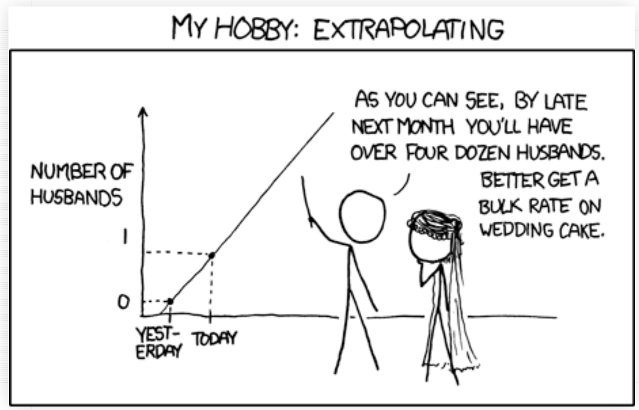
Linear Regression

M. Drew LaMar
September 09, 2016



https://xkcd.com/605/

Introduction to Quantitative Biology, Fall 2016

Class announcements

- Biology Seminar today, 4:00 pm in Millington 150
- To celebrate, NO READING QUIZ FOR MONDAY
- I will start grading stuff soon. Sorry for the delay!
- Homework #2
    - OpenStats, Chapter 4: 4.6.3 Hypothesis testing (p. 209) - #4.18, 4.20, 4.22, 4.24, 4.28, 4.30
    - OpenStats, Chapter 7: 7.5.1 Line fitting, residuals, and correlation (p. 356) - #7.1-7.10 (even)
    - OpenStats, Chapter 7: 7.5.2 Fitting a line by least squares regression (p. 362) - #7.24, 7.26, 7.30
    - OpenStats, Chapter 7: 7.5.4 Inference for linear regression (p. 367) - #7.36

Linear regression is a statistical model

Linear regression is a model formulation

Usually (but not always) it is reserved for situations where you assert evidence of causation (e.g. A causes B)

Correlation, in contrast, describes relationships (e.g. A and B are positively correlated)

Linear correlation coefficient

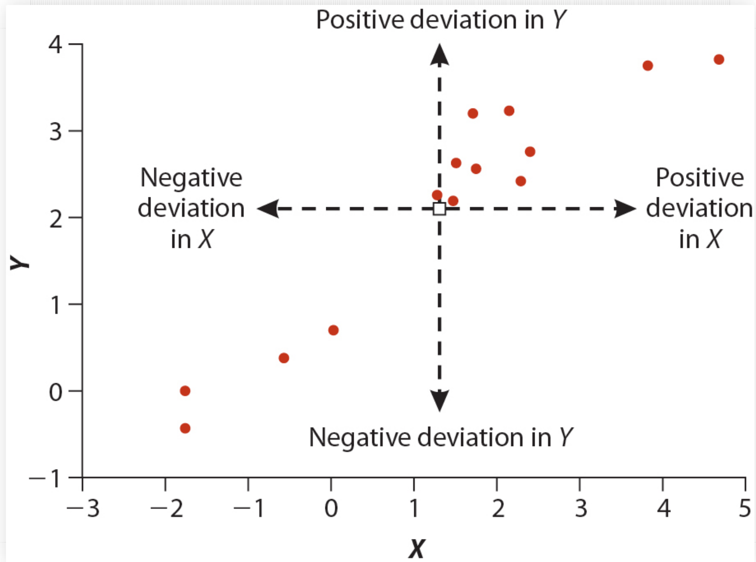
Variables: For a correlation, our data consist of two numerical variables (continuous or discrete).

Definition: The (linear) ***correlation coefficient*** ρ measures the strength and direction of the association between two numerical variables in a population.

The linear (Pearson) correlation coefficient measures the tendency of two numerical variables to **co-vary** *in a linear way*.

The symbol r denotes a sample estimate of ρ.

