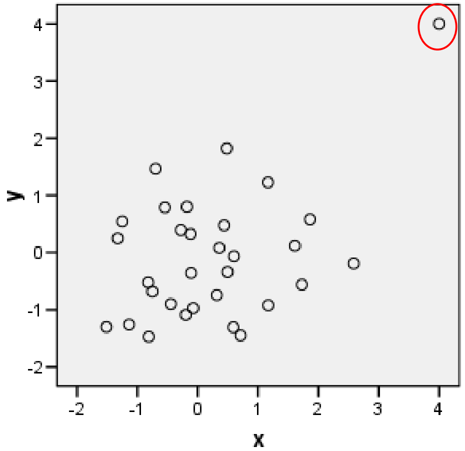
Correlations

			x	y
Spearman's rho	x	Correlation Coefficient	1.000	.821**
		Sig. (2-tailed)	.	.000
		N	19	19
	y	Correlation Coefficient	.821**	1.000
		Sig. (2-tailed)	.000	.
		N	19	19

**. Correlation is significant at the 0.01 level (2-tailed).

Another example of Spearman's rank correlation

Unrelated *x* and y, with one added outlying observation with both high *x* and *y* value.



Correlations

		x	y
x	Pearson Correlation	1	.445*
	Sig. (2-tailed)		.012
	N	31	31
y	Pearson Correlation	.445*	1
	Sig. (2-tailed)	.012	
	N	31	31

*. Correlation is significant at the 0.05 level (2-tailed).

Pearson correlation is sensitive to the outlier: relatively high correlation (and significantly different from 0)

Spearman correlation is not really sensitive to the outlier and consequently lower

Correlations

			x	y
Spearman's rho	x	Correlation Coefficient	1.000	.185
		Sig. (2-tailed)	.	.319
		N	31	31
	y	Correlation Coefficient	.185	1.000
		Sig. (2-tailed)	.319	.
		N	31	31

WAGENINGEN UR
For quality of life

Simple Linear Regression

Overview:

1) Define the model
2) Estimate the model
3) Inference on model parameters (by means of t-test and C.I.)
4) Test the model : ANOVA table
5) Checking model assumptions
6) Prediction by using the model

O&L Chapter 11 (11.1-11.6)

Example fish storage in ice

Storage of raw fish in ice is delayed by x hours, x = 0,3,6,9,12, each with 2 replicates. After a 7-day storage in ice the quality (y) of each fish is measured on a 10 point scale.

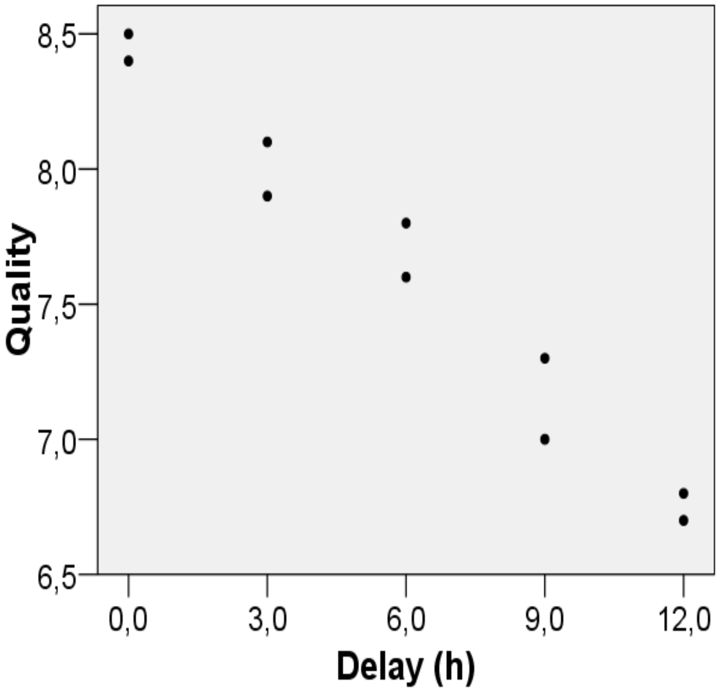
Delay (x)	0	3	6	9	12
Quality(y)	8.5	7.9	7.8	7.3	6.8
	8.4	8.1	7.6	7	6.7

Question: How does y depend upon delay x?

There are many types of relationship.
To create a framework for an answer:
we **assume** a linear relationship between **mean of *y*** and x :

$$\mu_y = \beta_0 + \beta_1 x$$

1. Individual values of y may deviate from the mean value on the line.
2. The problem simplifies to finding only two *parameters*: β_1 and β_0.



Linearity is an assumption, which needs checking. Does it seem reasonable here?

1. Simple linear regression model

- Model : $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ $\quad i = 1,2,\ldots,n$ $\quad \varepsilon_i \sim N(0,\sigma) \quad \varepsilon_i\text{'s}\ independent$

  ε_i 's are often called "errors".

  We can also write:

  $y_i \sim N(\mu_i, \sigma)\,, \quad y_i\text{'s are independent}$

- y is called response or dependent variable. It is *numerical/quantitative.*
- x is called regressor, independent variable or explanatory variable. It is usually *numerical.* It can be fixed (in experiment) or observed (random).
- The regression coefficients β_0 and β_1, and standard deviation σ are the (unknown) **parameters** of the regression model. What do they mean?

  β_0 = intercept = mean response when x = 0

  β_0 has a practical interpretation only if x = 0 is in experimental region.

  β_1 = slope = change in mean response when x increases by 1 unit.

  $\sigma = \sigma_\varepsilon$ = standard deviation of ε

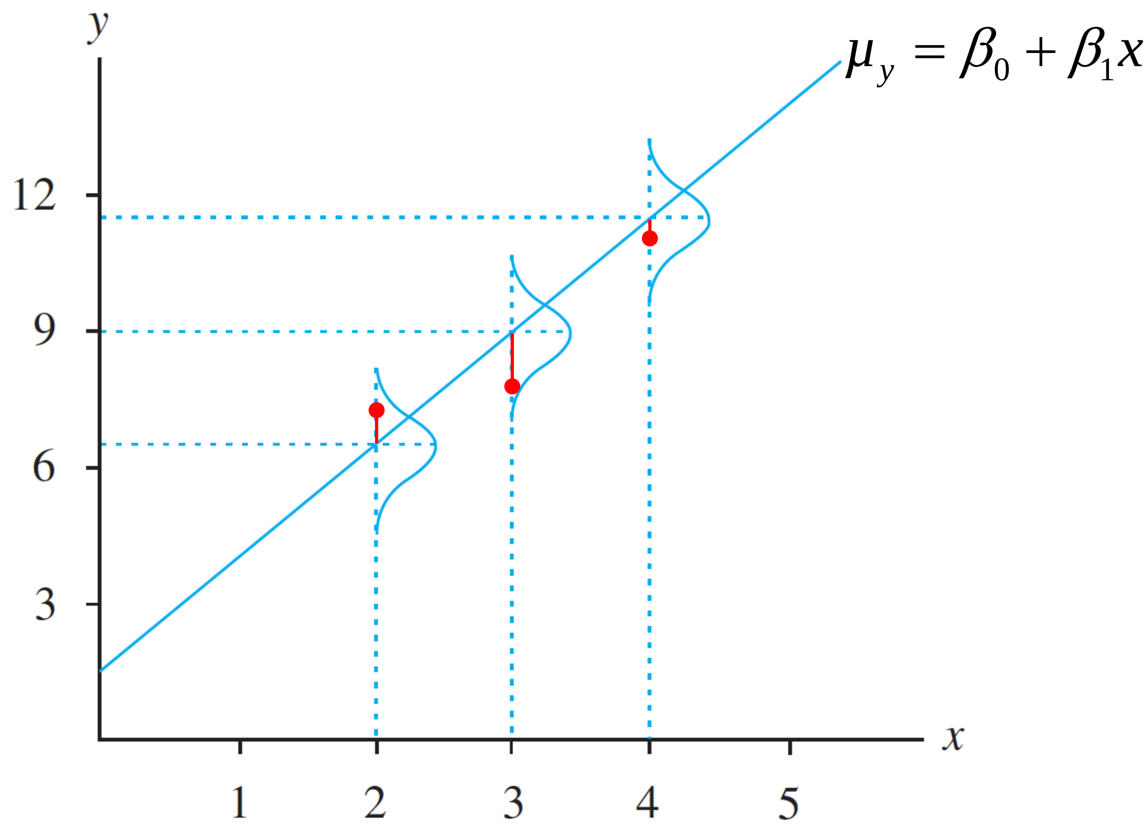
  = standard deviation of y corrected for x

