MASM22/FMSN30: Linear and Logistic Regression, 7.5 hp
FMSN40: ... with Data Gathering, 9 hp

Lecture 2, spring 2019

Confidence and prediction intervals, t-test, residuals

Mathematical Statistics / Centre for Mathematical Sciences
Lund University

27/3-19

Confidence intervals

The *true values* of (β_0, β_1) are completely unknown. Yes, we can estimate those, but how precise are these estimates?
A confidence interval (CI) is calculated from the observations, therefore in principle different from dataset to dataset. It "frequently" includes the *true unknown value* of the parameter of interest if the experiment is repeated.
But how frequently? This depends on the chosen "confidence level" $1 - \alpha$, which is a probability level. Here α is the "significance level" $(0 < \alpha < 1)$.

The choice of α is essentially arbitrary. You can choose any α. In practice it is heavily influenced by custom, e.g. $\alpha = 0.1, 0.05, 0.01$ are typically used.
Alternatives are available: e.g. Bayesian methods (lovely stuff, but not part of the course).

We will see that we can construct an interval based on the given data. However, if we repeat the experiment with new data we would obtain another interval. So...?
Statistical inference based on the Neyman-Pearson approach (1930's) proved that the intervals as given below are "optimal", in the sense that if we repeat the experiments say 100 times with new data, then $(1-\alpha)\%$ of those 100 intervals will include the true (β_0, β_1). *[will be emphasized again later when introducing hypthesis testing]*
A confidence interval with confidence level $1-\alpha$ is given by

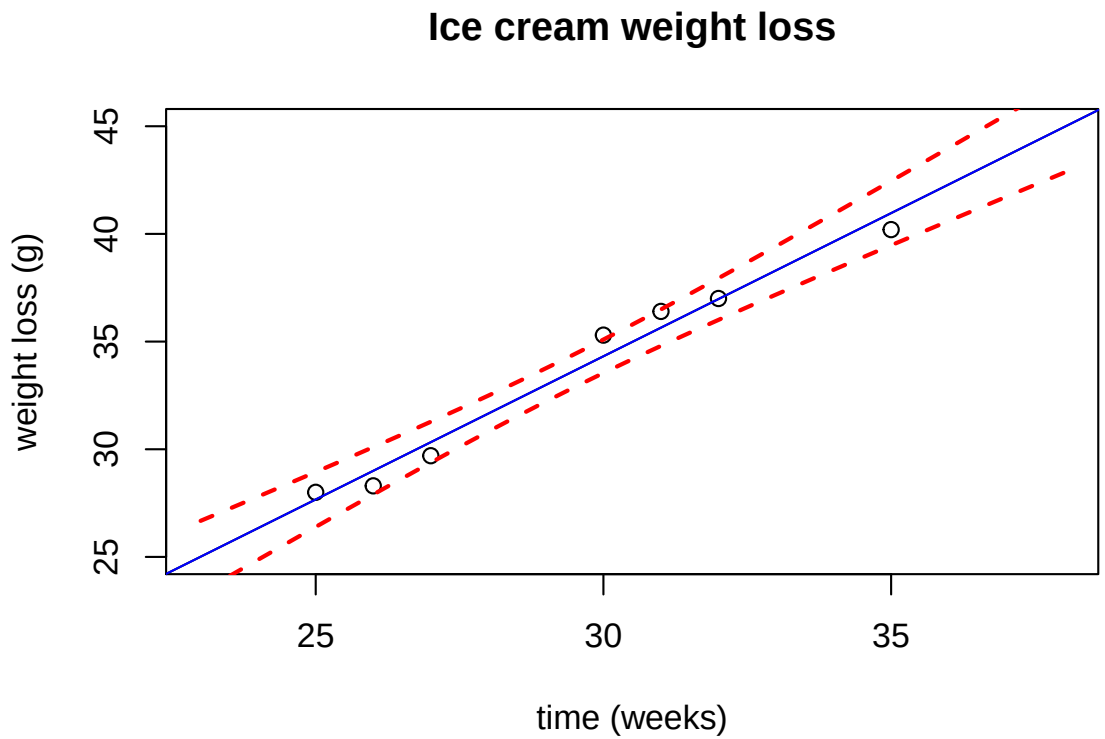
$$I_{\beta_0} = \left(\hat{\beta}_0 \pm t_{\alpha/2,n-2} \cdot s \sqrt{\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^n (X_i - \bar{X})^2}} \right)$$

$$I_{\beta_1} = \left(\hat{\beta}_1 \pm t_{\alpha/2,n-2} \cdot \frac{s}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2}} \right)$$

