

13. Linear Regression Models

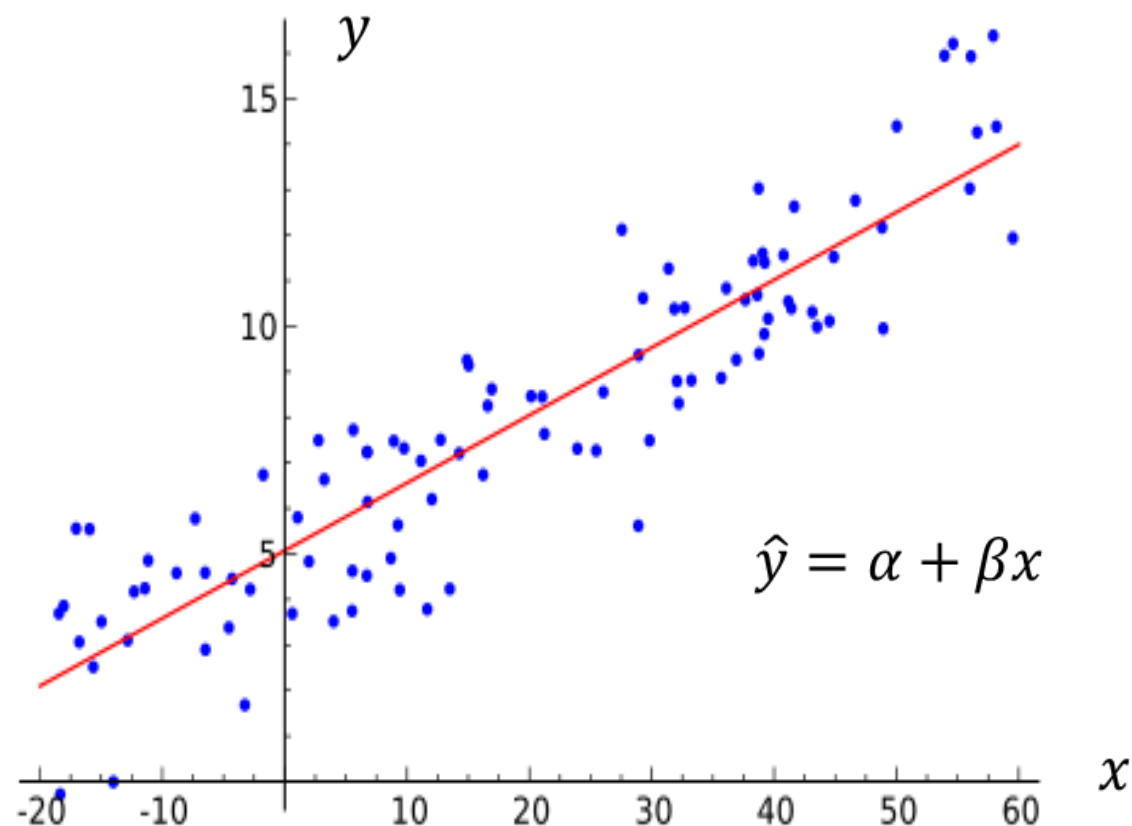
Gunvor Elisabeth Kirkelund

Lars Mandrup

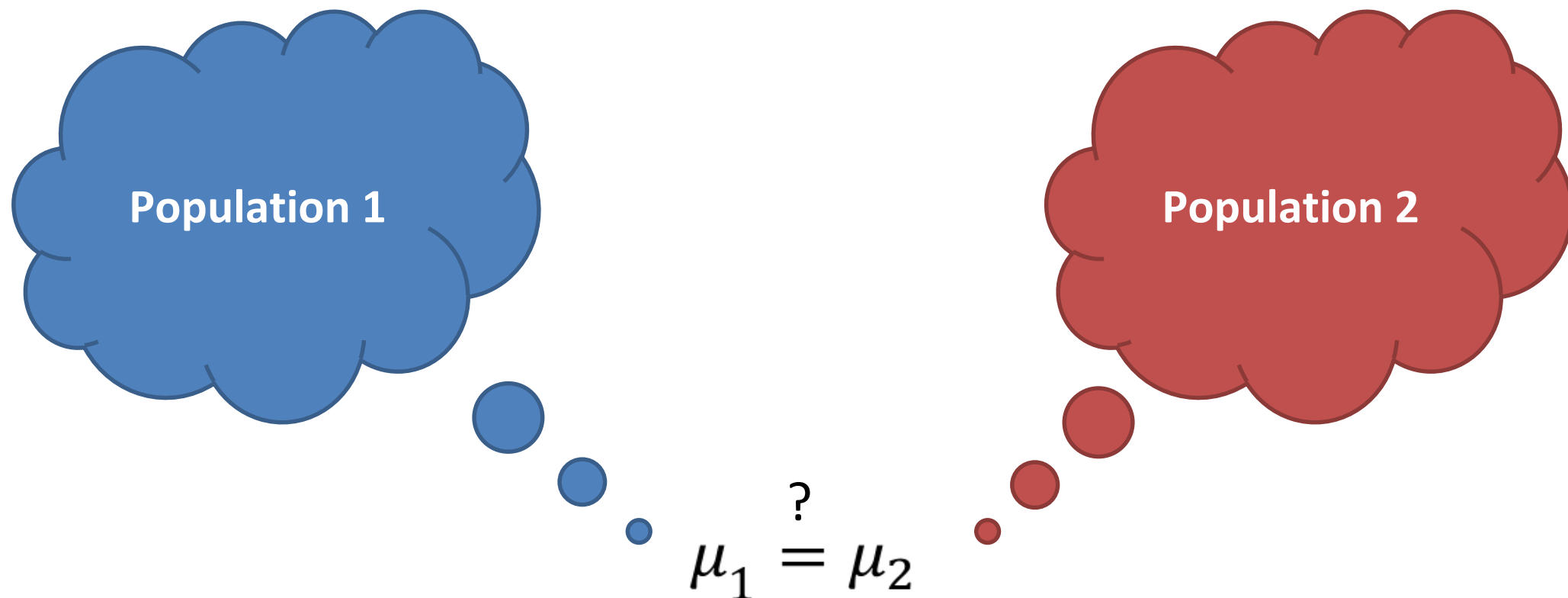
Slides and material provided in parts by
Henrik Pedersen

Today's Content

- ❖ Repetition from last time
- ❖ Linear models
- ❖ Linear regression



Comparing two population means



- Fx. The height of people from Funen (μ_1) and Jutland (μ_2)

Test Catalog for Comparing Two Means (known variance)

Statistical model:

- $X_{1i} \sim N(\mu_1, \sigma_1^2), i = 1, 2, \dots, n_1$ and $X_{2i} \sim N(\mu_2, \sigma_2^2) i = 1, 2, \dots, n_2$
- Parameter estimate: $\hat{\delta} = \bar{x}_1 - \bar{x}_2 \sim N(\mu_1 - \mu_2, \sigma_1^2/n_1 + \sigma_2^2/n_2)$
- Where the observation is $\bar{x}_1 - \bar{x}_2 =$ 'the difference between two sample means'.

Hypothesis test (two-tailed):

- $H_0: \mu_1 = \mu_2$
- $H_1: \mu_1 \neq \mu_2$
- Test size: $z = \frac{(\bar{x}_1 - \bar{x}_2)}{\sigma \sqrt{1/n_1 + 1/n_2}} \sim N(0, 1)$
- Approximate p-value: $2 \cdot |1 - \Phi(|z|)|$

Assuming equal variances: $\sigma_1 = \sigma_2 = \sigma$

95% confidence interval:

- $\delta_- = (\bar{x}_1 - \bar{x}_2) - 1.96 \cdot \sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$
- $\delta_+ = (\bar{x}_1 - \bar{x}_2) + 1.96 \cdot \sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$

Test Catalog for Comparing Two Means (unknown variance)

Statistical model:

- $X_{1i} \sim \mathcal{N}(\mu_1, \sigma_1^2)$, $i = 1, 2, \dots, n_1$ and $X_{2i} \sim \mathcal{N}(\mu_2, \sigma_2^2)$, $i = 1, 2, \dots, n_2$

- Parameter estimate: $\hat{\delta} = \bar{x}_1 - \bar{x}_2 \sim \mathcal{N}\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$

*the observed difference
between the two sample means*

$$s^2 = \frac{1}{n_1 + n_2 - 2} \left((n_1 - 1) \cdot s_1^2 + (n_2 - 1) \cdot s_2^2 \right)$$

*pooled variance of
the two samples*

Hypothesis test (two-tailed):