  = standard deviation of y "around the regression line".

Constant standard deviation σ_ε

Errors ε are normally distributed with expected value 0, and constant standard deviation σ_ε

assumed to be the same for all values of *x.*



2. Least Squares Estimation of β_0 and β_1

- Question: What is the best line through the points?

  = What are the best estimates for β_0 and β_1?

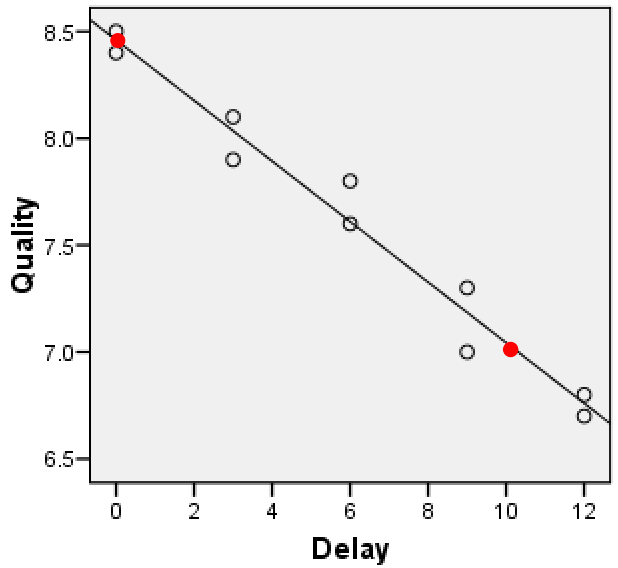
In the scatterplot this "best" line is shown. Guess the equation.

- To answer this, a *criterion* to be minimized is needed that combines the distances of the points to the line into one number
- The criterion generally chosen is:
  the 1) *sum* of 2) *squared* 3) *vertical* distances from the points to the line.
- This is called the *Least Squares Method*.
- Deviation = $e_i = y_i - \hat{y}_i = y_i - (b_0 + b_1 x_i)$

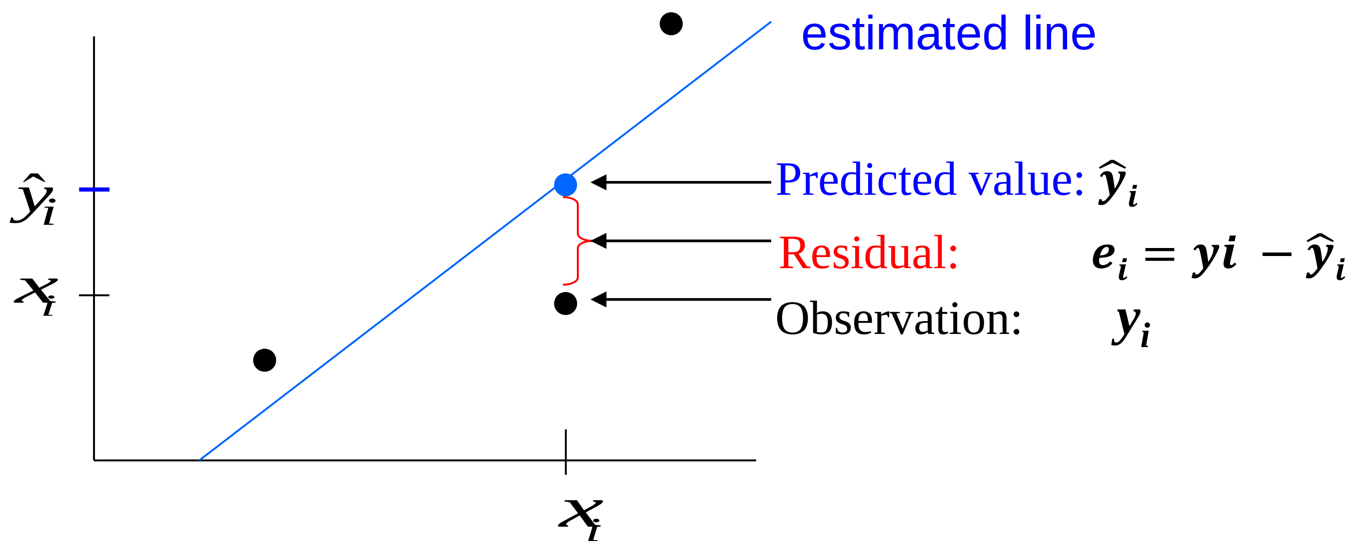
  SSE = $\Sigma_i e_i^2$. The $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimize SSE:

$$\hat{\beta}_1 = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sum (x-\bar{x})^2} = \frac{S_{xy}}{S_{xx}}$$

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



Predicted values and residuals



- Predicted or fitted value (model value) $\hat{y}_i$ (pronounce y-i- hat),

  predicted value $\hat{y}_i$ is the expected value of y according to the fitted regression

  line at the given x–value x_i.

- Residual e_i

  the difference between the observed y_i and the predicted value $\hat{y}_i$,

  the distance between the point and the line in the y–direction,

  and an "estimate" for error ε_i .

3. Inference for slope β_1 (and intercept β_0)

$\widehat{\beta_1}$ has a standard error $se(\hat{\beta}_1) = SE_{b_1} \left(= s_\varepsilon \sqrt{1/S_{xx}}\right)$ We read it from SPSS.

