Confidence intervals for logistic regression slopes
  Two methods - correspond to the two test methods
  Wald CI:
$$\left(\hat{\beta}_1 - z_{1-\alpha/2} \times se,\ \ \hat{\beta}_1 + z_{1-\alpha/2} \times se\right) = \hat{\beta}_1 + \pm z_{1-\alpha/2} \times se$$
    For 95% interval, use $z_{0.975} = 1.96$
    Donner: $-0.0665 \pm 1.96 \times 0.0322 = (-0.130, -0.0034)$
  Likelihood CI:
    No simple expression, computed numerically
    Donner: $(-0.140, -0.010)$
    As with the tests, Likelihood makes fewer assumptions
  These are intervals for the log odds ratio
  Usually simpler to report (and interpret) intervals for odds ratios
    Exponentiate the end points of the log odds intervals
    Donner, Wald: $(\exp -0.130,\ \exp -0.0034) = (0.88, 0.997)$
    Donner, Likelihood: $(\exp -0.140,\ \exp -0.010) = (0.87, 0.990)$

Reporting the association of age and P[surv]
  If this were an experimental study, could say:

    Increasing age by 1 year multiplies the odds of survival by 0.936, 95% ci (0.87, 0.99)

  But this is an observational study, so can't imply age reduced the survival

    The odds of survival of an individual is 0.936 (95% ci: 0.87, 0.99) times that for an individual one year younger.

    The odds of survival of an individual is 1.068 (95% ci: 1.01, 1.15) times that for an individual one year older

Multiple Linear Regression (MLR):
  More than one X variable

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + \varepsilon_i$$

  Three "hard" parts:
    Interpretation of coefficients
    Choosing the appropriate model comparison
    Choosing a model to answer study questions
  Plus some details (probably next week):
    Standardized residuals:
    Additional diagnostics: Cook's D, VIF
  And a major new topic: model selection (definitely next week)

Motivating example: Brain weights across mammal species, Chapter 9 Case study 2
  Brain weight is positively associated with body size
  Is it associated with other characteristics, e.g. gestation period or litter size?
    After accounting for body size

Plot the pairs of variables - see non-linear relationships
Log transform all variables
relationships now look like straight lines

Interpretation of coefficients
Intercept, β_0: mean Y when **all** X variables = 0
Slope, β_j: effect (or difference) when X_j increased by 1
and all other variables held constant
Estimated coefficients may depend on which other variables in model
Brain size case study: β for log litter size = -2.08 or -0.54 or -0.31
Answering different questions because holding different variables constant
SLR and MLR coefficients are the same only when X's are uncorrelated

The "X" matrix:
Write $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} \cdots + \beta_k X_{ki}$ as a matrix multiplication: $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$

$$\boldsymbol{X} = \begin{bmatrix} 1 & X_{11} & X_{21} & \cdots & X_{k1} \\ 1 & X_{12} & X_{22} & \cdots & X_{k2} \\ 1 & X_{13} & X_{22} & \cdots & X_{k2} \end{bmatrix} \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix}$$

Details of matrix multiplication not relevant
You need to know what the X matrix is, in case you read or hear about it
Point to know for later is that the intercept is an "X" variable with value = 1

Estimation, etc.: No new concepts, computers needed for almost all computation
Estimation: no simple non-matrix formula
betas: "simple" matrix expression, $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}$
Same equation for 5 variables or 500 variables - just more columns in $\boldsymbol{X}$
Also works for SLR (1 variable, 2 columns in $\boldsymbol{X}$)
Modern software uses the matrix approach even for SLR
Requires a computer for matrix inverse and matrix multiplication
error sd $\hat{\sigma} = s = \sqrt{\Sigma(Y_i - \hat{Y}_i)^2/(n-p)}$
p = # param, including intercept, so p = # X variables + 1
error df: $n - p$
Precision: simple matrix algebra expression, no simple formula
Var-Cov matrix of $\hat{\boldsymbol{\beta}} = s^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$
Depends on spread in X values, # obs, correlation between 2 (or more) X's
Inference: T tests on individual parameter - as usual
Or F test for Overall regression (next section)

Model comparisons:
Most of our model comparisons have been a full model vs intercept only
e.g., different means (full) vs equal means (only an intercept)

One exception: ANOVA lack of fit: regression model vs different means model
When more than 1 term in the model, many possible model comparisons
Some more useful than others
Overall Regression: model comparison between full model and $Y_i = \beta_0 + \varepsilon_i$
Null hypothesis: **all coefficients** = 0, except intercept; $\beta = 0$, except β_0
F test, often presented as an ANOVA table
F tests of individual terms:
More than one model comparison
Sequential tests: Type I SS and tests
Partial tests: type III SS and tests

