Linear Models with R

Review

→ (Brief) review of vectors/matrices, Probability, and Statistics

Preface

- objective is to learn what methods are available, and more importantly, when they should be applied

Ch. 1: Introduction

1.1: Before You Start

Statistical Approach to a Scientific Problem

- Ask a scientific question and formulate a statistical problem
- collect data
- initial, exploratory analysis
- Answer the question (inferential statistics)

Problem Formulation

- understand the physical background and the objective
  - Make sure you know what is wanted
  - Put the problem into statistical terms

Collect Data

- Determine the sampling population
- Sampling
  - Is this a random sample? Experiment or observational study?
  - Is there missing data? How is it handled?
  - Is there measurement error?
- If possible, always perform designed experiments

1.2: Initial (Exploratory) Data Analysis

- organize data
- Display data graphically
- Summarize data
- Be alert for unexpected (e.g. scaling, missingness, etc.)

1.3: When to Use Regression Analysis

Inferential Statistics

- Estimate parameters · Make predictions · Test hypotheses
- What did we learn? What is still uncertain? What may have gone wrong?

Regression Analysis

- Build a model to "explain" the relationship btwn a single variable Y and other variables $X_1, \dots, X_p$.

where { Y: response (dependent) var, output; X: predictor (independent) var, input, covariate; $p=1$: SLR, $p>1$: MLR }

Remark: Y: Multidimensional ⇒ Multivariate regression

Goals of Regression Analysis

- Describe data
- Make predictions
- Assessment of (association) effect of predictor(s) on response

Remark: Association does not imply causation

* Regression Analysis (alone) cannot establish causation

Types of variables

- Qualitative (categorical); can't say one is bigger than another
- Quantitative (numerical)
  → Discrete (counts) → Continuous measures

What we will cover

- X: continuous, discrete, categorical
- Y: continuous ⟶ linear reg.
- Y: Binary ⟶ logistic reg.
- Y: Discrete (counts) ⟶ Poisson reg.

} Generalized Linear Regression Models (GLM)

uses a linear combination of predictors to explain a dependent var. } random component; linear (systematic) component; link function

Emphasis of the course

- Practice using (generalized) linear regression models
- Learn what methods are available and their limitations
- Many examples w/ less mathematical theory
- More intuition w/ less derivation of formulas
- Will still learn mathematical foundations behind practical tools

Ch. 2: Estimation

Textbook	Lecture content
2.1: Linear Models	Regression Analysis
2.2: Matrix Representation	Linear Regression Analysis
2.3: Estimating β	Simple Linear Regression (SLR)
2.4: Least Squares Estimation	Multiple Linear Regression (MLR)
2.5: Example – Calculating β	Statistical Properties of $\hat{\beta}$ (and $\hat{\sigma}^2$)
2.6: Gauss-Markov Theorem	Goodness of Fit
2.7: Goodness of Fit	Gauss-Markov Theorem ($\hat{\beta}$: BLUE)
2.8: Example (R – done separately)	

Regression Analysis

y: response - output

$x = (x_1, \dots, x_p)$ - predictors - input

Goal: model the relationship btwn y and $x_1, \dots, x_p$

> **Note:** "regression to the mean" phenomena refers to unusually extreme events tending to be followed by events closer to the avg.

General Form: Let $X = (x_1, \dots, x_p)^T$, $y = (y_1, \dots, y_n)^T$

then, $\boxed{y = f(X) + \varepsilon}$

where $f(\cdot)$: underlying truth, ε - (statistical) error

- There is an unknown mean function $E(Y|X) = f(X)$
  - we want to estimate $f(X)$ w/ a statistical model (e.g. regression, NN, etc.)
- Error term $\varepsilon = Y - E(Y|X)$
  - measures variability w.r.t. the mean
  - assume: $E(\varepsilon|X) = 0$
- Usually we are given a set of data
  - n: sample size (# obs)
  - denoted $(x_{11}, \dots, x_{1p}, y_1), \dots, (x_{n1}, \dots, x_{np}, y_n)$
- can represent in matrix form, i.e.

$$\text{design matrix } (X) := \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix}$$

$$\text{response vector } (y) := (y_1, \dots, y_n)^T$$

Linear Regression Analysis

- There is no way to estimate $f(\cdot)$ directly given a finite number (e.g. n) of samples
  - functions represent infinite-dimension space
- We put some restrictions/structure on $f(\cdot)$, i.e.

$$\boxed{f(x) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p} \quad \text{for } n > p$$

where β_j: unknown (scalar) parameters, β_0: intercept

$\equiv$ regression coefficients

- Estimation of $f(\cdot)$ is reduced to estimation of $(p+1)$ β_j's

Additional Ex

(-) GLM s.t. $E(Y|X) = g(\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p)$

g: known (often nonlinear) function

e.g. logistic function for logistic regression

(-) NN

- often when $n \ll p$
- requires additional regularization

What Does "Linear" mean?

- Linear models is linear in parameters, not (necessarily) linear in predictors

Prop. Formally, a function g is linear in β if

$$g(a \cdot \beta + a^* \beta^*) = a \cdot g(\beta) + a^* g(\beta^*)$$

where $a, a^* \in \mathbb{R}$, $\beta, \beta^* \in \mathbb{R}^p$

Corollary: Equivalently, linearity satisfies the following prop's:

- Additivity: $f(x+y) = f(x) + f(y)$
- Homogeneity (of deg. 1): $f(\alpha x) = \alpha f(x)$ $\forall \alpha$.

Ex: $f(x) = \beta_0 + \beta_1 e^{x_1} + \beta_2 \ln(x_2) + \beta_3 x_1 x_3$ is a linear model

non-e.g.: whereas $f(x) = \beta_0 + \beta_1 x_1^{\beta_2}$ is not

Transformations

$f(x) = \beta_0 x_1^{\beta_1}$ is not a linear model.

However, notice that

$$\ln f(x) = \ln \beta_0 + \beta_1 \ln x_1$$

Hence, if we let $f^*(x) = \ln f(x)$, $\beta_0^* = \ln \beta_0$, $\beta_1^* = \beta_1$, we have

$$f^*(x) = \beta_0^* + \beta_1^* \ln x_1$$

which is a linear model

Implications

- Linear models are less restrictive than you might think
- they can be made very flexible by transformation of the response & predictors

Def (linear transformation): $f(x) = \beta_0 + \beta_1 g_1(x) + \cdots + \beta_p g_p(x)$

where $g_1, \dots, g_p$ are known functions

- Linear models are not necessarily straight lines, e.g.

$$y = ax^2 + bx + c$$

Simple Linear Regression

Let $p=1$: number of predictors

model: $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, $i = 1, \dots, n$ (*)

s.t. $E(Y_i | x_i) = \beta_0 + \beta_1 x_i$

Assumptions: $\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$

1) $E(\varepsilon_i | x_i) = 0$ $\Leftrightarrow$ $E(Y_i | x_i) = \beta_0 + \beta_1 x_i$ (+ zero mean, + linearity)

2) $V(\varepsilon_i | x_i) = \sigma^2$ $\Leftrightarrow$ $V(Y_i | x_i) = \sigma^2$ (+ constant variance; $V(\beta_0 + \beta_1 x_i + \varepsilon_i | x_i)$ by var. of constant)

3) Given x's, ε_i's are ind. (or uncorrelated)

4) Normality

Remark: For violations of assumptions, namely (2), there are modifications to model (*) we can make that are discussed later

Let $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)^T$, $X = (x_1, \dots, x_n)^T$

Thus, equivalently assume: (1) $E(\varepsilon | X) = \begin{pmatrix} E(\varepsilon_1 | x_1) \\ \vdots \\ E(\varepsilon_n | x_n) \end{pmatrix}_{n \times 1} = 0$.

$\varepsilon \overset{iid}{\sim} N(0, \sigma^2)$

(2) $Cov(\varepsilon | X) = \begin{pmatrix} V(\varepsilon_1 | x_1) & \cdots & \\ \vdots & & \\ Cov(\varepsilon_i, \varepsilon_j | x_i, x_j) & & \\ & & V(\varepsilon_n | x_n) \end{pmatrix}_{n \times n} = \begin{pmatrix} \sigma^2 & & 0 \\ & \ddots & \\ 0 & & \sigma^2 \end{pmatrix} = \sigma^2 I_n$

(3) Stronger assumption assumes ε_i's are independent

(4) Further assume ε_i's are normally distributed

Goal: Given (y_i, x_i), $i \in \{1, \dots, n\}$, estimate β_0, β_1, and σ^2

+ we do this by minimizing errors

$\Rightarrow$ one criterion is **least squares:** $\min_{\beta_0, \beta_1} \sum_{i=1}^n \varepsilon_i^2 = \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2$ (**)

$= \sum_{i=1}^n (y_i - E(Y_i | x_i))^2$

Remark: observe the special case s.t. $E(Y|X) = \beta_0 \Rightarrow \hat{\beta}_0 = \underset{\beta_0}{\operatorname{argmin}} \sum_{i=1}^n (y_i - \beta_0)^2$

$\Rightarrow \cdots \Rightarrow \hat{\beta}_0 = \bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$.

Next, we differentiate (**) w.r.t. β_0, β_1, set to zero, and solve, i.e.

$\frac{\partial}{\partial \beta_0} = (-2) \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i) = 0$

$\Rightarrow \sum y_i - n\hat{\beta}_0 - \hat{\beta}_1 \sum x_i = 0$

$\Rightarrow \hat{\beta}_0 = \frac{1}{n} \sum y_i - \hat{\beta}_1 \frac{1}{n} \sum x_i$

$\Rightarrow \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$

and

$\frac{\partial}{\partial \beta_1} = (-2) \sum_{i=1}^n x_i (y_i - \beta_0 - \beta_1 x_i) = 0$

$\Rightarrow \sum x_i y_i - \hat{\beta}_0 \sum x_i - \hat{\beta}_1 \sum x_i^2 = 0$

$\Rightarrow \sum x_i y_i - \hat{\beta}_0 n\bar{x} - \hat{\beta}_1 \sum x_i^2 = 0$

$\Rightarrow \sum x_i y_i - (n\bar{y} - \hat{\beta}_1 n\bar{x}) n\bar{x} - \hat{\beta}_1 \sum x_i^2 = 0$, substitution

$\Rightarrow \sum x_i y_i - n\bar{x}\bar{y} = -\hat{\beta}_1 \sum x_i^2 - \hat{\beta}_1 n\bar{x}^2$

$\Rightarrow \hat{\beta}_1 = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum x_i^2 - n\bar{x}^2}$ (a) (b)

where

a) is equivalent to $\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$:

$\sum (x_i - \bar{x})(y_i - \bar{y}) = \sum (x_i y_i - x_i \bar{y} - \bar{x} y_i + \bar{x}\bar{y})$

$= \sum x_i y_i - \bar{y} \sum x_i - \bar{x} \sum y_i + n\bar{x}\bar{y}$

$= \sum x_i y_i - \bar{y}(n\bar{x}) - \bar{x}(n\bar{y}) + n\bar{x}\bar{y}$

$= \sum x_i y_i - n\bar{x}\bar{y}$.

and b) is equivalent to $\sum_{i=1}^n (x_i - \bar{x})^2$:

$\sum (x_i - \bar{x})^2 = \sum (x_i^2 - 2\bar{x}x_i + \bar{x}^2)$

$= \sum x_i^2 - 2\bar{x} \sum x_i + n\bar{x}^2$

$= \sum x_i^2 - 2\bar{x} n\bar{x} + n\bar{x}^2$

$= \sum x_i^2 - 2n\bar{x}^2 + n\bar{x}^2$

$= \sum x_i^2 - n\bar{x}^2$

This implies that:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$\hat{\beta}_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

Remarks: · hat notation used for estimates

· $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$

$\Leftrightarrow \bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}$

$\Rightarrow (\bar{x}, \bar{y})$ lies on the estimated regression line

· $\hat{\beta}_1 = \dfrac{cov(X,Y)}{V(X)}$ | P. $cov(X,Y) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$

$\times V(X) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = s_x^2 = \dfrac{cov(X,Y)}{V(X)}$ ∎

· $\hat{\beta}_1 = \dfrac{cov(X,Y)}{V(X)} = cor(X,Y)\dfrac{s_y}{s_x}$ | P. $cor(X,Y) = \dfrac{cov(X,Y)}{s_x s_y} \Rightarrow cov(X,Y) = cor(X,Y)\, s_x s_y$

So, $\dfrac{cov(X,Y)}{V(X)} = \dfrac{cor(X,Y)\, s_x s_y}{s_x^2} = cor(X,Y)\cdot\dfrac{s_y}{s_x}$

Another interpretation

Formally, let $r = cor(X,Y)$, $s_y = SD(y)$, $s_x = SD(x)$, we can rewrite

$$\begin{aligned} y &= \hat{\beta}_0 + \hat{\beta}_1 x \\ &= \bar{y} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 x \quad \text{via substitution} \\ \Leftrightarrow\ y - \bar{y} &= \hat{\beta}_1 (x - \bar{x}) \\ &= r\frac{s_y}{s_x}(x - \bar{x}) \\ \Leftrightarrow\ \frac{y-\bar{y}}{s_y} &= \frac{\hat{\beta}_1 (x-\bar{x})}{s_x} \end{aligned}$$

Or, if x and y are standardized first (mean 0 SD 1), simply

$y = rx$ P. $\hat{\beta}_1 = \dfrac{cov(X,Y)}{V(X)} = \dfrac{cor(X,Y)\, s_x s_y}{V(X)} = cor(X,Y)$ when X, Y are standardized. ∎

recall: Standardizing; for $i = 1, \dots, n$

$x_i \longrightarrow \dfrac{x_i - \bar{x}}{s_x} = x_i^*$

$y_i \longrightarrow \dfrac{y_i - \bar{y}}{s_y} = y_i^*$

} s.t. x_i^* and y_i^*, $i \in \{1, \dots, n\}$ have (sample) means 0 and SD 1

Two regression lines

suppose x and y have both been standardized

· Regress y on x: $y = rx$

· Regress x on y: $x = ry$

} when $-1 \le r \le 1$ $\Leftrightarrow$ $-1 \le r \le 1$

regression effect: predictions always "regress" towards the mean

Q: why $x = ry$ and not $x = \frac{1}{r} y$

↑ two regressions minimize different error directions

Multiple Linear Regression

Model: $y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i$

Remarks: · $x_i \in \mathbb{R}^{(p+1)\times 1} = (1, x_{i1}, \cdots, x_{ip})^T$

· $\beta \in \mathbb{R}^{(p+1)\times 1} = (\beta_0, \beta_1, \cdots, \beta_p)^T$

· $E(Y_i | x_i) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$

$\Leftrightarrow E(Y_i | x_i) = x_i^T \beta$

· p predictors

· $(p+1)$ β parameters and (1) $\sigma^2 = var(\varepsilon_i | x_i)$ parameter

Assumptions: 1) (linear) mean function $E(Y_i | x_i) = x_i^T \beta \Leftrightarrow E(\varepsilon_i | x_i) = 0$

$\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$ 2) var function $var(Y_i | x_i) = \sigma^2 \Leftrightarrow var(\varepsilon_i | x_i) = \sigma^2$

} for $i \in \{1, \dots, n\}$

$\Leftrightarrow$

2) $\varepsilon_i, \varepsilon_j | X$ independent or uncorrelated $cov(\varepsilon_i, \varepsilon_j | X) = 0$ for $i \neq j$;

1) $E(\varepsilon_{n\times 1} | X) = 0$ for $\varepsilon_{n\times 1} = (\varepsilon_1, \cdots \varepsilon_n)^T$

2) $cov(\varepsilon_{n\times 1} | X) = \sigma^2 I_n$

Matrix Notation

Let $y = (y_1, \dots, y_n)^T$, $X \in \mathbb{R}^{n\times(p+1)} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix} = \begin{pmatrix} x_1^T \\ \vdots \\ x_n^T \end{pmatrix}$, $\beta \in \mathbb{R}^{(p+1)\times 1} = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_p \end{pmatrix}$, $\varepsilon \in \mathbb{R}^{n\times 1} = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$

Remarks: · x_{ij}: i^{th} row, $(j+1)^{th}$ column

· i^{th} row of X is x_i^T where $x_i = \begin{pmatrix} 1 \\ x_{i1} \\ \vdots \\ x_{ip} \end{pmatrix}$, $i = 1, \dots, n$

Thus, more compactly

$$y_{n\times 1} = X_{n\times(p+1)} \beta_{(p+1)} + \varepsilon_{n\times 1} \quad \Leftrightarrow \quad \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} x_1^T\beta \\ \vdots \\ x_n^T\beta \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

Estimating β

· observe y and X (n samples)

· want to minimize errors

Least Squares criteria: $\min_{\beta} \sum \varepsilon_i^2 = \varepsilon^T \varepsilon$

$= (y - X\beta)^T (y - X\beta)$, $y = X\beta + \varepsilon \Rightarrow \varepsilon = y - X\beta$.

, $\varepsilon(y_i - x_i^T\beta)^2$

$= y^T y - \beta^T X^T y - y^T X \beta + \beta^T X^T X \beta$ (FOIL)

where (cross product terms): $-y^T X\beta - (X\beta)^T y$

$= -\underbrace{y^T X \beta}_{1\times 1} - \underbrace{\beta^T X^T y}_{1\times 1}$, $(y^T X\beta)^T = \beta^T X^T y$

$= -2y^T X\beta$. ∎

Differentiating (*) wrt β and setting to 0:

$$0 = \nabla_\beta (y^T y - 2y^T X\beta + \beta^T X^T X\beta)$$

$$= -2(y^T X)^T + 2(X^T X)\beta, \text{ by (1),(2) (3) respectively}$$

$$0 = -2X^T y + 2(X^T X)\beta$$

$$\Rightarrow X^T y = (X^T X)\beta$$

$$\Downarrow$$

normal equation: $X^T X\hat{\beta} = X^T y$

$$\Downarrow$$

ordinary least squares (OLS) estimator

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

- assumes $X^T X$ invertible
- $X_{n\times(p+1)}$ full column rank $\Leftrightarrow X^T X$ invertible

note: if $n < (p+1)$, then singular (not invertible)

matrix-calculus formulas

Let $A \in \mathbb{R}^{m\times n}$, $B \in \mathbb{R}^{n\times m}$, $C \in \mathbb{R}^{n\times n}$ (symmetric)

recall: $\mathrm{tr}(A) = a_{11} + a_{22} + \cdots + a_{nn}$

1) $\nabla_B \mathrm{tr}(AB) = A^T$

2) $\nabla_B \mathrm{tr}(B^T C B) = 2CB$

(3) when $m = 1 \Rightarrow A \in \mathbb{R}^{1\times n}$, $B \in \mathbb{R}^{n\times 1}$ vectors

Note that $\mathrm{tr}(AB) = AB$

$\mathrm{tr}(B^T C B) = B^T C B$

Now,

1) $\nabla_B (AB) = A^T$

2) $\nabla_B (B^T C B) = 2CB$

Remark: $\|A\|_F = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n} |a_{ij}|^2} = \sqrt{\mathrm{tr}(A^T A)} = \sqrt{\mathrm{tr}(AA^T)} = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{m} |a_{ij}|^2}$,

i.e. the Frobenius norm: standard Euclidean norm of a matrix treated as a long vector.