Confidence interval for β_1: $\left(\hat{\beta}_1 \pm t_{\alpha/2, n-2}\, se(\hat{\beta}_1)\right)$

d.f. = (n–2), because 2 parameters (β_0, β_1) are estimated

T-test for H_0: $\beta_1 = 0$:
(SPSS gives <u>all</u> output)

$$TS:\ t = \frac{\hat{\beta}_1 - 0}{se(\hat{\beta}_1)},\ \text{when}\ \ H_0\ \text{is true}\ \ t \sim t_{n-2}$$

For e.g. H_0: $\beta_1 = 1.3$, use
(SPSS gives no t- or P-value)

$$t = \frac{\hat{\beta}_1 - 1.3}{se(\hat{\beta}_1)},\ \text{when}\ H_0\ \text{is true}\ \ t \sim t_{n-2}$$

Inference for β_0, also based on t_{n-2}-distribution, proceeds likewise.

***H_0*: $\boldsymbol{\beta_1}$= 0** can also be tested using an F-test, but **only for *H_a*: $\boldsymbol{\beta_1 \neq 0}$:**

Fish storage, SPSS output

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	8.460	.066		127.995	.000
	Delay (h)	-.142	.009	-.984	-15.750	.000

a. Dependent Variable: Quality

$$\begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix}$$

you should be able to interpret all output (except the standardized coefficients) and know by what principle it is obtained

Notation: we may use b_1 for $\widehat{\beta_1}$

Example of a test. Does more Delay reduce fish quality?

1) H_0: $\beta_1 = 0$ vs H_a: $\beta_1 < 0$.

2) TS: t= $b_1/se(b_1)$. *3)* Under H_0 t~ t_8 (n=10)

4/5) Under H_a t tends to smaller values, so we use LPV.

6) Outcome TS: t= -15.75

7) LPV = 0.000/2

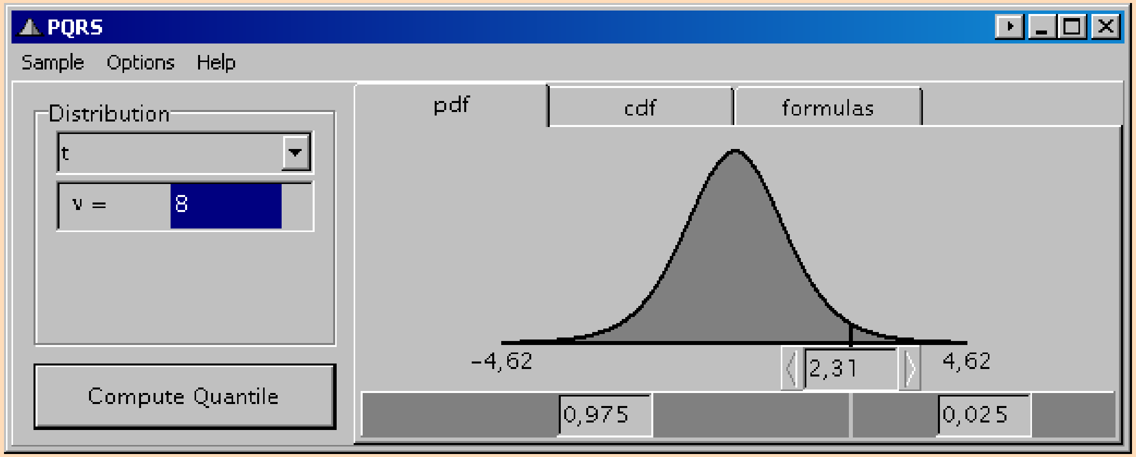
8) H_0 is rejected, H_a is proven. It is shown ($\alpha = 0.05$) that more delay leads to lower **mean** fish quality

Fish storage, two-sided confidence interval

two-sided 0.95-confidence interval for b_1 :

$$(b_1 \pm t_8(0.025) * \mathrm{SE}_{b1}) \rightarrow (-0.142 \pm 2.31 * 0.009)$$

so, 0.95-confidence interval is: (–0.163, –0.121)



SPSS summary output for regression: r_{yx}, R^2, s_ε

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.984[a]	.969	.965	.12068

a. Predictors: (Constant), Delay (h)

b. Dependent Variable: Quality

$R = |\, r_{yx} \,|$

s_ε

Coefficient of determination $R^2 = r_{yx}^2$

- When the values for x are chosen over a wider range (if this is possible in the design stage), R^2 will increase, but the intercept, slope and residual variance will remain about the same (apart from estimation error).
- *So, although* R^2 is quite popular, it's size depends on the choice of values of x, therefore, R^2 should be handled with care.
- Note that for a correlation we need a random sample of pairs (x, y), but for regression we are allowed to choose values for x, and observe the associated values for random variable y.

```
> x<-c(0,0,3,3,6,6,9,9,12,12)
> y<-c(8.5, 8.4, 7.9, 8.1, 7.8, 7.6, 7.3, 7, 6.8, 6.7)
> lm(y~x)

Call:
lm(formula = y ~ x)

Coefficients:
(Intercept)            x
     8.4600      -0.1417

> summary(lm(y~x))

Call:
lm(formula = y ~ x)

Residuals:
     Min       1Q   Median       3Q      Max
-0.18500 -0.06000  0.01500  0.05875  0.19000

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  8.460000   0.066097  128.00 1.55e-14 ***
x           -0.141667   0.008995  -15.75 2.64e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1207 on 8 degrees of freedom
Multiple R-squared:  0.9688,    Adjusted R-squared:  0.9649
F-statistic: 248.1 on 1 and 8 DF,  p-value: 2.638e-07
```

So ... Who feels the same way ?

4. ANOVA table for regression

- Up to now: What is the (best) line? Answer comes from LS-estimation.
- `How good is the fit? Answer comes from **ANOVA-table**.
  It splits observed total variation in y in two components:
  1) variation attributed to variation in *x*
  2) "error" variation attributed to chance (parameter σ)

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	3.613	1	3.613	248.069	.000[a]
	Residual	.117	8	.015		
	Total	3.729	9			

a. Predictors: (Constant), Delay (h)

b. Dependent Variable: Quality

$$\hat{\sigma}_{\varepsilon}^{2} = s_{\varepsilon}^{2} = MSE$$

$R^2 = \mathbf{SS_{Regression}} / \mathbf{SS_{Total}}$ = proportion 'explained' variation

