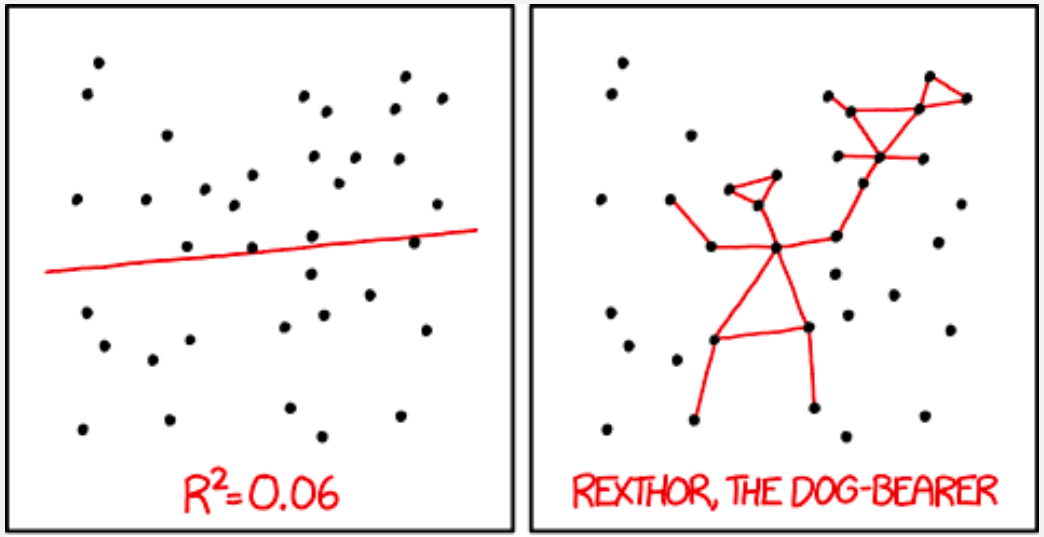
Regression

Biology 683

Lecture 6

Heath Blackmon



Last week

1. Midterm

2. P-values
3. Bayes Theorem
4. Reproducibility Crisis

Today

1) Simple Linear Regression
2) General Linear Models

Correlation vs Regression

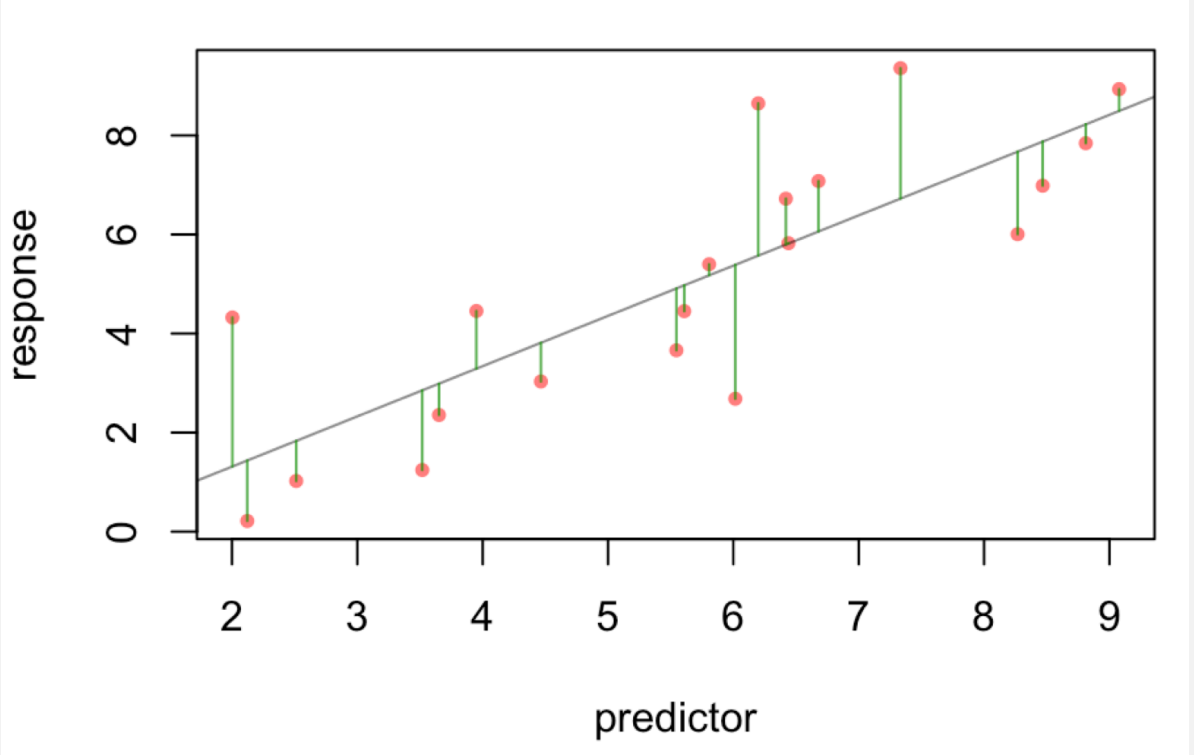
- Both methods are ways to explore contingency between variables.
- Regression describes the degree to which we can predict the value of one variable based on the value of another.
- Regression calculates a line that describes this relationship between a response and explanatory variable.
- Use regression when you believe there is a strong case for causation.

