Hypothesis Testing

Terminology: “tails” -- is the rejection region made up of one or two sides of the rejection region?

Example: Comparing two means:

- **two-tailed p-value**: $P(T > |t| \text{ or } T < -|t|) = 2*P(T > |t|)$?
  (when there is no assumption of direction of difference)
- **one-tailed p-value**: $P(T > t)$? (when $H_a$ posits the second mean is greater)
  $P(T < t)$? (when $H_a$ posits the second mean is less)
Resampling Techniques

“nonparametric” tests

The permutation test:

- $t_{\text{obs}}$ = Compute observed score
- passes = 0
- for 1 to $B$:
  - randomly permute the data, keeping the same sizes per class
  - $t_B$ = compute score on permuted data
  - if $t_B >$ (or $<$) $t_{\text{obs}}$: passes+=1
- $p\_\text{value} = \frac{\text{passes}}{B}$

Application: comparing two distributions, especially when they are unknown.
Linear Regression

Finding a linear function based on $X$ to best yield $Y$.

$X$ = “covariate” = “feature” = “predictor” = “regressor” = “independent variable”

$Y$ = “response variable” = “outcome” = “dependent variable”

Regression: $ r(x) = \mathbb{E}(Y \mid X = x)$

goal: estimate the function $r$
Linear Regression

Finding a linear function based on $X$ to best yield $Y$.

$X = \text{“covariate” = “feature” = “predictor” = “regressor” = “independent variable”}$

$Y = \text{“response variable” = “outcome” = “dependent variable”}$

Regression: $r(x) = \mathbb{E}(Y | X = x)$

\text{goal: estimate the function $r$}

Linear Regression (univariate version): $r(x) = \beta_0 + \beta_1 x$

\text{goal: find $\beta_0, \beta_1$ such that } r(x) \approx \mathbb{E}(Y | X = x)$
Linear Regression

Simple Linear Regression

\[ Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \]

where \( \mathbb{E}(\epsilon_i|X_i) = 0 \) and \( \mathbb{V}(\epsilon_i|X_i) = \sigma^2 \)

\[ r(x) = \beta_0 + \beta_1 x \]
Linear Regression

Simple Linear Regression

\[ Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \]

where \( E(\epsilon_i | X_i) = 0 \) and \( V(\epsilon_i | X_i) = \sigma^2 \)

- intercept
- slope
- error
- expected variance
Simple Linear Regression

\[ Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \]

where \( E(\epsilon_i | X_i) = 0 \) and \( V(\epsilon_i | X_i) = \sigma^2 \)

Estimated intercept and slope:

\[ \hat{\beta}(x) = \hat{\beta}_0 + \hat{\beta}_1 x \]

\[ \hat{Y}_i = \hat{\beta}(X_i) \]

Residual:

\[ \hat{\epsilon}_i = Y_i - \hat{Y}_i \]
Linear Regression

Simple Linear Regression

\[ Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \]

where \( E(\epsilon_i | X_i) = 0 \) and \( V(\epsilon_i | X_i) = \sigma^2 \)

Estimated intercept and slope:

\[ \hat{r}(x) = \hat{\beta}_0 + \hat{\beta}_1 x \]

\[ \hat{Y}_i = \hat{r}(X_i) \]

Residual:

\[ \hat{\epsilon}_i = Y_i - \hat{Y}_i \]

Least Squares Estimate. Find \( \hat{\beta}_0 \) and \( \hat{\beta}_1 \) which minimizes the residual sum of squares:

\[ RSS = \sum_{i=1}^{n} \hat{\epsilon}_i^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2 \]
Linear Regression

via Gradient Descent

Start with $\hat{\beta}_0 = \hat{\beta}_1 = 0$

Repeat until convergence:

Calculate all $\hat{Y}_i$

\[ \hat{\beta}_0 = \hat{\beta}_0 - \alpha \left( \sum_{i=1}^{n} \hat{Y}_i - Y_i \right) \]

\[ \hat{\beta}_1 = \hat{\beta}_1 - \alpha \left( \sum_{i=1}^{n} X_i (\hat{Y}_i - Y_i) \right) \]

Least Squares Estimate. Find $\hat{\beta}_0$ and $\hat{\beta}_1$ which minimizes the residual sum of squares:

\[ RSS = \sum_{i=1}^{n} \hat{e}_i^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2 \]
Linear Regression

