# Regression and Models with Multiple Factors

Biology 683

Lecture 6

Heath Blackmon

# Last week

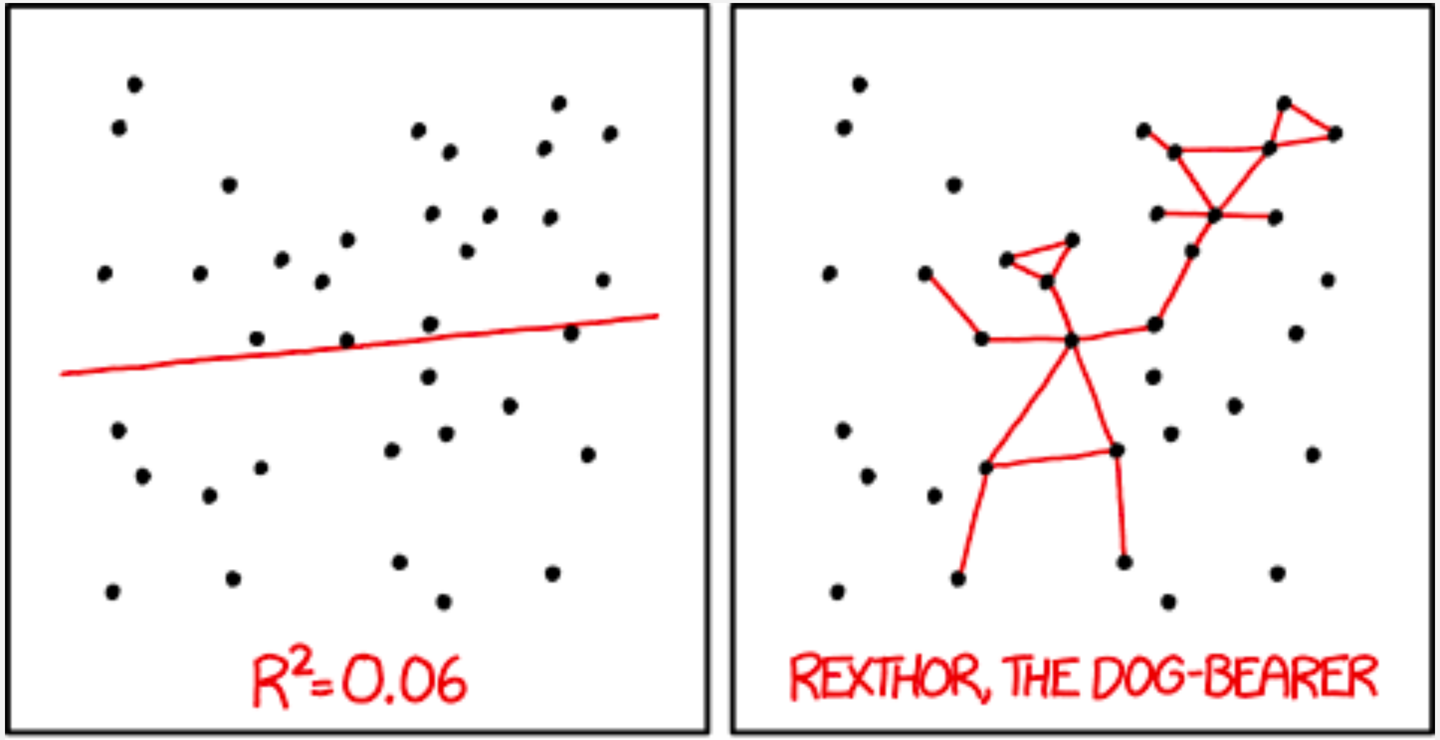
1. Give an example of blocking in your own field.

2. How can you recognize statistical interaction.

3. What are a couple of reasons for doing a power analysis?

# Today

1) Regression
2) Multiple Factors



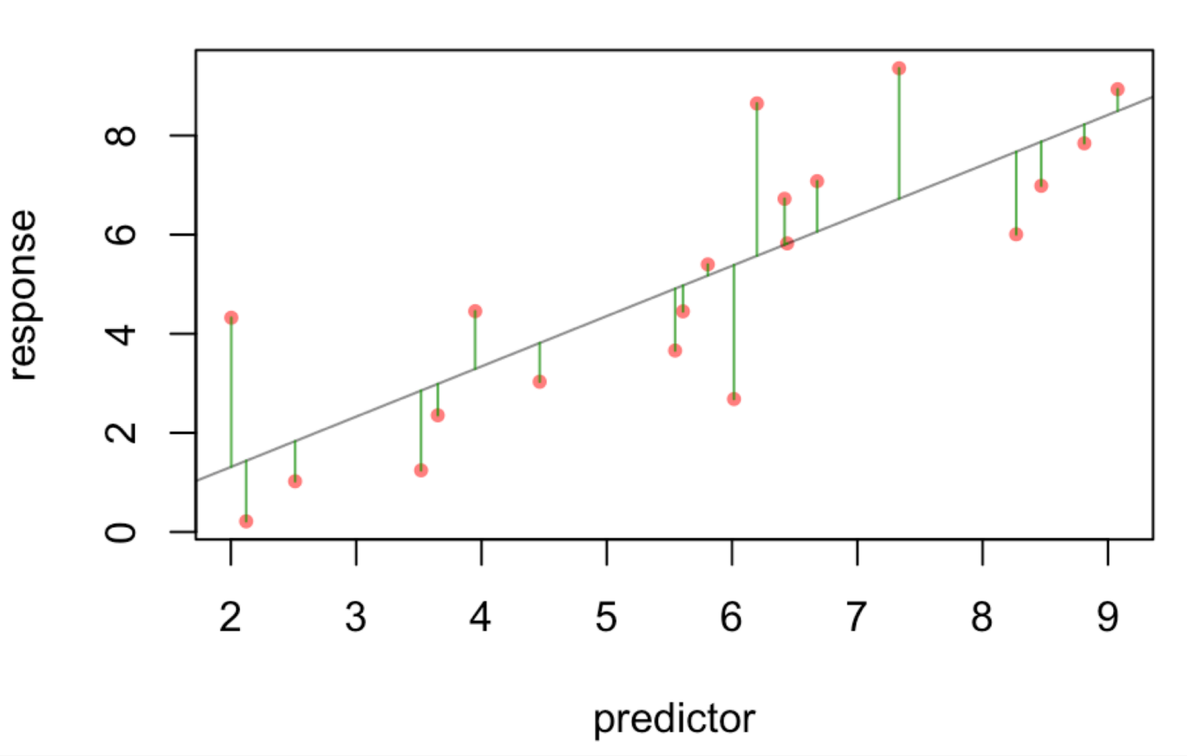
I DON'T TRUST LINEAR REGRESSIONS WHEN IT'S HARDER TO GUESS THE DIRECTION OF THE CORRELATION FROM THE SCATTER PLOT THAN TO FIND NEW CONSTELLATIONS ON IT.