- NULL hypothesis H_0 : $\delta = \mu_1 - \mu_2 = \delta_0$
- Alternative hypothesis H_1 : $\delta = \mu_1 - \mu_2 \neq \delta_0$
- Test size: $t = \frac{\hat{\delta} - \delta_0}{s \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t(n_1 + n_2 - 2)$
- p-value: $pval = 2 \cdot (1 - t_{cdf}(|t|, n_1 + n_2 - 2))$

OBS:

t-test (compared with Z-test)

- *Less knowledge*
- *Larger uncertainty*
- *Confidence interval larger*
- *More difficult to reject H_0*

Confidence interval $[\delta_-, \delta_+]$:

- $t_{\alpha/2, n-1} = t_{cdf}^{-1}(1 - \frac{\alpha}{2}, n_1 + n_2 - 2)$ (often just called t_0)

- Lower bound: $\delta_- = \hat{\delta} - t_{\alpha/2, n-1} \cdot s \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$

- Upper bound: $\delta_+ = \hat{\delta} + t_{\alpha/2, n-1} \cdot s \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$

Test Catalog for Paired Data

Statistical model:

- $d_i = X_{1i} - X_{2i} \sim \mathcal{N}(\delta, \sigma^2)$, $i = 1, 2, \dots, n$
- Parameter estimate:
$$\hat{\delta} = \bar{d} = \frac{1}{n} \sum_{i=1}^n d_i = \frac{1}{n} \sum_{i=1}^n (x_{1i} - x_{2i})$$
$$s_d^2 = \frac{1}{n-1} \sum_{i=1}^n (d_i - \bar{d})^2$$

the average of the observed difference between paired samples

sample variance of the observed difference

Hypothesis test (two-tailed):

- NULL hypothesis $H_0: \delta = \delta_0$
- Alternative hypothesis $H_1: \delta \neq \delta_0$
- Test size: $t = \frac{\hat{\delta} - \delta_0}{s_d / \sqrt{n}} \sim t(n-1)$
- p-value: $pval = 2 \cdot (1 - t_{cdf}(|t|, n-1))$

Confidence interval $[\delta_-, \delta_+]$:

- $t_{\alpha/2, n-1} = t_{cdf}^{-1}(1 - \frac{\alpha}{2}, n-1)$ (often just called t_0)
- Lower bound: $\delta_- = \bar{d} - t_{\alpha/2, n-1} \cdot \frac{s_d}{\sqrt{n}}$
- Upper bound: $\delta_+ = \bar{d} + t_{\alpha/2, n-1} \cdot \frac{s_d}{\sqrt{n}}$

OBS:

Paired vs Unpaired t-test

- *Elimination of disrupting factors*
- *Reducing uncertainty*
- *Confidence interval smaller*
- *Easier to reject H_0*

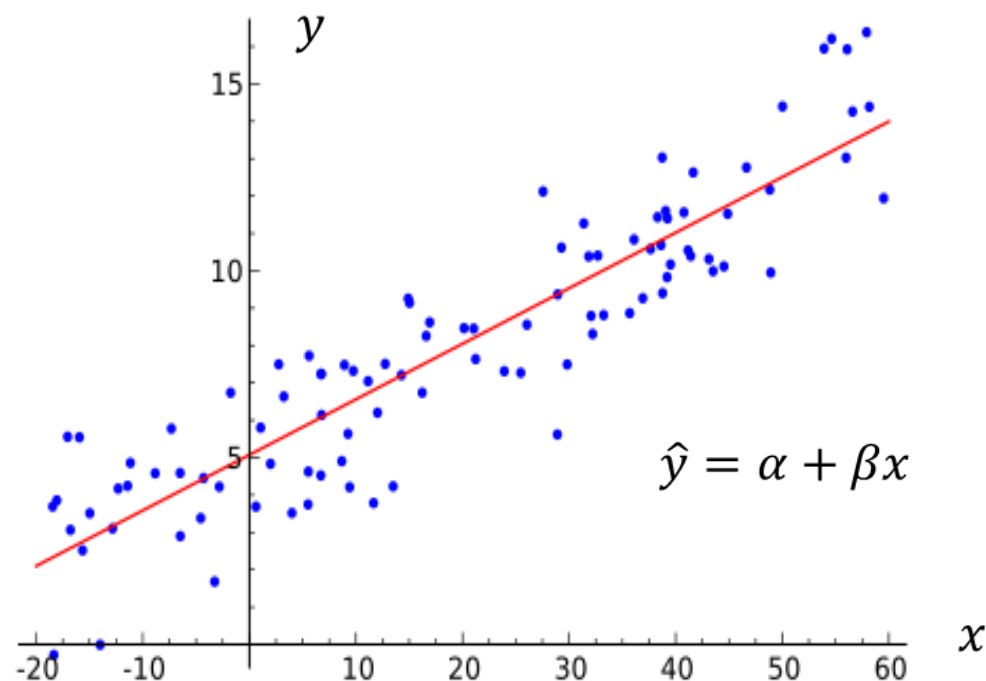
Linear Models

- Fits the set of n pairs of datapoints (x_i, y_i) to a straight line

$$y_i = \alpha + \beta x_i + \epsilon_i = \hat{y}_i + \epsilon_i$$

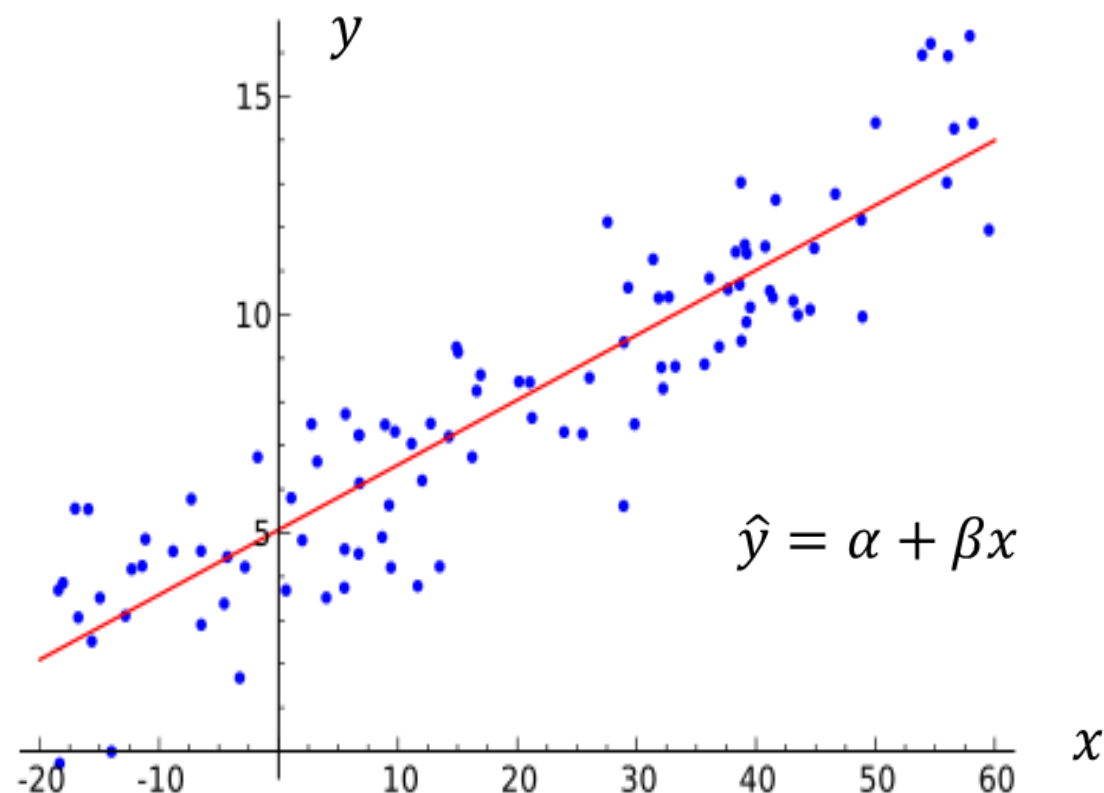
- where

- $\hat{y}_i = \alpha + \beta x_i$ are the estimated linear coherence (straight line)
- ϵ_i (called the residual) are a random variable modelling the errors in our linear assumption



Linear Models – When/Why?

- Can be used when the mean changes with some parameter (x).
- The variance should not change over time
- The mean is connected to the independent parameter (x) linearly!
- The simplest model – more advanced models not necessarily the best → start simple!



