MAT20306 - Advanced Statistics

Lecture 6: Multiple linear regression

WAGENINGEN UR
For quality of life

Biometris
Quantitative Methods brought to 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)

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

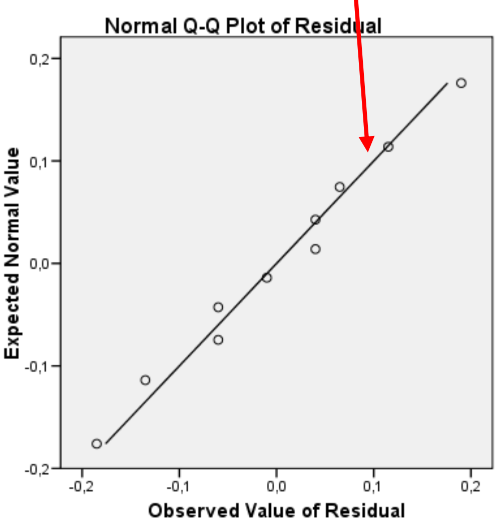
- In SPSS, store residuals and predicted values..

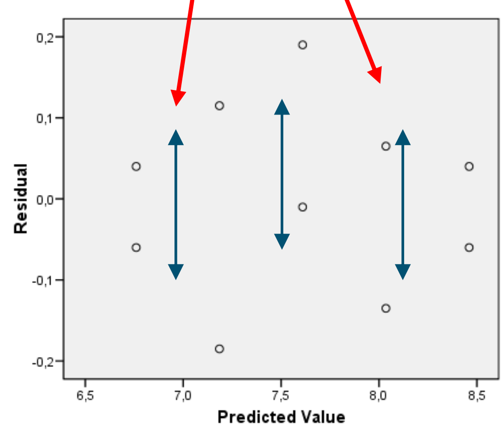
	Delay	Quality	PRE_1	RES_1
1	0	8.5	8.460	.040
2	0	8.4	8.460	-.060
3	3	7.9	8.035	-.135
4	3	8.1	8.035	.065
5	6	7.8	7.610	.190
6	6	7.6	7.610	-.010
7	9	7.3	7.185	.115
8	9	7.0	7.185	-.185
9	12	6.8	6.760	.040
10	12	6.7	6.760	-.060

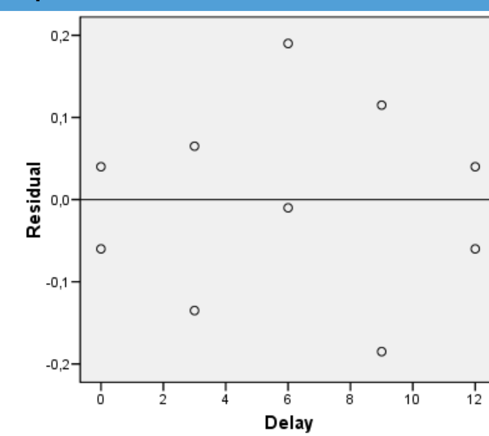
Normal QQ-plot: points approximately on straight line, so the assumption of normality is reasonable

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.

Inference for mean response μ_y when $x=x^*$

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

$$E(y \mid x^*) = \mu_y = \beta_0 + \beta_1 x^*$$

- 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^*:

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

WAGENINGEN UR
For quality of life

Inference 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

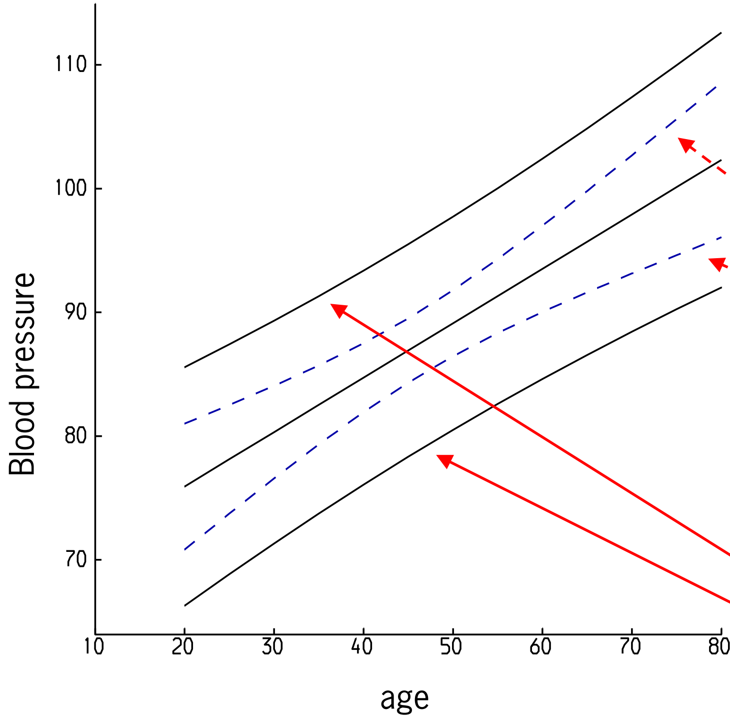
- Prediction interval for future individual response

