

Regression

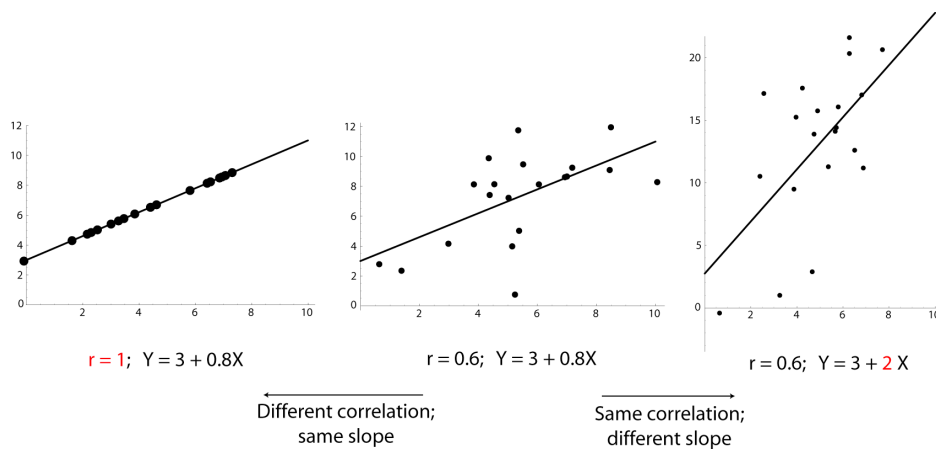
Chapter 17

Regression

Predicts Y from X

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

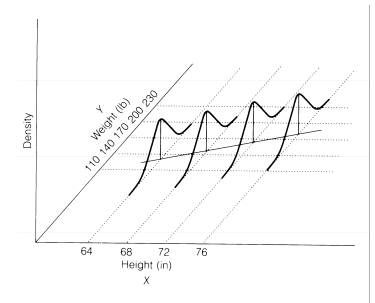
Correlation vs. regression



Regression assumes...

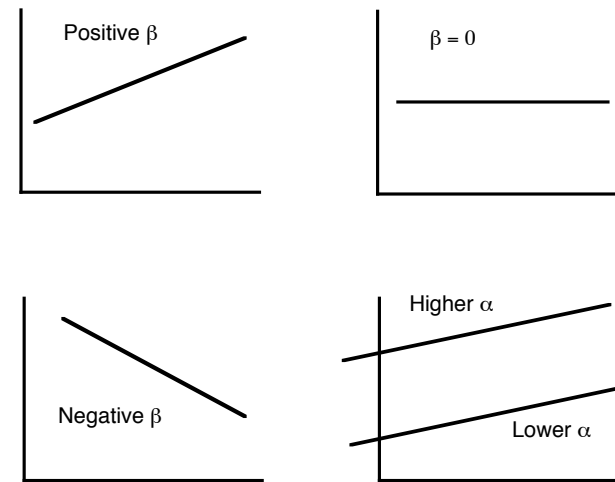
Random sample

Y is normally distributed with equal variance for all values of X



The parameters of linear regression

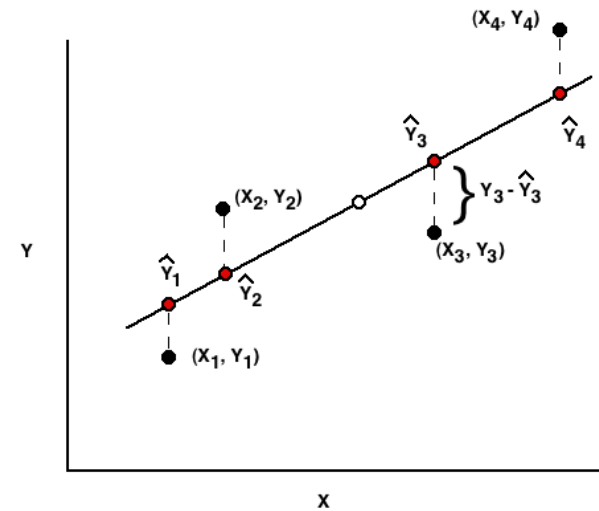
$$Y = \alpha + \beta X$$



Estimating a regression line

$$Y = a + b X$$

Nomenclature



Residual:

