Departure from underlying assumptions
\[ Y = X\beta + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2 I_n) \]

Underfitting

true: \[ Y = X\beta + Z\gamma + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2 I_n), \quad \gamma = 0 \]
fit: \[ Y \sim X\beta \]

\[ \hat{\beta} = (X^T X)^{-1} X^T Y \]
\[ E[\hat{\beta}] = (X^T X)^{-1} X^T (X\beta + Z\gamma) = \beta + (X^T X)^{-1} X^T Z\gamma \]
\( \hat{\beta} \) is biased unless \( X^T Z = 0 \). In this situation, the columns of \( X \) and \( Z \) are orthogonal.

\[ \text{e.g. balanced ANOVA} \]
\[ \text{Var}(\hat{\beta}) = \sigma^2 (X^T X)^{-1} \]

Let \( \hat{\beta}_c \) be the OLSE from \( Y \sim X\beta + Z\gamma \)

let \( W = (X, Z) \)
\[ (W^T W)^{-1} = \begin{pmatrix} (X^T X)^{-1} + LML^T & -LM \\ -ML^T & M \end{pmatrix} \]

where \( M = (Z^T (I - P_X) Z)^{-1} \), \( L = (X^T X)^{-1} X^T Z \)
\[ \text{Var}(\hat{\beta}_c) = \sigma^2 (X^T X)^{-1} + LML^T \]
\[ \text{Var}(\hat{\beta}_c) - \text{Var}(\hat{\beta}) = \sigma^2 LML^T \geq 0 \]

\[ \Rightarrow \text{Var}(\hat{\beta}) \leq \text{Var}(\hat{\beta}_c) \]

Underfitting leads to biased estimate but reduced variance.
\[ \hat{\sigma}^2 = \frac{\text{tr}((I-P_x)Y^T)(I-P_x)Y}{n-p} \]

\[ (n-p)E[\hat{\sigma}^2] = \text{tr}((I-P_x)\text{Var}(Y)) + [E(Y)]^T(I-P_x)E[Y] \]

\[ = \text{tr} (\sigma^2 (I-P_x)) + (X\beta + Z\gamma)^T(I-P_x)(X\beta + Z\gamma) \]

\[ = \sigma^2(n-p) + \gamma^T(I-P_x)Z\gamma \]

\[ E[\hat{\sigma}^2] = \sigma^2 + \frac{\gamma^T(I-P_x)Z\gamma}{n-p} \geq \sigma^2 \]

Underfitting leads to biased (upward) estimate of \( \sigma^2 \).

How about prediction?

Consider \( W_0 = (x_0^T, z_0^T)^T \)

If we use the correct model

\[ \hat{p}_{oc} = x_0^T \hat{\beta}_c + z_0^T \hat{\gamma}_c \]

\[ E[\hat{p}_{oc}] = x_0^T E[\hat{\beta}_c] + z_0^T E[\hat{\gamma}_c] = x_0^T \beta + z_0^T \gamma \]

\[ \text{Var}(\hat{p}_{oc}) = (x_0^T - z_0^T)(W_0W_0^T)^{-1}(x_0^T - z_0^T)\sigma^2 \]

\[ = \sigma^2\left[ x_0^T(x_0^T)^{-1}x_0 + (Lx_0 - z_0)^T M (Lx_0 - z_0) \right] \]

\[ \geq \sigma^2 x_0^T(x_0^T)^{-1}x_0 \]

\[ M = (Z^T(I-P_x)Z)^{-1} \]

Underfitting \( \hat{p}_0 = x_0^T \hat{\beta} \)

\[ E[\hat{p}_0] = x_0^T E[\hat{\beta}] = x_0^T (\beta + (X^T X)^{-1}X^T Z \gamma) \]

\[ = x_0^T \beta + x_0^T (X^T X)^{-1}X^T Z \gamma \]

\[ \text{Var}(\hat{p}_0) = x_0^T (X^T X)^{-1}x_0 \sigma^2 \leq \text{Var}(\hat{p}_{oc}) \]

Again, biased but reduced variance.
Overfitting

true \( Y = X\beta + \varepsilon \) \( \varepsilon \sim \mathcal{N}(0, \sigma^2 I) \)

fit \( Y = \tilde{X} \tilde{\beta} + \tilde{\varepsilon} \)

