Multiple Linear Regression

1.1 Examples.

- **Polynomial Regression**

$$E[Y|X] = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \cdots + \beta_k x_1^k = X\beta$$

$$X = [1, x_1, x_1^2 \cdots x_1^k] \qquad \beta = [\beta_0 \ \beta_1 \cdots \beta_k]^T$$

- **Multiple Regression**

$$E[Y|X] = \beta_0 + \beta_1 x_1 + \cdots \beta_k x_k = X\beta$$

$$X = [1, x_1, x_2 \cdots x_k] \qquad \beta = [\beta_0 \ \beta_1 \cdots \beta_k]^T$$

- **Model with interaction**

$$E[Y|X] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 = X\beta$$

$$X = [1 \ x_1 \ x_2 \ x_1 x_2] \qquad \beta = [\beta_0 \ \beta_1 \ \beta_2 \ \beta_{12}]^T$$

- **Model Transformation**

$$E[Y|X] = \beta_0 + \beta_1 \log(x_1) + \beta_2 \sin(x_2) = X\beta$$

$$X = [1 \ \log(x_1) \ \sin(x_2)] \qquad \beta = [\beta_0 \ \beta_1 \ \beta_2]^T$$

Model Setup

In all cases, we use a linear model

$E[Y|X] = X\beta$ where $X = [1 \ x_1 \ x_2 \cdots x_k] \in R^p$

$\beta = [\beta_0, \beta_1 \cdots \beta_k]$ is a $p \times 1$ column vector

$p = 1 + k$ ($p = 2$ for SLR)

Assumption

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. X without subscript refers to $X = [X_1\ X_2 \cdots X_k]$

② There is a single response Y
(multiple responses – multivariate)

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 $E(\varepsilon) = E(\varepsilon|X) = 0$ (zero mean)
and $\text{Var}(\varepsilon) = \text{Var}(\varepsilon|X) = \sigma^2$ (constant variance)

If we have n observations $(y_i, \underbrace{1, x_{i1} \cdots x_{ik}}_{x_i})$ $i = 1 \cdots n$

$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} \cdots \beta_k x_{ik}$
$= x_i \beta + \varepsilon_i$
$(\beta_0, \beta_1 \cdots \beta_k)$

In matrix form

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

$Y_{n\times 1} = X_{n\times p}\ \beta_{p\times 1} + \varepsilon_{n\times 1}$ $(p = k+1)$

where $E(\varepsilon|X) = 0_{n\times 1}$ ← assumption
$\text{Var}(\varepsilon|X) = \sigma^2 I_n$. < diag: $\text{Var}(\varepsilon_i)$; others: $\text{Cov}(\varepsilon_i, \varepsilon_j) = 0$

$\Rightarrow E(Y|X) = X\beta$
$\text{Var}(Y|X) = \sigma^2 I_n$

1.3 Parameter Interpretation

σ^2: the variance of the noise around the true regression line.

β_0: the expected value of Y when $x_1 = 0, x_2 = 0 \cdots x_k = 0$

$$E(Y \mid x_1 = 0 \cdots x_k = 0) = x\beta$$
$$= \beta_0 + 0 \cdot \beta_1 + \cdots + 0 \cdot \beta_k$$
$$= \beta_0$$

Note: In some case, $E(Y \mid x = 0) = \beta_0$ might not have a meaning when $x = 0$.

β_j for $j = 1 \cdots k$. Take β_1 as an example.

$$E(Y \mid x_1 = x_1 + 1, x_2 = x_2, \cdots x_k = x_k) - E(Y \mid x_1 = x_1 \cdots x_k = x_k).$$
$$= (\beta_0 + \beta_1(x_1 + 1) \cdots \beta_k x_k) - (\beta_0 + \beta_1 x_1 \cdots \beta_k x_k)$$
$$= \beta_1$$

If we select two sets of cases from the distribution of the data, where x_1 differs by 1. We expect Y to differ by β_1 on average

Note: $\beta_1 \neq Y \mid (x_1 = x_1 + 1, \cdots x_k = x_k) - Y \mid (x_1 = x_1, \cdots x_k = x_k)$

1.4 Least Squares Estimation

Estimates Provided with observed data $\{(y_1, x_1) \cdots (y_n, x_n)\}$ we estimate β using LS.

