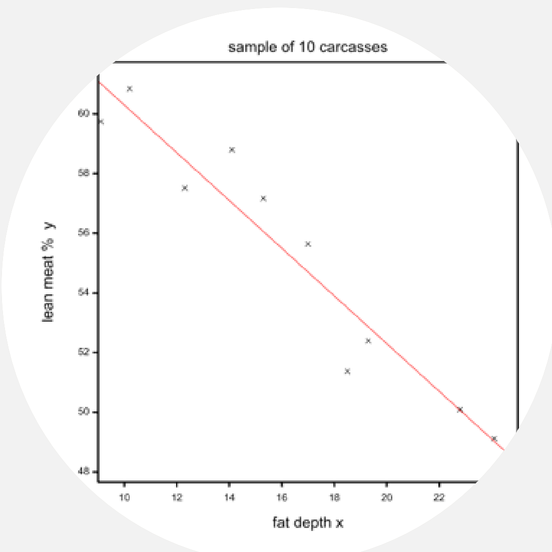
Linear and Generalized Linear Models (4433LGLM6Y)

Simple and Multiple Linear Regression

Meeting 1

Dr. Jos Hageman



WAGENINGEN UNIVERSITY
WAGENINGEN UR

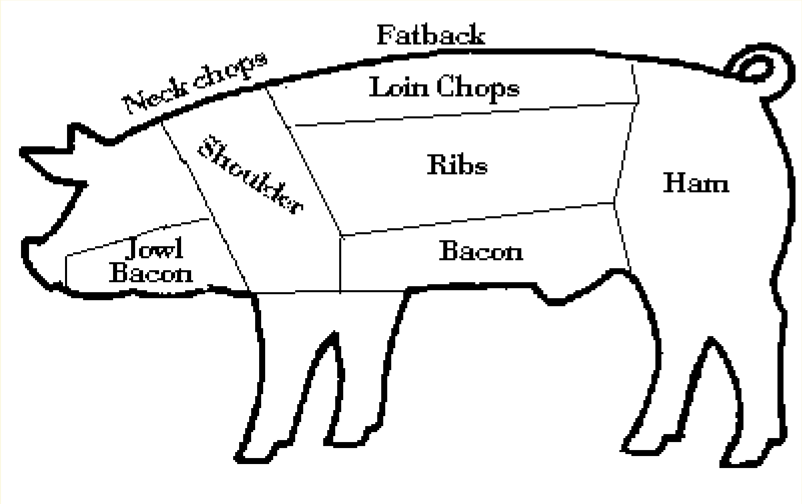
Simple regression
Example: prediction of the lean meat percentage

In slaughterhouses the percentage lean meat of a pig carcass must be determined for payment to the farmer.

To determine the lean meat percentage the carcass has to be dissected in meat, fat and bone.

This is costly and destructive.

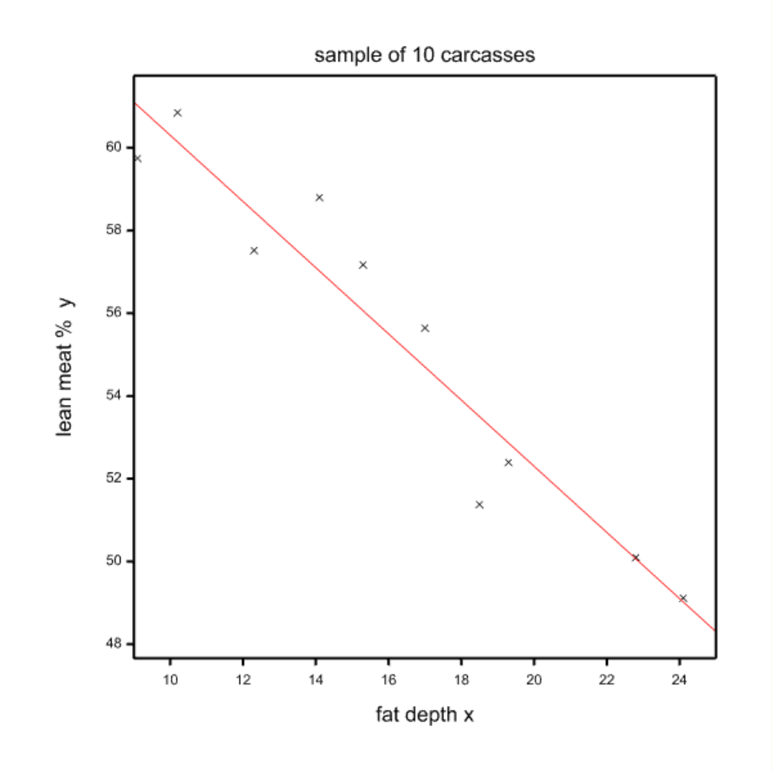
So, we predict the lean meat percentage, in order to keep the carcass intact.



