

LINEAR REGRESSION

Håkon K. Gjessing

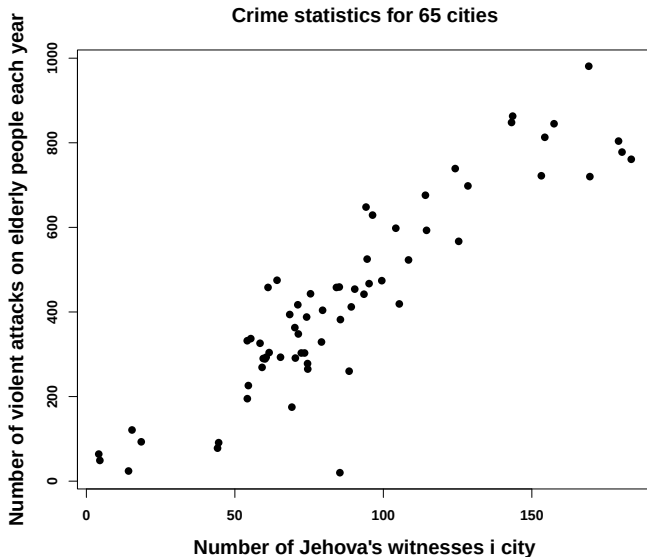
Professor/Principal Investigator

- Centre for Fertility and Health, Norwegian Institute of Public Health, Oslo
 - a Norwegian Research Council Centre of Excellence
- Department of Global Public Health and Primary Care, University of Bergen

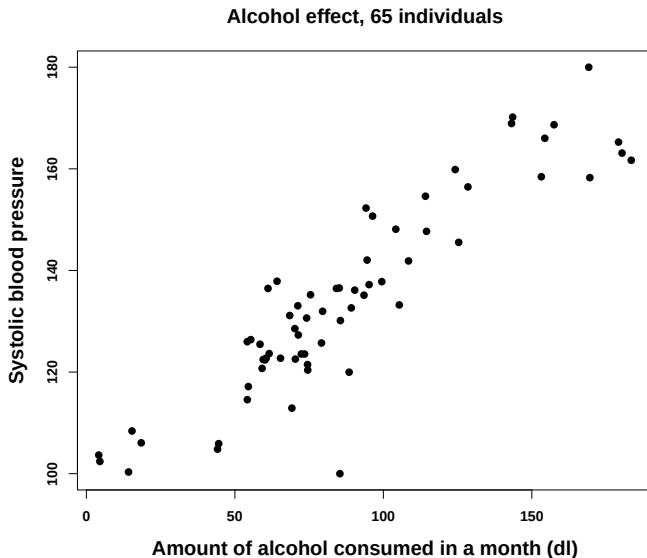
Kathmandu

Friday, 2 October 2026

THE SOURCE OF ALL EVIL...

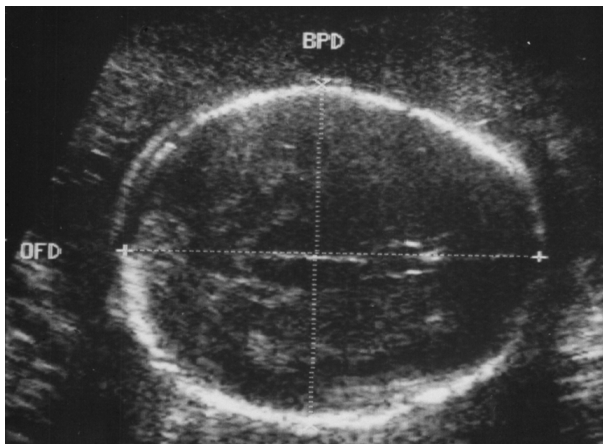


THE SOURCE OF ALL EVIL... (VERSION II)



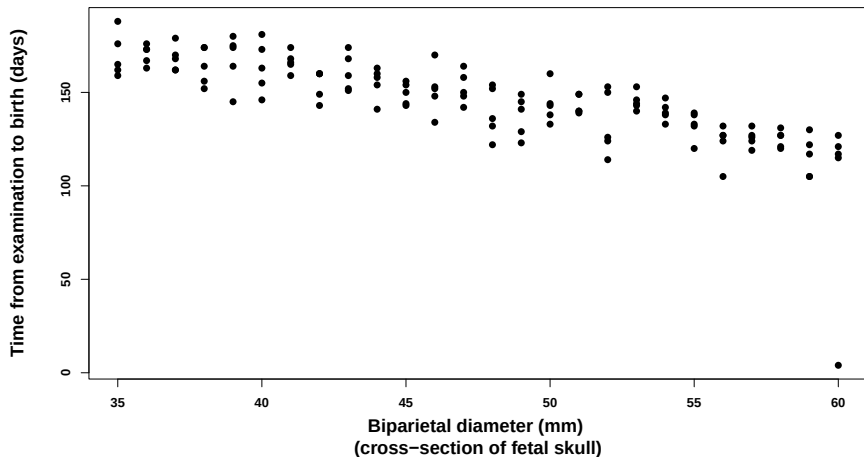
PROBLEM: ESTIMATE A TERM DATE FOR PREGNANT WOMEN

- At first routine ultrasound examination around week 18
Ultrasound measurements, such as **BiParietal Diameter (BPD)**