Remarks: - We discuss alternative methods for when $n < (p+1)$ later
- We also discuss (multi)collinearity/invertible violation later

Fitted Model

(Fitted) { value: $E(Y_i|X_i) = \hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_p x_{ip}$
model: $E(Y|X) = \hat{f}(X) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_p x_p$

Residuals: $\hat{\varepsilon}_i = e_i = y_i - \hat{y}_i$
$= y_i - \hat{E}(Y_i|X_i)$

Residual vector: $\hat{\varepsilon} = (\hat{\varepsilon}_1, \ldots, \hat{\varepsilon}_n)^T$; $\hat{\varepsilon}_{n\times 1} = Y_{n\times 1} - \hat{Y}_{n\times 1} = Y - X\hat{\beta}$

Properties: (1) sample mean of residuals is zero

$$\Leftrightarrow \bar{\hat{\varepsilon}} = \frac{1}{n}\sum_{i=1}^{n} \hat{\varepsilon}_i = 0$$

P. $X^T X\hat{\beta} = X^T Y \Leftrightarrow X^T Y - X^T X\hat{\beta} = 0 \Leftrightarrow X^T(Y - X\hat{\beta}) = 0 \Leftrightarrow X^T\hat{\varepsilon} = 0$

First row vector of X^T is $(1, \ldots, 1)_{1\times n}$

$\Rightarrow (1, \ldots, 1)\hat{\varepsilon}_{n\times 1} = 0 \Rightarrow \sum_{i=1}^{n} \hat{\varepsilon}_i = 0 \Rightarrow \bar{\hat{\varepsilon}} = 0$. ●

Remark: Normal eq. further implies that residual and predictor are uncorrelated to each other, i.e. X^T and $\hat{\varepsilon}$ are orthogonal ($X^T\hat{\varepsilon} = 0$).

Residual sum of squares (RSS): $\sum_{i=1}^{n} \hat{\varepsilon}_i^2$

Hat Matrix

$\hat{y} = X\hat{\beta} = X(X^T X)^{-1} X^T y = Hy$ s.t.

Hat matrix (HM): $H = X(X^T X)^{-1} X^T$

fitted values: $\hat{y} = Hy$

residuals: $\hat{\varepsilon} = y - \hat{y} = Y - (Hy) = (I - H)y$

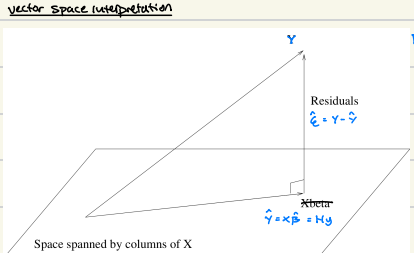
$\Rightarrow H$ is a projection matrix

Projection Matrix

Def: H is a projection matrix if
i) $H^T = H$ (symmetric)
ii) $HH = H$ (idempotent)

note: able to show that $X(X^T X)^{-1} X^T$ satisfies i) and ii)

Vector Space Interpretation



$\mathbb{R}^n$

$(p+1)$-dim if X has full column rank

★ the projection matrix H projects $y_{n\times 1}$ onto the column space of $X_{n\times(p+1)}$, which reduces to the vector space interpretation of least squares estimate
- subspace spanned by column vectors of $X_{n\times(p+1)}$

$\Leftrightarrow$ Let $C(X)$:= column space of X.
Then $P_{C(X)}(y) = Hy = \hat{y} = X\hat{\beta} = (X^T X)^{-1} X^T y$, i.e.
the hat matrix (H) is the projection operator $P_{C(X)}$
s.t. $H: y \to \hat{y}$

★ $\min_\beta (y - X\beta)^T (y - X\beta)$ minimizes the Euclidean dist. btwn y and the linear space spanned by columns of X

- $\hat{\varepsilon} \perp$ (linear space) $\Leftrightarrow X^T\hat{\varepsilon} = 0$
- $\hat{\varepsilon} \perp \hat{Y} \Leftrightarrow \hat{y}^T\hat{\varepsilon} = 0$
$\Leftrightarrow (X\hat{\beta})^T\hat{\varepsilon} = 0$
$\Leftrightarrow \hat{\beta}^T X^T\hat{\varepsilon} = 0$. ●

(Statistical) Properties of $\hat{\beta}$ (and $\hat{\sigma}^2$)

Here, we treat $\hat{\beta}$ as random variables

Assume that X (observed data) is fixed (constant), i.e. $E(Y|X)$.

$\boxed{\hat{\beta} = \beta + (X^TX)^{-1}X^T\varepsilon.}$

P. $\hat{\beta} := (X^TX)^{-1}X^TY \Rightarrow$ all 'randomness' comes from $Y \in \mathbb{R}^{n\times 1}$

$= (X^TX)^{-1}X^T(X\beta + \varepsilon) \Rightarrow$ all 'randomness' comes from ε (dist.)

$= (X^TX)^{-1}X^TX\beta + (X^TX)^{-1}X^T\varepsilon$

$= \beta + (X^TX)^{-1}X^T\varepsilon.$ ∎

$\boxed{\text{unbiased} := E(\hat{\beta}) = \beta}$

P. $E(\hat{\beta}) = E(\beta + (X^TX)^{-1}X^T\varepsilon)$

$= \beta + (X^TX)^{-1}X^T E(\varepsilon) = \beta.$ ∎ (with $E(\varepsilon) = 0$)

- $\hat{\beta}$ is an unbiased estimator for β.

Define covariance matrix $:= cov(\hat{\beta}) = var(\hat{\beta}) \in \mathbb{R}^{(p+1)\times(p+1)}$

Prop. Assuming $cov(\varepsilon) = var(\varepsilon) = \sigma^2 I \in \mathbb{R}^{n\times n}$,

$\boxed{var(\hat{\beta}) = \sigma^2 (X^TX)^{-1}}$

P. $Var(\hat{\beta}) = Var(\beta + (X^TX)^{-1}X^T\varepsilon)$

$= (X^TX)^{-1}X^T var(\varepsilon) X (X^TX)^{-1}$, Recall: $A \in \mathbb{R}^{m\times n}$, then $var(A\vec{\varepsilon}) = A\, var(\vec{\varepsilon})A^T$ (with $var(\varepsilon) = \sigma^2 I_n$)

$= \sigma^2 (X^TX)^{-1}X^TX(X^TX)^{-1}$, $(X^TX)^{-1}$ is symmetric $\Rightarrow ((X^TX)^{-1})^T = (X^TX)^{-1}$.

$= \sigma^2(X^TX)^{-1}.$ ∎

Remarks: • $(i+1, j+1)$-th entry of $var(\hat{\beta})$ is $\widehat{cov}(\hat{\beta}_i, \hat{\beta}_j)$ for $j = 0, \ldots, p$; $i = 0, \ldots, p$

$\boxed{\bullet \sqrt{var(\hat{\beta}_j)} \text{ is the SE of } \hat{\beta}_j}$

Prop. If we further assume $\varepsilon \sim N(0, \sigma^2 I_n)$, then $\boxed{\hat{\beta} \sim N(\beta, \sigma^2(X^TX)^{-1})}$

Estimating Variance

σ^2 can be estimated, i.e.

$$\boxed{\hat{\sigma}^2 = \frac{\sum_i (y_i - \hat{y}_i)^2}{n-(p+1)} = \frac{RSS}{n-(p+1)} \quad \text{for df}: n-(p+1)}$$

Remarks (df intuition):

Starting w/ an ideal case (unrealistic),

suppose we observe all the statistical errors $\varepsilon_1, \ldots, \varepsilon_n$

Then we can use sample var to estimate σ^2, i.e.

$\hat{\sigma}^2 = \frac{1}{n-1}\sum_i(\varepsilon_i - \bar{\varepsilon})^2.$

But, in practice, $\varepsilon_1, \ldots, \varepsilon_n$ are not observed

Thus, replace ε_i's by $\hat{\varepsilon}_i$'s, i.e.

- $\sum_i(\varepsilon_i - \bar{\varepsilon})^2 \leftarrow \sum_i(\hat{\varepsilon}_i - \bar{\hat{\varepsilon}})^2$, $\bar{\hat{\varepsilon}} = 0$

$= \sum \hat{\varepsilon}_i^2 = RSS$

- $(n-1) \leftarrow n-(p+1)$

Ex: Consider $Y_i = \beta_0 + \varepsilon_i$ $(p=0)$.

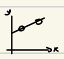
$\Rightarrow \hat{\beta}_0 = \bar{Y}.$

Thus, $\hat{\sigma}^2 = \frac{\sum_i(Y_i - \hat{Y}_0)^2}{n-1} = \frac{RSS}{n-(p+1)}$ for $p=0$: $= \frac{RSS}{n-1}$.

Here, all but one n observations are 'free', hence

$n-1$ df for $p=0$ parameters. ∎

Ex: $n = 2$, $p = 1$



$\hat{\varepsilon}_1 = \hat{\varepsilon}_2 = 0$

$RSS = 0$

$df = 0$; $n-(p+1)$

$= 2-(1+1) = 0.$ ∎

$\boxed{\text{unbiased} := E(\hat{\sigma}^2) = \sigma^2}$

Proof. seen in pg. 43

Proof. I will prove the unbiasedness of $\hat{\sigma}^2$, i.e. $E(\hat{\sigma}^2) = \hat{\sigma}^2$:

Let $Y = X\beta + \varepsilon$ assuming $\begin{cases} E(\varepsilon) = 0 \\ Var(\varepsilon) = \sigma^2 I_{n\times n} \end{cases}$ (*)

we define the projection matrices $H = X(X^TX)^{-1}X^T$ and $\tilde{H} = I - H$

with the following properties:
- $H = HH$
- $H = H^T$
- $\tilde{H} = \tilde{H}^T$
- $HX = X$
- $\tilde{H}X = 0$

Recall that e estimates ε such that
- $HY = \hat{Y}$
- $\tilde{H}Y = (I-H)Y = Y - \hat{Y} = e$
- $\tilde{H}Y = \tilde{H}(X\beta + \varepsilon) = \tilde{H}\varepsilon$

$\Rightarrow e = \tilde{H}\varepsilon.$

We define $RSS := \sum_{i=1}^{n} e_i^2 = e^Te = (\tilde{H}\varepsilon)^T(\tilde{H}\varepsilon) = \varepsilon^T\tilde{H}^T\tilde{H}\varepsilon = \varepsilon^T\tilde{H}\varepsilon.$

Following from the assumptions (*),

$E(\varepsilon\varepsilon^T) = \sigma^2 I_{n\times n}.$

Using the trace identity, i.e.

$tr(AB) = tr(BA)$

this implies that

$tr(H) = tr(X(X^TX)^{-1}X^T)$

$= tr(X^TX(X^TX)^{-1})$

$= tr(I_{(p+1)\times(p+1)})$

$= p+1$

and

$tr(\tilde{H}) = tr(I - H)$

$= tr(I) - tr(H)$

$= n-(p+1).$

Now, to show that $\hat{\sigma}^2$ is unbiased, first:

$E(RSS) = E(\varepsilon^T\tilde{H}\varepsilon) = E\{tr(\varepsilon^T\tilde{H}\varepsilon)\}$

$= E\{tr(\tilde{H}\varepsilon\varepsilon^T)\}$

$= tr(\tilde{H}(\sigma^2 I_{n\times n}))$

$= \sigma^2 tr(\tilde{H})$

$= \sigma^2(n-(p+1)).$

Then,

$E(\hat{\sigma}^2) = E\left[\frac{RSS}{n-(p+1)}\right] = \frac{\sigma^2(n-(p+1))}{n-(p+1)} = \sigma^2.$ ∎

Remark: $V(\hat{\sigma}^2)$ is more complex & beyond scope of this class

Goodness of Fit

- Need a measure of how well the model fits w/ the data
- One idea is to use the residual sum of squares, i.e. $\sum_i (y_i - \hat{y}_i)^2$

  + seems reasonable but consider units, i.e.

  need scaling so we have same measurement scale

Coefficient of Determination

$$\boxed{R^2 := 1 - \frac{RSS}{TSS} = \frac{\sum_i(y_i - \hat{y}_i)^2}{\sum_i(y_i - \bar{y})^2} = \frac{SS_{reg}}{TSS} = \frac{\sum_i(\hat{y}_i - \bar{y})^2}{\sum_i(y_i - \bar{y})^2}}$$

Interpretation: measures the proportion of variability of response (y) that can be explained by the association with its predictors (covariates) under the fitted regression model

Remarks:
- RSS: variation in y unexplained by model
- SS_{reg}: variation in y explained by model
  + sample var. of $\hat{y}_i$, $i = 1, \ldots, n := SS_{reg}/n-1$; $V(\hat{y}_i) = \frac{\sum(\hat{y}_i - \bar{y})^2}{n-1} = \frac{SS_{reg}}{n-1}$
  + It can be shown that

  $\bar{Y} = \frac{1}{n}\sum_i \hat{y}_i$; P. Assume $\sum(y_i - \hat{y}_i) = 0 \Rightarrow \sum y_i = \sum \hat{y}_i \Rightarrow \bar{Y} = \frac{1}{n}\sum \hat{y}_i.$ ∎

- It can be shown that

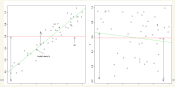
  $\underline{TSS = SS_{reg} + RSS}$

  P. $TSS = \sum(y_i - \bar{y})^2 = \sum(y_i - \hat{y}_i + \hat{y}_i - \bar{y})^2 = \sum[(y_i - \hat{y}_i)^2 + (\hat{y}_i - \bar{y})^2 + 2(y_i - \hat{y}_i)(\hat{y}_i - \bar{y})]$

  $= RSS + SS_{reg} + 2\sum(y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) = RSS + SS_{reg}.$ ∎ (with $\sum(y_i - \hat{y}_i) = 0$)

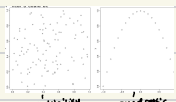
- $R^2 \in [0, 1]$

- R^2 closer to 1 indicates good fit



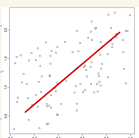
$R^2 \approx 1$ $\quad$ $R^2 \approx 0$

- $R^2 \approx 0$ for different reasons



uncorrelated $\quad$ quadratic trend

- Small R^2 does not necessarily mean that y and x are not linearly related; can have slight trend w/ high variance



$R^2 \approx 0.2$

- In SLR: $R^2 = r^2$

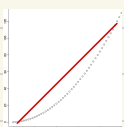
P. In SLR, $\hat{\beta}_1 = \frac{\sum (x_i-\bar{x})(y_i-\bar{y})}{\sum (x_i-\bar{x})^2}$ and $\hat{y}_i = \bar{y} + \hat{\beta}_1(x_i-\bar{x})$.

Now, $SS_{reg} = \sum(\hat{y}_i-\bar{y})^2 = \sum[\hat{\beta}_1(x_i-\bar{x})]^2 = \hat{\beta}_1^2 \sum (x_i-\bar{x})^2$

$= \left(\frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sum(x_i-\bar{x})^2}\right)^2 \sum(x_i-\bar{x})^2 = \frac{[\sum(x_i-\bar{x})(y_i-\bar{y})]^2}{\sum(x_i-\bar{x})^2}$.

Then, $R^2 = \frac{[\sum(x_i-\bar{x})(y_i-\bar{y})]^2}{[\sum(x_i-\bar{x})^2][\sum(y_i-\bar{y})^2]} = \left(\frac{cov(x,y)}{s_x s_y}\right)^2 = r^2$. ■

- $R^2 \approx 1$ does not mean the linear model is correct



$R^2 \approx 0.9$

Adjusted R^2

- adjust for the number of predictors in the model

$$\text{adj-}R^2 := 1 - \frac{RSS/(n-(p+1))}{TSS/(n-1)} = 1-(1-R^2)\frac{n-1}{n-(p+1)}$$

Remarks: · Penalizing (correction) factor for additional predictors
→ helps w/ model selection (comparison) using R^2

· R^2 will always increase (or stay the same) when adding more predictors
while adj-R^2 will not necessarily increase (probabilistic chance - but high likelihood)

Gauss-Markov Theorem

Q: why use the least-squares estimate $\hat{\beta}$?

Def: ∀ estimate $\tilde{\beta}$, we say ($\tilde{\beta}$ is) unbiased if $E(\tilde{\beta}) = \beta$

Def: we say ($\tilde{\beta}$ is) linear if it is a linear function in Y.
e.g. OLS $\hat{\beta} = (X^TX)^{-1}X^TY$ is linear

Thm (★) Suppose $y = X\beta + \varepsilon$, X - full (column) rank, $E(\varepsilon)=0$, $Var(\varepsilon) = \sigma^2 I$.

Consider $\psi = c^T\beta$, i.e. any linear combination of unknown $\beta \in \mathbb{R}^{p+1}$ and c is a known, pre-specified vector.

Then, among all unbiased linear estimates of ψ, $\hat{\psi} = c^T\hat{\beta}$ has the minimum variance and is unique

Remarks: · Let $c^T = (1, x_1, \dots, x_p)$. Then $\psi = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$

· OLS solution is the Best Linear Unbiased Estimate (BLUE)

· $E(\tilde{\psi}) = \psi \Rightarrow E(c^T\tilde{\beta}) = c^T\beta$.

· ∀ unbiased and linear estimator $\tilde{\beta}$,
$Var(c^T\hat{\beta}) \le Var(c^T\tilde{\beta})$
where $\hat{\beta}$: OLS and equality holds i.f.f. $\tilde{\beta} = \hat{\beta}$.

· Situations where estimators other than OLS should be used:
+ correlated errors and/or unequal variance: generalized least squares (Ch.1)
+ long-tailed error distribution: use robust estimates (Ch.4); typically not linear in y
+ collinear predictors: try biased estimators, e.g. ridge regression (ch.9)

Proof(★). Let $\tilde{\psi}$ be any other linear unbiased estimator for ψ.
Then $Var(\tilde{\psi}) \ge Var(\hat{\psi})$, and $Var(\tilde{\psi}) = Var(\hat{\psi}) \Rightarrow \tilde{\psi} = \hat{\psi}$.
Since $\tilde{\psi}$ is a linear estimator, it can be written as $\tilde{\psi} = d^Ty$ for $d \in \mathbb{R}^n$, d^T.
Thus, $\tilde{\psi} = d^Ty = d^TX\beta + d^T\varepsilon$.

Note that $E(\tilde{\psi}) = d^TX\beta$ and since $\tilde{\psi}$ is unbiased, $d^TX\beta = c^T\beta$.
Since this is true for any β, it must be that
$d^TX = c^T$.
Now let $q = d - X(X^TX)^{-1}c$. Note
$q^T = d^T - c^T(X^TX)^{-1}X^T$
and therefore, multiplying both sides by X,
$q^TX = d^TX - c^T(X^TX)^{-1}X^TX$
$= d^TX - c^T$
$= 0$
Next note $d^T = q^T + c^T(X^TX)^{-1}X^T$ and therefore
$Var(\tilde{\psi}) = Var(d^T\varepsilon)$
$= \sigma^2 d^Td$
$= \sigma^2[q^T + c^T(X^TX)^{-1}X^T][q + X(X^TX)^{-1}c]$
$= \sigma^2(q^Tq + c^T(X^TX)^{-1}c]$
$= \sigma^2 q^Tq + Var(\hat{\psi})$
≥ 0 and $q^Tq = 0 \Rightarrow q = 0 \Rightarrow \tilde{\psi} = \hat{\psi}$. ■

What can Go wrong?