We minimize $S(\beta) = \sum_{i=1}^{n} (y_i - x_i\beta)^2$

$$= (Y - X\beta)^T (X\beta - Y)$$
$$= \|Y - X\beta\|^2$$

$$\frac{\partial S}{\partial \beta_j} = -2 \sum_{i=1}^{n} (y_i - x_i\beta) x_{ij} = 0 \qquad j = 0, 1 \cdots k$$

We have $p = k + 1$ least squares normal equations.

which can also written in the matrix format

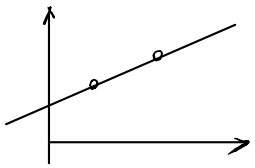
$$\nabla_\beta S(\beta) = -2X^T(Y - X\beta) = 0_p \Rightarrow X^TX\beta = X^TY$$

This gives the solution for the least squares estimator

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

Note: $(X^TX)^{-1}$ exists if

- $n \geq p+1$
- The predictors must be linear independent, i.e. No columns of X is a linear combination of the other columns. (To check column rank of $X = k+1$)



at least you can have a model (fit a regression)

Fitted value The fitted value of the regression model corresponding to $x_i = (1, x_{i1}, x_{i2} \cdots x_{ik})$ is

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_k x_{ik}, \quad i = 1 \cdots n$$

In the matrix format, it is $\hat{Y} = X\hat{\beta} = X(X^TX)^{-1}X^TY$

• new

Statistical Properties of Least-squares Estimators

For estimator $\hat{\beta}$, if X is non-random, then

$$E(\hat{\beta}) = E(\hat{\beta}|X) = \beta$$

$$Var(\hat{\beta}|X) = \sigma^2(X^TX)^{-1} = C \cdot \sigma^2$$

$C_{(k+1)\times(k+1)}$

Estimation of σ^2 - Estimate the variation of the data

The residual sum of squares for MLR is

$$\begin{aligned} SS_{Res} &= (Y - \hat{Y})^T(Y - \hat{Y}) \\ &= (Y - X\hat{\beta})^T(Y - X\hat{\beta}) \\ &= \|Y - X\hat{\beta}\|_2^2 \end{aligned}$$

For the estimator of σ^2, we have

$$\hat{\sigma}^2 = MS_{Res} = \frac{SS_{Res}}{n-p} = \frac{1}{n-p}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

$$= \frac{\|Y - X\hat{\beta}\|_2^2}{n-p}$$

In R, residual standard error $\hat{\sigma} = \sqrt{\hat{\sigma}^2}$

We can show that $\hat{\sigma}^2$ is an unbiased estimator.

$$E(\hat{\sigma}^2 | X) = \sigma^2$$

Student's t-test

Requires the additional GN-SLR assumptions,
We further assume that

$$\begin{cases} \varepsilon \sim N(0_n, \sigma^2 I_n) \\ \varepsilon \text{ is independent of } X \end{cases}$$

From these assumptions conditional on X,
Y has a M.V (multivariate) Gaussian distribution

$$Y|X \sim N(X\beta, \sigma^2 I_n)$$

We can show that, $\hat{\beta}$ has a sample distribution

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

• unknown

$$\hat{\beta}_j \sim N(\beta_j, [\sigma^2 (X^T X)^{-1}]_{(j+1)(j+1)}) \quad j = 0 \dots k$$

diagonal entry

To test $\begin{cases} H_0: \beta_j = 0 \\ H_1: \beta_j \neq 0 \end{cases}$ where $j = 0 \dots k$

Equivalently, $\begin{cases} H_0: Y = \beta_0 + \beta_1 X_1 + \cdots \beta_{j-1} X_{j-1} + \beta_{j+1} X_{j+1} \cdots + \beta_k X_k + \varepsilon \\ H_1: Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + \varepsilon \end{cases}$

If H_0 is rejected, it indicates that the predictor X_j is likely to be meaningful addition to the model

We use the statistic $T_j = \dfrac{\hat{\beta}_j - 0}{ese(\hat{\beta}_j)}$

where $ese(\hat{\beta}_j) = \sqrt{c_{jj}} = \sqrt{\hat{\sigma}^2 [(X^T X)^{-1}]_{(j+1)(j+1)}}$

• $MS_{Res} = \frac{SS_{Res}}{n-p}$

If under the GN-SLR, $\varepsilon \sim N(0, \sigma^2 I_n)$, and under the $H_0: \beta_j = 0$

$$T_j \sim t_{n-p}$$

We reject H_0 if

$$|T_j| > k \equiv t_{\frac{\alpha}{2}, n-p}$$

or use the p-value

$$P(T > |T_j|) \equiv p\text{-value} < \alpha.$$

F-test

Consider the regression $Y = X\beta + \varepsilon = \underbrace{[X_{(1)}, X_{(2)}]}_{X} \begin{bmatrix} \beta_{(1)} \\ \beta_{(2)} \end{bmatrix} \begin{matrix} \} p-r \\ \} r \end{matrix} + \varepsilon$