Regression in R

1) With linear regression we find the linear equation that best predicts the values of Y based on the values of X.

2) $y = bx + a$

3) Least-squares regression minimizes the squared deviations of the data points from that line.



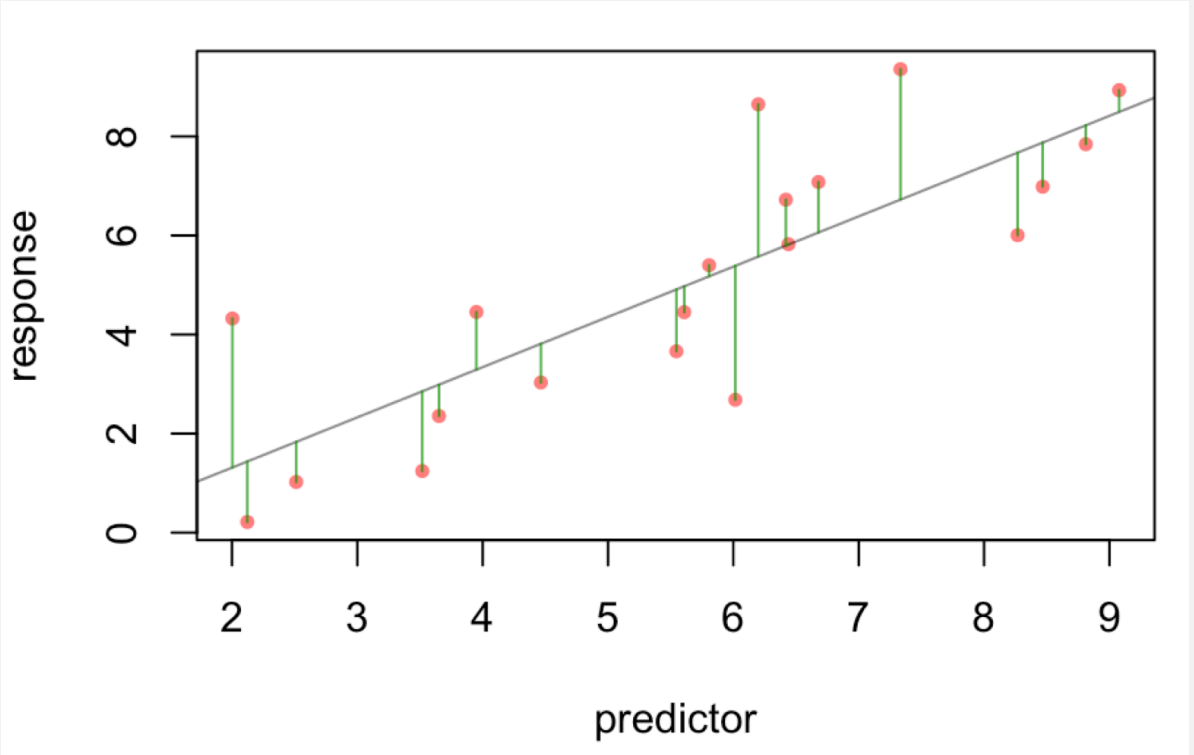
Example of regression

```
set.seed(3)
x <- runif(min = 1, max = 10, 20)
y <- rnorm(20, mean = x, sd = 2)
fit.xy <- lm(y ~ x)
summary(fit.xy)
```

$$y = bx + a$$

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

```
Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-2.7060 -0.9742 -0.4539  0.9479  3.0728

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.7173     1.0302  -0.696    0.495
x             1.0150     0.1708   5.943 1.27e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.625 on 18 degrees of freedom
Multiple R-squared:  0.6624,    Adjusted R-squared:  0.6437
F-statistic: 35.32 on 1 and 18 DF,  p-value: 1.267e-05
```



