Simple Linear Regression

same
① The distribution of X is arbitrary (and perhaps X is even non-random)
② $Y = \beta_0 + \beta_1 X_1 + \varepsilon$. If $X_1 = x_1$, then $Y = \beta_0 + \beta_1 x_1 + \varepsilon$ for some coefficients β_0, β_1 and random noise ε.
③ $\varepsilon \sim N(0,1)$ $\quad E(\varepsilon)=0$, $Var(\varepsilon)=\sigma^2$
④ ε is independent of X_1 and independent across observations. ε is uncorrelated with X_1.

Linear correlation $\Rightarrow$ those having 0 correlation might dependent non-linearly.

Optimal Linear Predictor

$$(\beta_0^*, \beta_1^*) = \underset{(\beta_0,\beta_1)}{\operatorname{argmin}} E_{X,Y}[(Y-(\beta_0+\beta_1 X))^2]$$

$$\begin{cases} \beta_0^* = E(Y) - \beta_1^* E(X) \\ \beta_1^* = \dfrac{Cov(X,Y)}{Var(X)} \end{cases} \Rightarrow \begin{cases} \hat\beta_0 = \bar y - \dfrac{S_{xy}}{S_{xx}}\bar x_1 \\ \hat\beta_1 = \dfrac{S_{xy}}{S_{xx}} \end{cases}$$

Least square Estimates

$$S(\beta) = \frac{1}{n}\sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_{i1})^2 = \frac{1}{n}(Y-X\beta)^T(Y-X\beta)$$

$$\begin{bmatrix}\hat\beta_0 \\ \hat\beta_1\end{bmatrix} = \begin{bmatrix} n & \sum x_{i1} \\ \sum x_{i1} & \sum x_{i1}^2 \end{bmatrix}^{-1}\begin{bmatrix}\sum y_i \\ \sum x_{i1}y_i\end{bmatrix} = (X^TX)^{-1}X^TY$$

$E(\hat\beta_1)=\beta_1$, $E(\hat\beta_0)=\beta_0$, $Var(\hat\beta_1)=\dfrac{\sigma^2}{S_{xx}}$, $Var(\hat\beta_0)=\sigma^2\left(\frac{1}{n}+\frac{\bar x_1^2}{S_{xx}}\right)$

$$\begin{cases} \hat\beta_1 = \sum \frac{x_{i1}-\bar x_1}{S_{xx}} y_i & \sum c_i = 0 \quad \sum c_i x_i = 1 \\ \hat\beta_0 = \sum (\frac{1}{n} - \bar x_1 c_i) y_i & \sum b_i = 1 \quad \sum b_i x_i = 0 \end{cases} \qquad \hat\beta = \underbrace{(X^TX)^{-1}X^T}_{A} Y$$

$$Var(\hat\beta|X) = A\,Var(Y|X)A^T = A\sigma^2 I A^T = \sigma^2 AA^T = \sigma^2(X^TX)^{-1}$$

$$\begin{bmatrix} Var(\hat\beta_0|X) & Cov(\hat\beta_0,\hat\beta_1|X) \\ Cov(\hat\beta_1,\hat\beta_0|X) & Var(\hat\beta_1|X)\end{bmatrix} = \frac{\sigma^2}{nS_{xx}}\begin{bmatrix}\sum x_{i1}^2 & -\sum x_{i1} \\ -\sum x_{i1} & n\end{bmatrix}$$

① The Law of Total Expectation
$E(\hat\beta_1) = E_X[E_{Y|X}(\hat\beta_1|X_{11}\cdots X_{1n})] = E_X(\beta_1) = \beta_1$
② The Law of Total Variance
$Var(\hat\beta_1) = E_X[Var_{Y|X}(\hat\beta_1|X_{11}\cdots X_{n1})] + Var[E_{Y|X}(\hat\beta_1|\cdots)]$

Residuals

$$e_i = y_i - \hat y_i = y_i - \hat\beta_0 - \hat\beta_1 x_{i1}$$

$$\begin{cases} \frac{\partial S}{\partial \beta_0} = \frac{1}{n}(-2)\sum_{i=1}^n (y_i - \hat\beta_0 - \hat\beta_1 x_{i1}) = 0 \\ \frac{\partial S}{\partial \beta_1} = \frac{1}{n}(-2)\sum_{i=1}^n (y_i - \hat\beta_0 - \hat\beta_1 x_{i1})x_{i1} = 0 \end{cases}$$