$= X_{(1)}\beta_{(1)} + X_{(2)}\beta_{(2)} + \varepsilon$

$n \times (p-r)$ $(p-r) \times 1$ $n \times r$ $r \times 1$

We would like to test

$H_0: \beta_{(2)} = 0_{r \times 1}$
$H_1: \beta_{(2)} \neq 0_{r \times 1}$

Equivalently,

$H_0: Y = X_{(1)}\beta_{(1)} + \varepsilon$ (reduced model)
$H_1: Y = X_{(1)}\beta_{(1)} + X_{(2)}\beta_{(2)} + \varepsilon$ (full model)

Then the null hypothesis $\beta_{(2)} = 0$ can be tested by the statistic

$$F_0 = \frac{\overline{SS_R}(\beta_{(2)} \mid \beta_{(1)})/r}{\overline{SS_{Res}}/(n-p)} = \frac{\overline{SS_R}(\beta_{(2)} \mid \beta_{(1)})/r}{MS_{Res}}$$ degree of freedom

- $\overline{SS_R}(\beta_{(2)} \mid \beta_{(1)}) = \overline{SS_R}(\beta) - \overline{SS_R}(\beta_{(1)})$ ($\overline{SS_R}(\beta)$: $df = p$; $\overline{SS_R}(\beta_{(1)})$: $df = p-r$; $\overline{SS_R}(\beta_{(2)} \mid \beta_{(1)})$: $df = r$)

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

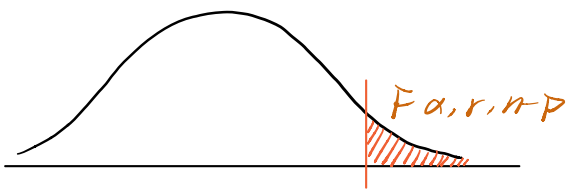
$\hat{\beta} = (X^TX)^{-1}X^TY$ $\quad \hat{\beta}_{(1)} = (X_{(1)}^T X_{(1)})^{-1} X_{(1)}^T Y$

Extra sums-of-squares due to $\beta_{(2)}$ when $\beta_{(1)}$ is in the model.
i.e. Extra contribution in $\overline{SS_R}$ due to the predictors in $X_{(2)}$ under $H_0: \beta_{(2)} = 0$.

- $MS_{Res} = \frac{SS_{Res}}{n-p} = \frac{Y^TY - \hat{\beta}X^TY}{n-p}$

Under $H_0: \beta_{(2)} = 0$, under GN-SLR, the F_0 follows F distribution with $df_1 = r$, $df_2 = n-p$, we reject H_0 if

$F_0 > F_{\alpha, r, n-p}$



left threshold of p-value

Sequential F-test

- Test 1

  $H_0: Y = \beta_0 + \varepsilon$

  $H_1: Y = \beta_0 + \beta_1 X_1 + \varepsilon$

- Test 2

  $H_0: Y = \beta_0 + \beta_1 X_1 + \varepsilon$

  $H_1: Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$

  drawback: Order matters!

- Test 3

  $H_0: Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$

  $H_1: Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$

To compute F-statistic, we need to compute $\overline{SSR}(\beta_1|\beta_0)$ for test 1, $\overline{SSR}(\beta_2|\beta_1,\beta_0)$ for test 2 and $\overline{SSR}(\beta_3|\beta_2,\beta_1,\beta_0)$ for test 3

$$F_1 = \frac{\overline{SSR}(\beta_1|\beta_0)/1}{SSRes(\beta_3,\beta_2,\beta_1,\beta_0)/(n-4)}$$ ending full model

↳ β_2, β_3 will be considered later.

$$F_1 = \frac{\overline{SSR}(\beta_1|\beta_0)/1}{SSRes(\beta_1,\beta_0)/(n-2)}$$ ⇒ results are equivalent when $n \to \infty$

$$F_2 = \frac{\overline{SSR}(\beta_2|\beta_1,\beta_0)/1}{SSRes(\beta_0,\beta_1,\beta_2,\beta_3)/(n-4)}$$

$$F_3 = \frac{\overline{SSR}(\beta_3|\beta_2,\beta_1,\beta_0)/1}{SSRes(\beta_0,\beta_1,\beta_2,\beta_3)/(n-4)}$$

Note: The denominators are always the same.