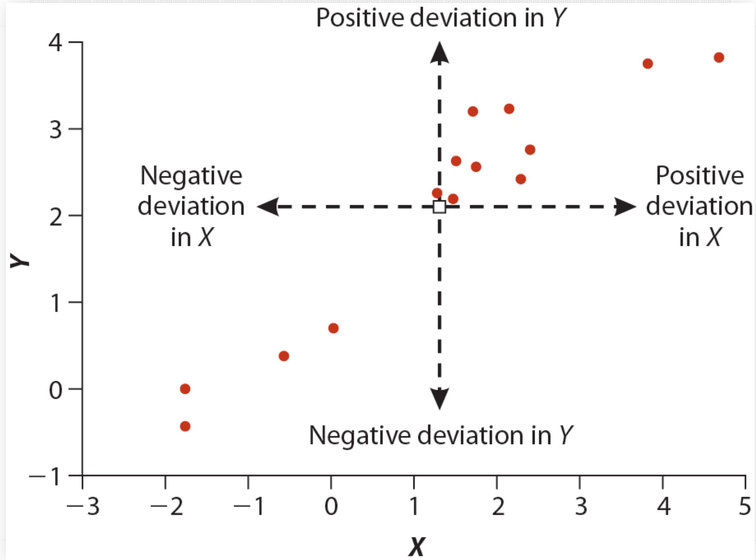
Sample correlation coefficient



$$r = \frac{\sum_i (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_i (X_i - \bar{X})^2}\sqrt{\sum_i (Y_i - \bar{Y})^2}}$$

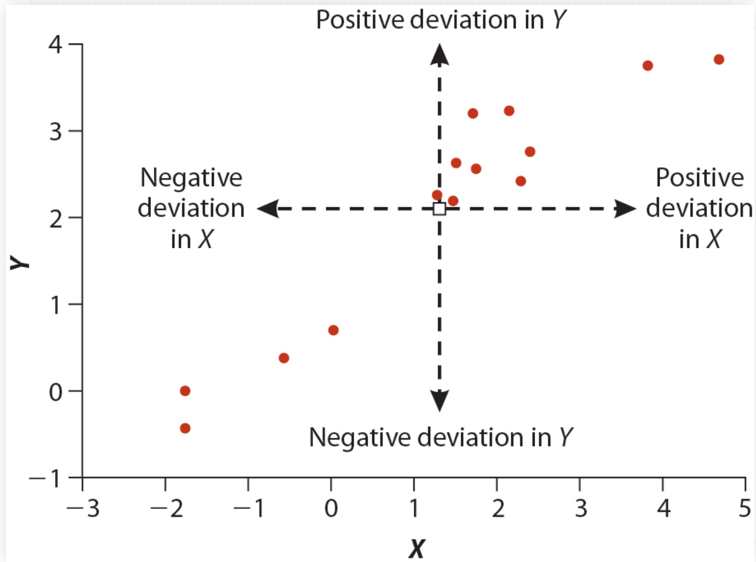
$$-1 \le r \le 1$$

Sample correlation coefficient



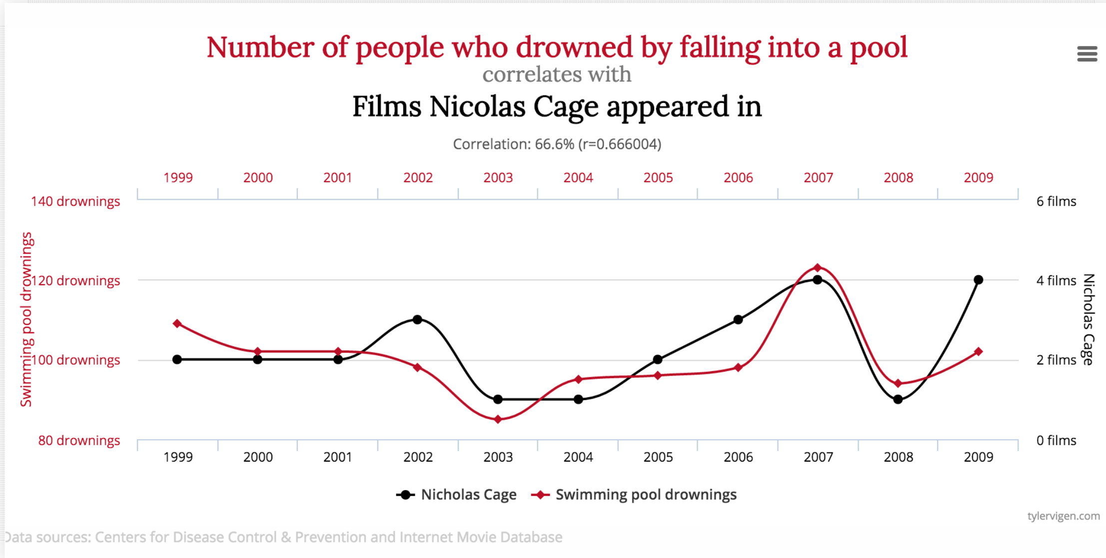
$$r = \frac{\frac{1}{n-1}\sum_i (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\frac{1}{n-1}\sum_i (X_i - \bar{X})^2}\sqrt{\frac{1}{n-1}\sum_i (Y_i - \bar{Y})^2}}$$