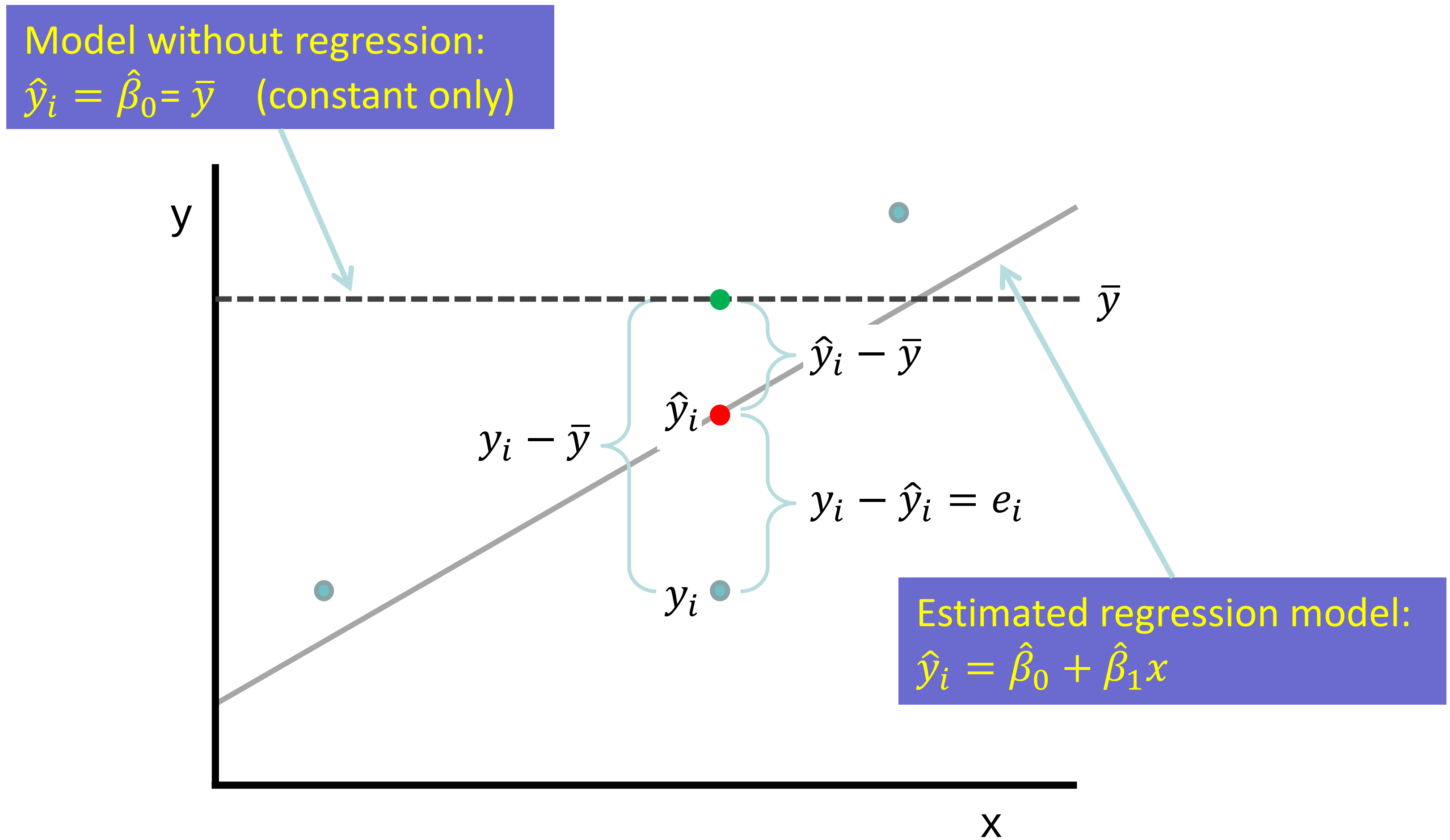
4. ANOVA table for regression

- Up to now: What is the (best) line? Answer comes from LS-estimation.
- `How good is the fit? Answer comes from **ANOVA-table**.
  It splits observed <u>total variation in y</u> in two components:
  1) systematic variation attributed to variation in *x*
  2) "error" variation attributed to chance (parameter σ)

```
> x<-c(0,0,3,3,6,6,9,9,12,12)
> y<-c(8.5, 8.4, 7.9, 8.1, 7.8, 7.6, 7.3, 7, 6.8, 6.7)
> anova(lm(y~x))
Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value    Pr(>F)
x          1 3.6125  3.6125  248.07 2.638e-07 ***
Residuals  8 0.1165  0.0146
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

R^2 = **SS**



Error for constant only: $y_i - \bar{y}$
Error for regression model: $y_i - \hat{y}_i$ → improvement given by $\hat{y}_i - \bar{y}$

$\sum(y_i - \bar{y})^2$: variation of all observations → TSS
$\sum(y_i - \hat{y}_i)^2$: variation attributed to error → SSE
$\sum(\hat{y}_i - \bar{y})^2$: variation explained by theregression model → SSR

ANOVA table for regression

The total variation in *y* (around the mean) is split into two sources: the systematic part (attributed to variation in *x*) and the random part (ε):

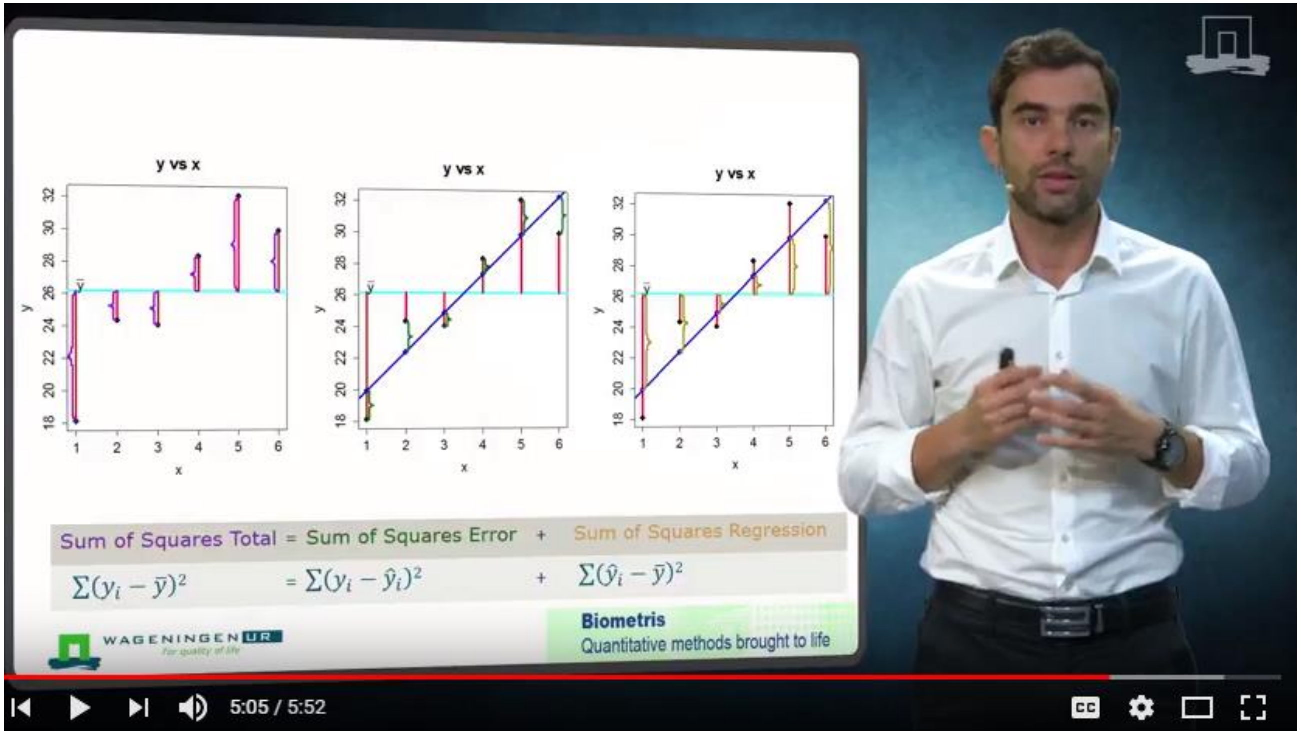
Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F
Regression	***SSR***	1	***MSR* = *SSR*/1**	*F* = ***MSR*/*MSE***
Error	***SSE***	***n*–2**	***MSE* = *SSE*/(*n*–2)**	
Total	*TSS*	*n*–1		

$$\hat{\sigma}_{\varepsilon}^{2} = s_{\varepsilon}^{2} = MSE$$

$$\begin{array}{ccccccccccc} \sum(y-\bar{y})^2 & = & \sum(y-\hat{y})^2 & + & \sum(\hat{y}-\bar{y})^2 & \Leftrightarrow & TSS & = & SSE & + & SSR \\ df_{Total} & = & df_{residual} & + & df_{regression} & \Leftrightarrow & n-1 & = & n-2 & + & 1 \end{array}$$

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

ANOVA table for regression



$$\sum (y-\bar{y})^2 = \sum (y-\hat{y})^2 + \sum (\hat{y}-\bar{y})^2 \Leftrightarrow TSS = SSE + SSR$$

$$df_{Total} = df_{residual} + df_{regression} \Leftrightarrow n-1 = n-2 + 1$$

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

