Simple Linear Regression

2016年10月3日　　0:01

Assumption

1. Uncorrelated/independent
2. Common variance σ^2, $\mathrm{Var}(Y_i) = \sigma^2$
3. Linear, $\mathbb{E}(Y_i|x_i) = \beta_0 + \beta_1 x_i$

$$\mathrm{s.e.}(\widehat{\beta_1}) = \sqrt{\mathrm{Var}(\widehat{\beta_1})} = \sqrt{\frac{\hat{\sigma}^2}{S_{XX}}}$$

$$\mathrm{s.e.}(\widehat{\beta_0}) = \sqrt{\mathrm{Var}(\widehat{\beta_0})} = \sqrt{\hat{\sigma}^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{XX}}\right)}$$

Least Squares Estimation

$$Q = \sum_{i=1}^{n}\{y_i - E(Y_i|x_i)\}^2 = \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$$

$$\frac{\partial Q}{\partial \beta_0} = 0 \implies \widehat{\beta_0} = \bar{y} - \widehat{\beta_1}\bar{x}$$

$$\frac{\partial Q}{\partial \beta_1} = 0 \implies \widehat{\beta_1} = \frac{S_{XY}}{S_{XX}}$$

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

$$S_{XY} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n}(x_i - \bar{x})y$$

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

$\widehat{\beta_0}$ and $\widehat{\beta_1}$ are expressed as a function of the random variables Y_i instead of the obseved data y_i

Properties

$\widehat{\beta_0}$ and $\widehat{\beta_1}$ are unbiased and consistent estimators of β_0 and β_1, respectively

$$\mathbb{E}(\widehat{Y_i}) = \beta_0 + \beta_1\bar{x}$$

$$\mathbb{E}(\widehat{\beta_1}) = \beta_1$$

$$\mathbb{E}(\widehat{\beta_0}) = \beta_0$$

$$\mathrm{Var}(\widehat{\beta_1}) = \frac{\sigma^2}{S_{XX}}$$

$$\mathrm{Var}(\widehat{\beta_0}) = \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{XX}}\right)$$

$$\mathrm{cov}(\widehat{\beta_0}, \widehat{\beta_1}) = \mathrm{cov}\left(\sum\left(\frac{1}{n} - \frac{\bar{x}(x_i - \bar{x})}{S_{XX}}\right)Y_i, \sum\frac{x_i - \bar{x}}{S_{XX}}Y_i\right)$$

$$= \sum\left(\frac{1}{n} - \frac{\bar{x}(x_i - \bar{x})}{S_{XX}}\right)\left(\frac{x_i - \bar{x}}{S_{XX}}\right)\mathrm{Var}(Y_i) = -\frac{\bar{x}}{S_{XX}}\sigma^2$$

$$\mathrm{cov}(\bar{Y}, \widehat{\beta_1}) = \mathrm{cov}\left(\sum\frac{1}{n}Y_i, \sum\frac{x_i - \bar{x}}{S_{XX}}Y_i\right) = 0$$

estimate

Residual and Regression Sums of Squares

$$e_i = y_i - \widehat{y_i} = (y_i - \bar{y}) - \widehat{\beta_1}(x_i - \bar{x})$$

$$\mathrm{RSS} = \sum e_i^2 = S_{YY} - \frac{S_{XY}^2}{S_{XX}} = S_{YY} - \mathrm{regressionSS} = \mathrm{totalSS} - \mathrm{regressionSS}$$

$$\mathbb{E}(\mathrm{RSS}) = (n-2)\sigma^2$$

$$\hat{\sigma}^2 = s^2 = \frac{1}{n-2}\mathrm{RSS} = \frac{\mathrm{RSS}}{\mathrm{DF}} = \mathrm{residualMeanSquare}$$

$$Q = \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2 = S_{XX}\left(\beta_1 - \frac{S_{XY}}{S_{XX}}\right)^2 + n(\beta_0 - \bar{y} - \beta_1\bar{x})^2 + S_{YY} - \frac{S_{XY}^2}{S_{XX}}$$