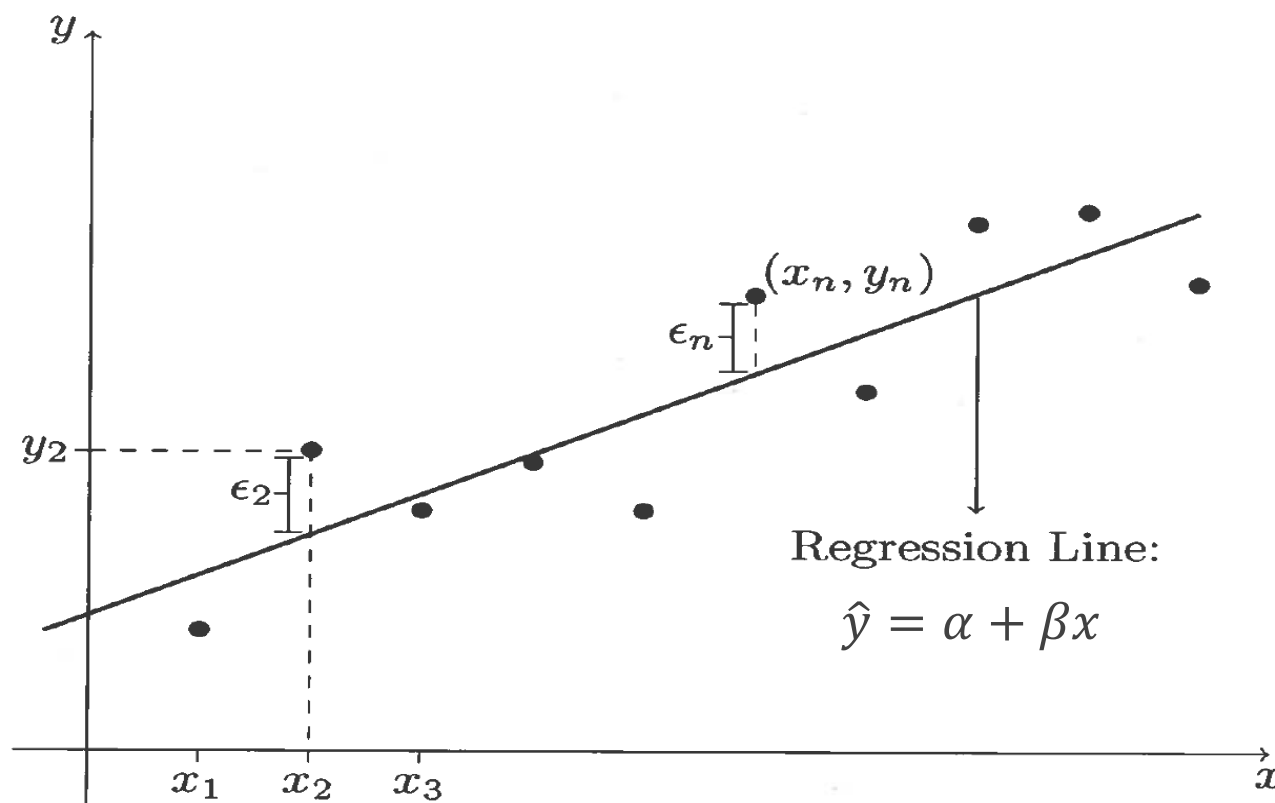
Residual

- A residual is the difference between the measured and predicted data:

$$\epsilon_i = y_i - \hat{y}_i = y_i - (\alpha + \beta x_i)$$

measured \nearrow \nwarrow *estimated*

- The residual ϵ_i is an independent random variable modelling the error in our model



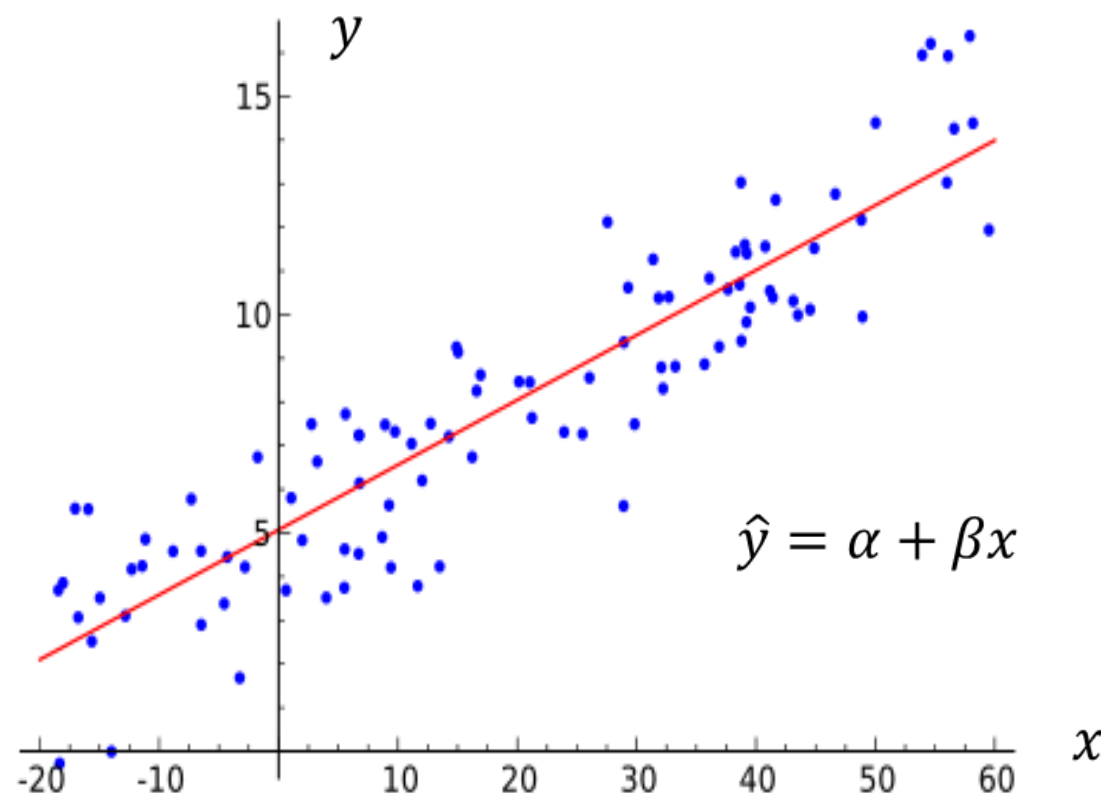
ϵ_i = the vertical distances between the points of data and the fitted line

Linear Regression

- Fits a straight line through the set of n points (x_i, y_i) :

$$y_i = \alpha + \beta x_i + \epsilon_i = \hat{y}_i + \epsilon_i$$

- ie. determine values of the slope β and interception α that makes the sum of squared residuals (errors) ($\epsilon^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 \sim \chi^2$) of the model as small as possible



Statistical Model

- Data comes in pairs:

fx. time t (x_i, y_i) for $i = 1, 2, \dots, n$

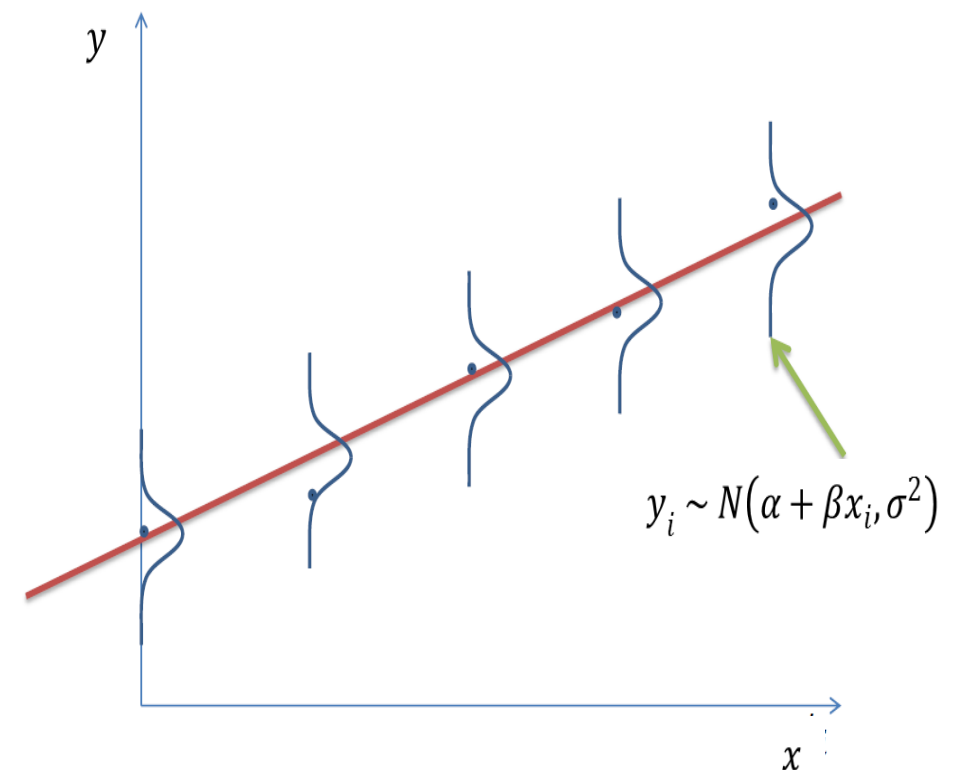
- where x is the independent (or predictor) variable and y is the dependent (or response) variable

- Statistical model:** $y_i = \alpha + \beta x_i + \epsilon_i$

- with

$$\epsilon_i \sim \mathcal{N}(0, \sigma^2) \quad \text{and} \quad y_i \sim \mathcal{N}(\alpha + \beta x_i, \sigma^2)$$


- where β is the slope of the straight line and α its intercept with the y -axis.