- X^TX could be singular (happens if predictors are linearly dependent or if $p > n$).
- the following assumptions could be violated: $\begin{cases} \text{var}(\varepsilon) = \sigma^2 I & \text{(constant var)} \\ E(\varepsilon) = 0 & \text{(linearity)} \end{cases}$

note: OLS best only among linear, unbiased estimators

Ch. 3: Inference

Textbook	Lecture content
3.1: HT's to compare models	Hypothesis Testing
3.2: Testing Examples	$\hat{\beta}$ distribution
3.3: Permutation tests (Skip)	(General) F-test and/or t-test
3.4: (Simultaneous) CI's for β	Testing more than one Predictor
3.5: CI's for Prediction	confidence intervals
3.6: Designed Experiments	Simultaneous confidence Regions
3.7: observational Data	CI's for Prediction
3.8: Practical Difficulties	Interpreting Parameter Estimates
	Experimental Design
	Randomization — orthogonality

Inference

Estimates: $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p$

- want to draw conclusions about $\beta_0, \beta_1, \dots, \beta_p$
- two main inference tools (HT and CI)

Hypothesis Testing

- use probability to decide whether data is consistent w/ hypothesis

Null Hypothesis: H_0 (e.g. $\beta = 0$)

Alternative Hypothesis: H_A (e.g. $\beta \neq 0$)

- Decide whether data is consistent w/ H_0: $\begin{cases} \text{if not} \Rightarrow \text{reject } H_0 \text{ \& accept } H_A \\ \text{otherwise} \Rightarrow \text{FTR } H_0 \end{cases}$

Errors in Hypothesis Testing

decision \ true state	H_0 true	H_0 false
FTR H_0	✓	T2E
reject H_0	T1E	✓ (P)

T1E := $P(\text{reject } H_0 \mid H_0 \text{ true})$: false positive (rate)

T2E := $P(\text{FTR } H_0 \mid H_0 \text{ false})$: false negative (rate)

(Statistical Power) := $1 - \text{T2E} = 1 - \text{FNR}$ = true positive (rate)

$= (\text{T2E})^C = P(\text{reject } H_0 \mid H_0 \text{ false})$; correctly detect effect when one exists

Procedure

Set significance level := $\alpha = P(\text{T1E})$

→ typically $\alpha = 0.05$ (default), 0.01

Compute the p-value := prob of observed or more extreme departure from H_0 (in favor of H_A) when H_0 is true

p-value $< \alpha \Rightarrow$ Reject H_0

Further Assumption on Errors

In addition to linearity, $E(\varepsilon) = 0$, we assume a distribution for the errors, i.e.

$$\boxed{\varepsilon \overset{iid}{\sim} N_n(0, \sigma^2 I_n)} \quad (\Leftrightarrow E(\varepsilon_i) = 0,\ \text{cov}(\varepsilon_i, \varepsilon_j) = V(\varepsilon_i) = \sigma^2 = \sigma^2 I_n)$$

Distribution of $\hat{\beta}$

recall: $\hat{\beta} = \beta + (X^TX)^{-1}X^T\varepsilon$

If $\varepsilon \overset{iid}{\sim} N_n(0, \sigma^2 I)$, then

$$\hat{\beta} \sim N_{p+1}(\beta, (X^TX)^{-1}\sigma^2)$$

$$\hat{\beta}_j \sim N(\beta_j, (X^TX)^{-1}_{(j,j+1)}\sigma^2)$$

where, in practice, we use the approximation s.t.

$$\widehat{SE}(\hat{\beta}_j) := \sqrt{(X^TX)^{-1}_{(j,j+1)}\hat{\sigma}^2}, \quad \hat{\sigma}^2 = \frac{\sum_i (y_i - \hat{y}_i)^2}{n-(p+1)} = \frac{RSS}{n-(p+1)}$$

under the normal assumption on the errors:

$$\frac{\hat{\beta}_j - \beta_j}{SE(\hat{\beta}_j)} \sim N(0,1) \quad \text{and} \quad \frac{\hat{\beta}_j - \beta_j}{\widehat{SE}(\hat{\beta}_j)} \sim t_{n-(p+1)}$$

Proof (Sketch). $\hat{\beta}_j - \beta \overset{d}{\sim} N(0,1)$ and $\widehat{SE}(\hat{\beta}_j) \overset{d}{\sim} \frac{\sum_{i=1}^{n-(p+1)} z_i^2,\ z_i \overset{iid}{\sim} N(0,1)}{n-(p+1)} = \frac{\chi^2_{n-(p+1)}}{n-(p+1)}$.

$$\Rightarrow \frac{\hat{\beta}_j - \beta}{\widehat{SE}(\hat{\beta}_j)} \simeq t_{n-(p+1)}. \ \square$$

recall: $t_{df} := \frac{N(0,1)}{\sqrt{\frac{\chi^2_{df}}{df}}} \ (\Leftrightarrow t_{n-(p+1)} = \frac{N(0,1)}{\sqrt{\frac{\chi^2_{n-(p+1)}}{n-(p+1)}}})$ where $\chi^2_{n-(p+1)} \overset{d}{\sim} \sum_{i=1}^{n-(p+1)} z_i^2,\ z_i \overset{iid}{\sim} N(0,1)$

t-distribution

Probability density function (pdf):

$$N(0,1) \sim \frac{1}{\sqrt{2\pi}} \exp\left[\frac{1}{2}\left(-\frac{x^2}{2}\right)\right]$$

$$t_n \sim \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\cdot\Gamma\left(\frac{n}{2}\right)} \left(1 + \frac{x^2}{n}\right)^{-\left(\frac{n+1}{2}\right)}$$

Remarks:
- Has a single parameter n (df)
- Symmetric around 0, "bell-shaped", heavier tails than normal
- As $n \to \infty$, $t_n \to N(0,1)$
- $P(|t_{df}| \geq t) = \begin{cases} P(t_{df} \geq t) \\ \text{or} \\ P(t_{df} \leq -t) \end{cases}$

Another (General) Approach: (F-test)

recall: $RSS = \sum_i \hat{\epsilon}_i^2$

Remark: H_0 model is nested (simpler, special case) to the H_A model

- Fit a model under H_0 and compute RSS_{H_0} (e.g. $\beta_{popPS} = 0$).

  e.g. $X_{popPS}: x_2$, $H_0: \beta_2 = 0 \iff E(Y|X) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$

- Fit another model under $H_0 \cup H_A$ and compute $RSS_{H_0 \cup H_A}$ (e.g. no restriction on β_{popPS})

  e.g. $H_A: \beta_2 \neq 0 \iff E(Y|X) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$

- Now compute

$$F = \frac{\left[\frac{RSS_{H_0} - RSS_{H_0 \cup H_A}}{df_{H_0} - df_{H_0 \cup H_A}}\right]}{\left[\frac{RSS_{H_0 \cup H_A}}{df_{H_0 \cup H_A}}\right]} \quad (\star)$$

where $(\star) = \hat{\sigma}^2$ under $H_0 \cup H_A$ model

If H_0 is true

$$F \sim F_{df_1, df_2} \text{ where } \begin{cases} df_1 = df_{H_0} - df_{H_0 \cup H_A} \\ df_2 = df_{H_0 \cup H_A} \end{cases}$$

compute p-value $= P(F_{df_1, df_2} > F)$

Summary

1. Define (larger) full model (FM); thought to be most appropriate for data
2. Define (smaller) reduced model (RM); described by H_0
3. Use F-stat to decide whether to reject smaller model in favor of H_A (larger, full model)

H_0: RM $= M_{H_0}$; $\beta_i = 0$ for i: number of parameters $\in$ FM, $\notin$ RM

H_A: FM $= M_{H_0 \cup H_A}$; $\beta_i \neq 0$

↳ reject H_0

↳ there is (at least) one significant linear relationship between β_i and Y

F-distribution

Let $Z_1, \dots, Z_n \overset{iid}{\sim} N(0,1)$. Then

$$U = Z_1^2 + \dots + Z_n^2 \sim \chi^2_n = \text{Gamma}\left(\frac{n}{2}, 2\right).$$

Prop: Suppose $U \sim \chi^2_n$, $W \sim \chi^2_m$ are independent. Then

$$\frac{U/n}{W/m} \sim F_{n,m}$$

Remarks: - $F_{df_1, df_2} \geq 0$

- $t^2_{df} \sim F_{1, df}$ P. when $n=1$, $U = Z_1^2 \Rightarrow \frac{U/n}{W/m} = \frac{Z_1^2}{W/m} = \left(\frac{Z_1}{\sqrt{W/m}}\right)^2 \sim t^2_m$.

anova(M_{H_0}, $M_{H_0 \cup H_A}$) →

	Res.DF	RSS	DF	Sum of Square	F	Pr(>F)
	df_{H_0}	RSS_{H_0}				
	df_{H_A}	RSS_{H_A}	$df_{H_0} - df_{H_A}$	$RSS_{H_0} - RSS_{H_A}$	F-stat	p-value

where $H_A: H_0 \cup H_A$

F-test and t-test

Prop: F-test and two-sided t-test are equivalent for testing a single predictor

Proof: Consider SLR model s.t. $E(Y|X) = \beta_0 + \beta_1 x$, $x \in \mathbb{R}$

Show that F-test and $(t\text{-test})^2$ are equivalent for testing

$H_0: \beta_1 = 0$ vs. $H_A: \beta_1 \neq 0$

i.e. prove 1) F-stat $= (t\text{-stat})^2$

$$\Leftrightarrow \frac{\left[\frac{RSS_{H_0} - RSS_{H_0 \cup H_A}}{df_{H_0} - df_{H_0 \cup H_A}}\right]}{\left[\frac{RSS_{H_0 \cup H_A}}{df_{H_0 \cup H_A}}\right]} = \left[\frac{\hat{\beta}_1}{\widehat{SE}(\hat{\beta}_1)}\right]^2$$

2) $F_{1, n-2} \sim (t_{n-2})^2$. □

Testing more than one Predictor

$H_0: \beta_1 = \beta_2 = 0$

H_A: not H_0, i.e. at least one $\beta_i \neq 0$ for $i = 1, 2$ } testing a pair

$H_0: \beta_1 = \beta_2 = \dots = \beta_p = 0 \iff E(Y|X) = \beta_0$

H_A: not H_0; i.e. at least one $\beta_i \neq 0$ for $i = 1, \dots, n$. } overall F-test (as seen in overall R Summary table)

Suppose we want to test the equality of two effects:

$H_0: \beta_1 = \beta_2 \Rightarrow \beta_1 x_1 + \beta_2 x_1 = \beta_1 (x_1 + x_2)$; ($H_0$ ⊂ $\text{Im}(Y \sim I(x_1 + x_2) + \dots)$) } testing a subspace

$H_A: \beta_1 \neq \beta_2$

Suppose we want to test an effect value, i.e.

$H_0: \beta_k = c_0 \Rightarrow$ (im $CY \sim \beta_0 + \cdots +$ offset$(c_0 * P_k)$)

$H_a: \beta_k \neq, >, < c_0$

} testing a subspace

CF-test approach)

$\Leftrightarrow$ Save H_0 and H_a

w/ $t_{obs} \sim \frac{\hat{\beta}_k - (\beta_{0,k} = c_0)}{\hat{SE}(\hat{\beta}_k)}$

confidence intervals

HLT comments:

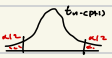
- only answers yes/no questions
- dependence on sample space
  - ↑ larger sample space → more likely to reject H_0
- notion of statistical significance vs. practical significance

e.g. assuming standardized data

$\hat{\beta} = 0.0001$, testing $H_0: \beta_i = 0$ vs. $H_a: \beta_i \neq 0$ → $p < 0.0001$ statistically significant but not necessarily practically significant

confidence intervals for β_j

recall: $\frac{\hat{\beta}_j - \beta_j}{\hat{SE}(\hat{\beta}_j)} \sim t_{n-(p+1)}$



Hence, $P\left(-t^{(\alpha/2)}_{n-(p+1)} \leq \frac{\hat{\beta}_j - \beta_j}{\hat{SE}(\hat{\beta}_j)} \leq t^{(\alpha/2)}_{n-(p+1)}\right) = 1-\alpha$

$\Leftrightarrow$ with prob. $(1-\alpha)$, i.e. $100(1-\alpha)\%$ confidence:

two-sided: $\hat{\beta}_j - t^{(\alpha/2)}_{n-(p+1)} \cdot \hat{SE}(\hat{\beta}_j) \leq \beta_j \leq \hat{\beta}_j + t^{(\alpha/2)}_{n-(p+1)} \cdot \hat{SE}(\hat{\beta}_j)$ or $(-\infty, \hat{\beta}_j + t^{\alpha}_{n-(p+1)} \cdot \hat{SE}(\hat{\beta}_j)]$ / $[\hat{\beta}_j - t^{\alpha}_{n-(p+1)} \cdot \hat{SE}(\hat{\beta}_j), \infty)$ one-sided

(result ← im $CY \sim x_1 + \cdots + x_p$)

confint(result)

output 95% 2-sided CI

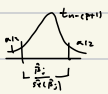
Remarks: $t^{(\alpha)}$ is the tail prob.: $P(\xi > t^{(\alpha)}) = \alpha$.

- choose $\alpha = 0.05 \Rightarrow 100(1-0.05) = 95\%$ CI
- $t^{0.025}_{n-(p+1)} \approx 2$; for large $n-(p+1)$, $t^{0.025}_{n-(p+1)} \to 1.96$

General Form: Estimate ± (critical value)(SE(estimate))

Corollary: $H_0: \beta_j = 0$, $H_a: \beta_j \neq 0$. Given α, $(1-\alpha)$ CI covers $\beta_j = 0 \Leftrightarrow$ p-value $> \alpha$.

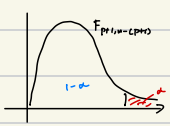
$\Leftrightarrow \left|\frac{\hat{\beta}_j}{\hat{SE}(\hat{\beta}_j)}\right| < t^{(\alpha/2)}_{n-(p+1)}$



simultaneous confidence regions

Prop. with prob $1-\alpha$, i.e. confidence $100(1-\alpha)\%$:

$(\hat{\beta}-\beta)^T X^T X (\hat{\beta}-\beta) \leq (p+1)\hat{\sigma}^2 F^{(\alpha)}_{p+1, n-(p+1)}$



Remarks: · Equivalently $P\left[(\hat{\beta}-\beta)^T X^T X (\hat{\beta}-\beta) \leq (p+1)\hat{\sigma}^2 F^{(\alpha)}_{p+1,n-(p+1)}\right] = 1-\alpha$

(*) $\frac{1}{p+1}(\hat{\beta}-\beta)^T (\hat{\sigma}^2 (X^T X)^{-1})^{-1} (\hat{\beta}-\beta) \sim F_{p+1, n-(p+1)}$ where $\widehat{var}(\hat{\beta}_{p+1}) = \hat{\sigma}^2 (X^T X)^{-1}$

$\Leftrightarrow \frac{(\hat{\beta}-\beta)^T X^T X (\hat{\beta}-\beta)}{(p+1)\hat{\sigma}^2} \sim F_{p+1, n-(p+1)}$

Prop: More generally, for any $S \subseteq \{0, 1, \ldots, p\}$

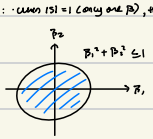
e.g. $S = \{1, 2\} \Leftrightarrow$ CI for $(\beta_1, \beta_2)^T$

$\frac{1}{|S|}(\hat{\beta}_S - \beta_S)^T [\widehat{var}(\hat{\beta}_S - \beta_S)]^{-1} (\hat{\beta}_S - \beta_S) \overset{d}{\sim} F_{|S|, n-(p+1)}$

where $|S|$: cardinality of set S

Remark: · when $|S| = 1$ (only one β), then this is equivalent to the 2-sided (ξ) CI

Note:



Note: the correlation btwn predictors and coefficients are often different in sign

→ two positively correlated predictors will attempt to perform the same job of explanation

- the more work one does, the less the other needs to do and hence a negative coefficient correlation

Proof (*). Given a single $\hat{\beta}_j$: $\frac{\hat{\beta}_j - \beta_j}{\hat{SE}(\hat{\beta}_j)} \sim t_{n-(p+1)}$

$\Rightarrow \frac{(\hat{\beta}_j - \beta_j)^2}{\hat{SE}^2(\hat{\beta}_j)} = \frac{(\hat{\beta}_j - \beta_j)^2}{\widehat{var}(\hat{\beta}_j)} \sim F_{1,df}$

$\Leftrightarrow (\hat{\beta}_j - \beta_j)(\widehat{var}(\hat{\beta}_j))^{-1}(\hat{\beta}_j - \beta_j) \sim F_{1,df}$ (*)

Now consider $\hat{\beta}_{p+1}$ s.t. $\hat{\beta}_{p+1} - \beta_{p+1} \sim N(0, \sigma^2 (X^T X)^{-1})$

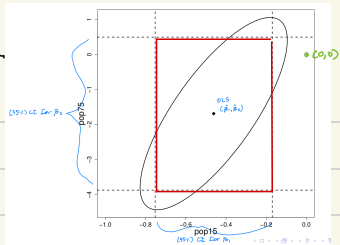
Generalizing (*) to the multivariate case:

$\frac{1}{p+1}(\hat{\beta}-\beta)^T (\hat{\sigma}^2 (X^T X)^{-1})^{-1} (\hat{\beta}-\beta) \sim F_{p+1, n-(p+1)}$

($\widehat{var}(\hat{\beta}_{p+1})$)

Thus, $\frac{(\hat{\beta}-\beta)^T X^T X (\hat{\beta}-\beta)}{(p+1)\hat{\sigma}^2} \sim F_{p+1, n-(p+1)}$

where ellipsoid a result of stretching/shrinking by variance (and hence correlation) correction factor



• (0,0) $H_0: \beta_1 = \beta_2 = 0$
H_A: not H_0

Remark: • simultaneous CI region more complex than intersection of respective 95% CI's (seen in red)
→ ellipsoid region accounts for simultaneous region; further details are beyond scope of class

• (0,0) outside of confidence region ⟺ p-value < α.

• direction (slanted/tilted) ellipse influenced by pairwise (anti/non) correlation
vertical/horizontal shape ⟺ orthogonal predictors

Better to consider joint CI's when possible - especially when $\hat{\beta}$ are heavily correlated

CI's for Prediction

(Training) Data: (x_i, y_i), $i = 1, \dots, n$.

Given new predictors, x_0, what is the predicted response?

i.e. given $Y_0 = x_0^T\beta + \varepsilon_0$, find $\hat{y}_0 = x_0^T\hat{\beta}$.

Two types of predictions

i) Prediction (1-α) of a future observation: $= \hat{y}_0 \pm t^{(\alpha/2)}_{n-(p+1)} \hat{\sigma} \sqrt{1 + x_0^T (X^T X)^{-1} x_0}$

remark: w.r.t. $Y_0 = x_0^T\beta + \varepsilon_0$

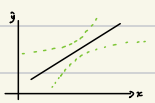
2) Prediction (1-α) of the future mean response: $= \hat{y}_0 \pm t^{(\alpha/2)}_{n-(p+1)} \hat{\sigma} \sqrt{x_0^T (X^T X)^{-1} x_0}$

remark: w.r.t. $x_0^T\beta = E(Y_0 | x_0)$

• Prediction (future observation) intervals vs. confidence (future mean response) intervals

```
result <- lm(Y ~ X1 + ... + Xp)
X0 <- data.frame(X1 = c1, ..., Xp = cp)
predict(result, X0, interval = "confidence")
predict(result, X0, interval = "prediction")
```

Prediction Band Plot:



• CI's for predictions get wider as we move away from data range

Proof (1). For a future obs., where $Y_0 = x_0^T\beta + \varepsilon_0$.

