Chapter 8 Correlation and simple regression

Kolbjørn Engeland

koe@nve.no

UNIVERSITY OF OSLO

What is correlation?

- Measures linear relationships between variables
- Auto-correlation
- Correlogram

UNIVERSITY OF OSLO

Pearson correlation coefficient

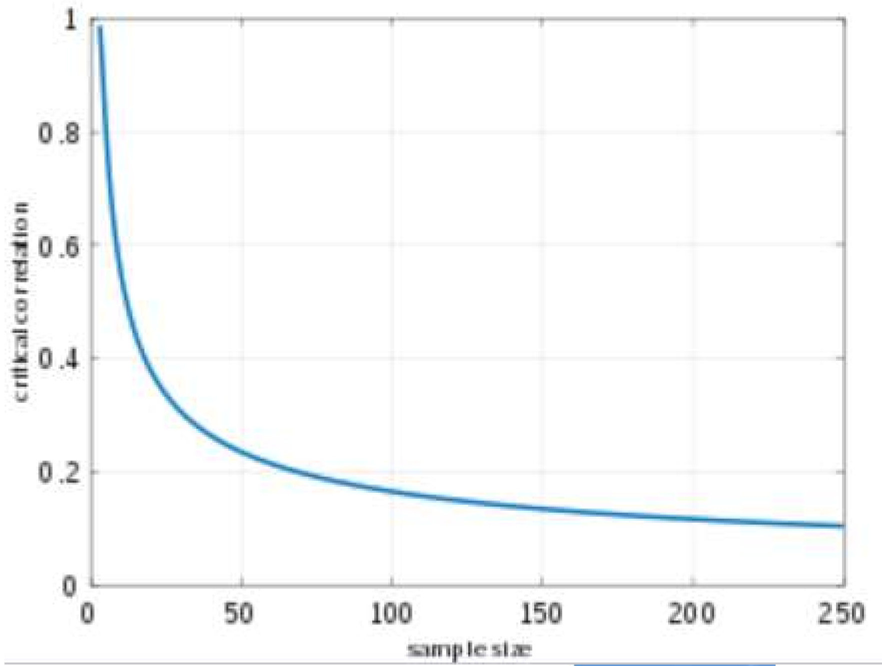
$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

- Significance of correlation: t-test with n-2 degress of freedom

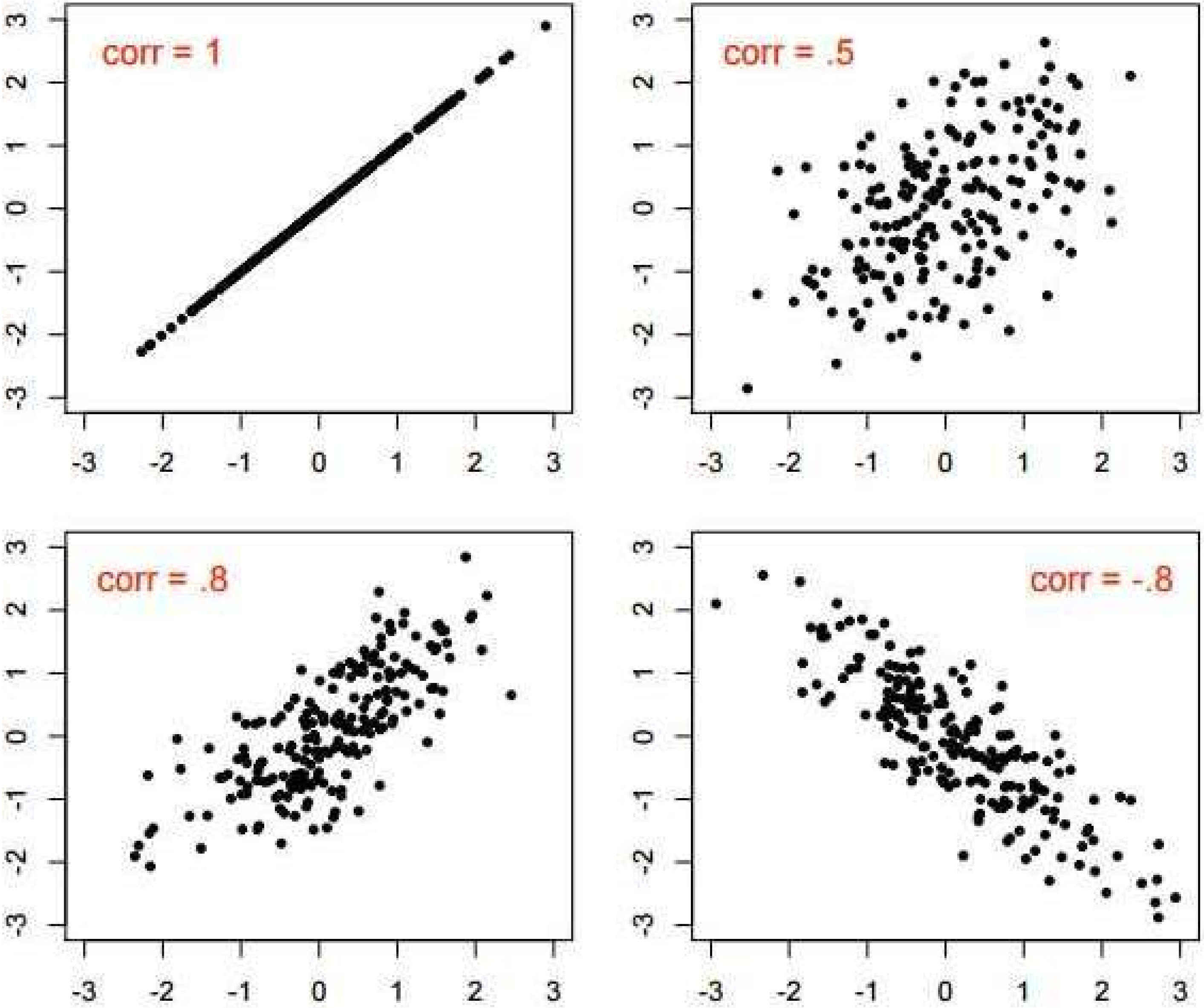
$$t = r\sqrt{\frac{n-2}{1-r^2}}$$

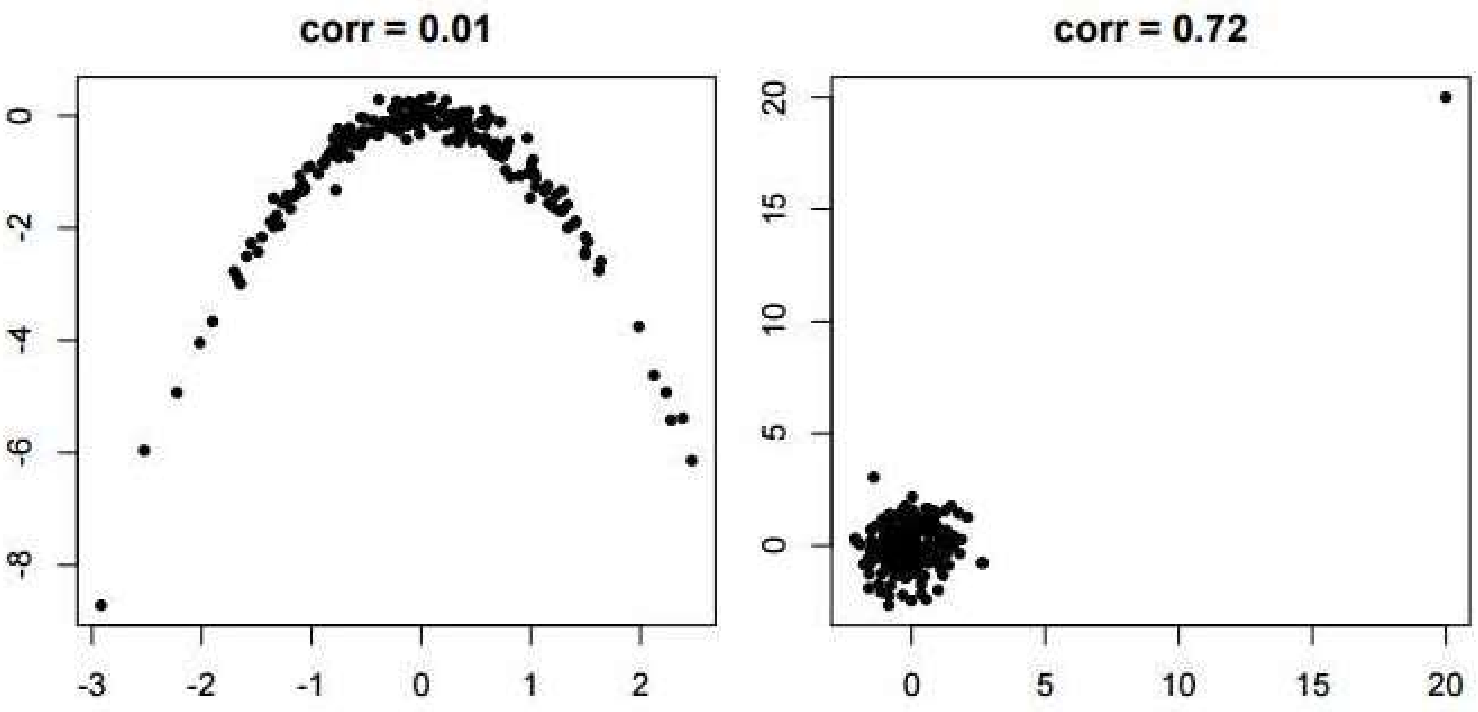
- Critical correlation as a function of sample size:

$$r = \frac{t}{\sqrt{n-2+t^2}}.$$



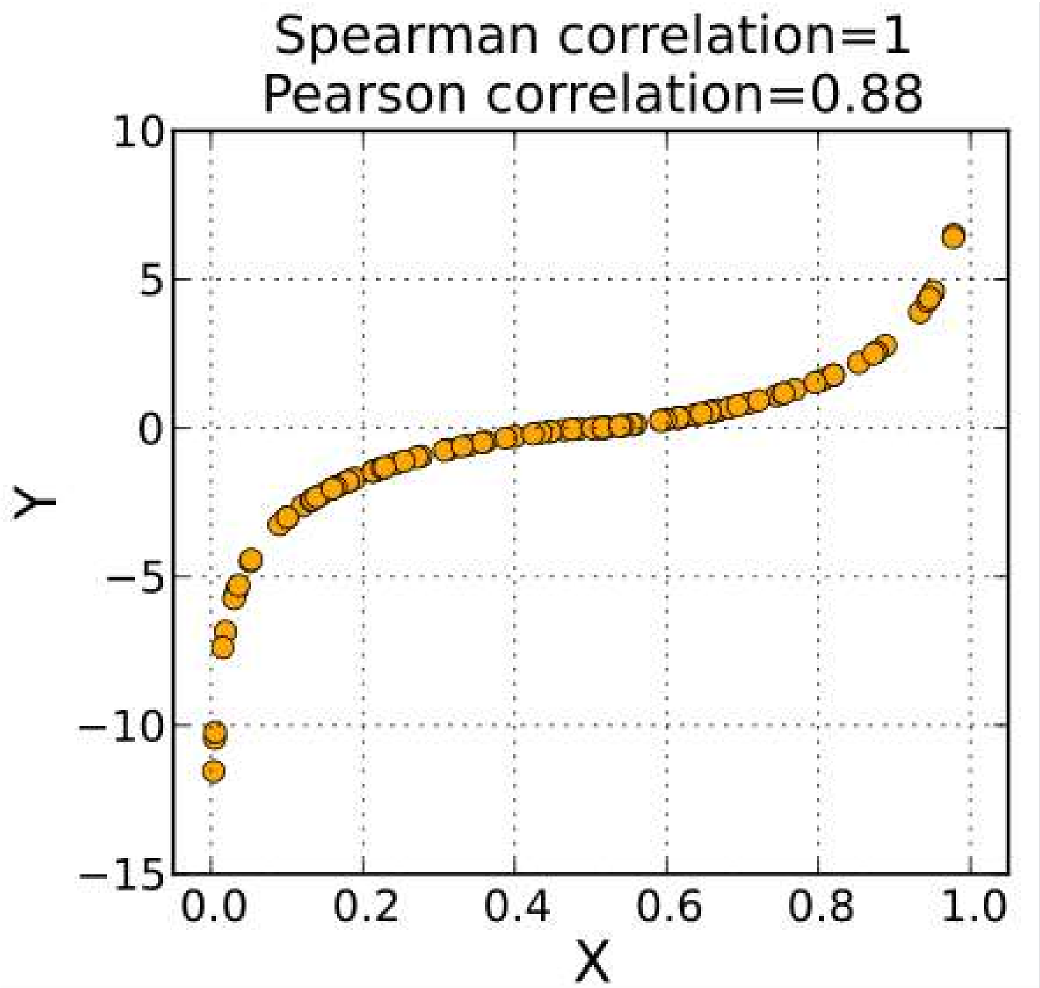
UNIVERSITY OF OSLO





Also be careful with influential observations.

Spearman rank correlation



Mann Kendall

$$\tau = \frac{(number\ of\ concordant\ pairs) - (number\ of\ discordant\ pairs)}{\left(\frac{n(n-1)}{2}\right)}$$

$$\tau = \frac{2\sum_{i<j} sgn(x_i - x_j) sgn(y_i - y_j)}{n(n-1)}$$

$$\mathrm{E}(\tau) = 0 \qquad \mathrm{Var}(\tau) = \frac{2(2n+5)}{9n(n-1)}$$

The regression equation

- $\boldsymbol{y} = a + b\boldsymbol{x} + \varepsilon$
- Standard assumption: $\varepsilon \sim N(0, \sigma^2)$

- A prediction of y is:
- $\hat{y}_i = \hat{a} + \hat{b}x_i + \varepsilon$

- The variance of the prediction is
- $Var(\hat{y}) = Var\left(\hat{a} + \hat{b}x_i\right) + \sigma^2$

Why simple regression?

- Analyze
  - Understand hydrology
    - Trends
    - What is controlling key variables?
      - Catchment properties controlling flood, droughts
      -
  - NB: Correlation does not imply causality
- Predict
  - In time
  - In space

What is simple regression?

- Conditional expectation
  - E(Y|X)
- Best linear fit between two variables

$$M = \sum e_i^2 = \sum (Y_i - \hat{Y}_i)^2 = \sum (Y_i - a - bX_i)^2 \qquad \frac{\partial M}{\partial a} = 0, \quad \frac{\partial^2 M}{\partial a^2} > 0$$

$$\frac{\partial M}{\partial b} = 0, \quad \frac{\partial^2 M}{\partial b^2} > 0$$

Estimate of the regression coefficients

$$\hat{b} = \frac{Cov(\boldsymbol{x}, \boldsymbol{y})}{Var(\boldsymbol{x})}$$

$$\hat{a} = \bar{y} - \hat{b}\bar{x}$$

$$\hat{\sigma} = \sqrt{\frac{SSE}{n-2}}$$

UNIVERSITY OF OSLO

Standard error of coefficients

$$\hat{b} \sim t_{n-2}\left(\hat{b}, \hat{\sigma}_b^2\right) \qquad \hat{a} \sim t_{n-2}(\hat{a}, \hat{\sigma}_a^2)$$

$$\hat{\sigma}_b^2 = \frac{\hat{\sigma}^2}{\sum(x_i - \bar{x})^2} \qquad \hat{\sigma}_a^2 = \hat{\sigma}^2\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum(x_i - \bar{x})^2}\right)$$

$$Cov\left(\hat{a}, \hat{b}\right) = -\frac{\overline{\hat{\sigma}^2 x}}{\sum(x_i - \bar{x})^2}$$

Hypotheis testing of coefficients:

- Test if b is significantly different from zero!
- Use t-test (n-2 degrees of freedom) since the variance is unknown.

Standard error of regression line

- Variance for the regression line:

$$\hat{\sigma}_{y_r}^2 = \hat{\sigma}^2\left(\frac{1}{n} + \frac{(x-\bar{x})^2}{\sum(x_i-\bar{x})^2}\right)$$

- Prediction: Need to add residual variance:

$$\hat{\sigma}_{y_r}^2 = \hat{\sigma}^2\left(1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{\sum(x_i-\bar{x})^2}\right)$$

Maximum likelihood estimation

- Maximize the likelihood of the observations

Assumptions

- Linearity
- Normality
- Homoscedasity
- Independence
- iid !!!!

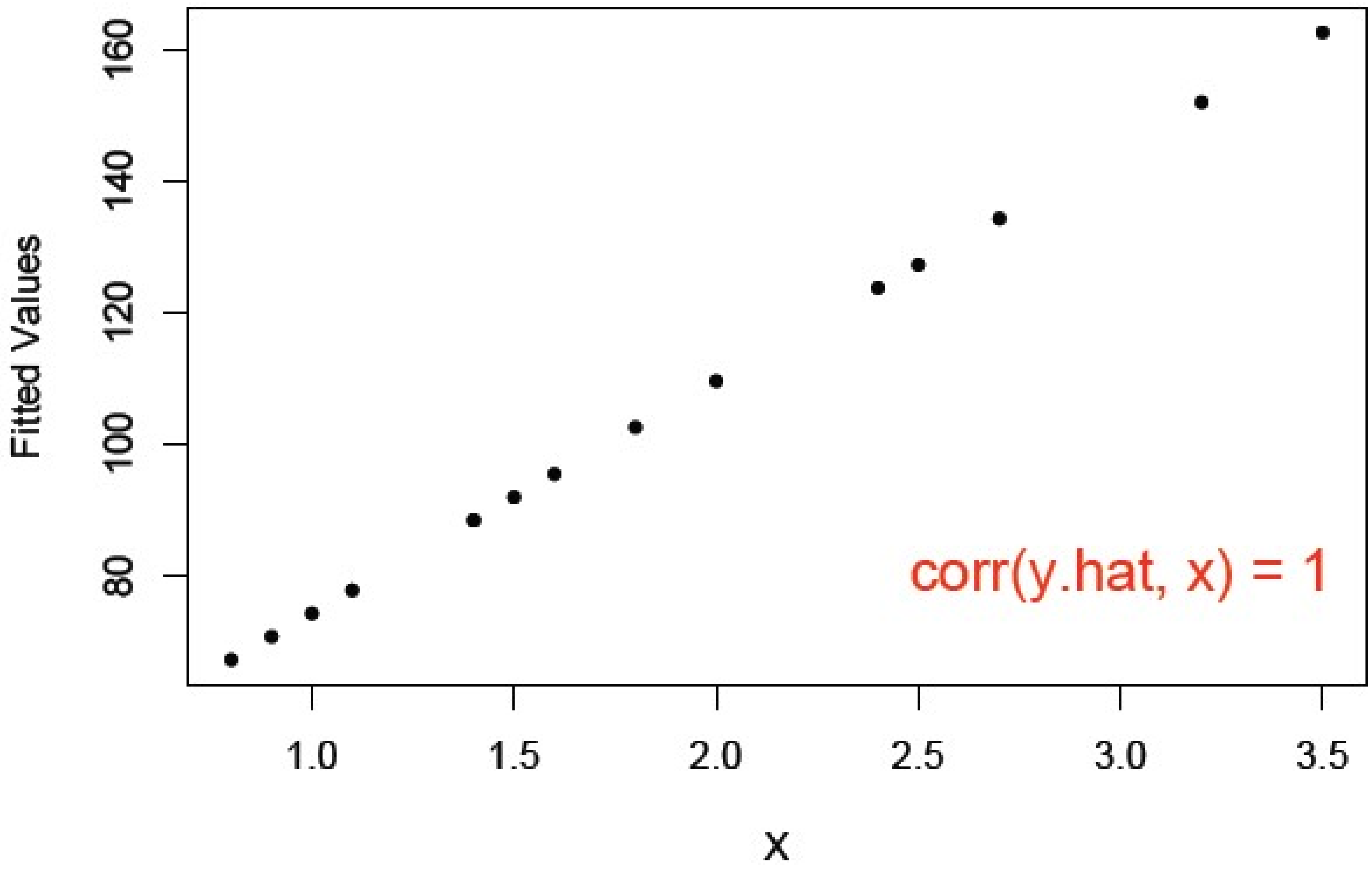
How to evaluate assumtions?

