Linear modeling II and logistic regression

STEPHANIE J. SPIELMAN, PHD

BIO5312, FALL 2017

General linear models

`lm(Numeric response ~ <predictors>)`

Single numeric predictor: Regression

Single categorical predictor: ANOVA

Multiple numeric predictors: multiple regression

Multiple categorical predictors: *n*-way ANOVA

Single categorical and *n* numeric predictors: ANCOVA

Multiple categorical and *n* numeric predictors: linear model

How does each predictor affect the response?
Goal is to model the response (*predict outcomes*) with a set of explanatory variables

General linear models

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \varepsilon$$

Interpreting coefficients depends on **type of variable**

- Categorical predictors: Increase/decrease in Y *relative to first category*
- Numeric predictors: Increase/decrease in Y for every 1 unit increase in X

Linear model coefficients

Single numeric predictor

```
> summary( lm(Sepal.Length ~ Sepal.Width, data = setosa) )

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.6390     0.3100   8.513 3.74e-11 ***
Sepal.Width   0.6905     0.0899   7.681 6.71e-10 ***
```

Single categorical predictor (with three levels)

```
> summary(lm(calcium ~ group, data = tidy.data))

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)         938.33      74.02  12.676 2.04e-09 ***
grouposteopenia    -138.33     104.69  -1.321 0.206168
grouposteoporosis  -540.00     104.69  -5.158 0.000117 ***
```

Linear models with multiple predictors

Additive effects consider the *independent* effect of each predictor on the response

$$Y = \beta_o + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

```
lm(Numeric response ~ predictor1 + predictor2, data = data)
```

Linear models with multiple predictors

Interaction effects consider the interaction between potentially non-independent predictors

$$Y = \beta_o + \beta_1 X_1 + \beta_2 X_2 + \beta_3 (X_1 * X_2) + \varepsilon$$

```
lm(Numeric response ~ predictor1 * predictor2, data = data)
```

```
lm(numeric response ~ predictor1 + predictor2 + predictor1:predictor2, data=data)
```

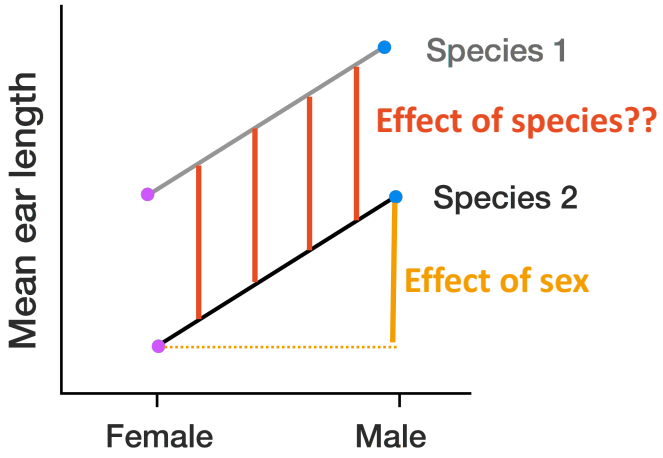
Wikipedia's "real world example": Adding sugar to coffee and stirring the coffee. Neither of the two individual variables has much effect on sweetness but a combination of the two does.

Interaction plots

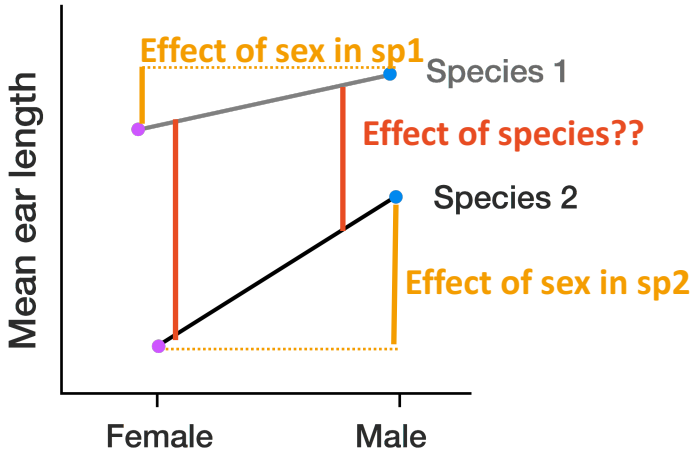
Visualize the interaction between **two categorical predictors**

- How do species and sex influence ear length in rabbits?
- Does ear length differ between different rabbit species, *controlling for sex*?
- Does ear length differ between different rabbit sexes, *controlling for species*?