Example: 3 X variables called A, B, and C: E $Y_i = \beta_0 + \beta_1 A_i + \beta_2 B_i + \beta_3 C_i$

Type I = sequential tests:
Drop in error SS when term added to model with "previous" terms
full model = previous terms + this one
reduced model = previous terms in equation

Term	Reduced	Full
A	β_0	$\beta_0 + \beta_1 A_i$
B	$\beta_0 + \beta_1 A_i$	$\beta_0 + \beta_1 A_i + \beta_2 B_i$
C	$\beta_0 + \beta_1 A_i + \beta_2 B_i$	$\beta_0 + \beta_1 A_i + \beta_2 B_i + \beta_3 C_i$

Order of the terms in the equation matters
because each term added to preceding terms

Different order of variables: E $Y_i = \beta_0 + \beta_3 C_i + \beta_2 B_i + \beta_1 A_i$

Term	Reduced	Full
A	$\beta_0 + \beta_1 B_i + \beta_2 C_i$	$\beta_0 + \beta_1 A_i + \beta_2 B_i + \beta_3 C_i$
B	$\beta_0 + \beta_1 C_i$	$\beta_0 + \beta_1 C_i + \beta_2 B_i$
C	β_0	$\beta_0 + \beta_1 C_i$

Two different tests for C
Different "previous" variables
Almost always different results
Same only when all X's are uncorrelated with each other

Type III = partial tests:
Drop in error SS when term added
full model = all terms in model
reduced model = all other terms in equation (i.e. omitting this term)

Term	Reduced	Full
A	$\beta_0 + \quad\quad \beta_2 B_i + \beta_3 C_i$	$\beta_0 + \beta_1 A_i + \beta_2 B_i + \beta_3 C_i$
B	$\beta_0 + \beta_1 A_i + \quad\quad \beta_3 C_i$	$\beta_0 + \beta_1 A_i + \beta_2 B_i + \beta_3 C_i$
C	$\beta_0 + \beta_1 A_i + \beta_2 B_i$	$\beta_0 + \beta_1 A_i + \beta_2 B_i + \beta_3 C_i$

Note: Type I test for last term in model always same as type III test

Sequential and partial described using SSE, for regressions with normally distributed errors
  Linear regression (all terms have 1 df):
    Type III F tests correspond to T tests on individual parameters
    F statistic = (T statistic)^2, have same p-value
  Concepts of which models are compared apply to ANOVA models
    when model has more than one term
  Logistic regression:
    one parameter: Type III model comparisons test same hypothesis as Z test
    Answers are similiar but not identical
      Difference between Wald and likelihood ratio (drop in deviance) tests

Type II:
  Same as Type III when no interactions in the model
    Interactions will be discussed soon
    When there are interactions, type III preferred
  My opinion: just use type III instead of type II
  But some software calls it type II, even though it's better known as type III test

Type IV:
  There is also a type IV, has historical interest only.
    Was developed for a specific difficult situation.
      Missing cells in factorial ANOVA
    Didn't actually work as intended. Never used today

Which approach should I use?
  When all X variables are uncorrelated, type I = type III
    Very rare in regression problems
    Does happen in ANOVA with equal sample sizes per treatment
  US practice: use Type III tests almost all the time
    These answer the most interesting questions
    And, don't have to decide the "correct" order of terms
  ANOVA lack of fit: Type I because the sequence matters
    regression, then means model. Other way around (means then reg) is junk
  Some other parts of the world: Type I
    Folks who designed the R lm/anova functions preferred type I
    BEWARE: anova() gives you type I (sequential) tests
      Output doesn't tell you that you're getting type I
      probably not what you want
    until a few years ago, hard to get type III tests from R
      now: use functions in add-on libraries (emmeans, car, lmerTest)

Fun with models (part 1): Constructing a model to answer various sorts of questions.
If you want to "control" for important confounding variables
Your focus is relationship between X_1 and Y.
But you know that Y may be related to X_2, X_3,
Add X_2, X_3, ... to model
β_1 is relationship between Y and X_1 when all others held constant
Ex: litter size and brain weight, controlling for body weight,
All variables log transformed for linearity
log litter size = β_0 + β_1 log body weight + β_2 log brain weight
Human nutrition:
Standard practice to include age, gender and sometimes BMI in models
Report results without those variables and with those variables in model

If you want to allow lines to curve (classical approaches):
quadratic regression:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \varepsilon_i$$

