M1 Simple Linear Regression

Estimation Basics & Simple Linear Regression Notation

Linear relationship $y_i = ax_i + b$, slope a, intercept b

Functional relationship all y_i are determined by the line (fits line perfectly)

Statistical relationship all y follow overall trend but w/ error/deviations

written as $y_i = ax_i + b + e_i$, where e_i is the difference between each y_i & trend $ax_i + b$

Inference on Population Mean

Population $Y \sim N(\mu, \sigma^2)$, σ^2 unknown

Sample data $y_1, \cdots, y_n$

Estimate μ, $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$, s^2 with sample

Variability due to possible samples is represented w/ sampling distribution $\frac{\bar{y} - \mu}{\frac{s}{\sqrt{n}}} \sim T_{n-1}$

Inference on a Linear Trend

Population trend $Y = \beta_0 + \beta_1 X + \varepsilon$, a statistical relationship

- Y random response variable
- X fixed predictor variable
- ε random error given by $\varepsilon = Y - \beta_0 - \beta_1 X$

Functional part $\mathbb{E}[Y|X] = \beta_0 + \beta_1 X$

Sample data: pairs $(x_1, y_1), \cdots, (x_n, y_n)$

Need to estimate β_0 & β_1 from sample to estimate means/trend

Errors & Varition

Population errors (unknown) $\varepsilon_i = Y_i - (\beta_0 + \beta_1 X_i)$

Total error in population trend is

$$\sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} [Y_i - (\beta_0 + \beta_1 X_i)]^2 = \sum_{i=1}^{n} (Y_i - \mathbb{E}[Y_i|X_i])^2$$

similar to sd. numerator $\sum_{i=1}^{n} (y_i - \bar{y})^2$

Estimated trend $y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{e}_i$

sample errors are residuals $\hat{e}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i) = y_i - \hat{y}_i$

fitted values are estimated means $\hat{y}_i$

Measure total error around estimated trend using Residual Sum of Squares

$$RSS = \sum_{i=1}^{n} \hat{e}_i^2 = \sum_{i=1}^{n} (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2$$

Ordinary Least Squares Estimation Process

"Line of best fit" should fit snuggly among data points (least amount of distance error)

Instead of minimizing individual errors, minimize Residual Sum of Errors

- total variation around line in the data
- RRS is called estimating equation used for estimating unknown population trend

Use to find estimates for β_0 & β_1 to make RRS as small as possible

find line that makes all residuals as small as possible

Why squared?

Prevents cancellation of ⊕ & ⊖ residuals, easier to work w/ algebraically

Penalizes distant points more than closer

All possible lines/models in model space along w/ predictor info, response in a diff. space

model/line that is shortest distance from all points = orthogonal projection from y to model space

Ordinary Least Squared Steps

1. Set up estimating equation for given model w/ parameters
2. Take partial derivatives of estimating equation w/ respect to each unknown parameter
3. Set each derivative to 0
4. Solve for each unknown parameter

$$RSS = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

$$\frac{\partial RSS}{\partial \beta_0} = -2\sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

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

$$\frac{\partial RSS}{\partial \beta_1} = -2\sum_{i=1}^{n} x_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \rightarrow \hat{\beta}_1 = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum x_i^2 - n\bar{x}^2} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

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

deviations away from mean of x & mean of y

total variation in predictor

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

no relationship means $\hat{\beta}_0 = \bar{y}$

ONLY used for simple linear models (1 predictor)

If predictor is different form ($\hookrightarrow x^2$), need to rederive

Interpretation of Simple Linear Regression Estimates

Coefficients

$\mathbb{E}[Y|X] = \beta_0 + \beta_1 X$ estimating means

Intercept $\hat{\beta}_0$ is mean/average response @ predictor is 0 ——— is it meaningful/realistic?

Slope $\hat{\beta}_1$ is change in mean/average/expected for one-unit increase in value of predictor

NOT same as all responses change in value by $\hat{\beta}_1$ @ predictor increases by 1 unit

What makes a relationship linear?

"Linear" refers to the coeffients/parameters, not predictor

relationship between Y & X may not appear linear, but maybe Y & X^2 is?

$\hookrightarrow y_i = \beta_0 + \beta_1 \sin x_i + \varepsilon_i$, $y_i = \beta_0 + \beta_1 \mathbb{I}(\text{answer} = \text{yes}) + \varepsilon_i$

Non-linear: $y_i = \log(\beta_0 + \beta_1 x_i + \varepsilon_i)$, $y_i = \beta_0 e^{\beta_1 x_i} + \varepsilon_i$

Any relationship that is a linear combination of the coeffients is a linear relationship

Matrix Form

$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ for $i = 1, \cdots, n$

$$Y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}, \quad X = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}, \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

$n \times 1$ $\quad 2 \times 1$ $\quad n \times 2$ $\quad n \times 1$

Each row of $Y = X\beta + \varepsilon$ is equal to algebraic form

$$\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

MULTIPLE LINEAR REGRESSION (MLR) MODELS

Involves more than one predictor $y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i$

Functional relationship $\mathbb{E}[Y \mid X_1, \ldots, X_p] = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$

The mean response is dependent on the value of p predictors

Sample data: n sets of $(y_i, x_{i1}, \ldots, x_{ip})$

Need to estimate p-dimensional surface of best fit by estimating $\beta_0, \beta_1, \ldots, \beta_p$

MATRIX FORM

Algebraic form: $y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i$ for $i = 1, \ldots, n$

$$Y = \underset{n \times 1}{\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}}, \quad \beta = \underset{(p+1) \times 1}{\begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}}, \quad X = \underset{n \times (p+1)}{\begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix}}, \quad \varepsilon = \underset{n+1}{\begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}}$$

The matrix expression of the multiple linear regression trend is simply $Y = X\beta + \varepsilon$

PREDICTOR MATRIX

Each column represents info of one predictor, first column always 1's for intercept

Suppose one predictor (X_1) is discrete, but one is qualitative (Yes, No, Maybe)

Use indicator variables to indicate value of each entry

$$X_2 = \begin{cases} 1 & \text{if Yes} \\ 0 & \text{if otherwise} \end{cases} \qquad X_3 = \begin{cases} 1 & \text{if No} \\ 0 & \text{if otherwise} \end{cases}$$

↳

$$X = \begin{pmatrix} 1 & 2 & 1 & 0 \\ 1 & 4 & 0 & 1 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 3 & 0 & 1 \end{pmatrix} \begin{matrix} \text{Yes} \\ \text{Maybe} \\ \\ \text{No} \end{matrix}$$