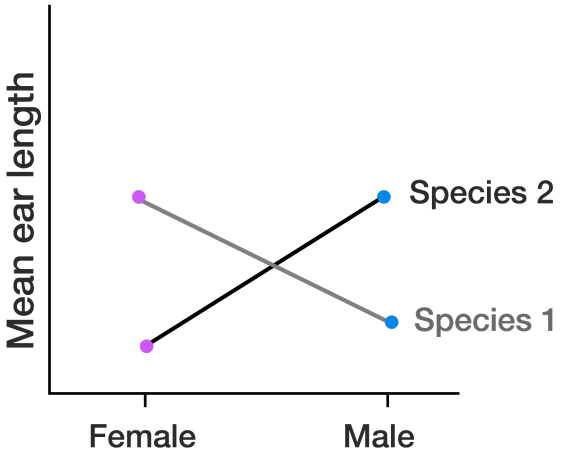
Independent



Interaction



Interaction



Two-way ANOVA = linear model with two categorical predictors

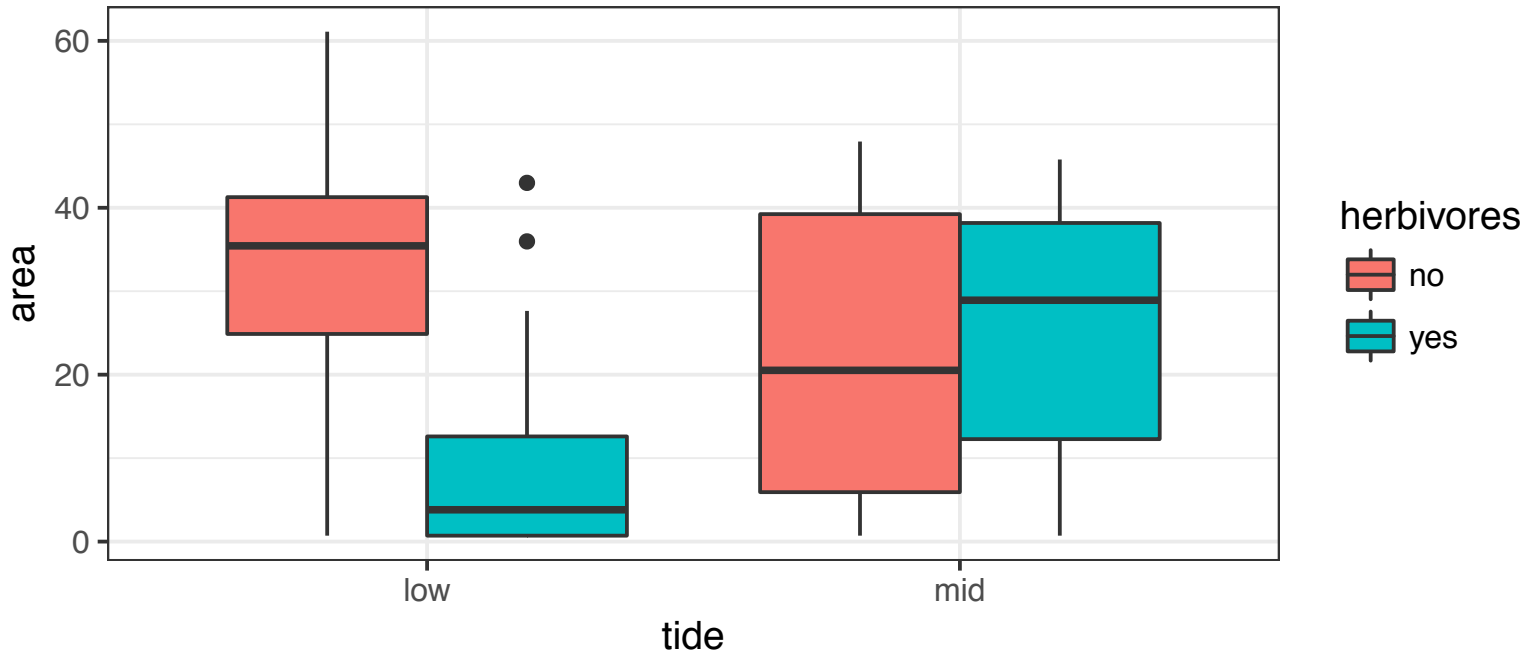
How do herbivores and intertidal zone affect the abundance of algae living in an intertidal habitat?

- With and without herbivore predation
- Low-tide vs mid-tide zone
- Measured surface area of algae after treatment

```
> head(algae)
  tide herbivores      area
1  low         no  9.405573
2  low         no 34.467736
3  low         no 46.673485
4  low         no 16.642139
5  low         no 24.377498
6  low         no 38.350604
```

Visualize the data distributions

```
> ggplot(algae, aes(x = tide, y = area, fill = herbivores)) + geom_boxplot()
```

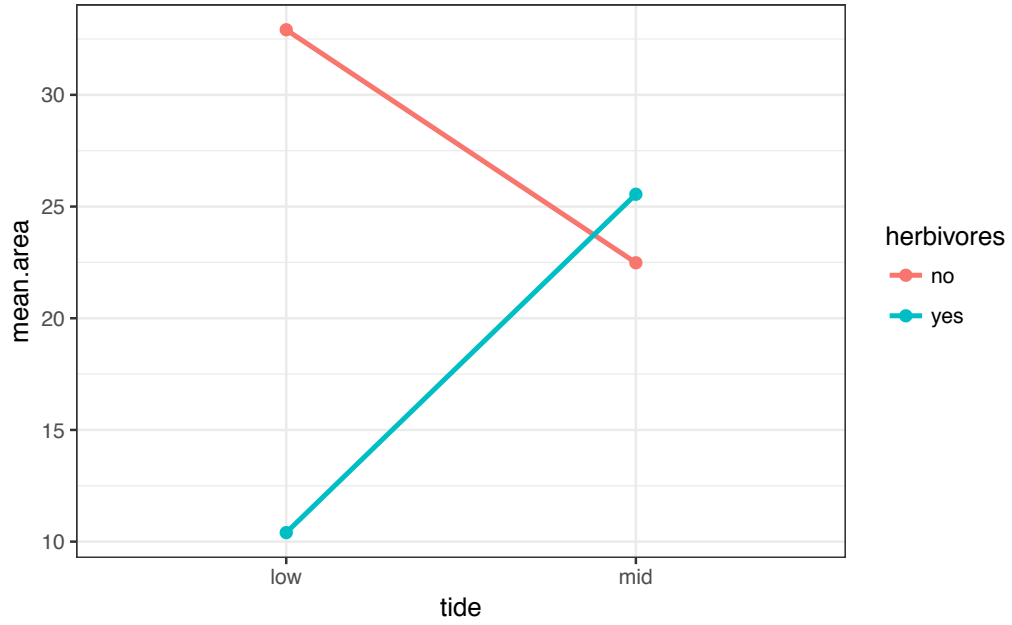


Visualize the data as interaction plot

```
> algae %>%
      group_by(tide, herbivores) %>%
      summarize(mean.area = mean(area)) -> algae.int

> algae.int
   tide herbivores mean.area
  <fctr>     <fctr>     <dbl>
1    low         no  32.91450
2    low        yes  10.40376
3    mid         no  22.48360
4    mid        yes  25.55094
```

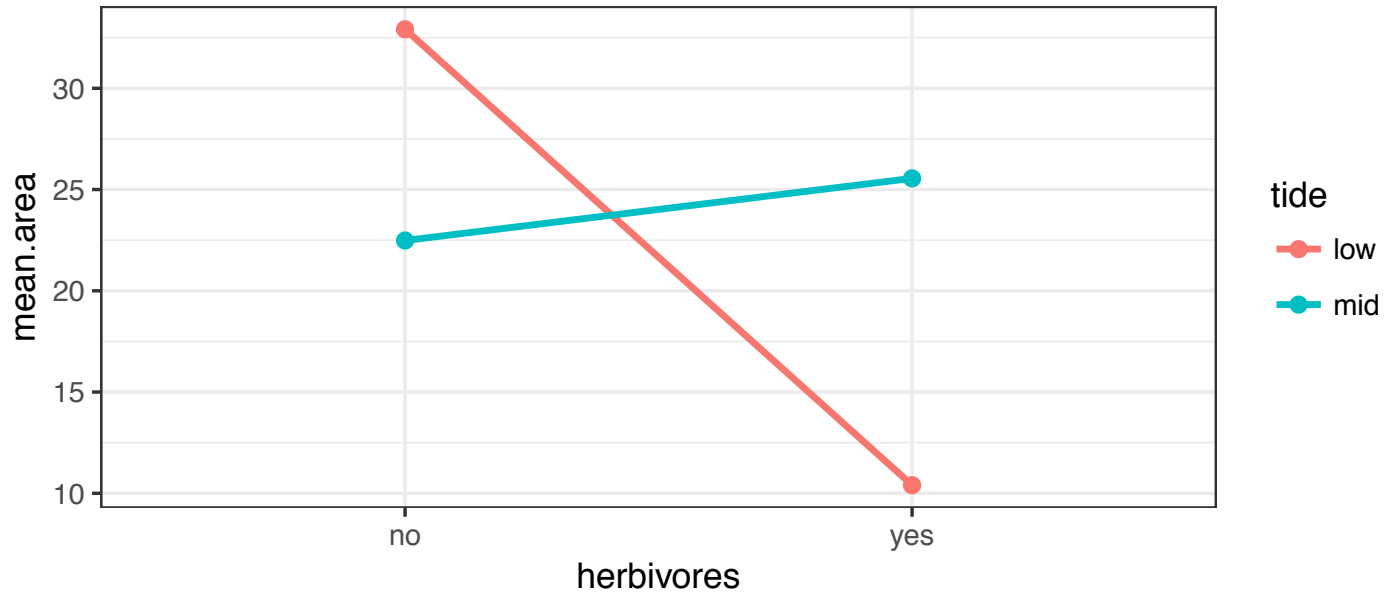
Visualize the data as interaction plot

```
> ggplot(algae.int, aes(x=tide,
                        y=mean.area,
                        group=herbivores,
                        color=herbivores)) + geom_point() + geom_line()
```