Sample correlation coefficient



$$r = \frac{\text{Covariance}(X, Y)}{s_X s_Y}$$

Spurious correlations



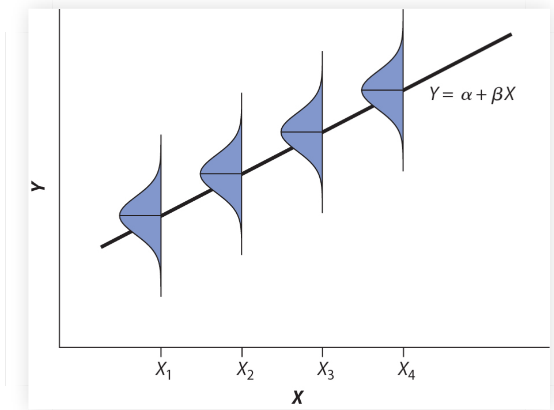
http://www.tylervigen.com/spurious-correlations

Important!

Technically, the linear regression equation is

$$\mu_{Y \mid X=X^*} = \alpha + \beta X^*,$$

were $\mu_{Y \mid X=X^*}$ is the mean of Y in the sub-population with $X = X^*$ (called *predicted values*).

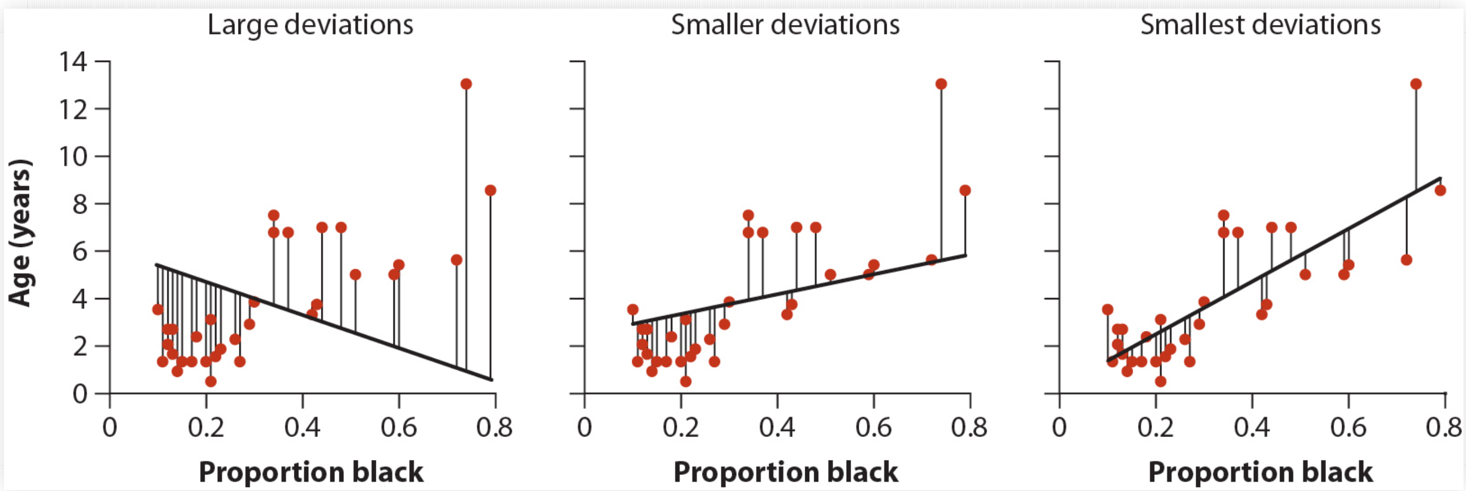


You are predicting the *mean* of Y given X.

