MULTIPLE LINEAR REGRESSION (MLR)

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + \varepsilon_i$$

Response surface: hyperplane in k+1 dimensions

$$\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}_{n\times 1} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1k} \\ 1 & x_{21} & \cdots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & & x_{nk} \end{bmatrix}_{n\times(k+1)} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}_{(k+1)\times 1}$$

$Y = X\beta$

$X^T Y = X^T X \beta$

$\hat{\beta} = (X^T X)^{-1} X^T Y$ → Moore-Penrose pseudoinverse

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

$\hat{Y} = HY$

Partial correlation

Effect of X_1 on Y without X_2

$$r_{YX_1 \cdot X_2} = \frac{r_{YX_1} - r_{YX_2} r_{X_1X_2}}{\sqrt{(1 - r^2_{YX_2})(1 - r^2_{X_1X_2})}}$$

semi-partial correlation

Effect of X_1 on Y w/o X_2 effect on X_1

$$sr_{Y X_1 \cdot X_2} = \frac{r_{YX_1} - r_{YX_2} r_{X_1X_2}}{\sqrt{(1 - r^2_{X_1X_2})}}$$

On adding new variable:

Increase in R^2 = SR^2 wrt newly added var and Y

Standardised reg coeffs

If data standardised → coeff also stor.

Otherwise:

$$\hat{\beta}_{std} = \hat{\beta} \times \frac{S_{X_i}}{S_Y}$$

For 1 SD change in X_i, $\hat{\beta}$ captures how many changes of SD Y will undergo

Qualitative / Categorical Variables

One-hot encode, introduces new dummy variables for each.

For n categories, $n-1$ dummy vars.

Encode 0 as something

OTHERWISE: perfect multicollinearity

REGRESSION

SLR:

$Y = \beta_0 + \beta_1 X + \varepsilon$

MLR:

$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n + \varepsilon$

$Y = \beta_0 + \beta_1 X^2 + \beta_2 X_3 X_2 + \varepsilon$

Non-linear:

$Y = \beta_0 + X_1^{\beta_1} + \frac{1}{\beta_1 + \beta_2 X_2} + \varepsilon$

ε: residual/random errors not explained by model. deviation from true value

SLR PROCESS (5)

1. Collect data
2. Preprocess
3. Train-test split
4. Descriptive analysis
5. Define functional form of regression (→ use scatter plot for it)
6. Estimate model parameters (→ OLS)
7. Regression model diagnostics (→ validate assumptions)
8. Test on validation
9. Deploy

OLS ASSUMPTIONS (5)

1. Regression model is linear wrt parameters
2. Independent var X is deterministic
3. Residuals ε are normally distributed
4. Conditional expected value of residual is zero: $E[\varepsilon_i | X_i] = 0$
5. Variance of residuals is constant: homoscedasticity: $Var(\varepsilon_i | X_i) = \text{constant} \ \forall X_i$
6. Residuals are uncorrelated

SOLUTION: OLS

Provides Best Linear Unbiased Estimate (BLUE):

$E[\beta - \hat{\beta}] = 0$

OLS: minimise SSE

$SSE = \sum_{i=1}^{n} \varepsilon_i^2$

Model: $Y = \beta_0 + \beta_1 X_1 + \varepsilon$

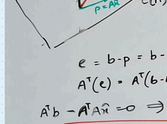
$\Rightarrow SSE = \sum_{i=1}^{n} (Y - \beta_0 - \beta_1 X_i)^2$

OLS PROOF

Linear Algebra

$Ax = b$

$\begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$



b does not lie on C(A) so we find best approx by projecting

$e = b - p = b - A\hat{x}$

$A^T(e) = A^T(b - A\hat{x}) = 0$ ← perpendicular

$A^T b - A^T A\hat{x} = 0 \Rightarrow A^T b = A^T A\hat{x}$

$\hat{x} = (A^T A)^{-1} A^T b$ → Moore-Penrose pseudoinverse

Calculus

Set partial derivatives of SSE = 0 and solve

$SSE = \sum_{i=1}^{n} (Y - \beta_0 - \beta_1 X_i)^2$

$\hat{\beta}_0 = \bar{Y} - \beta_1 \bar{X}$

$\hat{\beta}_1 = \frac{\overline{XY} - \bar{X}\bar{Y}}{\overline{X^2} - (\bar{X})^2}$

$\hat{\beta}_1 = \frac{Cov(XY)}{\sigma_x^2} = r \frac{\sigma_y}{\sigma_x}$

MLE

Assume error $\sim N(0, \sigma^2)$

$\Rightarrow y \sim N(\beta_0 + \beta_1 x, \sigma^2)$

$f(y | x; \beta_0, \beta_1, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(y - (\beta_0 + \beta_1 x_i))^2}{2\sigma^2}}$

Likelihood = $\prod_{\text{for } X} f(y|x)$ for $(x,y) \in X$

Log likelihood and maximise

-ve LL → minimise. Sub. $\beta_0 + \beta_1 x$ as y. you will get SSE.