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

Finding the "least squares"
regression line

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

Best estimate of the slope

$$b = \frac{\text{Covariance}(X, Y)}{\text{Variance}(X)}$$

Best estimate of the slope

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

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

Finding a

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

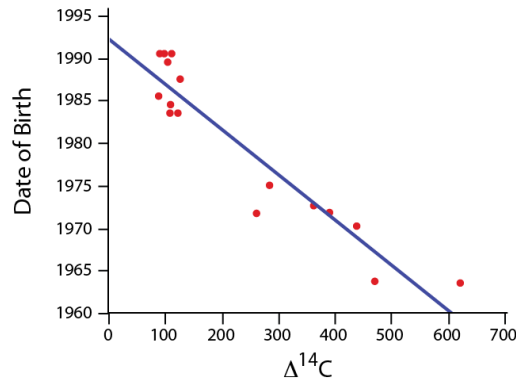
So..

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

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 ^{14}C ?



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

Teeth data:

$\Delta^{14}\text{C}$	Date of Birth	$\Delta^{14}\text{C}$	Date of Birth
89	1985.5	622	1963.5
109	1983.5	262	1971.7
91	1990.5	471	1963.7
127	1987.5	112	1990.5
99	1990.5	285	1975
110	1984.5	439	1970.2
123	1983.5	363	1972.6
105	1989.5	391	1971.8

Teeth data:

Let X be the $\Delta^{14}\text{C}$, and Y be the year of birth.

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

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

$$\sum Y^2 = 62704042$$

$$n = 16$$

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

Remember the shortcuts:

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

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

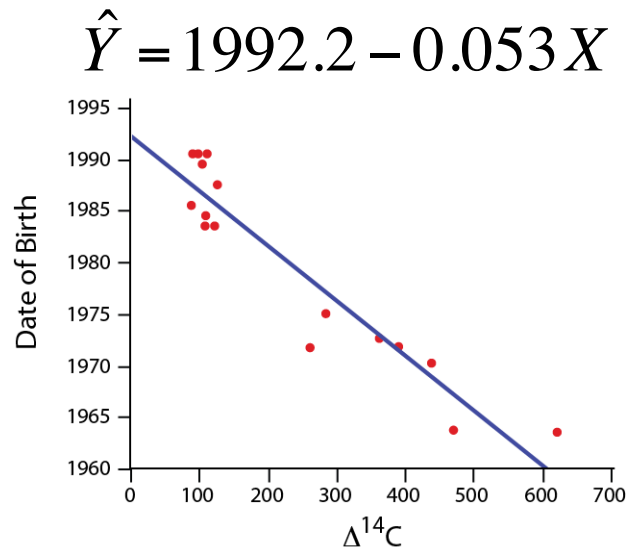
$$\begin{aligned}\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y}) &= \left(\sum X_i Y_i \right) - \frac{\sum X_i \sum Y_i}{n} \\ &= 7495223 - \frac{(3798)(31674)}{16} = -23393\end{aligned}$$

$$\begin{aligned}\sum_{i=1}^n (X_i - \bar{X})^2 &= \sum (X_i^2) - \frac{\left(\sum X_i \right)^2}{n} \\ &= 1340776 - \frac{(3798)^2}{16} = 439226\end{aligned}$$

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

Calculating a

$$\begin{aligned}a &= \bar{Y} - b\bar{X} \\ &= 1979.63 - (-0.053)237.375 = 1992.2\end{aligned}$$



