Linear regression: Interpretation of coefficients when one or both variables log transformed:
- Y is log transformed: $\log Y_i = \beta_0 + \beta_1 X + \varepsilon_i$
  - Similar to ANOVA on log Y Adding 1 to X adds β_1 to $\log Y$
  - So median Y multiplied by $\exp \beta_1$
- X is log transformed: $Y_i = \beta_0 + \beta_1 \log X_i + \varepsilon_i$
  - example: meat pH data: X is log hours.
  - Increasing X by $\log 2 \approx 0.693$ is a doubling of hours ($1 \to 2$ or $3 \to 6$).
  - So $\log 2 \times \beta_1 = 0.693 \times \beta_1$ is increase in mean Y when double the hours.
- Can have log-log regression: $\log Y_i = \beta_0 + \beta_1 \log X_i + \varepsilon_i$
  - Combine $\log X$ with $\log Y$: doubling X multiplies median Y by $\exp(0.693\beta_1)$

Estimating β_0 and β_1:
- Concept: find β_0 and β_1 so that predicted values are close to all observed values
- Define closeness by sum of squared residuals = SSE,
  - find $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimize SSE.

$$\begin{aligned} \hat{\beta}_1 &= \frac{\Sigma(X_i - \overline{X})Y_i}{\Sigma(X_i - \overline{X})^2} \\ \hat{\beta}_0 &= \overline{Y} - \hat{\beta}_1\overline{X} \end{aligned}$$

History:
- Procedure often called “least squares” or ordinary least squares (OLS)
  - Credited to Gauss (1795 or 1809) or Legendre (1805)
- Called regression because of Galton 1896
  - “Regression to mediocrity”: now called heritability,
  - but regression has stuck as the name for fitting Galton’s line

Connection to linear trend contrast:
- Linear regression estimated slope, fit to observations:

$$\hat{\beta}_1 = \frac{\Sigma(X_i - \overline{X})Y_i}{\Sigma(X_i - \overline{X})^2}$$

- Data in groups, calculate $\overline{Y}_{i.}$ for each unique X
- Fit regression to group means $(X_i, \overline{Y}_{i.})$

$$\hat{\beta}_1 = \frac{\Sigma(X_i - \overline{X})\overline{Y}_{i.}}{\Sigma(X_i - \overline{X})^2} = \Sigma\left(\frac{X_i - \overline{X}}{\Sigma(X_i - \overline{X})^2}\right)\overline{Y}_{i.}$$

- Linear trend contrast is the numerator of the slope estimate:

$$\hat{\beta}_1 = \Sigma(X_i - \overline{X})\overline{Y}_{i.}$$

- can get the slope as a contrast (by including the denominator)
- test of slope = 0 and test of linear trend contrast = 0 have the same numerator

have different se's because s^2 estimated differently
almost always very, very similar

Estimating error variance, s^2:
s is the sd of observations around the best fitting line
Assume straight line fits the data
residual $= Y_i - \hat{Y}_i$, where $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$
mean square error $= s^2 = \Sigma\, (Y_i - \hat{Y}_i)^2/$error df
error df: $N - 2$. Why 2? need to estimate 2 parameters, $\hat{\beta}_0$ and $\hat{\beta}_1$

Precision of estimates:
As expected, more obs increases precision but two other features
Slope:

$$\text{se } \hat{\beta}_1 = s\,\sqrt{\frac{1}{(N-1)s_X^2}}$$

s_X^2 is variance in X values. more spread out X's increase precision
Intercept:

$$\text{se } \hat{\beta}_0 = s\,\sqrt{\frac{1}{N} + \frac{\overline{X}^2}{(N-1)s_X^2}}$$

larger $\overline{X}$ decreases precision
If X's close to 0, intercept more precise
If X's a long way from $X = 0$, intercept less precise

Inference: (very familiar once have est. and se)
$(\hat{\beta} - \beta)/\text{se } \hat{\beta}$ has a T distribution with $N - 2$ df
You know how to construct tests and confidence intervals for individual parameters.

