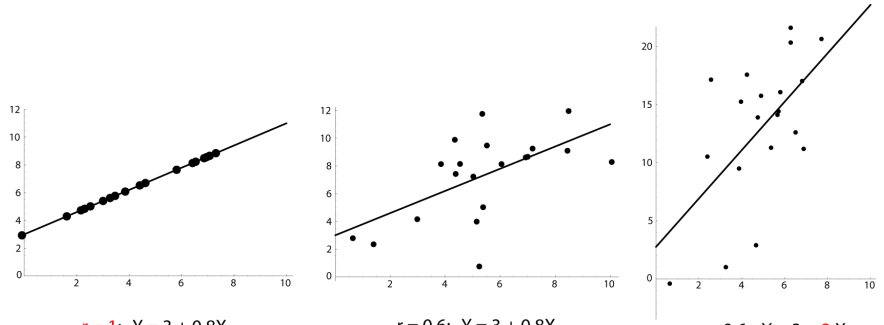
Regression

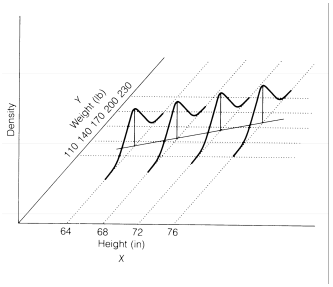
- Predicts *Y* from *X*
- Linear regression assumes that the relationship between *X* and *Y* can be described by a line

Correlation vs. regression



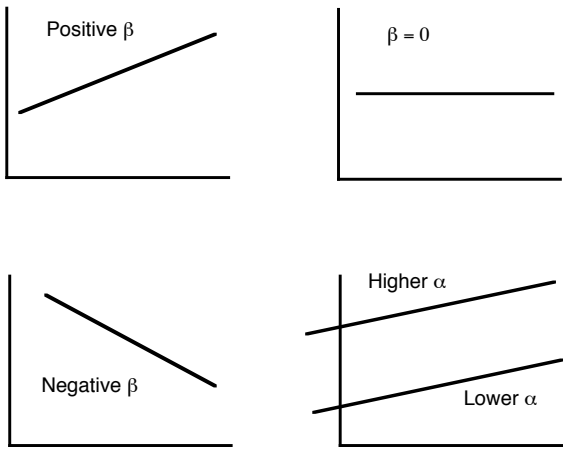
Regression assumes...

- Random sample
- *Y* is normally distributed with equal variance for all values of *X*



The parameters of linear regression

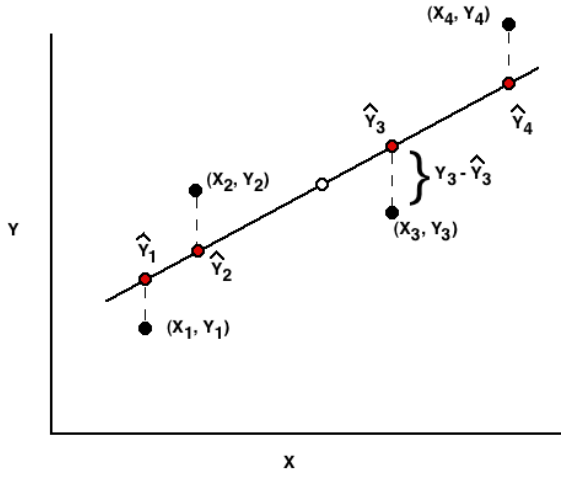
$$Y = \alpha + \beta X$$



Estimating a regression line

$$Y = a + b X$$

Nomenclature



Residual:

$$Y_i - \hat{Y}_i$$

Finding the "least squares" regression line

Minimize: $$SS_{residual} = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i \right)^2$$

Best estimate of the slope

$$b = \frac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2}$$

(= "Sum of cross products" over "Sum of squares of X")

Remember the shortcuts:

$$\sum_{i=1}^{n}\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right) = \left(\sum X_i Y_i\right) - \frac{\sum X_i \sum Y_i}{n}$$

$$\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2 = \sum\left(X_i{}^2\right) - \frac{\left(\sum X_i\right)^2}{n}$$

Finding *a*

$$\bar{Y} = a + b\bar{X}$$

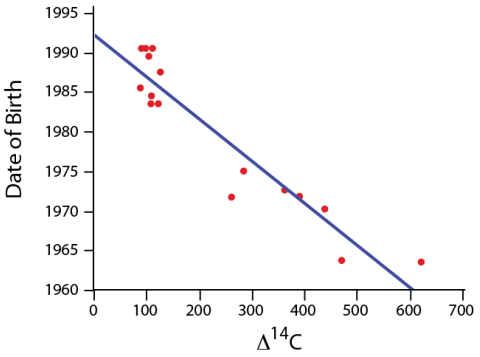
So..

$$a = \bar{Y} - b\bar{X}$$

Example: Predicting age based on radioactivity in teeth

Many above ground nuclear bomb tests in the '50s and '60s may have left a radioactive signal in developing teeth.