CONDITIONAL MEAN & DISTRIBUTION OF RESPONSES

SLR conditions on the value of one predictor

At each X value, we have a distribution of Y responses w/ mean $\mathbb{E}[Y \mid X]$

MLR conditions on the values of p predictors

↳ at values (x_1, x_2) of predictors (X_1, X_2), we have a distribution of responses Y with a mean $\mathbb{E}[Y \mid x_1, x_2]$

ESTIMATION VIA LEAST SQUARES

want a "snug" surface through data, minimize errors

Same Residual Sum of Squares as SLR, but more predictors & coeffients

$$RSS = \hat{e}^T \hat{e} = \sum_{i=1}^{n} \left(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_p x_{ip}) \right)^2$$

Finds estimates $\hat{\beta}$ that make all residuals $\hat{e} = Y - X\hat{\beta} = Y - \hat{Y}$ as small as possible

LEAST SQUARES PROCEDURE

1. Set up estimating equation for given model w/ parameters
2. Take partial derivatives of estimating equation w/ respect to each unknown parameter
3. Set each derivative to 0
4. Rearrange, solve for unknowns

$$RSS = \hat{e}^T \hat{e} = (Y - X\hat{\beta})^T (Y - X\hat{\beta})$$
$$= Y^T Y - 2Y^T X\hat{\beta} + \hat{\beta}^T X^T X \hat{\beta}$$
$$\frac{\partial RSS}{\partial \hat{\beta}} = -2X^T Y + 2(X^T X)\hat{\beta} = 0$$
$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

assuming inverse exists

WORKING W/ MATRIX ESTIMATORS

Data stored in X & Y matrices, so we use component matrices

$$X^TX = \begin{pmatrix} n & \sum x_{i1} & \sum x_{i2} & \cdots & \sum x_{ip} \\ \sum x_{i1} & \sum x_{i1}^2 & \sum x_{i1}x_{i2} & \cdots & \sum x_{i1}x_{ip} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sum x_{ip} & \sum x_{i1}x_{ip} & \sum x_{i2}x_{ip} & \cdots & \sum x_{ip}^2 \end{pmatrix}, \quad X^TY = \begin{pmatrix} \sum y_i \\ \sum x_{i1}y_i \\ \vdots \\ \sum x_{ip}y_i \end{pmatrix}$$

Fitted values $\hat{Y} = X\hat{\beta} = X(X^TX)^{-1}X^TY$ are found by multiplying responses Y by hat matrix $X(X^TX)^{-1}X^T$

hat matrix is a projection matrix, has same properties of symmetrix, idempotent

APPLICATION

$n = 12$

$\sum_{i=1}^{12} y_i = 90$ $\quad \sum_{i=1}^{12} x_{i1} = 52$ $\quad \sum_{i=1}^{12} x_{i2} = 102$ $\quad \sum x_{i1}x_{i2} = 536$

$\sum_{i=1}^{12} x_{i1}^2 = 296$ $\quad \sum_{i=1}^{12} x_{i2}^2 = 1004$ $\quad \sum_{i=1}^{12} y_i x_{i1} = 482$ $\quad \sum y_i x_{i2} = 872$

$$\hat{\beta} = (X^TX)^{-1}X^TY$$

$$= \text{(given)}\ X^TY = \begin{pmatrix} 12 & 52 & 102 \\ 52 & 296 & 536 \\ 102 & 536 & 1004 \end{pmatrix}^{-1} \begin{pmatrix} 90 \\ 482 \\ 872 \end{pmatrix} = \begin{pmatrix} 5.375 \\ 3.012 \\ -1.285 \end{pmatrix}$$

$$\hat{y}_i = 5.375 + 3.012x_{i1} - 1.285x_{i2}$$

INTERPRETATIONS

Let's say we have 2 predictors X_1 & X_2, we can fit 3 models:

Simple w/ X_1 only: $\hat{y}_i = 1.86 + 1.30x_{i1}$

Simple w/ X_2 only: $\hat{y}_i = 0.86 + 0.78x_{i2}$

2-predictor: $\hat{y}_i = 5.375 + 3.012x_{i1} - 1.285x_{i2}$

Why is the relationship between Y & X_2 positive in one model but negative in the other?

Estimation & interpretation of coefficients in multi-predictor models conditions on all other models

Only one fixed value of all other predictors @ estimating/interpreting coefficient of interest.

In simple model, coeffients interpreted as: $\hat{\beta}_0$ mean response @ predictor = 0

$\hat{\beta}_1$ change in mean response for 1 unit ↑ of value of predictor

Each coefficient is interpreted individually, change observed in response is attributed ONLY to that predictor

$\hat{\beta}_0$ is mean response @ ALL predictor values are 0

$\hat{\beta}_j$ is average/mean/expected change in response for a 1-unit increase in X_j, @ all others are fixed

Interactions allow relationship between response & one predictor to vary according to values of second one

↳ $\hat{\beta}_4$ SepalWidth * $\mathbb{I}$(Versicolor) — $\hat{\beta}_4 = 0.36055$

change in mean response for a 1 unit ↑ in Sepal Width for Versicolor compared to Setosa w/ fixed sepal length

ASSUMPTIONS OF LINEAR MODELS

@ fitting a model, we make the assumptions EVERY time

LINEARITY of relationship (Mean Zero Errors) assumptions

$\mathbb{E}[\varepsilon|X]=0$, $\mathbb{E}[Y|X]=X\beta$, $Y=X\beta+\varepsilon$ (true relationship in population)

Implies 2 things about population relationship
- true relationship is linear in coefficients
- true relationship is exactly $Y=X\beta+\varepsilon$ w/
  - no predictors omitted from X that should be present
  - no predictors included in X that should not be present
  - no predictors in X that are in the wrong functional form

violations
↳ omitting a predictor known to influence response, fitting linear model @ truth is log, including x @ it should be x^2

Ensures estimate coefficients unbiasedly

UNCORRELATED ERRORS (Independence)

$Cov(\varepsilon_i, \varepsilon_j)=0$, $Cov(y_i, y_j)=0$

Each data point in population must not be related / connected to any other data point
knowing info about one does not give any info. about another
↳ stock price data, weather data, measurements on same person

Ensure correct precision of estimates

CONSTANT ERROR VARIANCE (Homoskedasticity)

$Var(\varepsilon|X)=\sigma^2 I$, $Var(\varepsilon_i|X)=Var(y_i|X)=\sigma^2$

Each conditional distribution must have the same spread
only difference between each one is that the mean changed by a specific amount