How do you find the ”best fit” line?

Method of least squares

Definition: The ***least-squares regression*** line is the line for which the sum of all the ***squared*** deviations in Y is smallest.



How do you find the "best fit" line?

The method of least-squares leads to the following estimates for intercept and slope:

$$b = \frac{\sum_i (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_i (X_i - \bar{X})^2}$$

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

Note:

$$b = \frac{\text{Covariance(X, Y)}}{s_X^2} = r\,\frac{s_Y}{s_X}\,,$$

where r is the correlation coefficient!

Example: Biting lizards

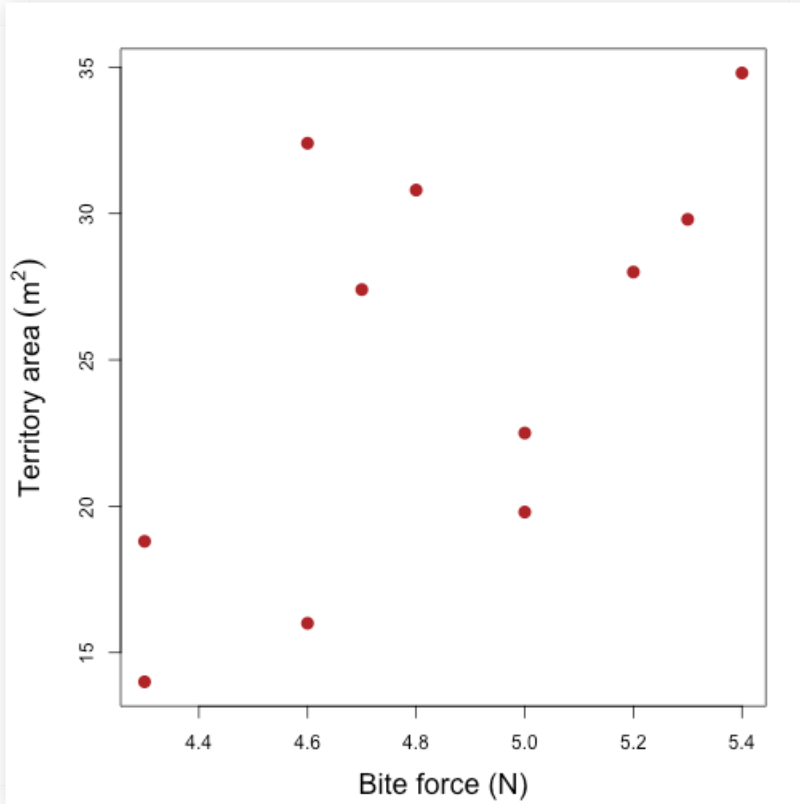
Example: Biting lizards

Example: Biting lizards

Example: Biting lizards

Male lizards in the species *Crotaphytus collaris* use their jaws as weapons during territorial interactions. Lappin and Husak (2005) tested whether weapon performance (bite force) predicted territory size in this species.

Example: Biting lizards



Example: Biting lizards

Compute best-fit line: Slope

$$b = \frac{\text{Covariance(X, Y)}}{s_X^2}$$

```
# Slope
(b <- cov(biteData$bite,
biteData$territory.area)/var(biteData$bite))
```

```
[1] 11.6773
```

Example: Biting lizards

Compute best-fit line: Intercept

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

```
# Intercept
(a <- mean(biteData$territory.area) -
b*mean(biteData$bite))
```

```
[1] -31.53929
```

Example: Biting lizards

Faster!!! Use `lm`…

```
(biteRegression <- lm(territory.area ~ bite,
data = biteData))
```

```


Call:
lm(formula = territory.area ~ bite, data =
biteData)

Coefficients:
(Intercept)          bite
     -31.54         11.68
```

Example: Biting lizards

Bonus!!! With `lm`, can add best-fit line to plot.