$E(Y_0) = E(\hat{y}_0) = x_0^T\beta \iff E(\hat{y}_0 - Y_0) = 0$

$\Rightarrow \text{var}(\hat{y}_0 - Y_0) = \text{var}(x_0^T\hat{\beta} - x_0^T\beta - \varepsilon_0)$

$= \text{var}(x_0^T(\hat{\beta} - \beta)) + \text{var}(\varepsilon_0)$ since $\hat{\beta}$ and ε are ind. (zero covariance)

$= x_0^T \text{var}(\hat{\beta}) x_0 + \sigma^2$, higher dim analog for $V(aX) = a^2 V(X)$

$= \sigma^2 x_0^T (X^T X)^{-1} x_0 + \sigma^2$

$= \hat{\sigma}^2 (1 + x_0^T (X^T X)^{-1} x_0)$ and $\dfrac{\hat{y}_0 - Y}{\hat{se}(\hat{y}_0 - Y_0)} \sim t_{n-(p+1)} \Rightarrow 1.$ ∎

Proof (2). We know $E(Y_0 | x_0) = x_0^T\beta$

Note that

$E(\hat{y}_0) = x_0^T\beta$ as above

$\widehat{\text{var}}(\hat{y}_0 - x_0^T\beta) = \widehat{\text{var}}(x_0^T\hat{\beta} - x_0^T\beta) \overset{1}{=} \hat{\sigma}^2 x_0^T (X^T X)^{-1} x_0$

and $\dfrac{\hat{y}_0 - x^T\beta}{\hat{se}(\hat{y}_0 - x_0^T\beta)} \sim t_{n-(p+1)} \Rightarrow (2).$ ∎

Extrapolation

• occurs when we try to predict the response for values of the predictor which lie outside the range of the original data; types:

1) Quantitative
⇒ check whether new x_0 is w/in range of original data
2) Qualitative
⇒ check whether the new x_0 is drawn from the same population from which the original sample was drawn

Interpreting Parameter Estimates

Q: What does $\hat{\beta}_j$, $j \in \{1, \dots, p\}$ mean?

A: a unit increase in x_j, w/ the other (known) predictors held constant will produce a change of $\hat{\beta}_j$ in the response y.

• may be problematic in practice
e.g. idea of holding vars constant may not make sense for
+ observational data (not under our control)
+ $Y \sim X + X^2$

note: marginal effect of β_j, on Y may change based on presence/absence of other predictors in model
(Simpson's Paradox - more on this later)

Q: $\hat{\beta}_0$: expected value of response Y when all x_j's = 0
• may not make literal sense depending on context

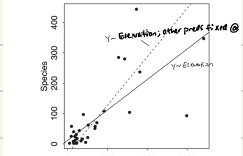


Figure 5.1 The fit for the simple model with just elevation as a predictor is shown as the solid line. The predicted response as elevation varies and the other four predictors held at their mean values is shown as a dashed line.

Effect Plot: visualizing the mean of the model for a given predictor
+ shows relationship btwn predictor and response for (larger) model
+ compute predicted responses for observed values of predictor, fixing the other predictors at their means

- observational Data
  - causal conclusion problematic; steps to make stronger case for causality:
    + try to include all relevant vars
    + try a variety of models and see whether a similar effect is observed
    + multiple studies under different conditions can help confirm a relationship
  - unmeasured lurking var Z may be the real cause and/or be a confounding variable

Simpson's Paradox: regression coef of a predictor may change sign when adding/deleting another predictor;

e.g. $E(Y|X,Z) = \beta_0 + \beta_1 X + \beta_2 Z$

Data only have X and Y. F.t. $Y \sim X$:

$$E(Y|X) = E(\beta_0 + \beta_1 X + \beta_2 Z|X) = \beta_0 + \beta_1 X + \beta_2 E(Z|X)$$

Suppose a) $X \perp Z$ (e.g. randomization) $\Rightarrow E(Z|X) = E(Z)$.

$$\Rightarrow E(Y|X) = [\beta_0 + \beta_2 E(Z)] + \beta_1 X$$

b) X and Z correlated (e.g. observational)

and assume $Z = \gamma_0 + \gamma_1 X + \varepsilon \Rightarrow E(Z|X) = \gamma_0 + \gamma_1 X$

$$\Rightarrow E(Y|X) = (\beta_0 + \beta_2 \gamma_0) + (\beta_1 + \beta_2 \gamma_1) X$$

Experimental Design

Q: How can we establish causal conclusion(s)?

- In designed experiments, our control over the experimental conditions allows us to make stronger conclusions from the analysis

\# Randomization can reduce the effect of unknown predictors not included in model

+ only reliable way to ensure data not unbalanced that favors either treatment/control
+ randomly assign experimental units to chosen values of the predictors
+ still doesn't necessarily ensure success; bad model, unrealistic controlled expt., etc.

\# Remark (Randomization): Let X_1: treatment (or not), X_2: unobserved (Data mean=0, var=1)

$$\Rightarrow Cov(X_1, X_2) = 0 \Rightarrow \text{sample } \gamma_1 \text{ and } \gamma_2 \approx \text{orthogonal}$$

$\Rightarrow$ Motivates how randomization can achieve causal conclusions

Orthogonality

- orthogonal predictors (by design) can also simplify the problem
  + allows us to more easily interpret the effect of one predictor w/o regard to another

Suppose we partition $X_{n\times(p+1)} = [\underset{n\times(p+1-g)}{X_{(1)}} \mid \underset{n\times g}{X_{(2)}}]$ s.t. $X_{(1)}^T X_{(2)} = 0$.

Now,

$$Y = X\beta + \varepsilon = \underset{(p+1-g)}{X_{(1)}}\beta_{(1)} + \underset{g}{X_{(2)}}\beta_{(2)} + \varepsilon$$

and

$$X^T X = \begin{pmatrix} X_{(1)}^T X_{(1)} & X_{(1)}^T X_{(2)} \\ X_{(2)}^T X_{(1)} & X_{(2)}^T X_{(2)} \end{pmatrix} = \begin{pmatrix} X_{(1)}^T X_{(1)} & 0 \\ 0 & X_{(2)}^T X_{(2)} \end{pmatrix}$$

which gives

$$\hat{\beta}_{(1)} = (X_{(1)}^T X_{(1)})^{-1} X_{(1)}^T Y \quad \text{and} \quad \hat{\beta}_{(2)} = (X_{(2)}^T X_{(2)})^{-1} X_{(2)}^T Y$$

$$\Leftrightarrow \hat{\beta} = (X^T X)^{-1} X^T Y = \begin{pmatrix} (X_{(1)}^T X_{(1)})^{-1} & 0 \\ 0 & (X_{(2)}^T X_{(2)})^{-1} \end{pmatrix} \begin{pmatrix} X_{(1)}^T \\ X_{(2)}^T \end{pmatrix} Y$$

$$= \begin{pmatrix} (X_{(1)}^T X_{(1)})^{-1} X_{(1)}^T Y \\ (X_{(2)}^T X_{(2)})^{-1} X_{(2)}^T Y \end{pmatrix}.$$

\# note: $\hat{\beta}_{(1)}$ will be the same regardless of whether $X_{(2)}$ is in model/not

Practical Difficulties

- Nonrandom samples
- Choice of Range of Predictors
- Model Misspecification
- Publication and Experimenter Bias
- Practical and Statistical Significance

Ch. 4: Diagnostics

Textbook	Lecture content
4.1: checking Error Assumptions	checking Error Assumptions
4.2: Finding unusual observations	Practical Error checking & Diagnostics
4.3: checking Structure of Model	(some) Handling of violated Assumptions

Diagnostics

- checking error assumptions
- Finding unusual points
- checking the model structure

Checking error assumptions

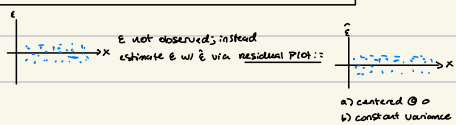
Assume:

$\varepsilon | X \overset{iid}{\sim} N(0, \sigma^2 I)$

a) $E(\varepsilon|X) = 0$
b) $\text{var}(\varepsilon|X) = \sigma^2 I$
c) ε's are ind.; identically distributed, Normal

remark: typically check in this order

a/b):



ε not observed; instead estimate ε w/ $\hat{\varepsilon}$ via residual plot ::

a) centered @ 0
b) constant variance (spread) of $\hat{\varepsilon}$

corollary: It can be shown that $\hat{\varepsilon} = (I_n - H)Y = (I_n - H)\varepsilon$

Proof. $\hat{\varepsilon} = Y - \hat{Y} = Y - HY$, $H = X(X^TX)^{-1}X^T \Rightarrow \hat{Y} = X\hat{\beta} = X(X^TX)^{-1}X^TY = HY$

$\Rightarrow \hat{\varepsilon} = (I_n - H)Y$

and

$HY = H(X\beta + \varepsilon) = H X\beta + H\varepsilon$

$= X\beta + H\varepsilon$, since $HX = X$

thus, $\hat{\varepsilon} = \overbrace{X\beta + \varepsilon}^{Y} - \overbrace{(X\beta + H\varepsilon)}^{\hat{Y}} = \varepsilon - H\varepsilon$

$\Rightarrow \hat{\varepsilon} = (I_n - H)\varepsilon$. ∎

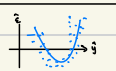
Residual Plot: Plot $\hat{\varepsilon}$ against x or any linear combination of x's (i.e. $\hat{y} = x^T\hat{\beta}$) to check for:

i) Non-linearity
→ expect a constant at mean function equal to 0 when model is correct, i.e.

ii) Homoscedasticity (Equal variance)

iii) Heteroscedasticity (unequal variance)

Remarks (violations): i)



- possible quadratic (non-linear) relationship

ii)

- fanning effect
- Heteroscedasticity

note: we want variance of residuals to be constant (iii) across all levels of the independent variable(s)

Prop (Residual Properties).

Treat $\hat{\varepsilon}$: random variable

(i): $E(\hat{\varepsilon}|X) = 0 \Leftrightarrow E(Y|X) = X\beta$

(ii): $\text{var}(\hat{\varepsilon}|X) = \sigma^2(I_n - H)$ if $\text{var}(\varepsilon|X) = \sigma^2 I_n$

(iii): $\text{cov}(\hat{\varepsilon}, \hat{Y}|X) = 0$

Treat $\hat{\varepsilon}$: samples. then

i) mean of $\hat{\varepsilon} = 0$, i.e. $\bar{\hat{\varepsilon}} = 0$

where $\bar{\hat{\varepsilon}} = \frac{1}{n}\sum_{i=1}^{n} \hat{\varepsilon}_i = 0$

ii) Sample cov. of $\hat{\varepsilon}$ and $\hat{Y} = 0$, i.e. $\widehat{\text{cov}}(\hat{\varepsilon}, \hat{y}) = 0$

remarks: • (i) - (3) hold if model assumptions satisfied;
• ii) and iii) always hold - even in spite of violated assumptions but not used for diagnosis

Q: Suppose SLR s.t. $\hat{\varepsilon} \sim \hat{y}$.

then $\hat{\beta}_1 = 0$ and $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$

$= \bar{y}$

$= 0$.

Proof (1). $E(\hat{\varepsilon}|X) = E\left[(I_n - H)\varepsilon | X\right] = (I_n - H)E(\varepsilon|X) = 0$. ∎

Proof (2). $\text{Var}(\hat{\varepsilon}|X) = \text{var}((I_n - H)\varepsilon|X)$

$= (I_n - H)\underbrace{\text{var}(\varepsilon|X)}_{\sigma^2 I_n}(I_n - H)$, since $(I_n - H)$: symmetric

$= \sigma^2(I_n - H)(I_n - H)$

$= \sigma^2(I_n - 2H + H^2)$

$= \sigma^2(I_n - H)$, since $H^2 = H$. ∎

Proof (3). $\text{cov}(\hat{\varepsilon}, \hat{Y}|X) = \text{cov}((I_n - H)Y, HY)$

$= (I_n - H)\underbrace{\text{var}(Y|X)}_{\text{var}(\varepsilon|X) = \sigma^2 I_n} H$

$= \sigma^2(I_n - H)\cdot H$

$= \sigma^2(H - H^2)$

$= 0$.

Proof (i). $\widehat{\text{cov}}(\hat{\varepsilon}, \hat{y}) = \frac{1}{n-1}\sum_{i=1}^{n}(\hat{\varepsilon}_i - \bar{\hat{\varepsilon}})(\hat{y}_i - \bar{\hat{y}})$

where $\bar{\hat{\varepsilon}} = \frac{1}{n}\sum_{i=1}^{n}\hat{\varepsilon}_i = 0$ by (i)

and $\bar{\hat{y}} = \frac{1}{n}\sum_{i=1}^{n}\hat{y}_i$

$= \frac{1}{n-1}\sum_{i=1}^{n}\hat{\varepsilon}_i(\hat{y}_i - \bar{\hat{y}})$

$= \frac{1}{n-1}\left(\sum_{i=1}^{n}\hat{\varepsilon}_i\hat{y}_i - \underbrace{\left(\sum_{i=1}^{n}\hat{\varepsilon}_i\right)}_{0 \text{ by (i)}}\bar{\hat{y}}\right)$

$= \frac{1}{n-1}\hat{\varepsilon}^T\hat{y}$

$= \frac{1}{n-1}\hat{\varepsilon}^T X\hat{\beta}$

$= 0 \in \mathbb{R}^{(n)\times 1}$ by normal eq. $X^T\hat{\varepsilon} = 0 \in \mathbb{R}^{(p+1)\times 1}$ ∎

Assumption checking (Practicality)

1) zero mean function assumption (linearity),

e.g. fit $\hat{\epsilon} \sim \hat{y} + \hat{y}^2$ and check (t-test) whether $\hat{y}^2$ is significant

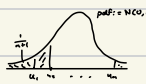
2) Fit $|\hat{\epsilon}| \sim \hat{y}$ and check (t-test) whether slope term is significant (constant variance)
(evidence for non-constant variance (heteroscedasticity))

3) Q-Q plot (Normality)

a) sort the residuals $\hat{\epsilon}_{(1)} \leq \hat{\epsilon}_{(2)} \leq \cdots \leq \hat{\epsilon}_{(n)}$

b) compute $u_i = \Phi^{-1}\left(\frac{i}{n+1}\right), i = 1, \ldots, n$ where Φ: cdf of $N(0,1)$ s.t.
↳ theoretical normal quantiles

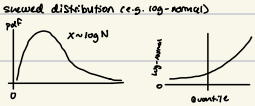
$\Phi(x) = P(N(0,1) < x)$:



$P(N(0,1) < u_1) = \frac{1}{n+1}$
$P(u_1 < N(0,1) < u_2) = \frac{1}{n+1}$
⋮
$P(N(0,1) > u_n) = \frac{1}{n+1}$

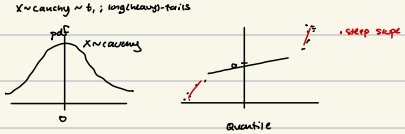
c) Plot $\hat{\epsilon}_{(i)}$ against u_i

Ex (Non-Normality): i) Skewed distribution (e.g. log-normal)



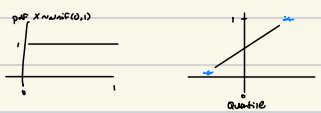
ii) Long-tailed dist. (e.g. cauchy)

X ~ cauchy ~ t_1 ; long(heavy)-tails



iii) Short-tailed dist. (e.g. uniform)

X ~ Unif ; short(light)-tail



Shapiro-Wilk test for normality: H_0: residuals are normal
- not very helpful or used in practice (QQ plot better)
- Small n: little power
- Large n: non-normality is less important ; non-normality mitigated by large n

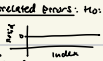
4) Correlated Errors

In temporally related data:
- Plot $\hat{\epsilon}$ against time
- Plot $\hat{\epsilon}_i$ against $\hat{\epsilon}_{i-1}$
- Time series analysis may be more appropriate

Remark: If there is no temporal relationship or other ordering in the variables, checking independence is more challenging

Durbin-Watson test for correlated errors: H_0: uncorrelated errors

- can create an index plot:
↳ look for successive patterns as evidence for correlated residuals



5) Finding Unusual Points

outliers: do not fit the model well

influential (points): affect the fit of the model substantially

note: a point can be none, one, or both of these

Leverage: Recall $H = X(X^TX)^{-1}X^T$ ($\hat{y} = Hy$)

leverage of point i := $h_i = H_{ii}$

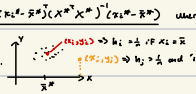
- h_i only depends on X
- $V(\hat{\epsilon}_i) = \sigma^2(1-h_i)$; recall: $\text{var}\hat{\epsilon}|X) = \sigma^2(I_n - H)$
- $\frac{1}{n} \leq h_i \leq 1$; By definition, $h_i = x_i'(X^TX)^{-1}x_i$ where $x_i' = \begin{pmatrix} x_{i1} \\ \vdots \\ x_{ip} \end{pmatrix}$

$V(\hat{\epsilon}_i) \geq 0$,
$\sigma^2(1-h_i) \geq 0$,
$h_i \leq 1$.

Let $x_i^{*T} = \begin{pmatrix} x_{i1}' \\ \vdots \\ x_{ip}' \end{pmatrix}$. Then,

$h_i = \frac{1}{n} + (x_i^* - \bar{x}^*)'(X^{*T}X^*)^{-1}(x_i^* - \bar{x}^*)$ where $\bar{x}^* = \frac{1}{n}\sum x_i^*$, $X^* = \begin{pmatrix} (x_1^* - \bar{x}^*)^T \\ \vdots \\ (x_n^* - \bar{x}^*)^T \end{pmatrix}$

Suppose $x_i^* = \bar{x}^*$, i.e.



- $\sum h_i = p+1$; Pf $\sum h_{ii} = \text{tr}(H) = \text{tr}(X(X^TX)^{-1}X^T)$
$= \text{tr}((X^TX)^{-1}X^TX)$
$= \text{tr}(I_{p+1}) = p+1$.

- Rule of Thumb: leverages greater than $\frac{2(p+1)}{n}$ are considered 'high'

  note: $\overline{h_{ii}} = \frac{1}{n}\sum h_{ii} = \frac{1}{n}(p+1)$.

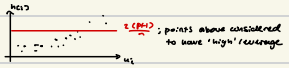
Remark: If $h_{ii} \to 1$, then

$$\text{var}(\hat{\varepsilon}_i) = \sigma^2(1-h_{ii}) \to 0 \Rightarrow \hat{\varepsilon}_i \to 0$$

Half-normal Plot

- Sort $h_{(1)} \le h_{(2)} \le \cdots \le h_{(n)}$
- compute $u_i = \Phi^{-1}\left(\frac{n+i}{2n+1}\right)$

- Plot $h_{(i)}$ against u_i



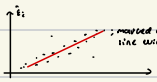
; points above considered to have 'high' leverage

Studentized Residuals

Since $\text{var}(\hat{\varepsilon}_i) = \sigma^2(1-h_i)$, define

(internally) studentized residuals: $r_i = \dfrac{\hat{\varepsilon}_i}{\hat{\sigma}\sqrt{1-h_i}}$ where $\text{se}(\hat{\varepsilon}_i) = \hat{\sigma}\sqrt{1-h_i}$

- Better to use studentized residuals for diagnostic plots (e.g. QQ-plot and testing constant variance)
- Practically, results are often similar (as compared w/ using $\hat{\varepsilon}_i$)



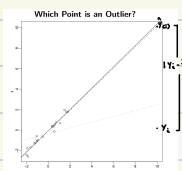
; marked deviation from straight line evidence of influential points)

Outliers

Q: How do we distinguish between truly unusual points and large residuals

- Exclude point i, recompute $\hat{\beta}_{(i)}$ and hence $\hat{y}_{(i)} = x_i^T\hat{\beta}_{(i)}$
- If $|y_i - \hat{y}_{(i)}|$ is large, then obs. i is an outlier

Q: How large is large



$\Rightarrow$ excluding i obs

$|y_i - \hat{y}_{(i)}|$ 'large'; evidence i is an outlier

(Externally) Studentized Residuals

It turns out

$$t_i = \frac{y_i - \hat{y}_{(i)}}{\hat{\sigma}_{(i)}\sqrt{1 + x_i^T(X_{(i)}^T X_{(i)})^{-1}x_i}} \quad (*)$$

where $\sqrt{\text{var}(y_i - \hat{y}_{(i)})}$ Same as prediction interval for a new obs. (ch. 4)

$$= r_i\left(\frac{n-(p+1)-1}{n-(p+1)-r_i^2}\right)^{1/2} \quad (**)$$

$\sim t_{n-(p+1)-1}$ under H_0: i^{th} obs. not outlier

Multiple Hypothesis Tests

- If $|t_i|$ is too large (small p-values), reject H_0 and conclude obs. i is an outlier
- For each obs. i, compare $|t_i|$ with $t^{\alpha/2}_{n-(p+1)-1}$

Note: When (simultaneously) conducting more than one hypotheses,

- we will reject too many points

Bonferroni correction

Type 1 Error (rate) $= P_{H_0}$(Reject at least one test)

$$\le \sum_{i=1}^{n} P_{H_0}(\text{reject } i\text{-th test})$$

$$= n\alpha$$

Thus, introduce the Bonferroni correction: test each hypothesis at level α/n

Remarks on Outliers

- two/more outliers could hide each other
- cluster of outliers: consider using robust methods
- Examine the context - what could it mean?
  - + data entry errors
  - + lurking variables
  - + Something gone wrong, e.g. fraudulent credit card use
  - + a new unknown effect
  - + some patterns just have exceptions

Influential Points

Def: one whose removal from the dataset would cause a large change in the fit

Outlier vs. High Leverage

· outlier & influential · outlier but not influential · influential but not outlier

· measure the influence via

+ $\hat{\beta} - \hat{\beta}_{(i)}$ — $\hat{\beta}$: OLS using all data; $\hat{\beta}_{(i)}$: OLS excluding i-th obs.