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

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

WAGENINGEN UR
For quality of life

The two intervals in one plot

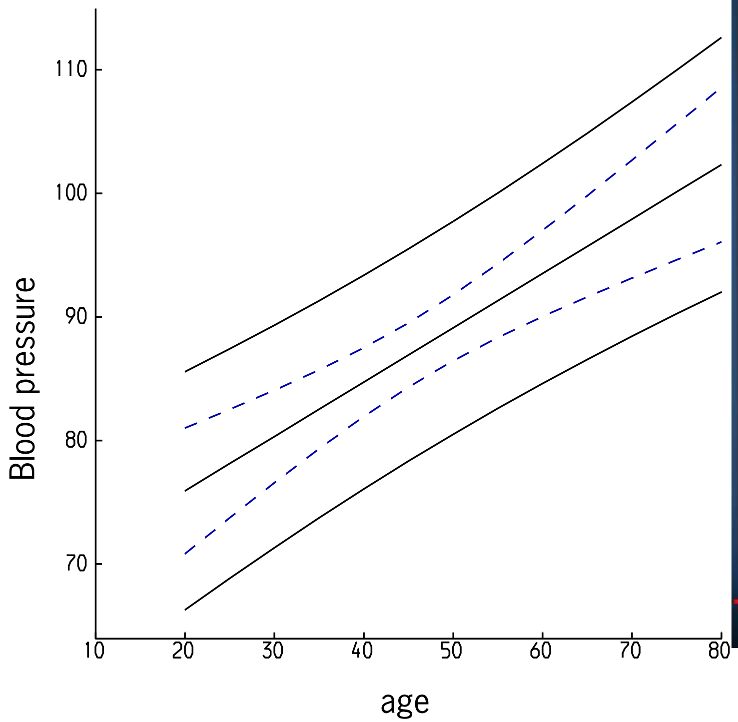


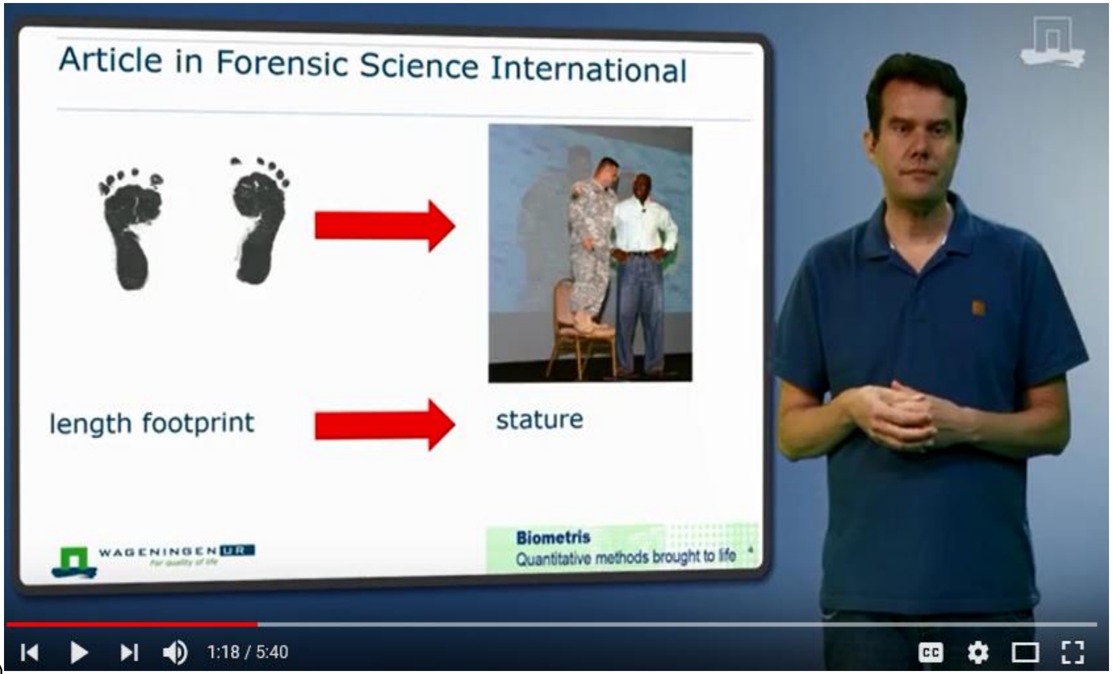
LS regression line

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

WAGENINGEN UR
For quality of life

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:

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

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:

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

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!

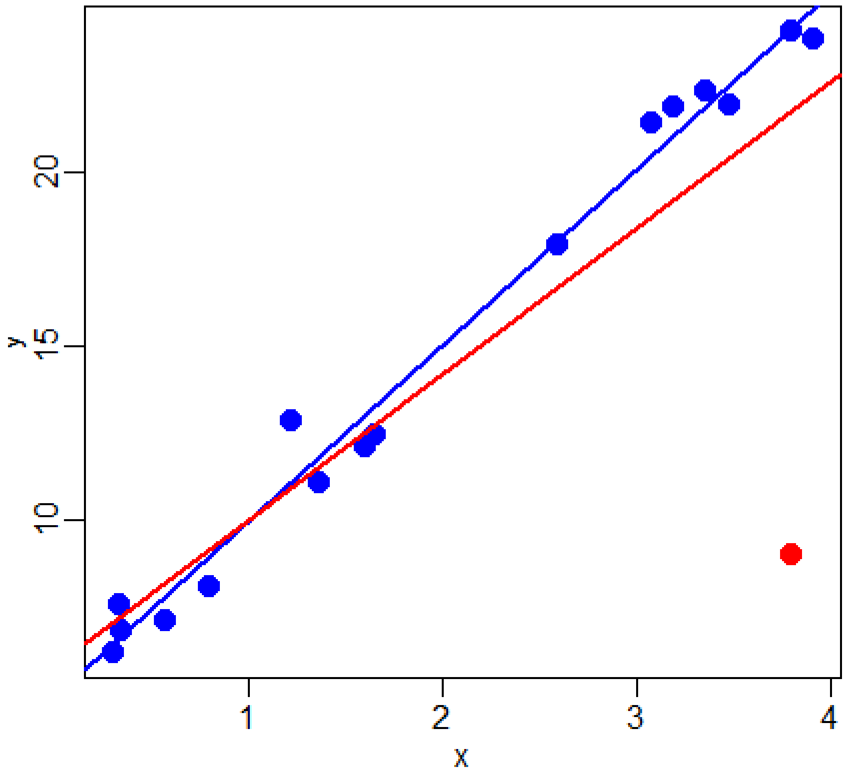
3. Standard error of estimator of mean quality at delay of 7 h:

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

WAGENINGEN UR
For quality of life

Outlier, leverage and influence

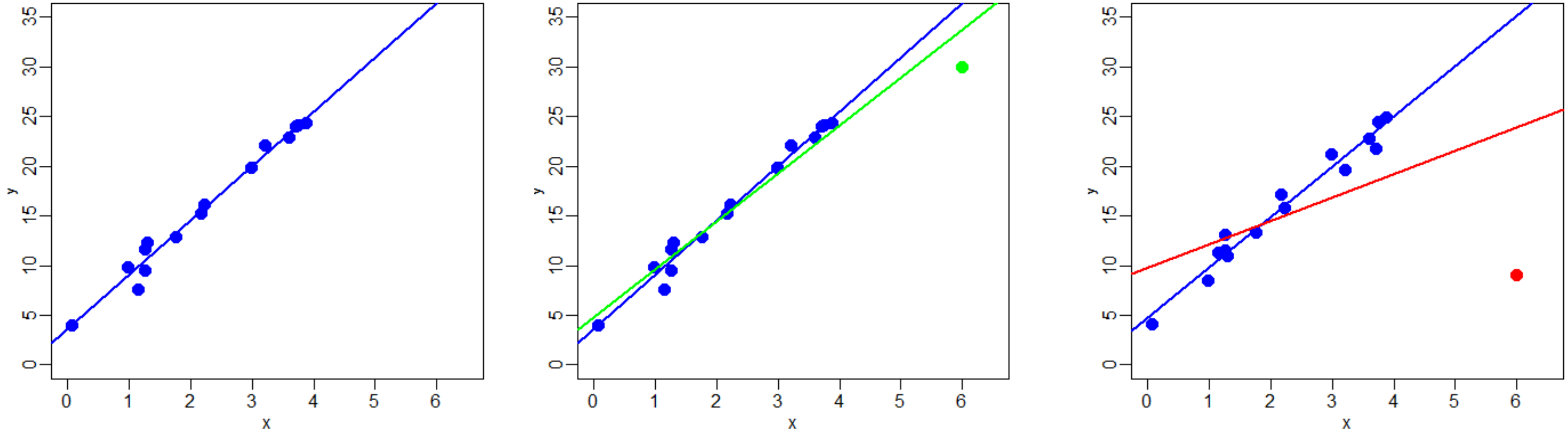
Outlier: observation with extreme y-value (compared to other observations with similar x-values)



Outlier, leverage and influence

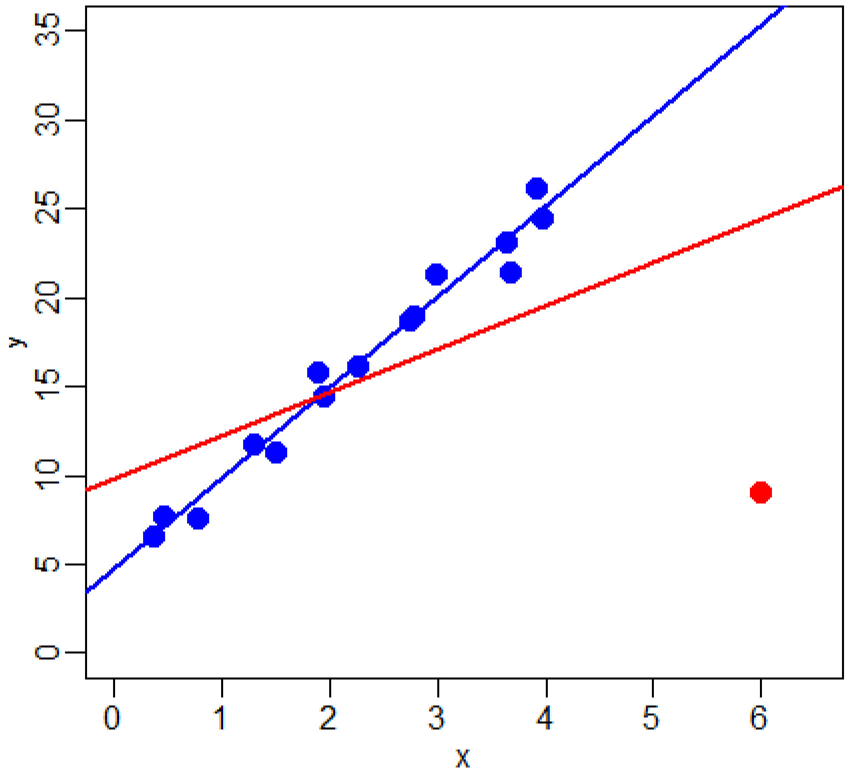
High leverage point: observation with extreme x-value(s).

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.

Multiple Linear Regression

Overview:

1) Define the model: - $\beta_0, \beta_1, \beta_2, \ldots, \sigma_\varepsilon$
2) Estimate the model
3) Test the model : ANOVA table
4) Inference on model parameters (by means of t-test and C.I.)
5) Checking model assumptions
6) Prediction by using the model

O&L Sections 12.1, 12.3, 12.4, 12.5 (12.9)

Example Weight Loss of Compound (Example 12.5, p676)

A compound is exposed to air for 4 different exposure times, and 3 different levels of humidity.

Both exposure time and relative humidity are experimental factors, with values fixed by the experimenter, and weight loss is measured.

Response: y = weight loss

Two explanatory variables:

x_1 = exposure time (hours)