Visualize the data as interaction plot, 2

```
> ggplot(algae.int, aes(x=herbivores,
                        y=mean.area,
                        group=tide,
                        color=tide)) + geom_point() + geom_line()
```



Fit the model

```
> model.additive    <- lm(area ~ herbivores + tide, data = algae)
> model.interaction <- lm(area ~ herbivores * tide, data = algae)


> tidy(model.additive)
           term  estimate std.error  statistic      p.value
1   (Intercept) 26.519979  3.602368  7.3618178 5.526594e-10
2 herbivoresyes -9.721701  4.159657 -2.3371402 2.273087e-02
3       tidemid  2.358142  4.159657  0.5669078 5.728570e-01


> tidy(model.interaction)
                   term  estimate std.error statistic      p.value
1           (Intercept)  32.91450  3.855532  8.536955 5.979662e-12
2         herbivoresyes -22.51075  5.452546 -4.128484 1.145511e-04
3               tidemid -10.43090  5.452546 -1.913034 6.051935e-02
4 herbivoresyes:tidemid  25.57809  7.711064  3.317064 1.548555e-03
```

Interaction model preferred

Examine the model in full

```
> tidy(model.interaction)

                 term  estimate std.error statistic      p.value
1         (Intercept)  32.91450  3.855532  8.536955 5.979662e-12
2       herbivoresyes -22.51075  5.452546 -4.128484 1.145511e-04
3             tidemid -10.43090  5.452546 -1.913034 6.051935e-02
4 herbivoresyes:tidemid  25.57809  7.711064  3.317064 1.548555e-03
```

Ignore additive effects when interaction is sig.

```
> glance(model.interaction)
  r.squared adj.r.squared    sigma statistic     p.value df    logLik      AIC
1 0.2281462     0.1895535 15.42213  5.911644 0.001329155  4 -263.8382 537.6765
       BIC deviance df.residual
1 548.4709 14270.52          60
```

Examine the ANOVA table (if you want)

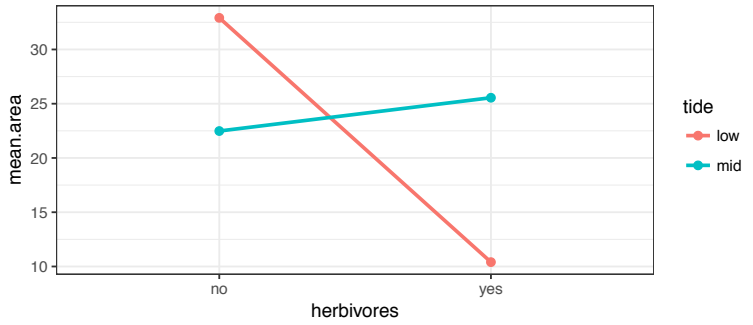
```
> anova(model.interaction)

Analysis of Variance Table


Response: area
                Df  Sum Sq Mean Sq F value   Pr(>F)
herbivores       1  1512.2 1512.18  6.3579 0.014360 *
tide             1    89.0   88.97  0.3741 0.543096
herbivores:tide  1  2617.0 2616.96 11.0029 0.001549 **
Residuals       60 14270.5  237.84
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Results and conclusions

We find a significant interaction effect between tidal zone and herbivore presence on algae growth area. On average, herbivores marginally increase algae growth in mid-tide, and herbivores greatly decrease algae growth in mid-tide.

Our model has a significant R^2=0.189, meaning that tidal zone and herbivore presence explain ~18.9% of the variation seen in algae growth area.

IS THIS A GOOD MODEL?



Hypothetically, let's say interaction was NS

```
> tidy(model.additive)

           term  estimate std.error  statistic      p.value
1   (Intercept) 26.519979  3.602368  7.3618178 5.526594e-10
2 herbivoresyes -9.721701  4.159657 -2.3371402 2.273087e-02
3       tidemid  2.358142  4.159657  0.5669078 5.728570e-01
```

The mean growth area is 26.52 under low-tide, no-herbivore conditions

Herbivores decrease algae growth by a factor of 9.722, compared to no herbivores.

~~Mid-tide increases algae growth by a factor of 2.358, compared to low-tide.~~