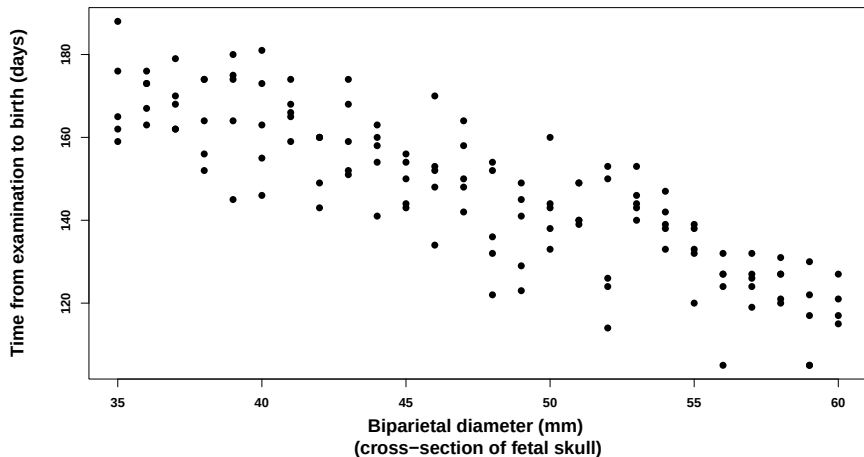
ULTRASOUND AT 18 WEEKS OF PREGNANCY

130 (129) children (random sample from 49302), Trondheim 1987 – 2004,
ultrasound examination at about 18 weeks of pregnancy



ULTRASOUND AT 18 WEEKS OF PREGNANCY (WITHOUT OUTLIER)

130 (129) children (random sample from 49302), Trondheim 1987 – 2004,
ultrasound examination at about 18 weeks of pregnancy



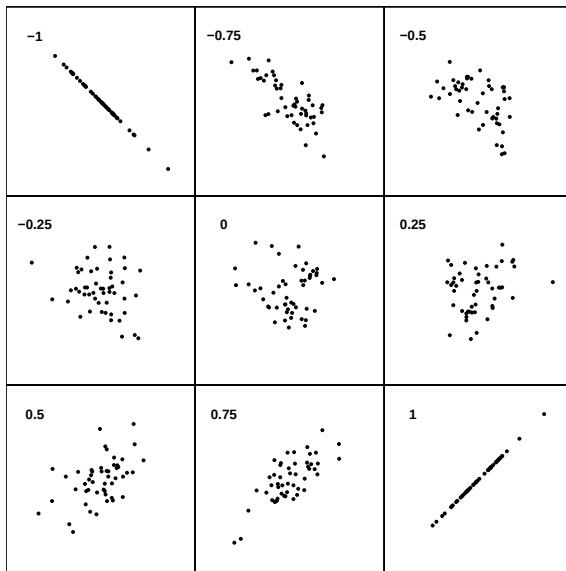
PEARSON CORRELATION COEFFICIENT

- X and Y are two continuous random variables
(... and a common assumption: They should follow a “bivariate normal distribution”)
- The correlation $\rho(X, Y)$ (rho) measures
linear relationship between X and Y

Properties:

- $-1 \leq \rho \leq 1$
- ρ positive: *Large* values of X relates to *large* values of Y, and small to small
- ρ negative: *Small* values of X relates to *large* values of Y, and vice versa
- $\rho = 0$: No clear relationship between X and Y

CORRELATION



Correlation ρ is not affected by *linear* changes in measurement scale

- **Advantage:**

What measurement scale being used

centimeter $\Leftrightarrow$ meter

month $\Leftrightarrow$ year

kilo $\Leftrightarrow$ gram

etc.

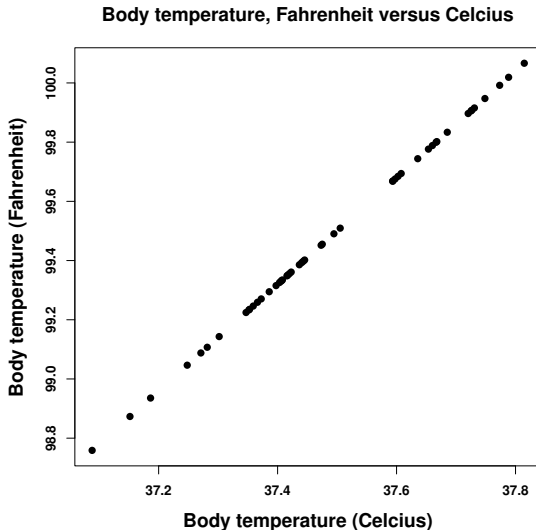
does not matter

- **Disadvantage:**

Correlation is thus not well suited for measuring *agreement*

CORRELATION: CELCIUS VERSUS FAHRENHEIT

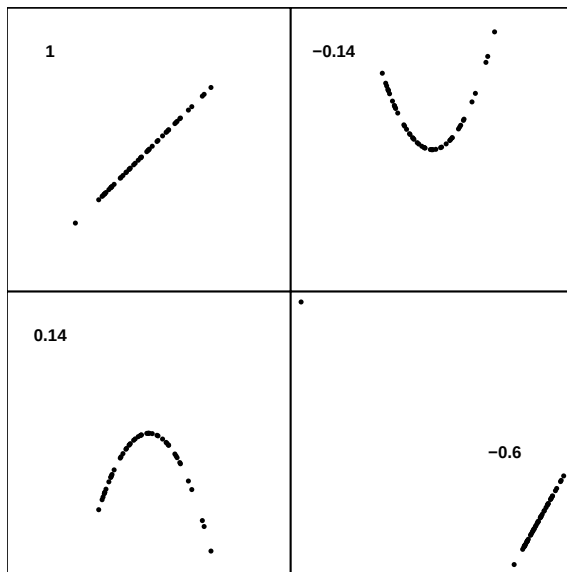
HIGH CORRELATION, LOW AGREEMENT



CORRELATION

- Correlation not well suited for *non-linear* relationships
(Spearman rank correlation can sometimes work better)
- High correlation does *not* prove causality!
- The correlation coefficient is perhaps the *most mis-used* measure ever!