Ensures we obtain reasonable estimates of variability for all conditional means

NORMAL ERRORS

$\varepsilon|X \sim N_n(0, \sigma^2 I)$, $Y|X \sim N_n(X\beta, \sigma^2 I)$, $\varepsilon_i \sim N(0, \sigma^2)$, normally distributed responses/errors

Each conditional prob. distribution must have the same shape
harder to verify in small samples
not needed to estimate coefficients by least squares

Allows us to utilize properties of Normal random variables for inferential purposes
↳ computing confidence intervals

Nothing stops us from fitting an invalid model w/ violated assumptions

VERIFYING ASSUMPTIONS USING RESIDUAL PLOTS

Fit model to data, extract fitted / predicted values ($\hat{y}_i$), extract residuals

Residuals vs. fitted for linearity, uncorrelated errors, constant variance

Residuals vs. predictor for linearity, uncorrelated errors, constant variance

- linearity violation: any systematic pattern, especially curves / other functions of predictors
- uncorrelated errors: large clusters of many points, pattern across times, some other sequencing information
- constant variance: any systematic pattern, especially fanning w/ increasing / decreasing spread

Normal Q-Q plot shows distribution of data against expected normal distribution

- normality: stark deviations from the diagonal line (minimal deviations is ok)

EXPLORE & UNDERSTAND DATA

Assumptions formally checked using residual plots, knowing data can help

Always conduct exploratory data analysis before fitting model

- skew in RV → issue w/ normality, linearity
- skews in predictor vars. → issue w/ linearity

Use existing research / literature, think about how data was collected / sample

ADDITIONAL CONDITIONS FOR MLR

Recall: interpretation of coefficients involved holding other predictors fixed

MLR estimates relationship using predictors jointly

Patterns in plots cannot be used to identify a specific violation, can give misleading conclusions

CONDITIONAL MEAN RESPONSE CONDITION

mean responses are a single function of a linear combination involving β

$\mathbb{E}[Y_i \mid X = x_i] = g(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip})$

↳ $\mathbb{E}[Y \mid X] = \log(\beta_0 + \beta_1 x_i)$ satisfies, $\mathbb{E}[Y \mid X] = \frac{\beta_1 x_{i1}}{\beta_2 x_{i2}} = \frac{g_1(x_1)}{g_2(x_i)}$ violates

look for random diagonal scatter, easily identifiable non-linear trend in response vs. fitted

CONDITIONAL MEAN PREDICTOR CONDITION

the mean of each predictor is related to each other predictor in no more complicated way than linearly

$$\mathbb{E}[X_i \mid X_j] = \alpha_0 + \alpha_1 X_j$$

Linear/no relationship satisfy condition, anything else violates

look for lack of curves/other non-linear patterns in all pairwise scatterplots of predictors

MITIGATING VIOLATED ASSUMPTIONS

Way to correct/improve violated assumption is via a transformation on relevant variables

↳ use $X^* = f(X)$ instead of X

Transformations can be made into any numerical variable

↳ natural logarithm (ln), square root improves most right skews

squares, cube roots, other logs can improve left skews

VARIANCE STABLIZING TRANSFORMATION

Variance Stabilizing Transformations specifically target violation of constant variance

applied only to the response

Choice depends on exact situation & data

identifiable data type (e.g. count data) / specific variable context

experience from similar contexts / data types

Works by exploiting connection between mean & variance of response

non-constant var. implies both $\mathbb{E}[Y|X]$ & $\text{Var}[Y|X]$ change w/ same X

Consider first-order Taylor series expansion of $f(Y)$ around mean

$$f(Y) = f(\mathbb{E}[Y]) + f'(\mathbb{E}[Y])[Y - \mathbb{E}[Y]) + \cdots$$

Take var. of both sides $\text{Var}[f(Y)] = V[f(\mathbb{E}[Y])] + \text{Var}[f'(\mathbb{E}[Y])(Y - \mathbb{E}[Y])]$

$f(\mathbb{E}[Y])$ is fixed, var( ) = 0

$f'(\mathbb{E}[Y])$ taken out, squared

$\text{Var}[Y - \mathbb{E}[Y]] = \text{Var}[Y]$, $\mathbb{E}[Y]$ is constant?

Therefore we get $\text{Var}[f[Y]] = f'(\mathbb{E}[Y])^2 \text{Var}[Y]$

undoes the non-constant variance of Y

↳ $Y \sim \text{Poi}(\lambda)$, $\mathbb{E}[Y|X] = \text{Var}[Y|X] = \lambda$ normally violates normality

Try to apply sqr root transformation $f(Y) = \sqrt{Y} \Rightarrow \frac{d}{dy}\sqrt{Y} = \frac{1}{2}Y^{-1/2}$

$$\text{Var}(Y^{\frac{1}{2}}) = \left(\frac{1}{2}\lambda^{-\frac{1}{2}}\right)^2 \lambda = \frac{1}{4}\lambda^{-1}\lambda = \frac{1}{4}$$

Now var. of $f(Y)$ is constant, using $f(Y) = \sqrt{Y}$ should satisfy nomality

BOX-COX TRANSFORMATIONS FOR NORMALITY

Uses maximum likelihood to estimate the power transformation

$$\log(L(\beta_0, \beta_1, \sigma^2|Y)) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\log\left(\underbrace{\sum(y_i - \beta_0 - \beta_1 x_i)^2}_{\text{RSS}}\right)$$

Find a transformation on Y by modifying RSS

$$RSS = \sum(\psi_M(Y, \lambda) - \beta_0 - \beta_1 x_i)^2$$

$\psi_M(Y, \lambda)$ is modified power transformation applied to Y

$\psi_M(Y, \lambda)$ looks disgusting, so once we get a max. likelihood estimate for λ, take Y^λ as our transformation

when $\lambda = 0$, use $\ln(Y)$ as our transformation

Box-Cox in Practice

Using simplier powers as transformations is much easier to interpret & apply

Pick something simpler than estimated λ:

↳ $\lambda = 0.103$ use ln(Y) since close to 0

$\lambda = 0.5$ use square root

$\lambda = 0.33$ use cube root

$\lambda = 0.25$ fourth root

$\lambda = -0.5$ reciprocal square root

$\lambda = -1$ reciprocal (inverse)

Function will report confidence intervals, can use range of possibilities

Application & Interpretation of Transformed Models

To interpret models that use transformations, always remember to incorporate variables as they are now

↳ "for one unit increase in x^2"

Once transformed, predicted values only represent conditional means on transformed scale, not original

↳ The expected log of Revenue due to Advertising @ log Ad Pages, log Subscription revenue, & log Newsstand Revenue are 0