Is it possible to predict a person's age based on dental C^{14}?



Data from 1965 to present from Spalding et al. 2005. Forensics: age written in teeth by nuclear tests. ***Nature*** *437: 333–334.*

Teeth data:

$\Delta^{14}C$	Date of Birth
89	1985.5
109	1983.5
91	1990.5
127	1987.5
99	1990.5
110	1984.5
123	1983.5
105	1989.5

$\Delta^{14}C$	Date of Birth
622	1963.5
262	1971.7
471	1963.7
112	1990.5
285	1975
439	1970.2
363	1972.6
391	1971.8

Teeth data:

Let X be the estimated age, and Y be the actual age.

$$\sum X = 3798, \quad \sum Y = 31674$$

$$\sum X^2 = 1340776, \quad \sum (XY) = 7495223$$

$$\sum Y^2 = 62704042$$

$$n = 16$$

$$\bar{X} = 237.375 \qquad \bar{Y} = 1979.63$$

$$\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}) = \left(\sum X_i Y_i\right) - \frac{\sum X_i \sum Y_i}{n}$$

$$= 7495223 - \frac{(3798)(31674)}{16} = -23393$$

$$\sum_{i=1}^{n}(X_i - \bar{X})^2 = \sum (X_i^2) - \frac{\left(\sum X_i\right)^2}{n}$$

$$= 1340776 - \frac{(3798)^2}{16} = 439226$$

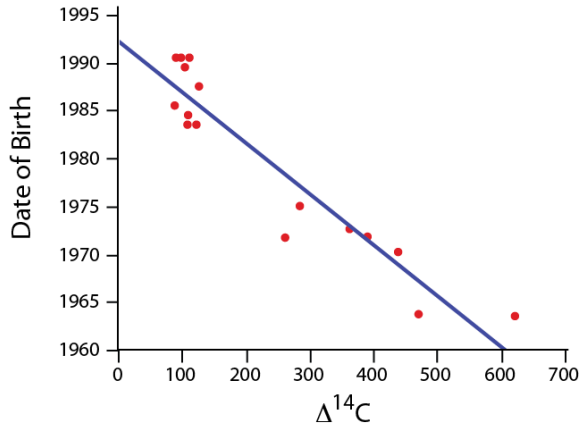
$$b = \frac{-23393}{439226} = -0.053$$

Calculating a

$$a = \bar{Y} - b\bar{X}$$

$$= 1979.63 - (-0.053)237.375 = 1992.2$$

$$\hat{Y} = 1992.2 - 0.053X$$



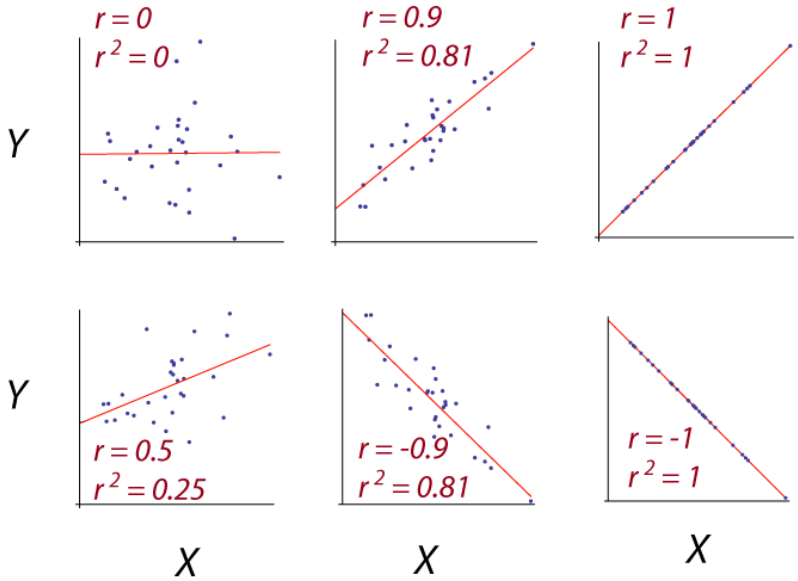
Predicting *Y* from *X*

If a cadaver has a tooth with $\Delta^{14}C$ content equal to 200, what does the regression line predict its year of birth to be?