Lean meat percentage data

y	x
59.7451	9.1
60.8473	10.2
57.5190	12.3
58.7997	14.1
57.1727	15.3
55.6435	17.0
51.3749	18.5
52.3963	19.3
50.0982	22.8
49.1171	24.1

y = percentage lean meat of a pig carcass
x = fat depth measurement
sample of 10 carcasses



Prediction of the lean meat percentage

Population = population of slaughter pigs

Experimental units = slaughter pigs

response variable y $\xleftarrow{?}$ explanatory variable x

Can we construct a prediction formula, i.e. an expression in terms of x (measured on the carcass) that gives a prediction for y (not measured)?

Refresher of simple regression

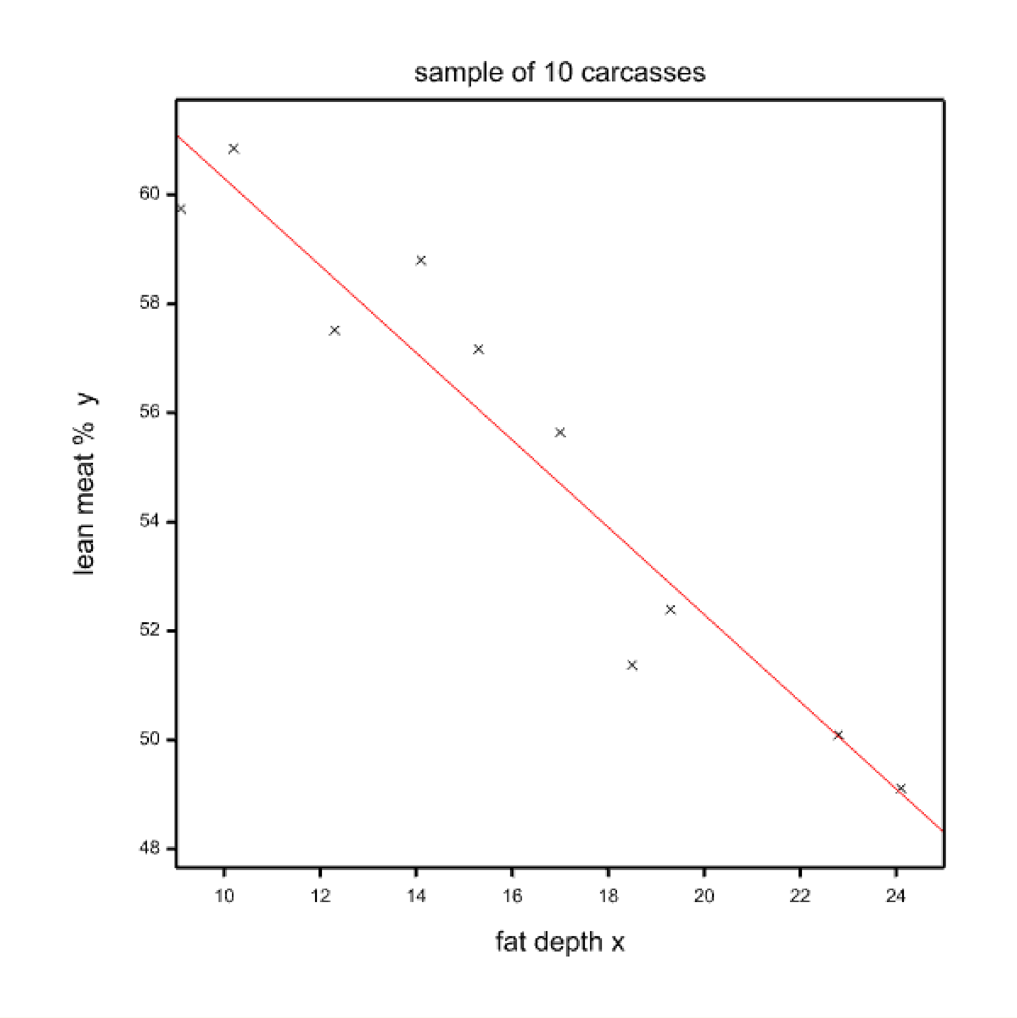
a statistical model for the data

$$y = \underbrace{\beta_0 + \beta_1 x}_{} + \epsilon$$

population mean for % lean
of all carcasses with fat depth x

departure from the mean
for individual carcass
=
biological variation in % lean
between carcasses with same
fat depth x

