Linear regression:
  Evaluate relationships between two continuous variables
  Three major uses
    Describe / test relationships
    Predict Y at new X, also (sometimes) predict X at new Y
    Test hypotheses about predictions
  Will focus on Describe and Predict

Requires a model for how Y depends on X
  Simplest, but uninteresting model, $Y_i = \mu + \varepsilon_i$: constant mean
  Simplest non-trivial model: $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$
    Simple linear regression model
    Book writes as predicted (aka expected) value of Y given X: $\mu(Y \mid X) = \beta_0 + \beta_1 X$
    β_0 and β_1 are just symbols, could also write a line as $Y = a + bX$ or $Y = m\,X + b$.
    Statisticians prefer β's because extends easily to 2, 3, or many different X variables

Interpretation of coefficients:
  β_0: intercept, predicted Y when $X = 0$, same as estimated mean of Y when $X = 0$
  β_1: slope, estimated change in mean Y when X increases by 1
  peanuts: increase by 1 unit is not interpretable (data from 99.65 to 99.9)
    More general: increase X by ΔX, on average Y increases by $\Delta X \times \beta_1$
  meat: X is log hours.
    Increasing X by $\log 2 \approx 0.693$ is a doubling of hours ($1 \rightarrow 2$ or $3 \rightarrow 6$).
    So $\log 2 \times \beta_1 = 0.693 \times \beta_1$ is increase in mean Y when double the hours.

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 new X_0:
Two different quantities
Predicting mean Y at a specified X
Predicting individual Y for one observation at a specified X
Same prediction, different uncertainty

Predicting mean Y: confidence interval for a predicted Y
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
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
Estimate $\hat{Y}_{pred} = \hat{\beta}_0 + \hat{\beta}_1 X_0$
Has 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 $\hat{X}_0 = (\text{se } \hat{Y}_{obs})/\hat{\beta}_1 \approx s/\hat{\beta}_1$
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

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

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 $\Rightarrow$ 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)