Regression Introduction and Estimation Review

Dr. Frank Wood
Quick Example - Scatter Plot

Use *linear_regression/demo.m*
Linear Regression

- Want to find parameters for a function of the form

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

- Distribution of error random variable not specified
Quick Example - Scatter Plot
Formal Statement of Model

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

- \( Y_i \) value of the response variable in the \( i^{th} \) trial
- \( \beta_0 \) and \( \beta_1 \) are parameters
- \( X_i \) is a known constant, the value of the predictor variable in the \( i^{th} \) trial
- \( \epsilon_i \) is a random error term with mean \( E\{\epsilon_i\} = 0 \) and finite variance \( \sigma^2\{\epsilon_i\} = \sigma^2 \)
- \( i = 1, \ldots, n \)
Properties

- The response $Y_i$ is the sum of two components
  - Constant term $\beta_0 + \beta_1 X_i$
  - Random term $\epsilon_i$

- The expected response is

\[
E\{Y_i\} = E\{\beta_0 + \beta_1 X_i + \epsilon_i\} \\
= \beta_0 + \beta_1 X_i + E\{\epsilon_i\} \\
= \beta_0 + \beta_1 X_i
\]
Expectation Review

▶ Definition

\[ E\{X\} = E\{X\} = \int XP(X)dX, \ X \in \mathcal{R} \]

▶ Linearity property

\[ E\{aX\} = aE\{X\} \]
\[ E\{aX + bY\} = aE\{X\} + bE\{Y\} \]

▶ Obvious from definition
Example Expectation Derivation

\[ P(X) = 2X, \ 0 \leq X \leq 1 \]

Expectation

\[
E\{X\} = \int_0^1 XP(X)\,dX
\]

\[
= \int_0^1 2X^2\,dX
\]

\[
= \left. \frac{2X^3}{3} \right|_0^1
\]

\[
= \frac{2}{3}
\]
Expectation of a Product of Random Variables

If \( X, Y \) are random variables with joint distribution \( P(X, Y) \) then the expectation of the product is given by

\[
E\{XY\} = \int_{XY} XYP(X, Y)\,dX\,dY.
\]
Expectation of a product of random variables

What if $X$ and $Y$ are independent? If $X$ and $Y$ are independent with density functions $f$ and $g$ respectively then

$$E\{XY\} = \int_{XY} XYf(X)g(Y)\,dX\,dY$$

$$= \int_X \int_Y XYf(X)g(Y)\,dX\,dY$$

$$= \int_X Xf(X)\left[\int_Y Yg(Y)\,dY\right]\,dX$$

$$= \int_X Xf(X)E\{Y\}\,dX$$

$$= E\{X\}E\{Y\}$$
Regression Function

- The response $Y_i$ comes from a probability distribution with mean

$$E\{Y_i\} = \beta_0 + \beta_1 X_i$$

- This means the regression function is

$$E\{Y\} = \beta_0 + \beta_1 X$$

Since the regression function relates the means of the probability distributions of $Y$ for a given $X$ to the level of $X$. 
Error Terms

- The response $Y_i$ in the $i^{th}$ trial exceeds or falls short of the value of the regression function by the error term amount $\epsilon_i$.
- The error terms $\epsilon_i$ are assumed to have constant variance $\sigma^2$. 
Response Variance

Responses $Y_i$ have the same constant variance

\[
\sigma^2\{Y_i\} = \sigma^2\{\beta_0 + \beta_1X_i + \epsilon_i\} \\
= \sigma^2\{\epsilon_i\} \\
= \sigma^2
\]
Variance ($2^{nd}$ central moment) Review

- **Continuous distribution**

\[ \sigma^2\{X\} = E\{(X - E\{X\})^2\} = \int (X - E\{X\})^2 P(X) dX, \ X \in \mathbb{R} \]