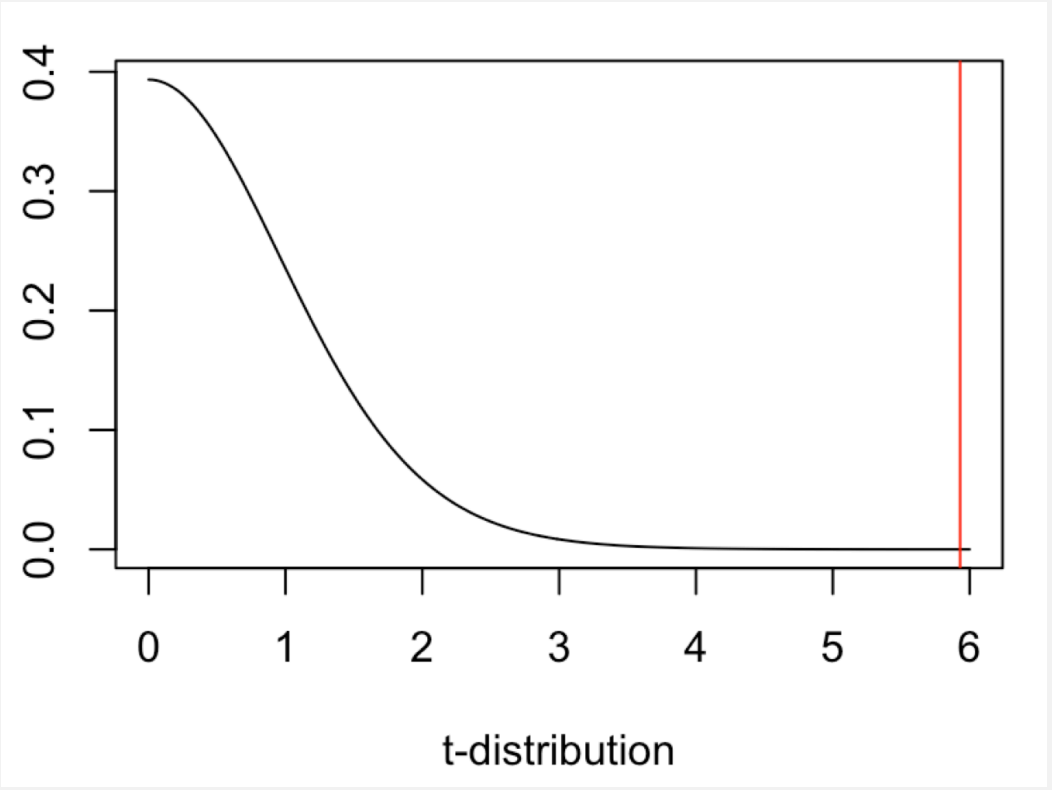
Example of regression

```
set.seed(3)
x <- runif(min = 1, max = 10, 20)
y <- rnorm(20, mean = x, sd = 2)
fit.xy <- lm(y ~ x)
summary(fit.xy)
```

$$y = bx + a$$

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

```
Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-2.7060 -0.9742 -0.4539  0.9479  3.0728

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.7173     1.0302  -0.696    0.495
x             1.0150     0.1708   5.943 1.27e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.625 on 18 degrees of freedom
Multiple R-squared:  0.6624,    Adjusted R-squared:  0.6437
F-statistic: 35.32 on 1 and 18 DF,  p-value: 1.267e-05
```

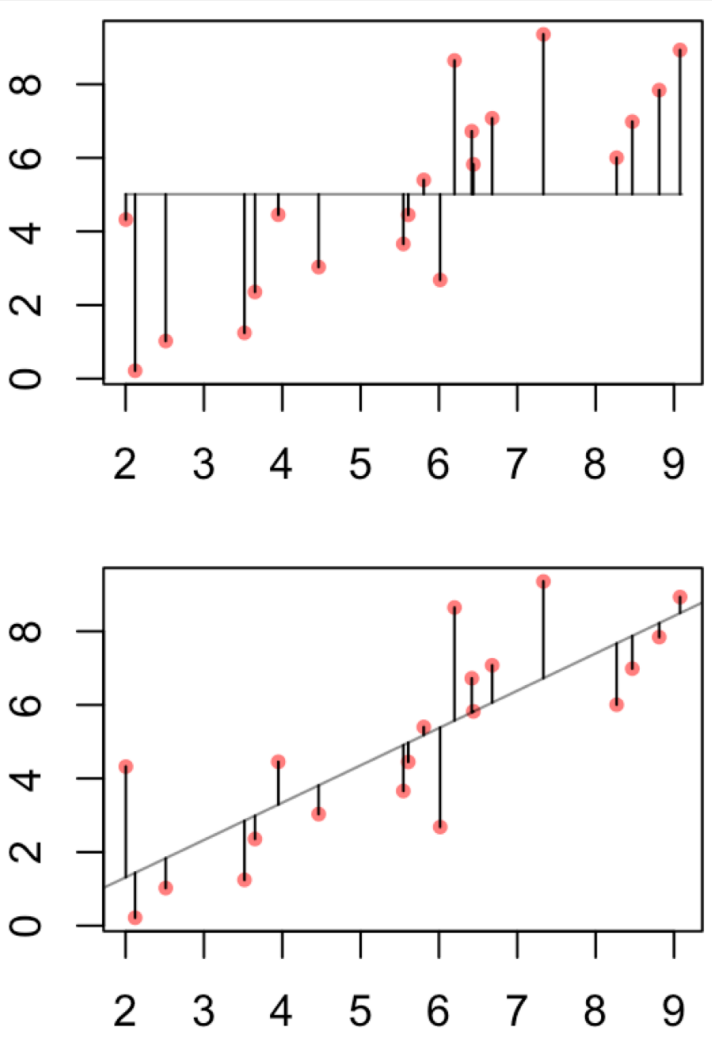


Example of regression

```
set.seed(3)
x <- runif(min = 1, max = 10, 20)
y <- rnorm(20, mean = x, sd = 2)
fit.xy <- lm(y ~ x)
summary(fit.xy)
```