For 1 unit increase in log of Advertising Pages, the slope is expected change in log of Ad Revenue for a fixed log Subscription Revenue & log Newsstand Revenue

Impact of Violations on Sampling Distributions

Any time we define a sampling distribution, we utilize assumptions about population

Assumptions give us properties of this distribution, mean, shape

@ assumptions don't hold properties no longer true

Sampling Distribution of Estimated Coefficents

Assumptions say $Y|X \sim N(X\beta, \sigma^2 I)$

Recall estimates found $\hat{\beta} = (X^TX)^{-1}X^TY$, $\hat{\beta}$ are function of Y

Assume assuptions hold in our population, for our model, we can find sampling distribution of $\hat{\beta}$

property of linearity of Normal random variables

$AY \sim N(A\mu_Y, A\Sigma A)$ $\quad \Sigma a_i Y_i \sim N(\Sigma a_i \mu_i, \Sigma a_i^2 \sigma_i^2)$ where A is a matrix of constants

Using linearity of Normals we get

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

@ our assumptions hold: our estimates are: unbiased

generally correlated

use same constant variance as errors

What Happens @ Shit is Violated?

Linearity mean is no longer β, estimates biased

Constant variance no longer have a single σ^2 as part of variance

Uncorrelated errors variance in estimators will be under- or over-estimated b/c we are borrowing individuals/measurements

Normality no Normal sampling distribution, large samples approx. Normal

INFERENCE ON LINEAR REGRESSION COMPONENTS

SAMPLING DISTRIBUTION OF COEFFICENTS

Any time we define a sampling distribution, we utilize assumptions about population to determine its properties

Sampling distribution is our reference for what is considered normal variation to expect

pivotal quality $\frac{\text{estimator - truth}}{\text{standard error}}$ is basis of CI & hypothesis test

Compared to sampling distribution to determine if whether estimated value is reasonable

PROPERTIES OF SAMPLING DISTRIBUTIONS

Assumptions say $Y|X \sim N(X\beta, \sigma^2 I)$, our estimates are $\hat{\beta} = (X^TX)^{-1}X^TY$

Using linearity of Normal's, our sampling distribution is $\hat{\beta} \sim N(\beta, \sigma^2(X^TX)^{-1})$

$$\begin{aligned}\mathbb{E}[\hat{\beta}|X] &= \mathbb{E}[(X^TX)^{-1}XY|X] \\ &= (X^TX)^{-1}X^T\mathbb{E}[X\beta+\varepsilon|X] \\ &= (X^TX)^{-1}X^T\{X\beta+\mathbb{E}[\varepsilon|X]\} \\ &= (X^TX)^{-1}X^TX\beta = \beta\end{aligned}$$

The LS estimators of β are unbiased

$$\begin{aligned}\text{Cov}(\hat{\beta}|X) &= \text{Cov}((X^TX)^{-1}X^TY|X) \\ &= (X^TX)^{-1}X^T\,\text{Cov}(X\beta-\varepsilon|X)\,X(X^TX)^{-1} \\ &= (X^TX)^{-1}X^T\,\text{Cov}(\varepsilon|X)\,X(X^TX)^{-1} \\ &= (X^TX)^{-1}X^T\sigma^2 I\,X(X^TX)^{-1} = \sigma^2(X^TX)^{-1}\end{aligned}$$

COVARIANCE MATRICES

Combine information about variance of each variable, how any two random variables vary together

Covariance measures joint variability of two random variables

$$\text{Cov}(\varepsilon|X) = \begin{pmatrix} \text{Var}(\varepsilon_1|X) & \text{Cov}(\varepsilon_1,\varepsilon_2|X) & \cdots & \text{Cov}(\varepsilon_1,\varepsilon_n|X) \\ \text{Cov}(\varepsilon_1,\varepsilon_2|X) & \text{Var}(\varepsilon_2|X) & \cdots & \text{Cov}(\varepsilon_2,\varepsilon_n|X) \\ \vdots & \vdots & \ddots & \vdots \\ \text{Cov}(\varepsilon_1,\varepsilon_n|X) & \text{Cov}(\varepsilon_2,\varepsilon_n|X) & \cdots & \text{Var}(\varepsilon_n|X) \end{pmatrix}$$

$$\text{Cov}(\hat{\beta}|X) = \begin{pmatrix} \text{Var}(\hat{\beta}_0|X) & \cdots & \text{Cov}(\hat{\beta}_0,\hat{\beta}_p|X) \\ \text{Cov}(\hat{\beta}_0,\hat{\beta}_1|X) & \cdots & \text{Cov}(\hat{\beta}_1,\hat{\beta}_p|X) \\ \vdots & \ddots & \vdots \\ \text{Cov}(\hat{\beta}_0,\hat{\beta}_p|X) & \cdots & \text{Var}(\hat{\beta}_p|X) \end{pmatrix}$$

MLR elements harder to express than SLR

$$\text{Cov}(\hat{\beta}|X) = \sigma^2(X^TX)^{-1} = \frac{\sigma^2}{\sum(x_i-\bar{x})^2}\begin{pmatrix} \frac{1}{n}\sum x_i^2 & -\bar{x} \\ -\bar{x} & 1 \end{pmatrix}$$

Non-zero covariance ⇓ everything is conditional in regression

UNKNOWN ERROR VARIANCE

$\hat{\beta} \sim N(\beta, \sigma^2(X^TX)^{-1})$ contains unknown parameter σ^2

Need to compute margins of error / standardized test statistics

Estimate σ^2 with $S^2 = \frac{RRS}{n-p-1} = \frac{\hat{e}^T\hat{e}}{n-p-1}$

↳ SLR $\hat{e}^T\hat{e} = \sum(y_i-\hat{\beta}_0-\hat{\beta}_1x_i)^2$

In practice, use $s^2(X^TX)^{-1}$ as variance matrix

Related distribution used to adjust $\frac{\hat{\beta}-\beta}{\sqrt{s^2(X^TX)^{-1}}} \sim T_{n-p-1}$

CONFIDENCE INTERVALS

General form CI : estimate ± (critical value)(standard error)

$$\frac{\hat{\beta}-\beta}{\sqrt{s^2(X^TX)^{-1}}} \sim T_{n-p-1}$$

Chance that your sample was one of the $(1-\alpha)$% CIs that overlapped the truth

INFERENCE ON INDIVIDUAL COEFFICENTS FOR β_j

$(1-\alpha)$% CI for β_j : $\hat{\beta}_j \pm t_{\frac{\alpha}{2}, n-p-1}\, s\sqrt{(X^TX)^{-1}_{(j+1,j+1)}}$