+ $X(\hat{\beta} - \hat{\beta}_{(i)}) = \hat{y} - \hat{y}_{(i)}$ (matrix)

Cook's Distance

$$\text{Cook statistic} := D_i = \frac{(\hat{y} - \hat{y}_{(i)})^T (\hat{y} - \hat{y}_{(i)})}{(p+1)\hat{\sigma}^2} = \frac{(\hat{\beta} - \hat{\beta}_{(i)})^T X^T X (\hat{\beta} - \hat{\beta}_{(i)})}{(p+1)\hat{\sigma}^2}$$

where recall: $\text{var}(\hat{\beta}) = \hat{\sigma}^2 (X^T X)^{-1}$

$$= \frac{1}{p+1} r_i^2 \frac{h_{ii}}{1-h_{ii}}, \quad \begin{cases} r_i := \text{standardized (studentized) residuals} \\ t_i := \text{externally studentized residuals} \end{cases}$$

+ combination of residual and leverage effect

+ $D_i \approx 1$ is a reference value for potential influential points

Checking the Structure of the Model

· Plot $\hat{\varepsilon}$ against $\hat{y}$ and x_j - but other predictors impact the relationship; want to isolate the effect of x_j on y (conditional on the other predictors)

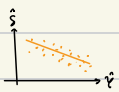
consider:

(i): Partial Regression Plots (Added Variable Plots)

i) Regress y on all x except x_j; get residuals $\hat{\delta}$

ii) Regress x_j on all x except x_j; get residuals $\hat{\gamma}$

iii) Plot $\hat{\delta}$ against $\hat{\gamma}$



· slope is $\hat{\beta}_j$; look for non-linearity, outliers, and influential points

· linearity satisfied

· no outliers or influential points

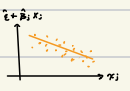
note: ↓ Same as $\hat{\beta}_j$; when fitting $Y \sim$ all X's; (i)-(iii) help address more diagnostics

(ii): Partial Residual Plots: Plot $\hat{\varepsilon} + \hat{\beta}_j x_j$ against x_j

Q: where does this come from?

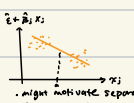
A: $Y = \sum_{j' \neq j} x_{j'} \hat{\beta}_{j'} + x_j \hat{\beta}_j + \hat{\varepsilon}$

$\Rightarrow Y - \sum_{j' \neq j} x_{j'} \hat{\beta}_{j'} = x_j \hat{\beta}_j + \hat{\varepsilon}$



· linearity satisfied

vs.



· might motivate separate regression for x_j < a and x_j > a (grouping analysis)

Summary of Diagnostics

· Just fitting a model is not good enough

· Graphical diagnostics are more informative but also more subjective

· Diagnostics often suggest a change in the model and then the whole process is repeated

· time-consuming but necessary

What to Do w/ Violated Assumptions

· Heteroscedasticity
  + nothing
    · use other methods for inference, e.g. bootstrap
  + weighted least squares (Ch. 8)
  + transformation of the response (Ch. 9)

· Nonlinearity: change the model { · add new predictors, e.g. $x^2, x^3, \ldots$ · transformation (Ch. 9)

· Non-Normality
  + nothing
    → especially for large n via CLT and applicable for short-tailed dist. - NOT heavy tail dist (e.g. Cauchy)
  + transformation of the response (skewed errors)
  + Robust methods (long-tailed dist.)
  + Inference based on other distributions

· Correlated Errors
  + Generalized least squares (Ch. 6)

Ch. 5: Problems with the Predictors

Textbook	Lecture content
5.1: Errors in the Predictors 5.2: changes of scale 5.3: collinearity	(MSmt) Error in Predictors change of scale Standardizing variables collinearity

Errors in Predictors: Suppose $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$.

then let $\tilde{x_i} = x_i + \delta_i$, δ_i - (measurement) error

- measurement error in X (preds) usually would lead to bias
- (if $\text{var}(\delta) >> \text{var}(\beta)$), then the bias is small and can practically be ignored

change of scale:
- Predictors of similar magnitude are easier to compare
- Numerical stability
- can aid interpretation

consequences:
- rescaling x_j leaves the t and F tests and $\hat{\sigma}^2$ and R^2 unchanged

Suppose $x_j \to \frac{x_j + a}{b}$, j-th predictor

while $\hat{\beta}_j \to b\hat{\beta}_j$

$\sqrt{\widehat{var}(\hat{\beta}_j)} = \widehat{SE}(\hat{\beta}_j) \to b\,\widehat{SE}(\hat{\beta}_j) = \sqrt{b^2 \widehat{var}(\hat{\beta}_j)}$

$Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_j x_j + \cdots + \beta_p x_p + \varepsilon$

since $x_j \to \frac{x_j + a}{b}$ and $\beta_j x_j = (\beta_j b)\frac{x_j + a}{b} - \beta_j a$

$\Rightarrow \beta_j \to \beta_j \cdot b$ and $\beta_0 \to \beta_0 - \beta_j a$, other β_j's, ε's unchanged

- Rescaling y leaves the t and F tests and R^2 unchanged but both $\hat{\sigma}$ and $\hat{\beta}_0$ rescaled by b; $\hat{\beta}_0$ is both shifted by a and rescaled by b:

Suppose $Y \to \frac{Y + a}{b}$.

then, $\frac{Y+a}{b} = \frac{\beta_0 + \beta_1 x_1 + \cdots + \beta_j x_j + \cdots + \beta_p x_p + \varepsilon + a}{b}$

$\Rightarrow \beta_0 \to \frac{\beta_0 + a}{b}$, $\beta_1 \to \frac{\beta_1}{b}, \ldots, \beta_p \to \frac{\beta_p}{b}$, $\varepsilon \to \frac{\varepsilon}{b}$

Standardizing variables:
- convert all variables to standard units (mean=0, sd=1), i.e.

for $i = 1, \ldots, n$, $j = 1, \ldots, p$, $x_{ij} \to \frac{x_{ij} - \bar{x}_j}{\text{sd of } x_{\cdot j}}$

for $i = 1, \ldots, n$, $Y_i \longrightarrow \frac{Y_i - \bar{Y}}{\text{sd}(Y_1, \ldots, Y_n)}$

- can compare coefficients directly
- Helps numerical stability, e.g. $(X^TX)^{-1}$ computation
- interpretation is nicer

note: Suppose X and Y standardized. then,

$\bar{Y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}_1 + \cdots + \hat{\beta}_p \bar{x}_p$

but all $\bar{x}_1, \ldots, \bar{x}_p$ are zero

$\Rightarrow \hat{\beta}_0 = 0$.

Collinearity: X^TX close to singular (not invertible)

cause: some columns are (almost) linear combos of others

Detection
- correlation matrix: large pairwise correlation
  - may not work for collinearity including more than two predictors, e.g. $x_1 \simeq x_2 + x_3$
- Regress x_j on other preds - get R^2_j.
  - $\to R^2$ close to 1 (e.g. $R^2_j > 0.9$) indicates a problem
- condition number of X^TX: $= \boxed{\kappa = \sqrt{\frac{\lambda_1}{\lambda_{p+1}}}}$

where λ_1: largest eigenvalue of X^TX

λ_{p+1}: smallest eig-val of X^TX

e.g. Suppose $X^TX = \begin{pmatrix} 1 & & \\ & \ddots & \\ & & 1 \end{pmatrix} \Rightarrow \kappa = \sqrt{1/1} = 1$

$X^TX = \begin{pmatrix} 1 & & \\ & \ddots & \\ & & 0.01 \end{pmatrix} \Rightarrow \kappa = \sqrt{1/0.01} = \sqrt{100} = 10$

* large κ indicates (evidence) of collinearity

$\Rightarrow$ practically, $\kappa > 20$ strong (presence) of evidence

consequences: • imprecise estimate of β, i.e. $\widehat{SE}(\hat{\beta}_j)$ large

• t-test fails to reveal significant preds: $\widehat{SE}(\hat{\beta}_j)$ large $\Rightarrow$ $t_j = \frac{\hat{\beta}_j}{\widehat{SE}(\hat{\beta}_j)} \to 0$ → large p-value → FTR H_0

• sensitivity to measurement errors

• Numerical instability

Variance Inflation Factor: Let $S_{x_j} = \sum_i (x_{ij} - \bar{x}_j)^2$. Then, it can be shown that

$$Var(\hat{\beta}_j) = \sigma^2 \left(\frac{1}{1-R_j^2}\right) \frac{1}{S_{x_j}}$$

• If x_j is independent (orthogonal) from other predictors, $Var(\hat{\beta}_j) = \frac{\sigma^2}{S_{x_j}}$

• If $R_j^2 \to 1$, $Var(\hat{\beta}_j) \to \infty$

and we define: $\boxed{VIF_j := \frac{1}{1-R_j^2}}$

↑ 'large' VIF indicates a problem

e.g. $VIF > 10$ ($\Leftrightarrow R_j^2 > 0.9$)

How to resolve collinearity: • If you mostly care about inference, drop highly correlated predictors

• variable selection may be used

• If interpretation is important and you must keep all predictors, do not use least squares; use some other estimation method, e.g. ridge regression

Ch.6: Problems with the Error

Textbook	Lecture content
6.1: Generalized Least Squares	Weighted Least Squares
6.2: Weighted Least Squares	Generalized Least Squares
6.3: Testing for Lack of Fit	Robust Regression
6.4: Robust Regression	

What can go wrong w/ the errors

Assume $\varepsilon \sim N(0, \sigma^2 I)$, i.e.

1) unequal variance: $Var(\varepsilon_i) \neq Var(\varepsilon_j)$

2) correlated: $Corr(\varepsilon_i) \neq Cor(\varepsilon_j)$

3) Heavy-tailed

Weighted Least Squares

Errors uncorrelated, but unequal variance, i.e.

$$\varepsilon \sim N(0, \sigma^2 W^{-1}) \quad \text{where } \sigma^2 W^{-1} = \begin{pmatrix} \sigma^2/w_1 & & 0 \\ & \sigma^2/w_2 & \\ 0 & & \ddots \\ & & & \sigma^2/w_n \end{pmatrix}_{n \times n}$$

$$\text{and } W^{-1} = diag\left(\frac{1}{w_1}, \dots, \frac{1}{w_n}\right), \; W = \begin{pmatrix} w_1 & & 0 \\ & \ddots & \\ 0 & & w_n \end{pmatrix}_{n \times n}$$

Remark: if $\underbrace{w_i = 1 \text{ for } i = 1, \dots, n}_{W = I} \Rightarrow Var(\varepsilon) = \sigma^2 I$ (OLS)

Ex: • Error variance proportional to a predictor, i.e. $Var(\varepsilon_i) \propto x_{i2} \Rightarrow w_i = x_{i2}^{-1}$

• Suppose y_i is the average of n_i observations. Then $w_i = n_i$

Proof. $Var(y_i) = Var\left(\frac{1}{n_i} \sum_{j=1}^{n_i} x_{ij}\right)$

assuming x_{ij} independent, $Var(x_{ij}) = \sigma^2$

$= \frac{\sigma^2}{n_i} \Rightarrow w_i = n_i$. ∎

Estimates

Under (WLS) model: $Y_i = \beta^T x_i + \varepsilon_i$, $i = 1, \dots, n$, $Var(\varepsilon_i) = \sigma^2/w_i$

$\Leftrightarrow Y = X\beta + \varepsilon$, $\varepsilon \sim N(0, \sigma^2 W^{-1})$

Note that

$Var(W^{1/2} \varepsilon_i) = Var(\sqrt{w_i}\, \varepsilon_i) = \sqrt{w_i}^2 \, Var(\varepsilon_i) = w_i \frac{\sigma^2}{w_i} = \sigma^2$

Now, assume the following

transformation: $\begin{cases} y_i \to \sqrt{w_i}\, y_i \\ x_i \to \sqrt{w_i}\, x_i \end{cases}$

Thus, $\underbrace{\sqrt{w_i}\, y_i}_{y_i^*} = \beta^T \underbrace{\sqrt{w_i}\, x_i}_{x_i^*} + \underbrace{\sqrt{w_i}\, \varepsilon_i}_{\varepsilon_i^*}$

which satisfy OLS assumptions.

$\Leftrightarrow Y^* = X^* \beta + \varepsilon^*$

$W^{1/2} Y \quad W^{1/2} X \quad W^{1/2} \varepsilon$ with $Var(\varepsilon^*) = \sigma^2 I$

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

$= (X^T\omega^{1/2}\omega^{1/2}X)^{-1}X^T\omega^{1/2}\omega^{1/2}Y$

$= (X^T\omega X)^{-1}X^T\omega Y$ (1).

Similarly, $\text{var}(\hat{\beta}) = \sigma^2((X^{*T}X^*)^{-1}$

$= \sigma^2(X^T\omega X)^{-1}$ (2)

Letting $\hat{\varepsilon} = Y - X\hat{\beta}$. So,

$\hat{\varepsilon}^* = Y^* - X^*\hat{\beta} = \omega^{1/2}(Y - X\hat{\beta}) = \omega^{1/2}\hat{\varepsilon}$.

$RSS = \hat{\varepsilon}^{*T}\hat{\varepsilon}^* = \hat{\varepsilon}^T\omega\hat{\varepsilon}$.

Thus, $\hat{\sigma}^2 = \frac{1}{n-(p+1)}\cdot\hat{\varepsilon}^T\omega\hat{\varepsilon}$ (3) ; Thus:

estimates: $\begin{cases} \hat{\beta} = (X^T\omega X)^{-1}X^T\omega Y \\ \text{var}(\hat{\beta}) = (X^T\omega X)^{-1}\sigma^2 \\ \hat{\sigma}^2 = \frac{\hat{\varepsilon}^T\omega\hat{\varepsilon}}{n-(p+1)} \end{cases}$

Remarks: • non-null weights can be used to indicate that different obs. have different variances

• otherwise, unbalanced groups can cause biased estimates

Generalized Least Squares

Errors (potentially) correlated and unequal variance, i.e.

In general,

$\varepsilon \sim N(0, \sigma^2\Sigma)$ where Σ is a general covariance matrix that satisfies {Positive Def. / Symmetric}

e.g. $\Sigma^{-1/2}\varepsilon \sim N(0, \sigma^2 I)$ since $\text{var}(\Sigma^{-1/2}\varepsilon) = \Sigma^{-1/2}\sigma^2\Sigma\Sigma^{-1/2} = \sigma^2 I_n$

Practically, $\Sigma^{-1/2}$ is not uniquely defined. Instead,

choose $\Sigma = SS^T$

where S: lower triangular matrix (the Cholesky decomposition)

and

$S^{-1}\varepsilon \sim N(0, \sigma^2 I) \Rightarrow \text{var}(S^{-1}\varepsilon) = \sigma^2 I_n$

So, we have the following

transformation: $\begin{cases} y \to S^{-1}y \\ x \to S^{-1}y \\ \varepsilon \to S^{-1}\varepsilon \end{cases} \Rightarrow Y = X\beta + \varepsilon,\ \text{var}(\varepsilon) = \sigma^2\Sigma$

$\downarrow$

$S^{-1}Y = S^{-1}X\beta + S^{-1}\varepsilon$

$Y^* \quad X^* \quad \varepsilon^*$, $\text{var}(\varepsilon^*) = \sigma^2 I_n$, i.e. OLS

With

estimates: $\begin{cases} \hat{\beta} = (X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}Y \\ \text{var}(\hat{\beta}) = (X^T\Sigma^{-1}X)^{-1}\sigma^2 \\ \hat{\sigma}^2 = \frac{1}{n-(p+1)}\hat{\varepsilon}^T\Sigma^{-1}\hat{\varepsilon} \end{cases}$

Remarks: • WLS is a special case of GLS with $\Sigma = \omega^{-1}$

• Practically, we estimate Σ from data

considering correlated Errors: • data collected over time: errors could be correlated

• one of the simplest correlation structures over time is the

autoregressive model: $\varepsilon_{i+1} = \rho\varepsilon_i + \delta_i$

assuming $\begin{cases} AR(1), \text{ i.e. today's statistical error only depends on yesterday's} \\ \text{Stationarity, i.e. } \varepsilon_i \text{ has same mean and variance, } i = 1, \dots, n \\ \delta_i \overset{iid}{\sim} N(0, \tau^2), \ |\rho| < 1 \end{cases}$

Then, $E(\varepsilon_{i+1}) = \rho E(\varepsilon_i) + E(\delta_i) = 0$ $\quad V(\varepsilon_{i+1}) = \rho^2 V(\varepsilon_i) + V(\delta_i)$

where $E(\delta_i) = 0$ $\quad$ where $\text{var}(\varepsilon_{i+1}) = \text{var}(\varepsilon_i) = \sigma^2$

and $E(\varepsilon_{i+1}) = E(\varepsilon_i) = m$ $\quad \sigma^2 = \rho^2\sigma^2 + \tau^2$

$m = \rho m \Rightarrow m(1-\rho) = 0$ $\quad \sigma^2(1-\rho^2) = \tau^2$

$\Rightarrow E(\varepsilon_i) = 0$ $\quad \Rightarrow \text{var}(\varepsilon_i) = \frac{\tau^2}{1-\rho^2}$

and $\text{cov}(\varepsilon_i, \varepsilon_j) = \text{cor}(\varepsilon_i, \varepsilon_j) = \rho^{|i-j|}$ (4)

$\Leftrightarrow$

$$\Sigma = \begin{pmatrix} \frac{\tau^2}{1-\rho^2} & \cdots & (4) \\ \vdots & \ddots & \vdots \\ (4) & \cdots & \frac{\tau^2}{1-\rho^2} \end{pmatrix}$$

notice: Σ depends on unknown τ^2, ρ parameters

Prop. can test $H_0: \rho = 0$

$H_a: \rho \neq 0$; evidence for correlated errors, i.e. $\text{corr}(\varepsilon_i, \varepsilon_j) \neq 0$

(4) CI for ρ does not contain 0

Robust Regression

Main concern: heavy-tailed error distribution

1) M-estimation (R)- what we cover

2) Least-trimmed estimation (in textbook if interested)

M-Estimation: Find β to minimize $\sum_{i=1}^{n} L(y_i - x_i^T\beta)$

Remarks: • $L(\cdot)$ is the loss function, possible loss functions:

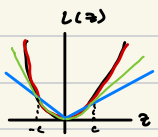
+ $L(a) = a^2$: least squares (LS)

Robust Estimation: less sensitive to large (residual) values

- should plot distribution of residuals

+ $L(z) = |z|$: least absolute regression (LAD); equivalently called median regression

$$\underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n} |Y_i - x_i^T\beta|$$



(c) $\text{median}(Y|x) = x_i^T\beta$

note: if dist. of ϵ_i is symmetric, then median = mean

Remark: in OLS: $\min_{\beta} \sum_{i=1}^{n} (y_i - \beta_0)^2 \Rightarrow \hat{\beta} = \bar{y}$

For LAD: $\min_{\beta} \sum_{i=1}^{n} |Y_i - \beta| \Rightarrow \hat{\beta} = \text{median}(Y_i)$

+ Huber's method: = (HM)

$$L(z) = \begin{cases} z^2/c, & |z| \le c \\ c|z| - c^2/c, & \text{otherwise} \end{cases}$$

- c should be a robust estimate of σ, e.g. median of $|\hat{\epsilon}_i|$

Ch. 7: Transformation

Textbook	Lecture content
7.1: Transforming the Response	Transforming the Response (Box-Cox method)
7.2: Transforming the Predictors	Transforming the Predictors

Outline

Transforming the response
- Box-Cox method

Transforming the predictors
- Polynomials
- Regression splines

Reasons to try transformations

- Nonlinearity

  Suppose $Y \sim X$: nonlinear but $g(Y) \sim X$: linear for g: transformation function

- Heteroscedasticity
- may improve fit
- incorporate a physical law (or some other domain relationship)

e.g. 