① $\sum_{i=1}^n (y_i - \hat y_i) = \sum_{i=1}^n e_i = 0$
② $\sum_{i=1}^n e_i x_{i1} = 0$
③ $\sum_{i=1}^n y_i = \sum_{i=1}^n \hat y_i$
④ $\sum_{i=1}^n \hat y_i e_i = 0$

$$\hat\sigma^2 = \frac{SS_{RES}}{n-2} = MS_{RES}$$ residual mean square

$$X\hat\beta = \underbrace{X(X^TX)^{-1}X^T}_{\text{Hat matrix}}Y = HY$$

① H is symmetric
② H is idempotent $H^TH = H$
③ $I_n - H$ is idempotent

$SS_{RES} = (Y-X\hat\beta)^T(Y-X\hat\beta) = Y^T(I_n - H)Y$
If v is a random vector, $E(v)=\mu$, $Var(v)=\Sigma$
For a constant matrix A, $E(v^TAv) = trace(A\Sigma) + \mu^TA\mu$

$E(SS_{RES}|X) = \sigma^2(n - trace(H))$
$trace(H) = trace(X(X^TX)^{-1}X^T) = trace(X^TX(X^TX)^{-1}) = p$

standard error (se) $= \sqrt{Var(\hat\beta_i)}$
estimated standard error (ese)

$$\begin{cases} ese(\hat\beta_0) = \sqrt{\hat\sigma^2\left(\frac{1}{n} + \frac{\bar x_1^2}{S_{xx}}\right)} \\ ese(\hat\beta_1) = \sqrt{\hat\sigma^2\frac{1}{S_{xx}}} \end{cases}$$

GN-SLR

① $\hat\beta_1 \sim N(\beta_1, \frac{\sigma^2}{S_{xx}})$ $\quad \hat\beta_0 \sim N(\beta_0, \sigma^2(\frac{1}{n} + \frac{\bar x_1^2}{S_{xx}}))$ $\quad \frac{(n-2)\hat\sigma^2}{\sigma^2} \sim \chi^2_{n-p}$
② $T_1 = \frac{\hat\beta_1 - \beta_1}{ese(\hat\beta_1)} \sim t_{n-p}$ $\quad T_0 = \frac{\hat\beta_0 - \beta_0}{ese(\hat\beta_0)} \sim t_{n-p}$
A $1-\alpha$ confidence interval $CI(\beta_i) = [\hat\beta_i \mp k\cdot ese(\hat\beta_i)]$
③ $\begin{cases} H_0: \beta_1 = c \\ H_1: \beta_1 \neq c \end{cases}$ We reject H_0 if $|T_1| \geq k \equiv t_{\frac{\alpha}{2}, n-2}$ $\quad T_1 = \frac{\hat\beta_1 - c}{ese(\hat\beta_1)}$

When H_0 is attained $\Rightarrow$ β_1 is statistically undistinguishable from 0

④ $CI(m(x_{01})) = [\hat m(x_{01}) \mp k\sqrt{MS_{RES}(\frac{1}{n} + \frac{(x_{01}-\bar x_1)^2}{S_{xx}})}]$

$PI(y) = [\hat m(x_{01}) \mp k\sqrt{MS_{RES}(1 + \frac{1}{n} + \frac{(x_{01}-\bar x_1)^2}{S_{xx}})}]$

Multiple Linear Regression [Multiple Regression]

① There are k predictors $X_1, X_2 \cdots X_k$. We make no assumption about their distribution. They may or may not be dependent.
② There is a single response Y.
③ Linear model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \varepsilon = X\beta + \varepsilon$ for some constant $\beta_0, \beta_1 \cdots \beta_k$
④ The noise has $E(\varepsilon) = E(\varepsilon|X) = 0$ and $Var(\varepsilon) = Var(\varepsilon|X) = \sigma^2$

β_0: The expected value of Y when $X_1 = \cdots X_k = 0$
β_i: If we select two sets of the cases from the distribution of the data, where x_i differs by 1, we expect Y to differ by β_i on average

variance of the noise around the true regression line.