α is our chosen significance level, $1-\alpha$ is our confidence level

The critical value corresponds to the $\alpha/2$ quantile of T distribution w/ $n-p-1$ degrees of freedom

$(X^TX)^{-1}_{(j+1,j+1)}$ refers to $j+1$ element of main diagonal

CONCLUDING INFERENCES

$(1-\alpha)$% of all intervals computed using data repeatedly obtained from the same population would contain the true β_j

Our interval represents plausible values of β_j with $(1-\alpha)$% confidence

We see the true β_j in the window $(1-\alpha)$% of the time

T-TEST

General form $\frac{\text{estimate - truth}}{\text{standard error}}$ $\frac{\hat{\beta}-\beta}{\sqrt{s^2(X^TX)^{-1}}} \sim T_{n-p-1}$

Chance of seeing a value as extreme assuming the hypothesized truth holds

INFERENCE ON INDIVIDUAL COEFFICENTS

Hypothesis test $H_0: \beta_j = \beta_j^0$ $t^* = \frac{\hat{\beta}_j - \beta_j^0}{s\sqrt{X^TX^{-1}_{(j+1,j+1)}}}$

β_j^0 is the value we hypothesize is the truth (usually 0)

Standard error $s\sqrt{(X^TX)^{-1}_{(j+1,j+1)}}$ used to standardize difference between estimate & hypothesized value

Use same statistics regardless of which $\hat{\beta}_j$ or which β_j^0 we test, or even which alternative

CONCLUDING INFERENCES

Testing whether $H_0: \beta_j = 0$, $H_A: \beta_j \neq 0$ is most common

Tests the null that there is no linear relationship exists between X_j & Y (in presence of other predictors)

Conclude test by comparing to sampling distribution

if $|t^*| > t_{\frac{\alpha}{2}, n-p-1} \Rightarrow$ reject null

if $P(|T_{n-p-1}| \geq |t^*|) < \alpha \Rightarrow$ reject null

} Claim a significant linear relationship exists

EXAMPLE

Estimated SLR model relating # of rooms cleaned (Y) to size of cleaning crew (X) is $\hat{y}_i = 1.785 + 3.701x_i$

We have $n = 53$, $\bar{x} = 8.679$, $s_x^2 = \frac{\sum(x_i-\bar{x})^2}{53-1} = 23.068$, $RSS = 2744.796$ $(X^TX)^{-1} = \begin{pmatrix} 0.08166037 & -0.007235435 \\ -0.007235435 & 0.0008336479 \end{pmatrix}$

1. Find variances $s^2 = \frac{RSS}{n-p-1} = \frac{2744.796}{53-1-1} = 53.82$

$Var(\hat{\beta}_0) = s^2(X^TX)^{-1}_{(1,1)} = 53.82(0.8166) \simeq 4.395$ $Var(\hat{\beta}_1) = s^2(X^TX)^{-1} = 53.82(0.0008) \simeq 0.0449$

$= s^2\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum(x_i-\bar{x})^2}\right)$ $= \frac{s^2}{\sum(x_i-\bar{x})^2}$

2. Critical value/distribution $T_{n-p-1} = T_{51} \Rightarrow \left|t_{\frac{0.05}{2}, 51}\right| = 2.00$

3. Compute interval / test statistic

95% CI : $\hat{\beta}_0 \pm t_{0.025,51}\sqrt{s^2(X^TX)^{-1}_{(1,1)}} = 1.785 \pm 2.00\sqrt{4.395} = [-2.41, 5.98]$

Of the 95% CI of all intervals computed using data repeatedly obtained from the same population that contains the true β_0

$H_0: \beta_1 = 0$, $H_A: \beta_1 \neq 0$ $t^* = \frac{\hat{\beta}_1 - 0}{\sqrt{s^2(X^TX)^{-1}_{(2,2)}}} = \frac{3.701}{\sqrt{0.0449}} = 17.466$

Since $17.466 > 2.00$, we reject H_0, conclude a statistically significant linear relationship exists

CONFIDENCE INTERVAL FOR MEAN RESPONSES

Can make an inference on $\mathbb{E}[Y|X]$, estimated by $\hat{Y} = X\hat{\beta}$

Treat each mean response individually

for each $x_0^T = (1, x_1, x_2, \dots, x_p)$, estimate $\hat{y}_0 = \hat{\mathbb{E}}[Y \mid X = x_0^T] = x_0^T\hat{\beta}$, a single estimated mean

$\hat{y}_0$ is a linear combination of Y: $\hat{y}_0 = x_0^T\hat{\beta} = x_0^T(X^TX)^{-1}X^TY$, linearity of Normals gets $\hat{y}_0 | X, x_0 \sim N(x_0^T\beta, \sigma^2 x_0^T(X^TX)^{-1}x_0)$

If assumptions hold, then $\hat{y}_0$ is unbiased $\mathbb{E}[\hat{y}_0|X, x_0] = \mathbb{E}[x_0^T\hat{\beta}|X, x_0] = x_0^T\mathbb{E}[\hat{\beta}|X, x_0] = x_0^T\beta$

then covariance matrix is $\text{Cov}(\hat{y}_0 | X, x_0) = \sigma^2 x_0^T(X^TX)^{-1}x_0$

σ^2 is estimated by s^2, giving $$\frac{\hat{y}_0 - x_0^T\beta}{\sqrt{s^2 x_0^T(X^TX)^{-1}x_0}} \sim T_{n-p-1}$$

INFERENCE ON MEAN RESPONSE $x_0^T\beta$

pretty much the same thing

$(1-\alpha)$% CI for $y_0 = x^T\beta$

$$x_0^T\hat{\beta} \pm t_{\frac{\alpha}{2}, n-p-1}\, s\sqrt{x_0^T(X^TX)^{-1}x_0}$$

If working w/ SLR, $\sqrt{\text{Var}(\hat{y}_0 | X, x_0)} = \sqrt{s^2\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum(x_i - \bar{x})^2}\right)}$