Drawing a “good” line



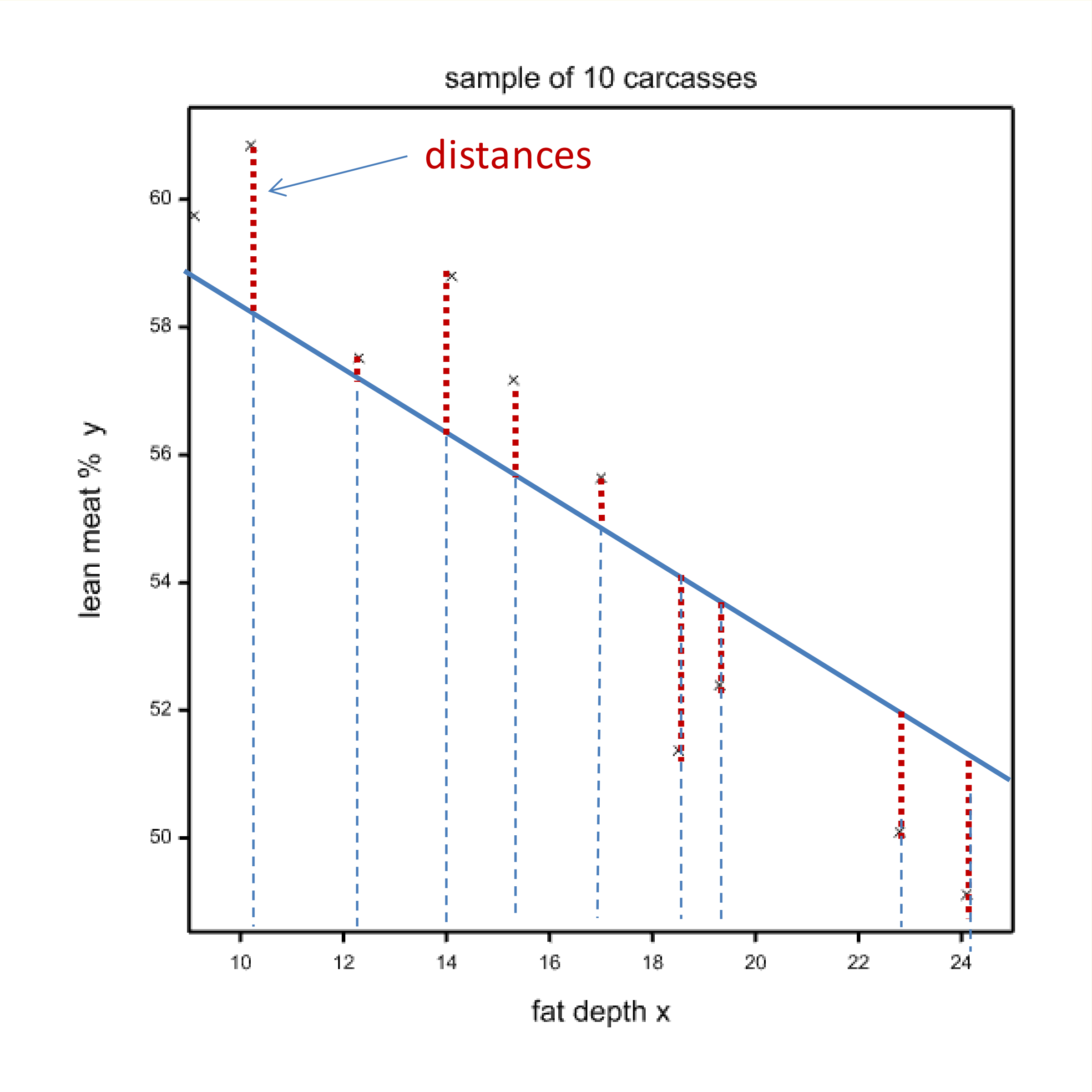
We showed a line passing through these points:

$$\hat{y} = 68.3 - 0.80 \, * x$$

Estimates are:

$\hat{\beta}_0 = 68.3$ and $\hat{\beta}_1 = 0.80$

But how were these estimates obtained?



Draw a line.

Look at distances of the points to the line.

Minimize these distances to find the ‘best’ line.

Least squares estimation

- remove minus signs of distances

- take sum of absolute values of distances?
- that is mathematically awkward to handle

- long ago decided, by Gauss among others, to take the sum of squared distances

Minimize sum of squares (SS) of distances

=

Least Squares Estimation

Carl Friedrich Gauss

Least squares estimates in R

$\vec{b} = (X^T X)^{-1} X^T \vec{y}$ $\quad \vec{b} - \vec{\beta} = (X^T X)^{-1} X^T \vec{\varepsilon}$ $\quad s^2 = \frac{1}{n-k} e^T e$

$\hat{\beta}_0$ = 68.34463 and $\hat{\beta}_1$= -0.80352

$Var(\vec{b}|X) = Var(\vec{b} - \vec{\beta}|X) = Var[(X^T X)^{-1} X^T \vec{\varepsilon}|X] = (X^T X)^{-1} X^T Var(\vec{\varepsilon}|X)[(X^T X)^{-1} X^T]^T$