CORRELATION (OTHER PATTERNS)



FORMULA FOR COMPUTING CORRELATION

Variances:

$$\text{var}(X) = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$

$$\text{var}(Y) = \frac{1}{n-1} \sum_{i=1}^n (Y_i - \bar{Y})^2$$

Standard deviations:

$$\text{SD}(X) = \sqrt{\text{var}(X)}$$

$$\text{SD}(Y) = \sqrt{\text{var}(Y)}$$

Covariance:

$$\text{covar}(X, Y) = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})$$

Correlation:

$$\rho(X, Y) = \frac{\text{covar}(X, Y)}{\text{SD}(X)\text{SD}(Y)}$$

DO YOU RECALL VECTOR GEOMETRY??

Length of a vector:

$$|\vec{X}| = \sqrt{x_1^2 + x_2^2 + x_3^2}$$
$$|\vec{Y}| = \sqrt{y_1^2 + y_2^2 + y_3^2}$$

Scalar product:

$$\vec{X} \cdot \vec{Y} = x_1 \cdot y_1 + x_2 \cdot y_2 + x_3 \cdot y_3$$

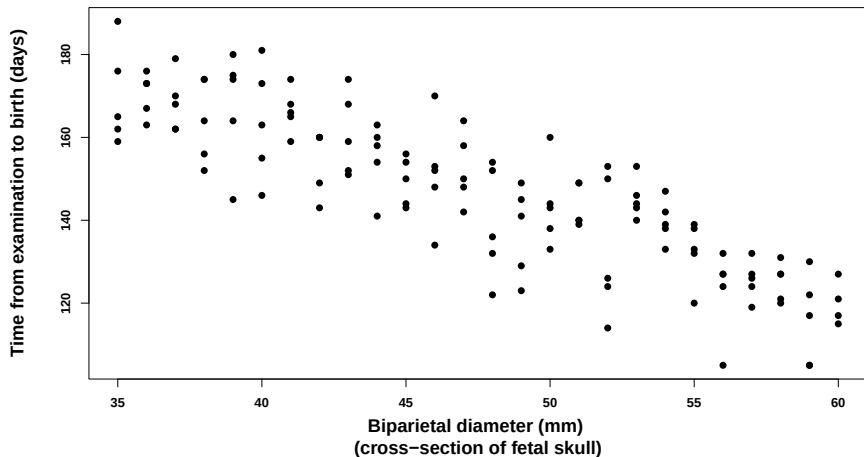
Cosine of angle between vectors:

$$\cos(\phi) = \frac{\vec{X} \cdot \vec{Y}}{|\vec{X}| |\vec{Y}|}$$

Something familiar?

ULTRASOUND AT 18 WEEKS OF PREGNANCY (WITHOUT OUTLIER)

130 (129) children (random sample from 49302), Trondheim 1987 – 2004,
ultrasound examination at about 18 weeks of pregnancy



TEST OF CORRELATION

SIGNIFICANCE AND CONFIDENCE INTERVAL

Test of correlation between

biparietal diameter (bpd) and remaining time of pregnancy (remain)

$$H_0 : \rho = 0, \quad H_A : \rho \neq 0$$

Pearson's product-moment correlation

```
data:  bpd and remain
```

```
t = -18.5925, df = 127, p-value < 2.2e-16
```

```
alternative hypothesis: true correlation is not equal to 0
```

```
95 percent confidence interval:
```

```
-0.8956341 -0.8006743
```

```
sample estimates:
```

```
cor
```

```
-0.8551727
```

(Printout from R, www.r-project.org)

CONFIDENCE INTERVAL FOR ρ

USING THE FISHERS Z-TRANSFORM

Values: $\rho = -0.855$ og $n = 129$.

Z-transform of ρ :

$$Z = \frac{1}{2} \log \frac{1 + \rho}{1 - \rho} = \frac{1}{2} \log \frac{1 - 0.855}{1 - (-0.855)} = -1.274.$$

Standard Error for the transformed value:

$$SE(Z) = \frac{1}{\sqrt{n-3}} = \frac{1}{\sqrt{129-3}} = 0.08909.$$

Confidence interval for Z : $-1.274 \pm 1.96 \cdot 0.08909 = (-1.449, -1.100)$.

Plug these in to the inverse formula (in place of Z):

$$\rho = \frac{e^{2Z} - 1}{e^{2Z} + 1},$$

... which gives the 95% confidence interval for ρ as $(-0.895, -0.800)$.