*OBS! In "Introduction to Probability, Statistics and Random Processes":
 $\alpha \rightarrow \beta_0$ and $\beta \rightarrow \beta_1$*

Empirical Variance

- Recall that $y_i \sim \mathcal{N}(\alpha + \beta x_i, \sigma^2)$
- The unbiased estimator of the variance is


$$s_r^2 = \frac{1}{n-2} \sum_{i=1}^n (y_i - (\hat{\alpha} + \hat{\beta} x_i))^2 = \frac{1}{n-2} \sum_{i=1}^n \epsilon_i^2$$


Two constraints $(\hat{\alpha}, \hat{\beta}) \rightarrow \div$ two degrees of freedom

- Statistical inference in linear regression concerns the parameter estimates: $\hat{\alpha}, \hat{\beta}$ and s_r^2 .

Model Fitting

- The goal of linear regression is to determine the values of the slope β and interception α that minimize the sum of residuals of the model:

$$R(\alpha, \beta) = \epsilon^2 = \sum_{i=1}^n (y_i - (\alpha + \beta x_i))^2$$


- The parameter estimate that minimize $R(\alpha, \beta)$ are:

$$\hat{\beta} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sum_{i=1}^n (x_i y_i - \bar{x} \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{s_{xy}}{s_{xx}}$$

$$\hat{\alpha} = \bar{y} - \hat{\beta} \bar{x}$$

- where $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$; $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$ (the average of x_i and y_i)
and $s_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2$; $s_{xy} = \sum_{i=1}^n (x_i y_i - \bar{x} \bar{y}) = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$

Derivation of the Intercept Parameter

- Partial derivative w.r.t. α and setting to zero:

$$\frac{\partial R(\alpha, \beta)}{\partial \alpha} = \frac{\partial}{\partial \alpha} \sum_{i=1}^n (y_i - \alpha - \beta x_i)^2 = -2 \sum_{i=1}^n (y_i - \alpha - \beta x_i) = 0$$

- It follows that:

$$n\alpha = \sum_{i=1}^n y_i - \beta \sum_{i=1}^n x_i = n\bar{y} - \beta n\bar{x}$$

\Downarrow

$$\hat{\alpha} = \bar{y} - \hat{\beta} \bar{x}$$

- where $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$; $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$ (the average of x_i and y_i)

Derivation of the Slope Parameter

- Partial derivative w.r.t. β and setting to zero:

$$\begin{aligned}\frac{\partial R(\alpha, \beta)}{\partial \beta} &= \frac{\partial}{\partial \beta} \sum_{i=1}^n (y_i - \alpha - \beta x_i)^2 = -2 \sum_{i=1}^n x_i (y_i - \alpha - \beta x_i) \\ &= -2 \left(\sum_{i=1}^n x_i y_i - \alpha \sum_{i=1}^n x_i - \beta \sum_{i=1}^n x_i^2 \right) = 0\end{aligned}$$

- Inserting the result $\alpha = \bar{y} - \beta \bar{x}$ and using that $\sum_{i=1}^n x_i = \sum_{i=1}^n \bar{x}$ we get:

$$\begin{aligned}\sum_{i=1}^n x_i y_i - (\bar{y} - \beta \bar{x}) \sum_{i=1}^n x_i - \beta \sum_{i=1}^n x_i^2 &= \sum_{i=1}^n (x_i y_i - \bar{x} \bar{y}) - \beta \sum_{i=1}^n (x_i^2 - \bar{x}^2) \\ &= \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) - \beta \sum_{i=1}^n (x_i - \bar{x})^2 = 0\end{aligned}$$

- It follows that:

$$\hat{\beta} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sum_{i=1}^n (x_i y_i - \bar{x} \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{s_{xy}}{s_{xx}}$$

- where: $s_{xy} = \sum_{i=1}^n (x_i y_i - \bar{x} \bar{y}) = n \cdot \text{Cov}(x, y)$ and $s_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2 = n \cdot \text{Var}(x)$

Example - Hubble's Law

- Hubble's law is the name for the observation in physical cosmology that objects observed in deep space are found to have a relative velocity away from the Earth that is approximately proportional to their distance from the Earth: $v = H \cdot x$
- Edwin Hubble's original measurements for 24 distant galaxies were (in Matlab notation)

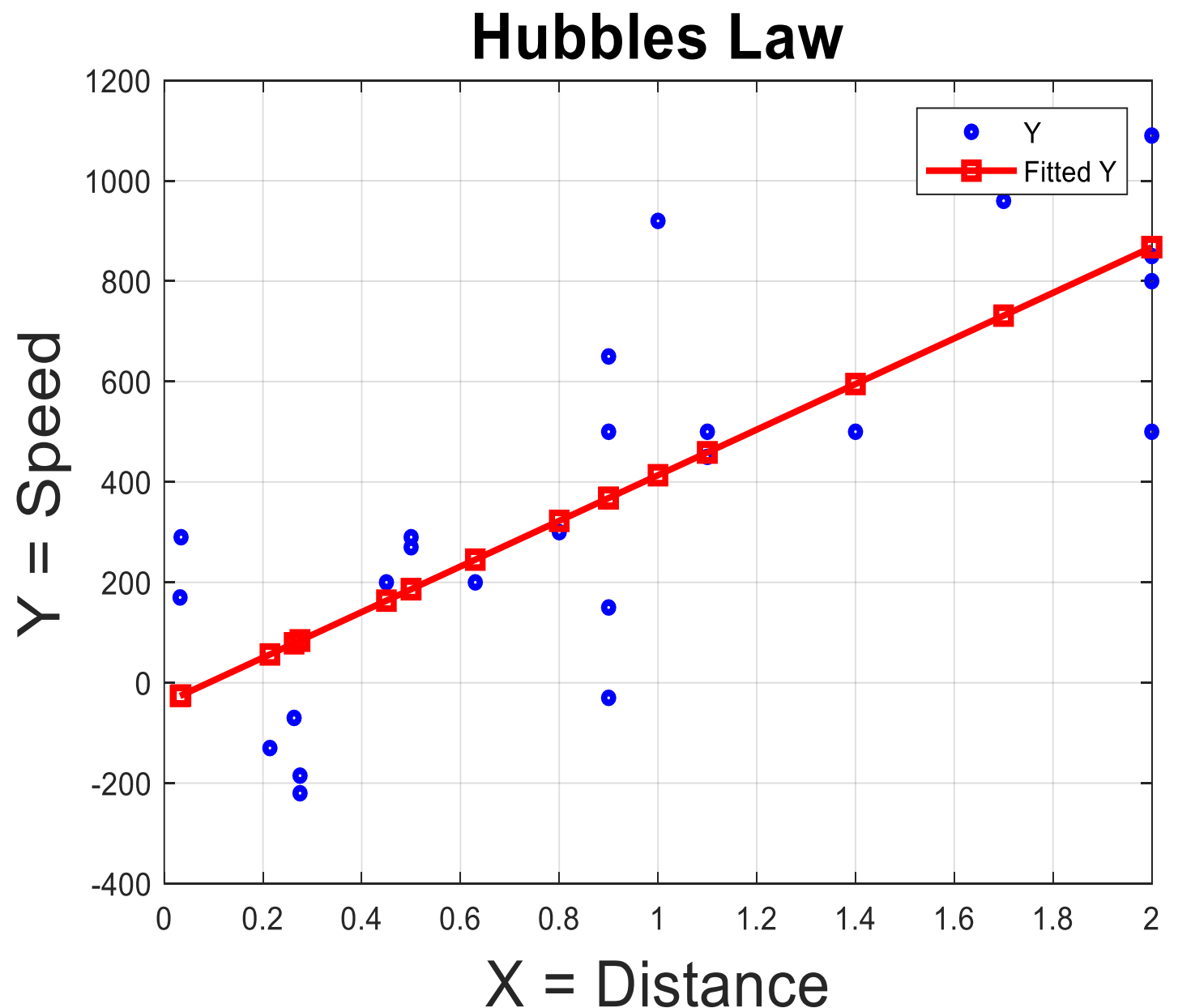
```
Distance = [ 0.032  0.034  0.214  0.263  0.275  0.275  ...  
            0.450  0.500  0.500  0.630  0.800  0.900  ...  
            0.900  0.900  0.900  1.000  1.100  1.100  ...  
            1.400  1.700  2.000  2.000  2.000  2.000 ];
```

```
Speed = [ 170  290 -130 -70 -185 -220 200 290 ...  
         270  200  300 -30  650  150 500 920 ...  
         450  500  500 960  500  850 800 1090 ];
```