$\hat\beta = (X^TX)^{-1}X^TY$
① $n \geq p$
② The predictors must be linearly independent (Check column rank of X: $k+1$)

$E(\hat\beta) = \beta$, $Var(\hat\beta) = \sigma^2(X^TX)^{-1}$ $\quad \hat\sigma^2 = \frac{\|Y - X\hat\beta\|_2^2}{n-p}$

$\begin{cases} \varepsilon \sim N(0_n, \sigma^2 I_n) \\ \varepsilon \text{ is independent of } X \end{cases} \Rightarrow Y|X \sim N(X\beta, \sigma^2 I_n)$
$\hat\beta \sim N(\beta, \sigma^2(X^TX)^{-1})$

① estimate
$\hat\beta_i \mp k\sqrt{\hat\sigma^2 \cdot C_{ii}}$ $\quad ese(\hat\beta_i)$ $\quad C = (X^TX)^{-1}$, $k = t_{\frac{\alpha}{2}, n-p}$

② mean response
$\hat m(x_0) \mp k\cdot\sqrt{\hat\sigma^2 \cdot x_0(X^TX)^{-1}x_0^T}$

③ new observation
$\hat m(x_0) \mp k\cdot\sqrt{\hat\sigma^2(1 + x_0(X^TX)^{-1}x_0^T)}$

$R^2 = \frac{SSR}{SST} = 1 - \frac{SS_{RES}}{SST}$ $\quad \hat\beta_1 S_{xy} = \hat\beta_1^2 S_{xx}$
$R^2 = 0$ if $\hat\beta_1 = 0$, $R^2 = 1$ if $SS_{RES} = 0$

$$\frac{\hat\beta_1^2 S_{xx}}{\hat\beta_1^2 S_{xx} + SS_{RES}} = \frac{\hat\beta_1^2 \frac{S_{xx}}{n}}{\hat\beta_1^2 \frac{S_{xx}}{n} + \frac{SS_{RES}}{n}} \xrightarrow{n\to\infty} \frac{\hat\beta_1^2 Var(x)}{\hat\beta_1^2 Var(x) + \sigma^2}$$

① We can manipulate the data that
- $R^2 = 0$ but the model is correct ($Var(x)$ small)
- $R^2 = 1$ but the model is wrong ($Var(x)$ large)

② only can be compared with same dataset with same untransformed response variable.

$R^2_{adj} = 1 - \frac{SS_{RES}/(n-p)}{SST/(n-1)}$
- $R^2 \to 1$ when $p \to n$
- $R^2_{adj} = R^2$ when $p = 1$

① low R^2 and low p-value (p-value $< \alpha$)
The model doesn't explain much of variation of the data but the model is significant
(Including x_i into the model is necessary but not sufficient)
② low R^2 and high p-value
The model doesn't explain much of the data and x_i is not useful. <worst>
③ High R^2 and low p-value
Model is sufficient, x_i is useful <best>

F-test

$y^T[I_n - H_1]y = y^T[I_n - H]y + y^T[H - H_1]y \quad (H_1 = \frac{1}{n}1_n 1_n^T)$

$\underbrace{\sum_{i=1}^n (y_i - \bar{y})^2}_{SST} = \underbrace{\sum_{i=1}^n (y_i - \hat{y}_i)^2}_{SSRes} + \underbrace{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}_{SSR}$

"systematic variation" in Y due to $Y \sim X$ linear

aggregate measure of fitted regression line relationship

$SST/\sigma^2 \sim \chi^2_{n-1} \quad SSRes/\sigma^2 \sim \chi^2_{n-p} \quad SSR/\sigma^2 \sim \chi^2_{p-1}$

ANOVA reject null if $F_0 > F_{\alpha, n-p}$ $(p < \alpha)$

Source	SS	df	MS	F
Regression	SSR	$p-1$	$MSR = SSR/(p-1)$	$\frac{SSR/(p-1)}{SSRes/(n-p)}$
Residual	SSRes	$n-p$	$MSRes = SSRes/(n-p)$	
Total	SST	$n-1$		

$(T_0)^2 = F_0$ [p-value same $P(F > F_0) = P(|T| > |T_0|)$]

1. If retain $H_0: \beta_1 = 0$ (don't find any significant share of variance associated with the regression).
   (a) The intercept-only model is correct
   (b) $\beta_1 \neq 0$, but the test don't have enough power to detect it ($\frac{\hat{\beta}_1}{se(\beta_1)}$)
   (c) The relationship is non-linear, but the best linear approximation has zero slope.
2. If reject H_0 (doesn't mean the SLR model is right, only that the non-zero slope linear model predict better than the intercept-only model).