- Linearity: scatter plot
- Normality: Histogram, qq-plot, KS test
- Independence: auto-correlation
- Homoscedadisity. Scatter plot of residuals

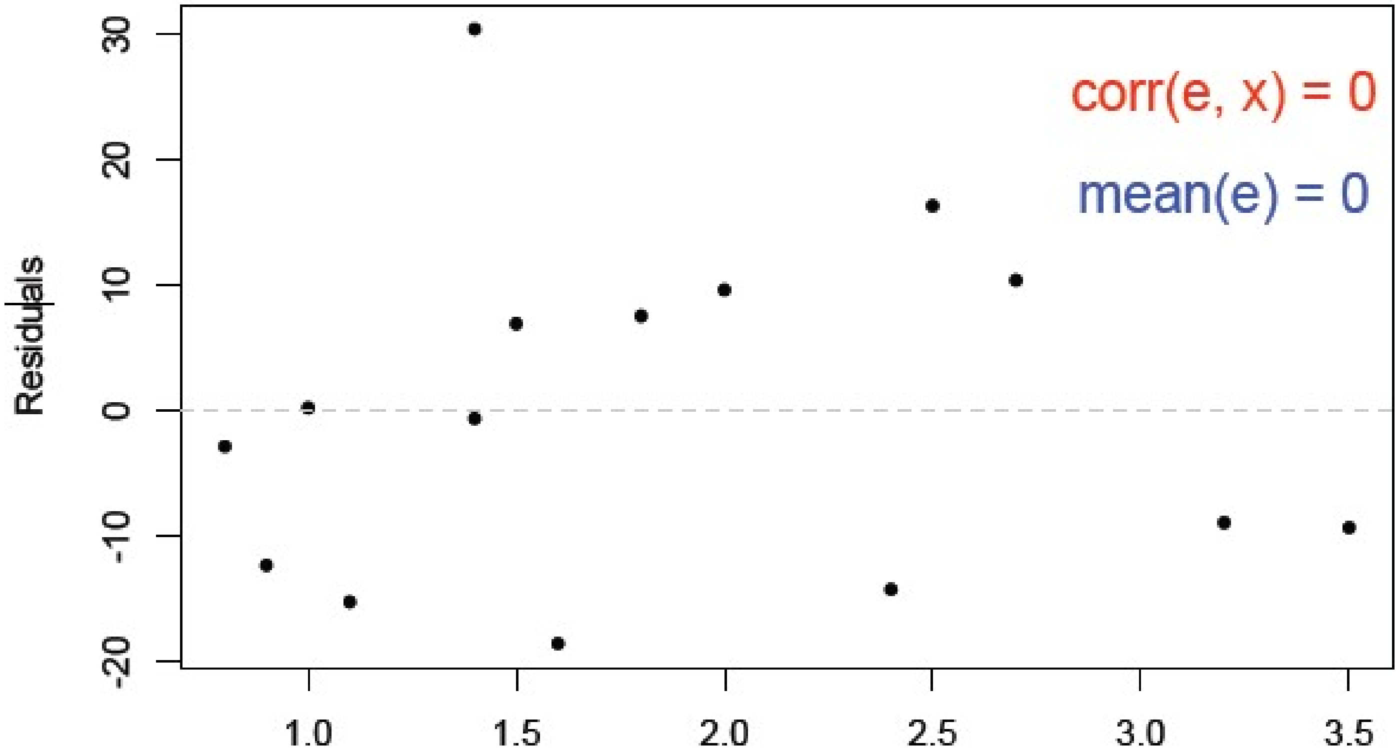
What if assumtions fails?

- Bias in predictions
- Wrong prediction variance

The Fitted Values and X



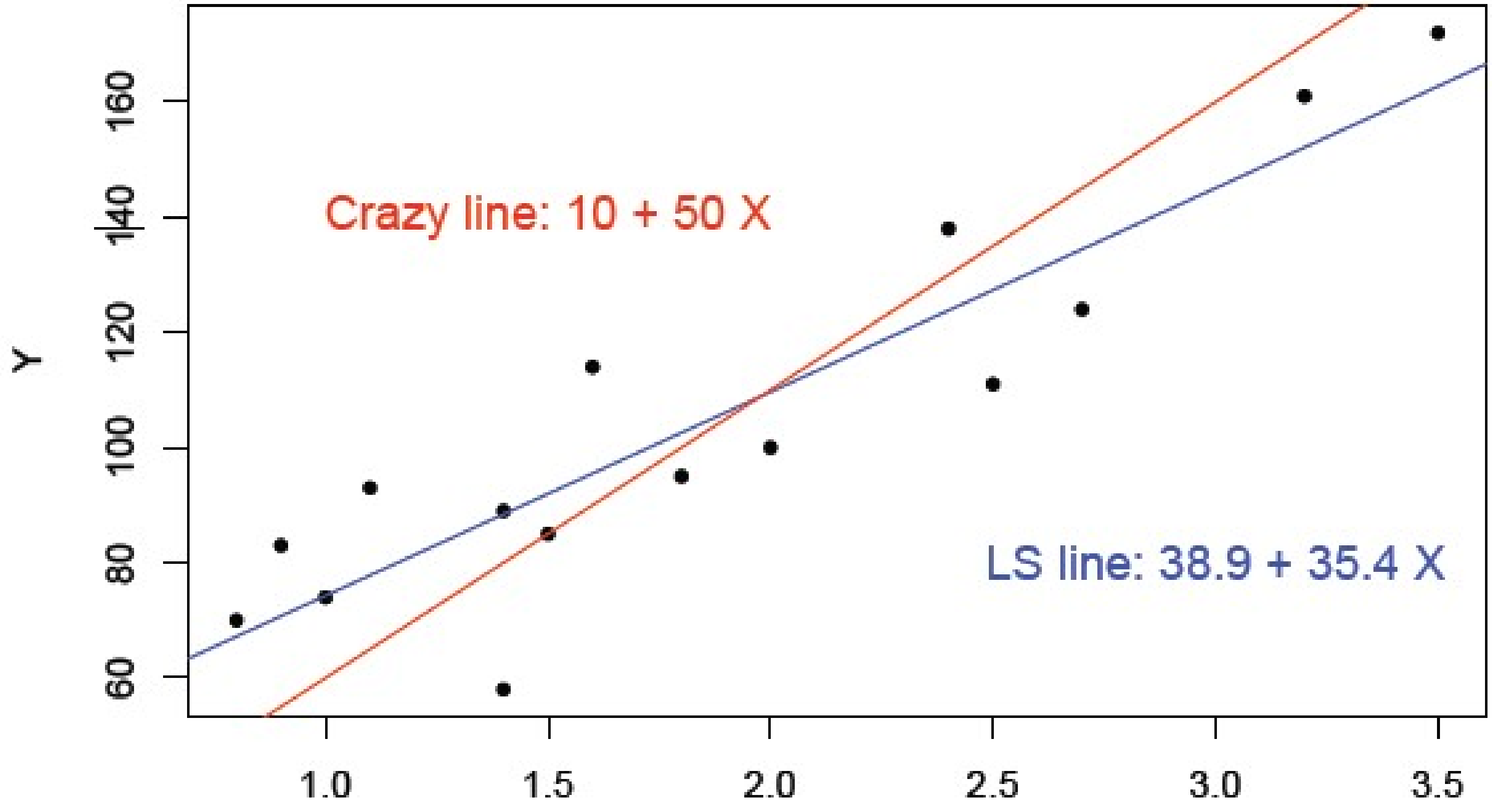
The Residuals and X



Why?

What is the intuition for the relationship between $\hat{Y}$ and e and

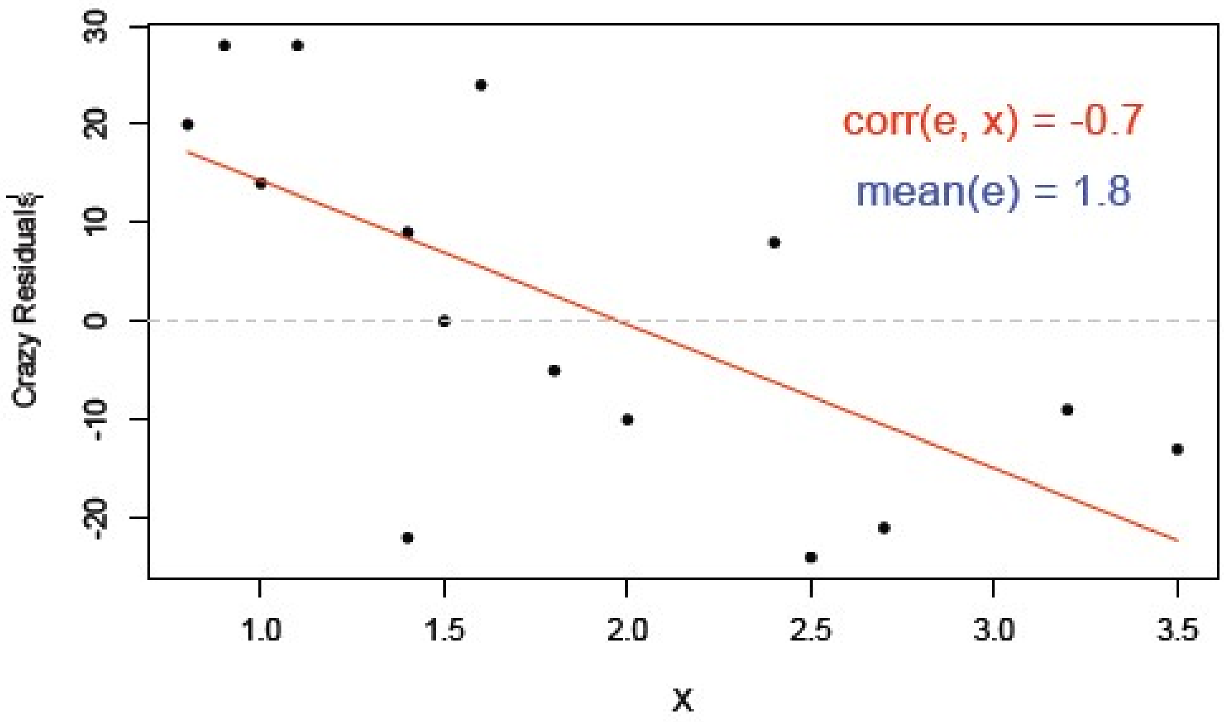
Lets consider some "crazy" alternative line:



Fitted Values and Residuals

This is a bad fit! We are underestimating the value of sm

and overestimating the value of big houses.



Clearly, we have left some predictive ability on the table!