x_2 = relative humidity

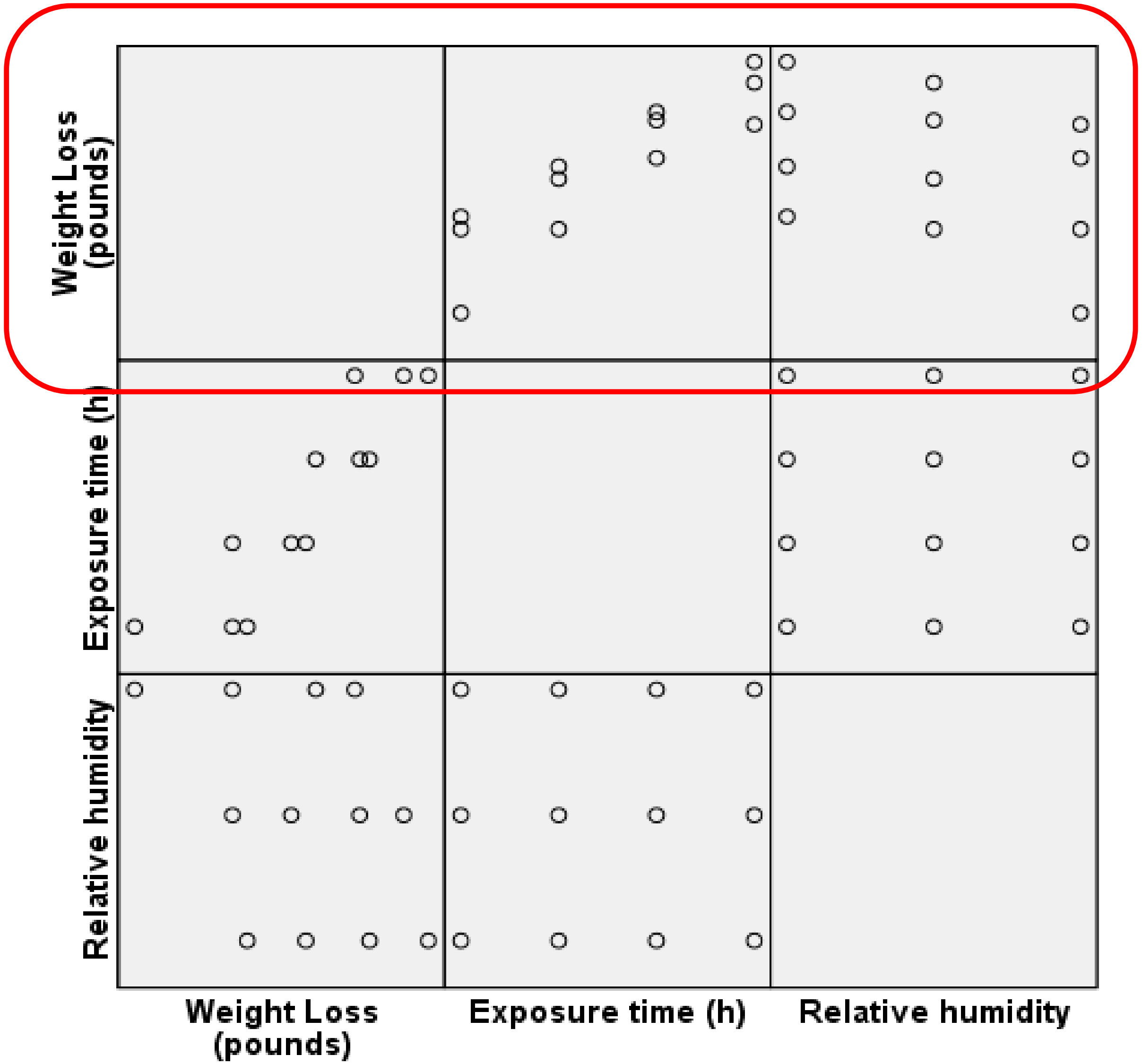
Weight_loss (y_i)	Time (x_{1i})	Humidity (x_{2i})
4.3	4	.20
5.5	5	.20
6.8	6	.20
8.0	7	.20
4.0	4	.30
5.2	5	.30
6.6	6	.30
7.5	7	.30
2.0	4	.40
4.0	5	.40
5.7	6	.40
6.5	7	.40

Model:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i,$$

ε_i's iid from $N(0, \sigma_\varepsilon)$, $i = 1 \ldots 12$

Example Weight Loss of a Compound, continued



Scatterplot matrix

1. Multiple linear regression model

Simple linear regression model

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \qquad \varepsilon_i\text{'s iid from } N(0, \sigma_\varepsilon),\ \ i = 1,..,n$$

Multiple linear regression model:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_k x_{ki} + \varepsilon_i, \quad \varepsilon_i\text{'s iid from } N(0, \sigma_\varepsilon)$$

Multiple Linear regression is **like simple** Linear Regression in several ways:

- In both cases there is **one numerical response** variable *y* which is explained by a *systematic* part and a *random* part.

- **Same assumptions** for *random* part (error terms ε_i):
  1. independent,
  2. normally distributed with mean 0, and
  3. constant variance σ^2_ε
- The *systematic* part is **linear in the parameters**, e.g.

  $\mu_{yi} = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}$

Multiple versus simple linear regression

Multiple linear regression is unlike simple linear regression, regarding:

- Interpretation of β_j :
  - effect on the mean response of increasing the j th regressor by 1 unit, keeping all other regressors constant.
  - β_j is called a partial regression coefficient.

- Problem of collinearity: two or more x-variables may be (strongly) correlated, which makes it difficult to separate the effects of these x-variables.

Example Weight Loss of a Compound, continued

- y (weight loss) seems to depend on x_1 (exposure time) and x_2 (relative humidity). A possible (first) model is:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

- Other possible model is e.g.:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \varepsilon$$

- Choosing a proper model can be difficult.
- Good practice of model building is to be led by:
  1. existing theory (knowledge),
  2. graphical summaries, and
  3. the principle of parsimony: keep the model as simple as possible, yet capturing the essence .

2. Least Squares again

- How do we get estimates for $\beta_0, \beta_1, \ldots, \beta_k$?

- We use residuals: $e_i = y_i - (b_0 + b_1 x_{1i} + \ldots + b_k x_{ki})$

- Least Squares: find $b_0 \ldots b_k$ such that $SSE = \Sigma_i\, e_i^2$ is minimal.
  - Regard SSE as a function of unknown parameters $b_0, b_1, \ldots, b_k$.
  - Set derivatives w.r.t. $b_0, b_1, \ldots, b_k$ equal to 0.
  - This yields (k+1) equations, with (k+1) unknown parameters, from which the LS estimates $b_0, b_1, \ldots, b_k$ can be solved.

- These are called the Normal Equations (see 12.3 O&L).

Weight loss and SPSS

- Let SPSS solve the normal equations...
- β_0, β_1, and β_2 estimated by the Least Squares method.

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	.667	.694		.960	.362
	Exposure time (h)	1.317	.100	.895	13.191	.000
	Relative humidity	-8.000	1.367	-.397	-5.853	.000

a. Dependent Variable: Weight Loss (pounds)

Fitted model for the weight loss example:

$$\hat{y} = 0.667 + 1.317\, x_1 - 8.00\ x_2$$

Weight loss and SPSS

- Let R solve the normal equations...