```
> glance(model.additive)
  r.squared adj.r.squared    sigma statistic    p.value df    logLik      AIC
1 0.0866022    0.05665475 16.63863  2.891804 0.06311441  3 -269.2263 546.4526
       BIC deviance df.residual
1 555.0881 16887.48          61
```

Results and conclusions

We find that herbivore presence significantly influences algae growth area. Herbivores decrease algae growth, on average, by a factor of 9.722, compared to no herbivores. Low vs. mid tide does not have a significant effect on algae growth. We did not detect an interaction effect between tide and herbivores.

Our model has a significant R^2=0.05665, meaning that tidal zone and herbivore presence explain ~5.7% of the variation seen in algae growth area.

The R^2 output

Model	R-squared	Adjusted R-squared
area ~ herbivores	0.0818	0.067
area ~ herbivores + tide	0.0866	0.05665

R^2 will **always** increase with more predictors

- It will fit noise if no signal
- When fitting a single model, consider this quantity

Adjusted R^2 accounts for presence of fitted noise

- When fitting multiple models and **selecting a model** based on R^2, consider this quantity, see next week for details

Breathe break

ANCOVA: **An**alayis of **Co**variance

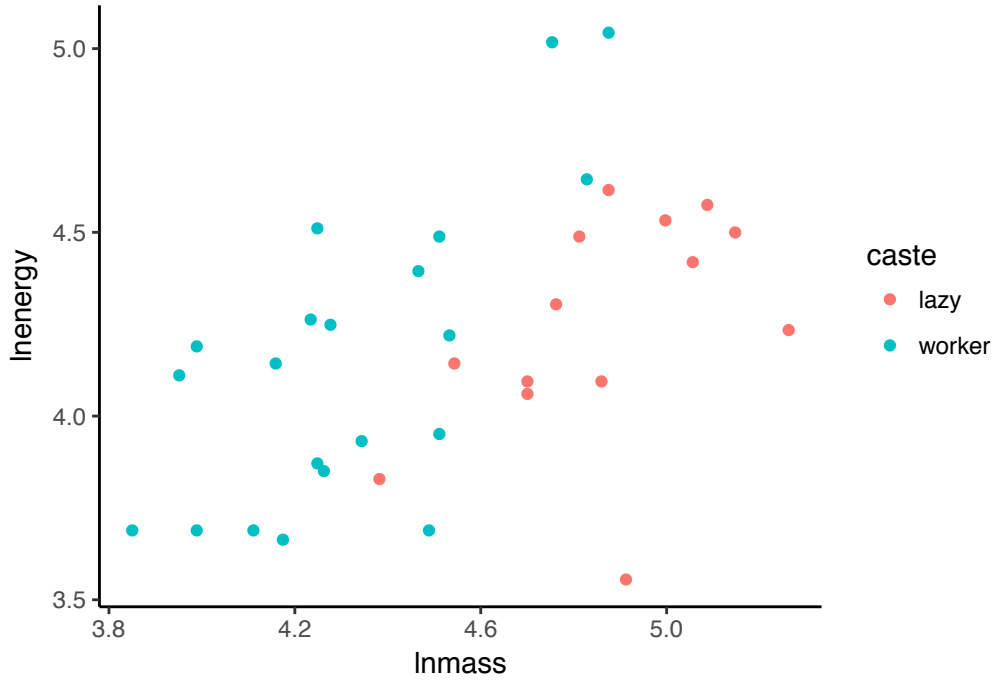
Mole rats have distinct social castes, where in a given colony only the single queen and a few males reproduce. The remaining males are workers. Researchers suspect that there may be also worker castes, with "frequent" and "infrequent" workers. They measured body mass and daily energy expenditure between the two groups of candidate castes.

Is energy expenditure different between castes, controlling for body weight?

Body weight is the **covariate**

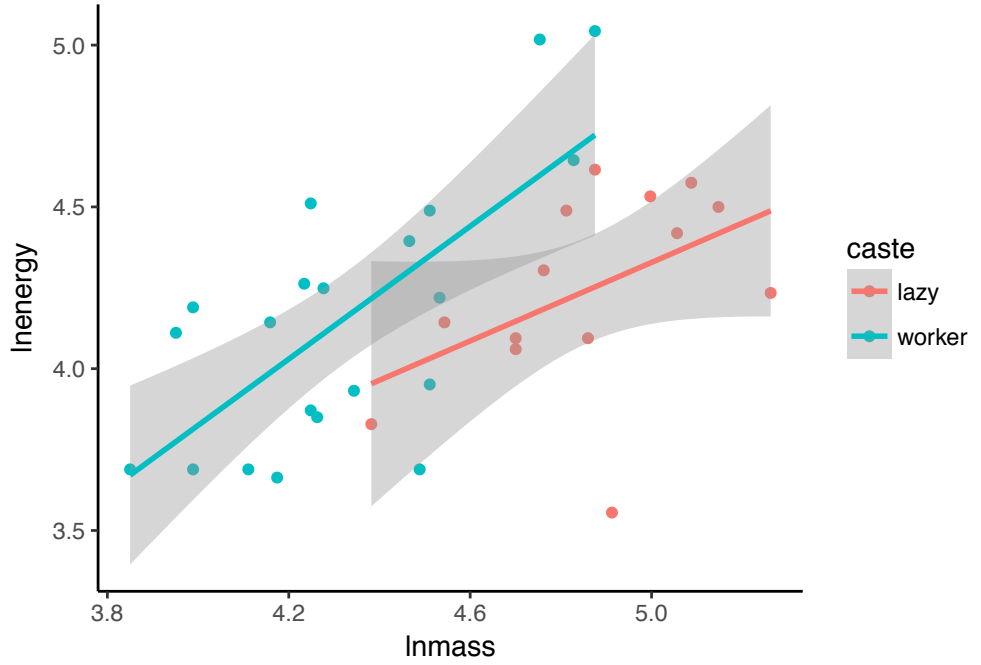
Visualize the data

```
> head(mole)
   caste   lnmass lnenergy
1 worker 3.850148 3.688879
2 worker 3.988984 3.688879
3 worker 4.110874 3.688879
4 worker 4.174387 3.663562
5 worker 4.248495 3.871201
6 worker 4.262680 3.850148

> ggplot(mole, aes(x = lnmass, y = lnenergy,
                   color = caste)) + geom_point()
```



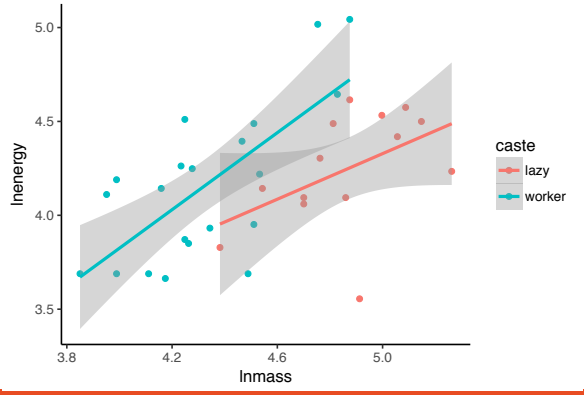
Visualize the data, 2

```
> ggplot(mole, aes(x = lnmass, y = lnenergy, color = caste)) +
      geom_point() +
      geom_smooth(method = "lm", aes(group = caste))
```



Fit the model, first with interaction effect

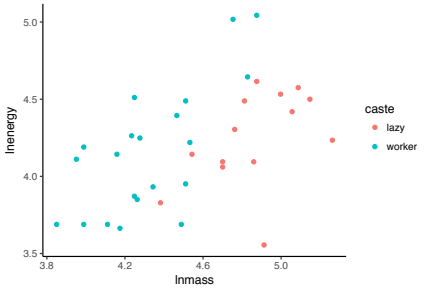
```
> model <- lm(lnenergy ~ lnmass * caste, data = mole)
> model <- lm(lnenergy ~ lnmass + caste + lnmass:caste, data = mole)