# Regression in R

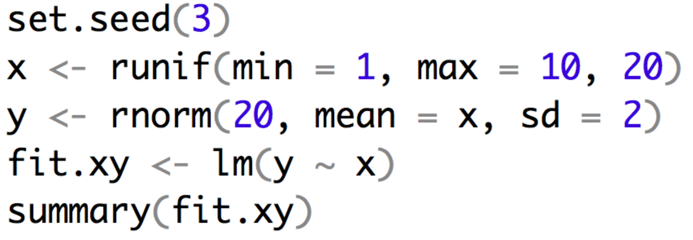
1) With linear regression we find the linear equation that best predicts the values of Y based on the values of X.

2) $y = bx + a$

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



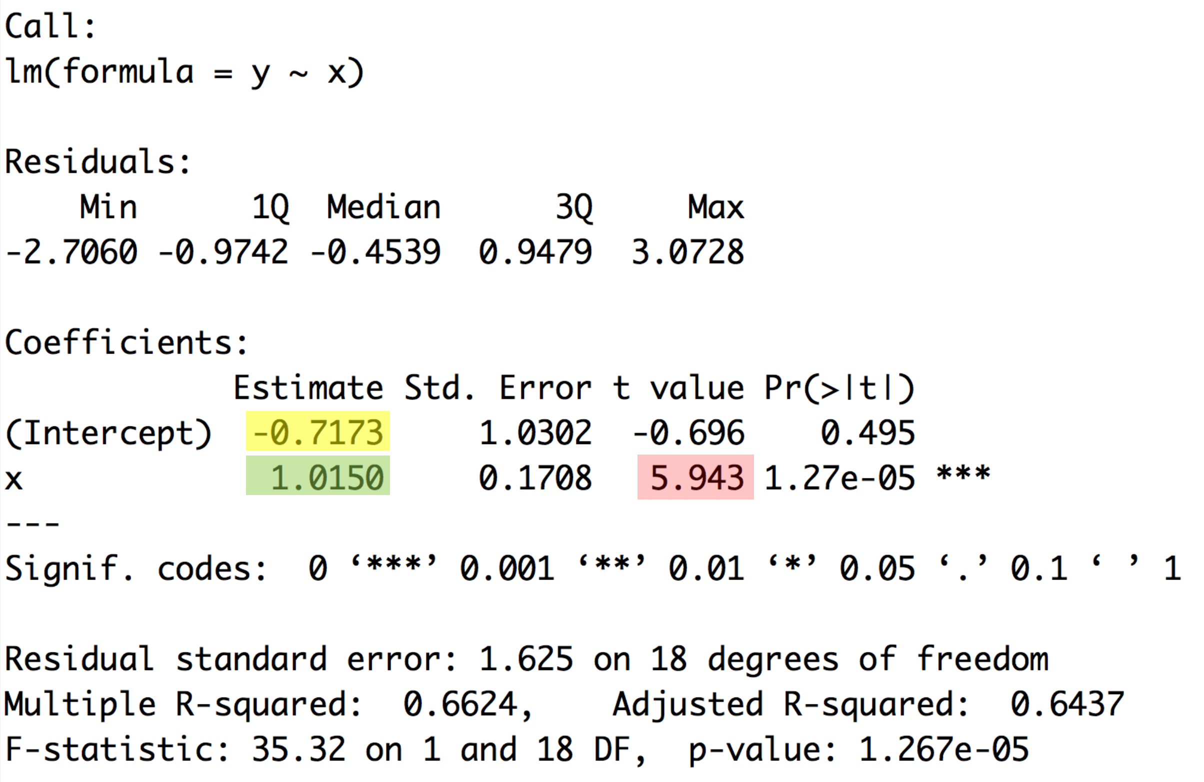
# Example of regression



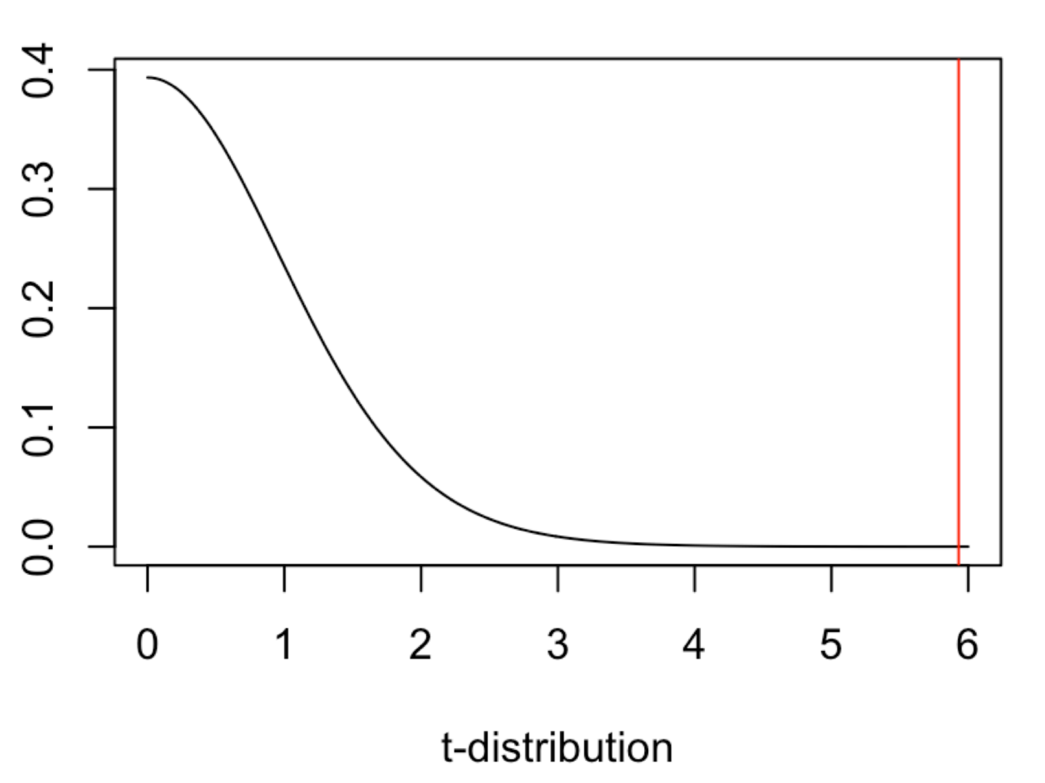
```
set.seed(3)
x <- runif(min = 1, max = 10, 20)
y <- rnorm(20, mean = x, sd = 2)
fit.xy <- lm(y ~ x)
summary(fit.xy)
```

$$y = bx + a$$

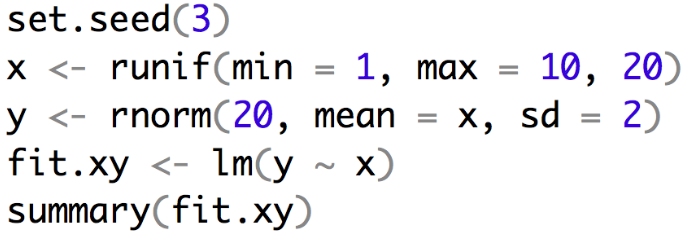
$$t = \frac{b - \beta_0}{SE_b}$$



```
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
```