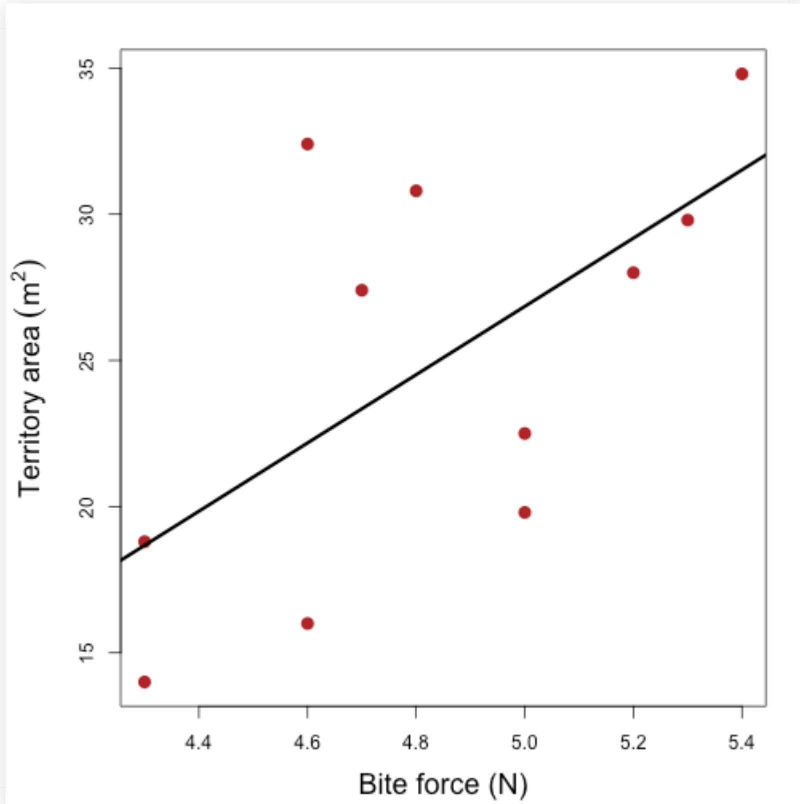
```
# Need to adjust margins to see axis labels
par(mar=c(4.5,5.0,2,2))

# Scatter plot
plot(biteData, pch=16, col="firebrick",
cex=1.5, cex.lab=1.5, xlab="Bite force (N)",
ylab=expression("Territory area" ~ (m^2)))

# Add in the best-fit line
abline(biteRegression, lwd=3)
```

Example: Biting lizards

Bonus!!! With `lm`, can add best-fit line to plot.



Predicted values and residuals

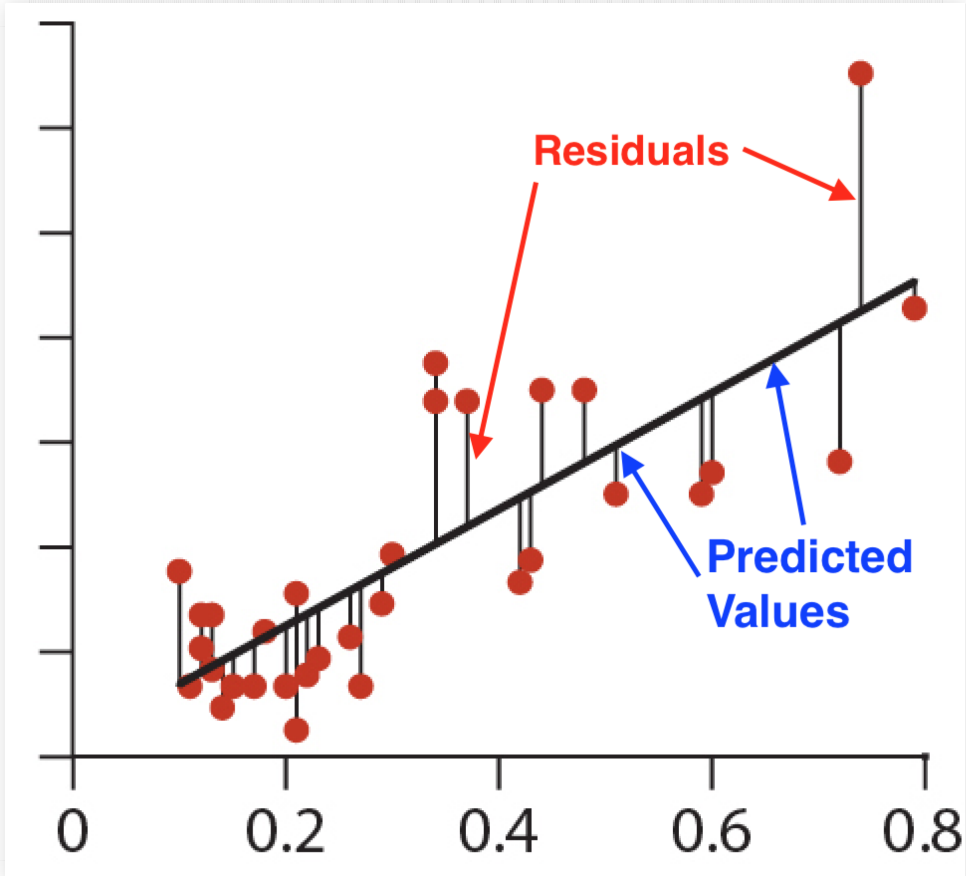
Definition: The ***predicted value*** of Y (denoted $\hat{Y}$, or $\mu_{Y \mid X}$) from a regression line estimates the mean value of Y for all individuals having a given value of X.