> tidy(model)
                 term   estimate std.error  statistic    p.value
1         (Intercept)  1.2939446 1.6691352  0.7752186 0.44408413
2              lnmass  0.6068898 0.3427563  1.7706163 0.08645869
3         casteworker -1.5712513 1.9518215 -0.8050179 0.42694079
4 lnmass:casteworker  0.4186224 0.4147347  1.0093740 0.32060937
```



Fit the additive model

```
> model <- lm(lnenergy ~ lnmass + caste, data = mole)
> tidy(model)
         term    estimate std.error  statistic      p.value
1 (Intercept) -0.09686842 0.9423033 -0.1027996 9.187635e-01
2      lnmass  0.89281513 0.1930335  4.6251819 5.886779e-05
3 casteworker  0.39334235 0.1461059  2.6921734 1.119835e-02

> glance(model)
  r.squared adj.r.squared     sigma statistic      p.value df
1 0.4090389     0.3721038 0.2965689  11.07454 0.0002212804  3
```

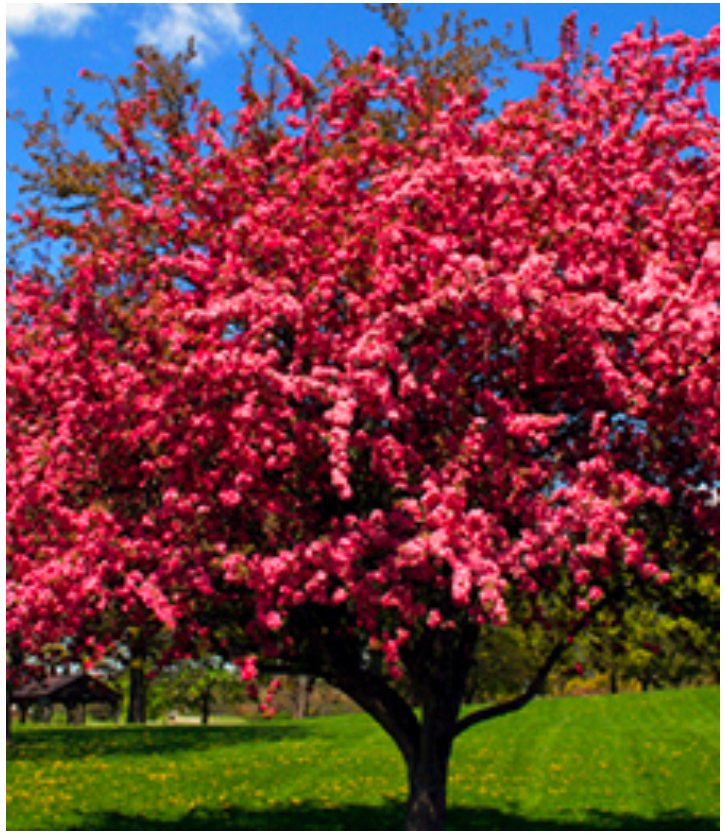
Live exercise: Interpret this model



Breathe break

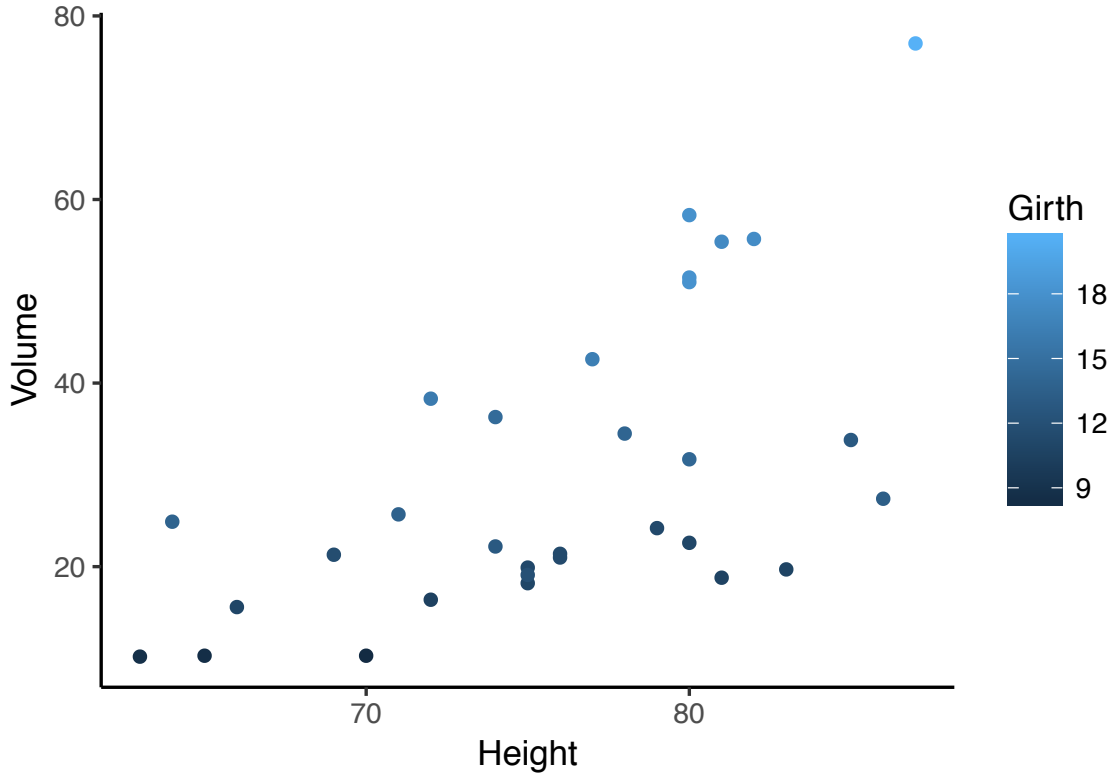
Multiple numeric predictors

```
> head(trees)
  Girth Height Volume
1   8.3     70   10.3
2   8.6     65   10.3
3   8.8     63   10.2
4  10.5     72   16.4
5  10.7     81   18.8
6  10.8     83   19.7

> nrow(trees)
[1]  31
```



Visualize the data

```
> ggplot(trees, aes(x = Height, y = Volume, color = Girth)) + geom_point()
```



Interpreting multiple numeric predictors