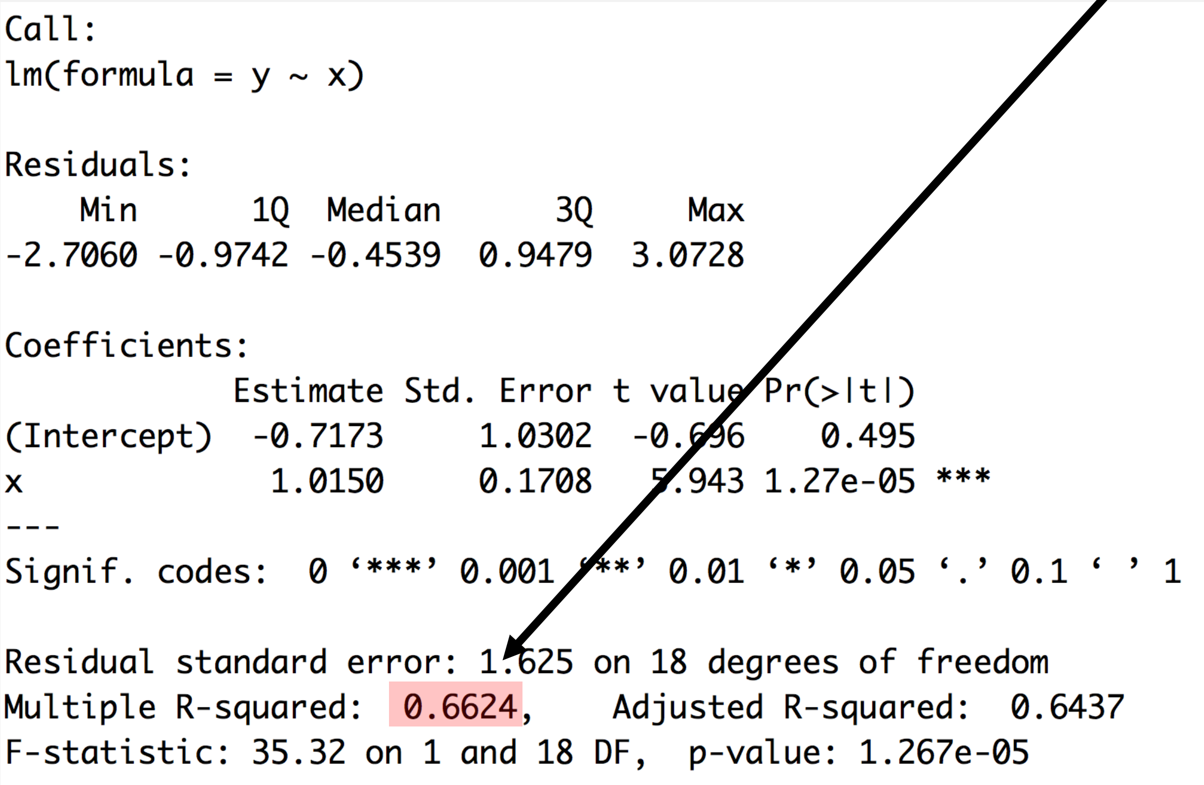


# Example of regression

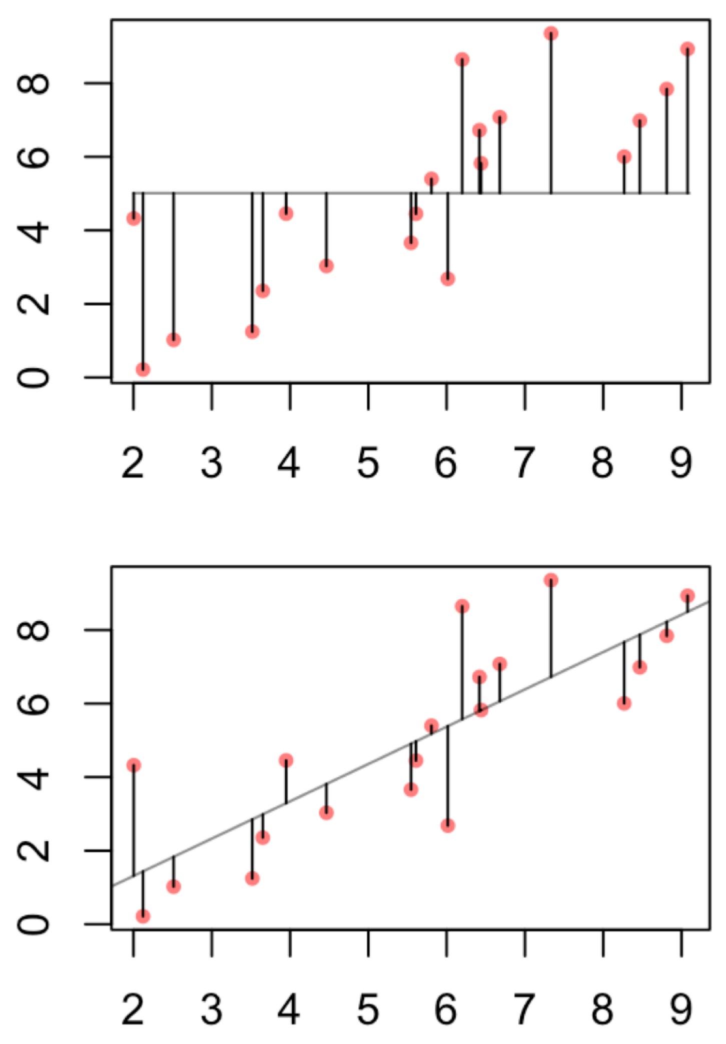


```r
set.seed(3)
x <- runif(min = 1, max = 10, 20)
y <- rnorm(20, mean = x, sd = 2)
fit.xy <- lm(y ~ x)
summary(fit.xy)
```

This can help to justify the biological importance assuming you have a regression that is significant rather than the statistical