Gradient descent

$SSE = \sum_{i=1}^{n} (Y - \beta_0 - \beta_1 X_i)^2$

$= \sum_{i=1}^{n} (Y - \hat{Y})^2$

$\frac{\partial}{\partial \beta_0}(SSE) = -2 \sum_{i=1}^{n} (Y - \beta_0 - \beta_1 X_i)$

$\frac{\partial}{\partial \beta_1}(SSE) = -2 \sum_{i=1}^{n} X_i (Y - \beta_0 - \beta_1 X_i)$

New update:

$\beta_0' = \beta_0 - \eta \frac{\partial (SSE)}{\partial \beta_0}$

$\beta_1' = \beta_1 - \eta \frac{\partial (SSE)}{\partial \beta_1}$

PARAMETER INTERPRETATION

$Y = \beta_0 + \beta_1 X + \varepsilon$

$\frac{\partial Y}{\partial X} = \beta_1$

β_0: expected value of Y when X is zero (X could be anything)

β_1: rate of change of Y wrt X. change in Y for unit X here (X could be anything, +ve for β_1)

VALIDATION OF SLR

① R^2: coeff of determination

measures proportion of variance captured by model

$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2$

$SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$

$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$

$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} = r^2$

② Regression coef. hypothesis test ← check for linear association b/w X, Y

$H_0: \beta_1 = 0$

$H_A: \beta_1 \neq 0$

$\hat{\beta}_1$ → normally distributed $N\left(\beta_1, \frac{\sigma^2}{\sum (x_i - \bar{x})^2}\right)$

But we don't know σ → replace w/ S_ε

test statistic becomes t-distributed

If $\beta_1 = 1$:

say $S_\varepsilon(\hat{\beta}) = 0.2 \Rightarrow 1.2 \times \frac{1}{5} = 1$

5 std errors away from 0

$S_\varepsilon = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n-2}}$ ← β_0, β_1 estimated

$S_\varepsilon(\hat{\beta}_1) = \frac{S_\varepsilon}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2}}$ ← horizontal variation of vertical variation

$t\text{-statistic} = \frac{\hat{\beta}_1 - \beta_1}{S_\varepsilon(\hat{\beta}_1)}$

dof: n-2

③ Test for overall model: ANOVA

H_0: no linear rel b/w Y and any X (all reg. coeff = 0)

H_1: Some linear rel exists

$F\text{-statistic} = \frac{MSR}{MSE} = \frac{SSR/1}{SSE/(n-2)}$

④ Residual analysis

(a) Normally distributed

- P-P plot b/w:
  - observed cumulative probability
  - expected cumulative probability
- Sort: $e_1 < e_2 < \cdots < e_k < \cdots < e_n$

Obs prob = $\frac{k}{n}$ for e_k

Expected prob = $\Phi\left(\frac{e_k - 0}{S_\varepsilon}\right)$

(b) Homoscedasticity

variance should be independent of x

(c) Functional form of regression

any pattern in pair. plot = bad

model will then need to be transformed

- → transform variables (log, sqaure, exp)
- → add new vars
- → non-linear model

⑤ Outlier analysis

Z-score

$Z = \frac{\hat{y}_i - \bar{y}}{\sigma_y} > 3 = \text{outlier}$

Mahalanobis Distance

1. Transform cols into uncorrelated vars
2. Scale cols to make variances 1
3. Finally calc Euclidean distance

$MD = \sqrt{(x - \bar{x})^T C^{-1} (x - \bar{x})}$ ← not same distribution

$MD > \chi^2_{\text{critical}}$ w/ dof = no of X's

$\Rightarrow$ outlier

SLR VALIDATION

① R^2 score

Been there done that

② Hypothesis testing with β_1

$\beta_1 = 0$ → no relationship b/w X, Y

$\Rightarrow H_0: \beta_1 = 0$; $H_A: \beta_1 \neq 0$ } two tailed t-test

$\hat{\beta}_1$ → sampled from a normal distri of all $\hat{\beta}_1$

$SE(\hat{\beta}_1)$ → SD of distri

$t\text{-statistic} = \frac{\hat{\beta}_1 - \beta_1}{S_\varepsilon(\hat{\beta}_1)} = \frac{\hat{\beta}_1}{S_\varepsilon(\beta_1)}$

where

$S_\varepsilon = \sqrt{\frac{\sum_{i=1}^{n} (X_i - \hat{X}_i)}{n-2}} = \sqrt{\frac{SSE}{n-2}}$