$$\overline{SSR}(\beta_3,\beta_2,\beta_1|\beta_0)$$

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

$$+ \underbrace{\frac{\overline{SSR}(\beta_3,\beta_2,\beta_1,\beta_0) - \overline{SSR}(\beta_2,\beta_1,\beta_0)}{\overline{SSR}(\beta_3|\beta_2,\beta_1,\beta_0)}}$$

$$= \overline{SSR}(\beta_1|\beta_0) + \overline{SSR}(\beta_2|\beta_1,\beta_0) + \overline{SSR}(\beta_3|\beta_2,\beta_1,\beta_0)$$

$$= \overline{SSR}(\beta_3,\beta_2,\beta_1,\beta_0) - \overline{SSR}(\beta_0)$$

Some technical details

Originally, we have the sum of squares decomposition

$$SS_T = SS_{Res} + SS_R$$

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

Now, we consider a different sum of squares.

$$\overline{SS_T} = SS_{Res} + \overline{SS_R}$$

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

Since $\sum_{i=1}^{n} y_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i + \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \sum_{i=1}^{n} \hat{y}_i^2 + 2\underbrace{\sum_{i=1}^{n} (y_i - \hat{y}_i)\hat{y}_i}_{0}$

Therefore,

$$\overline{SS_R} = \sum_{i=1}^{n} y_i^2$$

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

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

$$= \hat{\beta} X^T Y \qquad df = p$$

$1 \times p \quad p \times 1$

from normal equation $\underbrace{\sum_{i=1}^{n} (y_i - x_i\hat{\beta}) x_i \hat{\beta}}_{0}$

Confidence Interval

① Confidence interval for estimates

The $100(1-\alpha)\%$ confidence interval for the regression coefficients β_i, $i = 0, 1, 2 \cdots k$

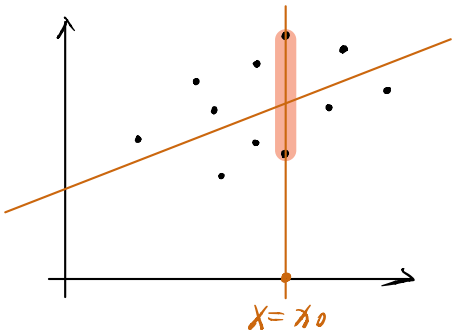
$$\hat{\beta}_i \pm k\sqrt{\hat{\sigma}^2 \cdot C_{ii}} \qquad k = t_{\frac{\alpha}{2}, n-p} \qquad C = (X^T X)^{-1}$$

$\sqrt{\hat{\sigma}^2 \cdot C_{ii}}$: $\text{ese}(\hat{\beta}_i)$

for $i = 0, 1 \cdots k$

Note: The confidence interval will change using different batch of data.

② Confidence interval for mean response



$m(x) = X\beta$ MLR

$= \beta_0 + \beta_1 x_1$ SLR

$x_0 = [1, x_{01}, x_{02} \cdots x_{0k}]$

The fitted value at this point is

$$\hat{m}(x_0) = x_0\hat{\beta} = [1, x_0 \cdots x_{0k}] \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}$$

This is the unbiased estimator of $E[Y \mid X = x_0]$, since

$$E[\hat{m}(x_0)] = x_0 E[\hat{\beta}] = x_0 \beta$$

and variance is

$$\text{Var}[\hat{m}(x_0)] = \sigma^2 x_0 (X^T X)^{-1} x_0^T.$$

Therefore, a $100(1-\alpha)\%$ confidence interval for the mean response at $X = x_0 = [1, x_{01}, x_{02} \cdots x_{0k}]$

$$CI(m(x_0)) = [\hat{m}(x_0) \pm k \cdot \sqrt{\hat{\sigma}^2 \cdot x_0 (X^T X)^{-1} x_0^T}]$$

② Prediction interval of new observation

A $100(1-\alpha)\%$ prediction interval for the future observation predicated at $X = x_0 = [1, x_{01}, x_{02} \cdots x_{0k}]$

$$PI(Y_0) = [\hat{m}(x_0) \pm k \cdot \sqrt{\hat{\sigma}^2 (1 + x_0 (X^T X)^{-1} x_0^T)}]$$

	MLR	GN-MLR
estimator $\hat{\beta}, \hat{\sigma}^2$	✓	
unbiaseness of variance, E and variance of $\hat{\beta}, \hat{\sigma}^2$	✓	
t-test		✓
F-test		✓
$CI(\beta_i)$		✓
$CI(m(x_0))$		✓
$PI(Y_0)$		✓

Model Adequacy Checking

The residuals are

$$e_i = y_i - \hat{y}_i = y_i - x_i \cdot \hat{\beta} \qquad i = 1 \cdots n$$

written in the matrix for $\quad e = (e_1, e_2 \cdots e_n)^T$

$$e = y - \hat{y} = y - X\hat{\beta} = y - \underbrace{X(X^TX)^{-1}X^T}_{H} y$$

$$= (I_n - H)y.$$

We have the following properties of e. MLR

① The distribution of e should have a center around 0.