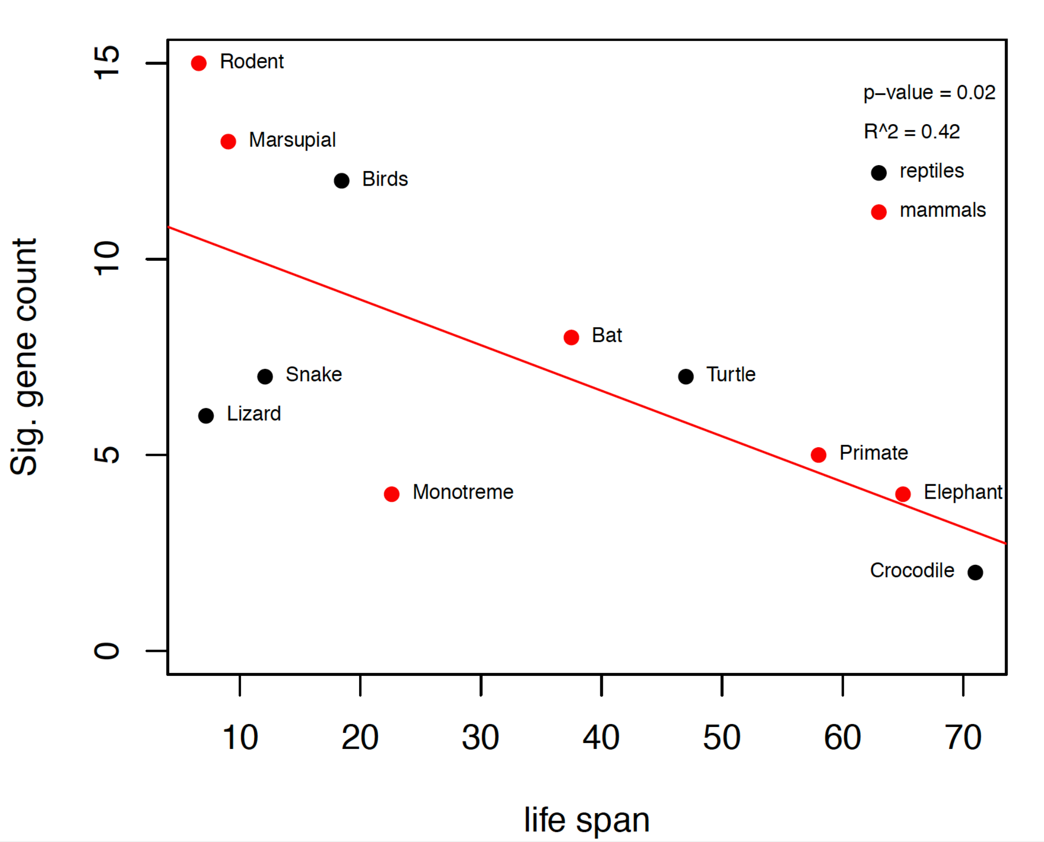


```
Call:
lm(formula = y ~ x)

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

1) We already know that the P53 network is important in guarding against cancer in long lived species.
2) So we might think that short lived species would be more likely to experience selection and change in these genes.
3) What problem might be present with this analysis?