\( \hat{\beta} = (X^T \hat{X})^{-1} X^T Y - L \tilde{\beta} = \bar{\beta} - \hat{\beta} \)

\( \hat{\gamma} = [Z^T (I-P_x) Z]^{-1} Z^T (I-P_x) Y \)

\( \mathbb{E} [\hat{\gamma}] = [Z^T (I-P_x) Z]^{-1} Z^T (I-P_x) X \beta = 0 \)

\( \mathbb{E} [\hat{\beta}] = \mathbb{E} [\hat{\beta}_o] - \mathbb{E} [\hat{\gamma}] = \beta - 0 = \beta \)

So overfitting gives unbiased estimates. What is the cost of overfitting?

\( \text{Var}(\hat{\beta}) = \sigma^2 (X^T X)^{-1} = \sigma^2 (X^T X)^{-1} = \text{Var}(\bar{\beta}) \)

\( \sigma^2 \)

\( \hat{\sigma} = \frac{X^T (I-P_x) Y}{n-p-q} \)

\( \mathbb{E}[\hat{\sigma}]^2 = \frac{1}{n-p-q} \left[ \frac{\text{tr}(I-P_x) \text{Var} \sigma^2 I_n}{} + (x\beta)^T (I-P_x) X \beta \right] \)

\( = \frac{1}{n-p-q} \left[ \sigma^2 (n-p-q) + 0 \right] = \sigma^2 \)

\( \hat{\sigma} = \frac{X^T (I-P_x) Y}{n-p-q} \)

\( X^T (I-P_x) X = X^T - X^T P_x X = X^T - X^T (X, (I-P_x) Z) (X^T)^{-1} (X^T) X \)

\( = X^T - X^T (P_x + (I-P_x) Z [Z^T (I-P_x) Z]^{-1} Z^T (I-P_x) X) \)

\( = X^T - X^T P_x X - O = O \)
Prediction
\[ \hat{\gamma}_0 = \gamma_0^* \hat{\beta} + \epsilon_0 \]
\[ E[\hat{\gamma}_0] = \gamma_0^* E[\hat{\beta}] + \epsilon_0 E[\hat{\gamma}] = \gamma_0^* \beta \]
\[ \text{Var}[\hat{\gamma}_0] = \text{Var}(\gamma_0^* \hat{\beta} + \epsilon_0 \hat{\gamma}) = \text{Var}(\gamma_0^* \hat{\beta} + (\epsilon_0 - \gamma_0^* \hat{\gamma})) \]
\[ \text{Var}(\gamma_0^* \hat{\beta}) + \text{Var}(\epsilon_0 - \gamma_0^* \hat{\gamma}) \]
\[ \text{Var}(\gamma_0^* \hat{\beta}) = \text{Var}(\hat{\gamma}_0) \]
unbiased but increased variance

Incorrect Variance Matrix
\[ \text{Var}(\epsilon) \neq \sigma^2 I_n \]
If the true variance is \( \text{Var}(\epsilon) = \sigma^2 V \)
\[ \text{If } \epsilon \text{ is independent, } \text{Var}(\epsilon) = 0 \]
Then \[ E[\hat{\beta}] = E[\gamma^* (X'X)^{-1} X'Y] = (X'X)^{-1} X'Y \beta = \beta \]
\[ \text{Var}(\hat{\beta}) = (X'X)^{-1} X' \sigma^2 V X (X'X)^{-1} = \sigma^2 (X'X)^{-1} \text{ in general} \]
\[ E[\hat{\sigma}^2] = \frac{1}{n-p} E[\epsilon' (I - P_{X,n}) \epsilon] = \frac{1}{n-p} \left[ \text{tr} [(I - P_{X,n}) \sigma^2 + (X'X)^{-1} \epsilon' \epsilon] \right] \]
\[ = \sigma^2 \frac{\text{tr} [(I - P_{X,n})]}{n-p} + \sigma^2 \text{ in general} \]

If \( \text{Var}(\epsilon_i) = w(u_i) \), where \( u_i = E[\epsilon_i] \). Consider transformation \( f(\cdot) \)
By Taylor's expansion to the first order \[ f(\gamma) \approx f(u) + (\gamma - u)f'(u) \]
we have \[ \text{Var}[f(\gamma)] \approx [f'(u)]^2 \text{Var}(\gamma) = [f'(u)]^2 w(u) \]
To stabilize variance, one needs to find $f$ such that $[f(x)]^2w(u)$ is a constant.