Predicting Y from X

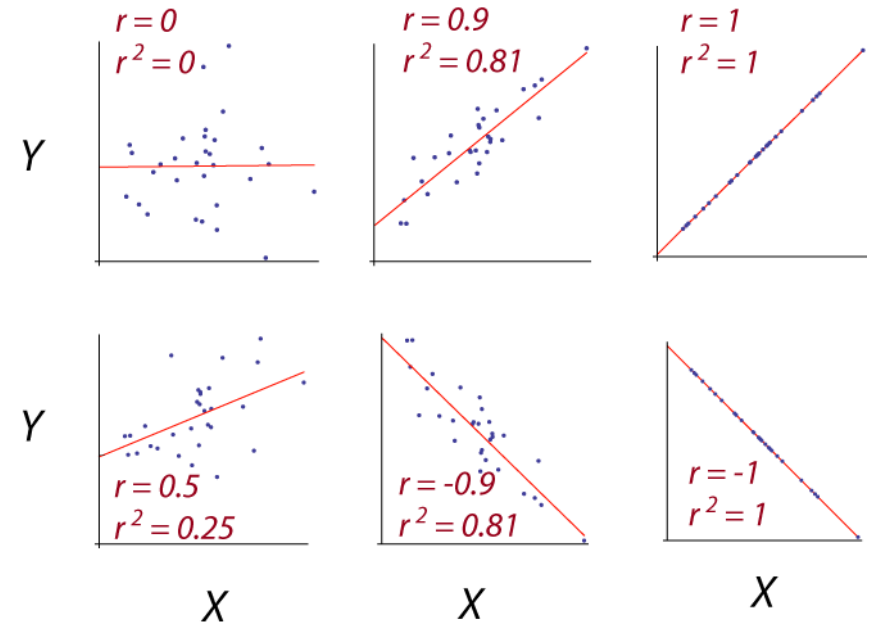
If a cadaver has a tooth with $\Delta^{14}\text{C}$ content equal to 200, what does the regression line predict its year of birth to be?

$$\begin{aligned}\hat{Y} &= 1992.2 - 0.053X \\ &= 1992.2 - 0.053(200) \\ &= 1981.6\end{aligned}$$

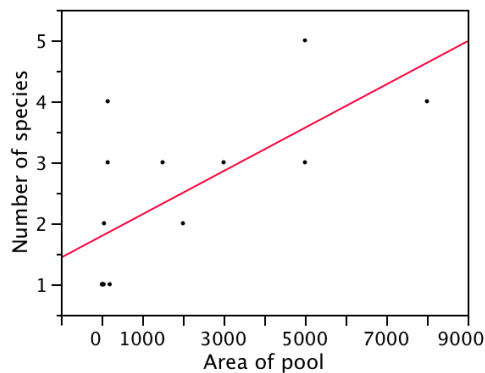
r^2 predicts the amount of variance in Y explained by the regression line

r^2 is the “coefficient of determination:

It is the square of the correlation coefficient r



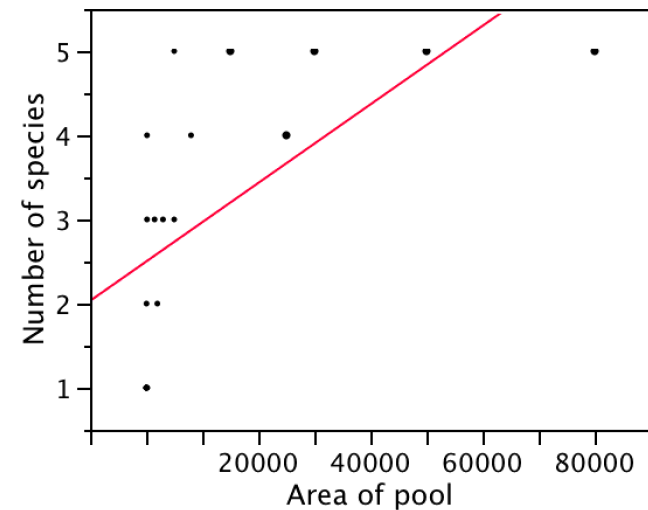
Caution: It is unwise to extrapolate beyond the range of the data.



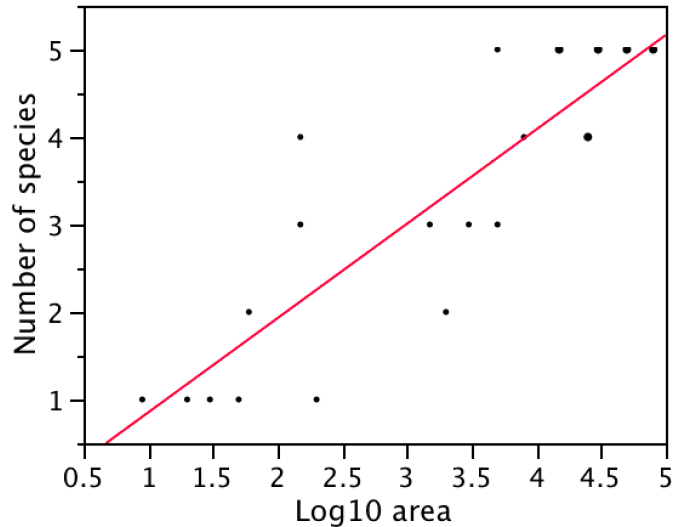
Number of species of fish as predicted by the area of a desert pool

If we were to extrapolate to ask how many species might be in a pool of 50000m², we would guess about 20.