# 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*
    - Note that variation around the line can make it very difficult to predict a value for a given individual with much confidence

# What are Residuals

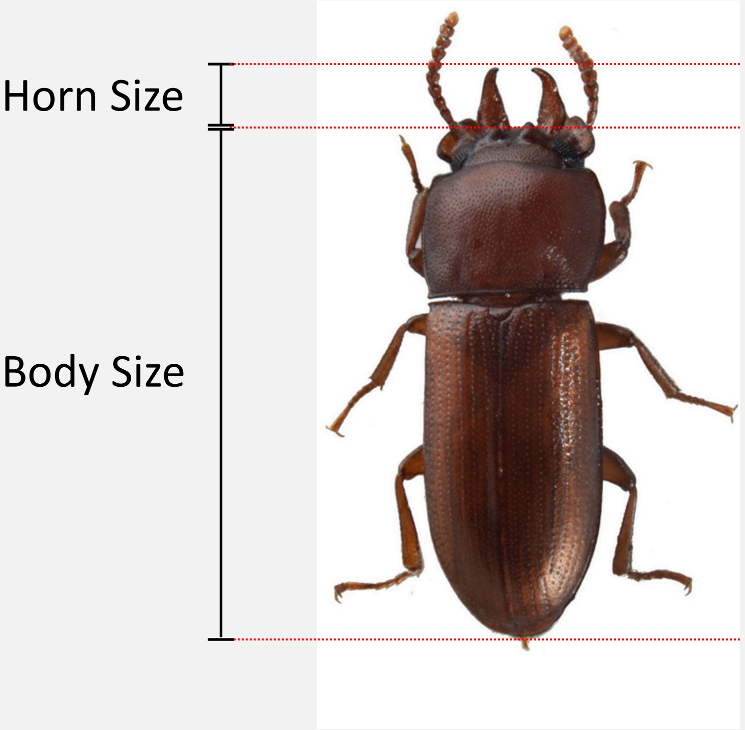
In general, the residual is the individual's departure from the value predicted by the model

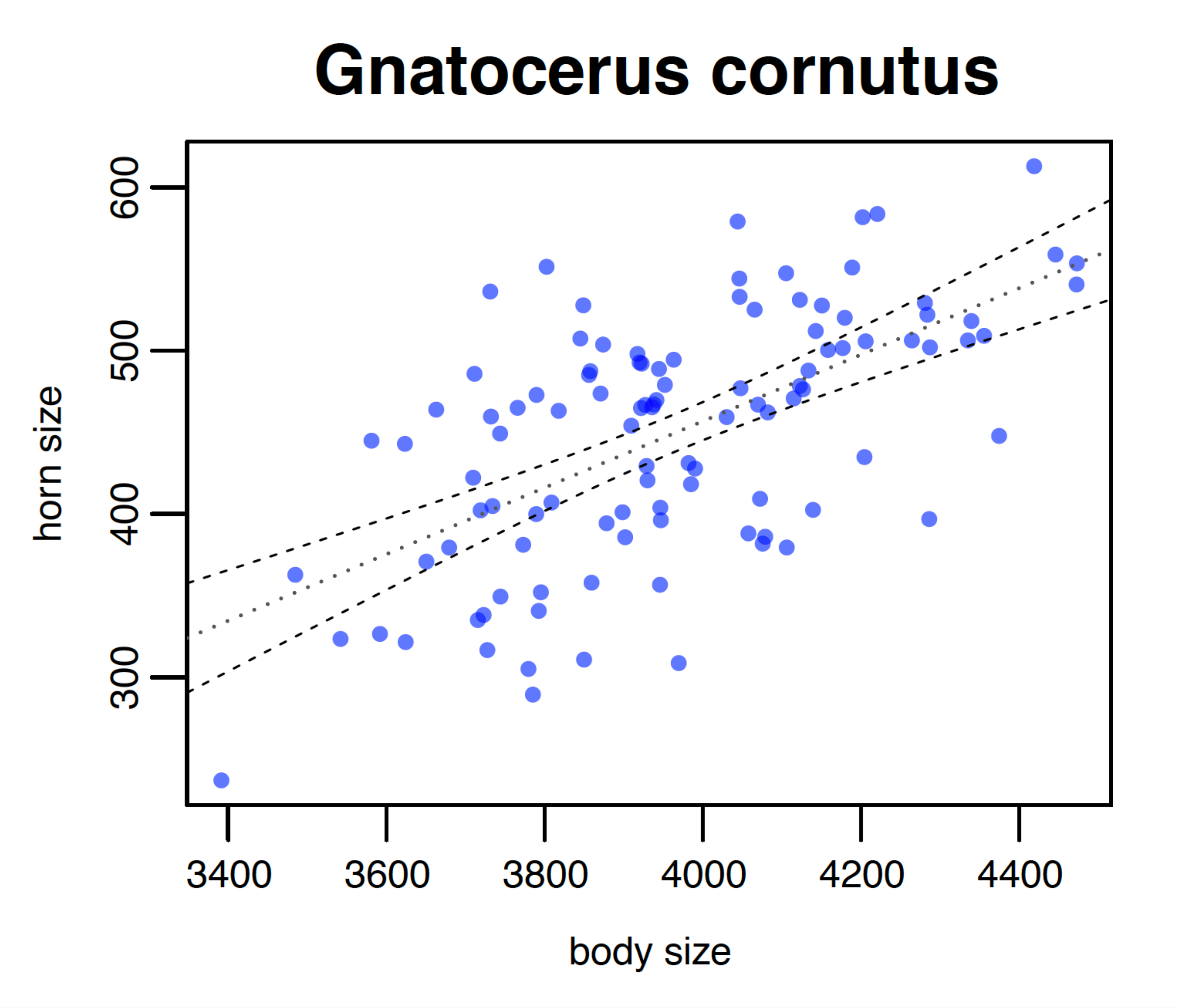