\[
f(x) = \int \frac{du}{\sqrt{w(u)}}
\]

E.g. $w(u) = u$

\[
f(u) = \int \frac{du}{u} = \ln u
\]

**Outliers**

There are different types of outliers:

1. Error outlier / regression outlier: a point with large $e_i$
2. Leverage point / $X$-outlier: a point with large $|x - \bar{x}|$
3. Influential points: a point that affects regression coefficient

Leverage

The hat matrix $H = P_x = X(X^TX)^{-1}X^T$ (X is full rank)

Properties of the hat matrix

- $H$ is a projection matrix, $\text{rank}(H) = \text{tr}(H) = p$
- $h_{ij} = x_i^T(x^Tx)^{-1}x_j$, where $x_i^T$ is the $i$th row of $X$
\( H^2 = H \Rightarrow h_{ii} = \frac{n}{\sum_{j=1}^{n} h_{ij}^2} = h_{ii}^2 + \sum_{j \neq i} h_{ij}^2 \Rightarrow h_{ii} \in (0, 1) \)

If \( h_{ii} = 1 \), \( h_{ij} = 0 \) for \( j \neq i \)

If the intercept is included in the model
\[
h_{ii} = \frac{1}{n} + \frac{M_D}{n-1}
\]
where \( M_D = (X_i - \bar{X})^T S^{-1} (X_i - \bar{X}) \), \( S \) is the sample variance of \( X \).

For the simple linear regression
\[
h_{ii} = \frac{1}{n} + \frac{(X_i - \bar{X})^2}{\sum_{j=1}^{n} (X_j - \bar{X})^2}
\]

The proof is in the next two pages.

\( h_{ii} \) measures the distance between the \( i \)th observation and the center of the observations.

\( \hat{y} = H \hat{y} \Rightarrow \hat{y}_i = \sum_{i=1}^{n} h_{ij} \hat{y}_j = h_{ii} \hat{y}_i + \sum_{j \neq i} h_{ij} \hat{y}_j \)

The larger (closer to 1) the \( h_{ii} \), the less the contribution of \( \hat{y}_j(j \neq i) \) on \( \hat{y}_i \) and \( h_{ii} \) measures to which extent \( \hat{y}_i \) is determined by \( \hat{y}_i \)

\( \text{Trace}(H) = p \Rightarrow \text{average } h_{ii} \text{ is } \frac{p}{n} \).

A rule of thumb for large leverage:
\( h_{ii} > 2 \frac{p}{n} \).
Start with the model with the intercept only

\( Y = 1 \beta_0 + \varepsilon \) \[ \text{OLS: } \hat{\beta}_0 = (X^T X)^{-1} X^T Y = \frac{1}{n} \bar{Y} \]

Update (**) by adding \( X \times p \): a projection matrix

\[ Y = 1 \beta_0 + X \beta + \varepsilon \]
\[ = 1 \beta_0 + 1 (\bar{X} X) \beta + \left( I - \frac{1}{n} \right) \beta + \varepsilon \]
\[ = 1 \left( \beta_0 + \frac{1}{n} \bar{X} \beta \right) + \left( I - \frac{1}{n} X \right) \beta + \varepsilon \]

\[ = \left( I - \frac{1}{n} X \right) \left( \begin{bmatrix} \bar{Y} \\ \beta \end{bmatrix} \right) + \varepsilon \]

\[ \text{OLS: } \begin{bmatrix} \hat{\beta} \\ \hat{\beta}_0 \end{bmatrix} = \left( W^T W \right)^{-1} W^T Y = \left( \begin{bmatrix} n & 0 \\ 0 & X^T \left( I - \frac{1}{n} X \right) \end{bmatrix} \right)^{-1} \left( \begin{bmatrix} 1^T \\ 0 \end{bmatrix} \right) \]
\[ = \left( \begin{bmatrix} 1^T \\ X^T \left( I - \frac{1}{n} X \right) \end{bmatrix} \right)^{-1} \left( \begin{bmatrix} 1^T \\ 0 \end{bmatrix} \right) \text{ where } R = I - \frac{1}{n} X \]