Coefficient of determination (`Multiple R-Squared`)

$$R^2 = \frac{\mathrm{regressionSS}}{\mathrm{totalSS}} = \frac{S_{XY}^2/S_{XX}}{S_{YY}} = \frac{S_{XY}^2}{S_{XX}S_{YY}}$$

ANOVA

SOURCE	DF	SS	MS	F	P
Regression	1	$\frac{S_{XY}^2}{S_{XX}}$	$\frac{S_{XY}^2}{S_{XX}}$	$\frac{S_{XY}^2}{S_{XX}\hat{\sigma}^2}$	
Residual	$n-2$	$S_{YY} - \frac{S_{XY}^2}{S_{XX}}$	$\hat{\sigma}^2$		
Total	$n-1$	S_{YY}			

$$\mathrm{MS} = \frac{\mathrm{SS}}{\mathrm{DF}}$$

$$F = \frac{\mathrm{regressionMS}}{\mathrm{residualMS}} = \frac{\mathrm{regressionSS}}{1}\Big/\frac{\mathrm{residualSS}}{n-2} \sim F_{1,n-2}$$

F can be used to test whether the expected response depends on the explanatory variable.

Alternative Formulation

$\mathbb{E}(Y_i|x_i) = \gamma + \beta_1(x_i - \bar{x})$

So that γ is the expected response at $x = \bar{x}$, rather than at $x = 0$.

Least Squares Estimation, $\hat{\gamma} = \bar{y}, \widehat{\beta_1} = \frac{S_{XY}}{S_{XX}}$

$\mathrm{Var}(\hat{\gamma}) = \frac{\sigma^2}{n}, \mathrm{cov}(\hat{\gamma}, \widehat{\beta_1}) = 0$

$$\mathrm{Var}(\widehat{\beta_0} + \widehat{\beta_1}x_*) = \mathrm{Var}\left(\hat{\gamma} + \widehat{\beta_1}(x_* - \bar{x})\right) = \sigma^2\left(\frac{1}{n} + \frac{(x_* - \bar{x})^2}{S_{XX}}\right)$$

Inference: Confidence Interval (Expected Response)

$\mathbb{E}(Y|x_*) = \beta_0 + \beta_1 x_*$

$\widehat{\mathbb{E}}(Y|x_*) = \widehat{\beta_0} + \widehat{\beta_1}x_* = \bar{Y} + \widehat{\beta_1}(x_* - \bar{x})$

Using () or properties of variance,

$$\mathrm{Var}\left(\widehat{\mathbb{E}}(Y|x_*)\right) = \sigma^2\left(\frac{1}{n} + \frac{(x_* - \bar{x})^2}{S_{XX}}\right)$$

$T \sim t_{n-2}$

Inference: Prediction Interval (Future Response)

$Y_* = \beta_0 + \beta_1 x_* + \varepsilon$, where $\varepsilon \sim N(0, \sigma^2)$

$$\mathrm{Var}(Y_*) = \sigma^2\left(1 + \frac{1}{n} + \frac{(x_* - \bar{x})^2}{S_{XX}}\right)$$

$T \sim t_{n-2}$

Inference (When Normal Distributed)

$$\widehat{\beta_1} \sim N\left(\beta_1, \frac{\sigma^2}{S_{XX}}\right) \longrightarrow \frac{\widehat{\beta_1} - \beta_1}{\sqrt{\sigma^2/S_{XX}}} \sim N(0,1) \longrightarrow \frac{\widehat{\beta_1} - \beta_1}{\sqrt{\hat{\sigma}^2/S_{XX}}} \sim t_{n-2}$$

Inference (When Normal Distributed)

$\widehat{\beta_1} \sim N\left(\beta_1, \frac{\sigma^2}{S_{XX}}\right)$ $\longrightarrow$ $\frac{\widehat{\beta_1}-\beta_1}{\sqrt{\sigma^2/S_{XX}}} \sim N(0,1)$ $\longrightarrow$ $\frac{\widehat{\beta_1}-\beta_1}{\sqrt{\hat{\sigma}^2/S_{XX}}} \sim t_{n-2}$