Explained variance

- *Decomposing the variance:*

$$Var(y) = Var(\hat{y}) + Var(\varepsilon)$$

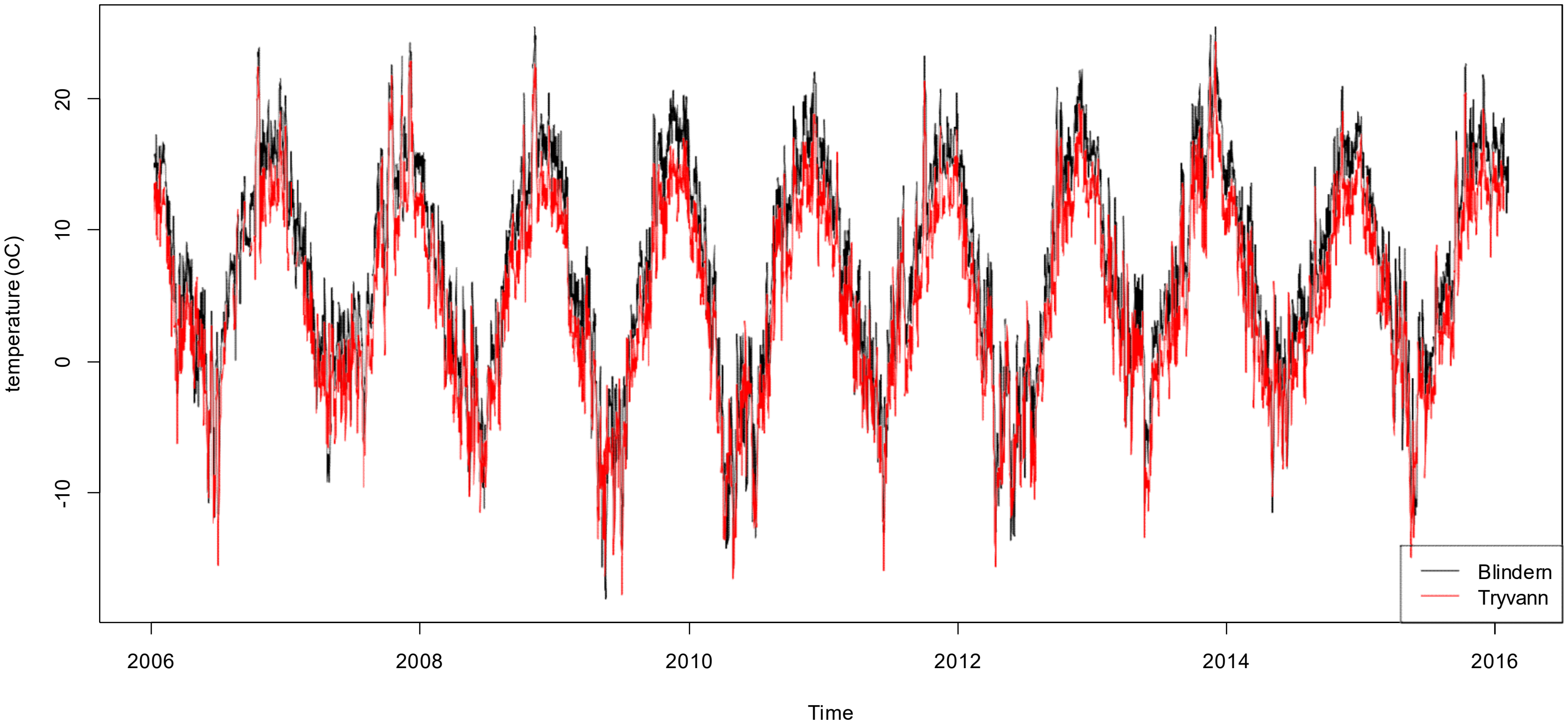
- SST = SSR + SSE

Explained variance

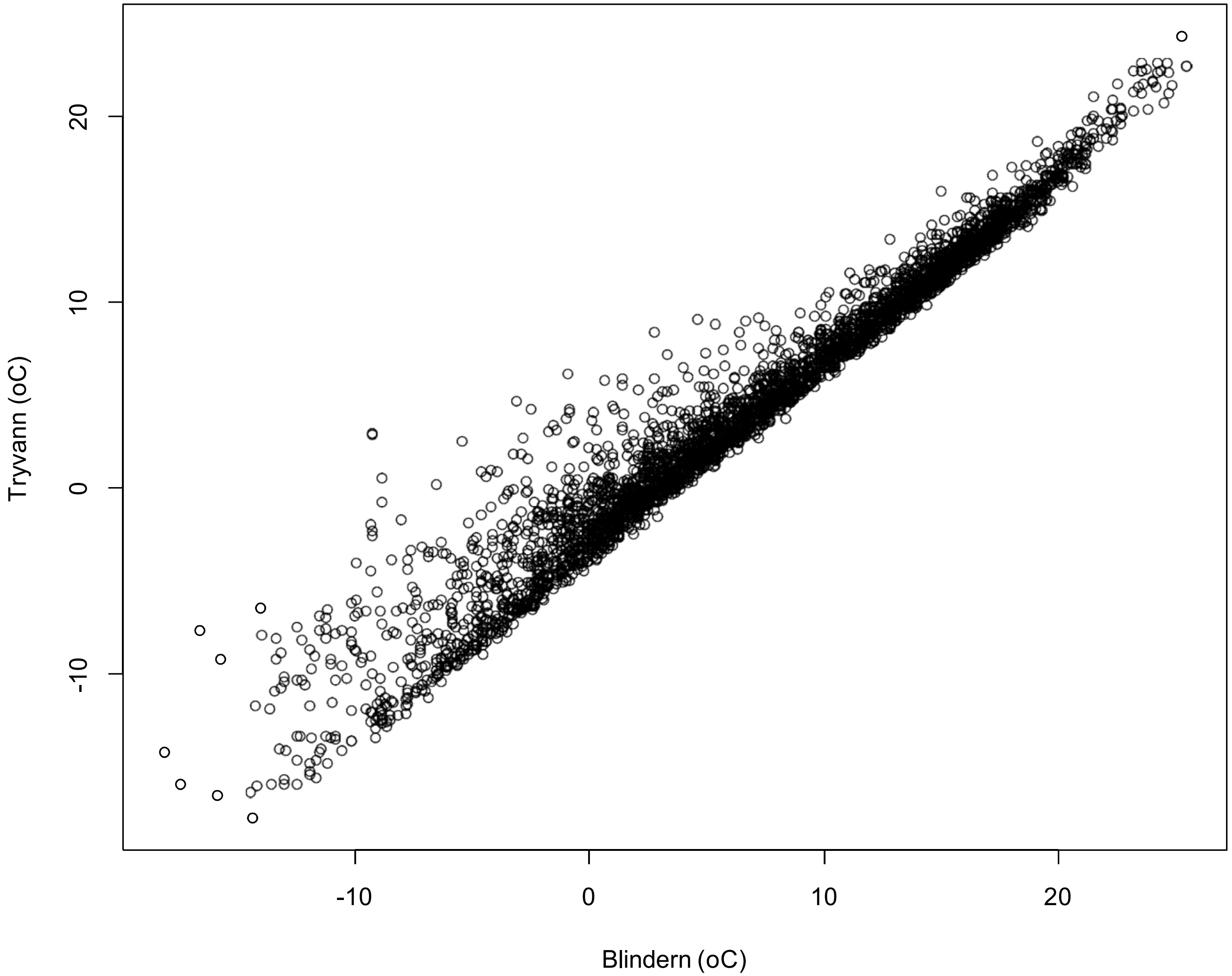
- Explained variation / total variation
- R^2 = SSR / SST = (SST – SSE) / SST = 1- SSE/SST
- Will be between 0 and 1
- Best model: $R^2 = 1$