β_2 quantifies the curvature ($\beta_2 = 0 \Rightarrow$ straight line)
Usual interpretation of β_1 and β_2 fails
Can't change X while holding X^2 constant
Test of $\beta_2 = 0$ tells you whether straight line adequate
Max/min Y at $X_m = -\hat{\beta}_1/(2\hat{\beta}_2)$
se X_m is hard; ci or tests even harder
polynomial regression, if quadratic isn't "wiggly" enough

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \beta_3 X_i^3 \{+\beta_4 X_i^4 \cdots\} + \varepsilon_i$$

Much less frequently used; much harder to interpret coefficients
Mostly used for predictions within range of X.
Extrapolations beyond range of X are usually very "wild" and untrustable

If you want to allow lines to curve (modern approaches):
non-parametric regression

$$Y_i = f(X_i) + \varepsilon_i$$

semi-parametric regression: Some X are curves, others are specified form

$$Y_i = f_1(X_{1i}) + \beta_2 X_{2i} + \varepsilon_i$$

Generalized additive model (GAM):

$$\mu_i = f_1(X_{1i}) + f_2(X_{2i}) \ \{+f_3(X_{3i}) \ldots\}$$

"Generalized": Y may be normal, Bernoulli, Binomial, Poisson, and others
left-hand side may include transformations, e.g. log, logit
Called the link function
"Additive": Effect of X_1 added to that from X_2 (and that from $X_3 \ldots$)
$f(X_i)$ is an arbitrary function relating Y_i to X_i.

Oodles of ways to estimate $f()$, including
  Splines
  kernel smoothing
  Support vector machines
  Neural networks
    (Feed forward NN, Convolution NN, Deep Learning, Multilayer Perceptron)
Data science studies all these, not in 587
All require a tradeoff:
  Very smooth (e.g. straight line): simple model, may not fit well, large error SS
  Very wiggly (e.g. connect-the-dots): complex model, fits very well, tiny error SS
    Likely to fit too well, bad predictions of new observations
All require choosing a "tuning parameter": balance between fit and complexity
All provide predictions of Y within the domain of X
  Any may provide "wild" predictions when asked to extrapolate
My go-to: splines
  Statistical theory to support a data-based tuning parameter
  R: mgcv library, `gam(Y ~ s(X1) + s(X2) + X3, family=binomial)`

Combining groups and continuous predictor variables (version 1, 2 groups)
  Define a new variable that indicates the group to which an observation belongs
  Have observations from men and women; sex variable has the values "man" or "woman"
    Define women = 0 when sex = 'man' and women = 1 when sex = 'woman'
    women is called an indicator variable: values of 0 or 1 indicate the group
  Two identical models:
    "T-test": $Y_{ij} = \mu_i + \varepsilon_{ij}$
    "regression on indicator variable": $Y_i = \beta_0 + \beta_1 \text{women}_i + \varepsilon_i$
  Predicted values from the regression:

Group	women	Regression mean	T-test mean
Man	0	β_0	μ_{man}
Woman	1	$\beta_0 + \beta_1$	μ_{women}

Notice that $\beta_1 = \mu_{women} - \mu_{man}$
  Think about how this relates to the definition of the slope
    and what increasing women by 1 "means"

Models with both groups (indicator variables) and continuous variables
  ANCOVA: analysis of covariance

$$Y_{ij} = \beta_0 + \beta_1\, group_i + \beta_2\, X_{ij} + \varepsilon_{ij}$$

  i indicates groups, j observation within group
  parallel lines
Heterogeneous regression lines

$$Y_{ij} = \beta_0 + \beta_1 women_i \beta_{2i}\, X_{ij} + \varepsilon_{ij}$$

each group (i) has a different slope   Pictures on the board for ANCOVA and heterogeneous regression lines models

Interaction:
All previous regression models have had additive effects
Example: model with sex (indicator for female) and age (continuous)
Additive model: difference (female - male) = sex effect same for all ages
plot of Y vs age has two parallel lines (same difference at all ages)
Interaction:
difference (female - male) depends on age, not constant
In general, effect of one X variable depends on level of a second
Heterogeneous regression lines have an interaction
Can be an interaction between
a grouping variable (e.g. sex) and a continuous one (e.g., age)
so slope relating Y to age is different for M and F
other examples are light/flowering time, bat echolocation
two continuous variables (e.g., litter size and body weight)
so slope relating brain size to litter size depends on body weight
two grouping variables (e.g., sex and ethnicity)
So difference between sexes, M-F, is not constant, depends on ethnicity