$\text{Area} = L \cdot W$

$\log \text{Area} = \log L + \log W$

Box-Cox Method

Transformation of the response $y \Rightarrow g_\lambda(y)$

A family of transformations indexed by λ when $y > 0$:

$$g_\lambda(y) = \begin{cases} \frac{y^\lambda - 1}{\lambda}, & \lambda \neq 0 \\ \ln(y), & \lambda = 0 \end{cases}$$

Remarks: · $\lambda = 1 \Rightarrow$ no transformation

· $\lim_{\lambda \to 0} \frac{y^\lambda - 1}{\lambda} = \log(y) = \ln(y)$

Prop. For $\lambda \neq 0$, $\frac{y^\lambda - 1}{\lambda} \cdot [G_m(y)]^{1-\lambda}$ makes the unit of $g_\lambda(y)$ the same for different λ

where $G_m(y) = \exp\left\{\frac{1}{n}\sum_{i=1}^{n} \log(y_i)\right\}$:= geom mean of $y_1, \ldots, y_n$

Goal: Find λ s.t. $g_\lambda(y) \sim X$ follows linear model, i.e.

we minimize RSS_λ from regression: $g_\lambda(y) \sim X$.

Equivalently, we can instead

compute likelihood of the data using the normal assumption

for any given λ.

Then,

choose λ to maximize $L(\lambda) = -\frac{n}{2}\ln\left(\frac{RSS_\lambda}{n}\right)$ (monotonic negative function)

- compute confidence intervals for λ using asymptotic dist. of the likelihood

Remarks: · may not choose λ that exactly maximizes $L(\lambda)$ (i.e. min. RSS_λ), but instead choose one that is easily interpreted, e.g.

$$\{-2, -1, 0, \pm\tfrac{1}{2}, +\tfrac{1}{3}, 1, 2\tfrac{1}{3}\}$$

- sensitivity to outliers, e.g.

$\hat{\lambda} = 5$ is extreme; practically, $\lambda \in [-2, 2]$ is often used

(likely due to outliers)

- If some $y_i \leq 0$, can add constant
  i.e. transform Y s.t. $Y \geq 0$ e.g. $\min Y = -1 \Rightarrow Y^* = Y + 1$ is now valid for BC-transformation
- Transformations of proportions, counts – elect to use generalized linear models (see later)
  + If y_i's are proportions (i.e. $\in (0,1)$), consider logit (transformation): $= \ln\left(\frac{y}{1-y}\right)$

note: log transformation is often used in practice if the range of a variable (Y or X) is over several magnitudes different (e.g. income)

Transforming the Predictors

Given $Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \varepsilon$ for predictors $\begin{cases} x_1 \\ \vdots \\ x_p \end{cases}$

Now suppose

$Y = \beta_0 + \beta_1 f_1(x) + \cdots + \beta_q f_q(x) + \varepsilon$ for (new) predictors $\begin{cases} f_1(x) \\ \vdots \\ f_q(x) \end{cases}$

Basis functions: $= f_j(x)$, e.g. $\begin{cases} \text{Polynomials} \\ \text{Regression splines} \end{cases}$

Polynomials: one Predictor case

$Y = \beta_0 + \beta_1 x_i + \cdots + \beta_p x_i^d + \varepsilon$ (R: lm(Y ~ X + I(X^2) + ... + I(X^d))

Q: How to choose d:

Forward: = keep adding terms until new term is not statistically significant

Backwards: = start w/ large d – keep eliminating the non-significant highest order term

Remarks (ISSUES): (1) as d↑, x_i^{d-1} and x_i^d may be highly correlated

(2) as d↑, x_i^d too large (if $|x| > 1$) or too small (if $|x| < 1$) $\Rightarrow (X^T X)^{-1}$ problematic

- usually don't eliminate lower order terms, even if not significant, when a higher order term is significant (Hierarchy)
- transforming predictor (e.g. $x_i = x_i - \bar{x}$) can affect significance of higher-order terms w.r.t. predictor (x_i)
  recall: Previously familiar w/ unchanged t-test(s) for transformed preds (w.r.t. exclusively linear terms, i.e. deg = 1)

orthogonal Polynomials

Suppose $p=1$, x_{i1}, $i = 1, \ldots, n$ for original predictors.

Assume polynomial original predictors (d degree), $i = 1, \ldots, n$ s.t. $x_i = \begin{pmatrix} x_{i0}=1 \\ x_{i1} \\ x_{i2} \\ \vdots \\ x_{id} \end{pmatrix}_{(d+1)\times 1}$

Then, for numerical stability, $\forall\, i = 1, \ldots, n$:

new predictors: $= \begin{cases} z_{i0} = 1 \text{ (intercept)} \\ z_{i1} = a_1 + b_1 x_{i1} \\ z_{i2} = a_2 + b_2 x_{i1} + c_2 x_{i1}^2 \\ \vdots \\ z_{id} = a_d + b_d x_{i1} + \cdots + h_d x_{i1}^d \end{cases} \Rightarrow z_i = \begin{pmatrix} z_{i0} \\ z_{i1} \\ \vdots \\ z_{id} \end{pmatrix}_{(d+1)\times 1}$

(R: lm(Y ~ Poly(X, d), "raw" = F)

default: orthogonal transform

note: "raw" = T returns higher-order terms for preds w/o orthogonal transform

where $a, b, c, \ldots$ are chosen s.t. $\tilde{z}_j^T \tilde{z}_i = 0$ (orthogonal) when $j \neq i$, i.e.

where $\tilde{z}_j = \begin{pmatrix} z_{1j} \\ z_{2j} \\ \vdots \\ z_{nj} \end{pmatrix}_{n\times 1}$ for $j = 0, \ldots, d$

and prediction matrix: $= Z_{n\times(d+1)} = \begin{pmatrix} | & & | \\ -\tilde{z}_0 & \cdots & -\tilde{z}_d \\ | & & | \end{pmatrix} = \begin{bmatrix} z_1^T \\ \vdots \\ z_n^T \end{bmatrix} = [\tilde{z}_0, \cdots \tilde{z}_j \cdots \tilde{z}_d]_{n\times(d+1)}$

Thus, for polynomial regression $Y \sim Z$, then

(new) $\hat{\beta}^* = \begin{pmatrix} \hat{\beta}_0^* \\ \vdots \\ \hat{\beta}_d^* \end{pmatrix} = (Z^T Z)^{-1} Z^T Y$

where Z is orthogonal s.t. $Z^T Z = \begin{pmatrix} \tilde{z}_0^T \tilde{z}_0 & & 0 \\ & \tilde{z}_1^T \tilde{z}_1 & \\ & & \ddots \\ 0 & & \tilde{z}_d^T \tilde{z}_d \end{pmatrix}$

$\Rightarrow \hat{\beta}_j^* = \frac{\tilde{z}_j^T Y}{(\tilde{z}_j^T \tilde{z}_j)}$ for $j = 0, \ldots, d$

Practically, we can further normalize over $\tilde{z}_j$ s.t.

$\tilde{z}_j^T \tilde{z}_j = 1$, $j = 0, \ldots, d$ (!!)

i.e., for e.g. $z_{i0} = \frac{1}{\sqrt{n}}$, $i = 1, \ldots, n$ (intercept)

normalizing constant z_{i0} (!!) holds

$\Rightarrow \hat{\beta}_j^* = \tilde{z}_j^T Y$ for $j = 1, \ldots, d$

note: Practically, easier to softly choose d as highest-order significant polynomial term
(orthogonal $\Rightarrow$ no correlation btwn different-order polynomial terms
$\Rightarrow$ hierarchical structure simplified, i.e.
more unlikely to have contradictory results as seen w/ non-orthogonal preds)

Polynomials in several predictors

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{12} x_1 x_2 + \varepsilon$$

($d=2$; interaction term)

R: `lm(y ~ polym(x1, x2, degree = 2))` — by default here still fitting orthogonal polynomials

Remarks:

- can use F-test to select d for, e.g.

  $H_0: d=1$ vs. $H_a: d=2$

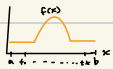
- For now, assume d is the same for all predictors we introduce (>1 pred)

Regression Splines (Nonparametric Statistics)

Disadvantage of polynomials: each data affects the fit globally

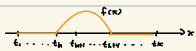
Remedy: B-spline

In general, suppose $E(Y|x) = f(x)$ where



cubic (B-spline) basis functions on interval (a, b) w/ pre-specified knots $t_1, \ldots, t_K$:

- non-zero on interval defined by four successive knots and zero elsewhere ⇒ local influence property



where, e.g.

$$\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \begin{matrix} = 3 < t_k \\ = 3.5 \in (t_k, t_{k+1}) \\ \\ = 10 > t_{k+1} \end{matrix} \Rightarrow \begin{bmatrix} f_k(x_1) \\ f_k(x_2) \\ \vdots \\ f_k(x_n) \end{bmatrix} \begin{matrix} = 0 \\ \neq 0 \\ \\ = 0 \end{matrix}$$

s.t. different k gives different predictors

where f_k takes (cubic) polynomial form that we skip

Then, fit $Y \sim f_1(x) + \cdots + f_K(x) + \cdots$ (the $f_k(x)$ terms: new predictors)

- cubic polynomial fit to each four successive knots
- smooth
- integrates to one

Other transformations

- Smoothing splines
- Generalized additive model ($p>1$)
- CART, MARS, MART, neural networks

Rule of thumb:

- For large datasets, complex models are better (w/ appropriate control of the number of parameters)
- For small datasets or high noise levels (e.g. social sciences), standard regression is more appropriate

Ch: Factors (Qualitative variables):

Material covered
Factors (Qualitative preds) · reg. w/ one factor · reg. w/ one factor and (other) cont. preds · main effects model
Multiple Factors
F-Test

Outline

- Regression w/ one factor only
- Regression w/ one factor and other (continuous) predictors

Factors

- Qualitative (categorical) predictors; can have two/more levels

One-Factor Models

- Factor vars can be included in a MLR mean function using dummy variables

Suppose $d=2$: · single dummy var: $\begin{cases} 0 : \text{category 1} \\ 1 : \text{category 2} \end{cases}$

+ assignment of labels to values is generally arbitrary, and will not change the outcome of the analysis

+ can be alternatively defined w/ different set of values, e.g. $\{1, -1\}$, $\{1, 2\}$, etc. (must have two distinct values)

Suppose $d=3$: j^{th} dummy variable U_j for the factor, $j = 1, \ldots, d$ has the i^{th} value u_{ij} for $i = 1, \ldots, n$, i.e.

$$u_{ij} = \begin{cases} 1 : \text{group}_i = j^{th} \text{ category of group} \\ 0, \text{ otherwise} \end{cases}$$

Ex (UN Data)

Observational Study; n = 199

Response: lifeExpF

Group: factor variable (d = 3): $\begin{cases} U_1 := \text{oecd} \\ U_2 := \text{other} \\ U_3 := \text{africa} \end{cases}$

where $U_1 + U_2 + U_3 = 1$

	Group	U_1	U_2	U_3
Afghanistan	other	0	1	0
Albania	other	0	1	0
Algeria	africa	0	0	1
Angola	africa	0	0	1
Anguilla	other	0	1	0
Argentina	other	0	1	0
Armenia	other	0	1	0
Aruba	other	0	1	0
Australia	oecd	1	0	0
Austria	oecd	1	0	0

$= \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}$

Remark: if we add an intercept to the mean function, the resulting model is overparameterized

b/c $U_1 + U_2 + U_3 = 1$, a col. of 1's, and the column of 1's is the regressor that corresponds to int.

Account for this problem by dropping one of the dummy vars (e.g. U_1):

$$E(Y) = \beta_0 + \beta_2 U_2 + \beta_3 U_3$$

Since $U_2 = U_3 = 0 \Rightarrow U_1 = 1$

$$\Rightarrow E(Y \mid U_2 = U_3 = 0, U_1 = 1) = \beta_0 + \beta_2(0) + \beta_3(0) = \beta_0$$

Similarly, $E(Y \mid U_1 = U_3 = 0, U_2 = 1) = \beta_0 + \beta_2(1) + \beta_3(0) = (\beta_0 + \beta_2)$

and $E(Y \mid U_1 = U_2 = 0, U_3 = 1) = \beta_0 + \beta_2(0) + \beta_3(1) = (\beta_0 + \beta_3)$

Prop (0): Let Y-response (numeric), X-explanatory (cat.)

lm(Y ~ factor(X))

Table 5.1 Regression Summary for Model (5.4)

	Estimate	Std. Error	*t*-Value	Pr(>|*t*|)
(Intercept), $\hat{\beta}_0$	82.4465	1.1279	73.09	0.0000
other, $\hat{\beta}_2$	−7.1197	1.2709	−5.60	0.0000
africa, $\hat{\beta}_3$	−22.6742	1.4200	−15.97	0.0000

$\hat{\sigma} = 6.2801$ with 196 *df*, $R^2 = 0.6191$.

→ $\hat{E}(Y \mid \text{Oecd} = 1) = 82.45$; testing $H_0: \mu_{\text{oecd}} = 0$ not practically meaningful

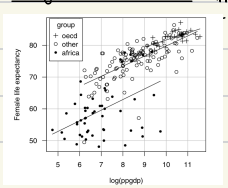
testing for $H_0: \mu_{\text{oecd}} = \mu_{\text{other}}$ and $H_0: \mu_{\text{oecd}} = \mu_{\text{Africa}}$ respectively

3 groups are

$$\hat{E}(\texttt{lifeExpF} \mid \texttt{group} = \texttt{oecd}) = \hat{\beta}_0 + \hat{\beta}_2 0 + \hat{\beta}_3 0 = 82.45$$
$$\hat{E}(\texttt{lifeExpF} \mid \texttt{group} = \texttt{other}) = \hat{\beta}_0 + \hat{\beta}_2 1 + \hat{\beta}_3 0 = 82.45 - 7.12 \quad (5.5)$$
$$\hat{E}(\texttt{lifeExpF} \mid \texttt{group} = \texttt{africa}) = \hat{\beta}_0 + \hat{\beta}_2 0 + \hat{\beta}_3 1 = 82.45 - 22.67$$

Remarks: df = n − (p+1); here, p = 2 ⇒ df = 199 − (2+1) = 196

Adding a continuous Predictor: Suppose we add log(ppgdp) as continuous var.



- can model w/ separate regressions for each factor level

+ three lines w/ separate int. and slope ∀ group

let group = j s.t.

	intercept	slope
j = 1 : oecd	γ_{01}	γ_{11}
j = 2 : other	γ_{02}	γ_{12}
j = 3 : africa	γ_{03}	γ_{13}

then $E(Y \mid \log(\text{ppgdp}) = x, \text{group} = j) = \gamma_{0j} + \gamma_{1j} x$

where $(\gamma_{0j}, \gamma_{1j})$ are the intercept and slope for level $j = 1, \ldots, d \Rightarrow$ 2d = 6 parameters

Instead, model is generally parameterized by

$$E(Y \mid \log(\text{ppgdp}) = x, \text{group}) = \beta_0 + \beta_{02} U_2 + \beta_{03} U_3 + \beta_1 x + \beta_{12} U_2 x + \beta_{13} U_3 x \quad (*)$$

Y ~ group + log(ppgdp) + group:log(ppgdp)

||

Y ~ group * log(ppgdp) where * : expands to include all main effects / interactions

1) OECD: $U_1 = 1, U_2 = U_3 = 0 \Rightarrow E(Y \mid x, \text{oecd}) = \beta_0 + \beta_1 x$

$\Leftrightarrow \gamma_{01} = \beta_0, \gamma_{11} = \beta_1$

2) other: $U_2 = 1, U_1 = U_3 = 0 \Rightarrow E(Y \mid x, \text{other}) = (\beta_0 + \beta_{02}) + (\beta_1 + \beta_{12}) x$

$\Leftrightarrow \gamma_{02} = \beta_0 + \beta_{02}, \gamma_{12} = \beta_1 + \beta_{12}$

3) africa: $U_3 = 1, U_1 = U_2 = 0 \Rightarrow E(Y \mid x, \text{africa}) = (\beta_0 + \beta_{03}) + (\beta_1 + \beta_{13}) x$

$\Leftrightarrow \gamma_{03} = \beta_0 + \beta_{03}, \gamma_{13} = \beta_1 + \beta_{13}$

Remark: parameters (β_0, β_1) are the int. and slope for the baseline level; remaining β's are the differences btwn the other levels & baseline

Table 5.3 Regression Summary for Model (5.7)

	Estimate	Std. Error	*t*-Value	Pr(>|t|)
(Intercept), $\hat{\beta}_0$	59.2137	15.2203	3.89	0.0001
1st order: `other`, $\hat{\beta}_{02}$	−11.1731	15.5948	−0.72	0.4746
1st order: `africa`, $\hat{\beta}_{03}$	−22.9848	15.7838	−1.46	0.1470
1st order: log(ppgdp), $\hat{\beta}_1$	1.5544	1.0165	1.53	0.1278
2nd ord: `other`: log(ppgdp), $\hat{\beta}_{12}$	0.6442	1.0520	0.61	0.5410
2nd ord: `africa`: log(ppgdp), $\hat{\beta}_{13}$	0.7590	1.0941	0.69	0.4887

$\hat{\sigma} = 5.1293$ with 193 *df*, $R^2 = 0.7498$.

Note: all terms are non-significant; likely due to correlation among terms

→ motivates removing higher-order (interaction) terms (i.e. main effects model)

Main Effects Model: Set (*) model ⟶ $E(Y|\log ppg = x, group) = \beta_0 + \beta_{02}U_2 + \beta_{03}U_3 + \beta_1 x$

$\beta_{12} = \beta_{13} = 0$ ⟶ all 3 groups have same slope

(~ group + log(ppdgp))

Main Effects Plot(s)

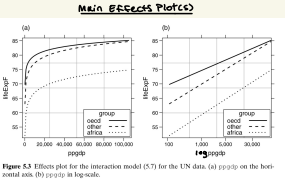


Figure 5.3 Effects plot for the interaction model (5.7) for the UN data. (a) `ppgdp` on the horizontal axis, (b) `ppgdp` in log-scale.