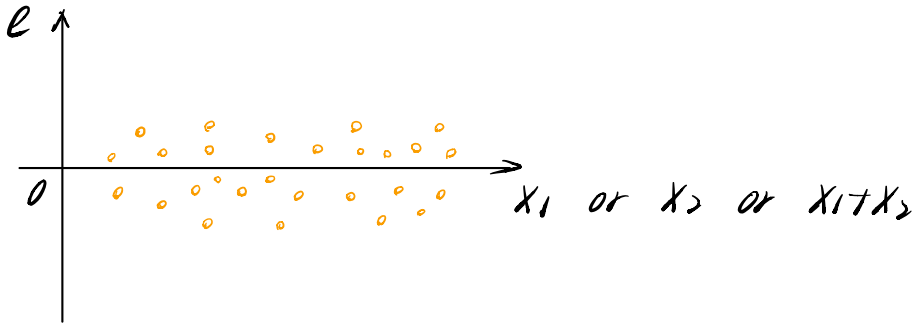
$$E[e|X] = 0_n$$

Since $E[e|X] = E[y - \hat{y}|X] = E[y|X] - E[\hat{y}|X]$

$$= X\beta - \underline{E[X\hat{\beta}|X]}$$

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

Plot e against any predictor or the linear combination of any predictor. This plot should have a constant mean equal to zero.



② The covariance between $\hat{y}$ and e is zero.

$$Cov(e, \hat{y}) = 0_{n \times n}$$

$$Cov(e, \hat{y}) = Cov((I_n - H)y, Hy)$$

$$= \sigma^2 \underline{H(I_n - H)}$$

$$= 0 \qquad H = HH \text{ idempotent.}$$

Plot e against $\hat{y}$, the plot should be patternless.

③ The variance of residual e is $\text{Var}(e|X) = \sigma^2(I_n - H)$

since $\text{Var}(e|X) = \text{Var}((I_n - H)Y|X)$

$= (I_n - H)^T (I_n - H)\,\text{Var}(Y|X)$

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

This implies that $e_1, e_2 \cdots e_n$ are correlated, not like $\varepsilon_1, \cdots \varepsilon_n$ which are uncorrelated.

In the scalar form

$$\text{Var}(e_i|X) = \sigma^2(1 - h_{ii}) \quad i = 1 \cdots n$$

Usually there is no outlier in the data, the h_{ii} will be small. The variance of e should be roughly constant.

Plot e against X_j, the plot should have constant variance.

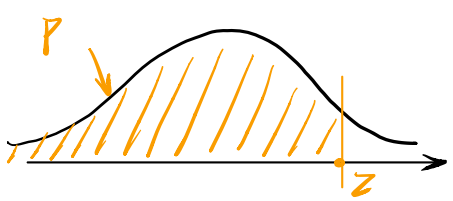
④ only for GN-MLR

If $Y|X$ is Gaussian, then $e = \begin{pmatrix} e_1 \\ \vdots \\ e_n \end{pmatrix}$ is also Gaussian with mean zero and variance $\sigma^2 I_n$. So each term $e_i \sim N(0, \sigma^2)$ for $i \cdots n$

- You can make a histogram of the distribution of the residuals, and compare that to a Gaussian
- quantile-quantile plot (qq plot)

For a continuous distribution, its CDF $F(z) = P(Z < z)$ has an inverse function.

$$F^{-1}(p) = z \quad \text{s.t.} \quad p = P(Z \le z)$$



If the distribution is $N(\mu, \sigma^2)$

$$F^{-1}(p) = \sigma \Phi^{-1}(p) + \mu$$

Φ^{-1}: inverse of CDF function for $N(0,1)$

If we plot $F^{-1}(p)$ against $\Phi^{-1}(p)$, we'll get a straight line.

But in practice, F or F^{-1} is unknown, but we can estimate quantile using data.

We have n residuals $e_1, e_2 \cdots e_n$
Put them in an increasing order $e_{(1)}, e_{(2)} \cdots e_{(n)}$, where $e_{(i)}$ is greater to a fraction $\frac{i}{n}$ or the residuals.

e.g. $e_{(2)} \geq e_{(1)}, e_{(1)}$ ($\frac{2}{n}$ of n obs.)

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

Now make a plot $\hat{F}^{-1}(\frac{i}{n}) = e_{(i)} \sim \Phi^{-1}(\frac{i}{n})$ to approximate $F(\frac{i}{n}) \sim \Phi^{-1}(\frac{i}{n})$ and hope to obtain a straight line if the distribution of e_i's indeed Gaussian.