[in R: qt$(1-\alpha/2, n-2)$ gives $t_{\alpha/2,n-2}$]

Confidence interval for the expected value $E(Y_0) = \beta_0 + \beta_1 X_0$ for a given value X_0:

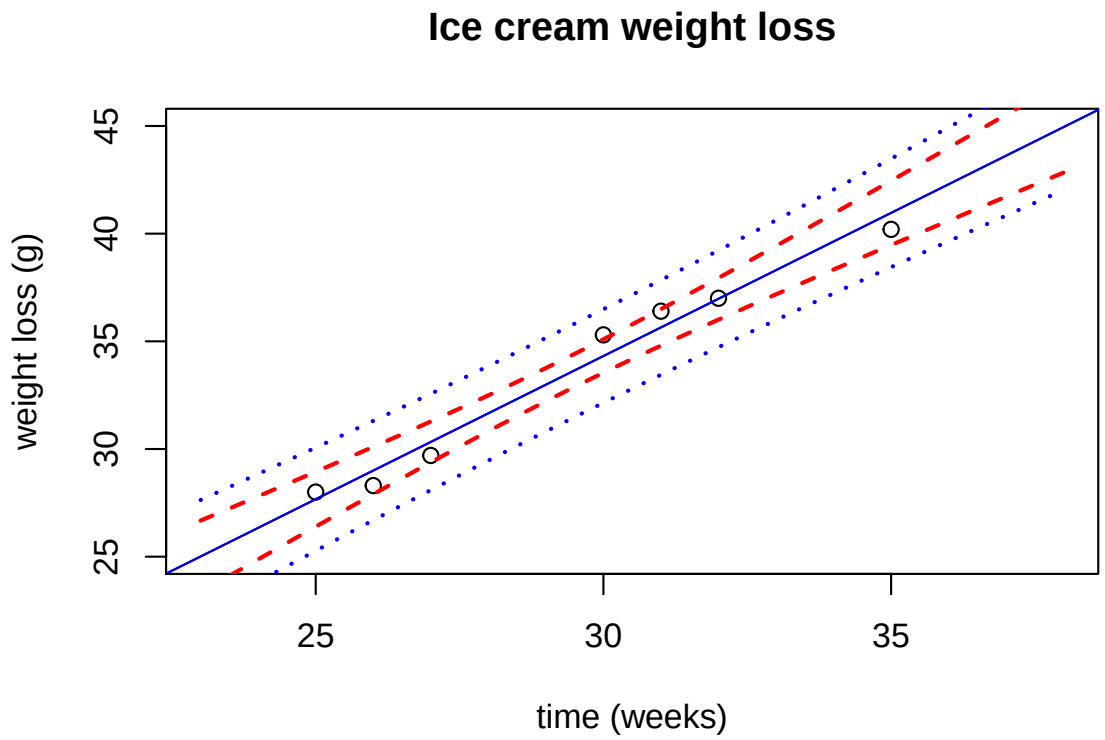
$$I_{E(Y_0)} = \left(\hat{\beta}_0 + \hat{\beta}_1 X_0 \pm t_{\alpha/2, n-2} \cdot s \sqrt{\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum_{j=1}^{n} (X_j - \bar{X})^2}} \right)$$



Prediction interval

Prediction interval for *future observation* $Y_{\text{pred}_0} = \beta_0 + \beta_1 X_0 + \epsilon_0$ for a given X_0:

$$I_{Y_{\text{pred}_0}} = \left(\hat{\beta}_0 + \hat{\beta}_1 X_0 \pm t_{\alpha/2, n-2} \cdot s \sqrt{1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum_{j=1}^{n} (X_j - \bar{X})^2}} \right)$$



Hypothesis testing

We can use the distributions previously derived to answer questions posted in term of *hypotheses*.