F-test for regression

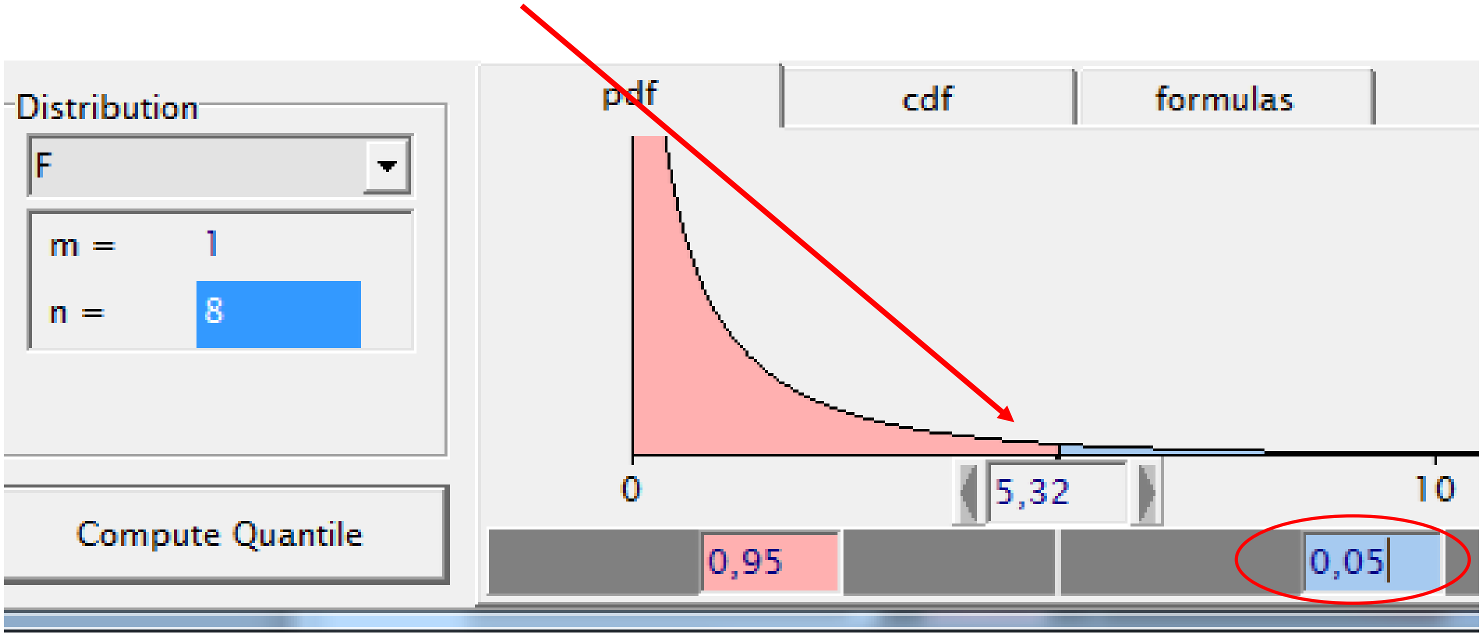
Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F
Regression	***SSR***	1	***MSR = SSR/1***	$F = MSR/MSE$
Error	***SSE***	***n*–2**	***MSE = SSE/(n–2)***	
Total	*TSS*	*n*–1		

- ***F*** compares regression mean square with residual mean square, to see if predictive value of the model (*x)* may be caused by chance alone.

- H_0: $\beta_1 = 0$, or: **model** (here: variable x) **has no predictive value** for y,
  H_a: $\beta_1 \neq 0$, or: **model** (here: variable x) **does have** predictive value
- TS: F= $MS_{Regression}$ / MS_{Error}
- Under H_0: $F \sim F(\mathbf{1}, \mathbf{n-2})$
  $df1 = df_{Regression} = 1$ (one parameter β_1 is involved) and
  $df2 = df_{Error} = (n - 2)$
  Under H_a F tends to large values, so we use RPV or right-sided RR.
- Critical values to determine RR are found in table 8. SPSS gives RPV.

F-test for regression, continued

- For the Fish storage example: n = 10, so $df1$ = 1, $df2$ = 10 – 2 = 8.
- So, RR for F: F> 5.32



- Outcome F statistic for the fish storage: 248

- NB.
  The F-test is only used for a **two-sided alternative** hypothesis H_a: $\beta_1 \neq 0$.
  For a **one-sided alternative** hypothesis, a t-test can be used.

5. Assumptions of simple linear regression model

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

Random part of the model

Systematic part μ_i of the model

- Assumptions

Random part of the model: errors ε_i are assumed:

1) independent,
2) normally distributed (with expected value 0), and
3) constant variance σ^2.

Systematic part of the model: expected value μ_i is assumed:

4) to be linearly related to x_i

5. Checking model assumptions

To check assumptions look at

residuals $$e_i = y_i - \hat{y}_i = y_i - (b_0 + b_1 x_i)$$

- Graphical checks are made, by plotting residuals in different ways:
  - Plot residuals versus expected quantiles of normal distribution to check normality assumption (check of 2): QQ – plot (Quantile – Quantile plot);
  - Plot residuals versus predicted values to check constant variance assumption (check of 3);