More data on fish in desert pools



Log transformed data:



Testing hypotheses about regression

$$H_0: \beta = 0$$

$$H_A: \beta \neq 0$$

b has a t distribution

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

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

Standard error of a slope

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

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

Sums of squares for regression

$$SS_{total} = \sum (Y_i - \bar{Y})^2$$

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

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

With $n - 2$ degrees of freedom for the residual

Teeth: Sums of squares

$$\begin{aligned} SS_{residual} &= SS_{total} - SS_{regression} \\ &= 1399.87 - 1245.90 \\ &= 154.0 \end{aligned}$$

$$\begin{aligned} df_{residual} &= n - 2 \\ &= 16 - 2 \\ &= 14 \end{aligned}$$

Radioactive teeth: Sums of squares

$$SS_{total} = \sum (Y_i - \bar{Y})^2 = 1399.87$$

$$SS_{regression} = b \sum (X_i - \bar{X})(Y_i - \bar{Y}) = 1245.9$$

Calculating residual mean squares

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

$$MS_{residual} = \frac{154.0}{14} = 11.0$$

Standard error of the slope

$$\begin{aligned} SE_b &= \sqrt{\frac{MS_{residual}}{\sum(X_i - \bar{X})^2}} \\ &= \sqrt{\frac{11.0}{439226}} \\ &= 0.005 \end{aligned}$$

95% confidence interval for slope with teeth example

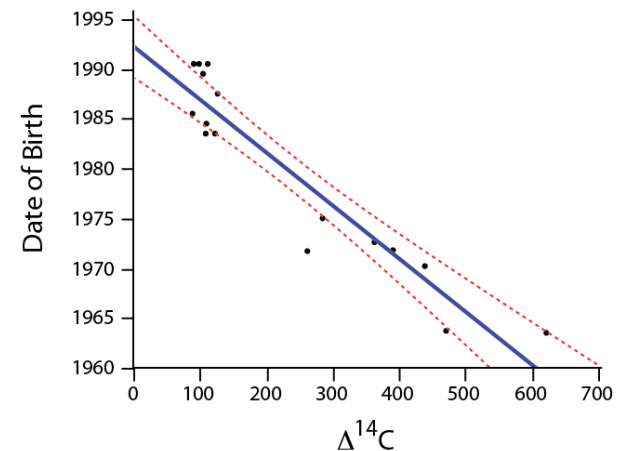
$$\begin{aligned} b \pm t_{\alpha[2],df} SE_b &= b \pm t_{0.05[2],14} SE_b \\ &= -0.053 \pm 2.14(0.005) \\ &= -0.053 \pm 0.011 \end{aligned}$$

b has a t distribution

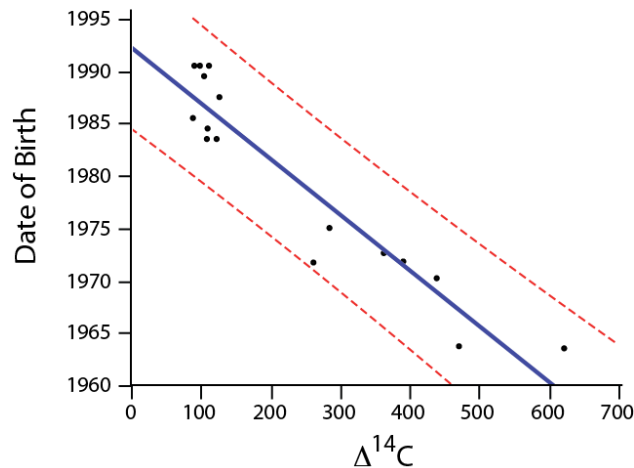
Confidence interval for a slope: $b \pm t_{\alpha[2],df} SE_b$