UNIVERSITY OF OSLO

Example, temperatures at Blindern and Tryvann



Corr=0.979

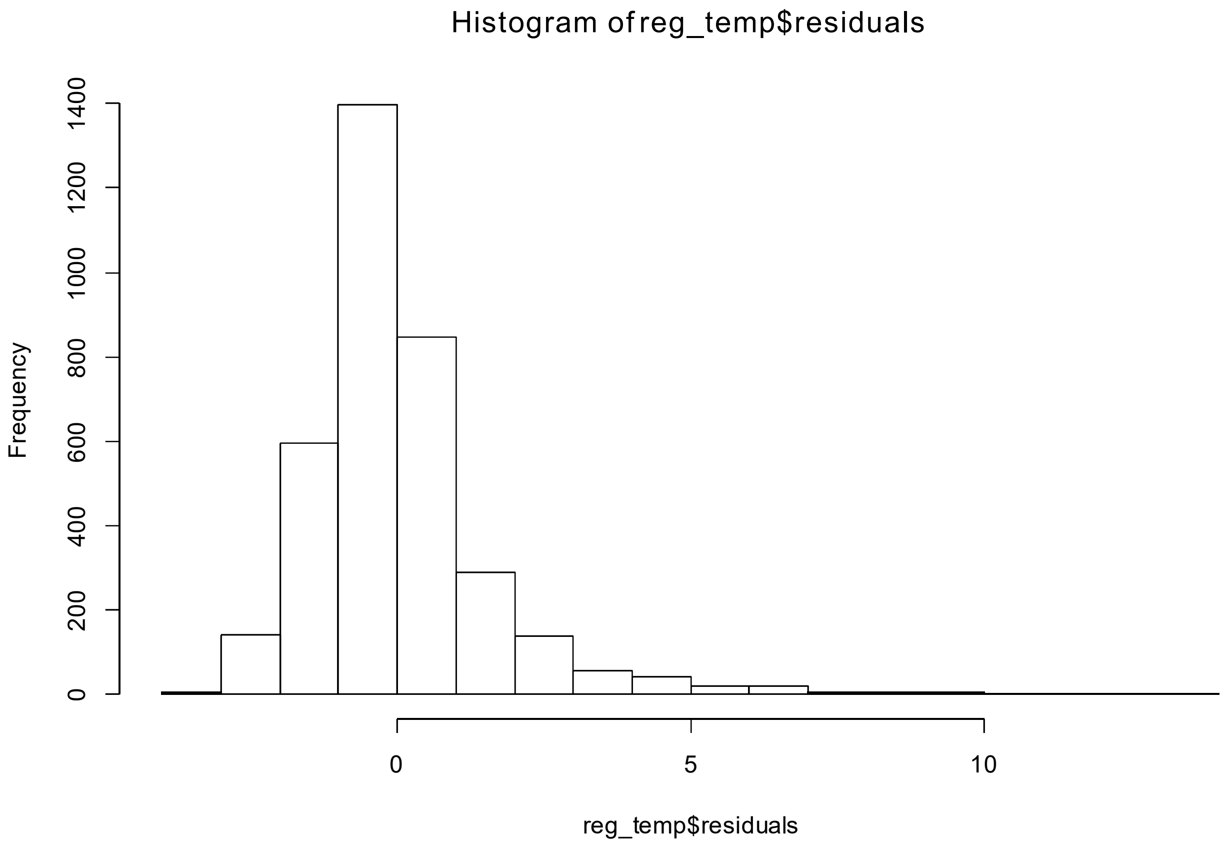


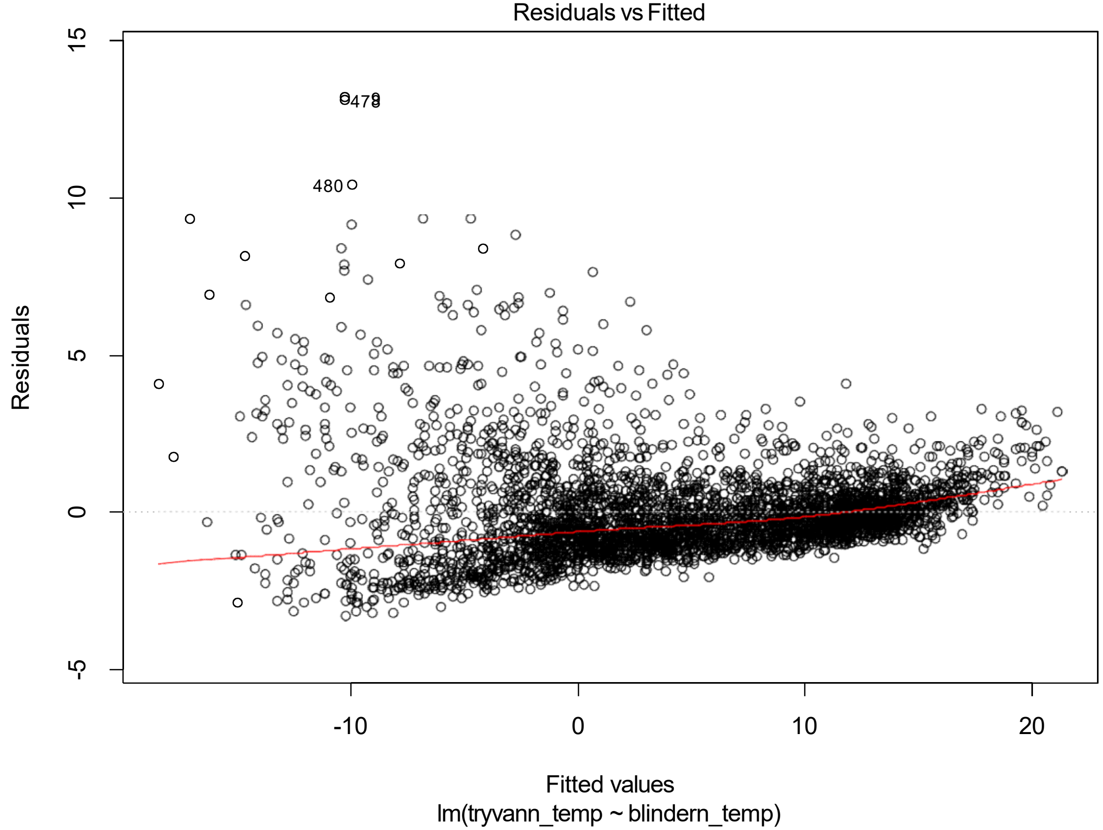
The fit

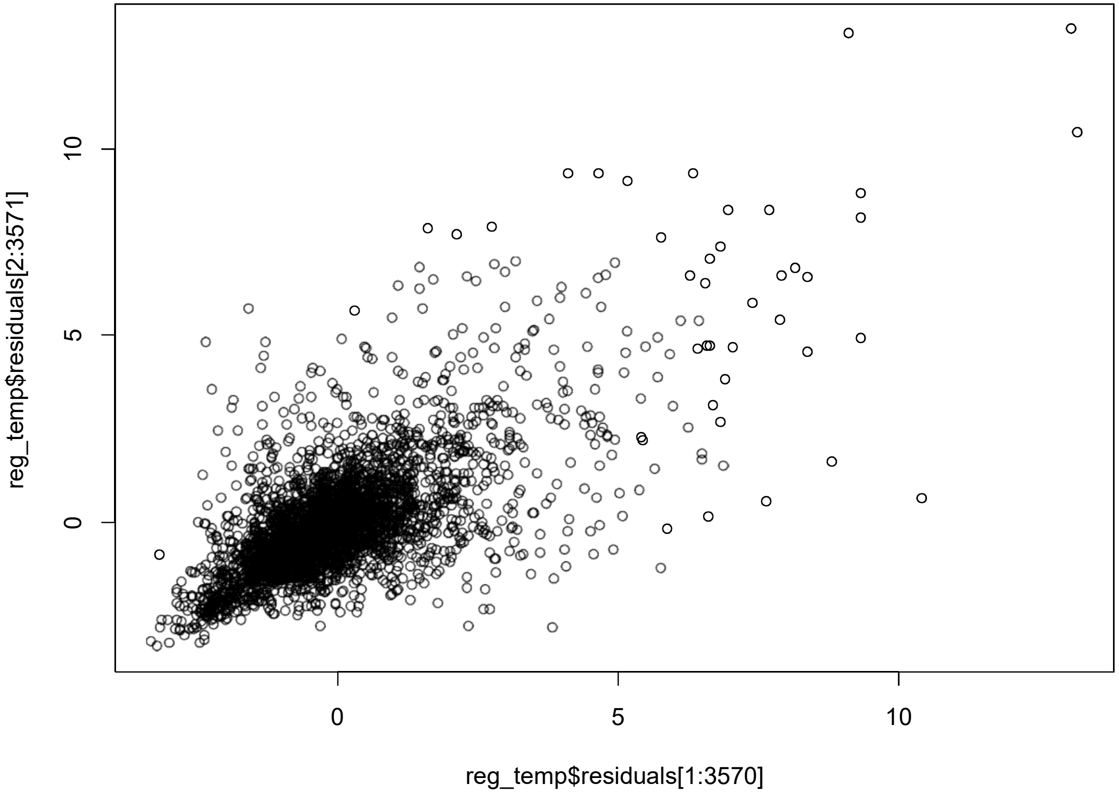
	Estimate	Std. Error	t value	Pr(>|t|)
a :	-1.888734	0.034181	-55.26	<2e-16 ***
b :	0.912494	0.003176	287.30	<2e-16 ***