  - Plot residuals versus *x* to check linearity assumption (check of 4).

- Independence assumption cannot be checked by using the data.

  It should follow from a proper experimental set-up or study design.

Example fish storage, checking model assumptions

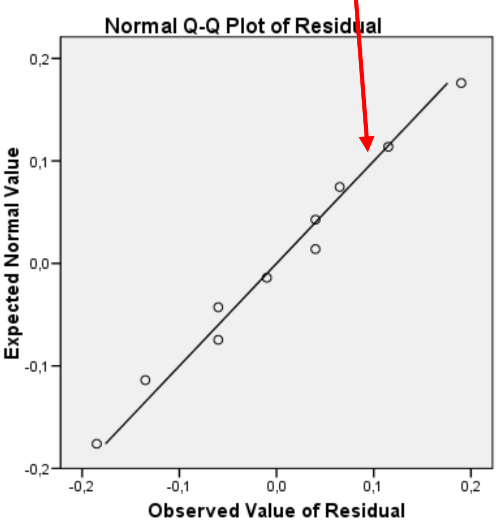
- In SPSS, store residuals and predicted values..

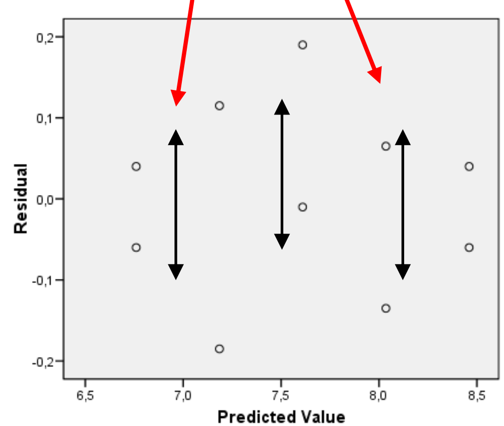
	Delay	Quality	PRE_1	RES_1
1	0	8.5	8.460	.040
2	0	8.4	8.460	-.060
3	3	7.9	8.035	-.135
4	3	8.1	8.035	.065
5	6	7.8	7.610	.190
6	6	7.6	7.610	-.010
7	9	7.3	7.185	.115
8	9	7.0	7.185	-.185
9	12	6.8	6.760	.040
10	12	6.7	6.760	-.060

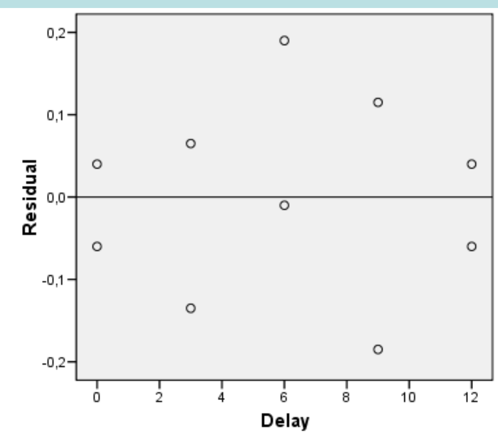
Normal QQ-plot: points approximately on straight line, so the assumption of normality is reasonable

Scatterplot of residuals on y-axis v.s. predicted values on x-axis: variation of residuals is approximately constant at different levels of the predicted value, so assumption of constant variance is reasonable.

Scatterplot of residuals (y-axis) v.s. regressor x (x-axis): residuals are approximately evenly spread around 0; they show no curve, so the assumption of a linear relationship is reasonable.







The last two plots are essentially identical, because $\hat{y} = \left(b_0 + b_1 x\right)$ and x differ only by a shift and multiplicative factor. This will change in multiple regression, later on.

6. Prediction for mean response μ_y when $x=x^*$

- simple linear regression model: $y = \mu_y + \varepsilon = \beta_0 + \beta_1 x + \varepsilon$
- Mean response at a specific level x^* is

$$\mu_y = \beta_0 + \beta_1 x^*$$

- Estimated mean response and standard error (replacing unknown β_0 and β_1 with estimates):

$$\hat{\mu}_y = \hat{\beta}_0 + \hat{\beta}_1 x^*, \quad se(\hat{\mu}_y) = s_\varepsilon \sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}}$$

- Confidence interval for mean response at x^*:

$$\left(\hat{\mu}_y \pm t_{\alpha/2, n-2}\, se(\hat{\mu}_y)\right)$$

6. Prediction for future individual response when $x = x^*$

- (Unknown) response at a specific level x^* is