$\begin{cases} H_0: \beta_{(2)} = 0_{r \times 1} \\ H_1: \beta_{(2)} \neq 0_{r \times 1} \end{cases} \Rightarrow \begin{cases} H_0: Y = X_{(1)}\beta_{(1)} + \varepsilon & \text{reduced model} \\ H_1: Y = X_{(1)}\beta_{(1)} + X_{(2)}\beta_{(2)} + \varepsilon & \text{full model} \end{cases}$

$F_0 = \frac{\overline{SSR}(\beta_{(2)}|\beta_{(1)})/r}{\overline{SSR}/(n-p)} \qquad \overline{SSR}(\beta) - \overline{SSR}(\beta_{(1)})$

Sequential F test

$\begin{cases} H_0: Y = \beta_0 + \varepsilon \\ H_1: Y = \beta_0 + \beta_1 X_1 + \varepsilon \end{cases} \quad F_1 = \frac{\overline{SSR}(\beta_1|\beta_0)/1}{\overline{SSRes}(\beta_2, \beta_1, \beta_0)/(n-4)}$ full

$\begin{cases} H_0: Y = \beta_0 + \beta_1 X_1 + \varepsilon \\ H_1: Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon \end{cases} \quad F_2 = \frac{\overline{SSR}(\beta_2|\beta_1, \beta_0)/1}{\overline{SSRes}(\beta_2, \beta_1, \beta_0)/(n-4)}$

$\overline{SSR}(\beta_2, \beta_1|\beta_0) = \overline{SSR}(\beta_2, \beta_1, \beta_0) - \overline{SSR}(\beta_0)$

$= \overline{SSR}(\beta_1|\beta_0) + SSR(\beta_2|\beta_1, \beta_0) + \overline{SSR}(\beta_2|\beta_1, \beta_0)$

global model $\overline{SSR}(\beta_{(1)}|\beta_{(1)}) = SSR(\beta)$

$\overline{SST} = SSRes + \overline{SSR}$

$\sum y_i^2 = \sum (y_i - \hat{y}_i)^2 + \sum \hat{y}_i^2 \Rightarrow \overline{SSR} = \sum_{i=1}^n \hat{y}_i^2 = \hat{\beta}^T X^T Y$

Model Adequacy Checking

residual $e = Y - \hat{Y} = (I_n - H)Y$

① The distribution of e should have a center around 0. $E(e|X) = 0_n$.

② The covariance between $\hat{Y}$ and e is zero. $Cov(e, \hat{Y}) = 0_{n \times n}$.

③ The variance of e is roughly constant. $Var(e|X) = \sigma^2(I_n - H)$

GN-MLR e is Gaussian with mean 0 and variance $\sigma^2 I_n$. $e_i \sim N(0, \sigma^2)$

quantile and quantile plot $F^{-1}(p) = \sigma\Phi^{-1}(p) + \mu \quad F \sim N(\mu, \sigma^2) \quad \Phi \sim N(0,1)$

$\Rightarrow \hat{F}^{-1}(\frac{i}{n}) = e_{(i)} \sim \Phi^{-1}(\frac{i}{n})$

Transformation

① variance-stability transformation $\sigma^2 \propto \begin{cases} E(Y) & Y = \sqrt{Y} \quad \text{Poisson} \\ E(Y)(1-E(Y)) & Y' = \sin^{-1}(Y) \quad \text{Binomial} \\ E(Y)^2 & Y' = \ln(Y) \end{cases}$

② linearity

Leverage

$\hat{Y}_i = \sum_{j=1}^n h_{ij} Y_i$ the amount of leverage exerted by the ith observation Y_i on the ith fitted value $\hat{Y}_i$.

$h_{ii} = x_i(X^TX)^{-1}x_i^T$ large h_{ii} $\Rightarrow$ ith observation has high leverage

Influence

Cook's Distance $D_i = \frac{(\hat{\beta}_{(i)} - \hat{\beta})^T X^T X (\hat{\beta}_{(i)} - \hat{\beta})}{c \cdot p \cdot MSRes}$ (X^TX: fitted estimate with ith observation removed)

$= \frac{e_i^2}{\hat{\sigma}^2} \frac{h_{ii}}{(1-h_{ii})^2} \frac{1}{p}$

(standard residual)^2 leverage

$X\hat{\beta}_{(i)} - X\hat{\beta} = \hat{Y}_{(i)} - \hat{Y} \Rightarrow D_i = \frac{(\hat{Y}_{(i)} - \hat{Y})^T (\hat{Y}_{(i)} - \hat{Y})}{p \cdot MSRes}$

Squared distance that the fitted value moves when the ith observation is removed.