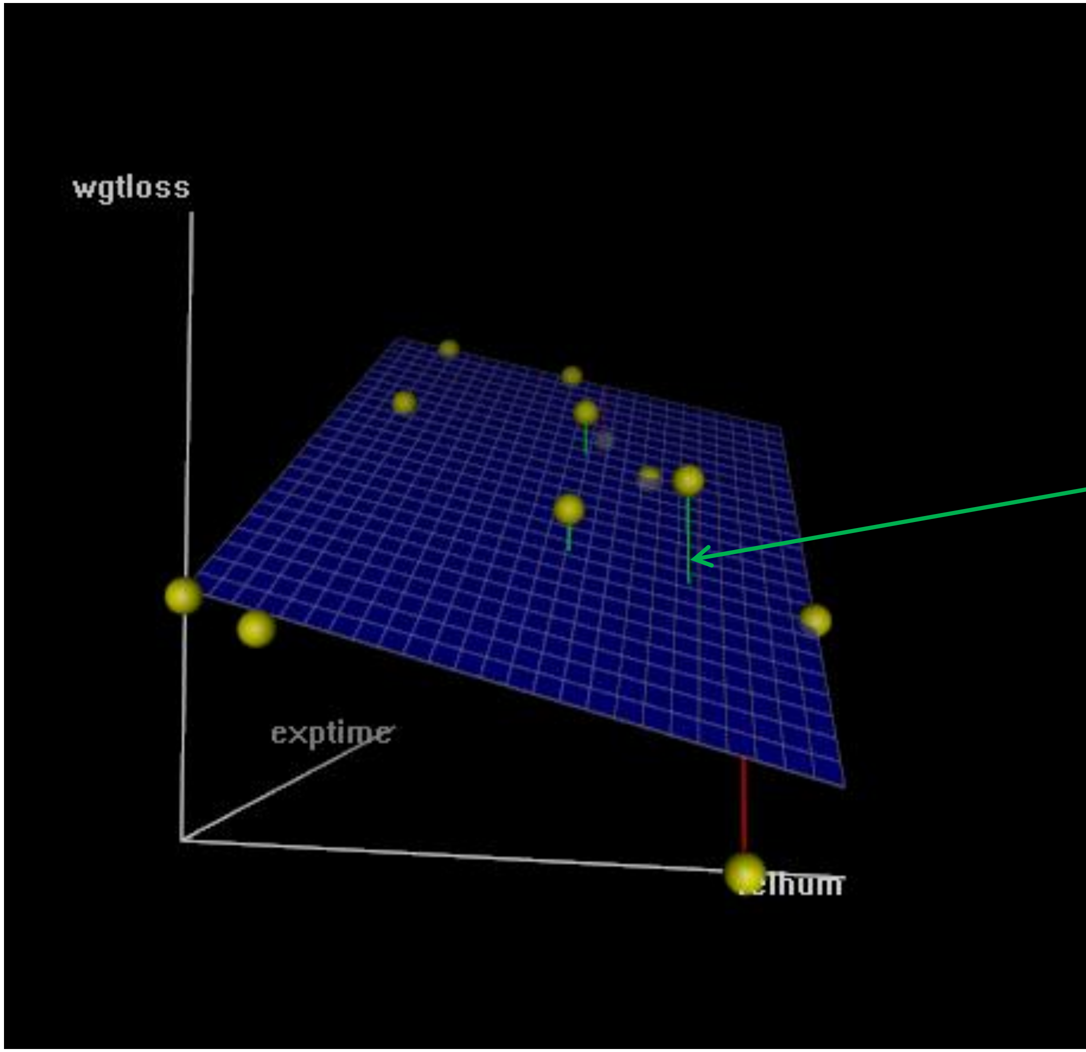
- β_0, β_1, and β_2 estimated by the Least Squares method.

```
> y<-c(4.3, 5.5, 6.8, 8, 4, 5.2, 6.6, 7.5, 2, 4, 5.7, 6.5)
> x1<-c(4, 5, 6, 7, 4, 5, 6, 7, 4, 5, 6, 7)
> x2<-c(0.2, 0.2, 0.2, 0.2, 0.3, 0.3, 0.3, 0.3, 0.4, 0.4, 0.4, 0.4)
> coefficients(lm(y ~ x1 + x2))
(Intercept)          x1          x2
  0.6666667   1.3166667  -8.0000000
```

Fitted model for the weight loss example:

$$\hat{y} = 0.667 + 1.317\, x_1 - 8.00\, x_2$$

The normal equations



In simple regression, residuals e were distances to a fitted line.

In multiple regression residuals e are distances to a fitted plane.

ANOVA table and σ_ε

- Again, ANOVA table shows how the total variation around the mean (TSS) is split into variation due to the systematic part (SSR) and the random part (SSE), of the model:

$$TSS = SSR + SSE$$

$SSRegr$ SSE TSS

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	31.124	2	15.562	104.133	.000[a]
	Residual	1.345	9	.149		
	Total	32.469	11			

a. Predictors: (Constant), humid, time

b. Dependent Variable: wt_loss

$SSRegr$ SSE

```
> anova(lm(y ~ x1 + x2))
Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value    Pr(>F)
x1         1 26.004 26.0042  174.01 3.425e-07 ***
x2         1  5.120  5.1200   34.26 0.0002428 ***
Residuals  9  1.345  0.1494
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- df's are:

$df_{Regression}$ = k = 2 (two slopes, instead of one slope),

df_{Total} = (n – 1) = 12 – 1 = 11 (same as before),

df_{Error} = *remaining df = 11- 2 = 9* (the difference).

Estimating the Standard Deviation of the Error

- Residuals

$$e_i = y_i - \hat{y}_i = y_i - (b_0 + b_1 x_{i1} + \Lambda + b_k x_{ik})$$

- Residual (error) sum of squares:

$$SSE = \sum e_i^2 = \sum \left(y_i - (b_0 + b_1 x_{i1} + \Lambda + b_k x_{ik})\right)^2$$

- Residual standard deviation: similar to simple linear regression, but notice residual (error) degrees of freedom $df_E = n - (k+1)$:

$$\hat{\sigma}_\varepsilon = s_\varepsilon = \sqrt{MSE} = \sqrt{SSE / df_E}$$

SPSS calls s_ε the "Standard Error of the Estimate", we call it: "residual (or error) standard deviation".

MSE can be found in the ANOVA table

Weight loss, continued

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	31.124	2	15.562	104.133	.000[a]
	Residual	1.345	9	.149		
	Total	32.469	11			

a. Predictors: (Constant), humid, time

b. Dependent Variable: wt_loss

```
> anova(lm(y ~ x1 + x2))
Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value    Pr(>F)
x1         1 26.004 26.0042  174.01 3.425e-07 ***
x2         1  5.120  5.1200   34.26 0.0002428 ***
Residuals  9  1.345  0.1494
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

$$\hat{\sigma}_\varepsilon^2 = s_\varepsilon^2 = MSE = SSE / df_E = SSE / \big(n - (k+1)\big) = 0.149$$

$$\hat{\sigma}_\varepsilon = s_\varepsilon = \sqrt{0.149} = 0.386.$$

3. Does the model have any predictive value?

- **Do the *x*-variables together have any predictive value?**
- H_0: $\beta_1 = 0$ and $\beta_2 = 0$ and … and $\beta_k = 0$ (H$_0$: no predictive value)
  H_a: at least one $\beta_j \neq 0$ (j =1,..,k).
- Test statistic:

$$F = \frac{\mathrm{MSRegr}}{MSE} = \frac{\mathrm{SSReg}/k}{SSE\,/\,dfE}$$

- Under H_0, test statistic F follows an F-distribution with $df_1 = k$, $df_2 = dfE$
- Large values are critical. RR for F is given by: F > F(k, dfE, 0.05). With output we prefer to use RPV.
- Output for this F-test can be found in the ANOVA-table.