LINEAR REGRESSION

We need to choose what is the:

Y : Outcome, dependent variable

X : Exposure, covariate, independent variable
medskip

We often think in terms of **causality**:

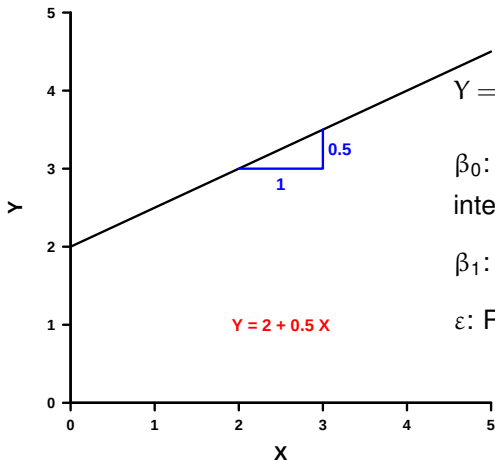
$$X \rightarrow Y$$

(BUT regression does not *guarantee* this!)

medskip **Purpose**:

- Find the effect of X on Y
for instance, alcohol dose on blood pressure
- Adjust (control) for other covariates, for instance smoking, physical activity etc.
that is, *multiple regression*
- Be able to predict the value of Y from the value of X

LINEAR REGRESSION



$$Y = \beta_0 + \beta_1 \cdot X + \varepsilon$$

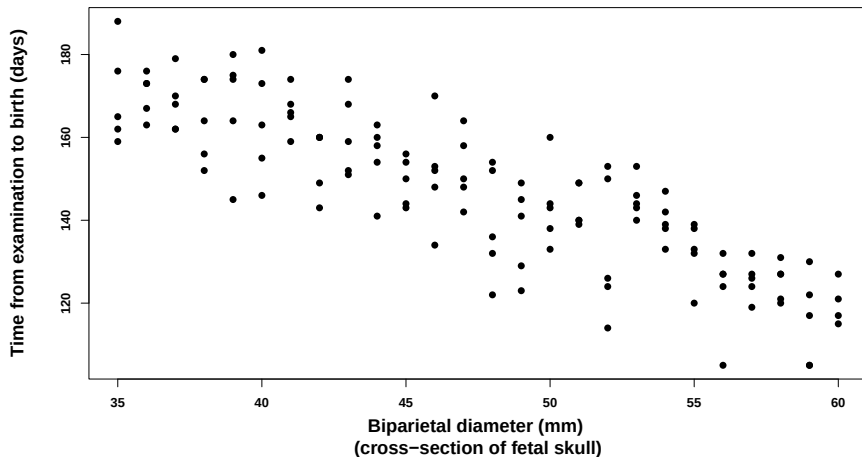
β_0 : Intercept (where regression line intercepts the Y-axis)

β_1 : Slope

ε : Random individual variation (Residual)

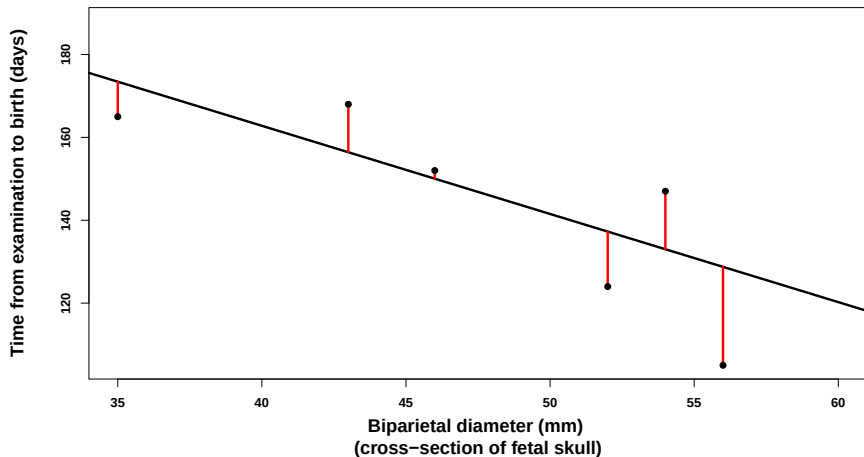
ULTRASOUND AT 18 WEEKS OF PREGNANCY (WITHOUT OUTLIER)

130 (129) children (random sample from 49302), Trondheim 1987 – 2004,
ultrasound examination at about 18 weeks of pregnancy

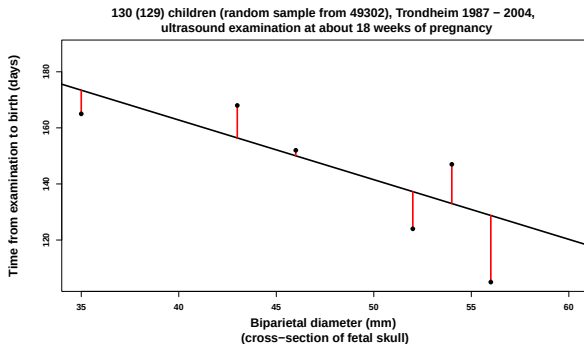


ULTRASOUND AT 18 WEEKS OF PREGNANCY: REGRESSION

130 (129) children (random sample from 49302), Trondheim 1987 – 2004,
ultrasound examination at about 18 weeks of pregnancy