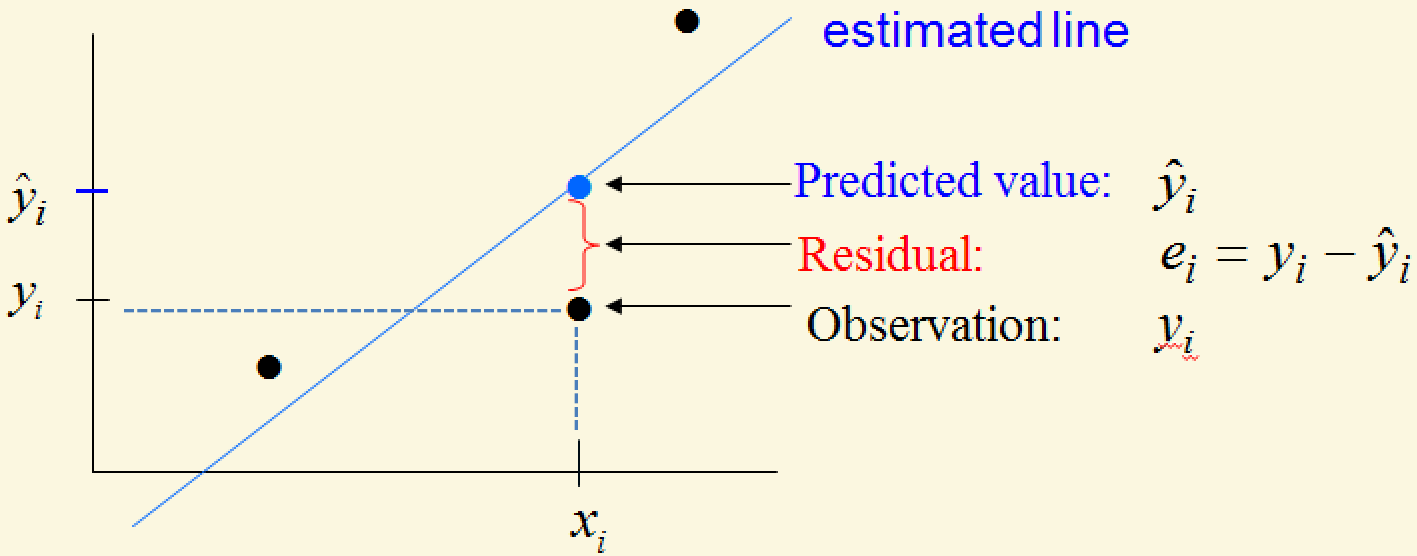
$= \sigma^2 (X^T X)^{-1} X^T X (X^T X)^{-1} = \sigma^2 (X^T X)^{-1}$

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 68.34463    1.43305  47.692 4.13e-11 ***
x           -0.80352    0.08451  -9.508 1.24e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.277 on 8 degrees of freedom
Multiple R-squared: 0.9187,     Adjusted R-squared: 0.9085
F-statistic:  90.4 on 1 and 8 DF,  p-value: 1.236e-05
```

$(X^{-1} X^T)^T$

Predicted or fitted values and residuals



Predicted value $\hat{y}_i$: point on the line
(also fitted value) estimate of population mean of all units with $x = x_i$

Residual e_i : difference between observation y_i and prediction $\hat{y}_i$
estimate of the error ϵ_i

Estimation of error variance σ_ϵ^2

If ϵ's were known, for the carcass data, an estimate for the variance σ_ϵ^2 would be:

$$\hat{\sigma}_\epsilon^2 = (\epsilon_1^2 + \ldots + \epsilon_{10}^2) \,/\, 10$$

We do not know error terms ϵ and use residuals e instead.

Because β_0 and β_1 are estimated from the data, there is the risk of underestimating σ_ϵ^2.
Therefore, because β_0 and β_1 are estimated, we have to subtract 2 and divide by 8:

$$\hat{\sigma}_\epsilon^2 = (e_1^2 + \ldots + e_{10}^2) \,/\, 8$$

Estimation of error variance σ_ϵ^2 in general

$$\hat{\sigma}_\epsilon^2 = (\underbrace{e_1^2 + \ldots + e_{10}^2}) \,/\, 8$$

This is the number of observations n reduced by 2: $(n-2)$

This is the *SS* that we minimized, called the sum of squares for error (or residual sum of squares), denoted by *SSE*.

Hence: $\hat{\sigma}_\epsilon^2 = (e_1^2 + \ldots + e_n^2)/(n-2) = SSE/(n-2)$

SSE / (*n* - 2) is called the mean square for error: *MSE*

The output from R again

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 68.34463    1.43305  47.692 4.13e-11 ***
x           -0.80352    0.08451  -9.508 1.24e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
' 1

Residual standard error: 1.277 on 8 degrees of freedom
Multiple R-squared: 0.9187,     Adjusted R-squared: 0.9085
F-statistic:  90.4 on 1 and 8 DF,  p-value: 1.236e-05
```

Estimate $\hat{\sigma}_\epsilon$ for σ_ϵ, the standard deviation of the ϵ's.

So: $\hat{\sigma}_\epsilon^2 = 1.277^2 = 1.63$.

ANOVA table

The ANOVA table is about three sums of squares

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

total sum of squares *SST*

regression sum of squares *SSR*

error (or residual) sum of squares *SSE*

ANOVA table for the lean meat data

ANOVA table lean meat data

	Source	df	SS	MS	F	P-value
SSR →	Regression	1	147.40	147.400	90.40	< 0.001
SSE →	Residual	8	13.04	1.631		
SST →	Total	9	160.44	17.827		

Total sum of squares *SST*

SST quantifies variation in y, ignoring x.