**Least Squares Estimate.** Find $\hat{\beta}_0$ and $\hat{\beta}_1$ which minimizes the residual sum of squares:

$$RSS = \sum_{i=1}^{n} \hat{\epsilon}_i^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2$$

**via Gradient Descent**

Start with $\hat{\beta}_0 = \hat{\beta}_1 = 0$

Repeat until convergence:

Calculate all $\hat{Y}_i$

$$\hat{\beta}_0 = \hat{\beta}_0 - \alpha \left( \sum_{i=1}^{n} \hat{Y}_i - Y_i \right)$$

$$\hat{\beta}_1 = \hat{\beta}_1 - \alpha \left( \sum_{i=1}^{n} X_i (\hat{Y}_i - Y_i) \right)$$

Learning rate

Based on derivative of $RSS$
Linear Regression

via Gradient Descent

Start with $\hat{\beta}_0 = \hat{\beta}_1 = 0$

Repeat until convergence:
- Calculate all $\hat{Y}_i$

$$\hat{\beta}_0 = \hat{\beta}_0 - \alpha \left( \sum_{i=1}^{n} \hat{Y}_i - Y_i \right)$$

$$\hat{\beta}_1 = \hat{\beta}_1 - \alpha \left( \sum_{i=1}^{n} X_i (\hat{Y}_i - Y_i) \right)$$

Least Squares Estimate. Find $\hat{\beta}_0$ and $\hat{\beta}_1$ which minimizes the residual sum of squares:

$$RSS = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2$$

via Direct Estimates (normal equations)

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (X_i - \bar{X})^2}$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$
Pearson Product-Moment Correlation

Covariance

\[ Cov(X, Y) = E(XY) - E(X)E(Y) = E((X - \bar{X})(Y - \bar{Y})) \]

via Direct Estimates (normal equations)

\[
\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}
\]

\[
\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1\bar{X}
\]
Pearson Product-Moment Correlation

**Covariance**

\[
Cov(X, Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y) = \mathbb{E}((X - \bar{X})(Y - \bar{Y}))
\]

**Correlation**

\[
r = r_{X,Y} = \frac{Cov(X, Y)}{s_X s_Y} = \frac{1}{n-1} \sum_{i=1}^{n} \left( \frac{X_i - \bar{X}}{s_X} \right) \left( \frac{Y_i - \bar{Y}}{s_Y} \right)
\]

via Direct Estimates (normal equations)

\[
\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}
\]

\[
\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}
\]
Pearson Product-Moment Correlation

Covariance

\[ Cov(X, Y) = E(XY) - E(X)E(Y) = E((X - \bar{X})(Y - \bar{Y})) \]

Correlation

\[ r = r_{X,Y} = \frac{Cov(X, Y)}{s_X s_Y} = \frac{1}{n-1} \sum_{i=1}^{n} \left( \frac{X_i - \bar{X}}{s_X} \right) \left( \frac{Y_i - \bar{Y}}{s_Y} \right) \]

via Direct Estimates (normal equations)

\[ \hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \]

\[ \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} \]

If one standardizes X and Y (i.e. subtract the mean and divide by the standard deviation) before running linear regression, then:

\[ \hat{\beta}_0 = 0 \quad \text{and} \quad \hat{\beta}_1 = r \]
Multiple Linear Regression

Suppose we have multiple independent variables that we’d like to fit to our dependent variable: \( Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_m X_{im} + \epsilon_i \)

If we include an \( X_{oi} = 1 \) for all \( i \) (i.e. adding the intercept to \( X \)). Then we can say:

\[
Y_i = \sum_{j=0}^{m} \beta_j X_{ij} + \epsilon_i
\]
Multiple Linear Regression

Suppose we have multiple independent variables that we’d like to fit to our dependent variable:  
\[ Y_i = \beta_0 + beta_1 X_{i1} + beta_2 X_{i2} + \ldots + beta_m X_{m1} + \epsilon_i \]

If we include and \( X_{oi} = 1 \) for all \( i \). Then we can say:

\[
Y_i = \sum_{j=0}^{m} \beta_j X_{ij} + \epsilon_i
\]

Or in vector notation across all \( i \):

\[
Y = X\beta + \epsilon
\]

Where \( \beta \) and \( \epsilon \) are vectors and \( X \) is a matrix.
Multiple Linear Regression

Suppose we have multiple independent variables that we’d like to fit to our dependent variable:

\[ Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_m X_{mi} + \epsilon_i \]

If we include and \( X_{oi} = 1 \) for all \( i \). Then we can say:

\[ Y_i = \sum_{j=0}^{m} \beta_j X_{ij} + \epsilon_i \]

Or in vector notation across all \( i \):

\[ Y = X \beta + \epsilon \]

Where \( \beta \) and \( \epsilon \) are vectors and \( X \) is a matrix.