\[ \Rightarrow \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta} \end{bmatrix} = \left( I - \frac{1}{n} X X^T \right) \begin{bmatrix} \bar{Y} \\ \beta \end{bmatrix} = \left( \frac{1^T}{n} - \frac{1}{n} X (X^T X)^{-1} X^T \right) \left( \frac{1}{n} - \frac{1}{n} X (X^T X)^{-1} X^T \right) \]

\[ (1, X^T) \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta} \end{bmatrix} = HY = (1, X^T) \left( \begin{bmatrix} \frac{1^T}{n} - \frac{1}{n} X (X^T X)^{-1} X^T \right) \left( \frac{1}{n} - \frac{1}{n} X (X^T X)^{-1} X^T \right) \right) Y \]

\[ \Rightarrow H = \frac{1}{n} \left( 1^T - (1, X) (X^T X)^{-1} X^T \right) \]
\[ = \frac{1}{n} \left( 1^T - (1, X) (X^T X)^{-1} X^T \right) \]
\[ = \frac{1}{n} \left( 1^T - (1, X) (X^T X)^{-1} X^T \right) \]
\[
\frac{1}{n} \mathbf{X} \mathbf{X}^T \mathbf{X}^T \mathbf{R} \mathbf{X}
\]

\[
\frac{1}{n} \mathbf{X} = \begin{pmatrix} x_1 & \cdots & x_p \end{pmatrix} = \begin{pmatrix} \bar{x}_1 & \cdots & \bar{x}_p \\
\vdots & & \vdots \\
\bar{x}_1 & \cdots & \bar{x}_p \end{pmatrix}
\]

\[RX = (x_1 - \bar{x}_1, \ldots, x_p - \bar{x}_p) : \text{centered } \mathbf{X} \]

\[\mathbf{X}^T \mathbf{X} = (n-1)S, \text{ where } S \text{ is the sample covariance} \]

So \[H = \frac{1}{n} \mathbf{X}^T + (n-1) \mathbf{S} \mathbf{X}^T \mathbf{X} \]
Residuals

Assume \( Y = X\beta + \epsilon, \epsilon \sim N(0, \sigma^2 I_n), \text{rank}(X_{n \times p}) = p \)

\[
e = Y - \hat{Y} = Y - H'Y = (I - H)Y
\]

\[
E[e] = 0, \quad \text{Var}(e) = \sigma^2 (I - H)
\]

\[
s^2 = \hat{\sigma}^2 = \frac{e'e}{n - p} = \frac{RSS}{n - p}
\]

\[
\text{Var}(e_i) = \sigma^2 (1 - h_{ii}), \quad \text{Cov}(e_i, e_j) = \sigma^2 I - h_{ij} \sigma^2
\]

**Semistudentized residuals**

\[
e_i^* = \frac{e_i}{\sqrt{s^2}}
\]

**Studentized residuals (internally studentized residuals)**

\[
r_i = \frac{e_i}{\sqrt{s_i^2 (1 - h_{ii})}}
\]

The motivating of this definition is that different \( e_i \) has different variance. One can show that

\[
\frac{r_i^2}{n - p} \sim \text{Beta}(\frac{1}{2}, \frac{1}{2(n - p - 1)})
\]

So \( r_i \)'s are identically distributed.

**Studentized deleted residuals (externally studentized)**

\[
t_i = \frac{e_i}{\sqrt{s_{i,0}^2 (1 - h_{ii})}}
\]

where \( s_{i,0}^2 \) is calculated with the \( i \)th case removed from the data.

We have shown earlier that \( \hat{\beta} - \beta_{i,j} = (X'X)^{-1}X_i e_i \)

The difference is called DFBETA.
\[(n-p-1)S_{11}^2 = \sum_{j=1}^{n} (y_j - \hat{y}_j (\hat{\beta}))^2 = \sum_{j=1}^{n} (e_j + \frac{h_j e_j}{1-h_{jj}})^2 = \sum_{j=1}^{n} (e_j + \frac{h_j e_j}{1-h_{jj}})^2\]

\[= \sum_{j=1}^{n} (e_j + \frac{h_j e_j}{1-h_{jj}}) - \frac{e_j^2}{1-h_{jj}}\]

\[= \frac{1}{1-h_{jj}} [ (1-h_{jj})e_j^2 + e_j^2 \sum h_j^2 + 2(h_{jj}e_j \sum h_j e_j) - \frac{e_j^2}{1-h_{jj}} ]\]

\[= (n-p)S_{11}^2 + \frac{e_j^2 h_j}{1-h_{jj}} - \frac{e_j^2}{(1-h_{jj})^2} = h_{jj} - 0\]

\[H_0: e_j = 0 \quad \Rightarrow \frac{1}{1-h_{jj}} = h_{jj}\]