Example: we want to test the hypothesis that for each additional week, the stored ice cream loses 1 gram ("null hypothesis" H_0) vs the hypothesis that it loses more then 1 gram ("alternative hypothesis" H_1).

This means testing:

$$\begin{cases} H_0 : \beta_1 = 1 \\ H_1 : \beta_1 > 1 \end{cases}$$

t-test for one β_j

We want to **test** some hypothesis (null, H_0) versus another hypothesis (alternative, H_1).
Motivation: hypothesis $H_0 : \beta_j = m$. Is the difference between $\hat{\beta}_j$ and m (= hypothesized value for the *unknown* β_j) larger than can be reasonably attributed to random variation ("chance")?

Remember the α used for constructing confidence intervals?

$$\alpha = \Pr(\text{accept } H_1 | H_0 \text{ is true})$$

So α is the probability of committing the error of accepting that $\beta_j \neq m$ when in reality it is $\beta_j = m$.
In simple regression we are typically interested in testing $H_0 : \beta_1 = 0$.

Accepting $H_0 : \beta_1 = 0$ implies that variables X and Y are unrelated.

Test H_0: $\beta_1 = 0$ (no linear relation) against H_1: $\beta_1 \neq 0$

If H_0 is true then

$$\frac{\hat{\beta}_1 - 0}{\sigma / \sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}} \sim N(0, 1)$$

$$t = \frac{\hat{\beta}_1 - 0}{s / \sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}} \sim t_{n-2}.$$

Thus reject H_0 at significance level α if $|t| > t_{\alpha/2, n-2}$

Test of β_0 in similar way.

This only works for hypotheses regarding a single β.
In multiple regression we might want to test H_0: $\beta_1 = \beta_2 = 0$ against H_1: "at least one of β_1 and $\beta_2 \neq 0$".

Exercise 1.9, book by Rawlings et al.

1.9. The following data relate biomass production of soybeans to cumulative intercepted solar radiation over an eight-week period following emergence. Biomass production is the mean dry weight in grams of independent samples of four plants. (Data courtesy of Virginia Lesser and Dr. Mike Unsworth, North Carolina State University.)

X	Y
Solar Radiation	*Plant Biomass*
29.7	16.6
68.4	49.1
120.7	121.7
217.2	219.6
313.5	375.5
419.1	570.8
535.9	648.2
641.5	755.6

Exercise 1.9, book by Rawlings et al.

Test a two sides hypothesis: $H_0 : \beta_1 = 0$ vs $H_1 : \beta_1 \neq 0$.

$$\hat{\beta}_1 = 1.269;$$

$$s = \text{Residual standard error} = \sqrt{\text{Residual Mean Squared Error}}$$

$$= \sqrt{\frac{\text{Residual Sum of Squares}}{df}} = \sqrt{\frac{6396}{6}} = \sqrt{1066} = 32.65;$$

$$\text{SE}(\hat{\beta}_1) = \frac{s}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}} \quad \text{(also called } \textbf{standard error of } \hat{\beta}_1\text{)}$$

$$= 32.65\sqrt{2.841082 \cdot 10^{-6}} = 0.05503;$$

$$|t| = \frac{|1.269 - 0|}{0.05503} = 23.063 > t_{0.025,6} = 2.45$$

H_0: $\beta_1 = 0$ should be rejected at $\alpha = 0.05$ (or even $\alpha = 0.0001$).

So solar radiation seems to be have a significant effect in *explaining* plant biomass.