Estimating \( \beta \):

\[ \hat{\beta} = (X^T X)^{-1} X^T Y \]
Multiple Linear Regression

Suppose we have multiple independent variables that we’d like to fit to our dependent variable: 

\[ Y_i = \beta_0 + \text{beta}_1 X_{i1} + \text{beta}_2 X_{i2} + \ldots + \text{beta}_m X_{i1} + \epsilon_i \]

If we include and \( X_{oi} = 1 \) for all \( i \). Then we can say:

\[ Y_i = \sum_{j=0}^{m} \beta_j X_{ij} + \epsilon_i \]

Or in vector notation across all \( i \):

\[ Y = X\beta + \epsilon \]

Where \( \beta \) and \( \epsilon \) are vectors and \( X \) is a matrix.

To test for significance of individual Coefficient, \( j \):

\[ t = \frac{\hat{\beta}_j}{se(\hat{\beta}_j)} = \frac{\hat{\beta}_j}{\sqrt{\frac{s_j^2}{\sum_{i=1}^{n} X_{ij} - X}}} \]

Estimating \( \beta \):

\[ \hat{\beta} = (X^T X)^{-1} X^T Y \]
Logistic Regression

What if $Y_i \in \{0, 1\}$? (i.e. we want “classification”)

$$p_i \equiv p_i(\beta) \equiv P(Y_i = 1|X = x) = \frac{e^{\beta_0 + \sum_{j=1}^{m} \beta_j x_{ij}}}{1 + e^{\beta_0 + \sum_{j=1}^{m} \beta_j x_{ij}}}$$
Logistic Regression

What if $Y_i \in \{0, 1\}$? (i.e. we want “classification”)

$$p_i \equiv p_i(\beta) \equiv P(Y_i = 1|X = x) = \frac{e^{\beta_0 + \sum_{j=1}^{m} \beta_j x_{ij}}}{1 + e^{\beta_0 + \sum_{j=1}^{m} \beta_j x_{ij}}}$$

$$\log \left( \frac{p}{1 - p} \right) = \beta_0 + \sum_{j=1}^{m} \beta_j x_{ij}$$
Logistic Regression

What if $Y_i \in \{0, 1\}$? (i.e. we want “classification”)

$$p_i \equiv p_i(\beta) \equiv P(Y_i = 1 | X = x) = \frac{e^{\beta_0 + \sum_{j=1}^{m} \beta_j x_{ij}}}{1 + e^{\beta_0 + \sum_{j=1}^{m} \beta_j x_{ij}}}$$

$$\log \left( \frac{p}{1 - p} \right) = \beta_0 + \sum_{j=1}^{m} \beta_j x_{ij}$$

To estimate $\beta$, one can use reweighted least squares:

1. Set $Z_i = \log \left( \frac{p}{1 - p} \right) + \frac{Y_i - p_i}{p_i(1 - p_i)}$, for $i = 1 \ldots n$
2. Let $W$ be a diagonal matrix, with $(i, i)$ equal to $p_i(1 - p_i)$
3. Set $\hat{\beta} = (X^T W X)^{-1} X^T W Y$
4. Repeat from 1 until $\hat{\beta}$ converges
Logistic Regression

What if $Y_i \in \{0, 1\}$? (i.e. we want “classification”)

$$p_i \equiv p_i(\beta) \equiv P(Y_i = 1 | X = x) = \frac{e^{\beta_0 + \sum_{j=1}^{m} \beta_j x_{ij}}}{1 + e^{\beta_0 + \sum_{j=1}^{m} \beta_j x_{ij}}}$$

$$\text{logit}(p) \quad \log \left( \frac{p}{1-p} \right) = \beta_0 + \sum_{j=1}^{m} \beta_j x_{ij}$$

To estimate $\beta$, one can use reweighted least squares:

1. Set $\beta_0...\beta_m = 0$
2. Let $W$ be a diagonal matrix, with $(i, i)$ equal to $p_i(1 - p_i)$
3. Set $\hat{\beta} = (X^TWX)^{-1}X^TWY$
4. Repeat from 1 until $\hat{\beta}$ converges