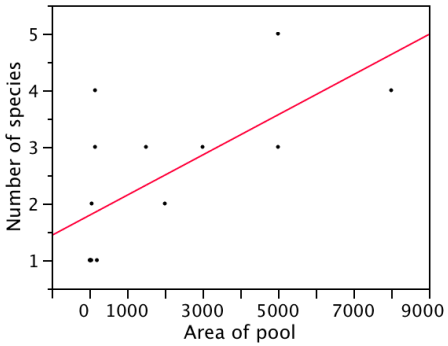
$$\hat{Y} = 1992.2 - 0.053X$$
$$= 1992.2 - 0.053(200)$$
$$= 1981.6$$

r^2 predicts the amount of variance in *Y* explained by the regression line

r^2 is the "coefficient of determination: it is the square of the correlation coefficient r



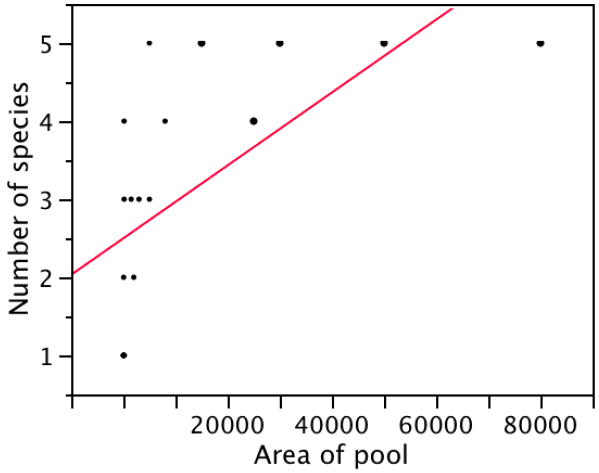
Caution: It is unwise to extrapolate beyond the range of the data.



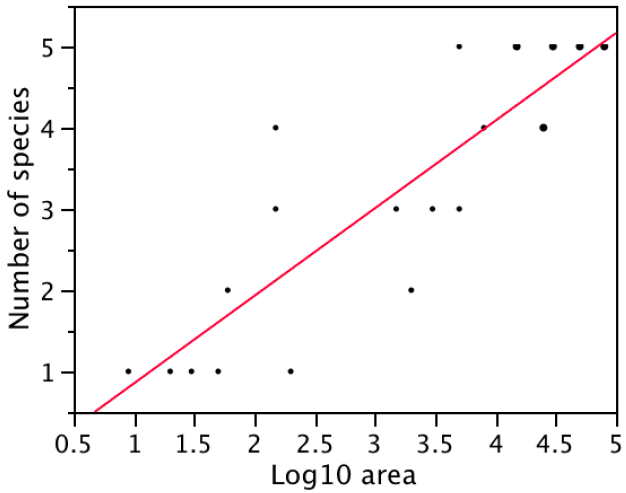
Number of species of fish as predicted by the area of a desert pool

If we were to extrapolate to ask how many species might be in a pool of 50000m^2, we would guess about 20.

More data on fish in desert pools



Log transformed data:



Testing hypotheses about regression

$H_0: \beta = 0$

$H_A: \beta \neq 0$

Sums of squares for regression

$$SS_{Total} = \sum Y_i^2 - \frac{\left(\sum Y_i\right)^2}{n}$$

$$SS_{regression} = b\sum\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)$$

$$SS_{residual} + SS_{regression} = SS_{Total}$$

With n - 2 degrees of freedom for the residual

Radioactive teeth: Sums of squares

$$SS_{Total} = \sum Y_i^2 - \frac{\left(\sum Y_i\right)^2}{n}$$

$$= 62704042 - \frac{(31674)^2}{16} = 1339.75$$

$$SS_{regression} = b\sum\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)$$