Weight loss, continued

- **Research question:** Does the model have any predictive value? Or : Do time or humidity or both have predictive value? Use α =0.05.

1) H_0: $\beta_1 = 0$ ánd $\beta_2 = 0$ versus H_a: at least one $\beta_j \neq 0$, $j = 1, 2$.

2) TS: F= MSRegr / MSE

3) Under H_0, F ~ F(2,9) 4/5) Use RPV

6) Outcome $F = 104.133$.

7) P-value = $P(F \geq 104.133) = 0.000 < 0.05$, so H_0 is rejected, H_a is proven. It is shown that the model has predictive value.

Can we see in the ANOVA table which variables have predictive value?

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	31.124	2	15.562	104.133	.000[a]
	Residual	1.345	9	.149		
	Total	32.469	11			

a. Predictors: (Constant), humid, time

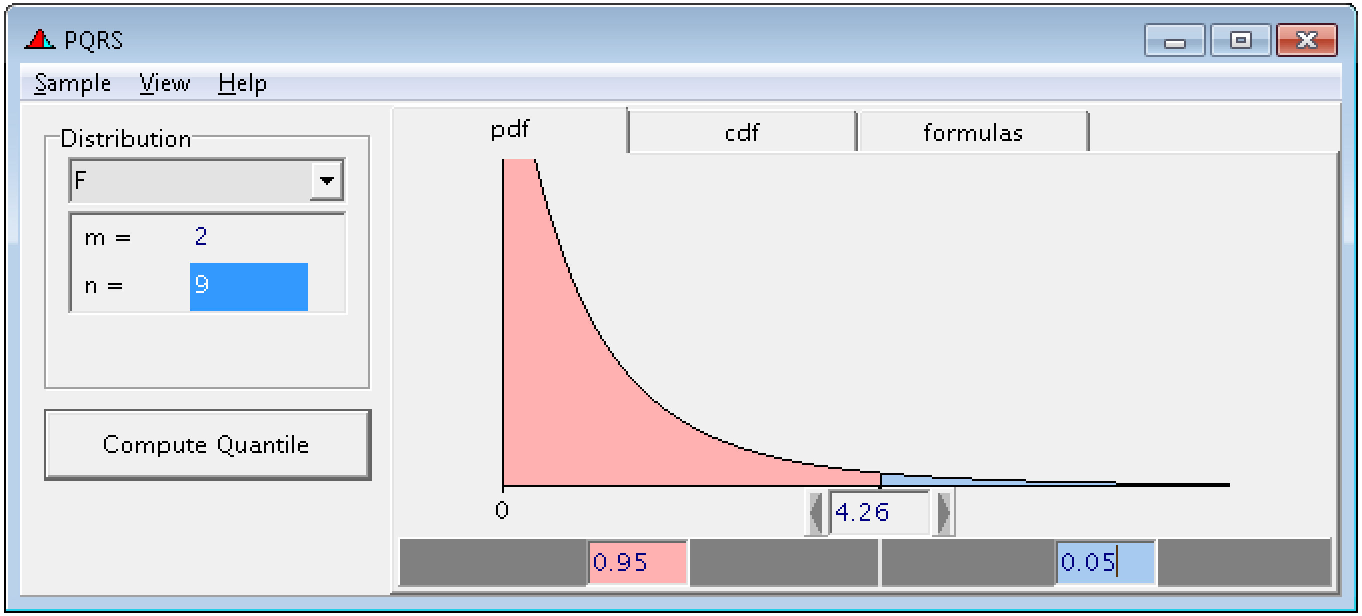
b. Dependent Variable: wt_loss

```
> anova(lm(y ~ x1 + x2))
Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value    Pr(>F)
x1         1 26.004 26.0042  174.01 3.425e-07 ***
x2         1  5.120  5.1200   34.26 0.0002428 ***
Residuals  9  1.345  0.1494
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Critical region for F-test

Or step 4/5: Use right-sided RR: **F > F(2,9, 0.05) = 4.26.**
You can find it using e.g. PQRS or table 8 (page 1181)



Or: Rejection Region or Critical region for F is: $(4.26, \infty)$

Residual Standard Deviation s_ε and "raw" standard deviation s_y

- Residual standard deviation s_ε is a measure of variability of y around it's expectation $\mu = \beta_0 + \beta_1 x_1 + \beta_2 x_2$ + ... from the regression.
- It is an absolute measure how good the regression model explains the variation in y: the smaller the better.
- s_y is the (ordinary) standard deviation of y, around estimate $\bar{y}$ for μ, and not around the estimate for $\mu = \beta_0 + \beta_1 x_1 + \beta_2 x_2$ + ...
- So, s_y is just an estimate for the standard deviation of y, as if you did not know the values of x_1, x_2, ...
- If s_ε is much smaller than s_y, apparently the x_j's help to obtain a better prediction for y, than the sample mean $\bar{y}$.

Weight Loss: compare some models by s_ε

Let's compare some models by s_ε :

1. Only intercept: $\mu_y = \beta_0$
2. $\mu_y = \beta_0 + \beta_1 x_1$
3. $\mu_y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$

Descriptive Statistics

Model: $\mu_y = \beta_0$

	Min.	Max.	Mean	Std. Dev.	Vari-ance
Weight Loss	2.00	8.00	5.5083	1.7181	2.952

For intercept-only model: ordinary standard deviation of y = residual standard deviation of y !

Model: $\mu_y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.979[a]	.959	.949	.38658

a. Predictors: (Constant), Relative humidity, Exposure time (h)

Model: $\mu_y = \beta_0 + \beta_1 x_1$

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.895[a]	.801	.781	.80405

a. Predictors: (Constant), Exposure time (h)

Extra sums of squares

- If a regressor *x* enters a regression model, the *SSE* will *decrease* and the *SSR* will *increase* with the same amount.
- Increase in *SSR* = decrease in *SSE* = *extra sum of squares* due to entering *x* into a given model.

- Starting model with *exposure time* only:

$$SSE_1 = 6.465$$

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	26.004	1	26.004	40.223	.000[a]
	Residual	6.465	10	.647		
	Total	32.469	11			

a. Predictors: (Constant), time

b. Dependent Variable: wt_loss

- Now add humidity:

$$SSE_2 = 1.345$$

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	31.124	2	15.562	104.133	.000[a]
	Residual	1.345	9	.149		
	Total	32.469	11			

a. Predictors: (Constant), humid, time

b. Dependent Variable: wt_loss

- Extra SS of *humidity* (after *exposure time*) = 6.465–1.345=**5.120** (= 31.124 –26.004)

- Generally, the extra sum of squares depends on the order of model terms, e.g. SS of x_1 first and x_2 after x_1 is generally not the same as SS for x_2 and SS for x_1 after x_2.

Coefficient of Determination R^2, judging the fit of a model

- ANOVA table: $TSS = SSR + SSE$

SSR → Regression Sum of Squares; TSS → Total Sum of Squares

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	31.124	2	15.562	104.133	.000[a]
	Residual	1.345	9	.149		
	Total	32.469	11			