Remarks:
- dropping int. terms from (*)
- allows each group to have its own intercept
- all groups have the same slope
  - + difference between the levels of the factor are the same for every fixed value of the cont. regressor
  - + the effect of the cont. regressor is the same for all levels of the factor

Table 5.4 Regression Summary for Model (5.8)

	Estimate	Std. Error	*t*-value	Pr(>|*t*|)
(Intercept), $\hat{\beta}_0$	49.5292	3.3996	14.57	0.0000
1st order: `other`, $\hat{\beta}_{02}$	−1.5347	1.1737	−1.31	0.1926
1st order: `africa`, $\hat{\beta}_{03}$	−12.1704	1.5574	−7.81	0.0000
1st order: log(ppgdp), $\hat{\beta}_1$	2.2024	0.2190	10.06	0.0000

$\hat{\sigma} = 5.1798$ with 195 *df*, $R^2 = 0.7422$. — note: now significant

Many Factors (>1)

- Increasing # of factors on a continuous Y in a mean function can add complexity, does not raise new fundamental issues

EXCWOOL-EXP.2

Table 5.5 The Wool Data

Variable	Definition
Y — `len`	Length of test specimen (250, 300, 350 mm)
X_1 — `amp`	Amplitude of loading cycle (8, 9, 10 mm)
X_2 — `load`	Load put on the specimen (40, 45, 50 g)
X_3 — log(`cycles`)	Logarithm of the number of cycles until the specimen fails

note: there are $3^3 = 27$ possible combinations of the 3 factors used

Main effects mean fn: $Y \sim X_1 + X_2 + X_3$

where X_i: 2 dummy vars, $i = 1, 2, 3$

⇒ Int. + 2(3) = 7 params

Let $X_1: u_{11}, u_{12}, u_{13}$, $X_2: u_{21}, u_{22}, u_{23}$, $X_3: u_{31}, u_{32}, u_{33}$

where u_{ij}: baseline (ref) group, $i = 1, 2, 3$

$$E(Y|x_1, x_2, x_3) = \beta_0 + \underbrace{\beta_{12}u_{12} + \beta_{13}u_{13}}_{x_1} + \underbrace{\beta_{22}u_{22} + \beta_{23}u_{23}}_{x_2} + \underbrace{\beta_{32}u_{32} + \beta_{33}u_{33}}_{x_3}$$

⇒ $E(Y|u_{12}=1, u_{21}=1, u_{31}=1) = \beta_0$ (all 3 factors set to low level)

⇒ $E(Y|u_{12}=1, u_{21}=1, u_{31}=1) = \beta_0 + \beta_{12}$, i.e. β_{12}: diff. btwn u_{12} and u_{11} w.r.t. X_1, holding low level values constant for X_2 and X_3

Full second-order mean fn: $Y \sim X_1 + X_2 + X_3 + X_1X_2 + X_1X_3 + X_2X_3$

- adds all the two-factor interactions to the mean function

(7 + 3(4)) = 19 params

Full third-order mean fn: $Y = X_1 + X_2 + X_3 + X_1X_2 + X_1X_3 + X_2X_3 + X_1X_2X_3$

- includes 3-factor interactions

(19 + 8) = 27 params

F-Tests (Ch. 3)

- For any linear regression model, the RSS measures the amount of variation in response not explained by regressors
- if the NH were false, then the $RSS_{AH} \ll RSS_{NH}$ under AH model
  - ⇒ Basis of F-test: enough evidence against the NH if difference ($RSS_{NH} - RSS_{AH}$) is large enough

NH: $E(Y|X_1=x_1, X_2=x_2) = x_1'\beta$, $\Leftrightarrow \beta_2 = 0$

AH: $E(Y|X_1=x_1, X_2=x_2) = x_1'\beta_1 + x_2'\beta_2$

anova(M_{H_0}, $M_{H_0 \cup H_A}$) ⇒

	Res.DF	RSS	DF	Sum of Square	F	Pr(>F)
H_0	df_{H_0}	RSS_{H_0}				
H_A	df_{H_A}	RSS_{H_A}	$df_{H_0} - df_{H_A}$	$RSS_{H_0} - RSS_{H_A}$	F-stat	p-value

where $H_0 : H_0 \cup H_A$

Note: t-test are appropriate for testing individual predictors (not, for e.g. interaction terms)

Remarks: F-test requires nested structure, i.e. cannot test:

~~$H_0 : Y \sim X_1 + X_2$ vs. $H_A : Y \sim X_1 + X_3$~~

Ch. 10: Variable Selection

Material covered

- Testing-based approaches
  - Backward elimination
  - Forward selection
  - Stepwise regression
- Criterion-based approaches
  - AIC and BIC
  - Adjusted R^2
  - Mallow's C_p

Testing-Based Approaches

Idea: test significance of predictors and eliminate in some principled fashion

- Based on individual p-values
- Multiple testing is not accounted for - but ranking is more important than the absolute size of p-values (not using TIE control - e.g. Bonferroni correction)
- Different methods use different rules to add/delete predictors

Backward Elimination (assumes: $n > p$)

1) Start w/ all predictors in the model
2) Remove the predictor w/ the highest p-value $> \alpha$
3) Refit model → return to 2
4) Stop when all p-vals $< \alpha$

Note: $\alpha > 0.05$ may be better if prediction is the goal

Forward Selection ("Better" in practice for large p)

1) Start w/ no pred. vars
2) ∀ predictor not in model, check the p-value if added to model
3) Add the one w/ the smallest p-value $< \alpha$
4) Refit model and return to (2)
5) Stop when no new predictors can be added

$Y \sim 1$ ⇒ { $Y \sim X_1$ ← smallest p-value, $Y \sim X_2$, ⋮, $Y \sim X_p$ } ⇒ { $Y \sim X_1 + X_2$, ⋮, $Y \sim X_1 + X_p$ } ⇒ ...

Stepwise regression: combination of backward elimination and forward selection (allows to add variables back after they have been removed)

Remarks:
- Greedy; may miss the optimal model,

  e.g. $Y \sim X_1 + \cdots + X_p \Rightarrow 2^p$ possible (subset) models

  forward or backward ≤ $p + (p-1) + (p-2) + \cdots + 1 = \frac{(p+1)p}{2}$ at most
- Remember not to take p-values at face value (multiple testing)
- Variables not selected can still be correlated w/ the response, but they do not improve the fit enough to be included
- tend to pick smaller models than desirable for prediction purposes

(Information) Criterion-Based Selection Procedure

Idea: choose the model that optimizes a criterion which balances goodness-of-fit (RSS) and model size (p)

- Implicitly uses either forward or backward selection
- No p-values
- Some theoretical guarantees
- Different methods use different goodness-of-fit measures and different penalties for model size

AIC and BIC

Akaike information criterion: $\boxed{AIC) := n \ln(RSS/n) + 2(p+1)}$

{step(..., k=2)}

- (theoretically), suppose candidate models not X_{true}, then choose AIC

Bayes information criterion: $\boxed{BIC) := n \ln(RSS/n) + (p+1)\ln n}$

- larger weight for p → often selects smaller models (minimizing)

{step(..., k=log(n))}

- (theoretical) selection consistency property holds: If $\underbrace{\text{candidate model}}_{x^{true}}$, choose BIC

Remark: Practically, could get different model choices

Likelihood: A general def. of AIC/BIC:

For model C, $AIC = -2\log(\text{max. likelihood of model C}) + 2(\#\text{ params model C})$

$BIC = -2\log(\text{max. likelihood of model C}) + \log n(\#\text{ param.s in C})$

Linear Regression Models:

$$\text{Model (True)} \begin{cases} Y_i = X_i^T\beta + \varepsilon_i \ , i = 1,\dots,n \\ \varepsilon_i \overset{iid}{\sim} N(0,\sigma^2) \text{ or } \varepsilon_{n\times 1} \sim N_n(0, \sigma^2 I_n) \end{cases}$$

$\Leftrightarrow Y_i | X_i \overset{ind}{\sim} N(X_i^T\beta, \sigma^2)$, $i = 1,\dots,n$

the joint prob. density function (pdf) of $Y = (Y_1,\dots,Y_n)^T$ given X:

$$p(Y_1,\dots,Y_n | X; \beta, \sigma^2) = \prod_{i=1}^n p(Y_i | X_i; \beta, \sigma^2)$$

$$= \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{\frac{-(Y_i - X_i^T\beta)^2}{2\sigma^2}\right\} =: \underbrace{L(\beta, \sigma^2 | X, Y)}_{\text{Data } (X,Y)} : \text{function of } \beta, \sigma^2 \text{ given}$$

The estimates Maximizing the likelihood function $L(\beta, \sigma^2 | X, Y)$ or, equivalently,

$$\text{log-likelihood} := \log L(\beta, \sigma^2 | X, Y) = \sum_{i=1}^n \left[\log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) - \frac{(Y_i - X_i^T\beta)^2}{2\sigma^2}\right] = \frac{-n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^n (Y_i - X_i^T\beta)^2$$

are called the maximum likelihood estimators (MLE) $:= (\hat\beta_{MLE}, \hat\sigma^2_{MLE})$

It can be shown that OLS: $\hat\beta = \hat\beta_{MLE}$;

to get $\hat\beta_{MLE}$, $\frac{\partial \log L(\beta, \sigma^2 | X, Y)}{\partial \beta} = \frac{\partial \sum_{i=1}^n (Y_i - X_i^T\beta)^2}{\partial \beta} = 0 \Rightarrow \hat\beta_{MLE} = \hat\beta_{OLS}$.

$\hat\sigma^2_{MLE}$: $\frac{\partial \log L}{\partial \sigma^2} = 0 \Rightarrow \frac{-n}{2}\cdot\frac{1}{\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum_{i=1}^n (Y_i - X_i^T\beta)^2 = 0$

$$\Rightarrow \hat\sigma^2_{MLE} = \frac{1}{n}\sum_{i=1}^n (Y_i - X_i^T\hat\beta_{MLE}) = \frac{RSS}{n}$$

$\neq \hat\sigma^2 = \frac{RSS}{n-(p+1)}$; but, $\hat\sigma^2 - \hat\sigma^2_{MLE} \xrightarrow{p} 0$ when: $\left|\begin{matrix} n \to \infty \\ p \text{ fixed} \end{matrix}\right.$

$E(\hat\sigma^2) = \sigma^2$, i.e. unbiased and bias: $\hat\sigma^2_{MLE} \to 0$.

Back to AIC/BIC:

For linear model C, $\log(\text{max. likelihood model C}) = \frac{-n}{2}\log(2\pi\hat\sigma^2_{MLE}) - \frac{1}{2\hat\sigma^2_{MLE}}\sum_{i=1}^n \underbrace{(Y_i - X_i^T\hat\beta_{MLE})^2}_{RSS_C}$

$$= \frac{-n}{2}\log\left(\frac{RSS_C}{n}\right) - \frac{n}{2} + \text{const.} \quad \left| \ x(-2) \right.$$

$$= n\log\left(\frac{RSS_C}{n}\right) + \text{const}$$

Adjusted R^2

Recall: $R^2 = 1 - \frac{RSS}{TSS}$

$$\text{adjusted } R^2 := R^2_a = 1 - \frac{RSS/(n-(p+1))}{TSS/(n-1)}$$

$$= 1 - \left(\frac{n-1}{n-(p+1)}\right)(1-R^2)$$

- adding a predictor will not necessarily increase R^2_a
- Maximizing R^2_a is equivalent to minimizing RSE $\hat\sigma$

Mallow's Cp

$$C_p := \frac{RSS_p}{\hat\sigma^2} + 2(p+1) - n$$

- $\hat\sigma^2$ is estimated from model w/ all predictors
- RSS_p is from the model w/ p predictors (candidate model)

Goal: minimize Cp

- $C_p \le p+1$ indicates good fit
- Cp estimates the (test data) MSE: $\frac{1}{n}\sum_i E(\hat Y_i - E(Y_i))^2$

[Suppose we have another set of data w/ same X_i's as original data but Y^*_i's new responses]; then, $MSE = \frac{1}{n}\sum_i (Y_i^* - \hat Y_i)^2$

Remark: Information criteria can be used together with: (AIC/BIC/R^2_a/Cp)
- Forward selection
- Backward selection
- Best subset selection

Variable Selection Summary

- sensitive to outliers
- Greedy, criterion-based methods are preferred
- It may happen that several models provide very similar fit
  - If models w/ similar fit lead to very different conclusions, the data are ambiguous
  - If conclusions are similar, choose a simpler model and/or predictors that are easier to measure

ch 11: Shrinkage Methods: | Outline
| Ridge Regression
| LASSO
| (Skipping PLS and PCA)

Ridge Regression

Penalizing the square of the coefficients:

$$\min_{\beta} \underbrace{\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2}_{RSS(\beta)} + \lambda \sum_{j=1}^{p}\beta_j^2 \quad (1)$$; Lagrangian function for l_2-see below)

- coefficients $\hat{\beta}^{ridge}$ are shrunken towards zero
- $\lambda \geq 0$ is tuning parameter (estimated)
- λ controls the amount of shrinkage: $\begin{cases} \lambda \to 0: \text{ OLS estimation, i.e. } \hat{\beta}^{ridge} \to \hat{\beta}^{OLS} \\ \lambda \to \infty: \hat{\beta}^{ridge} \to 0 \text{ for } j=1,\dots,p \end{cases}$ (since otherwise $\lambda \sum_j \beta_j^2 \to \infty$)

note: IF $\beta_j = 0$ for $j=1,\dots,p$, then (1) becomes $\min_{\beta_0} \sum_i (y_i - \beta_0)^2 \Rightarrow \hat{\beta}_0^{ridge} = \bar{y}$

$\Leftrightarrow \hat{\beta}_0^{ridge} \longrightarrow \bar{y}$ as $n \to \infty$

Equivalent Formulation: $\min_{\beta} (y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2$ subject to $\sum_{j=1}^{p}\beta_j^2 \leq s$ (2)

where $s \geq 0$ (tuning parameter) s.t.: $\begin{cases} s \to 0, \hat{\beta}^{ridge} \to 0 \text{ for } j=1,\dots,p \text{ and } \hat{\beta}_0^{ridge} \to \bar{y} \\ (\lambda > 0) \\ s \to \infty \ (\lambda = 0): \hat{\beta}^{ridge} = \hat{\beta}^{OLS} \end{cases}$

Why? - when there are many highly correlated variables (collinearity)

+ $\hat{\beta}^{OLS}$ may have large coefficient on one variable and a similarly large negative coefficient on its correlated variable (cancellation)
+ in ridge regression, the size constraint tries to avoid this phenomenon
- address (fit model) under high-dimensional problem ($p > n$) when OLS fails (X^TX not invertible)

Remark: ridge estimate is not equivariant under scaling of the predictors

- in practice, standardize the predictors first: $\tilde{x}_{ij} \leftarrow \frac{x_{ij} - \bar{x}_j}{sd(x_{ij})}$ for $\begin{cases} i=1,\dots,n \\ j=1,\dots,p \end{cases}$

Similarly, can center (or standardize) Y: $y_i \leftarrow y_i - \bar{y}$ or $\leftarrow \frac{y_i - \bar{y}}{sd(y_i)}$ for $i=1,\dots,n$

Ridge Estimation: Suppose X-standardized and Y-centered (standardized).

1) $\hat{\beta}_0^{ridge} = 0$ since $\frac{\partial \text{objective (1)}}{\partial \beta_0} = 0$

$\Rightarrow -2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_{i1} - \dots - \beta_p x_{ip}) = 0$

$\Rightarrow (\sum_i y_i) - n\beta_0 - \beta_1 \sum_i x_{i1} - \dots - \beta_p \sum_i x_{ip} = 0$

where Y-centered, X-standardized $\Rightarrow \sum y_i, \sum x_{ij} = 0$

$\Rightarrow \hat{\beta}^{ridge} = 0$.

2) The ridge problem is equivalent to the following w/o intercept:

$$\hat{\beta}_{p\times 1}^{ridge} = \underset{\beta_{p\times 1}}{\operatorname{argmin}} \sum_{i=1}^{n}(y_i - \beta^T x_i)^2 + \lambda \beta^T \beta$$ where $\beta = (\beta_1, \dots, \beta_p)^T$ (β_0 excluded); $x_i = (x_{i1}, \dots, x_{ip})^T$ ("1" term excluded)

$$= \underset{\beta_{p\times 1}}{\operatorname{argmin}} (Y - X\beta)^T(Y - X\beta) + \lambda \beta^T \beta$$ where $Y = (y_1, \dots, y_n)^T$, $X = (x_1^T, \dots, x_n^T)^T = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix}$

solving as done w/ OLS.

solution: $\hat{\beta}^{ridge} = (X^TX + \lambda I)^{-1}X^Ty$

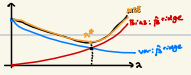
Remarks: - X-standardized, Y-centered (standardized)
- $\beta_{p\times 1}$ and $X_{n\times p}$ defined differently as above
- $\hat{\beta}$ is linear in y (as seen in OLS)
- $\hat{\beta}$ is biased (OLS: unbiased); $E(\hat{\beta}^{ridge}) \neq \beta$ (unless) for $\lambda > 0$
- even if X is not full rank, $(X^TX + \lambda I)$ is invertible
- $\hat{\beta}^{ridge}$ has smaller variance than OLS; thus, may have smaller MSE

MSE: For a new obs. (x^*, y^*),

$E(y^* - \hat{y}^*)^2 = Var(\hat{y}^*) + [Bias(\hat{y}^*)]^2 + \text{constant}$ (ignore $\hat{\sigma}^2$)

where $\hat{y}^* = x^*\hat{\beta}$ and $Bias(\hat{y}^*) = E(\hat{y}^*) - E(y^*) = (x^*)^T(E(\hat{\beta}) - \beta)$

- Emphasizes variance-bias trade-off:



Shrinkage in Ridge: Suppose orthonormal design ($X^TX = I$). Then, $\hat{\beta}^{ols} = X^Ty$ and

$$RSS(\beta) = (y - X\beta)^T(y - X\beta) = \underbrace{y^Ty}_{\text{constant}} - 2y^TX\beta + \beta^T\underbrace{X^TX}_{I}\beta$$

$$= \beta^T\beta - 2(\underbrace{\hat{\beta}^{ols}}_{\text{w/ } X^TX = I})^T\beta + \text{const.}$$

$$= (\beta - \hat{\beta}^{ols})^T(\beta - \hat{\beta}^{ols}) + \text{const}$$

$$= \text{const.} + \sum_{j=1}^{p}(\beta_j - \hat{\beta}_j^{ols})^2$$

$$\Rightarrow \text{(1) w/ an orthogonal design} = \sum_{j=1}^{p}(\beta_j - \hat{\beta}_j^{ols})^2 + \lambda\sum_{j=1}^{p}\beta_j^2$$

equivalent to component-wise minimization:

$$\min_{\beta_j}(\beta_j - \hat{\beta}_j^{ols})^2 + \lambda\beta_j^2 \rightarrow \frac{\partial(\cdots)}{\partial\beta_j} = 0 \Rightarrow \hat{\beta}_j^{ridge} = \frac{1}{1+\lambda}\hat{\beta}_j^{ols}$$

Remarks (Shrinkage in Ridge): · Shrink the estimate towards zero by a positive constant less than 1

$$\text{var}(\hat{\beta}_j^{ridge}) = \frac{1}{(1+\lambda)^2}\text{var}(\hat{\beta}_j^{ols}) \quad ; \quad \text{Bias}(\hat{\beta}_j^{ridge}) = \beta_j\left(\frac{-\lambda}{1+\lambda}\right)$$

2. Bias: $E(\hat{\beta}_j^{ridge}) - \beta_j$

$$= E\left(\frac{\hat{\beta}_j^{ols}}{1+\lambda}\right) - \beta_j$$

$$= \frac{\beta_j}{1+\lambda} - \beta_j = \beta_j\left(-\frac{\lambda}{1+\lambda}\right)$$