Hypothesis test for $H_0: y_0 = y_0^0$

$$t^* = \frac{\hat{y}_0 - y_0^0}{s\sqrt{x_0^T(X^TX)^{-1}x_0}}$$

Conclude based on sampling distribution T_{n-p-1}

No default value or special interpretation

EXAMPLE

Same model from before, estimate mean response for 5 crews

$$x_0^T = (1 \quad 5), \quad \hat{y}_0 = (1 \quad 5)\begin{pmatrix} 1.785 \\ 3.701 \end{pmatrix} = 20.29$$

$$\text{Var}(\hat{y}_0) = s^2 x_0^T(X^TX)^{-1}x_0 = 53.82\,(1 \quad 5)\,(X^TX)^{-1}\begin{pmatrix} 1 \\ 5 \end{pmatrix} = 1.623$$

$$\text{(SLR)} = s^2\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum(x_i - \bar{x})^2}\right) = 53.82\left(\frac{1}{53} + \frac{(5 - 8.679)^2}{(52)(23.068)}\right) = 1.623$$

critical value $|t_{0.025, 51}| = 2.00$

95% CI $x_0^T\hat{\beta} \pm t_{\frac{\alpha}{2}, n-p-1}\, s\sqrt{x_0^T(X^TX)^{-1}x_0} = 20.29 \pm 2.00\sqrt{1.623} = [17.74, 22.84]$

PREDICTION INTERVAL FOR ACTUAL RESPONSES

DIFFERENCE BETWEEN ACTUAL y & MEAN RESPONSE

Regression model gives predicted values ONLY for $\mathbb{E}[Y|X = x_0]$ —— parameter, fixed but unknown

Want to make a prediction about an individual person & respective y_0

Actual response is a realization of a random variable Y in population

Need to account for difference between actual value (want) & predict (regression value)

using prediction error $y_0 - \hat{y}_0 = x_0^T\beta + \varepsilon_0 - \hat{y}_0 = (x_0^T\beta - \hat{y}_0) + \varepsilon_0$

PREDICTION INTERVAL FOR ACTUAL RESPONSE

Since we can't predict an actual value, we instead create a prediction interval

gives range of possible values, not same as CI since we don't estimate a parameter

wider than CI for $\mathbb{E}[Y|X, x_0]$, due to extra σ^2

Interval centered at estimated mean response $\hat{y}_0$

$$(1-\alpha)\%\text{ PI}: \quad x_0^T\hat{\beta} \pm t_{\frac{\alpha}{2}, n-p-1}\, s\sqrt{1 + x_0^T(X^TX)^{-1}x_0}$$

SLR: $\hat{\beta}_0 + \hat{\beta}_1 x_0 \pm t_{\frac{\alpha}{2}, n-2}\, s\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum(x_i - \bar{x})^2}}$

EXAMPLE add a +1 in the $\text{Var}(\hat{y}_0)$, compute 95% PI based off that

DECOMPOSING VARIANCE

Fitting linear model minimizes RSS (variation around line)

Hypothesis test on slope in SLR also uses variation to decide on presence of significant linear relationship.

If $\hat{\beta}_1 - 0$ is much bigger than variation expected in estimating $\hat{\beta}_1$, then 0 is not a plausible value

If no linear relationship, then $\beta_1 = 0$, $\beta_0 = \bar{y}$ $\quad y_i = \beta_0 + \varepsilon$ is a horizontal line

Test statistic compares $\hat{\beta}_1 - 0$ to variation in estimating $\hat{\beta}_1$ $\quad$ includes $s^2 = \Sigma(y_i - \hat{y}_i)^2 / n-2$

For t-test to indicate significant linear $\hat{y}_i - \bar{y}$ across all observations, would need to be larger than residuals $y_i - \hat{y}_i$ overall

Aim to explain variation in response quantified by $SST = \Sigma(y_i - \bar{y})^2$

Each predictor added to model for y hopefully related to y, explains some of variation

There will always be some variation left unexplained

DECOMPOSITION OF SUMS OF SQUARES

3 sources of variation				df
	Total Sum of Squares	$SST = \Sigma(y_i - \bar{y})^2$	total amount prior to fitting model	$n-1$
	Regression Sum of Squares	$SS_{reg} = \Sigma(\hat{y}_i - \bar{y})^2$	variation explained by predictors/model	p
	Residual Sum of Squares	$RSS = \Sigma(y_i - \hat{y}_i)^2$	unexplained variation	$n-p-1$

$$SST = SS_{reg} + RSS$$

$$\Sigma(y_i - \bar{y})^2 = \Sigma(\hat{y}_i - \bar{y})^2 + \Sigma(y_i - \hat{y}_i)^2$$

ANOVA TEST FOR OVERALL SIGNIFICANCE

T-test on slopes of MLR can only say if predictor is linearly related to response in presence of others.

Cannot give overall notion whether linear relationship exists in model ⟶ use ANOVA

Significant linear relationship ∝ significant amount of variation in response explained by model

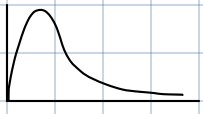
A good model would have $SS_{reg} > RSS$

Model where SS_{reg}/SST is large has RRS/SST small

Analysis of Variance Test of overall Significance compares RSS & SS_{reg},

$H_0: \beta_1 = 0$ all slopes are 0, $H_A: \beta_1 \neq 0$ at least one slope is not 0

Assuming H_0 is true, use test statistic $F^* = \dfrac{SS_{reg}/p}{RSS/n-p-1} \sim F(p, n-p-1)$ $\quad p = \#$ predictors



dividing each by its df standardizes quantity, gives us mean squares regression & mean squares residual

More extreme F^* values will be in right tail (larger values)

If p-value $< \alpha$ or $F^* > F_{(1-\alpha),(p,n-p-1)}$, then reject H_0, conclude a statistically significant linear relationship exists for at least one predictor.

EXAMPLE

3 predictor model (X_1, X_2, X_3) fit to response Y using sample size $n = 30$, $p = 3$.

Response sample variance $s_y^2 = 376.6853$, model estimated error variance $S^2 = 50.555$.

$H_0: \beta_1 = \beta_2 = \beta_3 = 0$ $\quad H_A:$ at least one $\beta_j \neq 0$

1. Degrees of freedom $\quad df_{RSS} = n-p-1 = 26$, $df_{SS_{reg}} = p = 3$
2. Find sum of squares $\quad SST = (n-1)s_y^2 = 29(376.6853) = 10923.87$, $RSS = (n-p-1)S^2 = 26(50.555) = 1314.43$

   $SS_{reg} = SST - RSS = 9609.34$
3. Find mean squares $\quad MSR = RSS/df_{RSS} = \dfrac{1314.43}{26} = 50.555$, $MS_{reg} = \dfrac{SS_{reg}}{3} = \dfrac{9609.34}{3} = 3203.147$
4. Test statistic $\quad F^* = MS_{reg}/MSR = \dfrac{3203.147}{50.555} = 63.36$
5. Conclude $\quad 63.36 > 2.98$, reject H_0, conclude significant linear relationship exists for at least one predictor.