In this case the model is simple – the linear regression – but residuals also exist for more complex models

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

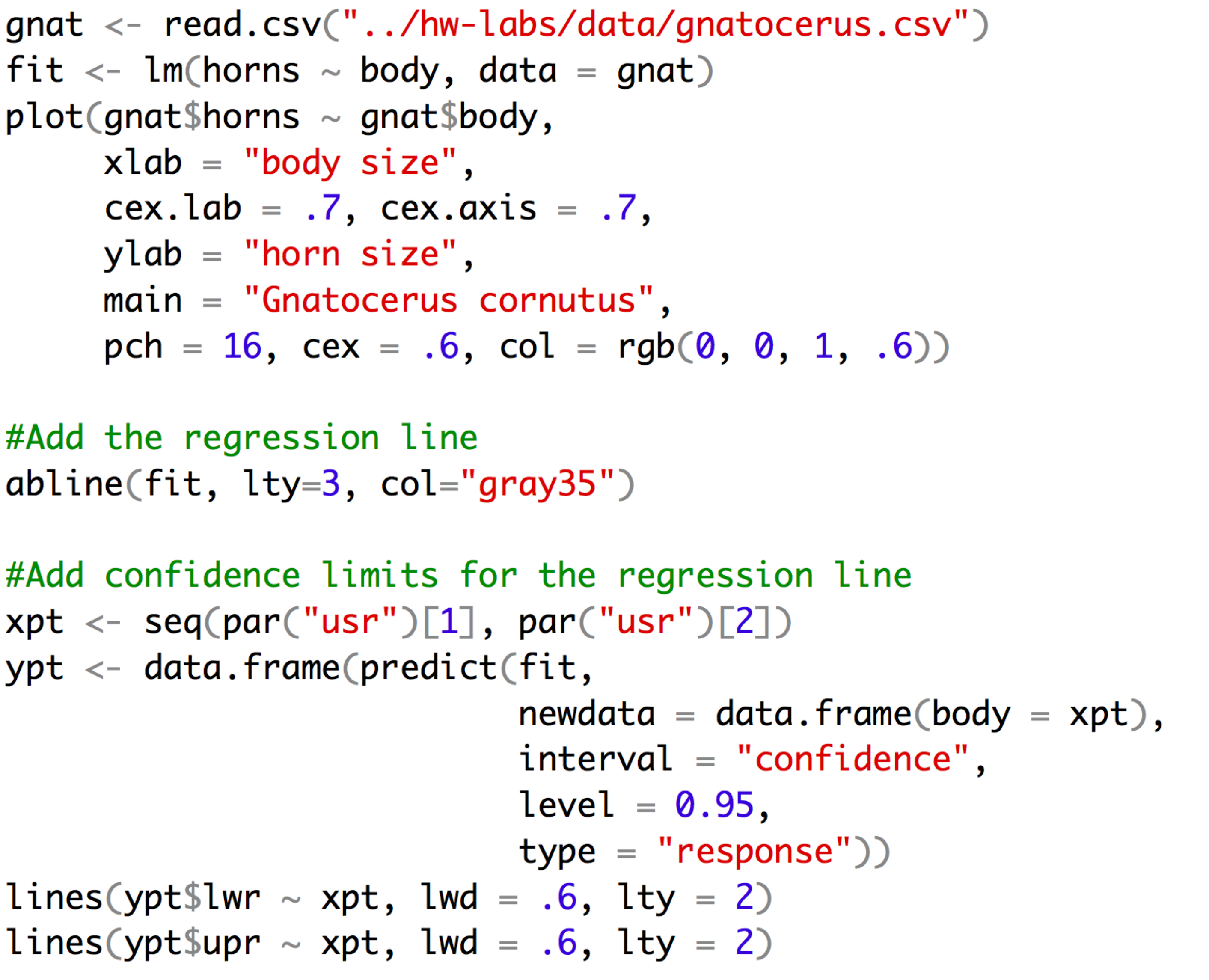
Residuals can be of interest in their own right, because they represent values that have been ***corrected*** for relationships that might be obscuring a pattern (e.g., the body weight-testes mass relationship)

# What are Residuals





# Making that plot