Transformation

Variance-stability transformation

We should consider the use of transformation when constant variance assumption is violated.

For example, $Y \sim Poi(\lambda)$. Then $E(Y) = Var(Y) = \lambda$.
Then if we model $E(Y|X=x) = X\beta = Var(Y|X=x)$
↑ not constant!

The mean of Y is related to X. To fix this, we can regress $Y' = \sqrt{Y}$ against X. Since the variance of $\sqrt{Y}$ is independent of the mean.

Variance-mean relationship	Transformation.
$\sigma^2 \propto$ constant	$Y' = Y$
$\sigma^2 \propto E(Y)$	$Y' = \sqrt{Y}$ (Poisson)
$\sigma^2 \propto E(Y)(1-E(Y))$	$Y' = \sin^{-1}(Y)$ (binomial)
$\sigma^2 \propto E(Y)^2$	$Y' = \ln(Y)$

Transformations to linearity

If we detect non-linearity using scatterplot or residual plots. In some case, a non-linear function can be linearized using a suitable transformation.

e.g. The Cobb-Douglas production function for observed economic data $i = 1 \dots n$

$$O_i = e^{\beta_0} L_i^{\beta_1} C_i^{\beta_2} u_i$$

O_i: total production output; L_i: labor input; C_i: capital input; u_i: random error term

Take natural log, we have that

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

$Y_i = \ln(O_i)$, $X_{i1} = \ln(L_i)$, $X_{i2} = \ln(C_i)$, ε_i: random error term

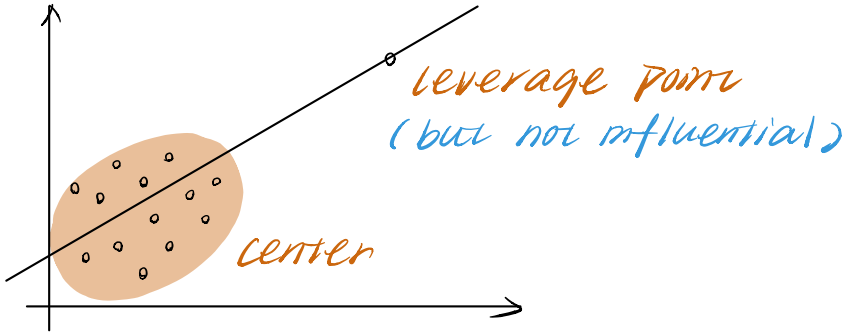
The transformation requires $\varepsilon = \ln(u) \sim N(0, \sigma^2)$

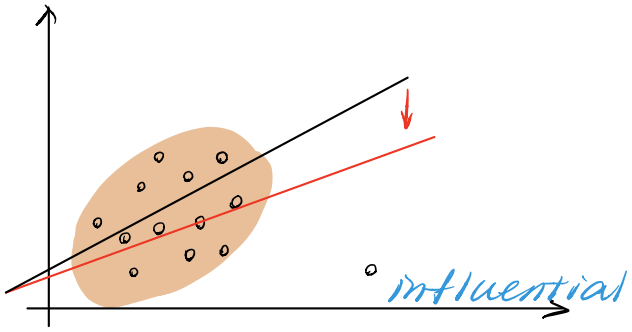
e.g. $Y = \beta_0 + \beta_1 \left(\frac{1}{X}\right) + \varepsilon$ can be linearized

$$X' = \frac{1}{X}, \quad Y = \beta_0 + \beta_1 X' + \varepsilon$$

Original function	Transformation	Linear form
$y = \beta_0 X^{\beta_1}$	$Y' = \log Y$, $X' = \log X$	$Y' = \log \beta_0 + \beta_1 X'$
$y = \beta_0 e^{\beta_1 X}$	$Y' = \log Y$	$Y' = \log \beta_0 + \beta_1 X$
$y = \beta_0 + \beta_1 \log X$	$X' = \log X$	$Y' = \beta_0 + \beta_1 X'$
$y = \frac{X}{\beta_0 X - \beta_1}$	$Y' = \frac{1}{Y}$ $X' = \frac{1}{X}$	$Y' = \beta_0 - \beta_1 X'$

Leverage and Influence Point





Recall that $\hat{Y} = HY$, $H = X(X^TX)^{-1}X^T$

$$Var(\hat{Y}|X) = \sigma^2 H$$

$$\Rightarrow Var(Y - \hat{Y}|X) = Var(e|H) = \sigma^2(I_n - H)$$

The element of H dictate how much each data point $Y_1 \dots Y_n$ affects the prediction $\hat{Y}_1 \dots \hat{Y}_n$. i.e. $\hat{Y}_i = \sum_{j=1}^{n} h_{ij} Y_j$

h_{ij} can be interpreted as the amount of leverage exerted by the jth observation Y_j on the ith fitted value $\hat{Y}_i$.