Sampling error over regression constant

$\frac{n}{\sqrt{12 \text{ dist } 12 \text{ closest}}} = \frac{S_1}{\sqrt{\sum_i (x_i - \bar{x}_i)^2}} = \frac{S_1}{\sqrt{SSE}}$

③ Hypothesis testing for whole model

ANOVA,

H_0: no linear relationship b/w Y and all X_i

H_A: linear relationship ✓

$F\text{-statistic} = \frac{MSR}{MSE} = \frac{SSR/1}{SSE/n-2}$

④ Residual analysis

- residuals ~ N → P-P plot
- homoscedasticity
  - plot standardized residuals against standardized $\hat{y}$
  - any pattern ≠ homoscedastic: variance depends on indep. var
- ensure functional form
  - residual plot pattern = y
  - improve by transform, adding more, or non-linear

⑤ Outlier analysis

- Z-score
  - ≥ 3 → outlier
- Mahalanobis distance
  - accounts for scaling and correlation

$M_d = \sqrt{(\vec{x} - \vec{\mu})^T C^{-1} (\vec{x} - \vec{\mu})}$

MULTIPLE LINEAR REGRESSION

$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_2 + \cdots + \beta_k X_{ki} + \varepsilon_i$

$\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1K} \\ 1 & x_{22} & \cdots & x_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & & x_{nK} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}$

Y X β ε

$Y = X\beta$

$X^T Y = X^T X \beta$

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

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

hat matrix H

$\hat{Y} = HY$

Partial corr

Corr b/w Y and X_i after removing effect of X_2 on both X_i and Y

$r_{YX_i \cdot X_2} = \frac{r_{YX_i} - (r_{YX_2})(r_{X_iX_2})}{\sqrt{(1 - r^2_{YX_2})(1 - r^2_{X_iX_2})}}$

Semi-partial corr

Same as before, but we remove effect of X_2 only from X_i

$sr_{X_i(Y \cdot X_2)} = \frac{r_{YX_i} - (r_{YX_2})(r_{X_iX_2})}{\sqrt{(1 - r^2_{X_iX_2})}}$

Standardized regression coefficients

→ Scales of diff vars vary.

→ You can standardize and then train model,

OR

$\text{Standardized } \beta = \hat{\beta} \times \left(\frac{\sigma_{X_i}}{\sigma_Y}\right)$

for 1 SD change in response, $\hat{\beta}_i$ captures no of SD changes in explanatory

Qualitative / Categorical Variables

One-hot encoding

For n values of categorical var → introduce $n-1$ β, dummy vars

MLR Validation

① Adjusted R^2 score

more vars → SSE decrease → R^2 value increases

SST constant

on adding new vars even if they do not contribute

$R^2 = 1 - \frac{SSE}{SST}$; $\frac{SSE}{n-k-1} \left(\frac{n-1}{SST}\right)$

$\text{Adj } R^2 = 1 - \frac{(SSE/n-k-1)}{(SST/n-1)} = 1 - \left(\frac{n-1}{n-k-1}\right)\left(\frac{SSE}{SST}\right) = 1 - \left(1 - R^2\right)\left(\frac{n-1}{n-k-1}\right)$

Adj $R^2 \leq R^2$ always

Only increases when new var contributes to explanatory power

k: no of predictors

② Hypothesis test for individual vars - t-test

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

linear func

follows Normal distri

$\sigma_{\hat{\beta}}$ → derived from sample

So for some X_i,

$H_0: \beta_i = 0$ (no relationship)

$H_A: \beta_i \neq 0$

two tailed t-test

$\text{t-test statistic} = \frac{\hat{\beta} - \beta}{S_\varepsilon(\hat{\beta})} = \frac{\hat{\beta}}{S_\varepsilon(\hat{\beta})}$

③ Verifying entire model using F-test

$F\text{-test statistic} = \frac{MSR}{MSE} = \frac{SSR/k}{SSE/n-k-1}$

$H_0: \beta_1 = \beta_2 = \cdots = \beta_K = 0$

H_A: Not all $\beta = 0$

NOTE: this does not tell us which β are non-zero

④ Verifying parts of the model with F-test

Full model

$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_K X_K$

Reduced model $p < K$

$Y = \alpha_0 + \alpha_1 X_1 + \cdots + \alpha_p X_p$

Objective of partial F-test: Check additional vars to X for significance

$H_0: \beta_{p+1} = \beta_{p+2} = \cdots = \beta_K = 0$

H_A: Some non-zero

$\text{F-test statistic} = \frac{(SSE_R - SSE_F)/(K-p)}{SSE_F/n-k-1}$

⑤ Residual analysis

- normal distri residuals → P-P plot
- homoscedasticity
  - plot std. residuals against std $\hat{y}$
  - pattern = heteroscedasticity
- ensure functional form (linear instead of log-linear)

⑥ Multicollinearity & Variance Inflation Factor.

indep vars corr b/w each other

Say $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$.