This is measured around the sample mean $\bar{y}$:

$$SST = (y_1 - \bar{y})^2 + (y_2 - \bar{y})^2 + \cdots + (y_{10} - \bar{y})^2$$

For the lean meat data: *SST* = 160.445

SST, degrees of freedom, and *MST*

Leeway (elbow room) to quantify variation in y, ignoring x:

10 differences $(y_1 - \bar{y})$, $(y_2 - \bar{y})$, $\dots (y_{10} - \bar{y})$ that add up to 0

basically 9 differences to quantify variation in $y \rightarrow$ so, 9 degrees of freedom

Estimated variance of *y*, ignoring x, is:

$$\hat{\sigma}_y^2 = \frac{SST}{n-1} = MST = \text{Mean Square for Total} =$$

$$\frac{1}{n-1}\{(y_1 - \bar{y})^2 + (y_2 - \bar{y})^2 + \cdots + (y_n - \bar{y})^2\} =$$

$$160.445 \,/\, 9 = 17.83$$

Splitting *SST* into *SSR* and *SSE*

In the ANOVA table (from R) *SST* is split into two parts.

```
Response: y
          Df  Sum Sq  Mean Sq  F value    Pr(>F)
x          1 147.401  147.401     90.4 1.236e-05
Residuals  8  13.044    1.631
```

SST = 160.445 = 147.401 + 13.044

SSR = 147.401 corresponds to systematic part of the model

SSE = 13.044 corresponds to random part of the model

Sum of squares for error *SSE*

- Minimum sum of squared distances
- Sum of squared distances to the line fitted by LS
- $(n-2)$ degrees of freedom
- $MSE = \frac{SSE}{n-2}$ is estimate for error variance σ_ϵ^2
- For the lean meat data: $\hat{\sigma}_\epsilon^2$= 13.044 / 8 = 1.63

Sum of squares for regression *SSR*

- *SSR* = *SST* – *SSE* = 160.445 - 13.044 = 147.401

- *SSR* reflects the part of *SST* that is "explained by x"

- 1 degree of freedom, because leeway for variation explained by x is through single parameter β_1.

- *MSR* = *SSR* / 1 =147.401 / 1 = 147.40

Multiple Linear Regression

Example: Weight loss of a compound

RQ: What is the effect of exposure to air over time and the humidity during expore, on weight loss of a compound?

y = weight loss of a chemical compound (pounds), this is the response

x_1 = exposure time to air (hours), values are chosen

x_2 = relative humidity during exposure, values are observed (differ from O&L)

$$y \xleftarrow{\quad ? \quad} x_1 \text{ and } x_2$$

response / dependent variable / y-variable / regressand

explanatory variables / independent variables / x-variables / regressors

Example: Weight loss of a compound

The data

	y	x_1	x_2
1	4.3	4	0.15
2	5.5	5	0.46
3	6.8	6	0.20
4	8.0	7	0.21
5	4.0	4	0.20
6	5.2	5	0.37
7	6.6	6	0.37
8	7.5	7	0.34
9	2.0	4	0.49
10	4.0	5	0.54
11	5.7	6	0.42
12	6.5	7	0.37

y = weight loss of a chemical compound (pounds), this is the response

x_1 = exposure time to air (hours), values are chosen

x_2 = relative humidity during exposure, values are observed (differ from O&L)

The multiple regression model

$$y = \underbrace{\beta_0 + \beta_1 x_1 + \beta_2 x_2}_{} + \epsilon$$

↑

response variable

=

observed weight loss of an individual experimental unit

↑

systematic part: population mean of weight loss for exposure time x_1 and relative humidity x_2

↑

random part: error term is departure of observed weight loss from the mean, represents variation around the mean

Multiple versus simple regression

In simple regression:
β_1 is expected change in y for unit change in x_1.

In multiple regression:
β_1 is expected change in y for unit change in x_1, while keeping all other x-variables constant.

Here:
β_1 is expected change in weight loss for one extra hour of exposure, while keeping relative humidity constant.

What is the interpretation of intercept β_0 ?

Least squares estimation

Find values $\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\beta}_2$ for β_0, β_1 and β_2 that minimize the sum of squared errors:

$$SS = \sum_{i=1}^{12}\left(y_i - (\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i})\right)^2$$

Same terminology as before:

$e_i = y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i}\right) = (y_i - \hat{y}_i)$ is a residual

$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta} x_{2i}$ is a fitted / predicted value

$\sum_{i=1}^{12}\left(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i})\right)^2 = \sum_{i=1}^{12} e_i^2 = SSE$ is error SS

Estimation of error variance σ_ϵ^2 in general

$$\hat{\sigma}_\epsilon^2 = (e_1^2 + \ldots + e_n^2) / (n - (k + 1))$$

Minimized sum of squares
= sum of squares for error
= residual sum of squares
= *SSE*.

degrees of freedom =
number of observations *n*
minus number of β
parameters (*k* slopes + 1
intercept)

Hence: $\hat{\sigma}_\epsilon^2 = SSE / (n - (k + 1))$

$SSE / (n - (k + 1))$ is the mean square for error (or residual): *MSE*

Some output (from R)