PARTIAL F TEST

Partial F test can account for conditionality to say if we can remove some predictors

If simple model as good as 3-predictor model, SS_{reg} & RSS should be similar

If RRS is too much bigger, can't remove everything at once

Compares 2 models — Complete / Full w/ p predictors — Reduced w/ p-k predictors

same data, same SST

Model w/ more predictors has smaller RSS, adjust for additional predictors by standardizing w/ df

$RSS_{reduced} > RSS_{full} \iff SS_{reg,reduced} < SS_{reg,full}$

Assume $H_0: \beta_2 = 0$ is true, $H_A: \beta_2 \neq 0$, where $\beta = (\beta_0, \beta_1, \beta_2)^T$

β_2 is vector of k coefficients that were removed to make reduced model

Quantify difference between RSS $RSS_{drop} = RSS_{reduced} - RSS_{full}$

Ratio of means of sums of squares $$F^* = \frac{RSS_{drop}/k}{RSS_{full}/(n-p-1)} \sim F(k, n-p-1)$$

If reject H_0, additional SS_{reg} from k predictors explain a lot of variation, want to keep. Otherwise don't keep.

Conclude there exists a significant linear relationship between Y & at least one of k predictors

EXAMPLE

Same set up as before. Model w/ only X_1 yields RSS = 1497.03. $H_0: \beta_2 = \beta_3 = 0$, H_A: at least one $\beta_j \neq 0$

1. Define model & values $RSS_{full} = (n-p-1)s^2 = 1314.43$, df = 26

   $RSS_{reduced} = 1497.03$

2. $RSS_{drop} = RSS_{reduced} - RSS_{full} = 182.6$

3. Test statistic $$F^* = \frac{RSS_{drop}/k}{RSS_{full}(n-p-1)} = 1.81$$

4. Conclude $1.81 < 3.37$. Fail to reject H_0, no significant relationship between Y & either X_2 or X_3, can be removed.

Cannot compare models from different datasets

Cannot compare models that are not subsets of one another

General order: ANOVA → T-test (to determine what to drop) → partial ANOVA

COEFFICIENT OF DETERMINATION R^2

Measure of goodness $$R^2 = \frac{SS_{reg}}{SST} = 1 - \frac{RSS}{SST} \longrightarrow R^2_{adj} = 1 - \frac{RSS/(n-p-1)}{SST/(n-1)}$$

Adjust b/c $SS_{reg,big} > SS_{reg,small}$, bigger model has larger R^2 by default

A bigger model is only better if SS_{reg} has increased enough to compensate for added complexity.

EXAMPLE

Same model as before. Compute R^2 and R^2_{adj}.

1. Collect given values $n = 30$, $p = 3$, $s_y^2 = SST/(n-1) = 376.6853$, $s^2 = RSS/(n-p-1) = 50.555$

2. Find decomposition pieces $SST = (n-1)s_y^2 = 10923.87$, $RSS = (n-p-1)s^2 = 1314.43$

3. Compute coefficients $$R^2 = 1 - \frac{RSS}{SST} \approx 0.88$$ 88% of variation in Y explained by model

   $$R^2_{adj} = 1 - \frac{RSS/(n-p-1)}{SST/(n-1)} \approx 0.866 < R^2$$

PROBLEMS WITH RELATED PREDICTORS

Conditional mean predictor condition mean of each predictor is related to each other in no more complicated way than linearly.

$$\mathbb{E}[X_i \mid X_j] = \alpha_0 + \alpha_1 X_j$$

Linear/no relationship is fine, anything else violates

Ideally no relationship. Linear means predictors are correlated / colinear

Correlation measures strength of linear association between 2 continuous values

Perfect correlation is -1 or +1

See if colums of matrix X are functions of one another. If yes, matrix is not full column rank, cannot find inverse $(X^TX)^{-1}$

↳ $X = \begin{pmatrix} 1 & 4 & 3 \\ 2 & 2 & 3 \\ 3 & 8 & 7 \end{pmatrix}$ Column 3 = column 1 + 0.5 (column 2)

MULTICOLINEARITY

Multicolinearity more than 2 predictors are related

correlation only gives us for 2 predictors

Possible problems include wrong estimated coefficients, contradictory significance, inflated variances

Sometimes we want multicolinearity

Plots show how we arrive at fitting a polynomial regression model

Using interaction term & main effect term in a model

To check for multicolinearity, compute variance inflation factor (VIF)

quantifies how much larger variance of a coefficient is due to multicolinearity.

↳ 3 predictors, $X_1 = \beta_0 + \beta_1 x_2 + \beta_2 x_3 + \varepsilon$.

R^2 measures strength of linear relationship between X_1 & both X_2, X_3

$$\text{Var}(\hat{\beta}_j) = \underbrace{\frac{1}{1 - R_j^2}}_{\text{VIF for } \beta_j} \times \frac{\sigma^2}{(n-1) S_{x_j}^2}, \quad j = 1, \ldots, p$$

The stronger the proportion of variation in X_j explained by remaining $p-1$ predictors, the stronger the inflation.

Generally use VIF > 5

PROBLEMATIC OBSERVATIONS

LEVERAGE

Leverage Observation observation that is very distant from center of X space that might change $\hat{Y}$

Potential to change how trend is estimated

Hat matrix H $\hat{y} = X\hat{\beta} = X(X^TX)^{-1}X^TY = HY$ projection matrix, project Y onto model space through X

$$\hat{y} = HY = \begin{pmatrix} h_{11} & h_{12} & \cdots & h_{1n} \\ h_{21} & h_{22} & \cdots & h_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ h_{n1} & h_{n2} & \cdots & h_{nn} \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_n \end{pmatrix} \qquad \hat{y}_i = h_{i1}y_1 + h_{i2}y_2 + \cdots + h_{in}y_n$$

Diagonal elements h_{ii} are **leverage** of observation i how much impact value of y_i has on $\hat{y}_i$ vs. n-1 others

SLR: $h_{ii} = \frac{1}{n} + \frac{(x_i - x_n)^2}{\sum_{j=1}^{n}(x_j - \bar{x})^2}$ ratio of distance of own x-value from center to total variation of predictor

$\sum_{i=1}^{n} h_{ii} = p+1 \Rightarrow$ average leverage $= \frac{p+1}{n}$

$\sum_{j=1}^{n} h_{ij}^2 = h_{ii}$ b/c H is idempotent

If $h_{ii} \approx 1$, other $h_{ij} \approx 0$. observation i really far from rest of data $\hat{y}_i = h_{ii}y_i + \sum_{j \neq i} h_{ij}y_j \approx y_i$

$0 \leq h_{ii} \leq 1$ tells us fraction of $\hat{y}_i$ due to y_i vs. other responses