a. Predictors: (Constant), humid, time

b. Dependent Variable: wt_loss

- R^2 is proportion of total variation in y–values (TSS) accounted for by the systematic part of the model (SSR), or

  R^2 = proportion variation in y "explained" by the variation in x–variables:

$$R^2 = \frac{SSR}{TSS} \quad \left(= 1 - \frac{SSE}{TSS}\right)$$

$$R^2 = \frac{31.234}{32.469} = 0.958$$

- 96% of variation in weight loss is "explained" by the variation in relative humidity and exposure time.

Some properties of R^2

- The higher R^2, the better the model fits the data.
- R^2 has values between 0 and 1.
  - Value 0 means $SSR= 0$, i.e. model explains nothing (more than intercept already does).
  - Value 1 means $SSR=TSS$, i.e. the regression model explains all variation of y.
- In simple linear regression R^2 is the square of the correlation coefficient r of y and (single) x.
- In multiple linear regression we have multiple x's, and R^2 equals the square of the correlation coefficient of y and predicted values $\hat{y}$ (called the multiple correlation coefficient).
- Compare R^2 values of different models only on the same data.
  - As in simple regression, when values for x-variables are chosen over a wider range (if this is possible in the design stage), R^2 will increase –
  - dependent on the design

.

R^2 and Adjusted R^2 (O&L, 13.2)

- R^2 always increases with an extra x-variable in the model, even if x is unimportant.
- Ideally, the preferred model should be simple / small and fit well.
- We want to take the number k of x-variables (= number of unknown β coefficients) into account as well, and compromise between "parsimony" and "fit".
- This can be done with the adjusted R^2 or R^2_{adj} :

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

$$R^2_{adj} = 1 - \frac{SSE/(n-(k+1))}{TSS/(n-1)} = 1 - \frac{MSE}{MST} = 1 - \frac{s_\varepsilon^2}{s_y^2}$$

- R^2_{adj} will *not* automatically increase with an extra x-variable.
- It will *only* increase, when the error mean square $MSE = SSE/df_E = s_\varepsilon{}^2$ (estimator for the residual variance) decreases.

4.1 Inference for a single regression coefficient

- 6.1 Hypothesis test for β_j , e.g. H_0: $\beta_j = 0$
- Test Statistic: $$t = \frac{\hat{\beta}_j - 0}{se(\hat{\beta}_j)}$$

This zero can be any value, it is the value under the H_0

Under H_0 $t \sim t_{dfE}$ (with dfE = $n - k - 1$)

- Meaning of "$\beta_3 = 0$" : "Including x_3 after the other x's does not improve the model" or

Keeping x_1, x_2 ... constant / corrected for effects of changes in x_1,x_2 ... there's no association of μ_y with x_3 / x_3 has no effect on μ_y.

Observational research

Experimental research

Example Weight Loss: t-test for β_1 or β_2

Coefficients[a]

t-tests for $H_0: \beta_j = 0$

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	.667	.694		.960	.362
	x1	1.317	.100	.895	13.191	.000
	x2	-8.000	1.367	-.397	-5.853	.000

a. Dependent Variable: y

- Does extra time increase the mean Weight Loss when humidity is kept constant? (use α =0.05)
- test for e.g. $H_0: \beta_1 = 0$ versus $H_a: \beta_1 > 0$
- Test Statistic: $t = \frac{\hat{\beta}_1 - 0}{se_{\hat{\beta}_1}}$ under H_0, t ~ t_9
- Under H_a t tends to larger values, so we use RPV.
- t = $= \frac{1.317-0}{0.1} = 13.191$ > 0 So RPV = 2-tailed PV / 2 = 0.000/2 < 0.05,
- H_0 is rejected, H_a is proven, it is shown that $\beta_1 > 0$.
- Extra time leads to larger mean Weight Loss ...
- [for time in observed range, and if humidity is kept constant, also within its observed range].

4.2 Confidence interval for β_j

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	.667	.694		.960	.362
	x1	1.317	.100	.895	13.191	.000
	x2	-8.000	1.367	-.397	-5.853	.000

a. Dependent Variable: y

- Table above gives the point estimates for β_j
- What about $(1-\alpha)$ confidence intervals for β_j?
- Two sided (1-α) CI : $\hat{\beta}_j \pm t_{dfE(\alpha/2)} \times se_{\hat{\beta}_j}$
- Give a two sided (1-α) CI for β_1:
- Limits are: $1.317 \pm t_9(0.025) * 0.100$

  Or: $1.317 \pm 2.262 * 0.100$ → (1.09, 1.54)

Example Weight Loss: t-test for β_1 or β_2

```
> summary(lm(y ~ x1 + x2))

