Types of outcomes
  - Continuous, binary, counts, ...

Dependence structure of outcomes
  - Independent observations
  - Correlated observations, repeated measures

Number of covariates, potential confounders
  - Controlling for confounders that could lead to spurious results

Sample size

These factors will determine the appropriate statistical model to use
Regression

▶ Technique used to model and analyze data:
  ▶ e.g. data with variables for age, sex, weight, height, socio-economic status, and diet.
▶ Exploits the relationships between variables to gain information about one of them through knowing values of the others.
  ▶ e.g. the relationship between weight and diet.
▶ Regression can be used for estimation, hypothesis testing, prediction, and modeling causal relationships.
Can variation in outcome $Y$ be explained by correlation with a different variable?

---

**Plot of $Y$**

- $Y$ and $Z$ are uncorrelated
- $Y$ and $X$ have a linear relationship
Do Psychedelics Affect Math Performance?


- Group of volunteers was given LSD, their mean scores on math exam and tissue concentrations of LSD were obtained at n=7 time points.
Math exam score = $\alpha + \beta \times \text{LSD concentration}$

- $\alpha$ is the intercept, $\beta$ is the slope.
- $\epsilon_i =$ distance between the regression line and the observed value for the $i$th data point.
Least Squares Regression

- Intuitively, we want a line that is “close” to as many data points as possible. → Minimize the total distance between the regression line and the data points.
- Let \( \hat{y}_i = \alpha + \beta x_i \) describe the linear relationship between predicted value \( \hat{y}_i \) and \( x_i \).
- Let \( e_i = \hat{y}_i - y_i \) be the difference between the predicted and observed values for the \( i \)th data point.
- Minimize the objective function

\[
\phi(\alpha, \beta) = \sum_{i=1}^{N} e_i^2 = \sum_{i=1}^{N} (\hat{y}_i - y_i)^2 = \sum_{i=1}^{N} (\alpha + \beta x_i - y_i)^2.
\]

- The quantity \( \sum_{i=1}^{N} e_i^2 \) is often called Residual or Error Sum of Squares (\( SS_{res} \)).
Least Squares Regression

Solve for $\alpha$ and $\beta$ that minimize

$$\phi(\alpha, \beta) = \sum_{i=1}^{N} (\alpha + \beta x_i - y_i)^2.$$  

Calculus: take partial derivatives, set equal to zero and solve:

$$0 = \frac{\partial}{\partial \alpha} \phi = 2 \sum_{i=1}^{N} (\alpha + \beta x_i - y_i)$$

$$\Rightarrow \alpha = \bar{y} - \beta \bar{x}.$$  

$$0 = \frac{\partial}{\partial \beta} \phi = 2 \sum_{i=1}^{N} (\alpha + \beta x_i - y_i)x_i$$

$$= (\bar{y} - \beta \bar{x}) \sum_{i=1}^{N} x_i + \beta \sum_{i=1}^{N} x_i^2 - \sum_{i=1}^{N} x_i y_i$$

$$\Rightarrow \beta = \frac{\sum_{i=1}^{N} x_i y_i - \frac{1}{N} \left( \sum_{i=1}^{N} y_i \right) \left( \sum_{i=1}^{N} x_i \right)}{\sum_{i=1}^{N} x_i^2 - \frac{1}{N} \left( \sum_{i=1}^{N} x_i \right)^2}.$$
What can you do with a least squares regression line?

- A 1 unit increase in LSD concentration results in an average decrease of $\approx 9$ points on the math score.
- 70% is the predicted math score for having tissue concentration of 2.12 nanograms of LSD-25/mL.
- We have quantified the relation between the explanatory variable and the outcome.
Regression with more than one covariate

- An outcome rarely ever depends on just a single explanatory variable.
- Might be interested in the effect of multiple variables or want to control for the effect of a confounding variable.

Example:
- Height determined by sex, age, genetics.
- The effect of individual genetic variants on gene expression is of primary interest but need to account for confounders age, sex, batch, etc.