Useful tests:
$\beta_0 = 0$: not often useful
$\beta_1 = 0$: does mean Y change with X? Ho: no linear relationship
T test using $\hat{\beta}_1$
Test Ho: $\beta_1 = 0$ using model comparison. Two models:
full: $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$
reduced: $Y_i = \beta_0 + \varepsilon_i$ (same as equal means model)
Reject Ho when full fits much better than reduced, i.e., slope $\neq 0$
Can compute F statistic directly, or use an ANOVA table
Same p-value as T test, and $F = t^2$, since hypothesis has 1 df

Predictions at specific X values:
Could be X's used to fit regression or new X's
Two different types of predictions
Predicting mean Y at a specified X
Predicting individual Y for one observation at a specified X

Same predicted value, different uncertainty

Predicting mean Y: confidence interval for a predicted mean
If β_0, β_1 known, then prediction = $\beta_0 + \beta_1 X_0$
No uncertainty! because β_0, β_1 known
Estimate: $\hat{Y}_0 = \hat{\beta}_0 + \hat{\beta}_1 X_0$
Uncertain because of uncertainty in β_0, β_1

$$\text{se } \hat{Y}_0 = s\sqrt{\frac{1}{N} + \frac{(X_0 - \overline{X})^2}{(N-1)s_X^2}}$$

se formula demonstrates:
1) se $\hat{\beta}_0$ = se $\hat{Y}_0$ when $X_0 = 0$
2) se $\hat{Y}_0$ not constant. depends on X_0
smallest se when $X_0 = \overline{X}$, increases as move away from $\overline{X}$.

Predicting Y for one observation: prediction interval for a new observation
If β_0, β_1 known, then prediction = $\beta_0 + \beta_1 X_0$
This has uncertainty, because Y values are not on the line
Standard deviation of observations around the line is s
Estimate $\hat{Y}_{pred} = \hat{\beta}_0 + \hat{\beta}_1 X_0$
Predicted new observations have two sources of variability:
1) variability in the mean, se $\hat{Y}_0$
2) variability around the line, se $Y \mid \hat{Y}_0$
Add variances
1) has variance $s^2\left(\frac{1}{N} + \frac{(X_0-\overline{X})^2}{(N-1)s_X^2}\right)$ when doing SLR
2) has variance s^2
For SLR:

$$\text{se } \hat{Y}_{pred} = s\sqrt{1 + \frac{1}{N} + \frac{(X_0 - \overline{X})^2}{(N-1)s_X^2}}$$

In general (need se $\hat{Y}_0$ from computer):

$$\text{se } \hat{Y}_{pred} = \sqrt{\left(\text{se } \hat{Y}_0\right)^2 + s^2}$$

Calibration:
When does meat pH drop to 6.0?
Easy if Y = time, X = pH, $X_0 = 6.0$
Choice of Y and X matters.
All error variation in Y direction
X assumed known without error
Meat: time known exactly (set by experimenter) so X = time
Need to predict X_0 for specified Y_0
Known as the "calibration" problem

because calibration curves are a common application
X = known concentration, Y = measured signal,
want to predict concentration given a measurement
Prediction:

$$\hat{X}_0 = \frac{Y_0 - \hat{\beta}_0}{\hat{\beta}_1}$$

Precision: Approx. se $\hat{X}_0$ = (se $\hat{Y}_{obs}$)$/\hat{\beta}_1 \approx s/\hat{\beta}_1$
Confidence intervals and better se estimates can be computed
But beyond this course.

How I choose which is X and which is Y for a regression:
Experimental study: X is the manipulated variable, no choice
Observational study: 3 approaches
X is the antecedant concept; Y is the consequent concept
X is the more precisely measured variable
What do you want to predict? That's Y

Regression Assumptions:
Usual 3: independence, equal variances, normality
Plus: have correct model for the mean, "no lack of fit".
Importance: depends on goal, prediction interval is the most demanding

Assumption	estimates	tests	prediction interval
linearity	***	***	***
independence	ok	***	***
equal variance	ok	*	***
normality	ok	ok	***

Diagnoses:
plot of residuals vs predicted values
usual: no outliers, no trumpet
new: smile or frown ⇒ lack of fit
formal tests of lack of fit
Fit a more complicated model (e.g., $Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \varepsilon_i$)
When have > 1 obs at same X's, can fit regression or ANOVA
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

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