# Linear Regression

Dustin Pluta

2018 Statistics Bootcamp
Department of Statistics
University of California, Irvine

# Linear Regression I

- $Y$ is an $n \times 1$ response vector
- $X$ is an $n \times p$ data matrix with full column rank
- $\beta$ is a $p \times 1$ coefficient vector
- $\varepsilon$ an $n \times 1$ random error vector with mean 0 and finite variance

The **linear regression** model writes $Y$ as the sum of a *systematic* component $X\beta$, and a *stochastic* component $\varepsilon$:

$$Y = X\beta + \varepsilon.$$

# Linear Regression I

- If we assume $\varepsilon \sim N(0, \sigma^2 I_n)$, the induced distribution on $Y$ is

$$Y \sim N(X\beta, \sigma^2 I_n),$$

  that is, $Y$ follows a multivariate normal distribution with mean vector $X\beta$ and variance $\sigma^2 I_n$.
- This linear regression model makes four assumptions:
  - Linearity: $\mathbb{E}Y$ can be expressed as a linear combination of the features in $X$;
  - Independence of errors across observations;
  - Normally distributed error terms;
  - Homogenous (constant) variance of error terms.

# Linear Regression Likelihood

The **likelihood** for the linear regression model is

$$\mathcal{L}(\beta, \sigma^2 | X, Y) = \left(2\pi\sigma^2\right)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}(Y - X\beta)^T(Y - X\beta)\right\}$$

The **log-likelihood** is

$$\ell(\beta, \sigma^2) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(Y - X\beta)^T(Y - X\beta).$$

# MLE of Parameters

$$\frac{\partial \ell}{\partial \beta} = -\frac{1}{\sigma^2} X^T (Y - X\beta)$$
$$\frac{\partial \ell}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2} (Y - X\beta)^T (Y - X\beta)$$

Setting these to 0 and solving yields the MLE estimates for the regression parameters

## MLE of Parameters

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$
$$\hat{\sigma}^2 = \frac{1}{n} (Y - X\beta)^T (Y - X\beta)$$

# Unbiased estimator of $\sigma^2$

Since the bias of $\hat{\sigma}^2$ as an estimator of $\sigma^2$ grows with the number of covariates $p$, it is important to instead use the unbiased variance estimator

$$s^2 = \frac{1}{n-p}(Y - X\beta)^T(Y - X\beta).$$

# $\hat{\beta}$ Bias and Variance

- $\hat{\beta}$ is unbiased:

$$\begin{aligned}\mathbb{E}\left[\hat{\beta}\right] &= (X^TX)^{-1}X^T\mathbb{E}[Y] \\ &= (X^TX)^{-1}X^TX\beta = \beta.\end{aligned}$$

- To find the variance, we need to use the fact that for $A$ an $n \times q$ matrix, and an $n \times 1$ random vector $Y$, we have $Var(A^TY) = A^TVar(Y)A$. Applying this to the equation for $\hat{\beta}$ gives

$$\begin{aligned}Var(\hat{\beta}) &= Var\left((X^TX)^{-1}X^TY\right) \\ &= (X^TX)^{-1}X^TVar(Y)X(X^TX)^{-1} \\ &= \sigma^2(X^TX)^{-1}\end{aligned}$$

# Distribution of $\hat{\beta}$

Since $\hat{\beta}$ is a linear transformation of the normal random vector $Y$, we know $\hat{\beta}$ is normally distributed, with the mean and variance we just computed:

## Distribution of $\hat{\beta}$

$$\hat{\beta} \sim N\left(\beta, \sigma^2(X^T X)^{-1}\right).$$

In practice, we usually replace $\sigma^2$ with $s^2$, and take the distribution as approximate

## Approximate Distribution of $\hat{\beta}$

$$\hat{\beta} \dot{\sim} N\left(\beta, s^2(X^T X)^{-1}\right).$$

# Distribution of $\hat{\beta}_k$

## Distribution of $\hat{\beta}$

$$\frac{\hat{\beta}_k - \beta_k}{\hat{\text{se}}(\hat{\beta}_k)} \sim t\left(n-p\right),$$

where $\hat{\text{se}}(\hat{\beta}_k) = \sqrt{e_k^T s^2 (X^T X)^{-1} e_k}$, with $e_k$ is a $p$-length vector with a 1 as the $k$th element and 0 elsewhere.

# Hypothesis Tests

Consider testing the hypothesis

$$\begin{cases} H_0 & \beta = \beta^0 \\ H_1 & \beta \neq \beta^0. \end{cases}$$

For this *global* test of any association we have many possible tests to choose from, for example the **Wald test** or **general** $F$**-test**.

# Wald Test

## Test Statistic

$$
\begin{aligned}
T &= (\hat{\beta} - \beta^0)^T Var(\hat{\beta})^{-1} (\hat{\beta} - \beta^0) \\
&= \frac{1}{s^2} (\hat{\beta} - \beta^0)^T (X^T X) (\hat{\beta} - \beta^0) \\
&\overset{H_0}{\sim} \chi^2_{n-p}
\end{aligned}
$$

# Wald Test

We can also use the Wald test for testing single coefficients, $H_0 : \beta_k = \beta_k^0$.

## Test Statistic

$$T = \frac{\hat{\beta}_k - \beta_k^0}{\hat{\mathsf{se}}(\hat{\beta}_k)} \overset{H_0}{\sim} t_{n-p}$$

# Wald Test

We can *also* use the Wald test for testing arbitrary linear combinations of coefficients, $H_0 : c^T\beta = c^T\beta^0$ for a $p \times 1$ contrast vector $c$.

## Test Statistic

$$\begin{aligned} T &= (c^T\hat{\beta} - c^T\beta^0)^T Var(c^T\hat{\beta})^{-1}(c^T\hat{\beta} - c^T\beta^0) \\ &= (c^T\hat{\beta} - c^T\beta^0)^T [c^T Var(\hat{\beta})c]^{-1}(c^T\hat{\beta} - c^T\beta^0) \\ &= \frac{1}{s^2}(c^T\hat{\beta} - c^T\beta^0)^T [c^T(X^TX)^{-1}c]^{-1}(c^T\hat{\beta} - c^T\beta^0) \\ &\overset{H_0}{\dot{\sim}} \chi^2_{n-p} \end{aligned}$$

# General $F$-test

Alternatively, we can test for general linear association of $X$ and $Y$ with the general $F$-test:

**Test Statistic**

$$F = \frac{[(\hat{Y} - \bar{Y})^T(\hat{Y} - \bar{Y})]/(p-1)}{(Y - \hat{Y})^T(Y - \hat{Y})/(n-p)} \sim F(p-1, n-p)$$

This is often used in ANOVA.

# Confidence Intervals

**$100(1-\alpha)$% Confidence Interval for $\beta_k$**

$$\hat{\beta}_k \pm \hat{\mathsf{se}}(\hat{\beta}_k) t_{1-\alpha/2}(n-p)$$

# Confidence Intervals

**Approximate $100(1-\alpha)\%$ Confidence Interval for $c^T\beta_k$**

$$c^T\hat{\beta} \pm z_{1-\alpha/2}\sqrt{c^T Var(\hat{\beta})c}$$

# Interpretation

Let's examine the expression for $\hat{\beta}$:

**MLE of Coefficient Vector**

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$