Example - Hubble's Law

- Choosing:
x = Distance
y = Speed
- From data:
 $\bar{x} = 0.91$, $\bar{y} = 373.1$
- Slope estimate:
$$\hat{\beta} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = 454.2$$
- Intercept estimate:
$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} = -40.8$$

➤ *Estimate of Hubble's law:*
$$v = 454,2 \cdot r - 40,8$$



Statistical Inference on the Regression Slope

- In general, the null hypothesis about the slope that we wish to test takes the following form:

$$H_0: \beta = \beta_0$$

- It can be shown that the estimator of the slope is normally distributed

$$\hat{\beta} \sim \mathcal{N}\left(\beta, \frac{s_r^2}{s_{xx}}\right)$$

- where $s_r^2 = \frac{1}{n-2} \sum_{i=1}^n (y_i - \hat{y})^2$ is the empirical variance and $s_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2$

- The appropriate test statistics for $\hat{\beta}$ is therefore:


$$t = \frac{\hat{\beta} - \beta_0}{\sqrt{s_r^2 / s_{xx}}} \sim t(n - 2)$$

- The p-value is:

$$p\text{-val} = 2 \cdot (1 - t_{cdf}(|t|, n - 2))$$

Example - Hubble's Law

Let us test whether the regression slope deviates significantly from zero:

- Null hypothesis: $H_0: \beta = 0$  *No relation between x (distance) and y (velocity)*
- Parameter estimate: $\hat{\beta} = 454.2$ and $\hat{\alpha} = -40.8$
- Empirical variance: $s_r^2 = \frac{1}{n-2} \sum_{i=1}^n (y_i - \hat{y})^2 = \frac{1}{n-2} \sum_{i=1}^n (y_i - (\hat{\alpha} + \hat{\beta}x_i))^2 = 54247$
and $s_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2 = 9.583$
- Test size: $t = \frac{\hat{\beta} - \beta_0}{\sqrt{s_r^2 / s_{xx}}} = \frac{454.2 - 0}{\sqrt{54247 / 9.583}} = 6.035 \sim t(24 - 2)$
- p-value: $p\text{-val} = 2 \cdot \left(1 - t_{cdf}(|t|, n - 2)\right) = 2 \cdot \left(1 - t_{cdf}(6.035, 22)\right) \approx 0$
- Since $p\text{-val} < 0.05$ we **reject the null hypothesis that $\beta = 0$** .
- In other words, the data suggest that the regression slope deviates significantly from zero and therefore **there is some dependency between the distance and the velocity of objects in the universe.**

Statistical Inference on the Regression Slope

- The 95% confidence interval for the slope is

$$\beta_- = \hat{\beta} - t_0 \cdot \sqrt{\frac{s_r^2}{s_{xx}}}$$

$$\beta_+ = \hat{\beta} + t_0 \cdot \sqrt{\frac{s_r^2}{s_{xx}}}$$

- where

$$t_0 = \text{tinv}\left(1 - \frac{0.05}{2}, n - 2\right) = \text{tinv}(0.975, n - 2)$$

and $s_r^2 = \frac{1}{n-2} \sum_{i=1}^n (y_i - \hat{y})^2$; $s_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2$

Example - Hubble's law

- 95% confidence interval:

$$t_0 = \text{tinv}(0.975, 22) = 2.0739$$

$$\beta_- = \hat{\beta} - t_0 \cdot \sqrt{\frac{s_r^2}{s_{xx}}} = 298.1$$

$$\beta_+ = \hat{\beta} + t_0 \cdot \sqrt{\frac{s_r^2}{s_{xx}}} = 610.2$$

- The null hypothesis $H_0: \beta = 0$ is not within the 95% confidence interval, so we reject the null hypothesis

Statistical Inference on the Regression Intercept

- In general, the null hypothesis about the intercept that we wish to test takes the following form:

$$H_0: \alpha = \alpha_0$$

- It can be shown that the estimator of the intercept is normally distributed

$$\hat{\alpha} \sim \mathcal{N}\left(\alpha, s_r^2 \cdot \left(\frac{1}{n} + \frac{\bar{x}^2}{s_{xx}}\right)\right)$$

- where $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$; $s_r^2 = \frac{1}{n-2} \sum_{i=1}^n (y_i - \hat{y})^2$; $s_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2$

- The appropriate test statistics for $\hat{\alpha}$ is therefore:


$$t = \frac{\hat{\alpha} - \alpha_0}{\sqrt{s_r^2 \cdot \left(\frac{1}{n} + \frac{\bar{x}^2}{s_{xx}}\right)}} \sim t(n-2)$$

- The p-value is:

$$p\text{-val} = 2 \cdot (1 - t_{cdf}(|t|, n-2))$$

Example - Hubble's Law

Let us test whether the regression intercept deviates significantly from zero:

- Null hypothesis: $H_0: \alpha = 0$  *The line go through origo (0,0)*
- Parameter estimate: $\hat{\beta} = 454.2$ and $\hat{\alpha} = -40.8$
- Statistics:
$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = 0.9114$$
$$s_r^2 = \frac{1}{n-2} \sum_{i=1}^n (y_i - \hat{y})^2 = \frac{1}{n-2} \sum_{i=1}^n (y_i - (\hat{\alpha} + \hat{\beta}x_i))^2 = 54247$$
$$s_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2 = 9.583$$
- Test size:
$$t = \frac{\hat{\alpha} - \alpha_0}{\sqrt{s_r^2 \cdot \left(\frac{1}{n} + \frac{\bar{x}^2}{s_{xx}}\right)}} = \frac{-40.8 - 0}{\sqrt{54247 \left(\frac{1}{24} + \frac{0.9114^2}{9.583}\right)}} = -0.4887 \sim t(24 - 2)$$
- p-value:
$$p\text{-val} = 2 \cdot \left(1 - t_{cdf}(|t|, n - 2)\right) = 2 \cdot \left(1 - t_{cdf}(0.4887, 22)\right) \approx 0.63$$
- Since $p\text{-val} > 0.05$ we **fail to reject the null hypothesis that $\alpha = 0$** .
- In other words, the data suggest that **the regression intercept does not deviates significantly from zero**.

Statistical Inference on the Regression Intercept

- The 95% confidence interval for the intercept α is:

$$\alpha_- = \hat{\alpha} - t_0 \cdot \sqrt{s_r^2 \cdot \left(\frac{1}{n} + \frac{\bar{x}^2}{s_{xx}} \right)}$$

$$\alpha_+ = \hat{\alpha} + t_0 \cdot \sqrt{s_r^2 \cdot \left(\frac{1}{n} + \frac{\bar{x}^2}{s_{xx}} \right)}$$

- where

$$t_0 = \text{tinv} \left(1 - \frac{0.05}{2}, n - 2 \right) = \text{tinv}(0.975, n - 2)$$

and $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$; $s_r^2 = \frac{1}{n-2} \sum_{i=1}^n (y_i - \hat{y})^2$; $s_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2$

Example - Hubble's law

- 95% confidence interval:

$$t_0 = \text{tinv}(0.975, 22) = 2.0739$$

$$\alpha_- = \hat{\alpha} - t_0 \cdot \sqrt{s_r^2 \cdot \left(\frac{1}{n} + \frac{\bar{x}^2}{s_{xx}} \right)} = -124.2$$