$\widehat{\beta_0} \sim N\left(\beta_0, \sigma^2\left(\frac{1}{n}+\bar{x}^2\right)\right)$ $\longrightarrow$ $\frac{\widehat{\beta_0}-\beta_0}{\sqrt{\sigma^2\left(\frac{1}{n}+\frac{\bar{x}^2}{S_{XX}}\right)}} \sim N(0,1)$ $\longrightarrow$ $\frac{\widehat{\beta_0}-\beta_0}{\sqrt{\hat{\sigma}^2\left(\frac{1}{n}+\frac{\bar{x}^2}{S_{XX}}\right)}} \sim t_{n-2}$

$\bar{Y} \sim N\left(\beta_0+\beta_1\bar{x}, \frac{\sigma^2}{n}\right)$, independently of $\widehat{\beta_1}$

$\frac{\text{RSS}}{\sigma^2} = \frac{(n-2)\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-2}$, independently of $\widehat{\beta_1}$ and $\bar{Y}$

$$\begin{cases} Z \sim N(0,1) \\ Y \sim \chi^2_n \end{cases} \rightarrow T = \frac{Z}{\sqrt{Y/n}} \sim t_n$$

$$T = \frac{\widehat{\beta_1}-\beta_1}{\sqrt{\hat{\sigma}^2/S_{XX}}} = \frac{\dfrac{\widehat{\beta_1}-\beta_1}{\sqrt{\sigma^2/S_{XX}}}}{\sqrt{\dfrac{(n-2)\hat{\sigma}^2}{\sigma^2}/(n-2)}} = \frac{Z}{\sqrt{Y/(n-2)}} \sim t_{n-2}$$

Leverage

$$\widehat{y_i} = \widehat{\beta_0}+\widehat{\beta_1}x_i = \bar{y}+\widehat{\beta_1}(x_i-\bar{x}) = \bar{y}+(x_i-\bar{x})\sum_{j=1}^{n}\frac{(x_j-\bar{x})y_j}{S_{XX}}$$

$$= \sum_{j=1}^{n}\left(\frac{1}{n}+\frac{(x_i-\bar{x})(x_j-\bar{x})}{S_{XX}}\right)y_j = \sum_{j=1}^{n} a_{ij}y_j$$

Influence Matrix: $A = \{a_{ij}\}$

Leverage is A_{ii}, a measure of how much that particular data point influences its own fitted value. *(potential)*

Large Residual + High Leverage = Strong Influence

Small Residual + High Leverage = not too great influence (while inconsistent)

Large Residual + Low Leverage = not much effect on the fit

Cook's distance

$$d_i = \frac{1}{3\hat{\sigma}^2}\sum_{j=1}^{n}\left(\widehat{y_{j(i)}}-\widehat{y_j}\right)^2$$

$\widehat{y_{j(i)}}$ is the i'th fitted value obtained when the j'th point is excluded when computing fthe regression model.

3 = p + 1, where p is the number of unknown parameters.

Cook's distance measures the influence that each data point is *in fact* exerting on the fit.

Large Cook's Distance = potential outliers

variable.

Analysis of Residuals

$\varepsilon_i = Y_i - \mathbb{E}(Y_i|x_i) = Y_i - \beta_0 - \beta_1 x_i$

Assumptions (derive from original assumptions):

1. Uncorrelated/independent, $\text{cov}(\varepsilon_i, \varepsilon_j|x) = 0$
2. Common variance σ^2, $\text{Var}(\varepsilon_i|x) = \sigma^2$
3. $\mathbb{E}(\varepsilon_i|x) = 0$
4. If Y_i is Normal, $\varepsilon_i \overset{\text{iid}}{\sim} N(0,\sigma^2)$

Estimate ε_i,

$e_i = y_i - \widehat{\beta_0} - \widehat{\beta_1}x_i = y_i - \bar{y} - \widehat{\beta_1}(x_i-\bar{x})$

$p_{ij} = \frac{1}{n}+\frac{(x_i-\bar{x})(x_j-\bar{x})}{S_{XX}}$

Then, (corresponding random variable)

$\text{cov}(E_j, E_j|x) = -pij\sigma^2$

$\text{Var}(E_i|x) = (1-p_{ij})\sigma^2 = \left(1-\frac{1}{n}-\frac{(x_i-\bar{x})^2}{S_{XX}}\right)\sigma^2$

$\mathbb{E}(E_i|x) = 0$

If Y_i are Normal, so are the E_i.