```r
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?

1. Based on a post-hoc observational study, construct a new hypothesis for a novel group or system that has not yet been studied
2. For example, given the primate data, a reasonable prediction is that residual testes mass in deer will be associated with mating system
3. Collect a new observational data set from the new group to test the hypothesis

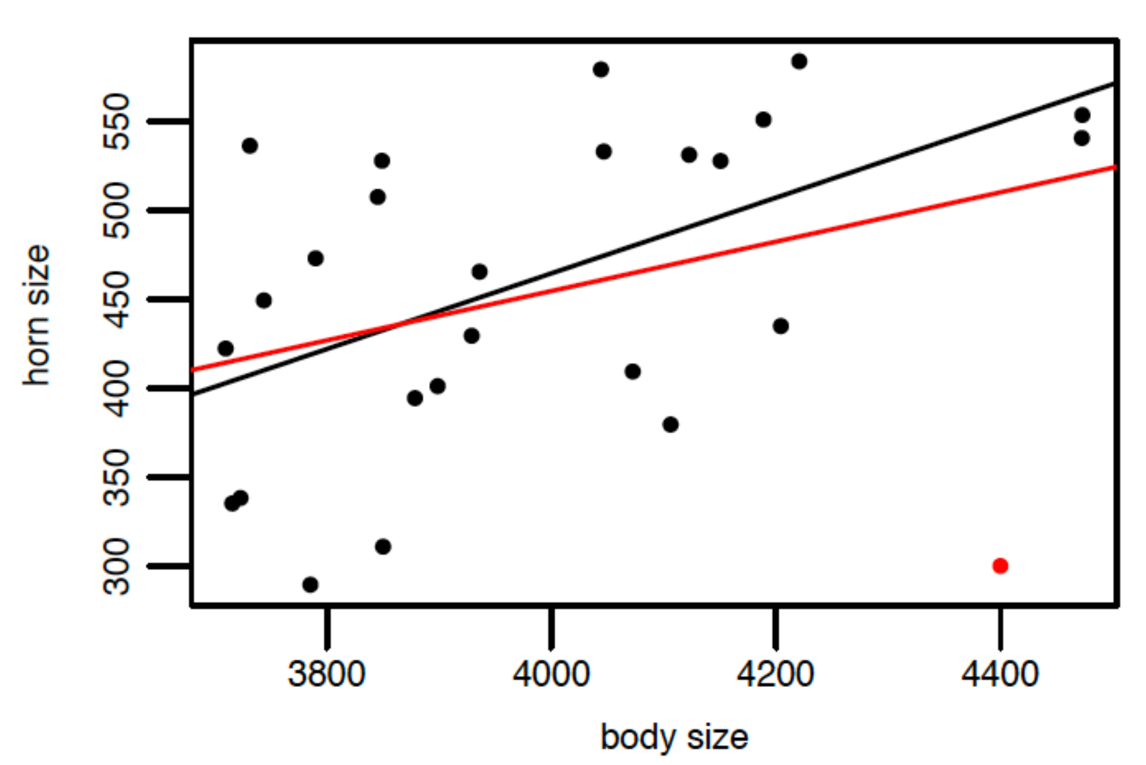
# Assumptions of Linear Regression

- 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***

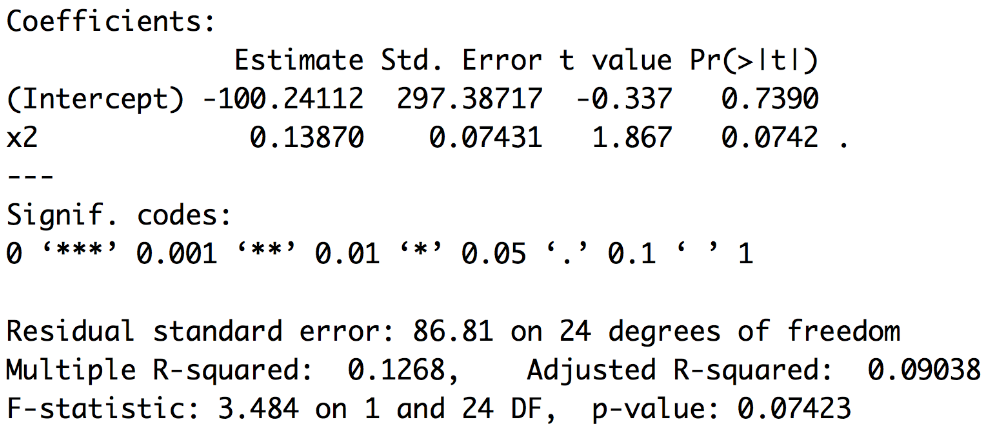
# Common Problems

- 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

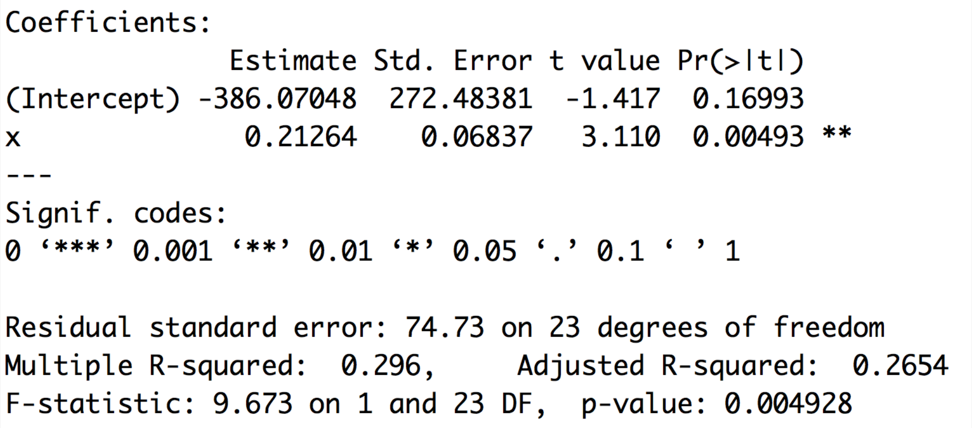
# Outliers



Leverage and cooks distance

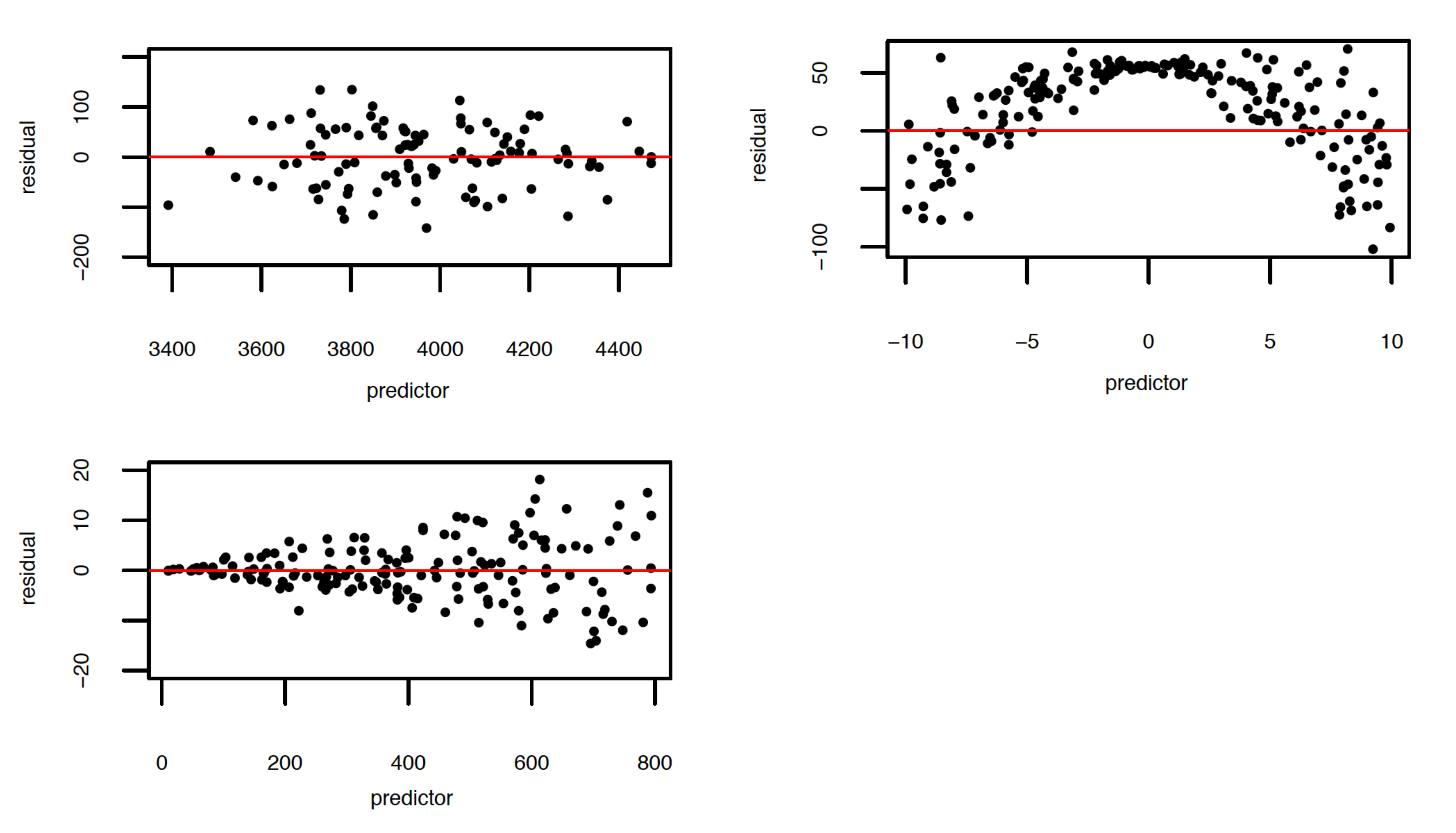


```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -100.24112  297.38717  -0.337   0.7390
x2             0.13870    0.07431   1.867   0.0742 .
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

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
```

# Residual plots



# Multiple Explanatory Variables

- 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***

# General Linear Models

- GLMs handle categorical factors and continuous factors in a single modeling framework

- **ANOVA** is a special case with just categorical explanatory variables

- **Linear regression** is a special case with just continuous explanatory variables

# Analysis of Covariance

- Used to test for a difference in means, while correcting for a variable that is correlated with the response variable

- The slopes must not differ in the two groups

  In other words, the mean comparison is only valid if the interaction term is not significant

- Also used to compare the slope of two regression lines

  If the interaction term is significant, then the conclusion is that the slopes are different

# Summary

- Statistical models can be quite complex, with potentially many factors and interaction terms

- The model is specified by something that looks like an equation: $Y = \mu + A + B + A*B$

- General linear models allow you to combine categorical and continuous factors into a single model
- Your sample size will limit the complexity of the model
- You will need to think about how you choose the model you estimated under (single model/model averaging)

# For Thursday

Read chapter WS 6-9

Bring laptop to class!

Heath Blackmon
BSBW 309
coleoguy@gmail.com
@coleoguy