$$= (-0.053)(-23393) = 1239.8$$

Teeth: Sums of squares

$$SS_{residual} = SS_{Total} - SS_{regression}$$

$$= 1339.75 - 1239.8$$

$$= 99.9$$

$$df_{residual} = 16 - 2 = 14$$

Calculating residual mean squares

$$MS_{residual} = SS_{residual} / df_{residual}$$

$$MS_{residual} = \frac{99.9}{14} = 7.1$$

Standard error of a slope

$$SE_b = \sqrt{\frac{MS_{residual}}{\sum\left(X_i - \bar{X}\right)^2}}$$

$$= \sqrt{\frac{7.1}{439226}} = 0.004$$

b has a *t* distribution

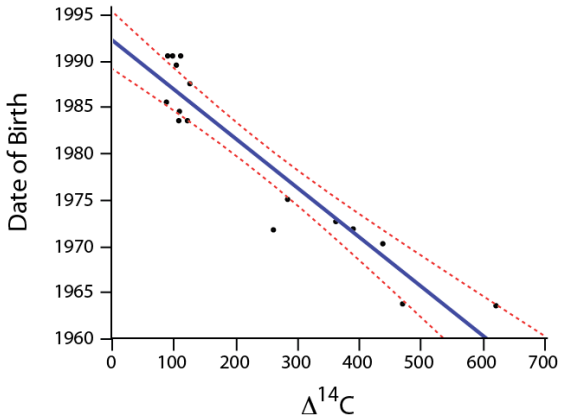
Confidence interval for a slope: $b \pm t_{\alpha[2],df} SE_b$

Hypothesis tests can use *t*: $t = \frac{b - \beta_0}{SE_b}$

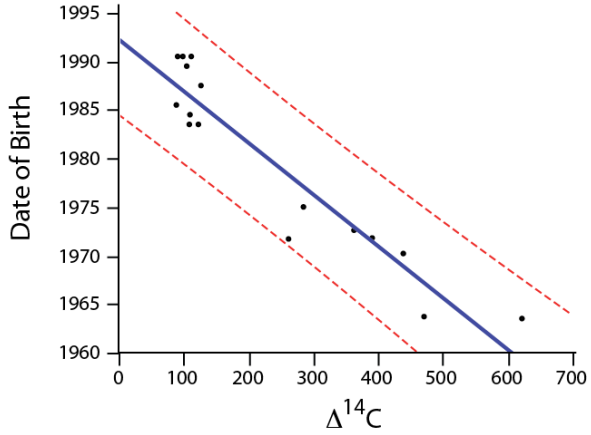
Example: 95% confidence interval for slope with teeth example

$$b \pm t_{\alpha[2],df} SE_b = b \pm t_{0.05[2],14} SE_b$$
$$= -0.053 \pm 2.14(0.004)$$
$$= -0.053 \pm 0.0018$$

Confidence bands: confidence intervals for predictions of mean *Y*



Prediction intervals: confidence intervals for predictions of individual *Y*



Hypothesis tests on slopes

H_0: β = 0

H_A: β ≠ 0

$$t = \frac{b - \beta_0}{SE_b}$$

$$t = \frac{-0.053 - 0}{0.004} = 13.25$$

$t_{0.0001(2),14}$ = ±5.36

So we can reject H_0, *P*<0.0001

Non-linear relationships

Transformations

Quadratic regression

Splines

Transformations

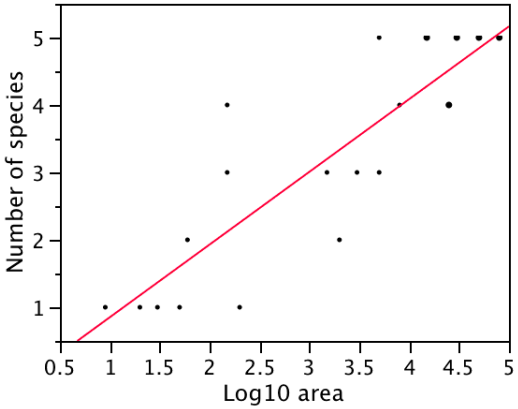
If $Y = aX^b$ then $\ln Y = \ln a + b \ln X$.

If $Y = ab^X$ then $\ln Y = \ln a + X \ln b$.

If $Y = a + \frac{b}{X}$ then set $X' = \frac{1}{X}$, and calculate $Y = a + bX'$.