This can help to justify the biological importance assuming you have a regression that is significant. It is the proportion of total variance explained by the regression.

```
Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-2.7060 -0.9742 -0.4539  0.9479  3.0728

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.7173     1.0302  -0.696    0.495
x             1.0150     0.1708   5.943 1.27e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.625 on 18 degrees of freedom
Multiple R-squared:  0.6624,    Adjusted R-squared:  0.6437
F-statistic: 35.32 on 1 and 18 DF,  p-value: 1.267e-05
```



Linear regression uses

- Depict the relationship between two variables in an eye-catching fashion
- Test the null hypothesis of no association between two variables
  - The test is whether or not the slope is zero
- Predict the average value of variable *Y* for a group of individuals with a given value of variable *X*
  - variation around the line can make it very difficult to predict a value for a given individual with much confidence
  - Predictions outside of the range of observed data is generally discouraged
- Used both for experimental and observational studies

What are Residuals

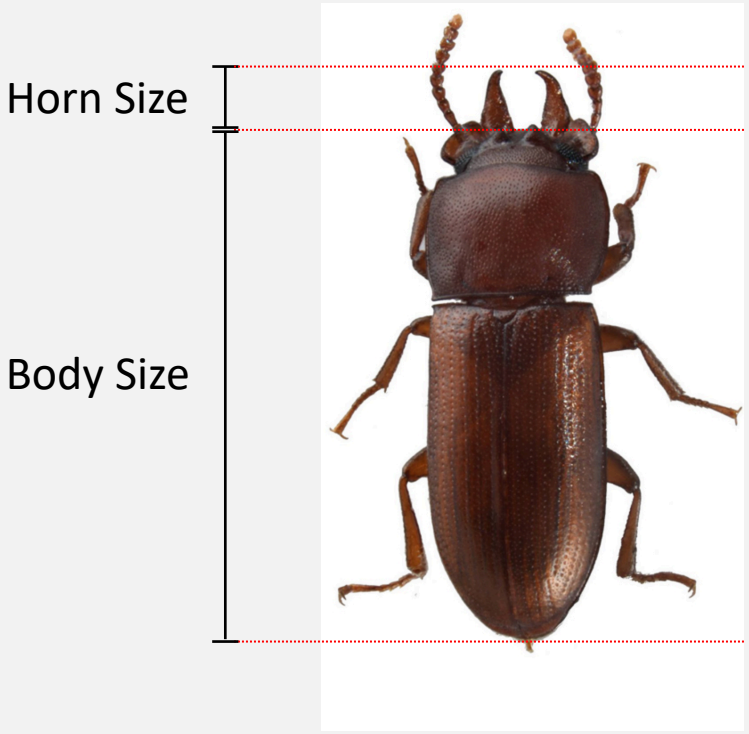
In general, the residual is the individual's departure from the value predicted by the model

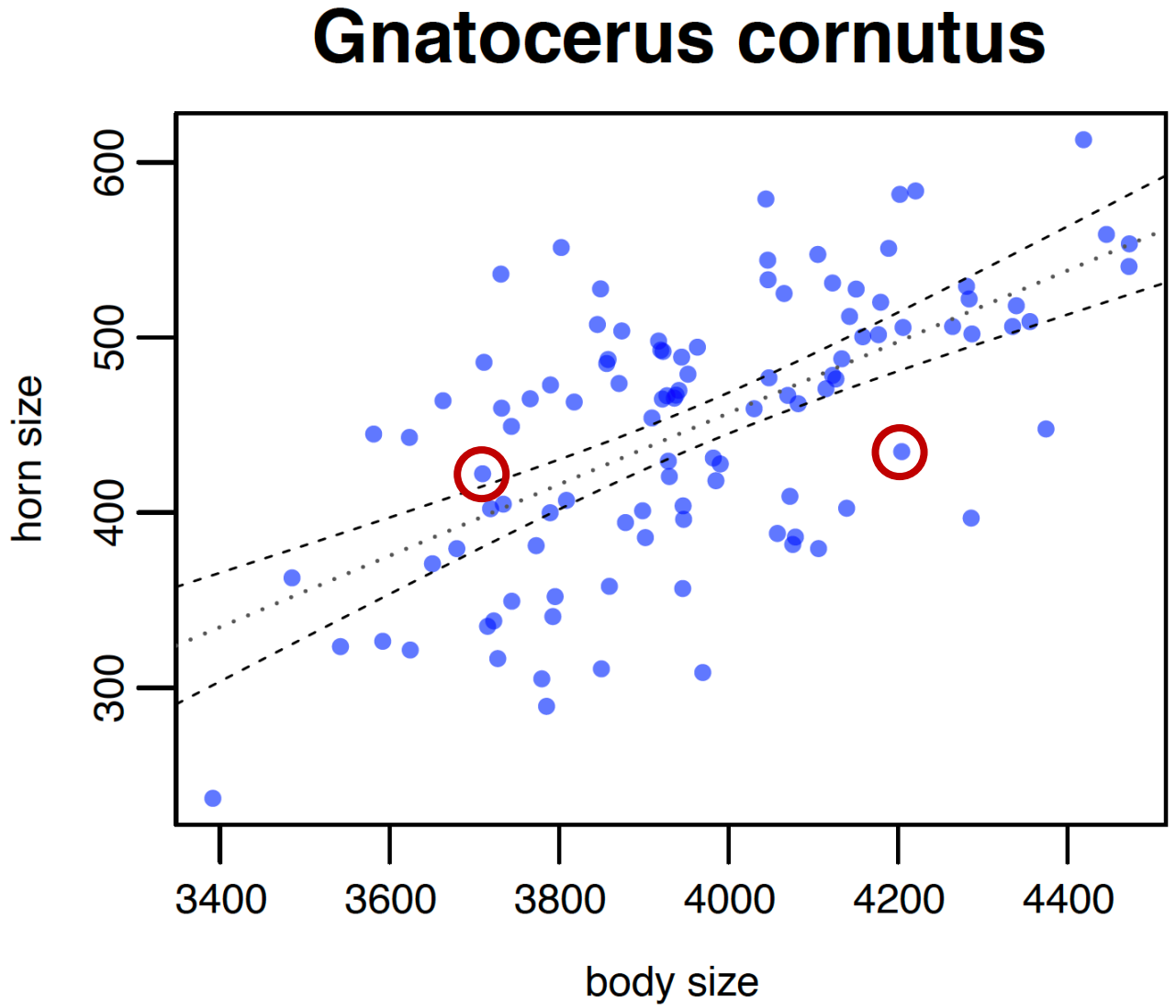
In this case the model is simple – the linear regression – but residuals also exist for more complex models