- Multiple linear regression models the mean outcome value on more than one explanatory variable.
- Provides the framework for interaction of explanatory variables.
Multiple linear regression

Linear regression with one continuous covariate $x_1$ and one dichotomous/binary covariate $x_2 \in \{0, 1\}$:

$$y_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2}.$$  

- The regression equation is two parallel lines, each with slope $\beta_1$.
- Data points with $x_2 = 0$ have intercept of $\alpha$.
- Data points with $x_2 = 1$ have intercept of $\alpha + \beta_2$.  

![Graphs showing two parallel lines with data points classified by $x_2$ values.]
Multiple linear regression

Linear regression with two continuous covariates $x_1$ and $x_2$:

$$y_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2}.$$ 

- Now the regression line is a plane through points in 3D space.
- Least squares distances are from the points to the surface of the plane.

![3D Scatterplot](http://www.statmethods.net/graphs/scatterplot.html)

*Plane with equation: $Y = \alpha + \beta_1 X_1 + \beta_2 X_2$*
Writing the least squares equations in matrix notation allows for derivation of parameters when there is more than one explanatory variable.

\[
y = \begin{bmatrix}
y_1 \\
y_2 \\
\vdots \\
y_N
\end{bmatrix}, \quad X = \begin{bmatrix}
1 & x_{11} & \ldots & x_{1p} \\
1 & x_{21} & \ldots & x_{2p} \\
\vdots & \vdots & \ddots & \vdots \\
1 & x_{N2} & \ldots & x_{Np}
\end{bmatrix}, \quad \beta = \begin{bmatrix}
\beta_0 \\
\beta_1 \\
\vdots \\
\beta_p
\end{bmatrix}, \quad \epsilon = \begin{bmatrix}
\epsilon_1 \\
\epsilon_2 \\
\vdots \\
\epsilon_N
\end{bmatrix}.
\]

Linear equation: \( y = X \beta + \epsilon. \)

- Matrix notation is much easier to code for computational estimation of regression parameters for arbitrary models.
- The same matrix algebra code works for 1 covariate or 100 covariates.
Residual sum of squares:

\[
S(\beta) = \sum_{i=1}^{N} \epsilon_i^2 = \epsilon^T \epsilon
\]

\[
= (y - X \beta)^T (y - X \beta)
\]

\[
= y^T y - 2\beta^T X^T y + \beta^T X^T X \beta.
\]

To minimize the \(SS_{res}\), take the derivative and set equal to 0:

\[
\frac{d}{d\beta} S = -2X^T y + 2X^T X \beta = 0.
\]

\[
\Rightarrow \hat{\beta} = \left(X^T X\right)^{-1}X^T y \text{ and } \hat{y} = X\hat{\beta}.
\]
Notice that we did not need a model to estimate the least squares regression line. It was simply geometry.

Least squares regression line allowed us to quantify the relationship.

But what if we want to make inference about the relationship?

That is, is the effect of LSD concentration on math test scores statistically significant?

**Statistical inference** is drawing conclusions about a population or parameter based on observed data.
We want to know if the variable $x$ has an effect on the outcome $Y$.

- In statistical terms: is $Y$ associated with $x$?

Linear model: $Y_i = \alpha + \beta x_i$.

- No association would correspond to the slope parameter $\beta = 0$ in the linear regression equation.
- The $\beta$ estimated from data is almost certainly no exactly zero.
- Is $\beta$ non-zero because of a true association between $x$ and $Y$, or simply because of unrelated variation in the outcome $Y$?
- We need a model to answer that question...
Define $Y_i = \alpha + \beta x_i + \epsilon_i$ to be the linear regression model that relates outcomes $Y$ to covariates $x$.