$$y_{x^*} = \mu_y + \varepsilon = \beta_0 + \beta_1 x^* + \varepsilon$$

- Predicted individual response

(replacing β_0 and β_1 by estimates, and replacing ε by its expected value 0):

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

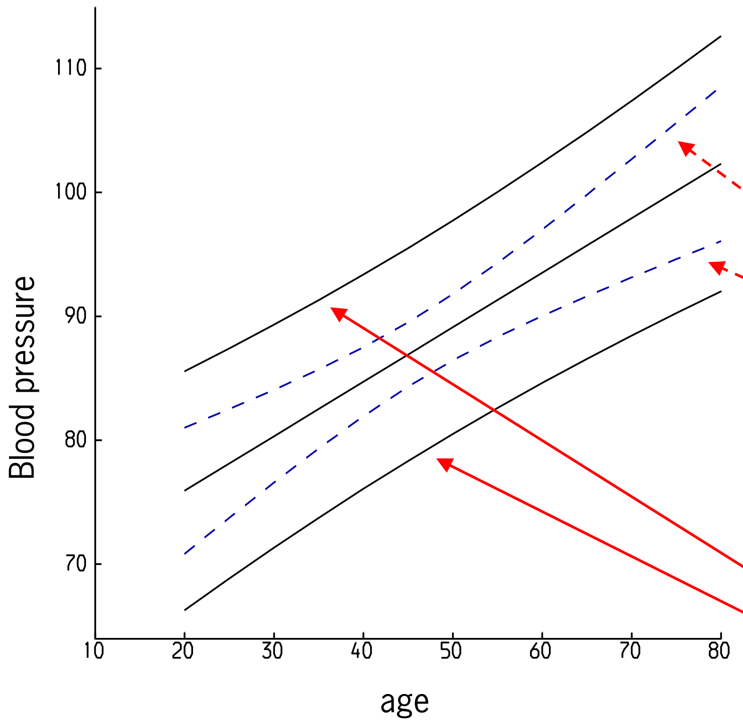
the same as the estimated mean response on the previous slide

- Prediction interval for future individual response

$$\left(\hat{y}_{x^*} \pm t_{\alpha/2,n-2}\ se(\hat{y}_{x^*})\right) = \left(\hat{y}_{x^*} \pm t_{\alpha/2,n-2}\, s_\varepsilon \sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}}\right)$$

the extra term 1, compared to se of estimated mean response, is due to the extra ε *in observation y*

The two intervals in one plot

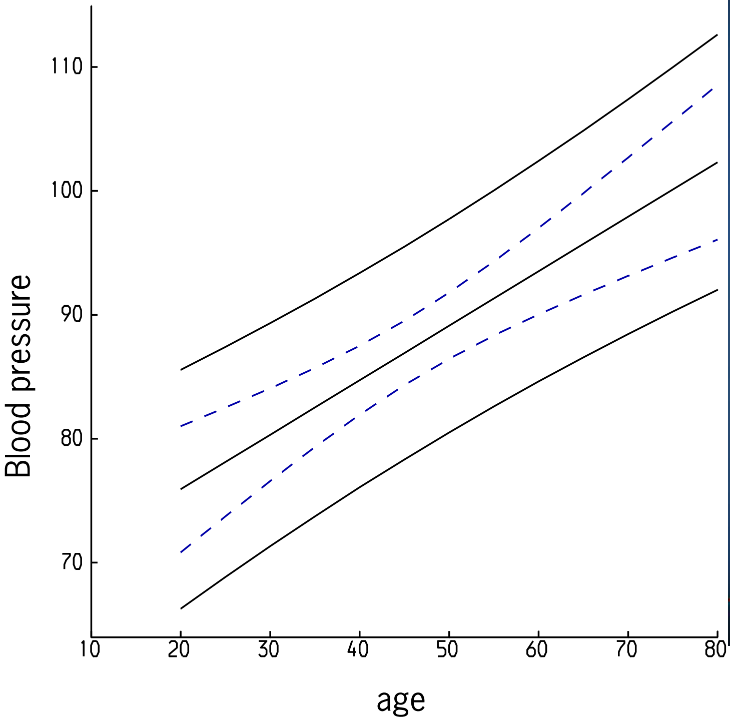


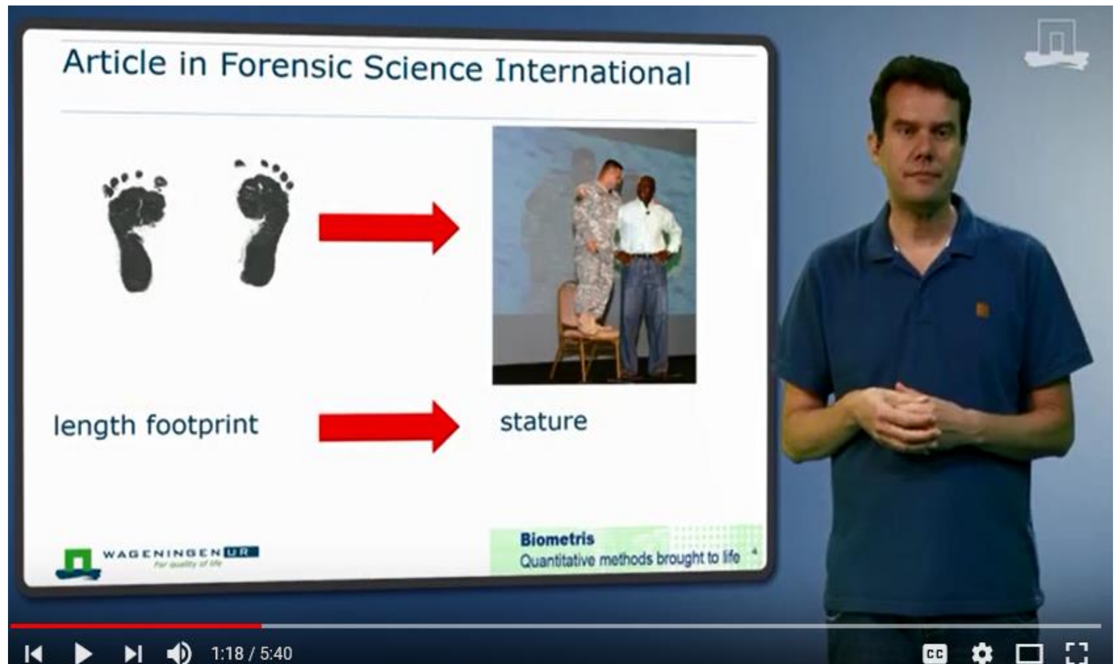
LS regression line

95% confidence limits for estimated mean response

95% prediction limits for predicted individual response; much wider interval!

The two intervals in one plot





Fish storage, continued SPSS output

x = delay (h) of fish storage in ice,

y = quality after subsequent 7-day storage in ice.

- estimate μ_y for delay x =7 (h) with associated se
- predict y if delay $x = 7$ (h)
- give 0.95-confidence interval for μ_y.
- give 0.95 prediction interval for y

which interval will be narrower?

Model: $y = \beta_0 + \beta_1 x + \varepsilon$,

$$\mu_y = \beta_0 + \beta_1 x$$

Two ways to proceed:

Hard way: fill in $x = 7$ in regression equation, calculate standard error and interval.

Easy way: let SPSS do the work:

(1) add an extra line $x = 7$ to the data

(2)in menu Regression ask for needed quantities and use Save

(3)interpret output in datafile

	Delay	Quality	PRE_1	SEP_1	LMCI_1	UMCI_1	LICI_1	UICI_1
1	.0	8.5	8.46	.066	8.31	8.61	8.14	8.78
2	.0	8.4	8.46	.066	8.31	8.61	8.14	8.78
3	3.0	7.9	8.04	.047	7.93	8.14	7.74	8.33
4	3.0	8.1	8.04	.047	7.93	8.14	7.74	8.33
5	6.0	7.8	7.61	.038	7.52	7.70	7.32	7.90
6	6.0	7.6	7.61	.038	7.52	7.70	7.32	7.90
7	9.0	7.3	7.19	.047	7.08	7.29	6.89	7.48
8	9.0	7.0	7.19	.047	7.08	7.29	6.89	7.48
9	12.0	6.8	6.76	.066	6.61	6.91	6.44	7.08
10	12.0	6.7	6.76	.066	6.61	6.91	6.44	7.08