This is because for our linear model β_1 is the coefficient that express the intensity of X. Such intensity is significantly different from zero (that's what our test using available data suggests).

p-values

Software typically report a measure known as p-value in order to decide between competing hypotheses.

$$\alpha = \Pr(\text{accept } H_1 \mid H_0 \text{ is true})$$

$$\text{p-value} = \Pr(|\text{Student's T}| \geq |t| \mid H_0 \text{ true})$$

The p-value is the probability of observing an experiment giving a result as the one produced by your data, *or an even more extreme result*, when in reality such result is due to chance only (H_0 true). If p-value $< \alpha$ our computed t is more extreme that would be expected by chance and H_0 is unlikely to be true.

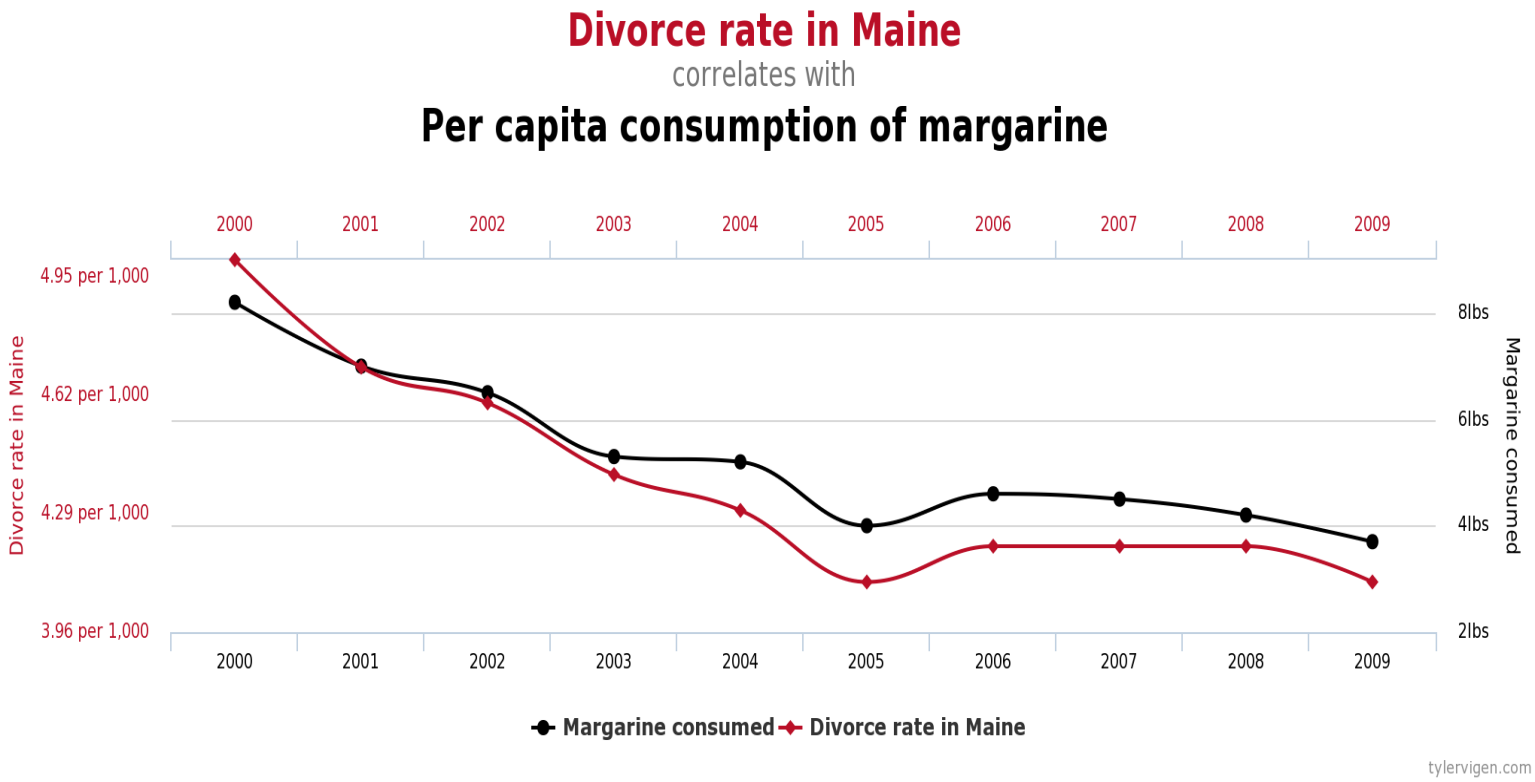
Test using p-value

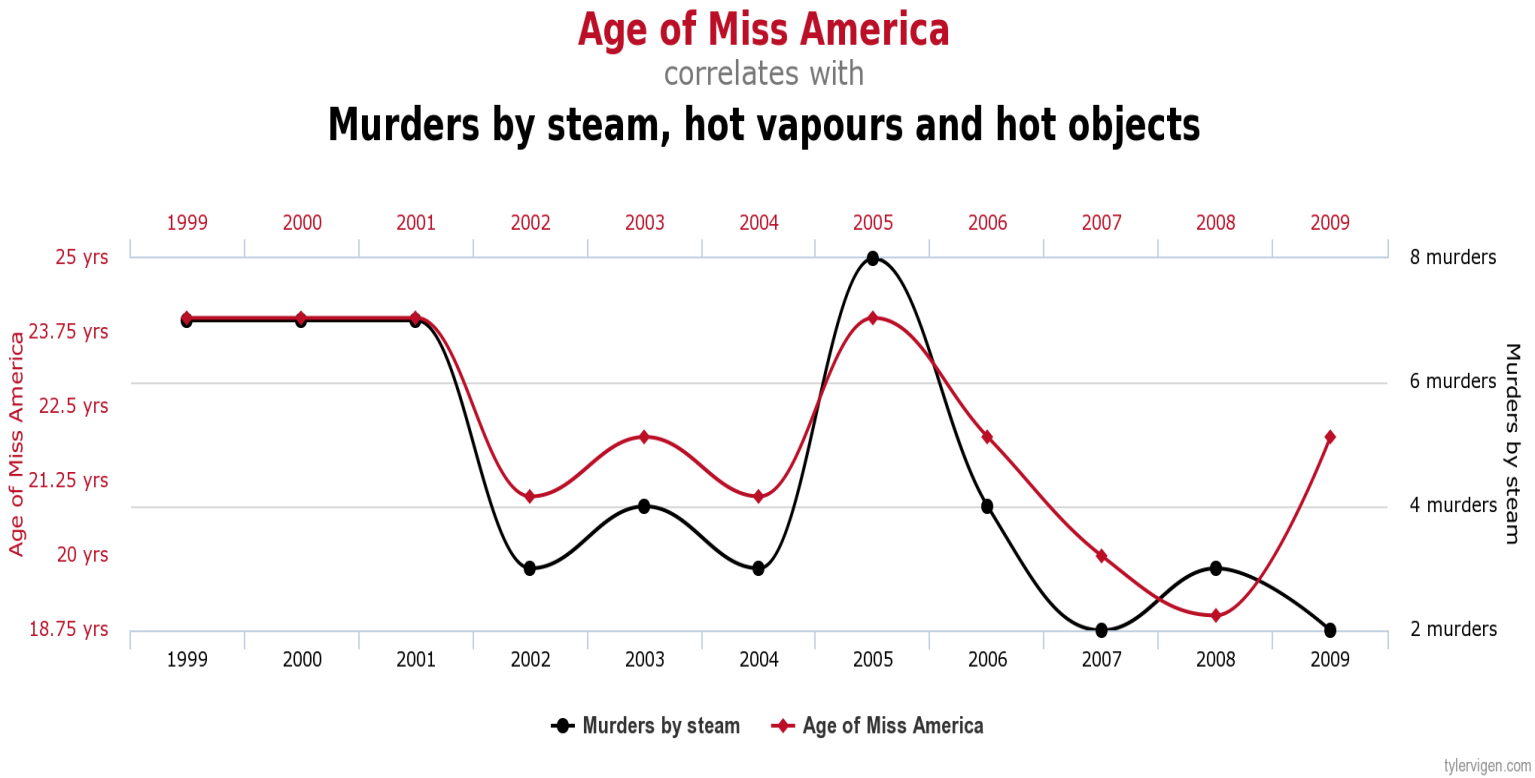
Reject H_0 at significance level α if p-value $< \alpha$.