- $\epsilon_i$ is called the **residual**, and is the quantity by which the observed value differs from the proposed model.
- Assume that $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$, that is, the residuals are normally distributed with mean zero and constant variance.
- Based on properties of the normal distribution, for a fixed covariate value $x_i$, the corresponding outcome is distributed as $Y_i \sim \mathcal{N}(\alpha + \beta x_i, \sigma^2)$.
- Goal is to formally test if the slope parameter $\beta$ is zero, that is we want to test the null hypothesis $H_0 : \beta = 0$. 
Linear regression statistical model

**Linear regression:** mean response depending linearly on quantitative $x$

- $x_i$ is the independent variable, i.e. it is not random.
- $Y_i$ is the dependent variable, it is random: $Y_i \sim \mathcal{N}(\alpha + \beta x_i, \sigma^2)$:
  - The expected value of $Y$ is a linear function of $x$, but for fixed $x$, the variable $Y$ differs from its expected value by a random amount.
Given the Normal distribution assumption on the residuals, we can write a likelihood function for the observed data $Y_i$ as follows:

- If each $Y_i \sim N(\alpha + \beta x_i, \sigma^2)$, then the likelihood function of all the observed data is

$$L(\beta, \alpha | Y) = \prod_{i=1}^{N} L(\beta, \alpha | Y_i),$$

- and the log-likelihood function is

$$\ell(\beta, \alpha | Y) = \sum_{i=1}^{N} \log L(\beta, \alpha | Y_i).$$
For each $Y_i$, the log-likelihood function is based on the pdf of a Normal random variable:

$$
\mathcal{L}(\beta, \alpha | Y_i) = \frac{1}{\sqrt{2\pi \sigma^2}} \exp\left\{ -\frac{(Y_i - \alpha - \beta x_i)^2}{2\sigma^2} \right\}.
$$

Then the likelihood for the full data is

$$
\ell(\beta, \alpha | Y) = \sum_{i=1}^{N} -\frac{1}{2} \log(2\pi \sigma^2) - \frac{(Y_i - \alpha - \beta x_i)^2}{2\sigma^2}
$$

$$
= -\frac{N}{2} \log(2\pi \sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{N} (Y_i - \alpha - \beta x_i)^2.
$$
Linear regression – likelihood approach

Given the log-likelihood function, we can compute the Maximum Likelihood Estimators (MLEs) for the parameters $\alpha$, $\beta$ and $\sigma^2$. Note that maximizing the log-likelihood

$$
\ell(\beta, \alpha|Y) = -\frac{N}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{N} (Y_i - \alpha - \beta x_i)^2
$$

is equivalent to minimizing

$$
\sum_{i=1}^{N} (Y_i - \alpha - \beta x_i)^2.
$$

This is exactly the least squares regression problem!

- The least squares regression solutions are the same as the MLEs.
- Maximum likelihood estimation has several nice large sample properties that accommodate hypothesis testing.
Analysis of Variance (ANOVA) partitions total variation in the observed outcomes into variation explained by a model and variation explained by random error.

For each observation, $Y_i - \bar{Y} = (\hat{Y}_i - \bar{Y}) + (Y_i - \hat{Y}_i)$.

Squaring, summing, and some algebra gives:

$$\sum_{i=1}^{N} (Y_i - \bar{Y})^2 = \sum_{i=1}^{N} (\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^{N} (Y_i - \hat{Y}_i)^2$$

$$SS_T = SS_R + SS_{res}$$

where $SS_T$ is the Total Sum of Squares, $SS_R$ is the Regression Sum of Squares, and $SS_{res}$ is the Residual Sum of Squares.
ANOVA table for testing the null hypothesis $H_0 : \beta = 0$

<table>
<thead>
<tr>
<th>Source of Variation</th>
<th>Sum of Squares</th>
<th>Degrees of Freedom</th>
<th>Mean Square</th>
<th>$F_0 \sim F_{1, N-2}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>$S_R$</td>
<td>1</td>
<td>$SS_R/1$</td>
<td>$SS_R/1$ ( SS_{res}/(N-2) )</td>
</tr>
<tr>
<td>Residual</td>
<td>$SS_{res}$</td>
<td>$N - 2$</td>
<td>$SS_{res}/(N - 2)$</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>$SS_T$</td>
<td>$N - 1$</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
ANOVA – graphical interpretation

\[ \hat{y} = \hat{\alpha} + \hat{\beta} x \]