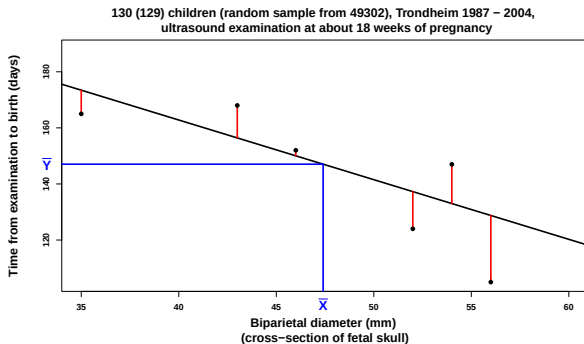
LINEAR REGRESSION, LEAST SQUARES METHOD



- Choose the values β_0 and β_1 that make the line “fit as best as possible” to the data
- I.e. those that make the Sum of Squares as small as possible:

$$\text{Sum of Squares} = \sum_i (Y_i - (\beta_0 + \beta_1 X_i))^2$$

LINEAR REGRESSION, LEAST SQUARES METHOD

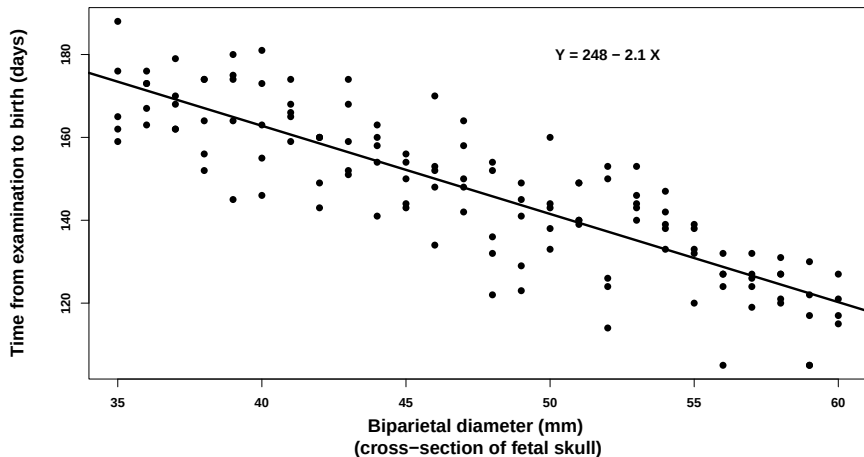


- Choose the values β_0 and β_1 that make the line “fit as best as possible” to the data
- I.e. those that make the Sum of Squares as small as possible:

$$\text{Sum of Squares} = \sum_i (Y_i - (\beta_0 + \beta_1 X_i))^2$$

ULTRASOUND AT 18 WEEKS OF PREGNANCY: REGRESSION

130 (129) children (random sample from 49302), Trondheim 1987 – 2004,
ultrasound examination at about 18 weeks of pregnancy



ULTRASOUND AT 18 WEEKS OF PREGNANCY: REGRESSION

Result:

$$Y = 248 - 2.1 \cdot X$$

$$\beta_0 = 248 \text{ days}, \quad \beta_1 = -2.1 \text{ days/mm}$$

Interpretation:

β_0 : Usually uninteresting

β_1 : Change, in days per mm BPD.

“How many days does a fetus take to grow 1 mm?”

Or: A fetus of e.g. 45 mm has 2.1 days more left of pregnancy than a fetus of 46 mm

ε : Difference between model predicted and observed Y-value

We assume ε is normally distributed:

Mean: 0, Standard deviation: σ (unknown), $\varepsilon \sim N(0, \sigma^2)$

ULTRASOUND AT 18 WEEKS OF PREGNANCY

CONFIDENCE INTERVALS

- Estimates from **sample** of 129 children:
- 248 and -2.1 (just forget about the 248!)
- Uncertainty? Confidence interval?
- P-value (needed, really??)

ULTRASOUND AT 18 WEEKS OF PREGNANCY

CONFIDENCE INTERVALS

- Estimates from **sample** of 129 children:
- 248 and -2.1 (just forget about the 248!)
- Uncertainty? Confidence interval?
- P-value (needed, really??)

Confidence interval (95%) for β_1 , from software:

-2.1 d/mm (-2.35 d/mm, -1.90 d/mm)

It is “likely” that the population value lies within this interval

ULTRASOUND AT 18 WEEKS OF PREGNANCY

CONFIDENCE INTERVALS

- Estimates from **sample** of 129 children:
- 248 and -2.1 (just forget about the 248!)
- Uncertainty? Confidence interval?
- P-value (needed, really??)

Confidence interval (95%) for β_1 , from software:

-2.1 d/mm (-2.35 d/mm, -1.90 d/mm)

It is “likely” that the population value lies within this interval

Full population, 49302 children

-1.90 d/mm

(By chance just on the border!)

ULTRASOUND AT 18 WEEKS OF PREGNANCY

PRINTOUT FROM SOFTWARE

Results regression (data sample):

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	247.9103	5.4915	45.14	<2e-16
bpd	-2.1278	0.1144	-18.59	<2e-16

Confidence intervals:

	2.5 %	97.5 %
(Intercept)	237.043633	258.777061
bpd	-2.354252	-1.901327

95% Confidence interval with the Normal distribution:

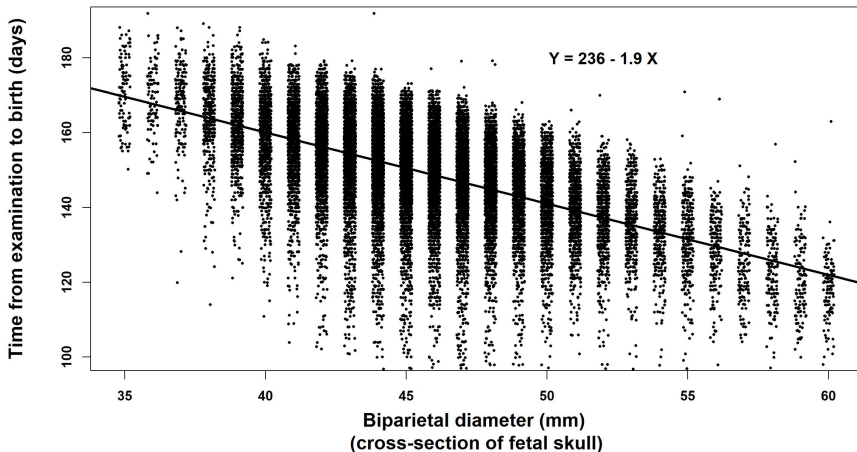
$$-2.1278 \pm 1.96 \cdot 0.1144 = (-2.352024, -1.903576)$$