- **Discrete distribution**

\[ \sigma^2\{X\} = E\{(X - E\{X\})^2\} = \sum_i (X_i - E\{X\})^2 P(X_i), \ X \in \mathbb{Z} \]
Alternative Form for Variance

\[ \sigma^2\{X\} = E\{(X - E\{X\})^2\} \]
\[ = E\{(X^2 - 2XE\{X\} + E\{X\}^2)\} \]
\[ = E\{X^2\} - 2E\{X\}E\{X\} + E\{X\}^2 \]
\[ = E\{X^2\} - 2E\{X\}^2 + E\{X\}^2 \]
\[ = E\{X^2\} - E\{X\}^2. \]
Example Variance Derivation

\[ P(X) = 2X, \, 0 \leq X \leq 1 \]

\[ \sigma^2\{X\} = E\{(X - E\{X\})^2\} = E\{X^2\} - E\{X\}^2 \]

\[ = \int_0^1 2XX^2 \, dX - \left(\frac{2}{3}\right)^2 \]

\[ = \frac{2X^4}{4}\bigg|_0^1 - \frac{4}{9} \]

\[ = \frac{1}{2} - \frac{4}{9} = \frac{1}{18} \]
Variance Properties

\[ \sigma^2\{aX\} = a^2 \sigma^2\{X\} \]
\[ \sigma^2\{aX + bY\} = a^2 \sigma^2\{X\} + b^2 \sigma^2\{Y\} \text{ if } X \perp \perp Y \]
\[ \sigma^2\{a + cX\} = c^2 \sigma^2\{X\} \text{ if } a, c \text{ both constant} \]

More generally

\[ \sigma^2\{\sum a_iX_i\} = \sum \sum a_ia_j \text{Cov}(X_i, X_j) \]
The covariance between two real-valued random variables $X$ and $Y$, with expected values $E\{X\} = \mu$ and $E\{Y\} = \nu$ is defined as

$$Cov(X, Y) = E\{(X - \mu)(Y - \nu)\}$$

Which can be rewritten as

$$Cov(X, Y) = E\{XY - \nu X - \mu Y + \mu \nu\},$$

$$Cov(X, Y) = E\{XY\} - \nu E\{X\} - \mu E\{Y\} + \mu \nu,$$

$$Cov(X, Y) = E\{XY\} - \mu \nu.$$
Covariance of Independent Variables

If $X$ and $Y$ are independent, then their covariance is zero. This follows because under independence

$$E\{XY\} = E\{X\}E\{Y\} = \mu \nu.$$ 

and then

$$\text{Cov}(XY) = \mu \nu - \mu \nu = 0.$$
Least Squares Linear Regression

Seek to minimize

$$Q = \sum_{i=1}^{n} (Y_i - (b_0 + b_1 X_i))^2$$

By careful choice of $b_0$ and $b_1$ where $b_0$ is a point estimator for $\beta_0$ and $b_1$ is the same for $\beta_1$

How?
Guess #1

![Graph showing a linear relationship between predictor/input and response/output with two regression lines: one for 'Guess, y = 0x + 21.2, mse: 37.1' and another for 'True, y = 2x + 9, mse: 4.22']
Guess #2

Graph showing a scatter plot with two lines:
- Black line: \( y = 1.5x + 13 \), \( \text{mse: 7.84} \)
- Red line: \( y = 2x + 9 \), \( \text{mse: 4.22} \)
Function maximization

- Important technique to remember!
  - Take derivative
  - Set result equal to zero and solve
  - Test second derivative at that point

- Question: does this always give you the maximum?

- Going further: multiple variables, convex optimization
Function Maximization

Find

$$\arg\max_x -x^2 + \ln(x)$$
Least Squares Max(min)imization

- Function to minimize w.r.t. $b_0$ and $b_1$, $b_0$ and $b_1$ are called point estimators of $\beta_0$ and $\beta_1$ respectively.

$$Q = \sum_{i=1}^{n} (Y_i - (b_0 + b_1 X_i))^2$$

- Minimize this by maximizing $-Q$.

- Either way, find partials and set both equal to zero.

$$\frac{dQ}{db_0} = 0$$
$$\frac{dQ}{db_1} = 0$$
Normal Equations

- The result of this maximization step are called the normal equations.

\[
\sum Y_i = nb_0 + b_1 \sum X_i
\]

\[
\sum X_i Y_i = b_0 \sum X_i + b_1 \sum X_i^2
\]

- This is a system of two equations and two unknowns. The solution is given by...
Solution to Normal Equations

After a lot of algebra one arrives at

\[
\begin{align*}
    b_1 &= \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2} \\
    b_0 &= \bar{Y} - b_1 \bar{X} \\
    \bar{X} &= \frac{\sum X_i}{n} \\
    \bar{Y} &= \frac{\sum Y_i}{n}
\end{align*}
\]