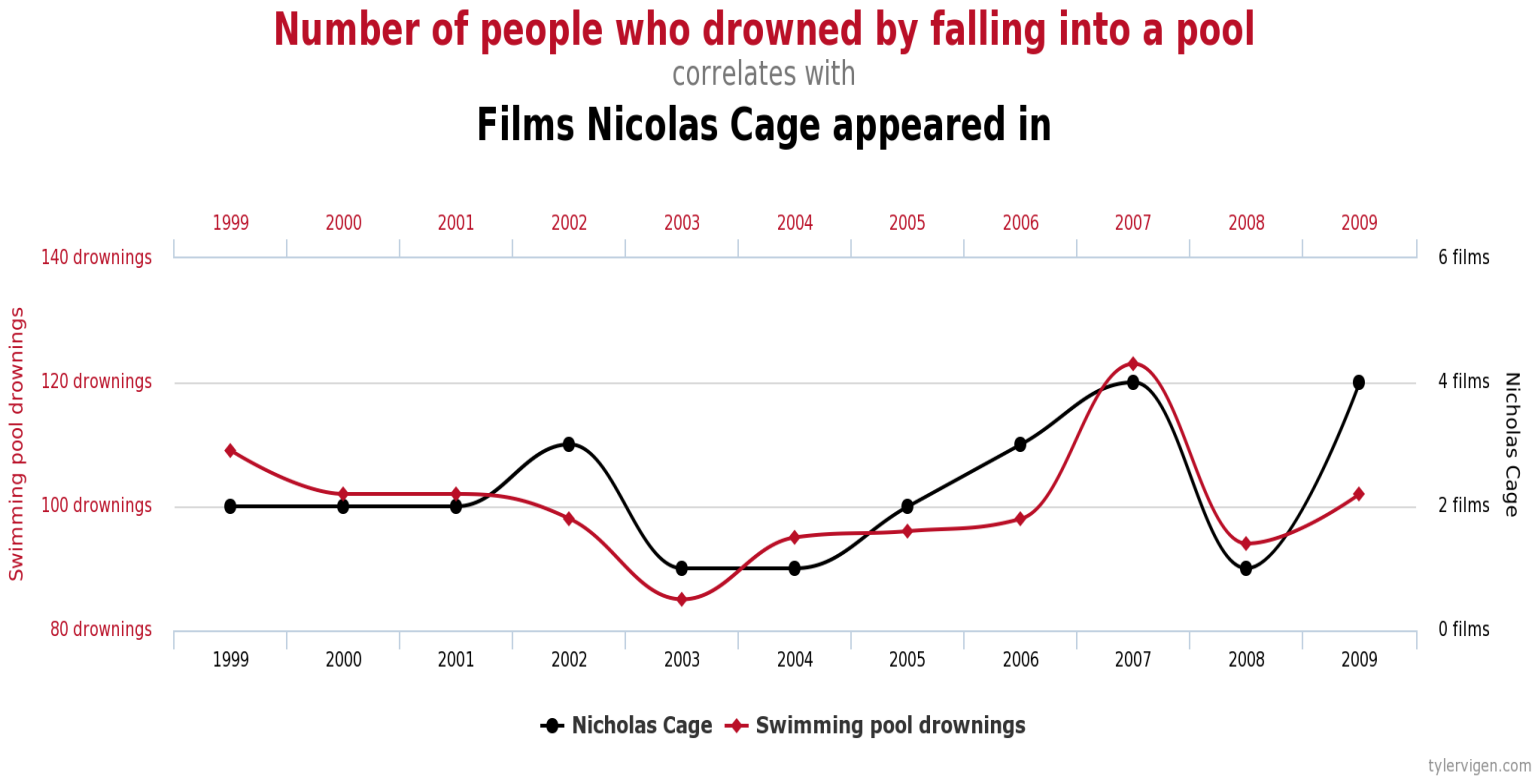
Relation between CI and hypothesis testing

- It is important to note that confidence intervals and hypothesis tests are built on exactly the same principles. The same statistic t is employed.
- This means that their conclusions are always consistent!
- Example: if we end up accepting that $H_0 : \beta_j = m$ for some m and significance level α, for sure the corresponding confidence interval for β_j **will contain** the value m. If we reject H_0 the confidence interval **will not contain** m.

