ANOVA lack of fit test
  ANOVA (different mean for each unique X) always fits
  regression may or may not fit
  Construct ANOVA table with full = ANOVA, reduced = regression
  Requires multiple observations with same X values (so can fit ANOVA)

Computing ANOVA lack of fit test:
  Need to compare two models:
    Regression: regression model describes the means at each X
    Separate means: need to model a unique mean for each X
  Can fit each model (regression, ANOVA) to get SS Error and df error for each
    Hand compute F statistic
    Or: `anova(regression, sepmeans)` in R will compare the two
    JMP Fit Model gives you the Lack of Fit test automatically
      results box may be minimized, if so, click the grey triangle to open it
  Easier way to compute the lof test in R or SAS (also works in JMP, but not necessary):
    make a copy of the X variable, call it Xc and declare it a factor/class variable/red bar
    write the model as:
      R: `y.lof <- lm(Y ~ X + Xc, data= ...)`,
      SAS: `model Y = X Xc`,
      JMP: put X then Xc into model effects box
  Type I SS (and tests) are "sequential" SS:
    change in fit when add Xc to a model already containing X
  Type III SS (and tests) are "partial" SS:
    change in fit when add any term to model with everything else
    Will talk a lot more about the difference soon
  The ANOVA lack of fit test requires Type I SS = sequential SS and tests
  How to get from software: In all cases, look at the Xc results (the factor version)
    R: anova(y.lof) gives you sequential SS and tests
    SAS: gives you both Type I and Type III tests - look for the Type I box
    JMP: Effect tests box is Type III tests,
      red triangle / Estimates / Sequential Tests adds the Type I tests

This is the end of material on midterm II

Correlation:
  What should I do when X and Y are equivalent?
    Could swap without changing "meaning"
    Almost always observational data
  Correlation between X and Y
    unitless measure of association between X and Y

$$r = \frac{\Sigma(Y_i - \overline{Y})(X_i - \overline{X})}{(N-1)s_X s_Y}$$

    1 = perfect positive, 0 = no linear association, -1 = perfect negative

Can test $\rho = 0$ and construct confidence intervals for ρ - Beyond this course
Connection to regression slope

$$r = \hat{\beta}_1 \frac{s_x}{s_y}$$

Test of $\rho = 0$ gives same p-value as test of $\beta_1 = 0$
but adds another assumption: (X,Y) is a simple random sample of individuals

"R-squared": r^2
takes values from 0 to 1
1 = perfect linear association (+ or -) between two variables
Compute as correlation coefficient squared
Can compute from regression ANOVA table:

$$r^2 = 1 - \frac{\text{full SSE}}{\text{c.total SSE}}$$

often reported for regressions
and interpreted as a measure of "goodness" of the regression
I hate this
1) meat pH: correlation between time (not log time) and pH: $r = -0.966$
$r^2 = 0.933$ Very large. Stupid regression: not linear
2) based on sample but interpreted as population quantity
depends on sampling design - often not a simple random sample
Collect data over small range of $X \Rightarrow$ small R^2
Collect data over large range of $X \Rightarrow$ large R^2
Even though relationship between X and Y is identical
I suggest R^2 has no meaning unless you have a simple random sample of observations
Not just simple random sample of Y at chosen X's

Better measures of "goodness" of a regression: all my opinion
Why are you fitting a regression?
To estimate a slope: how precise is that slope? report se $\hat{\beta}_1$ or ci for β_1
To predict new observations: how precise are those predictions: report se $\hat{Y}_{obs}$ or s
Not clear: I would report s

Logistic Regression for responses that are yes/no (1/0):
Example: Donner party data, case study 20.1.1
87 (90?) people trying to get to CA in 1846, stuck by snow, 40 (42?) died
Are women more likely to survive stressful situations?
Data: 45 individuals (age $\geq$ 15), sex, age, survived (1) or not (0)
20 survived, 25 died
Goal: compare P[survival] between sexes when compared at same age
Simpler goal: Is age associated with P[surv]? If so, how?

Simple situation: yes/no response, 1 X variable
Want to model P[yes] as a function of the X variable

Y_i is 0 or 1, π_i is P[yes] for i'th obs, X_i is X value for i'th obs
P[yes] is an average, so could try $\pi_i = \beta_0 + \beta_1 X_i + \varepsilon_i$
Issue: (besides unequal variance, non-normal errors)
Predicted values can be < 0 or > 1: not good!
Logistic regression: model log odds as function of X_i

$$\text{logit}(\pi_i) = \log \frac{\pi_i}{1 - \pi_i} = \beta_0 + \beta_1 X_i$$

Observe Y_i which is 1 with probability π_i
If draw π_i vs X_i, curve is sigmoid
same shape as logistic population growth curve in ecology
β_1 is increase in log odds when X increases by 1
$\exp \beta_1$ is odds ratio comparing $X + 1$ to X
How to estimate the coefficients?
Can not use least squares (for continuous responses)
Use maximum likelihood