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

Confidence bands:
confidence intervals for
predictions of mean Y



Prediction intervals: confidence intervals for predictions of individual Y



Hypothesis tests on slopes

$$H_0: \beta = 0$$

$$H_A: \beta \neq 0$$

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

$$t = \frac{-0.053 - 0}{0.005} = 10.6$$

$$t_{0.0001(2),14} = \pm 5.36$$

So we can reject H_0 , $P < 0.0001$

Regression in R

```
teethRegression <- lm(dateOfBirth ~ deltaC14, data = teethData)
summary(teethRegression)
```

```
Call:
lm(formula = dateOfBirth ~ deltaC14, data = teethData)

Residuals:
    Min       1Q   Median       3Q      Max
-6.6135 -2.1205  0.0113  2.8884  4.3598

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.992e+03  1.449e+00  1375.26 < 2e-16 ***
deltaC14     -5.326e-02  5.004e-03  -10.64  4.3e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.317 on 14 degrees of freedom
Multiple R-squared:  0.89, Adjusted R-squared:  0.8821
F-statistic: 113.3 on 1 and 14 DF, p-value: 4.296e-08
```

Regression in R

```
confint(teethRegression)
```

	2.5 %	97.5 %
(Intercept)	1989.16033201	1995.37440610
deltaC14	-0.06399224	-0.04252588

This generates the confidence interval for the estimate of the slope.

Non-linear relationships

Transformations

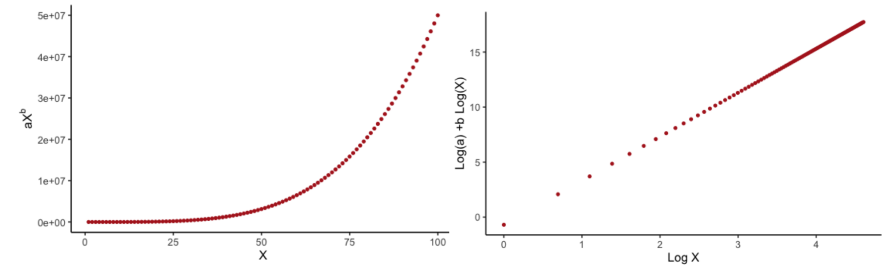
Quadratic regression

Splines

Sometimes transformations make non-linear relationships linear

$$Y = aX^b$$

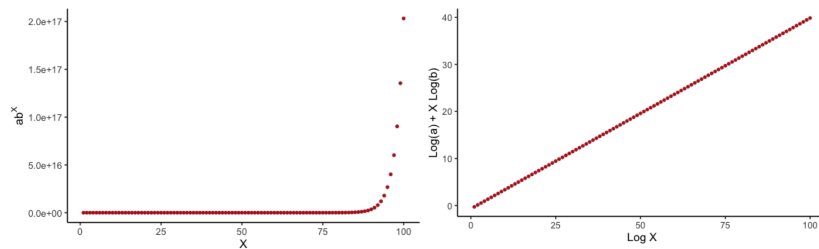
$$\log(Y) = \log(a) + b \log(X)$$



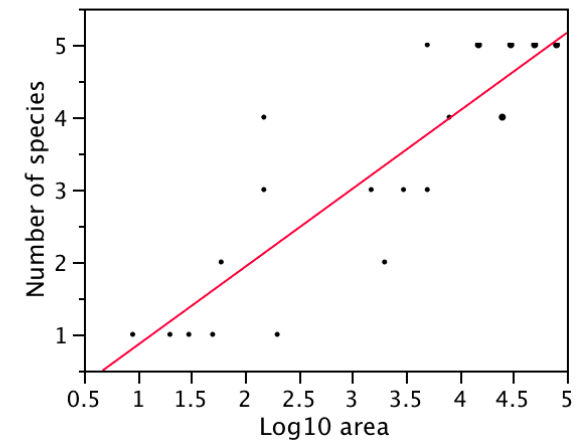
Sometimes transformations make non-linear relationships linear