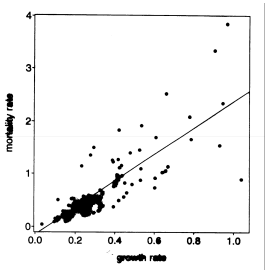
All of the equations on the right have the form Y=a+bX.

Non-linear relationship: Number of fish species vs. Size of desert pool

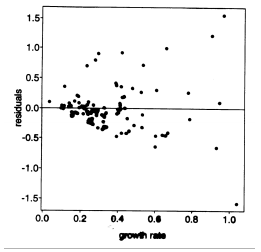


Residual plots help assess assumptions

Original:

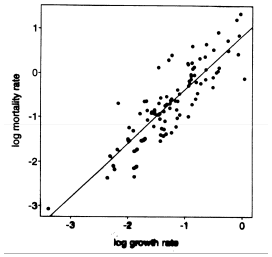


Residual plot

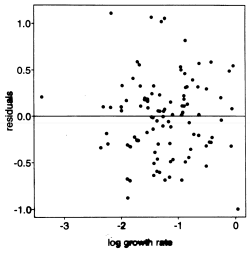


Transformed data

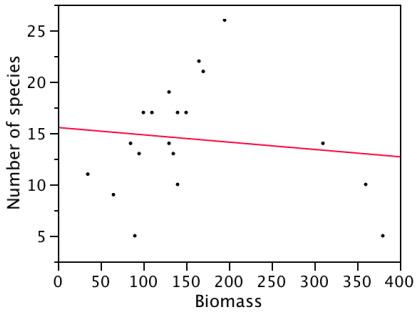
Logs:

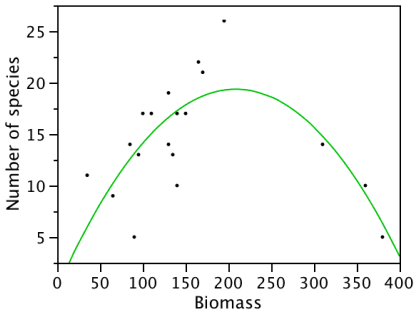


Residual plot



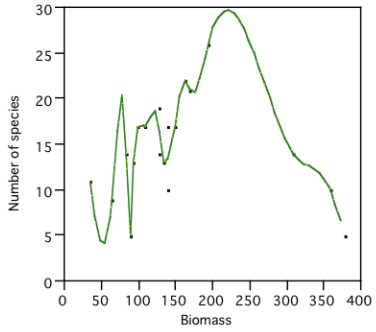
Polynomial regression





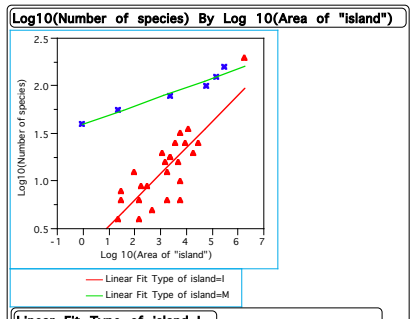
Number of species = 0.046 + 0.185 Biomass - 0.00044 Biomass^2

Do not fit a polynomial with too many terms (the sample size should be at least 7 times the number of terms)



Comparing two slopes

Example: Comparing species-area curves for islands to those of mainland populations



Hypotheses

H_0: $\beta_M = \beta_I$.

H_A: $\beta_M \neq \beta_I$.

The error in the difference of two slopes is normally distributed.

$$t = \frac{(b_1 - b_2) - (\beta_1 - \beta_2)}{SE_{b_1 - b_2}}$$

$df = n_1 - 2 + n_2 - 2$

$$\left(MS_{error}\right)_p = \frac{\left(SS_{error}\right)_1 + \left(SS_{error}\right)_2}{\left(DF_{error}\right)_1 + \left(DF_{error}\right)_2}$$

$$SE_{b1-b2} = \sqrt{\frac{\left(MS_{error}\right)_p}{\left(\sum\left(X-\bar{X}\right)^2\right)_1} + \frac{\left(MS_{error}\right)_p}{\left(\sum\left(X-\bar{X}\right)^2\right)_2}}$$

Analysis of covariance (ANCOVA)

Compares many slopes

H_0: $\beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 \ldots$

H_A: At least one of the slopes is different from another.

Logistic regression

Tests for relationship between a numerical variable (as the explanatory variable) and a binary variable (as the response).

e.g.: Does the dose of a toxin affect probability of survival?

Does the length of a peacock's tail affect its probability of getting a mate?