Definition: ***Residuals*** measure the scatter of points above and below the least-squares regression line, and are denoted by

$$r_i = \hat{Y}_i - Y_i,$$

where $\hat{Y}_i = a + bX_i$.

Predicted values and residuals



Prediction values

We can predict what the mean value of Y is for values of the explanatory variable X not represented in our data, *as long as we are within the range of values of the data.*

The function `predict` accomplishes this, and even gives us a standard error for our estimate.

```
(pred_5.1 <- predict(biteRegression,
data.frame(bite = 5.1), se.fit = TRUE))
```

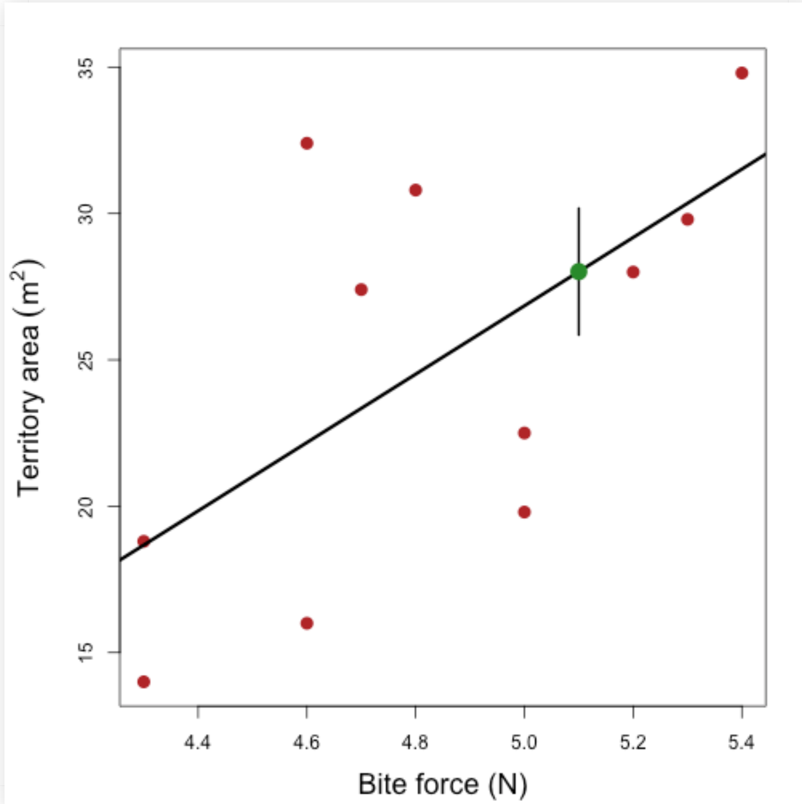
```
$fit
       1
28.01492

$se.fit
[1] 2.163259

$df
[1] 9

$residual.scale
[1] 5.788413
```