For a model that fits better, the residuals will be smaller on average

Residuals can be of interest in their own right, because they represent values that have been ***corrected*** for relationships that might be obscuring a pattern.

What are Residuals





Making that plot

```
gnat <- read.csv("../hw-labs/data/gnatocerus.csv")
fit <- lm(horns ~ body, data = gnat)
plot(gnat$horns ~ gnat$body,
     xlab = "body size",
     cex.lab = .7, cex.axis = .7,
     ylab = "horn size",
     main = "Gnatocerus cornutus",
     pch = 16, cex = .6, col = rgb(0, 0, 1, .6))

#Add the regression line
abline(fit, lty=3, col="gray35")

#Add confidence limits for the regression line
xpt <- seq(par("usr")[1], par("usr")[2])
ypt <- data.frame(predict(fit,
                          newdata = data.frame(body = xpt),
                          interval = "confidence",
                          level = 0.95,
                          type = "response"))
lines(ypt$lwr ~ xpt, lwd = .6, lty = 2)
lines(ypt$upr ~ xpt, lwd = .6, lty = 2)
```

Strong Inference for Observational Studies

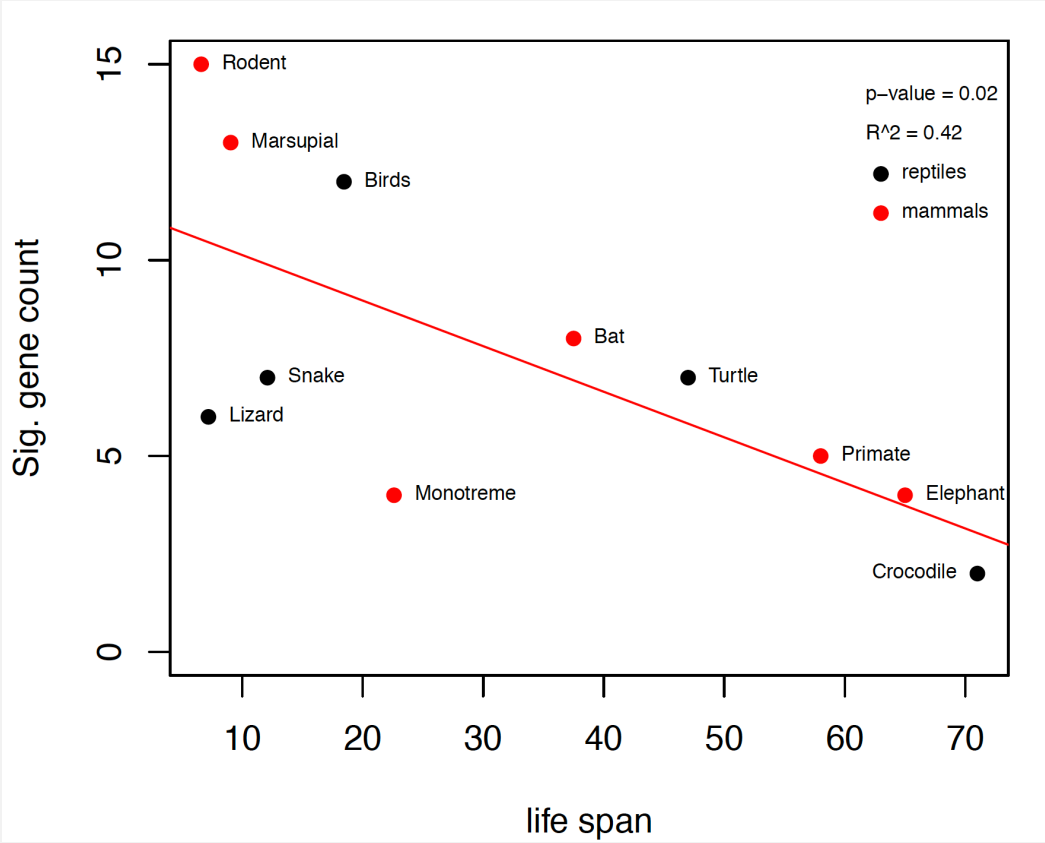
- Noticing a pattern in the data and reporting it represents a post hoc analysis
- This is not hypothesis testing
- The results, while potentially important, must be interpreted cautiously

What can be done?