- s : 1.54

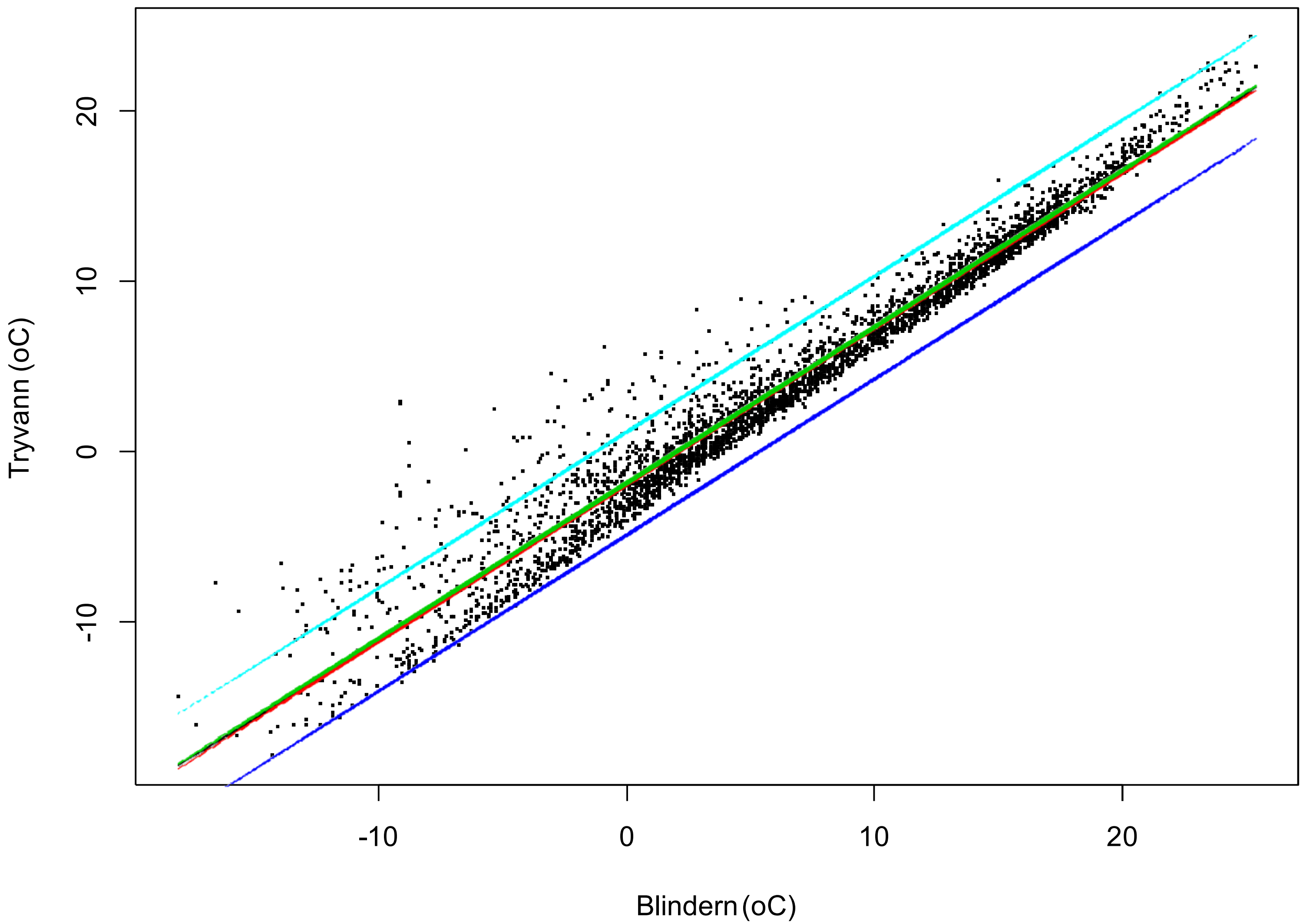
Histogram of residuals



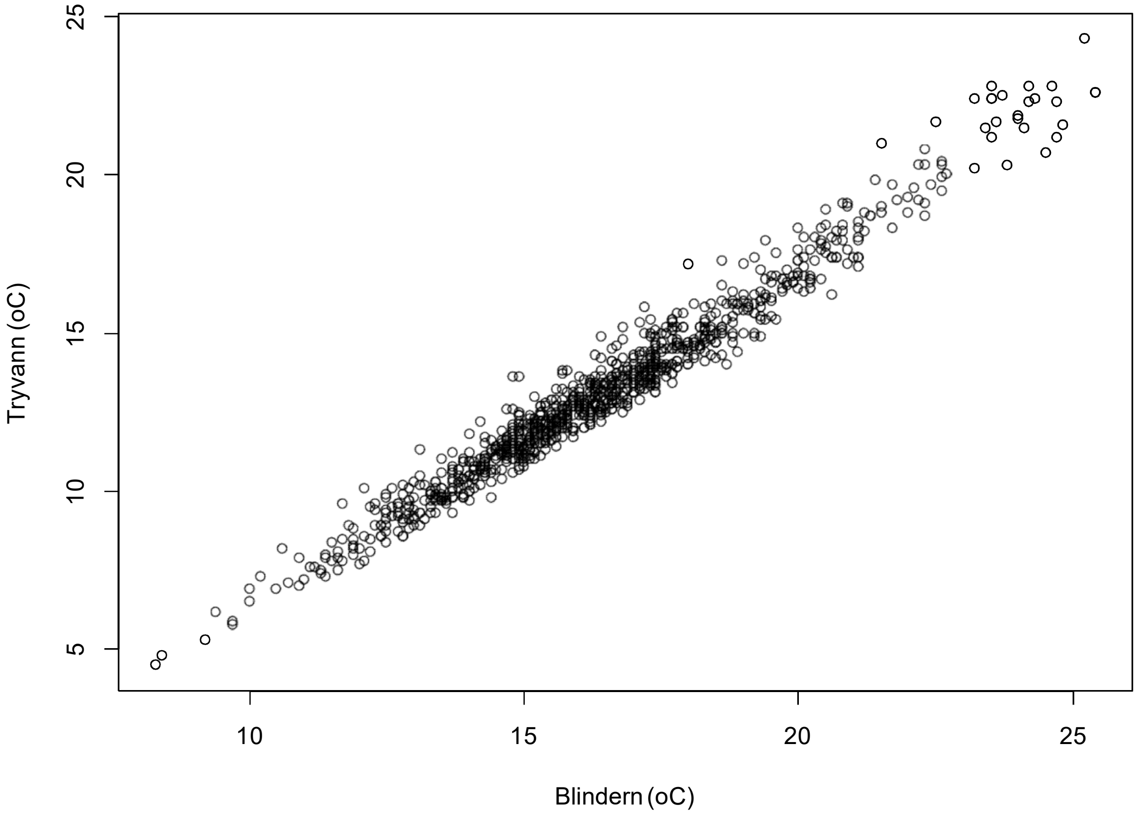

Residuals vs Fitted
Residuals
15
10
5
0
-5
-10
0
10
20
479
480
Fitted values
lm(tryvann_temp ~ blindern_temp)



reg_temp$residuals[2:3571]
reg_temp$residuals[1:3570]
0
5
10
0
5
10


Tryvann (oC)
20
10
0
-10
-10
0
10
20
Blindern (oC)