```
> model <- lm(Volume ~ Girth + Height, data = trees)
> tidy(model)
         term    estimate std.error statistic      p.value
1 (Intercept) -57.9876589 8.6382259 -6.712913 2.749507e-07
2       Girth   4.7081605 0.2642646 17.816084 8.223304e-17
3      Height   0.3392512 0.1301512  2.606594 1.449097e-02
```

A tree with girth and height of 0 will yield a timber volume of -57.99, on average.

For every unit increase in tree girth, timber volume increases by 4.708, on average.

For every unit increase in tree height, timber volume increases by 0.34, on average.

Interpreting multiple numeric predictors

```
> model <- lm(Volume ~ Girth + Height, data = trees)
> glance(model)
  r.squared adj.r.squared    sigma statistic      p.value df
1   0.94795     0.9442322 3.881832  254.9723 1.071238e-18  3
```

R^2 = 0.95, meaning that 95% of the variation in cherry tree timber volume can be explained by tree girth and height.

Interaction effect interpretation

```
> tidy(lm(Volume ~ Girth * Height, data = trees))
          term   estimate   std.error statistic      p.value
1  (Intercept) 69.3963156 23.83575455  2.911438 7.130666e-03
2        Girth -5.8558479  1.92133589 -3.047800 5.108696e-03
3       Height -1.2970834  0.30984288 -4.186262 2.699338e-04
4 Girth:Height  0.1346544  0.02437731  5.523759 7.484103e-06
```

There is a significant interaction effect between girth and height for modeling volume of cherry tree timber.

The effect of girth on timber volume increases by 0.135 for every unit increase of height.

Interaction effect interpretation

```
> glance(lm(Volume ~ Girth * Height, data = trees))
   r.squared adj.r.squared   sigma statistic      p.value df    logLik      AIC
1 0.9755642     0.9728491 2.70855  359.3122 7.290458e-22  4 -72.73458 155.4692
       BIC deviance df.residual
1 162.6391 198.0786          27
```

Our R^2 = 0.97, meaning that 97% of the variation in timber volume can be explained by the interaction between tree girth and height.

Always prefer the interaction model, if effect is significant.

Recap on interpreting coefficients

Categorical variable coefficients

- Increase in Y relative to other levels of X

Numeric variable coefficients

- Increase in Y for every unit increase in X

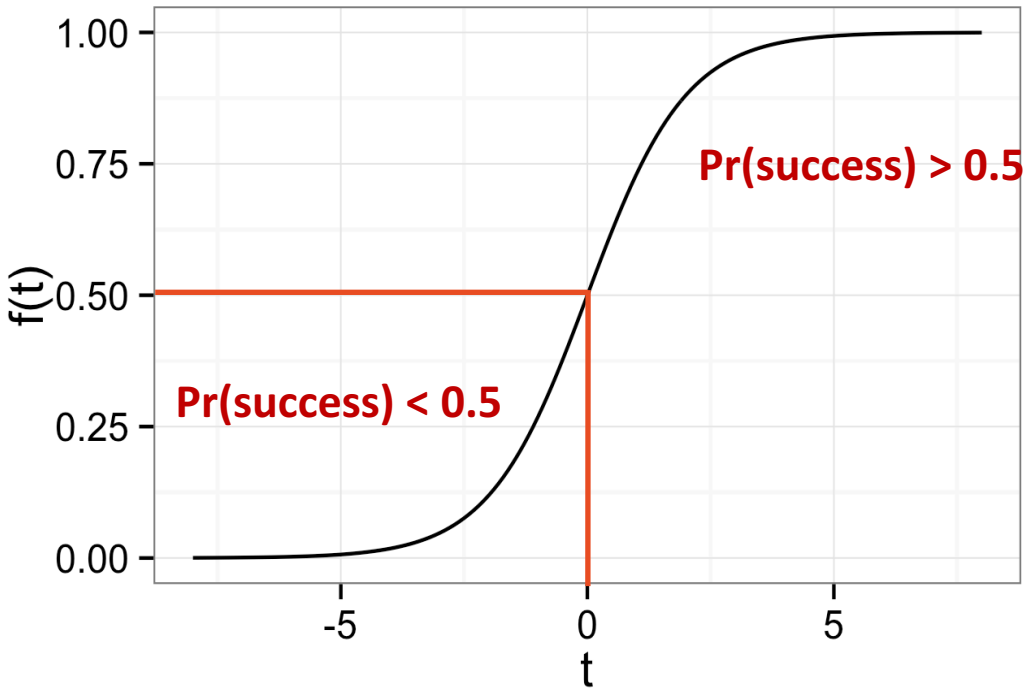
Interaction coefficients

- **Categorical-numeric**: Different slopes across categories
- **Numeric-numeric**: One numeric modulates the influence of the other

Exercise break

Logistic regression

Model a **binary response** instead of a numeric response by fitting a *logistic curve* to the data



$$f(t) = \frac{e^t}{1+e^t}$$

$$\Pr(success) = \frac{e^t}{1+e^t}$$

$$t = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \varepsilon$$

Logistic regression is a **classifier**

See machine learning classes for more details

Logistic regression with biopsy data

```
> head(biopsy)
  clump_thickness uniform_cell_size uniform_cell_shape marg_adhesion
1               5                 1                 1             1
2               5                 4                 4             5
3               3                 1                 1             1
4               6                 8                 8             1
5               4                 1                 1             3
6               8                10                10             8
  epithelial_cell_size bare_nuclei bland_chromatin normal_nucleoli mitoses
1                    2           1               3               1       1
2                    7          10               3               2       1
3                    2           2               3               1       1
4                    3           4               3               7       1
5                    2           1               3               1       1
6                    7          10               9               7       1
    outcome
1    benign
2    benign
3    benign
4    benign
5    benign
6 malignant
```

Running a logistic regression

glm(binary response ~ <predictors>)

```
> model <- glm(outcome ~ clump_thickness +
                         uniform_cell_size +
                         uniform_cell_shape +
                         marg_adhesion +
                         epithelial_cell_size +
                         bare_nuclei +
                         bland_chromatin +
                         normal_nucleoli +
                         mitoses,
               data=biopsy,
               family=binomial)
```

```
> model <- glm(outcome ~ ., data=biopsy, family=binomial)
```

Logistic regression output

```
> tidy(model)
                   term      estimate  std.error   statistic      p.value
1           (Intercept) -10.103942243 1.17487744 -8.59999681 7.971831e-18
2       clump_thickness   0.535014068 0.14201743  3.76724220 1.650608e-04
3     uniform_cell_size  -0.006279717 0.20907739 -0.03003537 9.760388e-01
4    uniform_cell_shape   0.322706496 0.23060065  1.39941710 1.616879e-01
5         marg_adhesion   0.330636915 0.12345089  2.67828703 7.399977e-03
6  epithelial_cell_size   0.096635417 0.15659236  0.61711452 5.371592e-01
7           bare_nuclei   0.383024572 0.09384327  4.08153469 4.473930e-05
8       bland_chromatin   0.447187920 0.17138238  2.60929928 9.072785e-03
9       normal_nucleoli   0.213030682 0.11287348  1.88734050 5.911454e-02
10              mitoses   0.534835631 0.32877389  1.62675821 1.037885e-01