Connecting regression and ANOVA (version 2, any number of groups):
ANOVA model: $Y_{ij} = \mu_i + \varepsilon_{ij}$
When k groups, k μ_i parameters. e.g. 3 groups, 3 μ_i parameters
Indicator variable:

$$X = I(\text{something}) \text{ means } X = \begin{cases} 1 & \text{when something is true} \\ 0 & \text{when something is false} \end{cases}$$

So I(group = 'b') is 1 when the group = 'b' and 0 when the group = 'a' or 'c'
Define 3 indicator variables, one for each group:
$X_{1i} = I(\text{i'th obs has group ='a'})$,
$X_{2i} = I(\text{i'th obs has group ='b'})$,
$X_{3i} = I(\text{i'th obs has group ='c'})$
Fit the model $Y_i = \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \varepsilon i$ (Note: no β_0, so no intercept)

group	X_{1i}	X_{2i}	X_{3i}	predicted value
a	1	0	0	$\beta_1 = \mu_a$
b	0	1	0	$\beta_2 = \mu_b$
c	0	0	1	$\beta_3 = \mu_c$

Add an intercept to previous model
Write as a regression using a column of 1's for β_0
Model is $Y_i = \beta_0 X_{0i} + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \varepsilon i$

group	X_{0i}	X_{1i}	X_{2i}	X_{3i}	predicted value
a	1	1	0	0	$\beta_0 + \beta_1 = \mu_a$
b	1	0	1	0	$\beta_0 + \beta_2 = \mu_b$
c	1	0	0	1	$\beta_0 + \beta_3 = \mu_c$

Nasty numerical problem: $\boldsymbol{X}$ has 4 columns, but 1 is redundant
Choose any three, fourth can be computed from them. fourth is not new information.
Called a "non-full rank" $\boldsymbol{X}$ matrix
Can not use the matrix equation $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}$ because $\boldsymbol{X}'\boldsymbol{X}$ has no unique inverse.
Errors/warnings like "$\boldsymbol{X}'\boldsymbol{X}$ matrix is singular" are telling you this

Software "fix" the problem differently
R: Drop the column for first group (X_1). Remaining three are full rank.
Can tell R to use other approaches, see contrasts() documentation,
especially the information in the See Also section
SAS: uses methods for non-full rank matrices, equiv. to dropping last column
JMP: uses "effects" coding, +1, 0 or −1 and drops the last column
Can request indicator parameterization (drop last column)

R:

group	X_{0i}	X_{1i}	X_{2i}	X_{3i}	predicted value
a	1		0	0	$\beta_0 = \mu_a$
b	1		1	0	$\beta_0 + \beta_2 = \mu_b$
c	1		0	1	$\beta_0 + \beta_3 = \mu_c$

SAS:

group	X_{0i}	X_{1i}	X_{2i}	X_{3i}	predicted value
a	1	1	0		$\beta_0 + \beta_1 = \mu_a$
b	1	0	1		$\beta_0 + \beta_2 = \mu_b$
c	1	0	0		$\beta_0 = \mu_c$

JMP:

group	X_{0i}	X_{1i}	X_{2i}	X_{3i}	predicted value
a	1	1	0		$\beta_0 + \beta_1 = \mu_a$
b	1	0	1		$\beta_0 + \beta_2 = \mu_b$
c	1	-1	-1		$\beta_0 - \beta_1 - \beta_2 = \mu_c$

Problem: All β's have different estimates in R, SAS, or JMP !!
Example: 3 groups, means are $\overline{Y}_1 = 5$, $\overline{Y}_2 = 10$, $\overline{Y}_3 = 9$

Parameter	JMP	R	SAS
β_0	8	5	9
β_a	-3	–	-4
β_b	2	5	1
β_c	–	4	–

NOT GOOD. Estimates of β's depend on arbitrary choice of parameterization
My advice: don't look at estimates of β's in ANOVA models
In R, don't look at summary() output
unless you understand how to interpret the coefficients
SAS and JMP: don't show the estimates unless you specifically request them

Estimable functions:
Good news: some quantities, such as group means, difference in means,
are same for all three choices (JMP, R, or SAS)
Estimable function: an estimate that does not depend on arbitrary choices
Some estimable functions:
$\mu_a, \qquad \mu_a - \mu_b, \qquad \mu_a - (\mu_b + \mu_c)/2$
Some non-estimable functions:
$\beta_1, \qquad \mu_a - (\mu_b + \mu_c)$
If software tells you 'non-est', either
wrote the wrong quantity (bad contrast or estimate statement)
wrote the wrong model
or the data is insufficient to fit the model