Model Selection

generalization error $G = E[(Y^0 - \hat{m}(X^0))^2]$

$G = E[\frac{1}{n}\sum_{i=1}^n (Y_i^0 - \hat{Y}_i)^2] \quad$ testing error

$= E[\frac{1}{n}\sum_{i=1}^n (Y_i - \hat{Y}_i)^2] + \frac{2}{n}\sigma^2 p \quad$ training error $\quad MSE \quad \frac{SSRes}{n}$

① Mallows C_p ↓

$C_p = \frac{1}{n}\sum_{i=1}^n (Y_i - \hat{Y}_i)^2 + \frac{2}{n}\hat{\sigma}^2 p$

② adjusted R^2 ↑

$R^2_{adj} = 1 - \frac{MSE \cdot \frac{n}{n-p}}{SST/n}$

$MSE \cdot \frac{1}{1 - p/n} \approx MSE(1 + \frac{p}{n}) = MSE + MSE \cdot \frac{p}{n} \xrightarrow{n \to \infty} MSE + \sigma^2 \cdot \frac{p}{n} = G$

③ AIC & BIC ↓

$AIC = n \cdot \ln(MSE) + 2p = n \cdot \ln(\frac{SSRes}{n}) + 2p$

$BIC = n \cdot \ln(MSE) + \ln(n) \cdot p$

Penalization

$\min_{\beta} \sum_{i=1}^n (y_i - (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik}))^2$ such that

$\|\beta\|_2^2 = \beta_1^2 + \beta_2^2 + \cdots + \beta_k^2 \leq t$ ridge penalization

$\Rightarrow \min_{\beta} \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_{i1} \cdots \beta_k x_{ik})^2 + \lambda \|\beta'\|_2^2$

No-intercept model $\hat{\beta} = (X^TX + \lambda I_p)^{-1} X^T Y$

① Control model complexity

② Deal with high collinearity between predictors

③ When $n \ll p$ (MLR will fail)

Factor Predictor

with intercept $x^{(l)} = \begin{cases} 1 & x = l \\ 0 & \text{otherwise} \end{cases}$ for $l = 0 \cdots M-1$

baseline level: $l = M$

$E[Y|x_i] = \beta_0 + \sum_{l=1}^{M-1} \beta_l x_i^{(l)}$

$\beta_1 = E[Y|X=1] - E[Y|X=M]$ and M

effect contrast between level $1 \cdots M-1$

$\beta_0 = E[Y|X=M]$

baseline level effect

no intercept $x^{(l)} = \begin{cases} 1 & x = l \\ 0 & \text{otherwise} \end{cases}$ for $l = 0 \cdots M$

$E[Y|x_i] = \sum_{l=1}^{M} \beta_l' x^{(l)} \quad \beta_l' = \begin{cases} \beta_0 & l = M \\ \beta_l + \beta_0 & l = 1 \cdots M-1 \end{cases}$ effect of level $m = 1 \cdots M$

Interaction

continuous $E[Y|X] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{12} X_1 \cdot X_2$ interaction

modification of the effect of outcome of X_1 on the presence of X_2 (VV)

factor ① $E[Y|X] = \beta_0 + \sum_{m=1}^{M_1-1} \beta_m x^{(m)} + \sum_{l=1}^{M_2-1} \beta_l x^{(l)} + \sum_{m=1}^{M_1-1}\sum_{l=1}^{M_2-1} \beta_{ml} x^{(m)} x^{(l)}$

② $E[Y|X] = \beta_0 + \sum_{m=1}^{M_1} \beta_m x^{(m)} + \sum_{l=1}^{M_2} \beta_l x^{(l)} + \sum_{m=1}^{M_1}\sum_{l=1}^{M_2} \beta_{ml} x^{(m)} x^{(l)}$

$Var(X \pm Y) = Var(X) + Var(Y) \pm 2Cov(X,Y)$

$E(X^2) = [E(X)]^2 + Var(X)$

$E(XY) = Cov(X,Y) + E(X)E(Y)$

$Cov(X,Y) = E[(X - E(X))(Y - E(Y))]$

$S_{XX} = \sum_{i=1}^n (x_{i1} - \bar{x}_1)^2 = \sum_{i=1}^n (x_{i1} - \bar{x}_1)x_{i1} = \sum_{i=1}^n x_{i1}^2 - n\bar{x}_1^2$

$S_{XY} = \sum (x_{i1} - \bar{x}_1)(y_i - \bar{y}) = \sum (x_{i1} - \bar{x}_1)y_i = \sum x_{i1}y_i - n\bar{x}\bar{y}$