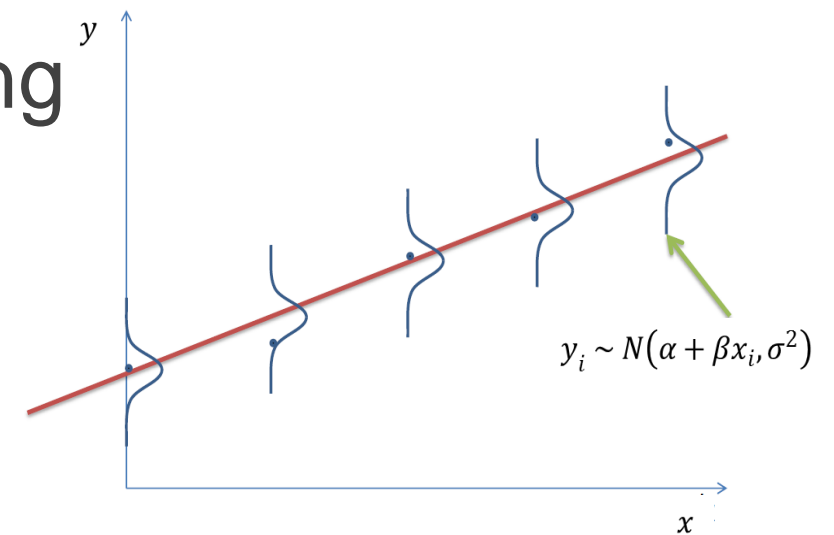
$$\alpha_+ = \hat{\alpha} + t_0 \cdot \sqrt{s_r^2 \cdot \left(\frac{1}{n} + \frac{\bar{x}^2}{s_{xx}} \right)} = 42.6$$

- The null hypothesis $H_0: \alpha = 0$ is within the 95% confidence interval, so we fail to reject the null hypothesis

Checking for Normality

- Recalling that the statistical model underlying linear regression is

$$y_i \sim \mathcal{N}(\alpha + \beta x_i, \sigma^2)$$



- the residual of the i'th sample should be normally distributed with zero mean and variance σ^2

$$\epsilon_i = y_i - (\alpha + \beta x_i) \sim \mathcal{N}(0, \sigma^2)$$

- Hence, a good way to check whether the assumption of linearity between x and y holds is to first fit the linear model and subsequently check that the residuals ϵ_i are normally distributed using a Q-Q plot.

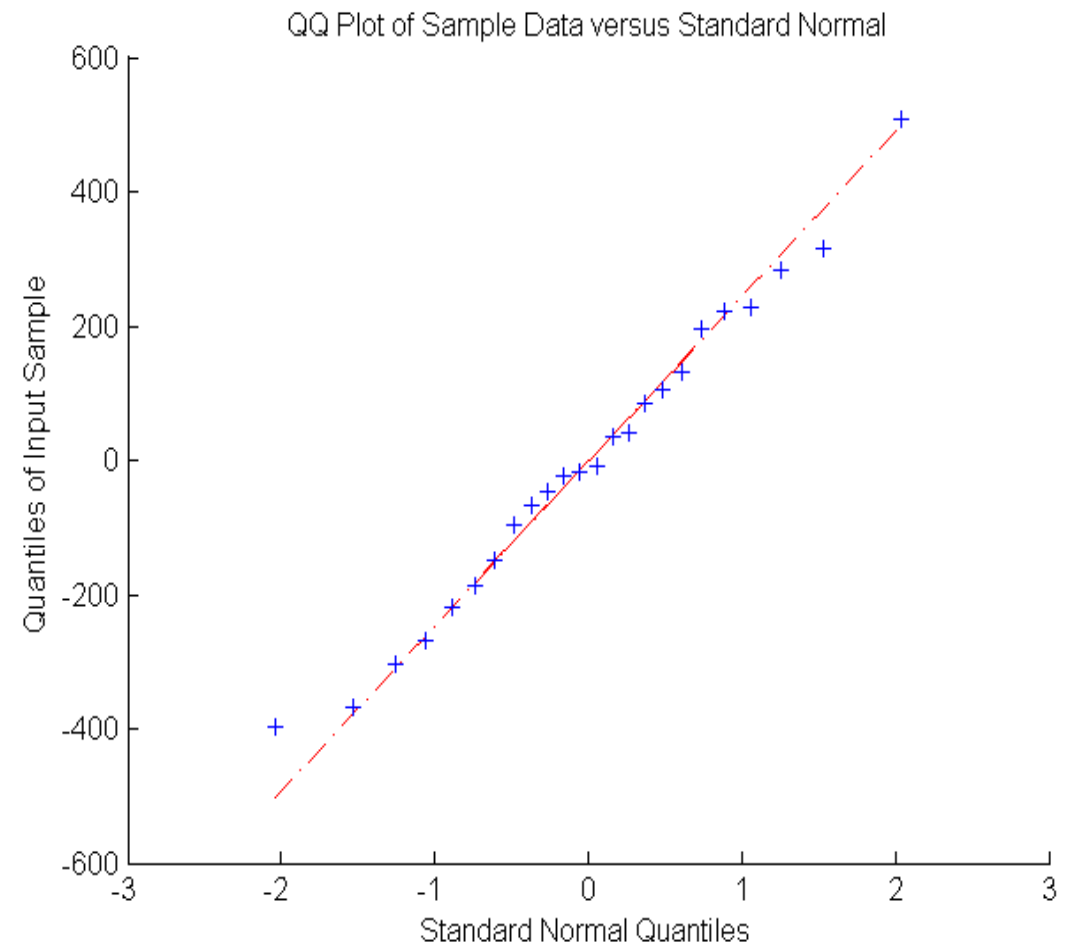
Checking for Normality Using Q-Q plot (Hubble's law)

- The residuals in Hubble's law example are

$$\text{res} = y - \alpha - \beta x$$

- The resulting Q-Q plot

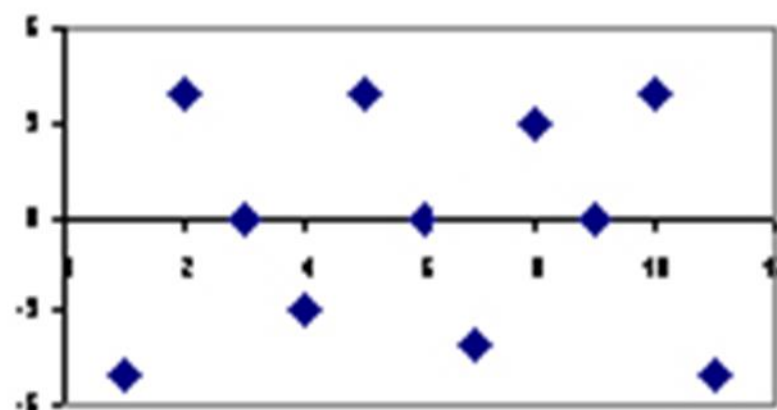
```
qqplot(res)
```



- shows that the residuals are approximately normally distributed, because the data points lie approximately on a straight line.
- Hence, it is safe to use simple linear regression to find the relation between the Speed and Distance of galaxies.

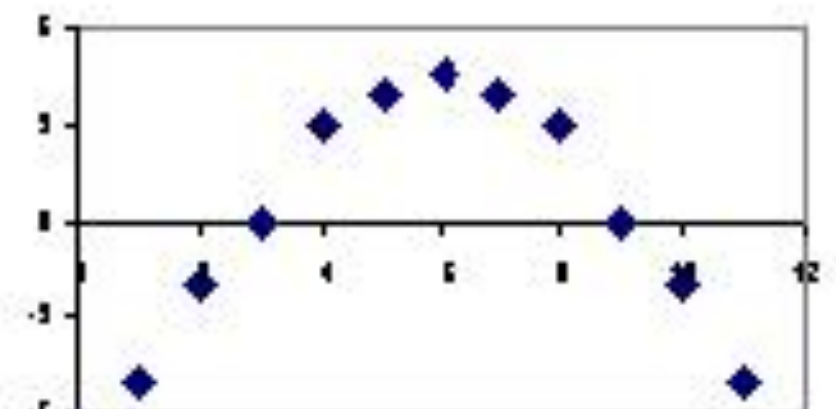
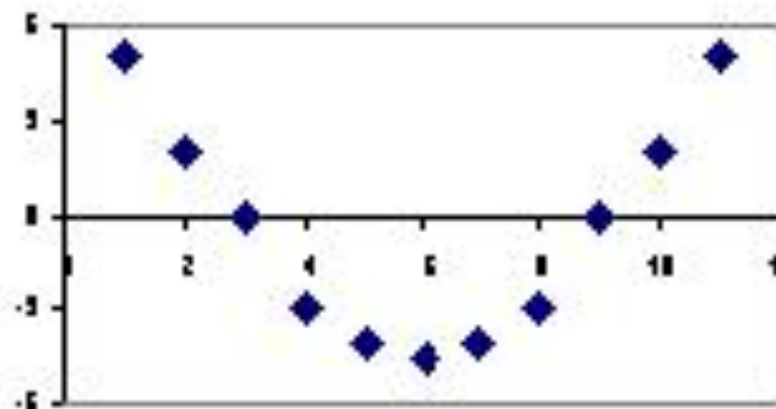
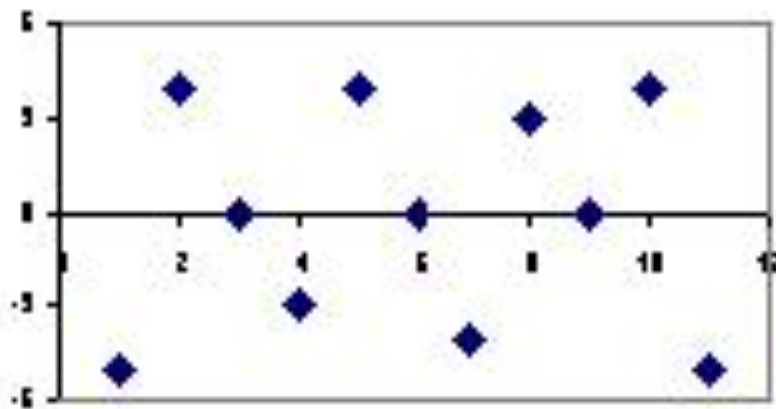
Residual Plots

- Another way to check the normality assumption is to make a so-called *residual plot*.
- A residual plot is a graph that shows the residuals on the vertical axis and the independent variable (x) on the horizontal axis.
- If the points in a residual plot are randomly dispersed around the horizontal axis, a linear regression model is appropriate for the data; otherwise, a non-linear model is more appropriate.