Likelihood:
Generalization of least squares to any statistical distribution
When $Y_i = 0$ or 1, Y_i has a Bernoulli distribution
Likelihood expresses how well a particular set of parameters fits the data
Maximum likelihood $\Rightarrow$ which β_0, β_1 fits best
Provides standard errors for estimates
which give tests and confidence intervals
Don't need to estimate pooled sd, so use normal distributions (Z scores)
Not T distributions Model comparison by comparing log likelihood for two models
Two names: likelihood ratio test and drop-in-deviance test
$-2(lnL_{reduced} - lnL_{full})$ has a known distribution when Ho true

Donner party, ignoring Sex: X_i = age of the individual
Shown graphically in the Donner age plots
Fitted equation: $\text{logit}(\pi_i) = 1.82 - 0.06647\text{Age}_i$
Interpretation of β_1:
Comparing two individuals differing in age by 1 year,
log odds of survival of the older one are 0.066 smaller
odds ratio is $\exp(-0.06647) = 0.936$
coefficient < 0, equivalently odds ratio < 1,
so probability of surviving decreases with age
Interpretation of β_0:
log odds of survival for an age 0 individual = 1.82
odds of survival for an age 0 individual = $\exp(1.82) = 6.17$
Probability of survival for an age 0 individual = $6.17/(1 + 6.17) = 0.86$

$$\pi = \frac{odds}{1 + odds}$$

But, age 0 is not relevant. Also an extrapolation.
Predicting survival probability for any age individual:
use fitted equation to get log odds,
then compute odds, then compute probability
Two individuals, age 50 and 51

Age	equation	log odds	odds	probability
51	$1.82 - 0.06647 \times 51$	-1.57	0.208	0.172
50	$1.82 - 0.06647 \times 50$	-1.50	0.222	0.182

Note: odds ratio = 0.208 / 0.222 = 0.936

Donner party: Is age associated with P[surv]?
Look at the estimated slope coefficient, $\hat{\beta}_1$
Estimate: -0.0665, se: 0.0322
Test method 1 (Wald test): Compute Z statistic

$$Z = \frac{estimate - parameter}{se}$$

Has an approximate standard normal distribution when H0 correct
Z = -2.063, p = 0.039
Test method 2 (likelihood ratio test):
Get the log likelihood (lnL) values for the full model and the reduced model
reduced model ($\beta_1 = 0$): lnL = -30.913, model has 1 parameter
full model ($\beta_1 \neq 0$): -28.145, model has 2 parameters
Calculate -2*change in lnL

$$C = -2(lnL_{reduced} - lnL_{full})$$

If Null hypothesis correct, $C \sim \chi^2$ with df = change in # parameters
$C = -2(-30.913 - -28.145) = -2 \times -2.768 = 5.54$
df = (2 - 1) = 1, p = 0.019
Deviance:
Book gives correct definition
Practical definition: $D = -2lnL$
So $C = D_{reduced} - D_{full}$
Two test methods give (slightly) different results
both are approximate
Prefer the LRT because it makes fewer assumptions
But often see the Wald test
Simpler to calculate when multiple parameters in the model
Differences usually not very large

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

Regression with count responses

Based on likelihood, but not Bernoulli distributions (0 or 1 on each individual)

Two types:

Fixed maximum: Binomial distribution

unlimited maximum: Poisson distribution

Example of fixed maximum: Modification of Vit C study

Response is whether or not you had a cold in Nov, Dec, Jan, Feb or Mar

Response is 0, 1, 2, 3, 4 or 5, (5 if had a cold in at each month)

Define π_i as probability have a cold in a month

Y_i is number of months for person i, has Binomial(5, π_i) distribution

possible events can be same or different for each individual

model log odds ($\log \pi_i/(1-\pi_i)$) as a function of the X variable

Example of unlimited maximum: another modification of the Vit C study

Response is number of colds you had during the winter season

No fixed limit, has values 0, 1, 2, $\cdots$, possibly large #

Define λ_i = average (or predicted) number of events for person i

λ_i can not be negative

So model $\log \lambda_i = \beta_0 + \beta_i$

Y_i is number of colds, has Poisson(λ_i) distribution

Fit either model by maximum likelihood

Overdispersion in models for Binomial or count data:
- Both Binomial and Poisson distributions: Var Y_i depends on mean Y_i
  - i.e., π_i or λ_i
- Sometimes the data are more variable than they "should" be
  - This is known as overdispersion
- Account for it by using a more complicated distribution for the data
  - Fixed maximum: Beta binomial distribution instead of Binomial
  - Unlimited maximum: Negative binomial distribution instead of Poisson
- My experience is that most ag/bio count data is overdispersed
  - Analysis of these data must account for overdispersion
  - Only exception is # bird eggs/clutch, which are less variable than expected