Some relations that are absurd and irrelevant




Age of Miss America
correlates with
Murders by steam, hot vapours and hot objects
1999 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009
25 yrs
23.75 yrs
22.5 yrs
21.25 yrs
20 yrs
18.75 yrs
Age of Miss America
8 murders
6 murders
4 murders
2 murders
Murders by steam
Murders by steam
Age of Miss America
tylervigen.com


Number of people who drowned by falling into a pool
correlates with
Films Nicolas Cage appeared in
1999 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009
140 drownings
120 drownings
100 drownings
80 drownings
Swimming pool drownings
6 films
4 films
2 films
0 films
Nicholas Cage
Nicholas Cage
Swimming pool drownings
tylervigen.com

Checking assumptions

We will actually go quite deep on checking assumptions in *multiple linear regression*. Here is a teaser.
Refer to the biomass data.

- linearity: easy to check `plot(y ~ x)` in simple linear regression. Very difficult in multiple regression.
- errors are randomly scattered around the zero: use **residuals** $e_i = y_i - \hat{y}_i$

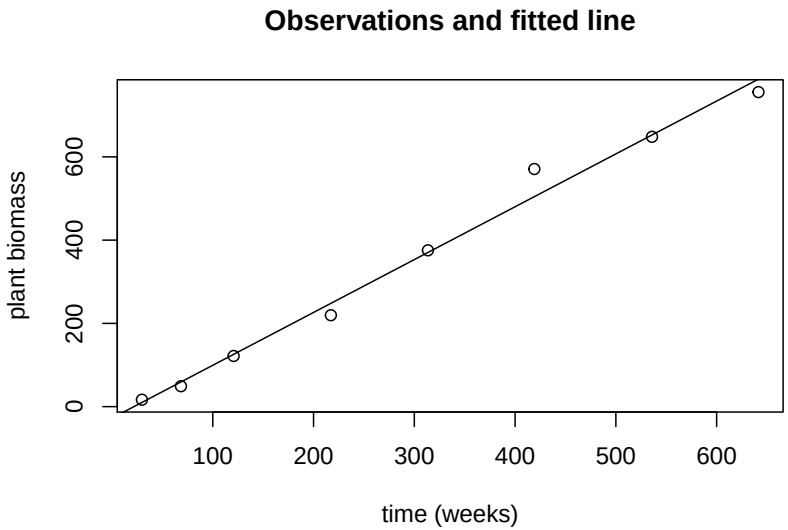
```
model1 <- lm(y ~ x)
e <- residuals(model1)
plot(e)
plot(e ~ x)
plot(e ~ model1$fitted.values)
abline(h = 0)
```