(The software uses the t-distribution)

ULTRASOUND AT 18 WEEKS OF PREGNANCY

REGRESSION, ALL DATA

All 49302 children, Trondheim 1987 - 2004,
ultrasound examination at about 18 weeks of pregnancy



ULTRASOUND AT 18 WEEKS OF PREGNANCY

PRINTOUT FROM SOFTWARE

Regression results (data sample):

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	247.9103	5.4915	45.14	<2e-16
bpd	-2.1278	0.1144	-18.59	<2e-16

Confidence intervals:

	2.5 %	97.5 %
(Intercept)	237.043633	258.777061
bpd	-2.354252	-1.901327

Regression results (full data):

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	236.21837	0.74393	317.5	<2e-16
bpd	-1.90307	0.01627	-117.0	<2e-16

Confidence intervals:

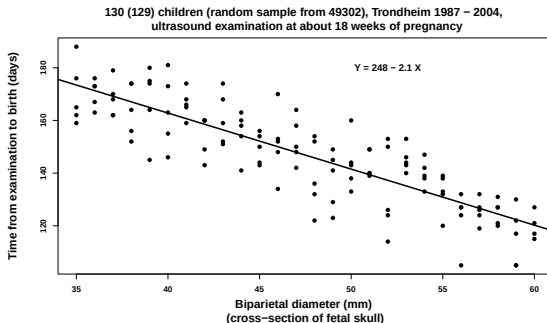
	2.5 %	97.5 %
(Intercept)	234.760243	237.676504
bpd	-1.934961	-1.871181

ULTRASOUND AT 18 WEEKS OF PREGNANCY: PREDICTION

What is our prediction (“best guess”) for e.g. a child with BPD 45 mm?

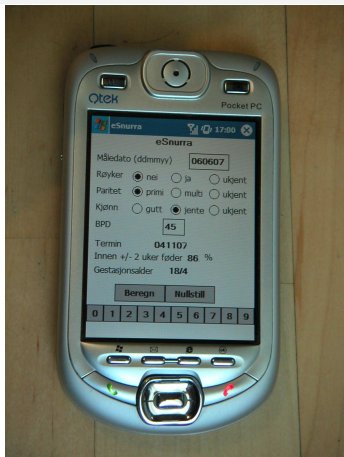
$$Y = 248 - 2.1 \cdot 45 = 153.5 \text{ days}$$

That is, we expect about 154 days (≈ 22 weeks) left until birth



NORWEGIAN PREDICTION SYSTEM

[HTTPS://WWW.NSFM.NO/ESNURRA](https://www.nsfm.no/esnurra)



ULTRASOUND AT 18 WEEKS OF PREGNANCY

MULTIPLE REGRESSION W/FEMUR LENGTH INCLUDED

Regression results:

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	247.9103	5.4915	45.14	<2e-16
bpd	-2.1278	0.1144	-18.59	<2e-16

Multiple R-squared: 0.7313

Regression results, femur length included:

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	248.4083	5.5442	44.805	<2e-16
bpd	-2.0348	0.1718	-11.845	<2e-16
femur	-0.1538	0.2115	-0.727	0.468

Multiple R-squared: 0.7324

ULTRASOUND AT 18 WEEKS OF PREGNANCY

FEMUR LENGTH ALONE

Regression results, femur length alone:

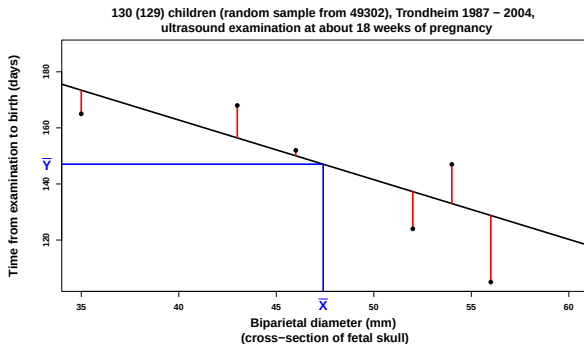
Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	211.4800	6.6387	31.856	<2e-16
femur	-2.0189	0.2044	-9.879	<2e-16

Multiple R-squared: 0.4345

(But femur length is better for prediction than this example would suggest)

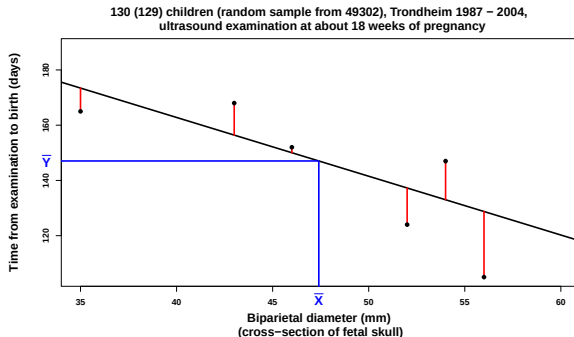
EXPLAINED VARIANCE: R^2 (R-SQUARED)