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.1916     1.0481  -0.183    0.859
x1            1.2952     0.1596   8.117 1.97e-05
x2           -4.1410     1.4670  -2.823    0.020
---
Residual standard error: 0.6173 on 9 degrees of freedom
Multiple R-squared:  0.8944,     Adjusted R-squared:
0.8709
F-statistic: 38.11 on 2 and 9 DF,  p-value: 4.043e-05
```

least squares estimates:

$\hat{\beta}_0$ = - 0.19 for β_0

$\hat{\beta}_1$ = 1.30 for β_1

$\hat{\beta}_2$ = -4.14 for β_2

estimate $\hat{\sigma}_\epsilon$ = 0.6173 for σ_ϵ,

so: $\hat{\sigma}_\epsilon^2 = 0.617^2 = 0.381$.

Prediction equation : $\hat{y}_i = -0.19 + 1.30\, x_{1i} - 4.14\, x_{2i}$

ANOVA table

Recall from simple regression:

ANOVA table is about three sums of squares (SS)

$$y = \beta_0 + \underbrace{\beta_1 x_1 + \beta_2 x_2}_{} + \epsilon$$

total sum of squares *SST*

regression sum of squares *SSR*

error (or residual) sum of squares *SSE*

ANOVA table - *SST*

SST = sum of squares of observations y minus sample mean $\bar{y}$,
degrees of freedom are: n - 1

MST = *SST* / (n - 1)

$\hat{\sigma}_y^2$ = *MST* = estimated variance of y ignoring x-variables

The same as in simple regression.

Source	*df*	*SS*	*MS*	*F*	*P-value*
Regression					
Error					
Total	11	32.47	2.95		

ANOVA table - *SSE*

SSE = minimized sum *SS* of squared distances,

= sum of squared residuals $e_1^2 + \ldots + e_n^2$

error (or residual) degrees of freedom are: $n - (k + 1)$

$MSE = SSE / (n - (k + 1))$

$\hat{\sigma}_\epsilon^2 = MSE$ = estimator of σ_ϵ^2 variation accounting for x -variables

Source	df	SS	MS	F	P-value
Regression					
Error	9	3.43	0.38		
Total	11	32.47	2.95		

$\hat{\sigma}_\epsilon^2 = MSE = 0.38$

MST & *MSE* two variance estimates

$$\hat{\sigma}_y^2 = MST =$$

estimated variance of y around sample mean $\bar{y}$ ignoring x_1, x_2.

$$\hat{\sigma}_\epsilon^2 = MSE =$$

estimated variance of y around fitted plane accounting for x_1, x_2

ANOVA table - *SSR*

$SSR = SST - SSE$

= part of *SST* "explained by" x-variables

degrees of freedom are number of slopes k

$MSR = SSR / k$

Source	*df*	*SS*	*MS*	*F*	*P-value*
Regression	2	29.04	14.52		
Error	9	3.43	0.38		
Total	11	32.47	2.95		

ANOVA table – F-test

F-test for $H_0 : \beta_1 = \beta_2 = 0$ "model has no predictive value"

test statistic is: $F = \frac{MSR}{MSE}$

reject H_0 for large values of F

Source	*df*	*SS*	*MS*	*F*	*P-value*
Regression	2	29.04	14.52	38.1	
Error	9	3.43	0.38		
Total	11	32.47	2.95		

F-test for predictive value – P-value

how large should $F = \frac{MSR}{MSE}$ be to reject $H_0\!: \beta_1 = \beta_2 = 0$?

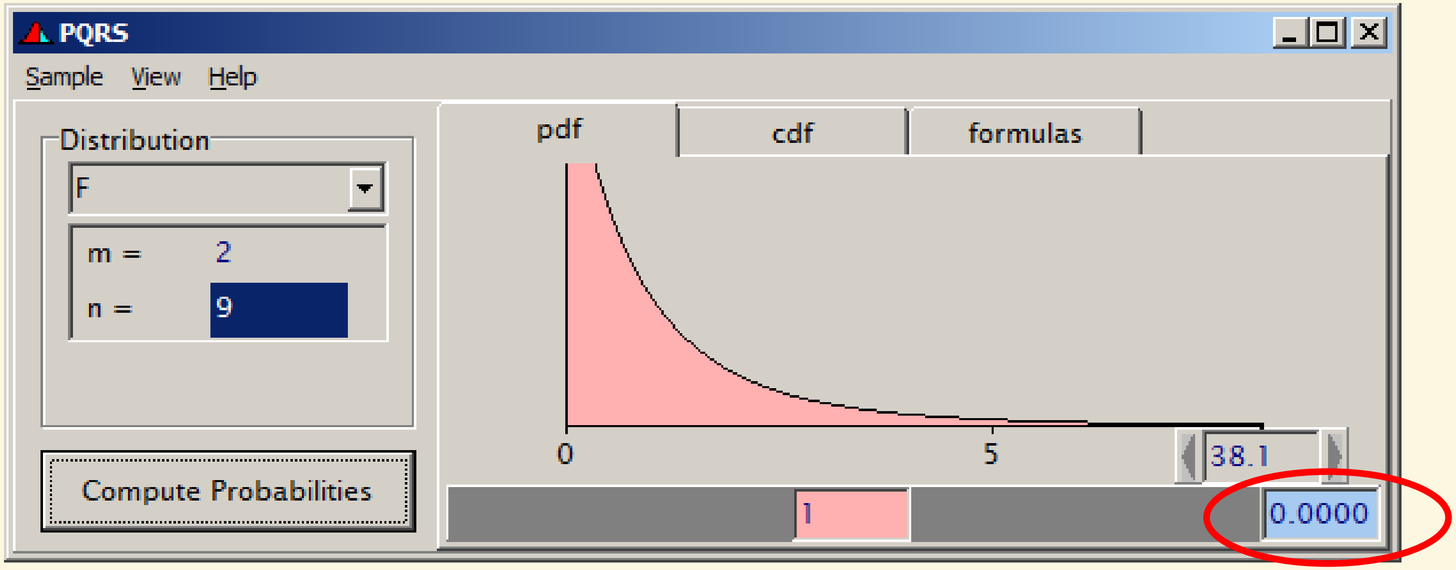
under H_0, *F* follows an F-distribution with df1 = *k*, df2 = *n* - (*k*+1)

df1 = 2 = df of *SSR* = number of β's involved in H_0

df2 = 9 = df of *SSE* = *n – (k + 1)* (--> leeway for estimation of σ_ϵ^2)

Source	*df*	*SS*	*MS*	*F*	*P-value*
Regression	2	29.04	14.52	38.1	0.00004
Error	9	3.43	0.38		
Total	11	32.47	2.95		

F-test - P-value, continued



P-value = area to the right of outcome 38.1 = 0.0000 .

This is smaller than 0.05.

Outcome 38.1 is too large to believe that H_0 is true: H_0 is rejected.

We conclude that either x_1, or x_2, or both x_1 and x_2 have predictive value for y, i.e. part of the variation in y is explained by x_1 and/or x_2.

t-tests & confidence intervals & prediction

t-test for single regression coefficient

e.g. $H_0: \beta_2 = 0$ (no humidity effect) vs. $H_a: \beta_2 \neq 0$

output from R