- Based on a post-hoc observational study, construct a new hypothesis for a novel group or system that has not yet been studied

Example

1) We already knew that the P53 network is important in guarding against cancer in long lived species.
2) We also knew that primates and elephants show rather little change in this network when compared to rodents.
3) Collect data on many more species and test apriori hypothesis that there will be a significant and negative regression coefficient.



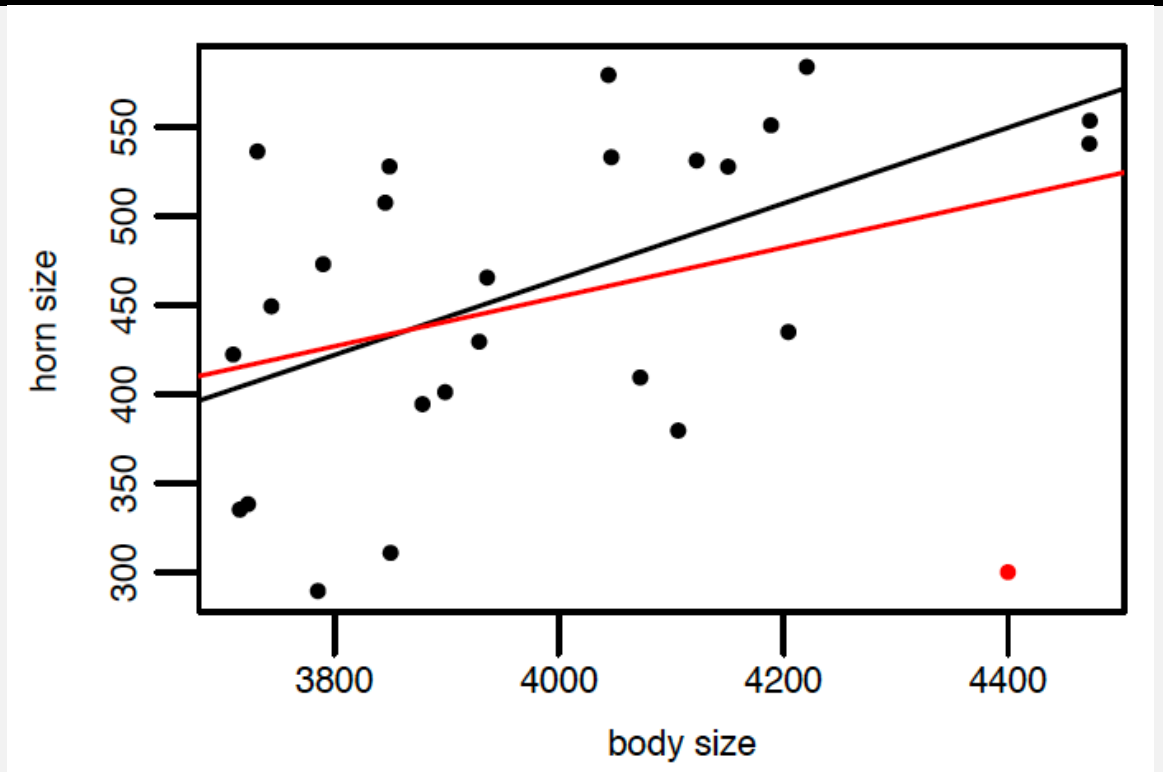
Assumptions of Linear Regression

- The true relationship must be linear
- At each value of *X*, the distribution of *Y* is normal (i.e., the residuals are normal)
- The variance in *Y* is independent of the value of *X*
- **Note that there are no assumptions about the distribution of *X***

Common Problems

- Outliers
  - Regression is extremely sensitive to outliers
  - The line will be drawn to outliers, especially along the x-axis
  - Consider performing the regression with and without outliers
- Non-linearity
  - Best way to notice is by visually inspecting the plot and the line fit
  - Try a transformation to get linearity [often a log transformation]
- Non-normality of residuals
  - Can be detected from a residual plot
  - Possibly solved with a transformation
- Unequal variance
  - Usually visible from a scatterplot or from a residual plot