- Choose the values β_0 and β_1 that make the line “fit as best as possible” to the data
- I.e. those that make the Sum of Squares as small as possible:

$$\text{Sum of Squares} = \sum_i (Y_i - (\beta_0 + \beta_1 X_i))^2$$

EXPLAINED VARIANCE: R^2 (R-SQUARED)



Sum of Squares, Total: $SS_T = \sum_i (Y_i - \bar{Y})^2$

Total variation in data (i.e. around $\bar{Y}$)

Sum of Squares, Error: $SS_E = \sum_i (Y_i - (\beta_0 + \beta_1 X_i))^2$

Variation around the line $\beta_0 + \beta_1 X$

EXPLAINED VARIANCE: R^2 (R-SQUARED)

Definition, explained variance:

$$R^2 = \frac{\text{Explained variance}}{\text{Total variance}} = \frac{SS_T - SS_E}{SS_T}$$

In our data:

$$SS_T = 44296$$

$$SS_E = 11901$$

$$R^2 = \frac{44296 - 11901}{44296} = 0.731 \approx 73\%$$

Note:

$$\rho = -0.855$$

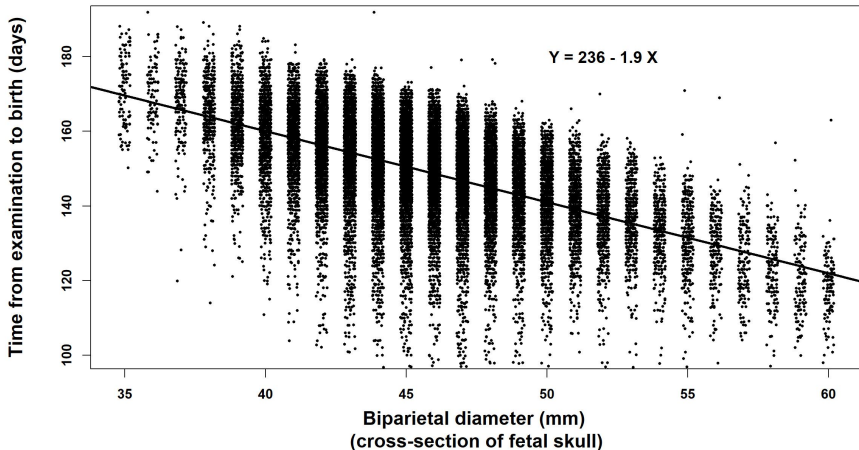
$$\rho^2 = 0.731!$$

... so R^2 can be calculated as the square of the correlation
(when you have only *one* X-variable)

SHORT NOTE ON NON-LINEAR REGRESSION

- No immediate reason to suspect non-linearity in our regression here:

All 49302 children, Trondheim 1987 - 2004,
ultrasound examination at about 18 weeks of pregnancy



SHORT NOTE ON NON-LINEAR REGRESSION

- No immediate reason to suspect non-linearity in our regression here:
- But we have a lot of data.... we might try!
- ... why not just do something simple?
- Let's use a second-degree polynomial:

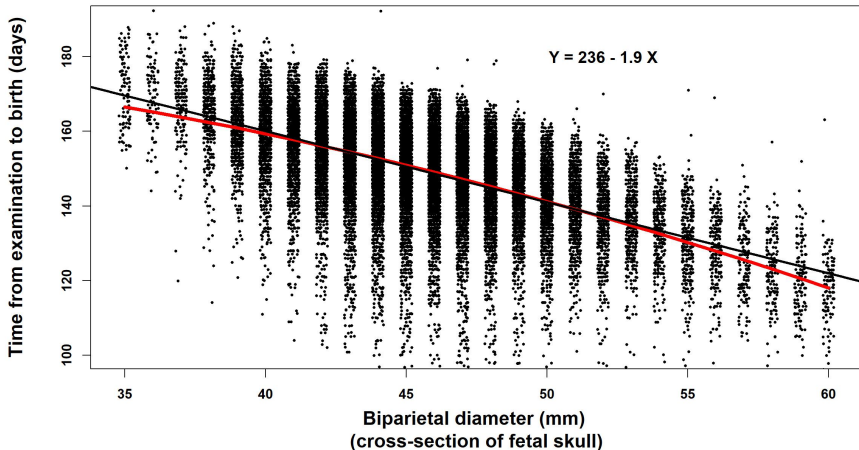
$$Y = \beta_0 + \beta_1 \cdot X + \beta_2 \cdot X^2$$

$$\text{remain} = \beta_0 + \beta_1 \cdot \text{bpd} + \beta_2 \cdot \text{bpd}^2$$

SHORT NOTE ON NON-LINEAR REGRESSION

- With second-degree polynomial regression:

All 49302 children, Trondheim 1987 - 2004,
ultrasound examination at about 18 weeks of pregnancy



SHORT NOTE ON NON-LINEAR REGRESSION

- Unfortunately, this is wrong!
- Big problem with polynomials...
- They are mostly determined by where there is a lot of data
- In this case, in the middle, around $BPD = 45\text{mm}$
- What happens other places will often be an artefact of the shape of the polynomial
- A second-degree polynomial (parabola) always curves either up or down

SHORT NOTE ON NON-LINEAR REGRESSION

Much better:

- Generalized Additive Model (GAM)

$$Y = \beta_0 + s(X)$$

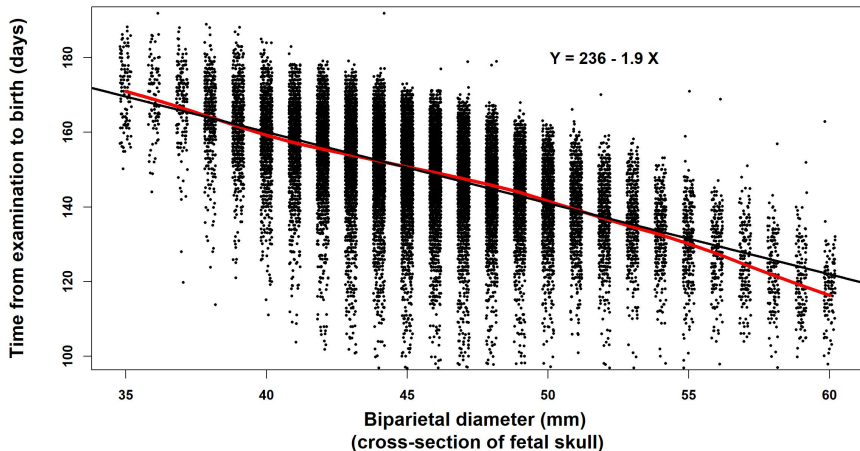
$$\text{remain} = \beta_0 + s(\text{bpd})$$

- where $s(X)$ is a "smooth spline" (or something like that) that is much more flexible than a polynomial
- And it is determined "locally", i.e. not only where there is a lot of data

SHORT NOTE ON NON-LINEAR REGRESSION

- With a GAM regression:

All 49302 children, Trondheim 1987 - 2004,
ultrasound examination at about 18 weeks of pregnancy

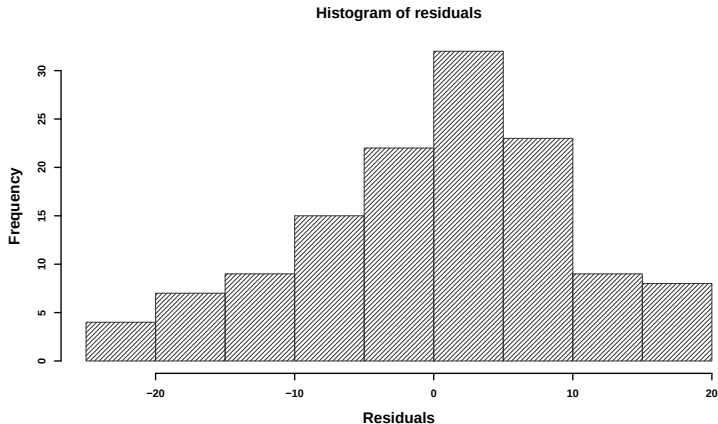


SHORT NOTE ON NON-LINEAR REGRESSION

- GAM-models are very useful and flexible
- Relatively easy to run
- But mostly in R

HISTOGRAM OF RESIDUALS, AS DIAGNOSTIC CHECK

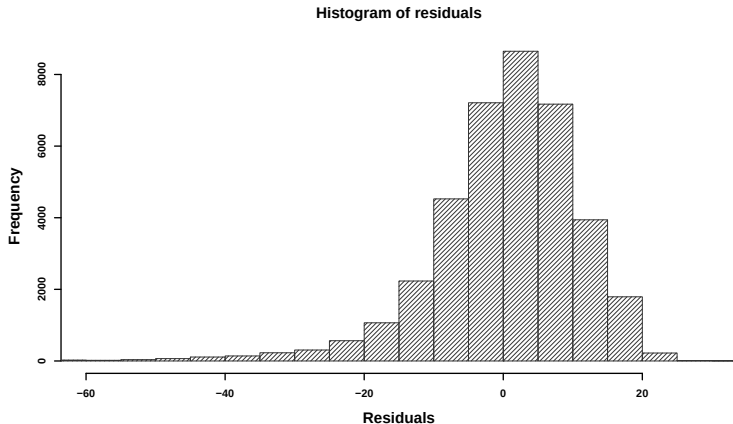
REDUCED DATA SET



Should be approximately Normally distributed!

HISTOGRAM OF RESIDUALS, AS DIAGNOSTIC CHECK

FULL DATA SET

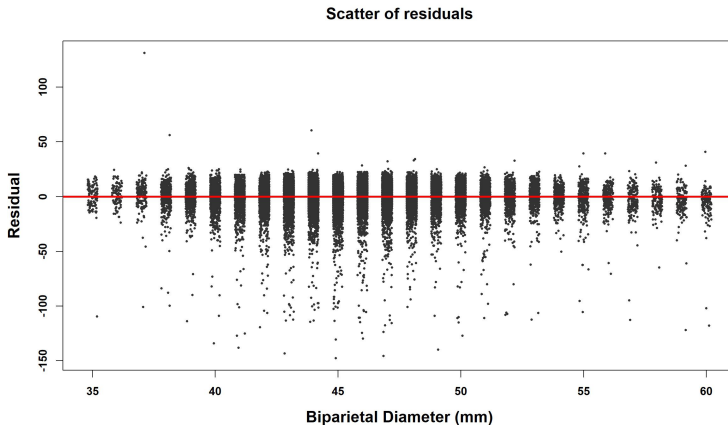


Should be approximately Normally distributed!

It's OK for this sample size!

SCATTER OF RESIDUALS, AS DIAGNOSTIC CHECK

FULL DATA SET



No signs of anything wrong, but outliers should be checked and removed

Note that large sample sizes make the plot seem wider