We focus on the diagonal element h_{ii} of H

$$h_{ii} = x_i (X^TX)^{-1} x_i^T$$

x_i: ith row of the matrix X.

If h_{ii} is large, then the ith observation has high leverage.

Influence (Cook's distance)

Each data point may have a different "influence" on the estimation of β

$$D_i = \frac{(\hat{\beta}_{(i)} - \hat{\beta})^T M (\hat{\beta}_{(i)} - \hat{\beta})}{c}$$

$M = X^TX$

c: PMSRes

$\hat{\beta}_{(i)}$ is the fitted estimate with ith observation removed.

We can also rewrite D_i as

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

(p = # predictors including intercept)

($\frac{e_i^2}{\hat{\sigma}^2}$: (standard residual)^2; $\frac{h_{ii}}{(1-h_{ii})^2}$: leverage)

$\Rightarrow$ D_i is affected by { residual $e_i \uparrow$ $D_i \uparrow$; leverage $h_{ii} \uparrow$ $D_i \uparrow$

D_i combines the residual for the ith observation and the location of that point to measure the influence of this point. A point with high D_i, potentially would be outlier.

There is another formula for D_i

Because $X\hat{\beta}_{(i)} - X\hat{\beta} = \hat{Y}_{(i)} - \hat{Y}$

$$D_i = \frac{(\hat{Y}_{(i)} - \hat{Y})^T (\hat{Y}_{(i)} - \hat{Y})}{p\,MSRes}$$

more robust

You can interpret it as the squared distance that the fitted value moves when the ith observation is removed.

$$\left\langle \min_{\beta} \|Y - X\beta\|_2 \right\rangle$$

Special Types of predictors in MLR

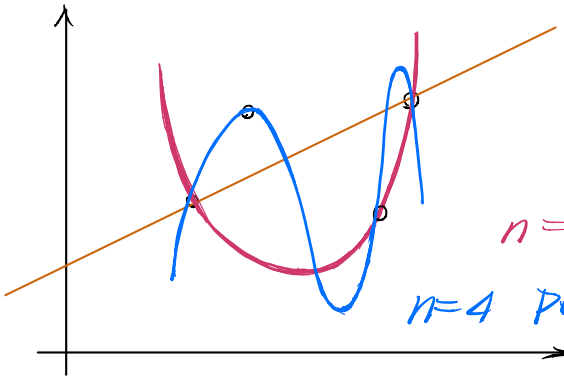
Polynomial Term

Suppose the model is

$$E[Y|X] = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \cdots + \beta_k X_1^k$$

$$= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$$

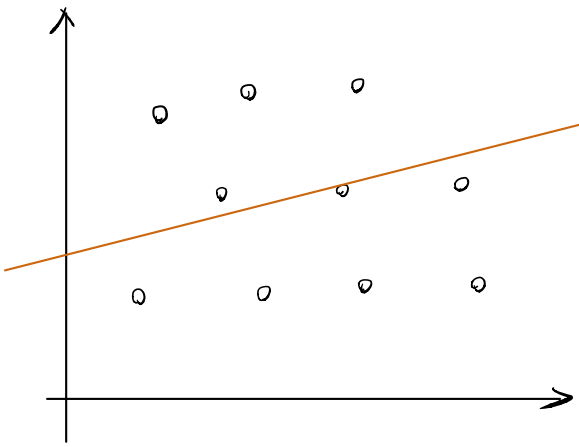
$k=3$ cubic function



$n=2$ perfect fit
$E[Y|X_1] = \beta_0 + \beta_1 X_1$

$n=3$ perfect fit $E[Y|X_1] = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2$

$n=4$ perfect fit $E[Y|X_1] = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \beta_3 X_1^3$



$k=1$, oversimplification

Data: $Y = \sin(X_1) + \varepsilon$

$k=3$ would be a good fit.

fig 1.7

Penalization

If we fit a polynomial regression with order k.

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_1^k + \varepsilon$$

or

$$E[Y|X_1] = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_1^k$$

Given n observations of data $\{y_i, x_{i1}\}_{i=k}^{n}$. You can estimate the value of $\beta_0, \beta_1 \cdots \beta_k$ by least squares. Denote $\beta = (\beta_0 \cdots \beta_k)^T$

$$\min_{\beta} \sum_{i=1}^{n} \underbrace{(y_i - (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{i1}^k))^2}_{\text{residual}}$$

If k is large, the model becomes too complicated or "overfitted" The corresponding fitted curve becomes wiggle.