> glance(model)
  null.deviance df.null   logLik      AIC      BIC deviance df.residual
1      884.3502     682 -51.4441 122.8882 168.1531 102.8882         673
```

Interpreting coefficients

For every unit increase in the predictor, the **log odds** of the response (malignancy) increases by the coefficient

- Log odds = Log(Pr(success)/Pr(failure))

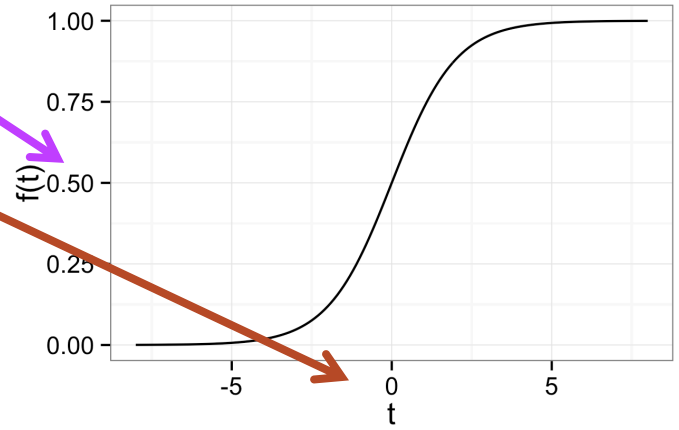
```
> tidy(model)
                  term      estimate  std.error   statistic      p.value
2      clump_thickness   0.535014068 0.14201743  3.76724220 1.650608e-04
5        marg_adhesion   0.330636915 0.12345089  2.67828703 7.399977e-03
7          bare_nuclei   0.383024572 0.09384327  4.08153469 4.473930e-05
8      bland_chromatin   0.447187920 0.17138238  2.60929928 9.072785e-03
```

Logistic regression output

```
> head( model$fitted.values )
          1           2           3           4           5           6
0.016046581 0.908808622 0.008137623 0.760934919 0.018166848 0.999973622

> head( model$linear.predictors )
         1          2          3         4         5         6
-4.116083  2.299174 -4.803086  1.157812 -3.989823 10.542969
```

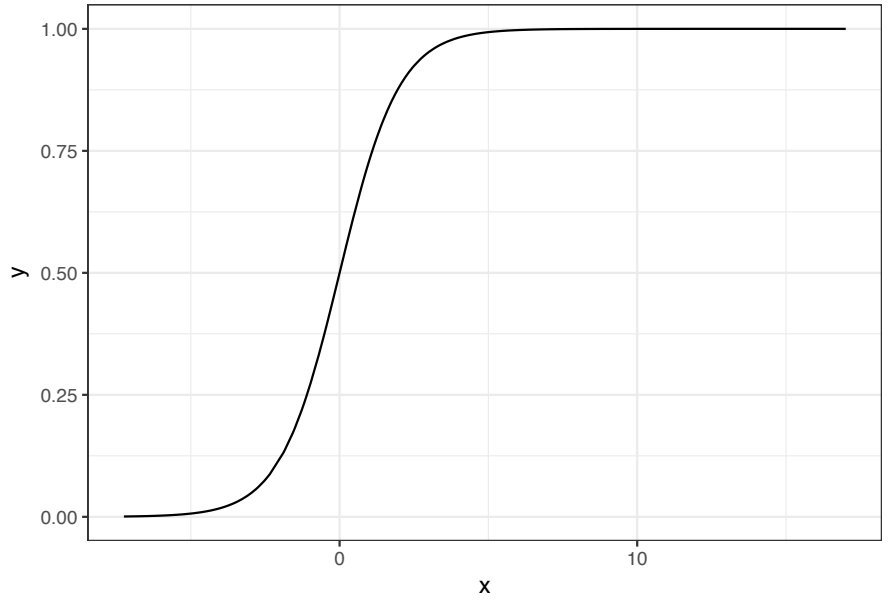
$$\Pr(success) = \frac{e^t}{1+e^t}$$

$$t = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \varepsilon$$



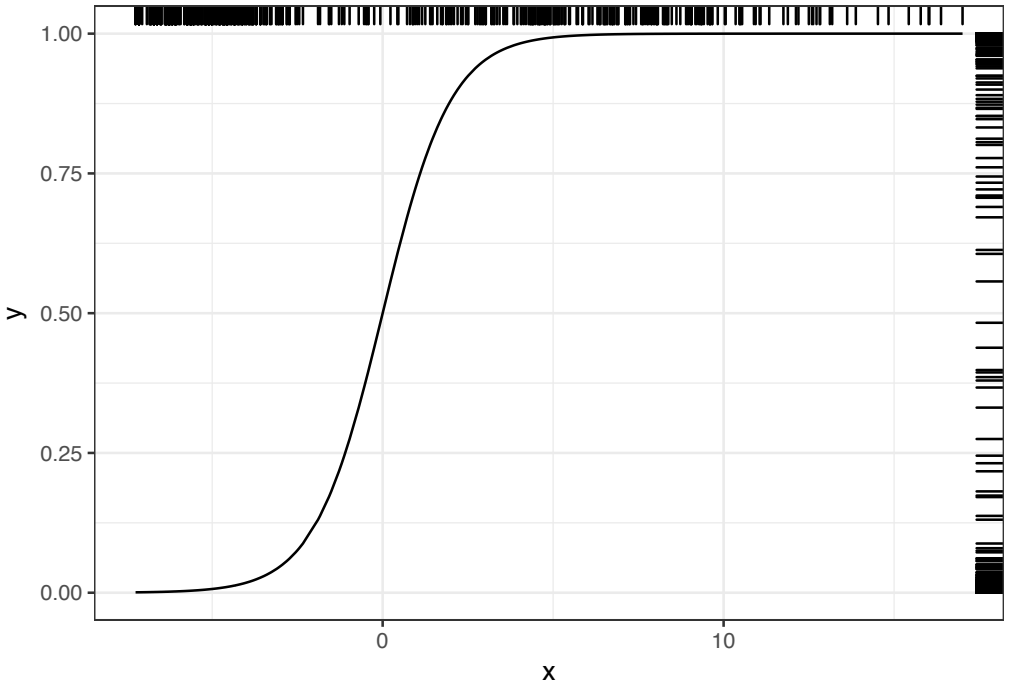
Visualize the logistic regression