$x_1 = \gamma_0 + \gamma_1 x_1$

Let R_{12}^2 be R^2 value for second model

$VIF = \frac{1}{1 - R_{12}^2}$ ← tolerance

$\sqrt{VIF}$ and by which → std errors inflates, t-stat deflates

$t_{\text{actual}} = \frac{\hat{\beta}}{S_\varepsilon(\hat{\beta})} \cdot \sqrt{VIF}$

Threshold (rule) = 4

⑦ Auto-Correlation: Durbin-Watson

errors correlated in time series → autocorr.

e_t, e_{t-k} → autocorr of lag k

→ underestimate std. errors → inflate t-stat

Let ρ be corr b/w (e_t, e_{t-1})

$H_0: \rho = 0$

$H_A: \rho \neq 0$

$\text{D-W test stat} = \frac{\sum_{i=2}^{n} (e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2} = 2\left(1 - \frac{\sum_{i=2}^{n} e_i e_{i-1}}{\sum_{i=1}^{n} e_i^2}\right)$

$D \in [0, 4]$

two critical values: D_L, D_U

$D < D_L$: +ve corr

$(4-D) < D_L$: -ve corr

$D > D_U$: no evidence for +ve corr

$(4-D) > D_U$: " " " -ve corr

$D_L < D < D_U$ } inconclusive

$D_L < (4-D) < D_U$ }

$D \approx 2$ → no autocorr

MLR VALIDATION

① Adjusted R^2 (coeff of multiple determination)

$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$

$\text{Adjusted } R^2 = 1 - \frac{SSE/n-(k+1)}{SST/n-1} = 1 - \frac{SSE}{n-k-1}\left(\frac{n-1}{SST}\right)$

$= 1 - \frac{SSE}{SST}\left(\frac{n-1}{n-k-1}\right)$

$\text{Adjusted } R^2 = 1 - (1 - R^2)\left(\frac{n-1}{n-k-1}\right)$

R^2 → non-decreasing

adj R^2 → can decrease or increase

adj $R^2 \leq R^2$

② Hypothesis test for reg. coefficient

Same idea as SLR.

$S_\varepsilon = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n-k-1}}$

$S_\varepsilon(\hat{\beta}_j) = \sqrt{S_\varepsilon^2 \left[(X^T X)^{-1}\right]}$

$t\text{-statistic} = \frac{\hat{\beta}_j - \beta_j}{S_\varepsilon(\hat{\beta}_j)}$

$H_0: \beta_j = 0$

$H_A: \beta_j \neq 0$

③ Significance of entire model: ANOVA

$F\text{-statistic} = \frac{SSR/k}{SST/n-1}$

$H_0: \beta_1 = \cdots = \beta_K = 0$

H_A: otherwise

④ Significance of part of the model

Full: $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_K X_K$

Reduced: $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$ $p < k$

$H_0: \beta_{p+1} = \beta_{p+2} = \cdots = \beta_K = 0$

H_1: otherwise

$F\text{-statistic} = \frac{(SSE_R - SSE_F)/(k-h)}{MSE_F} \sim \frac{SSE}{n-k-1}$

⑤ Residual analysis.

Same deal as SLR

MULTICOLLINEARITY: VARIANCE INFLATION FACTOR

$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$

To check for corr b/w X_1, X_2:

$X_1 = \alpha_0 + \alpha_1 X_2$ → R^2 here: R^2_{12}

$VIF = \frac{1}{1 - R^2_{12}}$

- VIF: amt by which variance of std error is inflated
- $\sqrt{VIF}$: " " " std. dev. " " " "
- " " " t-statistic of X_1 is deflated

$t_{\text{actual}} = \frac{\hat{\beta} - \beta^0}{S_\varepsilon(\hat{\beta})} \times \sqrt{VIF}$

$VIF_{\text{threshold}} = 4$

Remove var

Mean centre

PCA/SVD

LASSO/RIDGE

AUTOCORRELATION: Durbin-Watson Test

$Y_t = \beta_0 + \beta_1 X_t + \varepsilon_t$

ε_t and ε_{t-1} are related → std error deflated → t-statistic inflated

Say ρ = corr b/w $\varepsilon_t, \varepsilon_{t-1}$

$H_0: \rho = 0$

$H_A: \rho \neq 0$

$D = \frac{\sum_{i=2}^{n} (e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2} = 2\left(1 - \frac{\sum_{i=2}^{n} e_i e_{i-1}}{\sum_{i=1}^{n} e_i^2}\right)$

$D \in [0, 4]$

$D < D_L$: positive corr

$D > D_U$: no +ve corr

$D_L < D < D_U$: inconclusive

$(4-D) < D_L$: negative corr

$(4-D) > D_U$: no negative corr

$D_L < (4-D) < D_U$: inconclusive

$D \sim 2$ → no autocorr