11	7.0	.	7.47	.039	7.38	7.56	7.18	7.76

Example fish storage in ice, continued

	Delay	Quality	PRE_1	SEP_1	LMCI_1	UMCI_1	LICI_1	UICI_1
1	.0	8.5	8.46	.066	8.31	8.61	8.14	8.78
2	.0	8.4	8.46	.066	8.31	8.61	8.14	8.78
3	3.0	7.9	8.04	.047	7.93	8.14	7.74	8.33
4	3.0	8.1	8.04	.047	7.93	8.14	7.74	8.33
5	6.0	7.8	7.61	.038	7.52	7.70	7.32	7.90
6	6.0	7.6	7.61	.038	7.52	7.70	7.32	7.90
7	9.0	7.3	7.19	.047	7.08	7.29	6.89	7.48
8	9.0	7.0	7.19	.047	7.08	7.29	6.89	7.48
9	12.0	6.8	6.76	.066	6.61	6.91	6.44	7.08
10	12.0	6.7	6.76	.066	6.61	6.91	6.44	7.08
11	7.0	.	7.47	.039	7.38	7.56	7.18	7.76

5. 0.95-pred. int. of quality of an individual fish at delay of 7 h:

$$(\text{LICI\_1},\text{UICI\_1}) = \\ = \hat{y}_{x=7} \pm t_8(0.975)S\hat{E}(\hat{y}_{x=7}) = \\ = (7.18,\ 7.76)$$

1. Estimated mean quality of a fish at a delay of 7 h:

$$\text{PRE\_1} = \hat{\mu}_{y|x=7} = b_0 + b_1 \times 7 = 7.47$$

4. 0.95-conf. int. of mean quality at delay of 7 h:

$$(\text{LMCI\_1},\text{UMCI\_1}) = \hat{\mu}_{y|x=7} \pm t_8(0.975)S\hat{E}(\hat{\mu}_{y|x=7}) = \\ = 7.47 \pm 2.31 \times 0.039 = (7.38,\ 7.56)$$

2. Also predicted quality of individual fish at delay of 7 h:

$$\text{PRE\_1} = \hat{y}_{x=7} = b_0 + b_1 \times 7 + \hat{e} = 7.47 + 0 = 7.47$$

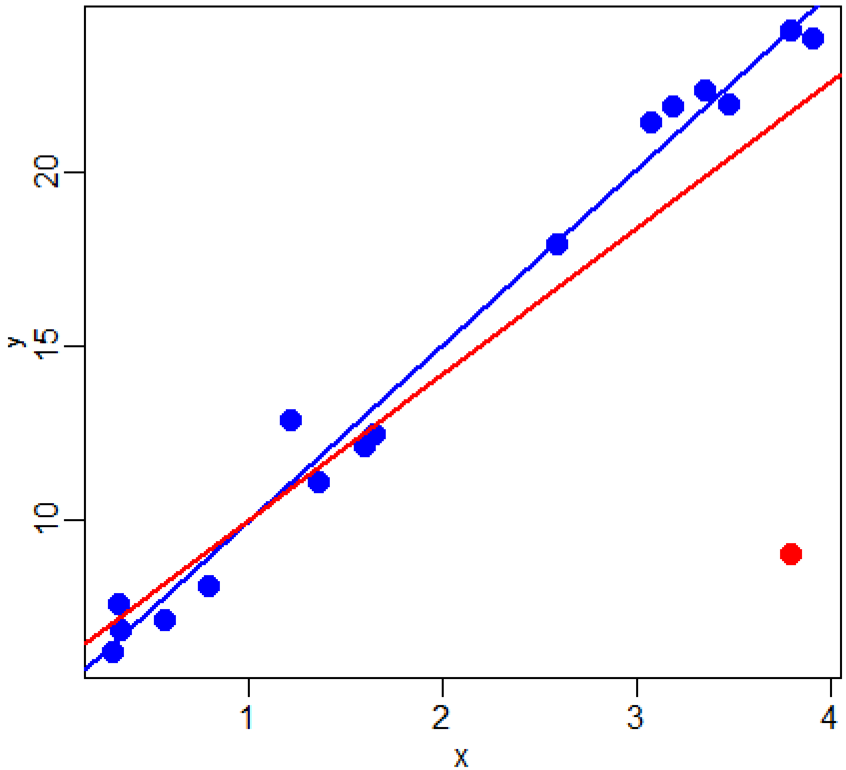
Same as estimated mean response!

3. Standard error of estimator of mean quality at delay of 7 h:

$$\text{SEP\_1} = S\hat{E}(\hat{\mu}_{y|x=7}) = s_\varepsilon \sqrt{\tfrac{1}{10} + \tfrac{(7-\bar{x})^2}{S_{xx}}} = 0.039$$

Outlier, leverage and influence

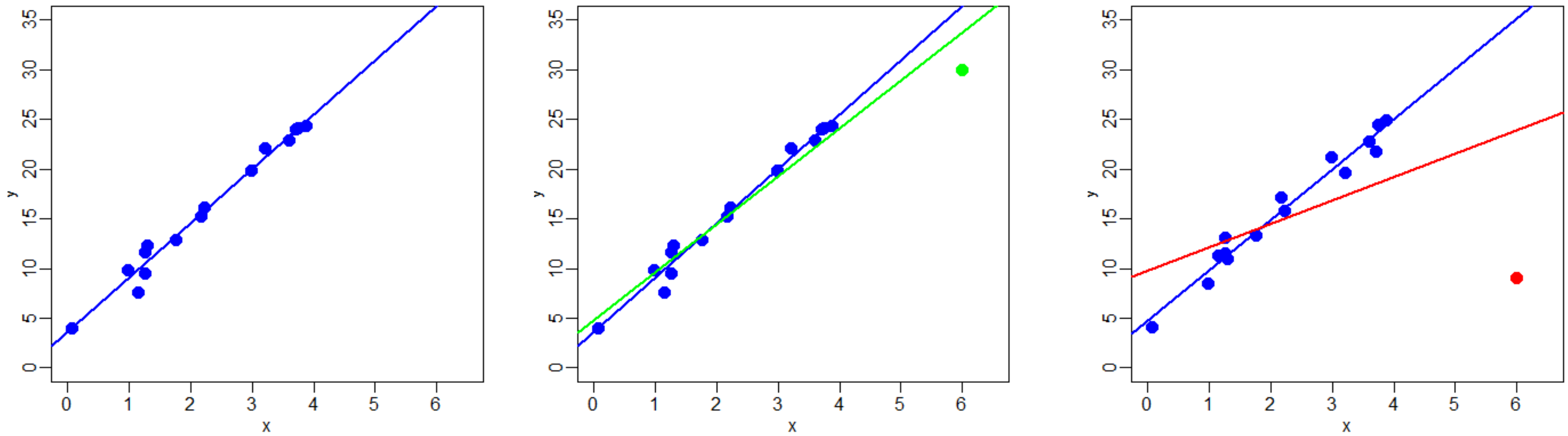
Outlier: observation with extreme y-value (compared to other observations with similar x-values)



Outlier, leverage and influence

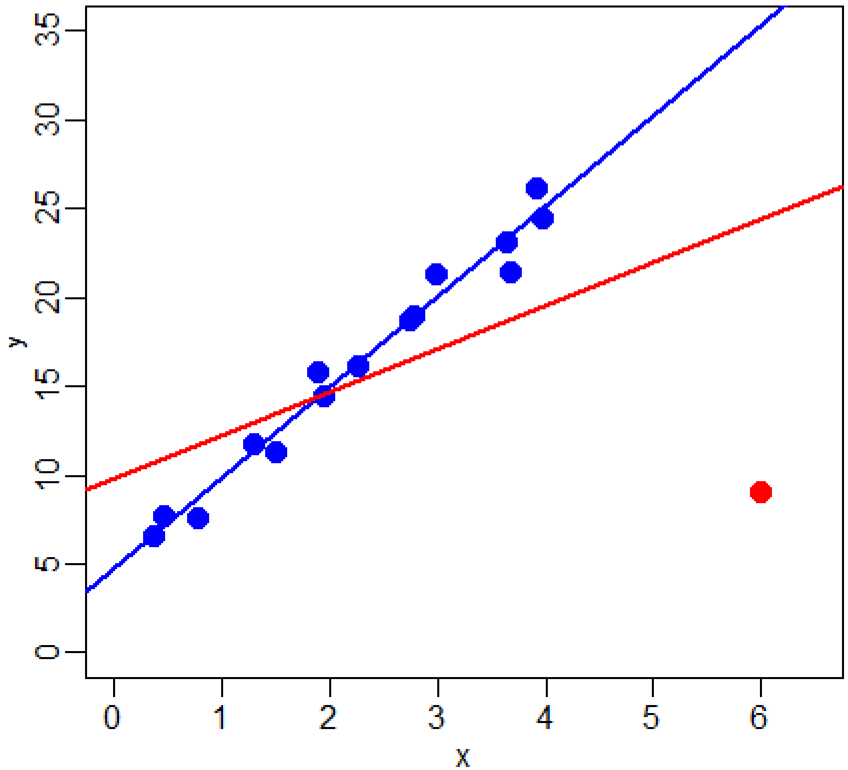
High leverage point: observation with extreme x-value(s).

May influence estimated coefficient(s).



Outlier, leverage and influence

Influential point: observation that strongly influences estimated regression coefficients(s).



Perform an analysis with and without the suspect observation(s) and see how much it matters for the conclusions.