Outliers



Leverage and cooks distance

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -100.24112  297.38717  -0.337   0.7390
x2             0.13870    0.07431   1.867   0.0742 .
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 86.81 on 24 degrees of freedom
Multiple R-squared:  0.1268,    Adjusted R-squared:  0.09038
F-statistic: 3.484 on 1 and 24 DF,  p-value: 0.07423
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -386.07048  272.48381  -1.417  0.16993
x              0.21264    0.06837   3.110  0.00493 **
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 74.73 on 23 degrees of freedom
Multiple R-squared:  0.296,     Adjusted R-squared:  0.2654
F-statistic: 9.673 on 1 and 23 DF,  p-value: 0.004928
```

Moving past simple models

- The reason ANOVA is so widely used is that it provides a framework to simultaneously test the effects of multiple factors
- ANOVA also makes it possible to detect ***interactions*** among the factors
- ANOVA is a special case of a ***general linear model***
- Linear regression is a special case of a ***general linear model***

GLM and LM function in R

- The GLM and LM function in R takes equations that can be described with the following operators

+	+X include this variable
:	X:Z include the interaction between these variables
*	X*Y include these variables and the interactions between them
^	(X + Z + W)^3 include these variables and all interactions up to three way

R versus the math implied

```
glm(y ~ X + W)
```

$$y_i = \beta_0 + \beta_1 X_i + \beta_2 W_i + \epsilon_i$$

```
glm(y ~ X * W)
```

$$y_i = \beta_0 + \beta_1 X_i + \beta_2 W_i + \beta_3 X_i W_i + \epsilon_i$$

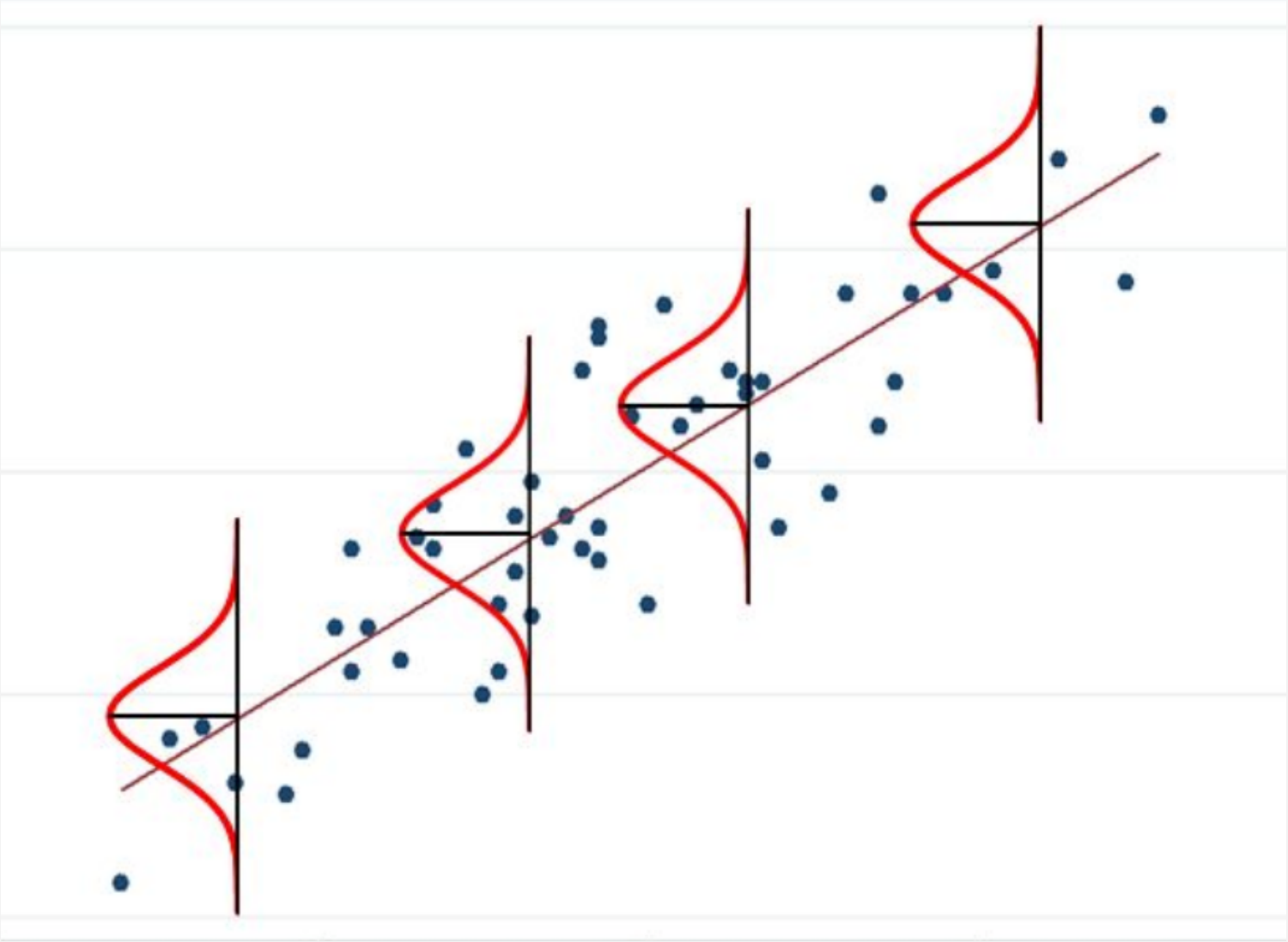
R versus the math oak example

```
Call:
glm(formula = specialist ~ temp * circ, data = oak)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.2804  -1.1295  -0.2256   0.9952   5.6787