Residual Plots

- Below, the residual plots show three typical patterns.
- The first plot shows a random pattern, indicating a good fit for a linear model.
- The other plot patterns are non-random (U-shaped and inverted U), suggesting a better fit for a non-linear model.
- Formally you must check the following two conditions:
 - The value of the residuals $\epsilon_i = y_i - (\alpha + \beta x_i)$ must not depend on x_i , but should lie randomly distributed around zero.
 - The variance of the residuals must not depend on x_i either.



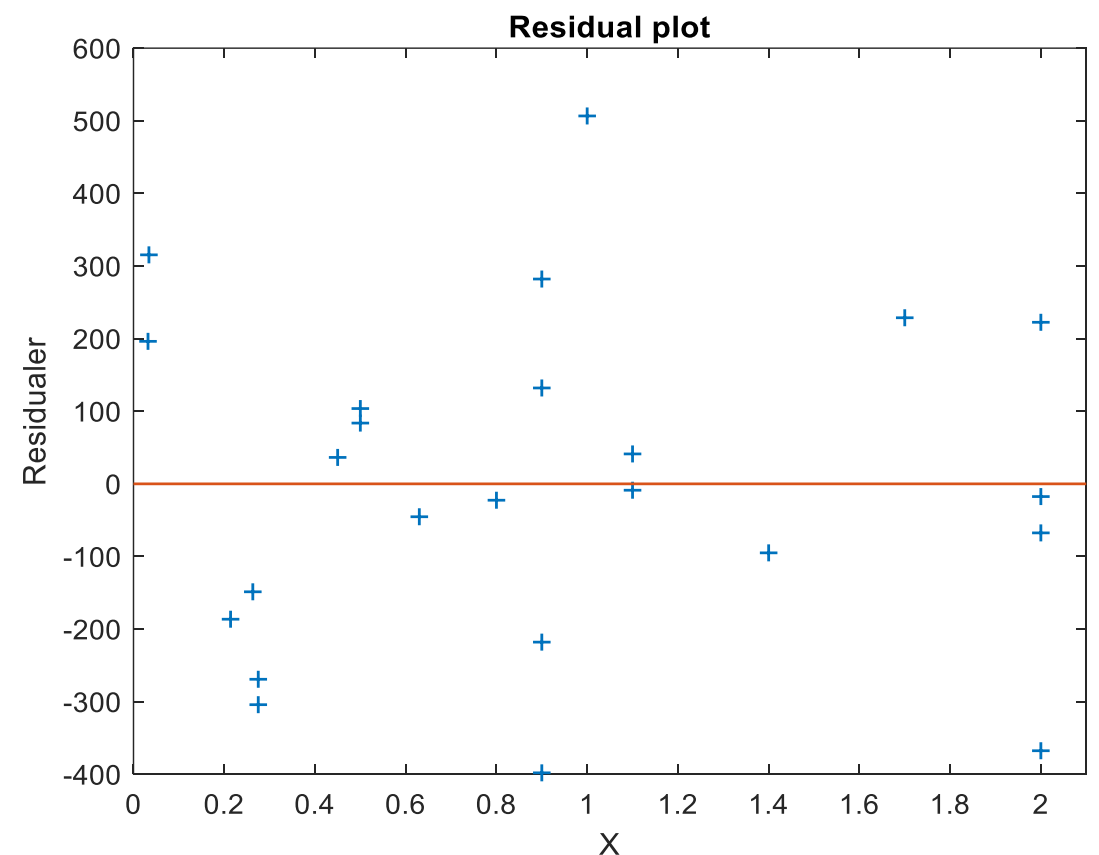
Checking for Normality Using Residual Plot (Hubble's law)

- The residuals in Hubble's law example are

$$\text{res} = y - \alpha - \beta x$$

- The resulting residual plot

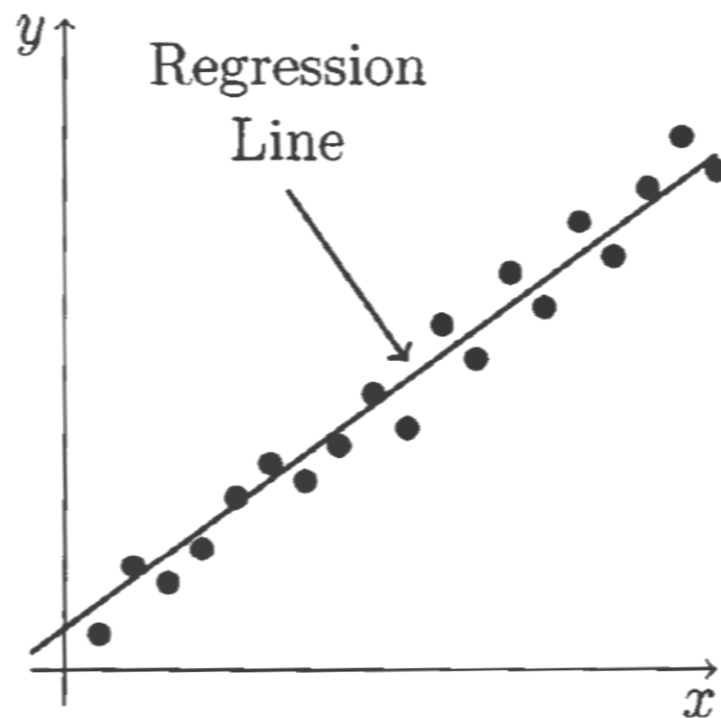
```
plot(x, res, '.', ...  
      [0 2.1], [0 0])  
axis([0 2.1 -400 600])  
xlabel('X')  
ylabel('Residuals')
```



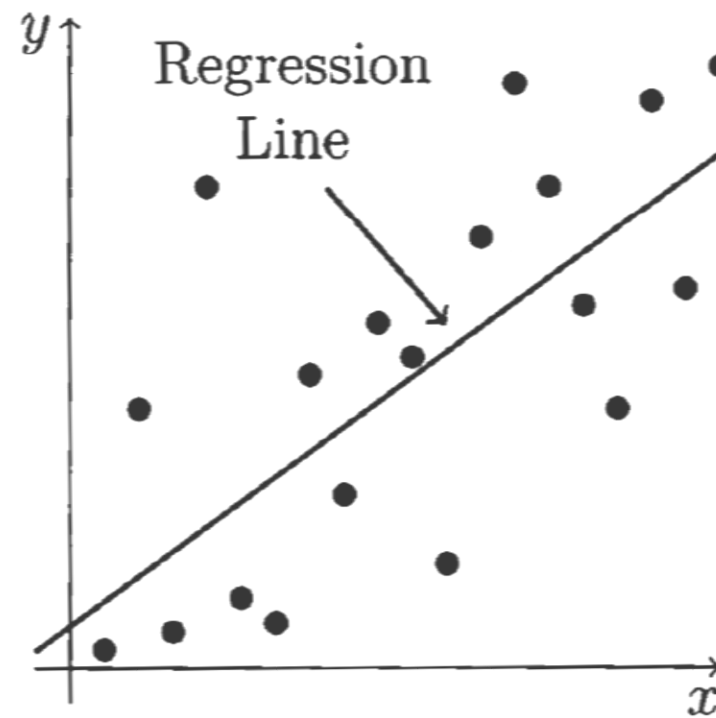
- Shows that the residuals are randomly distributed around zero and do not depend on x .
- Also, it appears that the variance of the residuals is independent of x .

How good is the linear model?