$$Y = ab^X$$

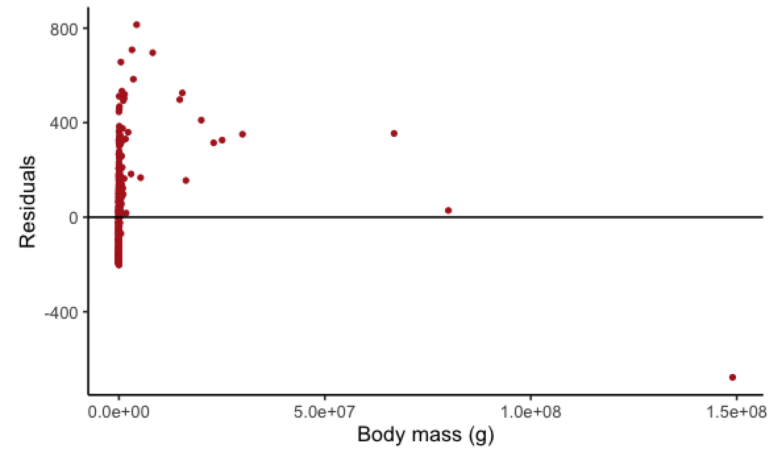
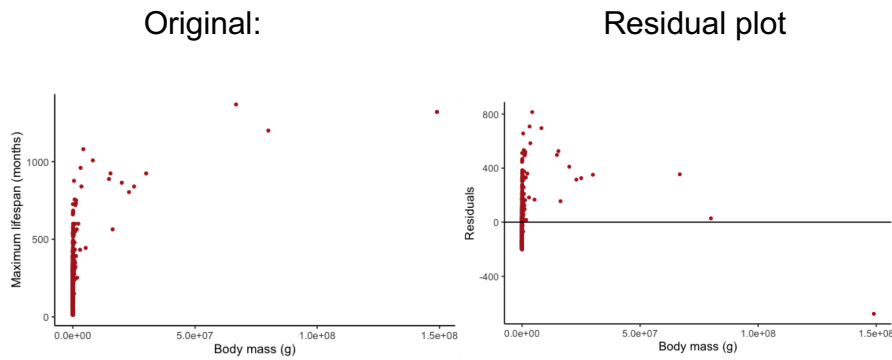
$$\log(Y) = \log(a) + X \log(b)$$



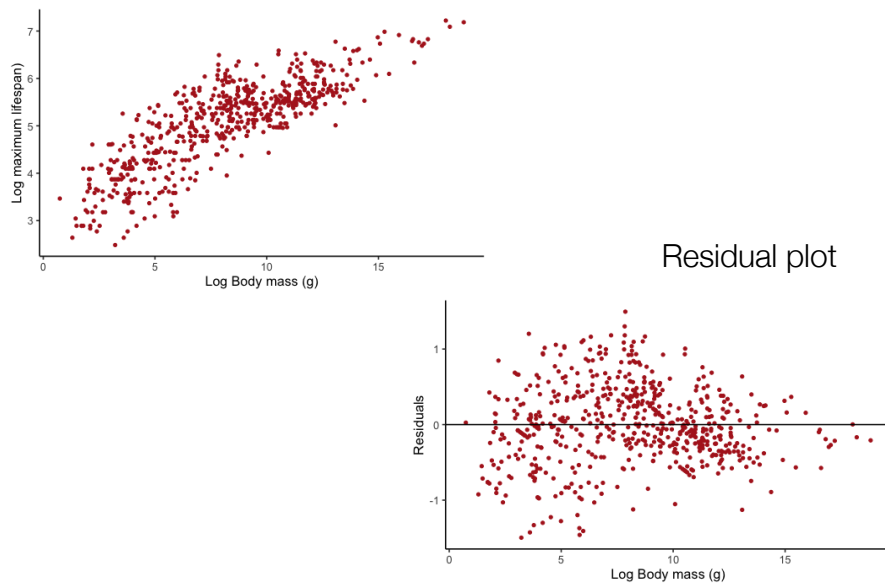
Non-linear relationship:
Number of fish species vs. Size of desert pool