OUTLIERS

Statistical Outlier far from outer quartiles

Regression Outlier far from trend / conditional means

Residual distance between observed response & trend / fitted value

$\hat{y} = X(X^TX)^{-1}X^TY = HY$ $\hat{e} = Y - \hat{Y} = Y - HY = (1-H)Y$

observed responses weighted by leverage, can be used to measure potential of each observation to attract regression line

$$\hat{e} = (1-H)Y = \begin{pmatrix} 1-h_{11} & -h_{12} & \cdots & -h_{1n} \\ -h_{21} & 1-h_{22} & \cdots & -h_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ -h_{n1} & -h_{n2} & \cdots & 1-h_{nn} \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

$\hat{e}_i = -h_{i1}y_1 - h_{i2}y_2 - \cdots + (1-h_{ii})y_i - \cdots - h_{in}y_n$

$\hat{e}_i = (1-h_{ii})y_i - \sum_{i \neq j} h_{ij}y_i$

Covariance of residuals depends on leverage $Cov(\hat{e}|X) = \sigma^2(1-H)$

Want residuals to have same properties we assume about errors, standardize each $\hat{e}_i$ by variance

$$r_i = \frac{\hat{e}_i}{s\sqrt{1-h_{ii}}}$$ have constant variance now

INFLUENCE

Both leverage points & outliers have potential to change estimated relationship

Problematic observation can influence how model is estimated by affecting how

- all fitted values are estimated
- its own fitted value is estimated
- at least one coefficient is estimated

To identify an **influential observation**, use 3 meaures of influence

Each is a **delete-one measure** (fit new models after deleting a single observation)

Measure influence using Cook's distance. Fit model using all n observations, then n-1.

Difference in estimated trend tells us influence of deleted observation

Either compare coefficients $\hat{\beta}$ or fitted values $\hat{Y}$

$$D_i = \frac{(\hat{Y}_{(i)} - \hat{Y})^T (Y_{(i)} - \hat{Y})}{(p+1)s^2} = \frac{(\hat{\beta}_{(i)} - \hat{\beta})^T (\hat{\beta}_{(i)} - \hat{\beta})}{(p+1)s^2}$$

i represents model w/ the i^{th} observation deleted

Just use

$$D_i = \frac{1}{(p+1)} \frac{h_{ii}}{(1-h_{ii})}$$

DFBETAS (difference in betas)

$$\left| DFBETAS_{j(i)} = \frac{\hat{\beta}_j - \hat{\beta}_{j(i)}}{\sqrt{s^2_{(j)} (X^T X)^{-1}_{j+1, j+1}}} \right| > \frac{2}{\sqrt{n}}$$

$\hat{\beta}_{j(i)}$ coefficient j from model w/o point i

Compare change in estimated value to expected variation in values due to sampling distribution

DFFITS (difference in fitted values)

$$\left| DFFITS_i = \frac{\hat{y}_i - \hat{y}_{i(i)}}{\sqrt{s^2_{(i)} h_{ii}}} \right| > 2\sqrt{\frac{p+1}{n}}$$

$\hat{y}_{i(i)}$ fitted value for observation i using model w/o i

use $DFFITS_i = \left(\frac{h_{ii}}{1-h_{ii}}\right)^{0.5} \frac{\hat{e}_i}{s_{(i)}\sqrt{1-h_{ii}}}$

MODEL SELECTION TOOLS

Used to help determine "best" model for a given purpose ↑ # predictors ⇒ ↓ bias, ↑ variance, possible over-fitting

Prediction extra predictions help explain more variation, better accuracy if not overfitted

Description too many predictors / complicated transformations hurt interpretability

LIKELIHOOD

Maximum likelihood used to determine estimators of parameters

Depends on Normality assumption $y_i | x_{i1}, \dots, x_{ip} \sim N(\beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip}, \sigma^2)$

Assuming independence, log likelihood is

$$\ln(L(\beta, \sigma^2 | Y)) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum(y_i - \beta_0 - \beta_1 x_{i1} - \dots - \beta_p x_{ip})^2$$

(RSS if β_j replaced w/ $\hat{\beta}_j$)

$$\sigma^2_{MLE} = \frac{RSS}{n} \qquad = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln\left(\frac{RSS}{n}\right) - \frac{n}{2}$$

Like R^2, log-likelihood only measures goodness w/o accounting for effect of additional predictors

Akaike's Information Criteria (AIC)

$$AIC = -2[l(\hat{\beta}, \hat{\sigma}^2 | Y) - (p+2)] \propto n\ln\left(\frac{RSS}{n}\right) + 2p$$

$n\ln(RSS/n)$ smaller the better

penalty $2p$ to control for added X's

If n is small, p is large fraction of n (ie. $n/(p+2) \leq 40$), use corrected AIC

$$AIC_c = AIC + \frac{2(p+2)(p+3)}{n-p-1}$$

smaller indicates better

Bayesian Information Criteria (BIC) has harsher penalty than AIC to favour simpler model

$$BIC = -2[\ln(L(\hat{\beta}, \hat{\sigma}^2 | Y)) - (p+2)\ln(n)] \propto n\ln\left(\frac{RSS}{n}\right) + (p+2)\ln(n)$$

smaller is better

ALL POSSIBLE SUBSETS SELECTION

4 main measures

adjusted R^2 $R^2_{adj} = 1 - \frac{RSS/(n-p-1)}{SST/(n-1)}$ larger better

smaller better:

- AIC $\propto n\ln(RSS/n) + 2p$
- Corrected AIC $\propto n\ln(RSS/n) + 2p + \frac{2(p+2)(p+3)}{n-p-1}$
- BIC $\propto n\ln(RSS/n) + (p+2)\ln(n)$

RSS drives idea of goodness; rest are penalties for # predictors

1. Compare models of each size using adjusted R^2
2. Use all 4 numerical criteria to pick the best of the best (R^2_{adj}, AIC, AIC_c, BIC)

PROS	CONS
all possible models fitted & compared	impratical for large # predictors
flexibility of defining the "best"	best of each subset does not account for issues
systematic approach	assumptions, multicolinearity, need to manually check

AUTOMATED SELECTION METHODS

At each step comput AIC / BIC, choose smallest value. Take a step, rinse & repeat until smallest AIC / BIC is obtained

Forward start w/ intercept model, add predictors

Backward start w/ full model, delete predictors

Stepwise iterate between forward & backward

PROS	CONS
less intensive than all possible subsets	all methods may not agree on preferred model
give idea of preferred model, though may not be best	violations (multicolinearity)
stepwise accounts for conditional nature of LR	do not consider data / question in decision-making