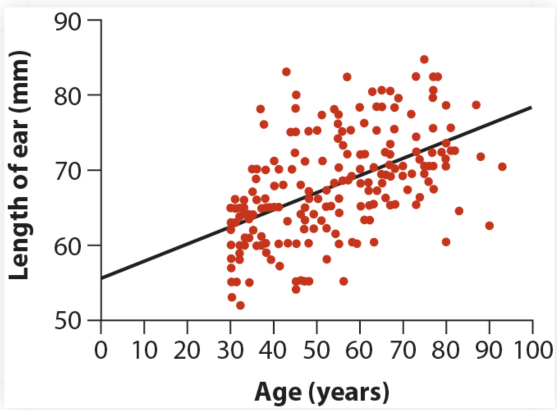
Prediction values



Prediction values - Extrapolation

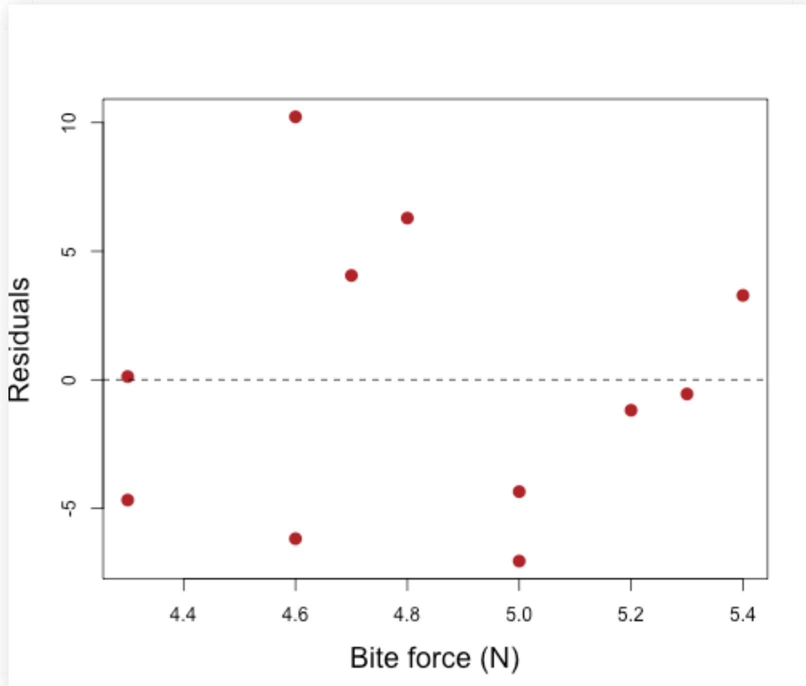
Definition: ***Extrapolation*** is the prediction of the value of a response variable outside the range of X-values in the data.

Regression should not be used to predict the value of the response values for an X-value that lies well outside the range of the data.



Residual plot

Definition: a ***residual plot*** is a scatter plot of the residuals $(\hat{Y}_i - Y_i)$ against the X_i, the values of the explanatory variable.



Residual plots

```
# Get residuals from regression output
biteData$res = resid(biteRegression)

# Plot residuals
plot(res ~ bite, data=biteData, pch=16,
cex=1.5, cex.lab=1.5, col="firebrick",
xlab="Bite force (N)", ylab="Residuals")

# Add a horizontal line at zero
abline(h=0, lty=2)
```

Residual plots to check assumptions

Linear regression

```
summary(biteRegression)
```

Linear regression

```
Call:
lm(formula = territory.area ~ bite, data =
biteData)

Residuals:
    Min      1Q  Median      3Q     Max
-7.0472 -4.5101 -0.5504  3.6689 10.2237

Coefficients:
            Estimate Std. Error t value
Pr(>|t|)
(Intercept)  -31.539     23.513  -1.341
0.2127
bite          11.677      4.848   2.409
0.0393 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*'
0.05 '.' 0.1 ' ' 1

Residual standard error: 5.788 on 9 degrees of
freedom
Multiple R-squared:  0.3919,    Adjusted
R-squared:  0.3244
```

```
F-statistic: 5.801 on 1 and 9 DF,  p-value:
0.03934
```

P-value is less than 0.05, so we can reject the null hypothesis that the slope $\beta = 0$.

Variation explained by explanatory variable

We can measure how well the line "fits" the data by estimating the R^2 value, i.e.

$$R^2 = \frac{\sigma^2_{\text{regression}}}{\sigma^2_{\text{response}}} = \frac{\sigma^2_{\text{response}} - \sigma^2_{\text{residual}}}{\sigma^2_{\text{response}}}.$$

This also can be said to measure the fraction of variation in Y that is "explained" by X.