- We can always find the best regression straight line to a set of datapoints (x_i, y_i)
- But it doesn't mean that the regression straight line fits the datapoints very well – ie. that the linear model is a good model



(a)



(b)

Sample Correlation Coefficient

- If we wish to quantify the strength of a linear relation, we can use the sample correlation coefficient

$$r = \rho = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} = \frac{s_{xy}}{\sqrt{s_{xx} \cdot s_{yy}}} = \frac{Cov(x, y)}{\sqrt{Var(x) \cdot Var(y)}}$$

- where $s_{xy} = \sum_{i=1}^n (x_i y_i - \bar{x} \bar{y})$; $s_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2$; $s_{yy} = \sum_{i=1}^n (y_i - \bar{y})^2$
- As we know from stochastic variables, the correlations coefficient (ρ) takes on values from -1 to 1.
- It can be shown that the estimate of the regression slope is lineary related to the correlation coefficient:

$$\hat{\beta} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{s_{xy}}{s_{xx}} = r \sqrt{\frac{s_{yy}}{s_{xx}}}$$

Large β \rightarrow large r \rightarrow strong correlation

Coefficient of Determination

- In simple linear regression, the *coefficient of determination*

$$R^2 = r^2,$$

- indicates how well the data fit the linear model.
- The coefficient of determination ranges from 0 to 1 with value close to 1 suggesting a strong linear relationship, and values close to 0 suggesting no linear relationship.
- The coefficient of determination in the example with Hubble's law is $R^2 = 0.6235$.
- To calculate the sample correlation coefficient between x and y in Matlab, use the command `corr2(x, y)`.

Outliers

- Outliers are data points that are separated from the rest of the data and potentially influential for the regression analysis.
- Outliers can have a dramatic on the sample correlation coefficient (and therefore the slope).
- Recalling the definition of the sample correlation coefficient,

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

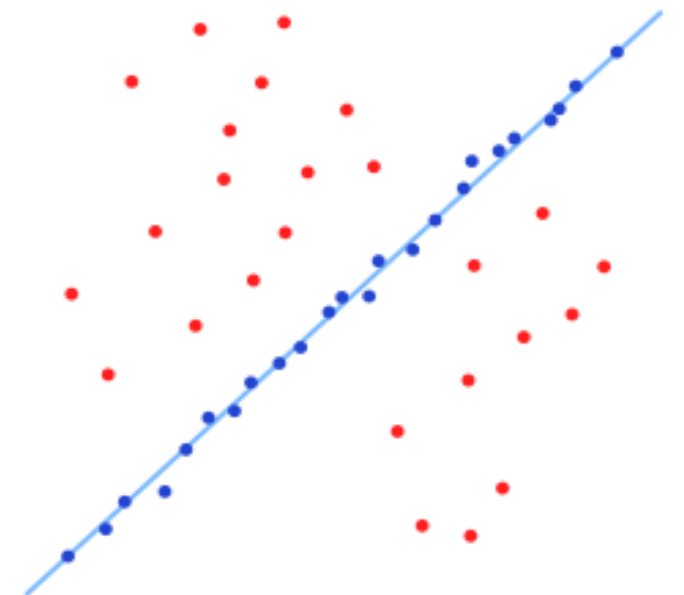


- an outlier is a point (x_i, y_i) , such that either $(x_i - \bar{x})$ or $(y_i - \bar{y})$, or both, is large.
- The extent of influence of any point can be judged in part by computing the correlation coefficient with and without that point.

Outliers

- Outliers must not be omitted:
 - just because they are outliers
 - from a conclusion

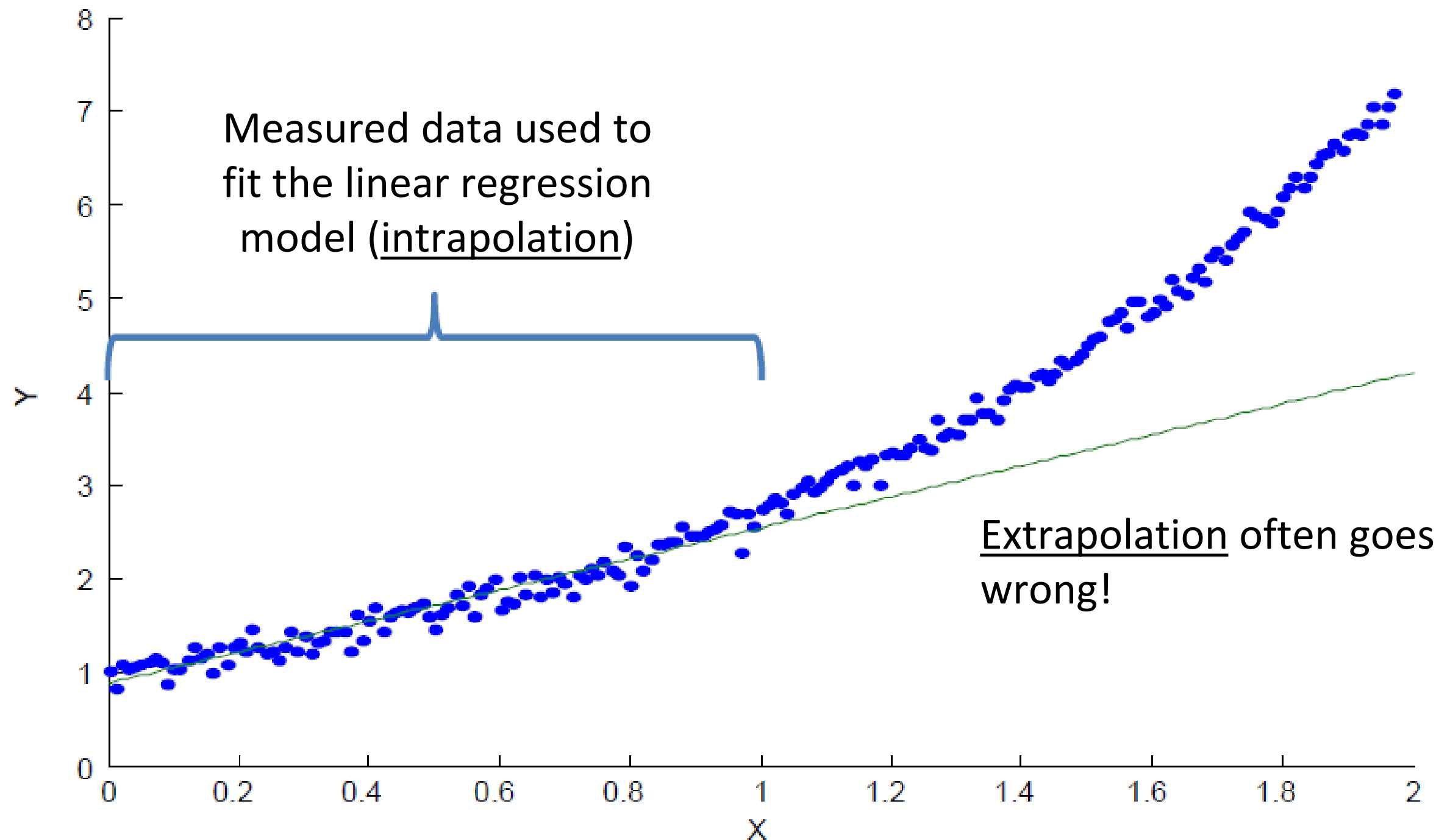
Fx. may new medication damage individual patients (allergy)
- Outliers can – with justification – be omitted from a linear regression
 - Several different methods/programs to omitting outliers – fx. RANSAC (Random Sample Consensus)



Usage of Linear Regression

- Linear regression is often used for prediction.
- Suppose, for instance, that the relationship between daily energy consumption of a power plant and the outside temperature is linear.
- Then, given the temperature of tomorrow (from a weather forecast), we can give an estimate of tomorrow's energy consumption of the power plant based on a linear model.
- When you use a **regression equation**, do not use values for the independent variable that are outside the range of values used to create the equation.
- That is called **extrapolation**, and it can produce unreasonable estimates.

Intrapolation / Extrapolation



Words and Concepts to Know

Linear Regression

Random Sample Consensus

Linear Model

Response

Intrapolation

Slope parameter

Sample Correlation Coefficient

Regression Intercept

Intercept parameter

Extrapolation

Predicted data

Outliers

Model Fitting

Slope parameter

Predictor

Residual

Empirical Variance

Residual plot

Inliers

RANSAC

Measured data

Coefficient of Determinaton

Kursus evaluering

- Åben ETSMP kursussiden på Blackboard
- Åben linket "Kursusevaluering" i venstre-menuen
- Udfyld evalueringsskemaet (gerne på dansk)
- Deadline søndag den 9. maj

- På næste onsdag samler vi op på evalueringerne