So \[t_1^2 = \frac{e_j^2}{S_{11}^2 (1-h_{jj})} = \frac{e_j^2 (n-p-1)}{(n-p)S_{11}^2 (1+h_{jj})} = \frac{h_{jj}^2 (n-p-1)}{n-p - r_i^2}\]

\[= \frac{B}{1-B} (n-p-1), \text{ where } B = \frac{r_i^2}{n-p} \sim \text{Beta}(\frac{1}{2}, \frac{1}{2}(n-p-1))\]

\[\sim F_{n-p-1, 1}\]

One can show that if \(B \sim \text{Beta}(\frac{1}{2}, \frac{1}{2})\), then

\[S_t \sim t_{n-p-1}\]

The \(t_i\)’s are identically distributed, but they are not independent.

\(t_i\) is considered large if \(|t_i| > 2\).

Here is another way of looking at \(t_i\). Consider model

\[y = x\beta + z\gamma + \epsilon\]

where \(z = (0, \ldots, 1 \ldots 0)\). Here the \(i\)th element of \(z\) is 1. Want to test \(H_0: \gamma = 0\). It can be shown that the \(F\)-statistic equals \(t_i^2\).
Leave-one-out methods for detecting influential points

DFBETA: \( \hat{\beta} - \hat{\beta}_{(i)} \)  change in \( \beta \)

DFFITS: \( x_i \hat{\beta} - x_i \hat{\beta}_{(i)} \)  change in fitted values

DFFITSS: standardized DFFITS, standardized by \( \frac{1}{\sqrt{\mathcal{S}(i)}} \)

COVRATIO: covariate ratio
\[
\frac{\det(\text{Var}(\hat{\beta}))}{\det(\text{Var}(\beta))} = \frac{\det(S_{ii})}{\det(s^2 (X^T X)^{-1})} \]

Cook’s D
Recall that a 100(1 - \alpha)\% confidence ellipse for \( \beta \) is
\[
\beta: (\beta - \hat{\beta})^T X X (\beta - \hat{\beta}) \leq \frac{ps^2}{\text{F}_{\frac{p}{n-p},1,0.9}}
\]
If the \( i \)th observation causes a substantial change in \( \beta \),
it is an influential point.
\[
D_i = \frac{(\hat{\beta}_{(i)} - \hat{\beta})^T X X (\hat{\beta}_{(i)} - \hat{\beta})}{ps^2}
\]
points with \( D_i > \text{F}_{\frac{p}{n-p},0.9} \) should be paid attention to.

Collinearity
- a problem unique to multiple regression
- not a violation of the assumptions, but is a serious problem
- occurs when one or several variables are linearly correlated with the other
- \( X^T X \) is almost singular
- doesn’t affect prediction seriously, but affects \( \text{Var}(\hat{\beta}) \)
- difficult to interpret \( \hat{\beta}_i \) and \( \hat{\beta}_j \) if \( x_i \) and \( x_j \) are linearly dependent
How to detect collinearity?

- Graphical methods: pairwise scatterplot. only provide information about pairs.
- Variance inflation factor (VIF): Let $R_j^2$ be the R-squared from regressing $X_j$ on the remaining predictors. The VIF for the $j$th predictor is defined by

$$VIF_j = \frac{1}{1-R_j^2}$$

$VIF > 10$ is considered large.

What can we do?
- remove redundant predictors
- ridge regression: replace $X^TX$ with $X^TX + \lambda I$, $\lambda > 0$, is small
- principal component analysis
- model selection

Normality

F-test is reasonably robust against nonnormality. Asymptotic theorem indicates that normality is not very important when $n$ is large.

Q-Q plot of residuals
- transformation
- there are robust methods for doing linear regression
Model Selection

We don't know the true model; the model should fit well. Parsimonious models are preferred. Trade off b/w underfitting and overfitting. Overfitting leads to large variance.