Variation explained by explanatory variable

Basic idea is:

- If R^2 is close to 1, then X is explaining most of the variation in Y, and any other variation which could be caused by other sources is negligible in comparison.
- If R^2 is close to 0, then X is explaining very little of the variation in Y, and the remaining variation is caused by other sources not accounted for or measured in the system of study.

Variation explained by explanatory variable

For the lizard example,

```
biteRegSummary <- summary(biteRegression)
biteRegSummary$r.squared
```

```
[1] 0.3919418
```

Thus, 39% of the variation in territory area is explained by bite force.

Annotate plot

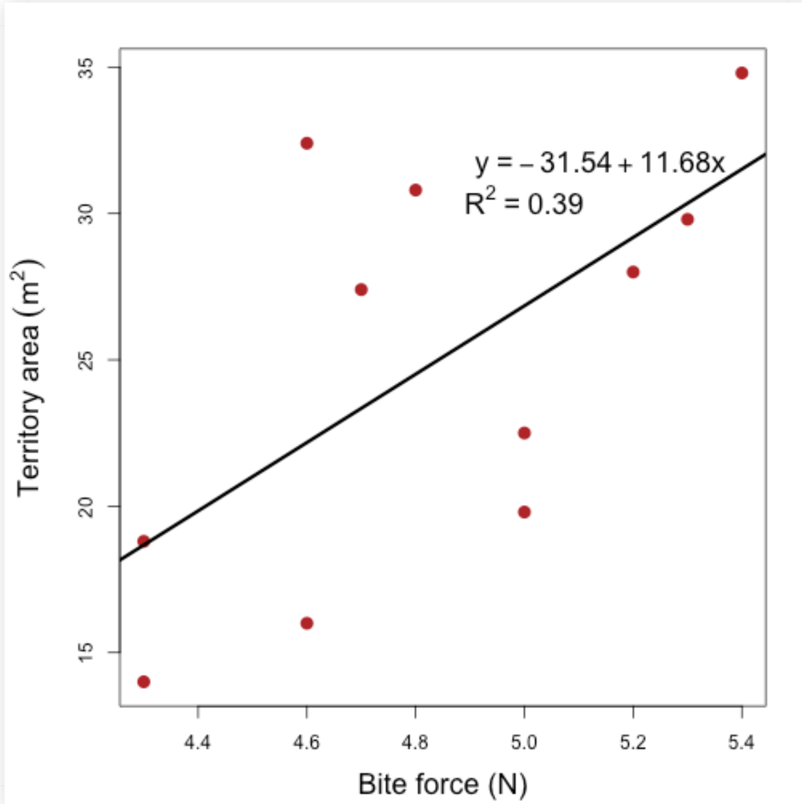
```
# Need to adjust margins to see axis labels
par(mar=c(4.5,5.0,2,2))

# Scatter plot
plot(territory.area ~ bite, pch=16,
col="firebrick", cex=1.5, cex.lab=1.5,
xlab="Bite force (N)",
ylab=expression("Territory area" ~ (m^2)),
data=biteData)

# Add in the best-fit line
abline(biteRegression, lwd=3)

# Text
text(5, 30.5, expression(R^{2} ~ "=" ~ 0.39),
cex=1.5)
text(5.14, 31.7, expression(y ~ "=" -31.54 +
11.68), cex=1.5)
```

Annotate plot



Summary of a regression in R

```
summary(biteRegression)
```

Summary of a regression in R

```
Call:
lm(formula = territory.area ~ bite, data =
biteData)

Residuals:
    Min      1Q  Median      3Q     Max
-7.0472 -4.5101 -0.5504  3.6689 10.2237

Coefficients:
            Estimate Std. Error t value
Pr(>|t|)
(Intercept)  -31.539     23.513  -1.341
0.2127
bite          11.677      4.848   2.409
0.0393 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*'
0.05 '.' 0.1 ' ' 1

Residual standard error: 5.788 on 9 degrees of
freedom
Multiple R-squared:  0.3919,    Adjusted
R-squared:  0.3244
```

```
F-statistic: 5.801 on 1 and 9 DF,  p-value:
0.03934
```