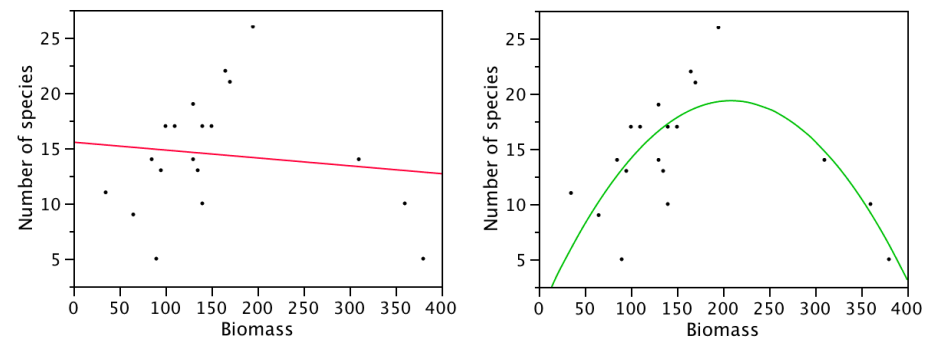
Residual plots help assess assumptions



Log transformation of both variables:

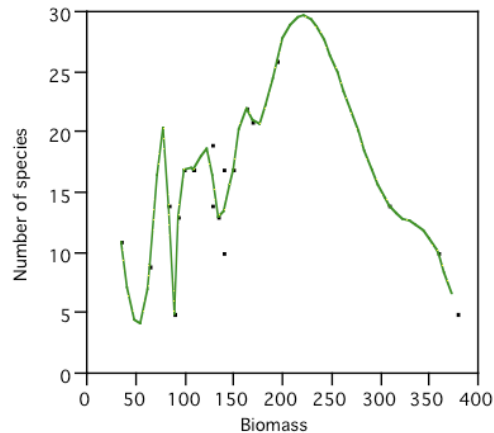


Polynomial regression



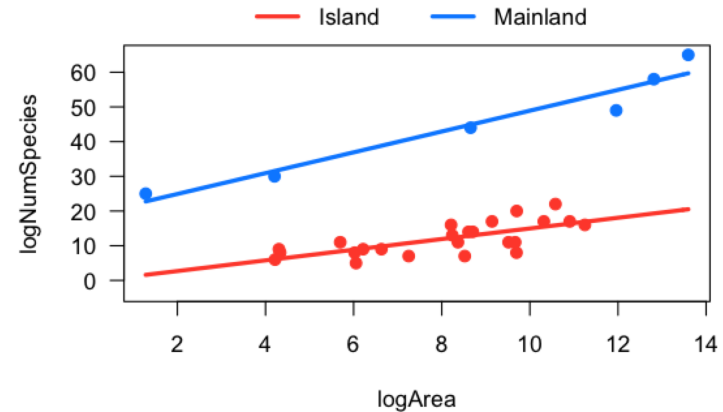
$$\text{Number of species} = 0.046 + 0.185 \text{ Biomass} - 0.00044 \text{ Biomass}^2$$

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



ANCOVA: 3 hypotheses

H_0 : Area has no effect on the number of species.

H_0 : Islands and mainland areas are the same for mean number of species.

H_0 : Area and island/mainland type do not interact in determining the number of species.

ANOVA table from R for ANCOVA

Anova Table (Type III tests)

Response: logNumSpecies

	Sum Sq	Df	F value	Pr(>F)
(Intercept)	0.21	1	0.017	0.8972079
IslandMainland	260.84	1	21.343	9.17e-05 ***
logArea	246.30	1	20.153	0.0001294 ***
IslandMainland:logArea	123.04	1	10.068	0.0038523 **
Residuals	317.76	26		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Analysis of covariance (ANCOVA)

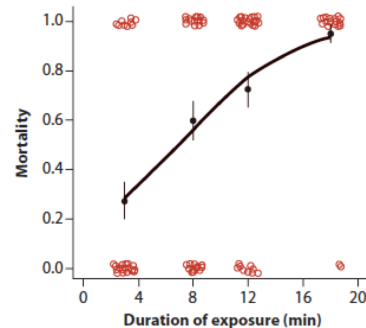
Compares many slopes

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

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

FIGURE 17.9-1

Mortality of guppies in relation to duration of exposure to a temperature of 5°C (data from Pitkow 1960). Treatments were 3, 8, 12, or 18 minutes of exposure, with 40 fish in each of the four treatments. Each point (red circle) indicates a different individual (points were offset using a random perturbation to reduce overlap). $Y = 1$ if the individual died, whereas $Y = 0$ if the individual survived. Black dots indicate the proportion of deaths (± 1 SE) in each treatment. The curve is the logistic regression predicting the probability of death.



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?