```
> model.fit <- tibble(x = model$linear.predictors,
                      y = model$fitted.values,
                      outcome = biopsy$outcome)

> ggplot(model.fit, aes(x = x, y = y)) + geom_line()
```



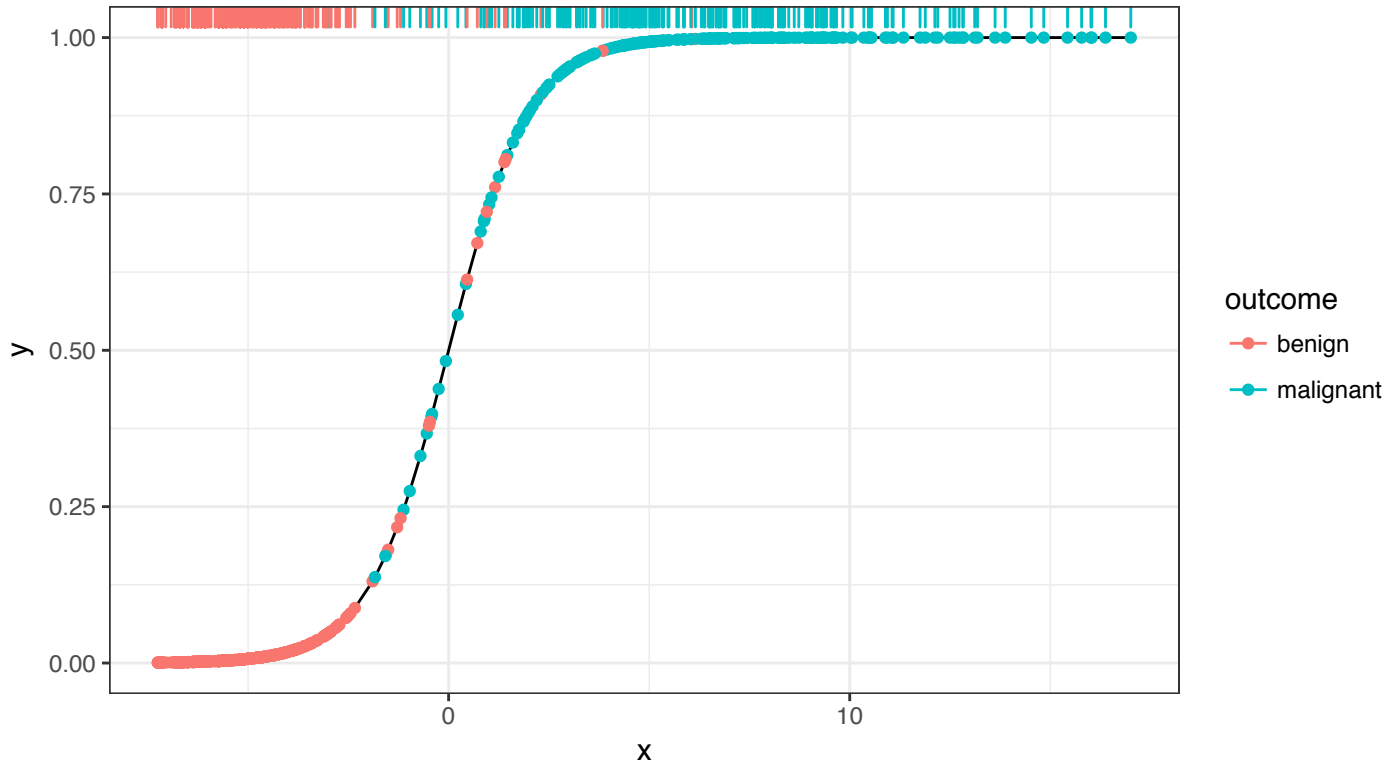
Visualize the logistic regression with a rug

```
> ggplot(model.fit, aes(x = x, y = y)) + geom_line() + geom_rug(sides = "tr")
```



Logistic regression clearly separated the groups

```
> ggplot(model.fit, aes(x = x, y = y, color = outcome)) +
      geom_line() +
      geom_point() +
      geom_rug(sides = "t")
```



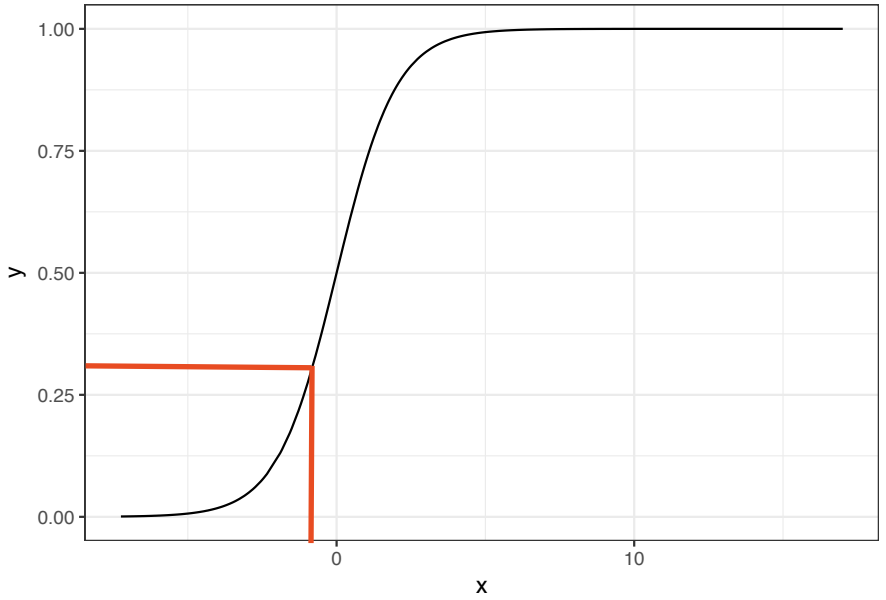
Using the logistic regression

$$\Pr(success) = \frac{e^t}{1+e^t}$$

$$t = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \varepsilon$$

```
> new.patient <- tibble(clump_thickness = 4,
                        uniform_cell_size = 2,
                              uniform_cell_shape = 7,
                              marg_adhesion = 3,
                              epithelial_cell_size = 8,
                              bare_nuclei = 1,
                              bland_chromatin =5,
                              normal_nucleoli = 2,
                              mitoses = 0)

> predict(model, new.patient)
         1
-0.9074803

> predict(model, new.patient, type = "response")
         1
0.2875157
```

Our logistic model gives a 28.7% probability that the new patient has a malignancy.



Next week we will evaluate models

How do we choose the best predictors, i.e. select the best model?

- This week's homework will teach you a very common, but very mediocre, approach

How can we evaluate model performance?

What procedures can we use to build models as robustly as possible?