**Residual:** \( y_i - \hat{y}_i \)

**Regression:** \( \hat{y}_i - \overline{y} \)

\[ SS_T = \sum (y_i - \overline{y})^2 \]
\[ SS_R = \sum (\overline{y} - \hat{y}_i)^2 \]
\[ SS_{RES} = \sum (y_i - \hat{y}_i)^2 \]
ANOVA – graphical interpretation

Y and Z are uncorrelated

Y and X are correlated
ANOVA – graphical interpretation

**Y and Z are uncorrelated**

\[ Y = 19 + (-0.7)Z \]

**Y and X are correlated**

\[ Y = 1.2 + (0.68)X \]
ANOVA – graphical interpretation

Y and Z are uncorrelated

Y and X are correlated

<table>
<thead>
<tr>
<th>Source</th>
<th>Sum of Squares</th>
<th>$F_0$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$SS_R$</td>
<td>42.47</td>
<td>0.36</td>
</tr>
<tr>
<td>$SS_{res}$</td>
<td>$1.2 \times 10^4$</td>
<td></td>
</tr>
<tr>
<td>$SS_T$</td>
<td>$1.2 \times 10^4$</td>
<td>$p = 0.55$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Source</th>
<th>Sum of Squares</th>
<th>$F_0$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$SS_R$</td>
<td>$10^4$</td>
<td>890.54</td>
</tr>
<tr>
<td>$SS_{res}$</td>
<td>1148.1</td>
<td></td>
</tr>
<tr>
<td>$SS_T$</td>
<td>$1.2 \times 10^4$</td>
<td>$p = 6 \times 10^{-51}$</td>
</tr>
</tbody>
</table>
Asumptions of linear regression

We have made multiple assumptions about the outcomes along the way.

▶ What exactly are they?
▶ How do we check them?
▶ What if they are violated?
  ▶ Inference can be incorrect.
  ▶ Extensions to simple linear regression

**Assumptions:**

1. Independence between observations.
2. Linearity between covariate and outcome.
3. Constant variance of errors (homoscedasticity).
4. Normally distributed errors.
Checking linearity assumption

- We are assuming a linear relationship between the outcome and covariate.
- Residual plot (residuals vs $X$) should look like a random cloud of points around the line $Y = 0$; Any pattern may indicate non-linearity.
- Try adding non-linear terms of the covariate (e.g. $x^2$, $\log x$, ...)

![Linear model fit on data with non-linear relation](image1)

![Residual plot shows clear patterns](image2)
Checking normality assumption

- We assume that the residuals are normally distributed.
- Can check normality assumption using a quantile-quantile plot of the residuals (QQ plot).
- A QQ plot is an ordered set of residuals plotted against the quantiles of a theoretical distribution.
- If the points follow the theoretical distribution, the points should form a line along $y = x$.
- A transformation of the outcome can often correct violations of the normality assumption (e.g. $\log Y$).
Relationship between $X$ and $Y$ is exponential

QQ plot from $Y = a + bX$

Log transform of $Y$ values creates linear relation

QQ plot from $\log(Y) = a + bX$
Checking constant variance assumption

- We assume that the residuals have constant variance across values of $X$.
- Look at plots of residuals versus predicted values, and plot of residuals vs. the covariate $X$.
- Use either a variable transformation or **weighted least squares regression** to fix.
Checking independence assumption

- We assume that the outcomes are independent.
- Usually this is determined by the experimental design:
  - Are measurements taken from the same people?
  - Are samples recruited from different hospitals? Are measurements from the same hospital more similar than measurements from different hospitals?
- Ignoring correlation between observations can lead to improper inference (your estimate of variance is wrong)
- **Linear mixed models** and **generalized estimating equations** allow you to take into account the correlation of the outcomes.
Summary

- Reviewed the basics of simple linear regression.

Hopefully, this lecture

- Improved your intuition on the theory behind simple linear regression.

- Made connections to ANOVA, likelihood theory, extensions to simple linear regression.