Include only June -August

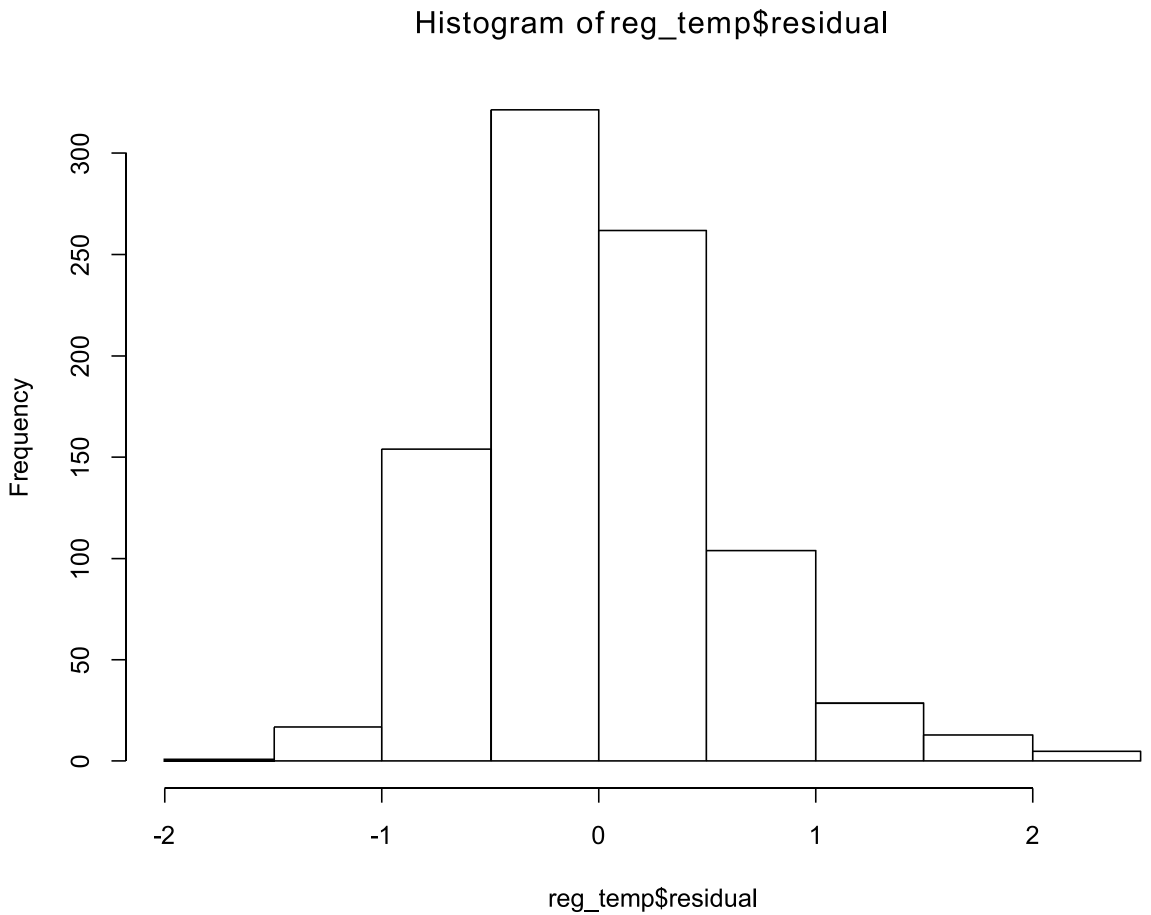


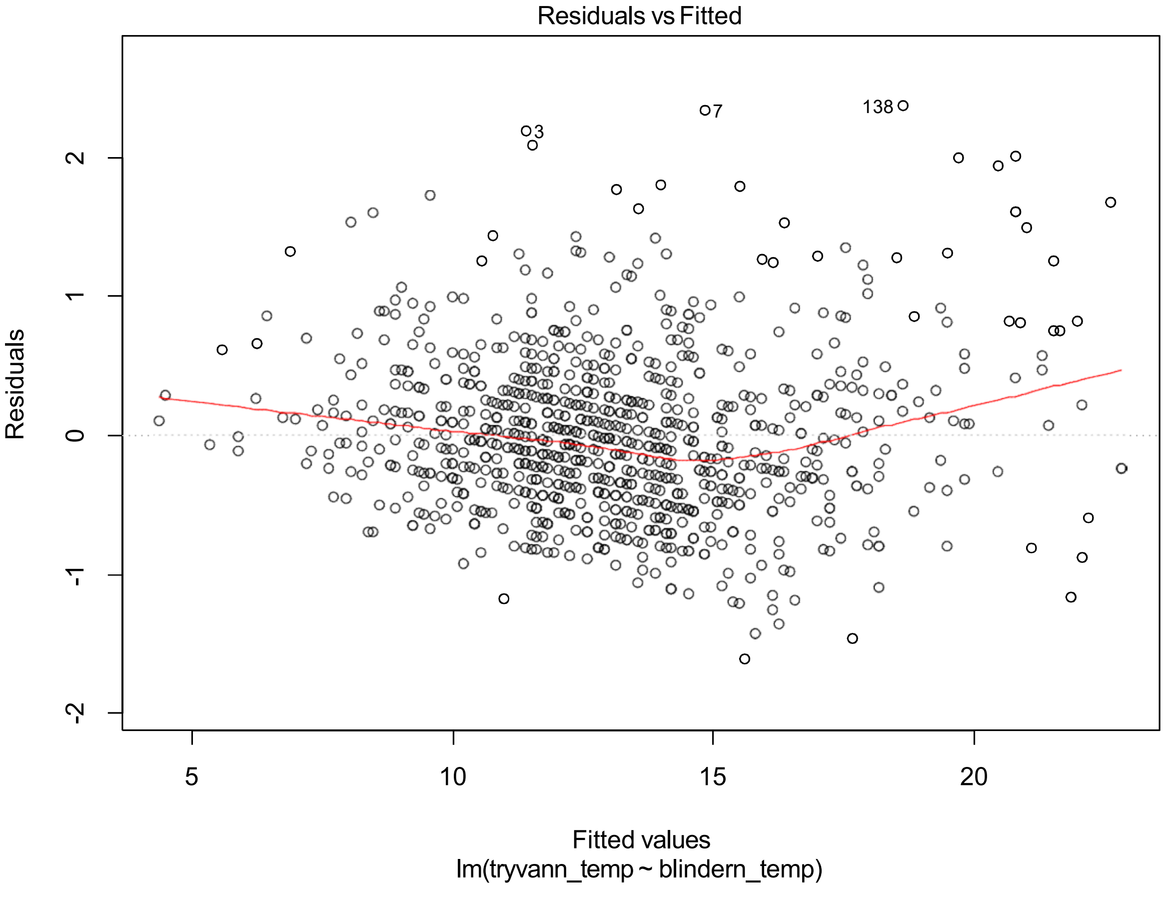
UNIVERSITY OF OSLO

The fit

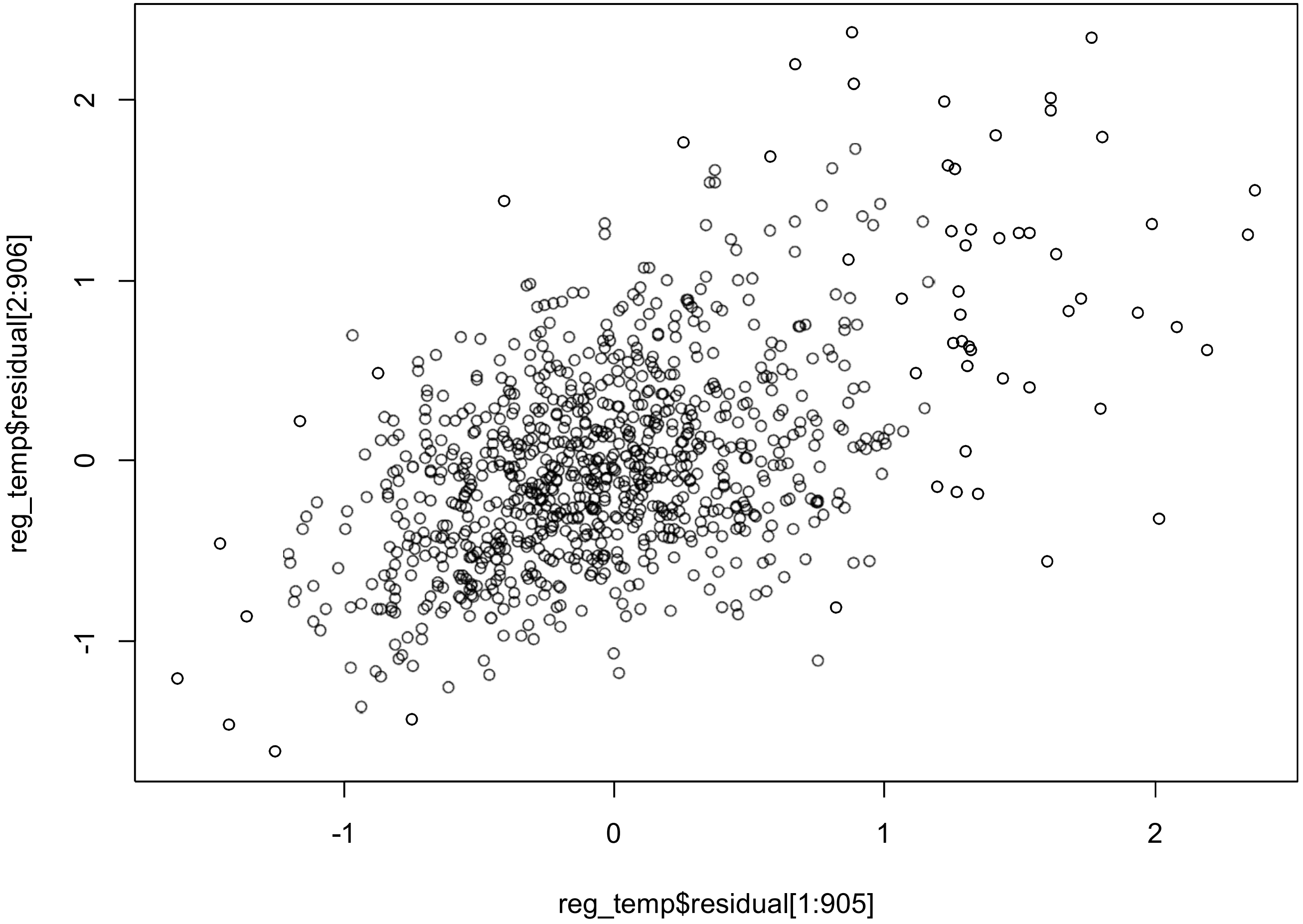
	Estimate	Std. Error	t value	Pr(>|t|)
a :	-4.5464	0.11587	-39.24	<2e-16 ***
b :	1.078	0.006954	155.02	<2e-16 ***

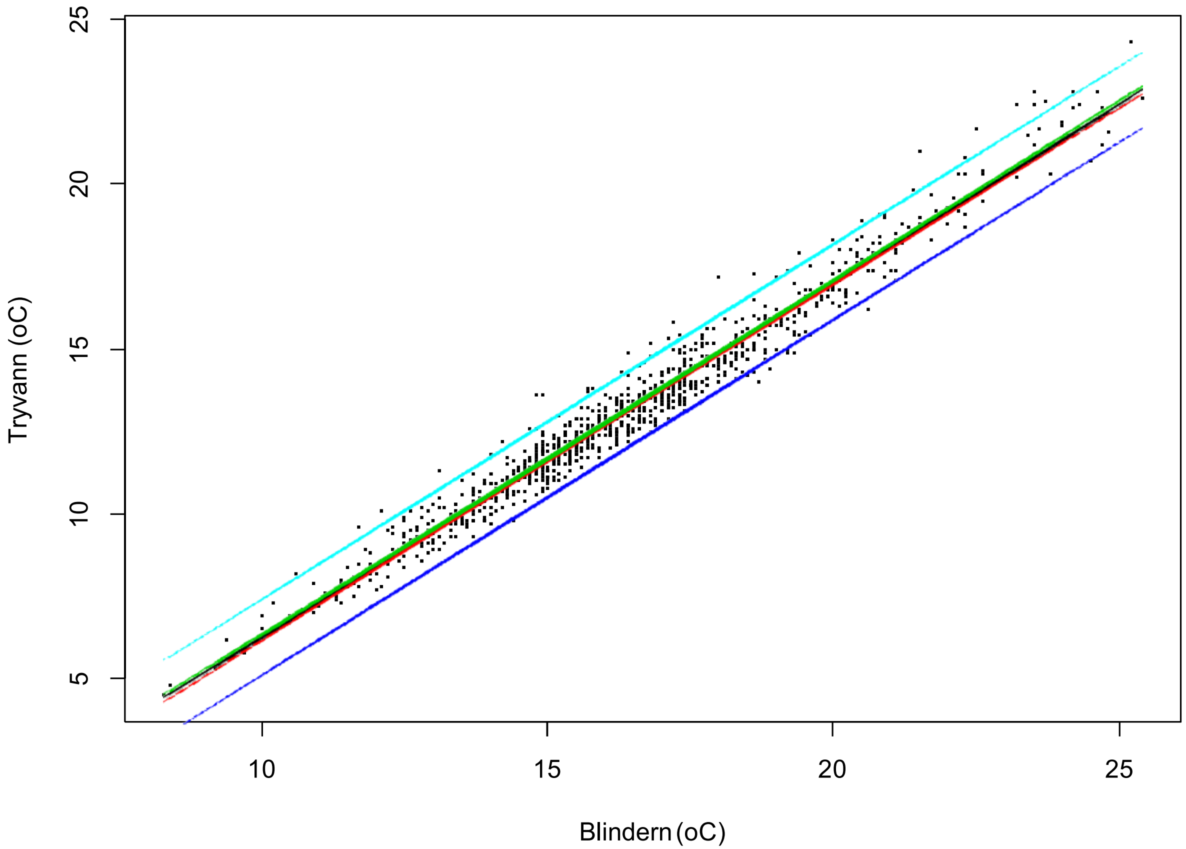
- s : 0.5827


Histogram of reg_temp$residual
Frequency
0
50
100
150
200
250
300
-2
-1
0
1
2
reg_temp$residual


Residuals vs Fitted
Residuals
2
1
0
-1
-2
5
10
15
20
Fitted values
lm(tryvann_temp ~ blindern_temp)
3
7
138

Cor=0.51




25
20
15
10
5
Tryvann (oC)
10
15
20
25
Blindern (oC)

The fit – use every second value

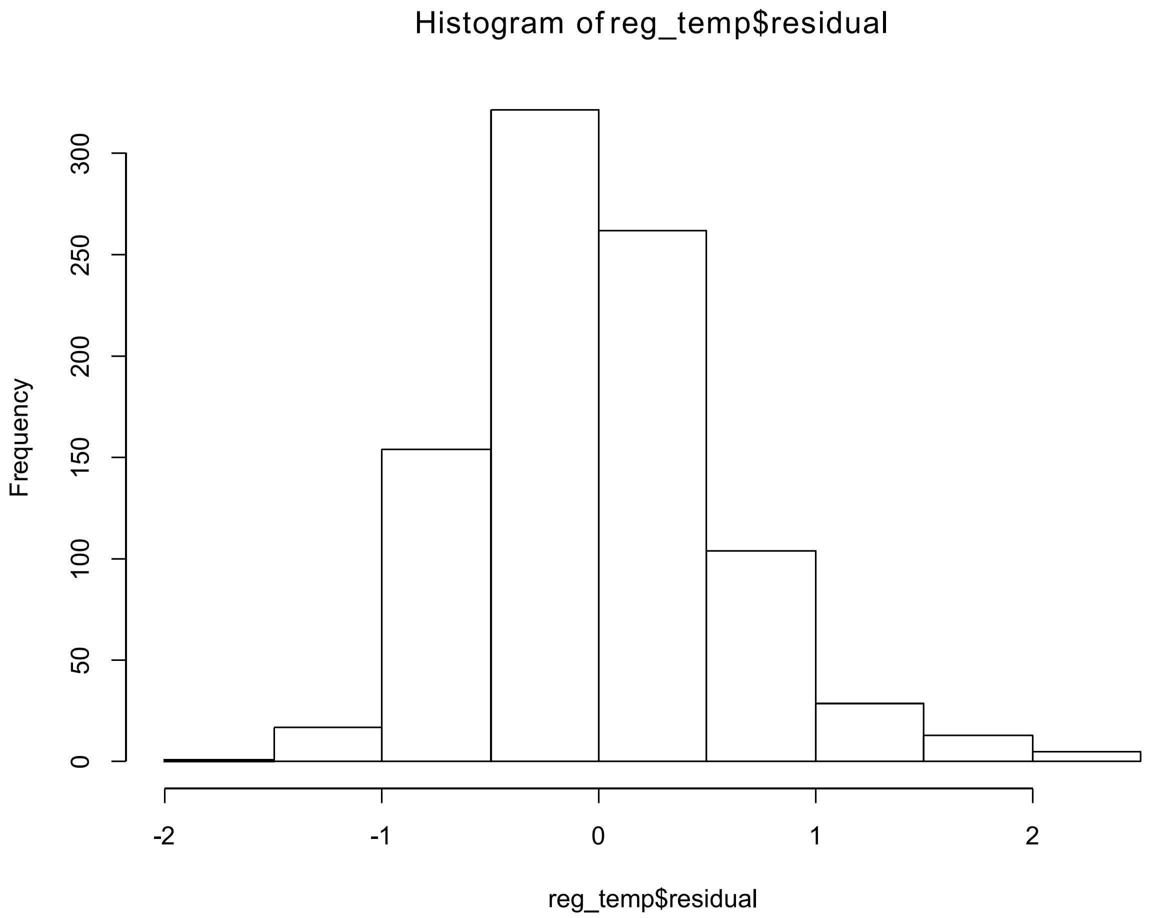
- Estimate Std. Error t value Pr(>|t|)
- a : -4.5464 0.11587 -39.24 <2e-16 ***
- b : 1.078 0.006954 155.02 <2e-16 ***