To fix this issue, we restrict the size of the coefficients β by using penalization, i.e.

$$\min_{\beta} \sum_{i=1}^{n} \underbrace{(y_i - (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{i1}^k))^2}_{\text{residual}}$$

such that $\|\beta'\|_2^2 \le t$ where

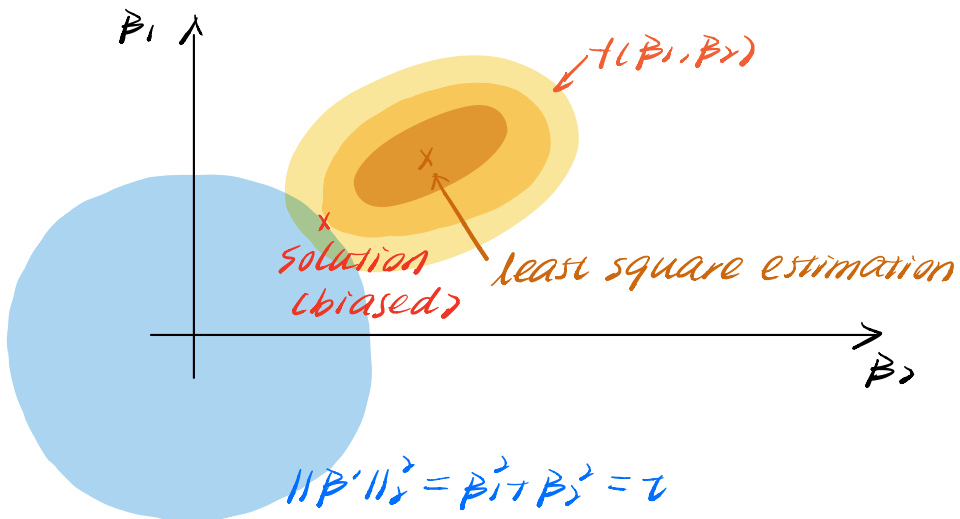
$$\|\beta\|_2^2 = \sqrt{\beta_1^2 + \beta_2^2 + \cdots + \beta_k^2}$$ usually β_0 is not penalized

e.g. Consider the model with only $\beta_0, \beta_1, \beta_2$

$$\min_{\beta_0, \beta_1, \beta_2} \sum_{i=0}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i1}^2)^2 = c$$

$f(\beta_1, \beta_2)$

s.t. $\|\beta'\| \le t$, $\beta' = (\beta_1, \beta_2)^T$



Mathematically, it is difficult to solve. Consider an alternative form.

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

$\lambda \geq 0$

For demonstration purpose. Consider a no-intercept model.

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

$$= \min_{\beta} \underbrace{\|y - X\beta\|^2}_{f(\beta)} + \lambda \|\beta'\|_2^2 \quad \text{ridge regression}$$

where $y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$ $\quad X = \begin{pmatrix} x_{11} & \cdots & x_{11}^{k} \\ \vdots & & \vdots \\ x_{n1} & \cdots & x_{n1}^{k} \end{pmatrix}$

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

We solve β

$$y^T X - X^T X \beta = \lambda \beta$$

$$(X^T X + \lambda I_p)\beta = y^T X$$

$$\Rightarrow \quad \hat{\beta} = (X^T X + \lambda I_p)^{-1} y^T X$$

Usage of ridge regression

① control model complexity

② deal with high collinarity between predictors.

y	x_1	x_2	x_6
	1	2	
	2	4	
	3	6	
	4	8	

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

becomes singular when two columns of x_j, x_i are perfectly correlated.

③ When $n \ll p$, n is the number of obvertions, p is the number of predictors.

$$n\left\{\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}\right. \underbrace{\begin{pmatrix} 1 & x_{11} & \cdots & x_{1k} \\ \vdots & & & \\ 1 & x_{n1} & \cdots & x_{nk} \end{pmatrix}}_{p}$$

MLR could fail in this case, but you should try ridge regression.

④ SVM with overlapped case (linear model) ≈ logistic regression with ridge penalization.

$$\text{SVM: } \sum_{i=1}^{n} (1 - y_i(\beta_0 + \beta^T x_i)) + \lambda \|\beta\|_2$$

$$\text{Logistic: } \sum_{i=1}^{n} \log(1 + e^{-y_i(\beta_0 + \beta^T x_i)}) + \lambda \|\beta\|_2$$