```
> m1<-lm(y ~ x1 + x2)
> coef(summary(m1))
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   -0.1916     1.0481  -0.183    0.859
x1             1.2952     0.1596   8.117    0.0000197
x2            -4.1410     1.4670  -2.823    0.020
```

$$t = \frac{estimate - value\ from\ H_0}{standard\ error} = \frac{-4.1410 - 0}{1.4670} = -2.823$$

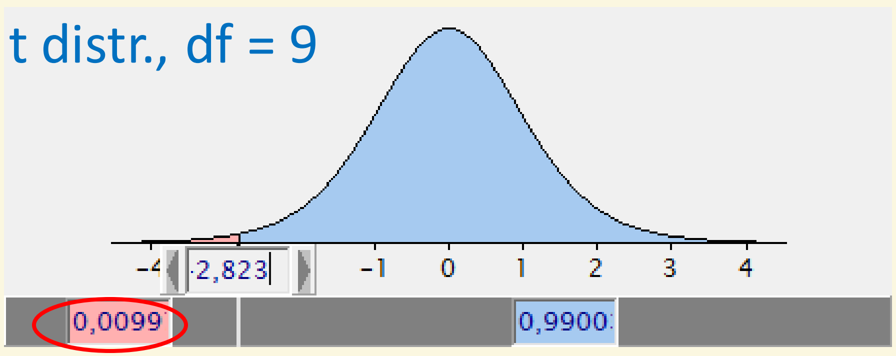
Compare with t-distribution with df = 9.

Again, df = 9 from SSE express leeway for estimation of σ_ϵ^2.

t-test, P-value

Coefficients:

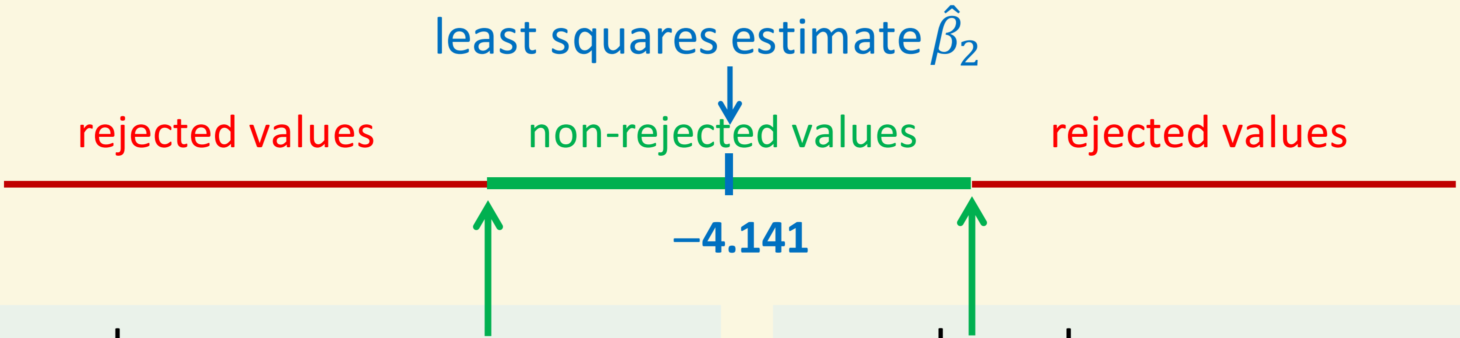
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.1916	1.0481	-0.183	0.859
x1	1.2952	0.1596	8.117	0.0000197
x2	-4.1410	1.4670	-2.823	0.020



Two-sided P-value is 2 * 0.0099 = 0.020 below 0.05, so reject H_0

Shown that humidity has a (negative) effect upon (expected) weight loss (in combination with exposure time).

Confidence interval for a slope, e.g. 0.95 CI for β_2



lower bound:

$\hat{\beta}_2 - constant * se(\hat{\beta}_2) =$

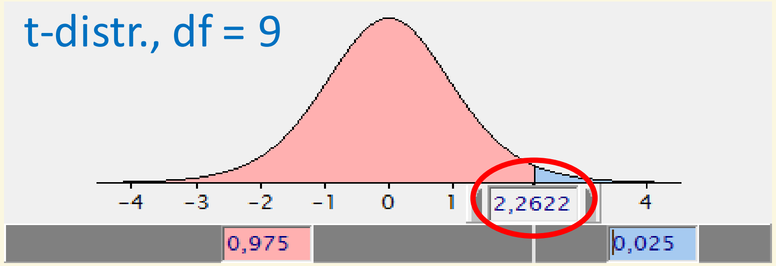
$-4.141 - 2.262 * 1.467 = -7.460$

upper bound:

$\hat{\beta}_2 + constant * se(\hat{\beta}_2) =$

$-4.141 + 2.262 * 1.467 = -0.823$