- s : 0.5827

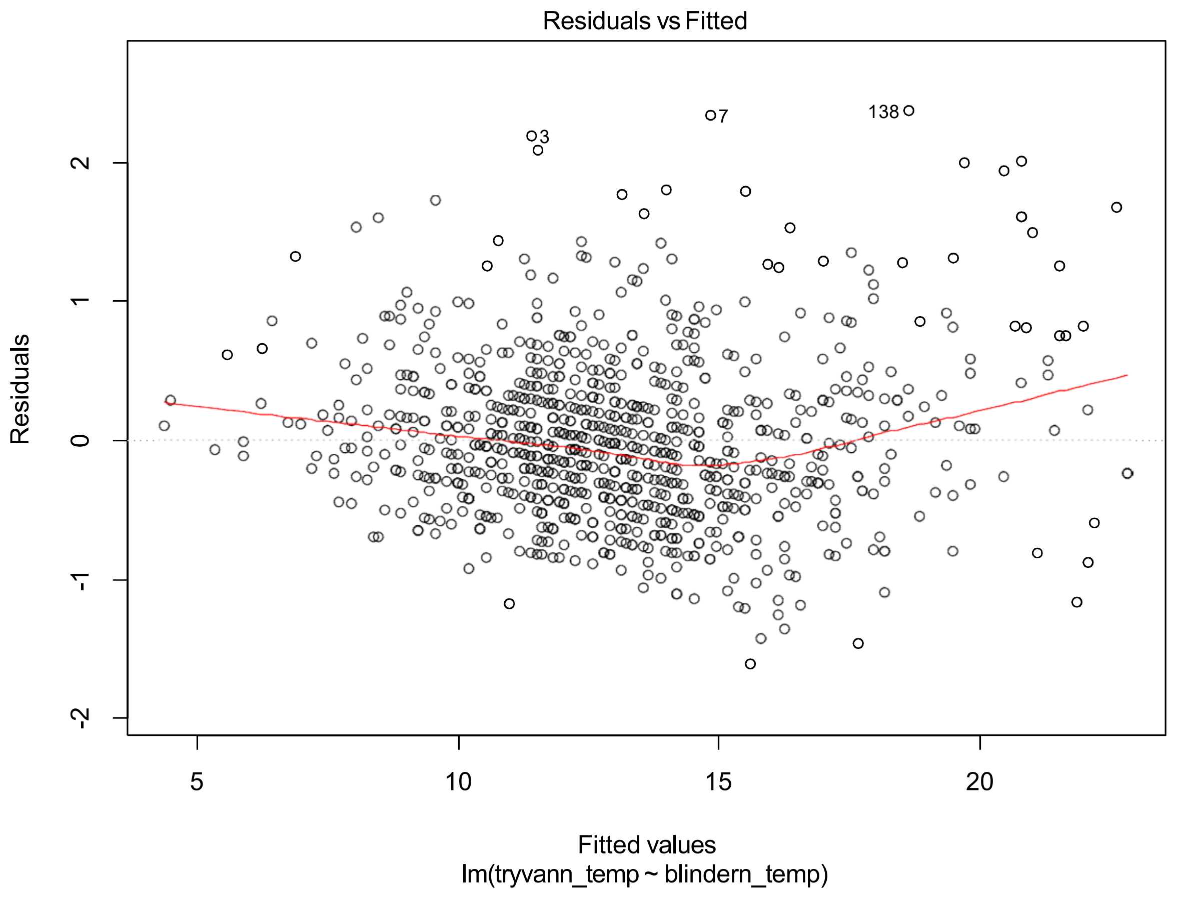
- Estimate Std. Error t value Pr(>|t|)
- a : -4.445 0.1644 -27.03 <2e-16 ***
- b : 1.0721 0.00986 108.71 <2e-16 ***

- s : 0.5827

UNIVERSITY OF OSLO

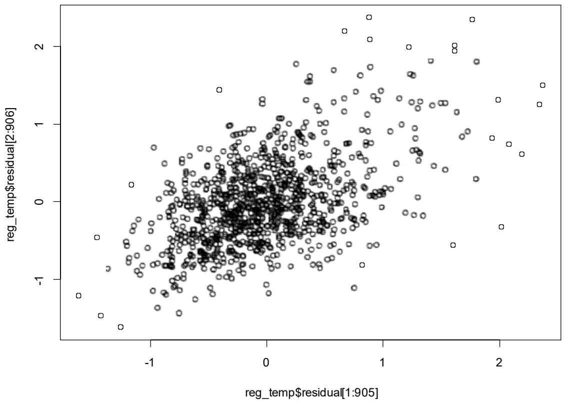


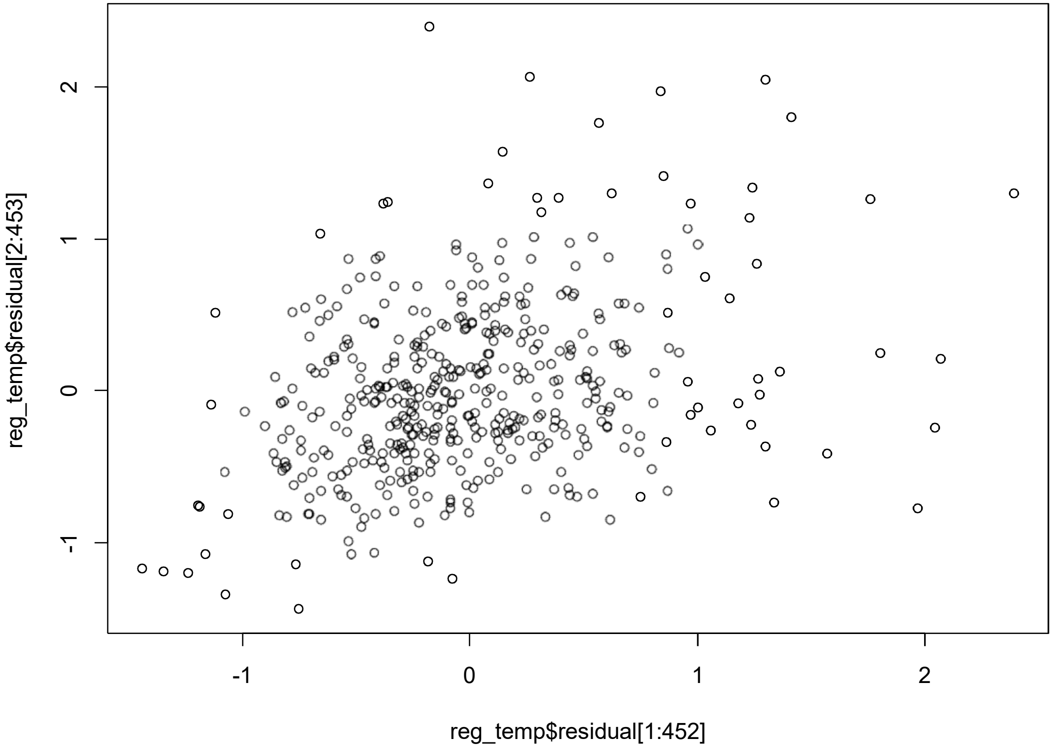
UNIVERSITY OF OSLO



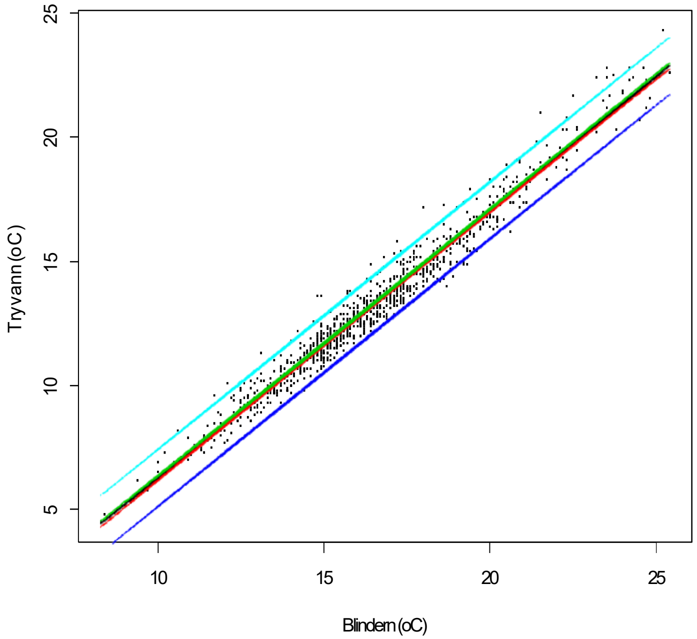
Cor=0.51

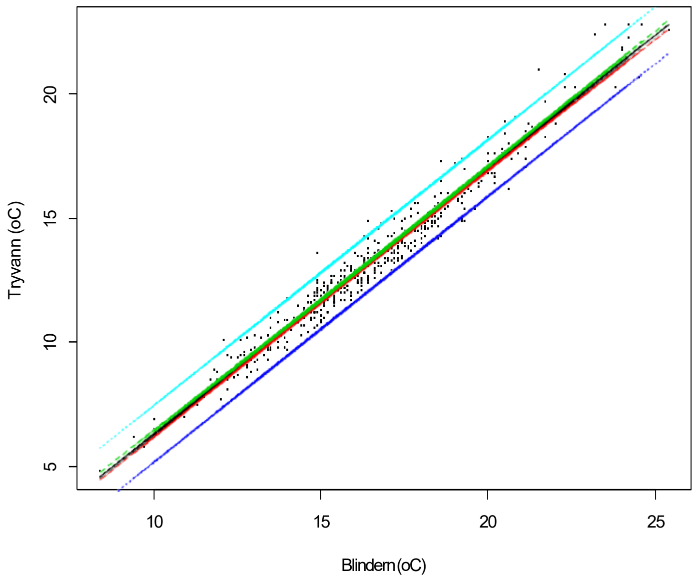
cor=0.33





A bit wider prediction intervals





Influential data points

- Single data points that affects estimates more than others
- http://omaymas.github.io/InfluenceAnalysis/
- Cooks distance:

$$D_i = \frac{\sum_{j=1}^{n} (\widehat{Y}_j - \widehat{Y}_{j(i)})^2}{(p+1)\,\widehat{\sigma}^2}$$