How to do model selection?

- Best subset regression: $2^p$ regression models
- Apply a sequence of hypothesis tests
  - Forward: start from no covariates, add one covariate at a time
  - Backward: start from the largest model, remove one covariate at a time
  - Stepwise: a forward selection followed by a backward elimination

In model selection, we want to select the subset of variables that leads to an optimal model based on a given standard:
- Goodness of fit
- Prediction error
- Number of nonzero coefficients
Goodness of fit (GOF)

- \( \text{RSS} = \| y - X\hat{\beta} \|^2 \) the smaller the better
- \( \frac{R^2}{\| y - \bar{y} \|^2} \) the larger the better (equivalent, leads to the largest model)

- To penalize the number of variables, define \( S^2 = \frac{RSS}{n-p} \) is unbiased for \( \sigma^2 \) the smaller the better

Adjusted \( R^2 \)

- \( R^2 = 1 - (1 - R^2) \frac{n}{n-p} \) the larger the better
- This method is equivalent to \( S^2 \)

\( (1 - R^2) \frac{n}{n-p} = \frac{RSS}{SSTO} \frac{n}{n-p} = S^2 \frac{n}{SSTO} \)

Prediction error

- Mallows' \( C_p \)
- training data \( Y = (Y_1, \ldots, Y_n)^T \), \( X = (X_1, \ldots, X_n)^T \), LSE: \( \hat{\beta} \)
- predict for \( n \) observations with \( X_0 = X \)

\[ \text{Prediction error} = E[Y] = E[Y_0] = \mu \]

\[ = E\|Y_0 - X\hat{\beta} - \mu\|^2 = E\|Y_0 - \mu + u - X\hat{\beta}\|^2 \]

\[ = E\|Y - u\|^2 + E\|u - X\hat{\beta}\|^2 + 2E(\text{tr}(X\hat{\beta}))^T E(\text{tr}(X\hat{\beta})) \]

\[ = n\sigma^2 + \|u - PY\|^2 + 2E(\text{tr}(X\hat{\beta}))^T E(\text{tr}(X\hat{\beta})) \]

\[ = n\sigma^2 + \|u - P\|^2 + E\|PE\|^2 \]

\[ = n\sigma^2 + u^T (I-P) u + \sigma^2 \|P\|_2 \]
\[ PE = u^T(I-p)u + pt^2 + n\sigma^2 \]

measures how well the model predicts the new responses.

\[ \frac{u^T(I-p)u + pt^2}{\sigma^2} = \frac{E[\text{RSS}_p]}{\sigma^2} + 2p + n \quad \text{as} \quad E[\text{RSS}] = E[u^T(I-p)u] = n^T(I-p)n + (n-p)\sigma^2 \]

\[ Cp = \frac{RSS_p}{\sigma^2} + 2p + n \]

It is common practice to estimate \( \sigma^2 \) using the largest model when the model fits well, \( u^T(I-p)u = 0 \), and \( E(C_p) \approx p \).

**Cross-Validation**

leave-one-out: predicted sum of squares \( \text{PRESS} = CV(1) = \frac{1}{n} \sum (y_i - x_i^T \hat{\beta}_{(i)})^2 \)

Analytical results exist for leave one out.

**AIC & BIC**

Suppose \( Y \sim N(X\beta, \sigma^2) \), \( X \) is full rank

\[ l(\beta, \sigma^2) = -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(y-x\beta)^T(y-x\beta) \]

\[ \text{mle} \quad l(\hat{\beta}, \hat{\sigma}^2) = -\frac{n}{2} \log(2\pi\hat{\sigma}^2) - \frac{n}{2} \]

Akaikke Information Criterion (AIC) is based on the well-known Kullback–Leibler measure for discrepancy \( KL(f|g) = \int \log \frac{f(y)}{g(y)} f(y) \, dy \)

\[ AIC = -2l(\hat{\beta}, \hat{\sigma}^2) + 2p + 1 = n \log(2\pi\hat{\sigma}^2) + n + 2(p+1) \]

It measures the true density \( f(Y) \) and \( N(\hat{Y}, \hat{\sigma}^2) \)

BIC: \( zp \rightarrow ap \) e.g. \( an = \log(n) \)