+ 2↑, shrinkage ↑, bias ↑, variance ↓

+ 2↓, shrinkage ↓, bias ↓, variance ↑

Model Assessment

objectives: (1) choose a value of a tuning parameter (λ) for a technique

(2) estimate the prediction performance of a given model

· Ran procedure on an independent test set (if available)

· one should use different test data for (1) and (2); a validation set for (1)
a test set for (2)

cross-validation

· often insufficient data to create a separate validation or test set; setting some data aside for validation is possible, but affects the accuracy of training estimates

K-Fold CV: 1) Divide the data into K disjoint subsets

2) Use subsets 2,...,K as training data and subset 1 as validation data
compute the MSE on subset 1, fitting over candidate models (different λ's)

3) Repeat for each subset

4) Average the result

Least Absolute Shrinkage and Selection operator (LASSO)

$$\min_{\beta}\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2 + \lambda\sum_{j=1}^{p}|\beta_j| \quad \Leftrightarrow \quad \min_{\beta}\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2 \text{ subject to } \sum_{j=1}^{p}|\beta_j| \le s$$

where $s \ge 0$ (tuning parameter) s.t.: $\begin{cases} s \to 0, \hat{\beta}^{lasso} \to 0 \text{ for } j = 1,\dots,p \text{ and } \hat{\beta}_0^{lasso} \to \bar{y} \\ s \to \infty\ (\lambda = 0): \hat{\beta}^{lasso} = \hat{\beta}^{ols} \end{cases}$

· Shrinkage: $\begin{cases} \lambda \to 0 & \hat{\beta}^{lasso} \to \hat{\beta}^{ols} \\ \lambda \to \infty & \hat{\beta}_j^{lasso} \to 0,\ j = 1,\dots,p \ \ \hat{\beta}_0^{lasso} = \bar{y} \end{cases}$

note: same as seen w/ Ridge regression

· Sparsity: some fitted coefficients are exactly zero

+ unique to lasso

+ motivates continuous variable selection

Soft Thresholding

Practically, we assume X-standardized, Y-centered (standardized)

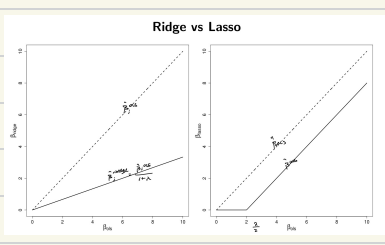
$\Rightarrow \hat{\beta}_0^{lasso} = 0$ and we don't consider the intercept term

When X is orthonormal ($X^TX = I$), we can minimize over β_j componentwise:

$$\min_{\beta_j}(\beta_j - \hat{\beta}_j^{ols})^2 + \lambda|\beta_j| \xrightarrow{\text{Solution}} \hat{\beta}_j^{lasso} = \begin{cases} \hat{\beta}_j^{ols} - \frac{\lambda}{2} & \text{if } \hat{\beta}_j^{ols} > \frac{\lambda}{2} \\ 0 & \text{if } |\hat{\beta}_j^{ols}| \le \frac{\lambda}{2} \\ \hat{\beta}_j^{ols} + \frac{\lambda}{2} & \text{if } \hat{\beta}_j^{ols} < -\frac{\lambda}{2} \end{cases} = \underbrace{\text{Sign}(\hat{\beta}_0^{ols}) \cdot (|\hat{\beta}_j^{ols}| - \frac{\lambda}{2})_+}_{\begin{matrix} 0 \text{ if } \le 0 \\ |\hat{\beta}_j^{ols} - \frac{\lambda}{2}| \text{ if } > 0 \end{matrix}}$$

· shrinks large coefficients by a constant ($\lambda/2$)

· truncates small coefficients to zero



Remarks: other popularly implemented approaches: group lasso (handling cat. issues all as one group)

Elastic Net (combine L1/L2 norm)

Summary

· Main reason to use Shrinkage: too many predictors or collinearity

· Interpretation is usually lost

· Ridge is still a linear model in original predictors but no inference } minimizing more nuanced loss function w/ standardized (centered) preds/response

· Prediction is usually improved by shrinkage

· All require selecting a tuning parameter

Ch: Binomial Data

outline

- Binomial data
- Generalized Linear Models (glm) for Binomial Data
- Inference for glm for binomial data
- odds Ratio
- overdispersion

Review: The Binomial Distribution

n independent trials: $z_1, \dots, z_n$

Suppose: $\begin{cases} P(z_i = 1) = p \text{ (success)} \\ P(z_i = 0) = 1-p \text{ (Failure)} \end{cases}$

Binomial rv: $= V = \sum_{i=1}^{n} z_i$ is the number of successes out of n iid trials

Prob Dist: $P(V=k) = \binom{n}{k} p^k (1-p)^{n-k}$, $k = 0, 1, \dots, n$

- $E(V) = np$, $V(V) = np(1-p)$; $\because E(z_i) = p$, $Var(z_i) = p(1-p)$ for $i = 1, \dots, n$
- as $n \to \infty$, Binomial $\to$ Normal: $\frac{V - np}{\sqrt{np(1-p)}} \to N(0,1)$ via CLT
- sample proportion (estimate of p): $\hat{p} = \frac{V}{n}$

Binomial Data: $i = 1, \dots, N$

- response y_i: number of successes out of n_i independent trials w/ prob. of success p_i

  where $y_i \sim \text{Binomial}(n_i, p_i)$, $E(y_i) = n_i p_i$

$x = (x_1, x_2, \dots, x_p)$: predictors (quantitative, Factors, or both)

assume: For all trials contributing to one response y_i, the predictors x_i have the same value (covariate class)

Goal: Model the relationship btwn y and $x_1, \dots, x_p$ via modeling the relationship btwn p_i and $x_1, \dots, x_p$

where $p_i = \frac{E(y_i)}{n_i}$

Binomial Regression

assume: $\begin{cases} y_i \sim \text{Binomial}(n_i, p_i) \\ y_i\text{'s are independent} \end{cases}$

want to construct a linear predictor, i.e.

$\eta_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$

Remark: cannot use $\eta_i = p_i$ since need $p \in (0,1)$

Idea: use a link transformation function

$\eta_i = g(p_i) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} \in \mathbb{R}$

Binomial Link functions:

logit: $= \eta = \log\left(\frac{p}{1-p}\right)$

$\Rightarrow$ logistic regression s.t. $\log\left(\frac{p_i}{1-p_i}\right) = x_i^T \beta$, $p_i = \frac{\exp\{x_i^T\beta\}}{1 + \exp\{x_i^T\beta\}}$ (sigmoid function)

where odds: $= \frac{p}{1-p}$ conducive to interpretation

Probit: $= \eta = \Phi^{-1}(p)$, Φ: cdf of $N(0,1)$

$\Rightarrow \Phi^{-1}(p_i) = x_i^T \beta$, $p_i = \Phi(x_i^T \beta)$

consider binary response, i.e. $n_i = 1$. Then, using probit link

is equivalent to $y_i = \begin{cases} 1 \text{ if } Y_i^* = x_i^T\beta + \varepsilon_i > 0 \\ 0 \text{ otherwise} \end{cases}$ where $\varepsilon_i \sim N(0,1)$

$\Rightarrow p_i = P(y_i = 1) = P(x_i^T\beta + \varepsilon_i > 0)$

$= P(-\varepsilon_i < x_i^T\beta)$, $\varepsilon_i \sim N(0,1)$

$= \Phi(x_i^T\beta)$, Φ: cdf of $N(0,1)$

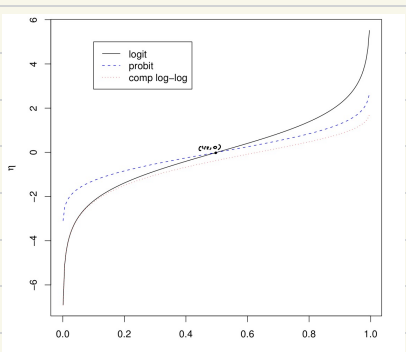
$\Rightarrow \Phi^{-1}(p_i) = x_i^T\beta$

* incorporates hidden variable that permits continuous response linear regression

complementary log-log: $\eta = \log(-\log(1-p))$

- less common in practice

note: all transform $p \in (0,1) \longrightarrow \eta \in (-\infty, \infty)$



Estimating parameters: Assume logistic model:

for $i=1,\dots,N$, $Y_i \sim \text{Binomial}(n_i, p_i)$ with $\log\left(\frac{p_i}{1-p_i}\right) = x_i^T\beta \Leftrightarrow p_i = \frac{\exp\{x_i^T\beta\}}{1+\exp\{x_i^T\beta\}}$

Maximum likelihood approach: find parameters (β in this case p_i) that maximize the likelihood of the data, i.e.

$$\prod_{i=1}^{N} P(Y_i = y_i) \text{ where } Y_i \sim \text{Binomial}(n_i, p_i)$$

$$= \prod_{i=1}^{N} \binom{n_i}{y_i} p_i^{y_i}(1-p_i)^{n_i - y_i}$$

$$\ell(p_1,\dots,p_N; y) = \sum_{i=1}^{N}\left[\log\binom{n_i}{y_i} + y_i \log(p_i) + (n_i - y_i)\log(1-p_i)\right]$$

and substitute p_i w/ corresponding sigmoid function: $\frac{e^{x_i^T\beta}}{1+e^{x_i^T\beta}}$

$\Rightarrow \ell(\beta; y)$ where β: only remaining unknown quantity

Need to maximize

$$\ell(\beta) = \sum_{i=1}^{N}\left[y_i(x_i^T\beta) - n_i \log(1+\exp(x_i^T\beta))\right] + \text{const (does not depend on } \beta)$$

obtain $\hat{\beta}_{MLE}$ via optimization (numerical) algorithm; no closed form solution

Note: If data is perfectly separable, $\hat{\beta}_{MLE}$ is not identifiable;
consider a shrinkage procedure (e.g. ridge/lasso) to introduce a penalty/constraint for solving

Inference

Z-test (wald test): $= \begin{cases} H_0: \beta_j = 0 \\ H_a: \beta_j \neq 0 \end{cases}$ for $j = 1,\dots,p$ predictors

where $Z_j := \frac{\hat{\beta}_j}{SE(\hat{\beta}_j)}$

note: similar to t-test in linear regression

Likelihood ratio test

- two nested models
- L is the larger model w/ ℓ parameters and likelihood L_L | $\hat{\beta}_{MLE}$; model L
- S is the smaller model w/ $s<\ell$ parameters and likelihood L_S | $\hat{\beta}_{MLE}$; model S
- the likelihood ratio test statistic $:= 2\log\frac{L_L}{L_S} = 2[\log L_L - \log L_S] \overset{H_0}{\sim} \chi^2_{\ell-s}$ ($\ell > s$)

note: similar to (general) F-test in OLS; measured difference in RSS under H_0 vs. H_a

For logistic regression, $LRT = 2\sum_{i=1}^{N}\left[Y_i \log\frac{\hat{p}_i^L}{\hat{p}_i^S} + (n_i - y_i)\log\frac{(1-\hat{p}_i^L)}{(1-\hat{p}_i^S)}\right]$

where $\hat{p}_i^L = \frac{\exp\{x_i^T\hat{\beta}^L\}}{1+\exp\{x_i^T\hat{\beta}^L\}}$,

$\frac{\hat{p}_i^L}{\hat{p}_i^S} = \frac{n_i\hat{p}_i^L}{n_i\hat{p}_i^S} = \frac{\hat{y}_i^L}{\hat{y}_i^S}$,

and $\frac{(1-\hat{p}_i^L)}{(1-\hat{p}_i^S)} = \frac{n_i - \hat{y}_i^L}{n_i - \hat{y}_i^S}$

Recall: $F_{\ell-s, n-(\ell+1)} := \frac{\chi^2_{\ell-s}}{\ell-s} \Big/ \frac{\chi^2_{n-(\ell+1)}}{n-(\ell+1)} \xrightarrow{n\to\infty} \chi^2_{\ell-s}$ (denominator $\to 1$ by LLN)

Deviance

- analog to RSS in OLS setting
- treat larger model to be the saturated model: $= Y_i \overset{ind}{\sim} \text{Binomial}(n_i, p_i)$, $i=1,\dots,N$;
  N parameters: $p_1,\dots,p_N$ to fit each data point perfectly,
  with fitted values $\hat{p}_i = y_i/n_i$
- in this case, the (test statistic) is called the deviance of S

$$D := 2\sum_{i=1}^{N}\left[Y_i \log\frac{y_i}{\hat{y}_i^S} + (n_i - Y_i)\log\frac{n_i - y_i}{n_i - \hat{y}_i^S}\right]$$

model S: (candidate (smaller) model compared to saturated model)

where $\hat{y}_i = n_i\hat{p}_i$, $\hat{p}_i$ are fitted probs from S

Prop. If Y_i's are truly binomial, independent, n_i are large then

$$D \overset{\cdot}{\sim} \chi^2_{n-s}$$

where: $\begin{cases} H_0: S \text{ model} \\ H_a: \text{saturated model} \end{cases}$

Remarks (uses of deviance): • D tests goodness-of-fit:

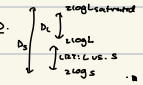
$$\text{p-value} = P(\chi^2_{n-s} > D)$$; small D ⇒ (+) good fit; large D ⇒ (-) poor fit

• compare two nested models, e.g. null (no predictors) and current model; in this case, use

$$2\log\frac{L_L}{L_S} = D_L - D_S \overset{\text{approx}}{\sim} \chi^2_{(n-s)-(n-l)} = \chi^2_{l-s}$$

where p-value: $P(\chi^2_{l-s} > D_L - D_S)$

Proof (visual):



• (If $n_i = 1$ (binary data), deviance cannot be used (for goodness-of-fit); but LRT can still be used to compare model S vs. L (S, L fixed))

• From R output: null deviance (from model w/ only intercept)
residual deviance (from fitted model)

note: AIC a general metric to compare all (including non-nested) models

Other measures of goodness-of-fit

• Help analyze contingency table (categorical data) analysis, e.g.

	Y	n − Y
i = 1	.	.
⋮	⋮	⋮
i = N	Y_N	$n_N - Y_N$
	Success	Failure

Here, $\chi^2 = \sum_{i=1}^{N}\left[\frac{(Y_i - \hat{Y}_i)^2}{\hat{Y}_i} + \frac{[(n_i - Y_i) - (n_i - \hat{Y}_i)]^2}{n_i - \hat{Y}_i}\right]$

NXZ

$$= \sum_{i=1}^{N}\frac{(Y_i - n_i\hat{p}_i)^2}{n_i\hat{p}_i(1-\hat{p}_i)}$$

Define the χ^2 goodness-of-fit statistic, i.e.

Pearson's $\chi^2 := X^2 = \sum_i \frac{(O_i - E_i)^2}{E_i}$

test: H_0: (fitted model)
H_a: saturated model (least restrictive)

where $\begin{cases} O_i: \text{observed count in each "bin" (e.g. } Y_i) \\ E_i: \text{expected count under tested model (e.g. } \hat{Y}_i) \end{cases}$

For binomial data, add successes and failures to get

$$X^2 = \sum_{i=1}^{N}\frac{(Y_i - n_i\hat{p}_i)^2}{n_i\hat{p}_i(1-\hat{p}_i)} \quad \& \quad \text{Pearson residuals: } r_i^P = \frac{(Y_i - \hat{Y}_i)^2}{\sqrt{\text{Var}(Y_i)}}$$

$$\Rightarrow X^2 = \sum_{i=1}^{N}(r_i^P)^2$$

Remark: typically $X^2 \approx$ deviance (D) and is practically used the same way, i.e.

$$X^2 \overset{H_0}{\sim} \chi^2_{N-3} = D$$

Evidence Intervals for Parameters

Asymptotically, $\hat{\beta}$ is normal, i.e. $\frac{\hat{\beta} - \beta_{\text{true}}}{SE(\hat{\beta})} \sim N(0,1)$

⇒ can use z-intervals; 95% CI: $\hat{\beta} \pm 1.96 SE(\hat{\beta})$

• Profile likelihood CI's are more accurate (based on considering the likelihood of one parameter w/ all others fixed)

Confidence Intervals for Predictions

• Predict probability of success $p(x_0) := \frac{\exp(x_0^T\hat{\beta})}{1+\exp(x_0^T\hat{\beta})}$ for a particular x_0

• No distinction btwn future observations & mean response

• Based on asymptotic normality of $\hat{\beta}$ and $x_0^T\hat{\beta}$

• Extrapolation will give unreliable predictions (as always)

Interpreting odds

$$\text{odds} = \frac{p}{1-p}$$

logistic regression (logit link) models log odds, i.e.

$$\log(\text{odds}) := \log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$

Interpretation: a unit increase in x_i w/ all other predictors held fixed leads to an increase of β_i in log-odds; equivalently, odds multiplied by e^{β_i} (where 1 is reference value)

• no such interpretation available for other link functions

What does a large deviance indicate?

• violation of model assumptions (outliers, non-linearity, model structure)

• Sparse data (small n_i)

• overdispersion

Overdispersion

the binomial model links the mean and the variance:

$\text{Var}(y_i) = n_i \hat{p}_i (1-\hat{p}_i)$; p_i parameter determines $E(Y_i)$ and $\text{Var}(Y_i)$
(not the case for normal data)

Overdispersion: = observed $\text{Var}(Y_i)$ is larger than the model postulates ($\frac{\sigma^2}{\phi} > 1$)

common causes: | terms are not independent
prob. of success is not constant

note: underdispersion is also possible but rare in practice

Estimating overdispersion

Introduce an additional **dispersion parameter** : $= \phi = \sigma^2$

s.t. $\text{Var}(Y_i) = \sigma^2 n_i p_i (1-p_i)$

can estimate σ^2 (as in linear regression) as

$$\hat{\sigma}^2 = \frac{X^2}{n-(p+1)}$$

Remarks:
- does not affect $\hat{\beta}$
- all standard errors must be multiplied by $\hat{\sigma}$
- Deviance can no longer be used to compare models (Binomial assumption no longer strictly met)

instead use an **approximate F-test** : $= \frac{(D_2 - D_1)/(df_2 - df_1)}{\hat{\sigma}^2}$ where $F_{df_2 - df_1, n-(p+1)}$

- Goodness-of-fit cannot be tested (same reason)
- Estimating overdispersion is only reasonable when n_i's are roughly equal

Summary

- with link functions, binomial data can be modeled easily
- approximate inference available for testing models and parameter values
- logit has advantages in interpretation

warnings:
- the estimation algorithm may not converge
- w/ small n_i, the X^2 approx. is poor
- overdispersion can be accounted for - but binomial assumption is sacrificed

---

Ch: Count Regression

Outline
Review of the Poisson Dist.
Poisson Regression
Inference via deviance
overdispersion
Ex: Galapagos data

Review: the Poisson Dist.

RV Y takes values 0, 1, 2, ...

Poisson Dist. has one parameter: $m > 0$

pdf: $P(Y=y) = \frac{e^{-m} m^y}{y!}$ $E(Y) = \text{Var}(Y) = m$

- can be used to approximate Binomial (n, p) if $\begin{cases} n \to \infty \\ p \to 0 \end{cases}$ then $np \to m > 0$
- if m is large, Y is approx normal, i.e. $\frac{Y-m}{\sqrt{m}} \to N(0,1)$ as $m \to \infty$
- if $Y_i \sim \text{Poisson}(m_i)$, Y_i - independent then $\sum_i Y_i \sim \text{Poisson}(m)$ where $m = \sum_i m_i$

Modeling count Data

Response y_i is a count; what dist. to assume?

- if count is bounded above, Binomial could be more appropriate
- if counts are large, normal approx. applies and regular linear regression may be used
- if count arises as number of 'failures' until a given number of 'successes', then negative binomial is appropriate