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 9.7621149  3.8327598   2.547   0.0114
temp        -0.5574479  0.2527323  -2.206   0.0282
circ        -0.0661544  0.0120692  -5.481 9.40e-08
temp:circ    0.0045895  0.0007887   5.819 1.61e-08
```

circ	temp	precip	specialist
592.0	15.8	257	3
680.0	14.7	455	1
340.0	14.5	458	1
310.0	14.5	458	4
260.0	14.5	458	2

$$y_i = \beta_0 + \beta_1 temp_i + \beta_2 circ_i + \beta_3 temp_i circ_i$$

When the response variable isn’t normal



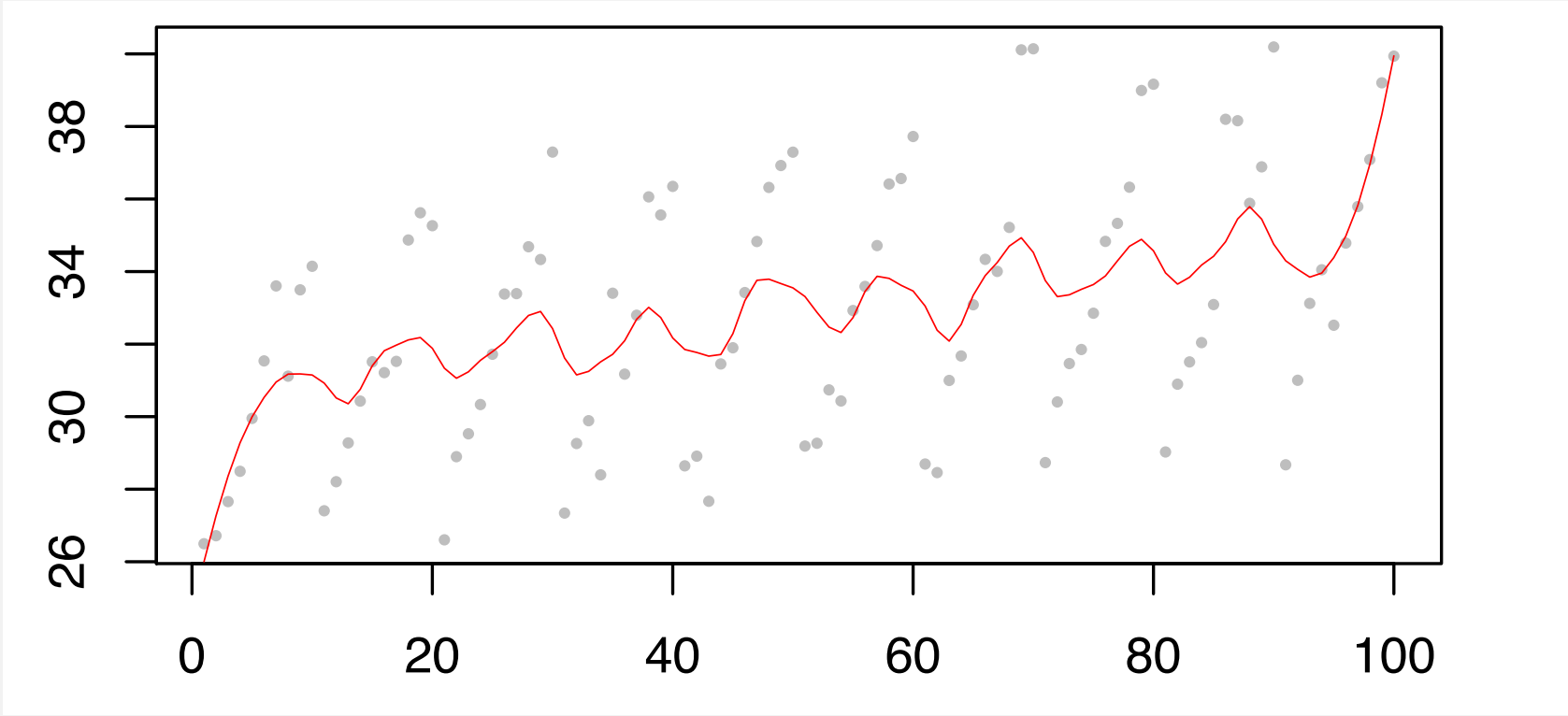
Other kinds of regression

Logistic regression allows us to fit a binary response variable (absent/present; alive/dead) with one or more categorical or continuous predictor variables.

Poisson regression allows us to fit a response variable that is Poisson distributed (number of extinctions in a unit of time, number of colonies per plate, (number of occurrences for rare events)) with one or more categorical or continuous predictor variables.

```
fit.logi <- glm(obs ~ pred2 , family="binomial")

fit.pois <- glm(obs ~ pred2, family="poisson")
```

Sometimes regression isn't best choice




```
lo <- loess(y~x, span=.2)
plot(y~x, pch=16, cex=.5, col="gray")
lines(predict(lo), col='red', lwd=.5)
```

For Thursday

Read chapter WS 17

Bring laptop to class!

Heath Blackmon
BSBW 309
coleoguy@gmail.com
@coleoguy

Models

"Remember that all models are wrong; the practical question is how wrong do they have to be to not be useful."

George Box