```
> coef(summary(m1))
            Estimate Std. Error ...
(Intercept)  -0.1916     1.0481 ...
x1            1.2952     0.1596 ...
x2           -4.1410     1.4670 ...
```



0.95 CI for β_2 is $(-7.5, -0.82)$, all “likely” values according to the data.

Confidence interval for a population mean μ

Estimate expected weight loss μ for $x_1 = 5.5$ and $x_2 = 0.50$:

$$\hat{\mu} = \hat{\beta}_0 + \hat{\beta}_1 * 5.5 + \hat{\beta}_2 * 0.50 =$$

$$-0.191 + 1.2952 * 5.5 - 4.1410 * 0.50 = 4.862$$

```
> pred.at <- data.frame(x1=5.5, x2=0.5)
> predict(m1, pred.at, se.fit=T,interval="confidence")
$fit
    fit         lwr     upr
1 4.862       4.205   5.518

$se.fit [1] 0.2903
```

R-output

estimate $\hat{\mu}$

standard error $\hat{\mu}$

lower& upper bound 0.95 CI for μ:
(4.862 $\pm$ 2.262 * 0.2903) = (4.205, 5.518)

Prediction of a single observation y

Prediction $\hat{y}$ for single y with $x_1 = 5.5, x_2 = 0.50$ is:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 * 5.5 + \hat{\beta}_2 * 0.50 + \hat{\epsilon}$$

$$= \hat{\mu} + \hat{\epsilon} = \hat{\mu} + 0 = 4.862 + 0 = 4.862,$$

since the best guess $\hat{\epsilon}$ for error term ϵ is 0.

So, (again) $\hat{\mu}$ and $\hat{y}$ are the same.

goodness of fit: R^2 & R^2_{adj}

R^2

R^2 = proportion of variation in y explained by x_1 and x_2.

Source	*df*	*SS*	*MS*	*F*	*P-value*
Regression	2	**29.040**	14.520	38.109	0.00004
Error	9	**3.429**	0.381		
Total	11	**32.469**	2.952		

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} = \frac{\mathbf{29.040}}{\mathbf{32.469}} = 1 - \frac{\mathbf{3.429}}{\mathbf{32.469}} = 0.894$$

89.4 % of variation in y is explained by x_1 and x_2.

R^2 (or R squared) is also called the coefficient of determination.

R^2, continued

R^2 is a measure for goodness of fit of the model.

R^2 is between 0 and 1; the closer to 1, the better the fit of the model.

R^2 depends upon how data were collected.

The wider the ranges of values for the explanatory variables, the larger R^2 will tend to be.

Only compare R^2 values of different models for the same data set.

R^2 and adjusted R^2

R^2 increases when a new x-variable is added to the model, regardless whether the extra variable has predictive value or not.

Therefore, an adjusted R^2 has been proposed:

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

R^2_{adj}, continued

$$R^2_{adj} = 1 - \frac{MSE}{MST}$$

MSE is the estimator for σ^2_ϵ (variance y accounting for x_1 and x_2).

MST is the estimator for σ^2_y (variance y ignoring x_1 and x_2).

So, R^2_{adj} is truly a proportion of explained variance.

When an x-variable is added to the model, R^2_{adj} only increases when $\hat{\sigma}^2_\epsilon = MSE$ decreases.