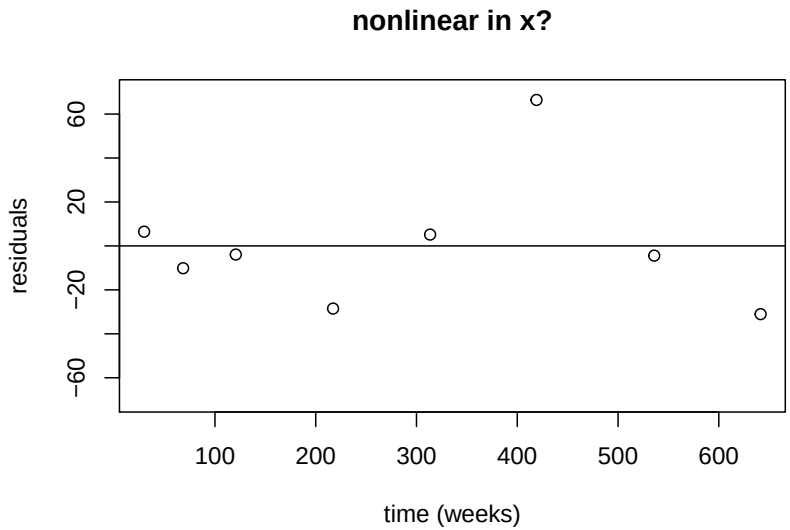
Checking assumptions

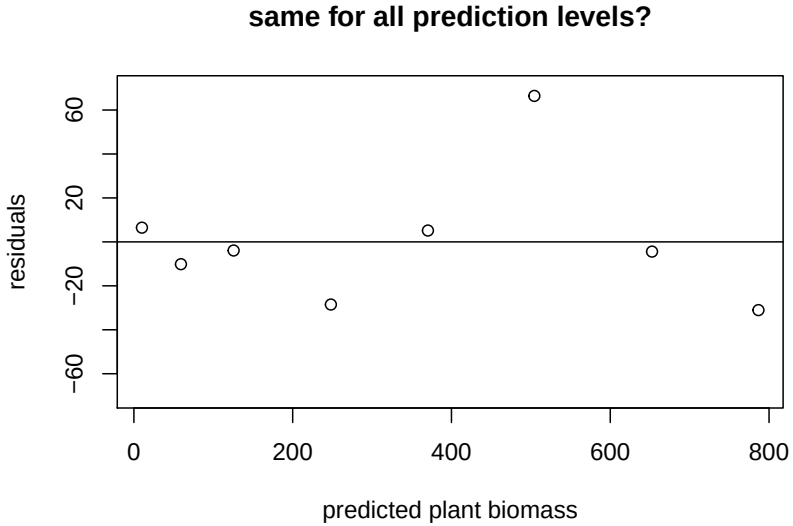
- Gaussianity of errors: difficult to check with few data points! Use qq-plots on residuals:

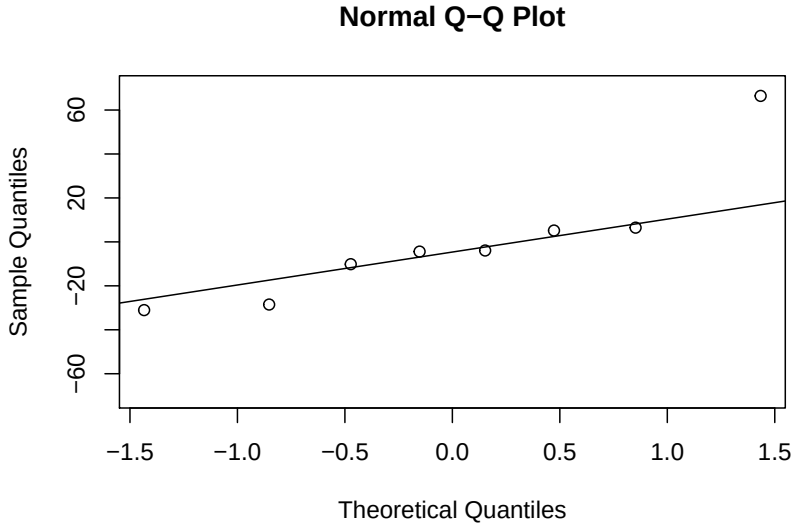
```
qqnorm(e)  # difficult to test with few points
qqline(e)
hist(e)
```

More discussion on these points in next lectures.


Observations and fitted line
plant biomass
0
200
400
600
100
200
300
400
500
600
time (weeks)


nonlinear in x?
residuals
−60
−20
20
60
100
200
300
400
500
600
time (weeks)


same for all prediction levels?
residuals
−60
−20
20
60
0
200
400
600
800
predicted plant biomass


Normal Q−Q Plot
Sample Quantiles
−60
−20
20
60
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
Theoretical Quantiles