Call:
lm(formula = y ~ x1 + x2)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.73333 -0.17083 -0.04167  0.33750  0.46667 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.66667    0.69423   0.960 0.361994    
x1           1.31667    0.09981  13.191 3.43e-07 ***
x2          -8.00000    1.36677  -5.853 0.000243 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3866 on 9 degrees of freedom
Multiple R-squared:  0.9586,	Adjusted R-squared:  0.9494 
F-statistic: 104.1 on 2 and 9 DF,  p-value: 5.993e-07
```

Confidence interval for β_1

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)	.667	.694		.960	.362	-.904	2.237
	Time	1.317	.100	.895	13.191	.000	1.091	1.542
	Humidity	-8.000	1.367	-.397	-5.853	.000	-11.092	-4.908

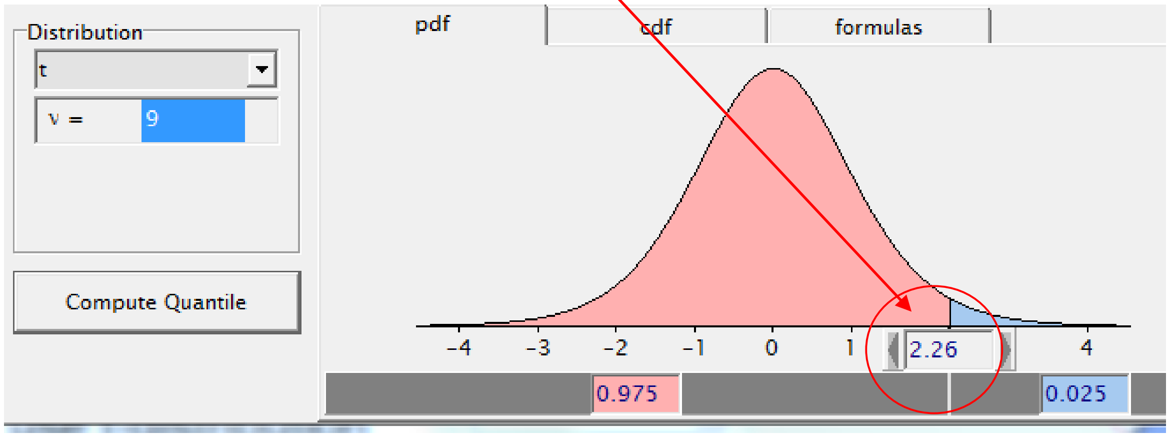
a. Dependent Variable: Weight_loss

0.95-confidence interval for β_1: $(1.317 \pm 2.262 * 0.100) = (1.09, 1.54)$

use t-distribution

$\alpha / 2 = 0.025$,

df = 12 – (2+1) = 9



5. Checking model assumptions

- Assumptions for ε_i's:
  1. Correct systematic part of model.
  2. Constant variance: Var(ε) = σ_ε^2
  3. Normality of ε's
  4. Independence

Note: mean of residuals *e* is always 0 (when there is an intercept in the model), even when the model fits poorly.

- Assumptions in short: ε_i iid from $\mathrm{N}(0,\sigma)$

- 2. Constant variance

  Plot residuals against $\hat{y}$. Cloud of points without structure is OK.

  Some patterns, e.g. loudspeaker form, indicate that constant variance assumption is possibly violated, e.g. variance increases with the mean.

  Same plots also may show outliers. Correct any obvious errors.

- Check if observations, including high leverage points, have no undue influence upon your conclusions. Only with good reason, leave out an observation.

Checking model assumptions, continued

Correct systematic part of model

Plot residuals against x_1, x_2, etc. separately.

A pattern in the mean vs one of the x-variables (e.g x_1) indicates that the model could be improved, e.g. by adding a squared term (x_1^2).

Normality

Q-Q plot of residuals.
Should look like a straight line (always some stragglers at the ends of the plot).

Equal Variances

By plotting residuals versus predicted values;

Independence

In general: can only be achieved by design, randomization, proper sampling.

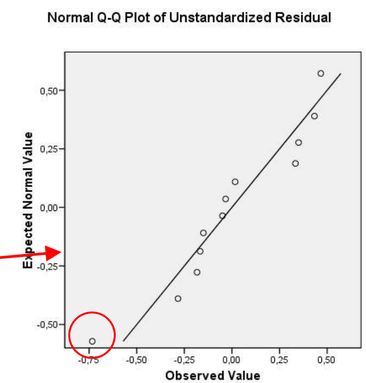
Weight Loss, checking model assumptions

Check assumptions:
1) normality
2) equal variances
3) linearity
4) independence

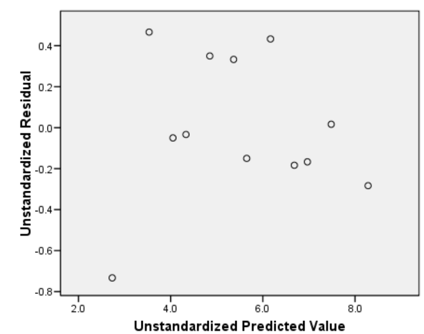
	time	humid	wt_loss	PRE_1	RES_1
1	4.0	.20	4.3	4.33333	-.03333
2	5.0	.20	5.5	5.65000	-.15000
3	6.0	.20	6.8	6.96667	-.16667
4	7.0	.20	8.0	8.28333	-.28333
5	4.0	.30	4.0	3.53333	.46667
6	5.0	.30	5.2	4.85000	.35000
			6.6	6.16667	.43333
			7.5	7.48333	.01667
			2.0	2.73333	-.73333
			4.0	4.05000	-.05000
			5.7	5.36667	.33333
			6.5	6.68333	-.18333

First ask for residuals and predicted values.

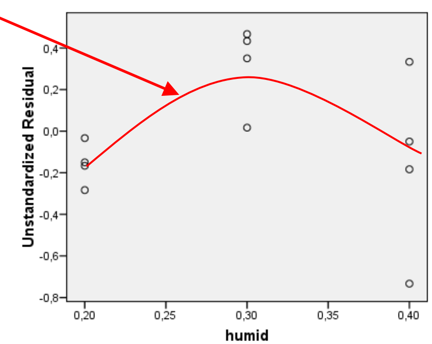
Then check
- normality by QQ-plot: points approximately on straight line? Strange observation?

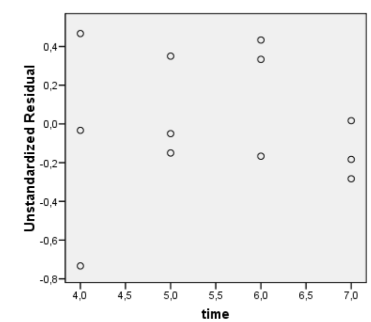


- equal variances by plotting residuals versus predicted values; evidence for non-constant variance in plot?



- linearity in x 's by plotting residuals versus individual x 's;
  curvature in plot versus humidity!
  relationship of y with humidity does not seem to be linear





- independence; cannot be checked graphically, should follow from study design we lack information in this example.