If n is large and none of the quantities $|x_i-\bar{x}|$ is large then the p_{ij} are small.

Standardized Residuals

$$r_i = \frac{e_i}{\text{estimated standard error of } E_i} = \frac{e_i}{\sqrt{\hat{\sigma}^2\left(1-\frac{1}{n}-\frac{(x_i-\bar{x})^2}{S_{XX}}\right)}}$$

i. Plot r_i against x_i, look for sign that $\mathbb{E}(R_i|x_i)$, $\text{Var}(R_i|x_i)$depends on x_i
ii. See if any unusually large values, $|r_i| > 2.5, > 3$
iii. Plot Normal probability of r_i

Multiple Regression

1. $\mathbb{E}(\mathbf{Y}|\mathbf{X}) = \mathbf{X}\beta$, $\mathrm{Var}(\mathbf{Y}|\mathbf{X}) = \sigma^2\mathbf{I}_n$
   Specially, in simple linear regression, $y_i = \beta_0 + \beta_1 x_i$,
   $$\mathbf{X} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix},\ \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix},\ \mathbb{E}(\mathbf{Y}|\mathbf{X}) = \mathbf{1}_n\beta_0 + \mathbf{x}\beta_1,\ \mathrm{Var}(\mathbf{Y}|\mathbf{X}) = \sigma^2\mathbf{I}_n$$

2. Least squares estimation: $Q = \sum_{i=1}^{n}\{y_i - \mathbb{E}(Y_i|\mathbf{X})\}^2 = \mathbf{y}^T\mathbf{y} - 2\mathbf{y}^T\mathbf{X}\beta + \beta^T\mathbf{X}^T\mathbf{X}\beta$
   Least squares *unbiased* estimator: $\hat{\boldsymbol{\beta}} = (\mathbf{X}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$
   $\mathbb{E}(\hat{\boldsymbol{\beta}}|\mathbf{X}) = \boldsymbol{\beta}$, $\mathrm{Var}(\hat{\boldsymbol{\beta}}|\mathbf{X}) = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$, $\mathrm{Var}(\mathbf{c}^T\hat{\boldsymbol{\beta}}|\mathbf{X}) = \sigma^2\mathbf{c}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{c}$

3. Vector of residuals: $\mathbf{e} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} = (\mathbf{I}_n - \mathbf{P_X})\mathbf{y}$,
   where $\mathbf{P_X} = \mathbf{X}(\mathbf{X}\mathbf{X})^{-1}\mathbf{X}^T$ is $n \times n$, symmetric, idempotent, rank p,
   $(\mathbf{I}_n - \mathbf{P_X})\mathbf{X} = \mathbf{0}$, $(\mathbf{I}_n - \mathbf{P_X})\mathbf{P_X} = \mathbf{0}$
   $\mathbb{E}(\mathbf{E}|\mathbf{X}) = (\mathbf{I}_n - \mathbf{P_X})\mathbb{E}(\mathbf{Y}|\mathbf{X}) = (\mathbf{I}_n - \mathbf{P_X})\mathbf{X}\boldsymbol{\beta} = \mathbf{0}$
   $\mathrm{Var}(\mathbf{E}|\mathbf{X}) = (\mathbf{I}_n - \mathbf{P_X})\sigma^2\mathbf{I}_n(\mathbf{I}_n - \mathbf{P_X}) = \sigma^2(\mathbf{I}_n - \mathbf{P_X})$
   $\mathrm{RSS} = \mathbf{e}^T\mathbf{e} = \mathbf{y}^T\mathbf{y} - \hat{\boldsymbol{\beta}}^T\mathbf{X}^T\mathbf{y}$, $\hat{\sigma}^2 = \dfrac{\mathrm{RSS}}{n-p} = \dfrac{\mathbf{y}^T\mathbf{y} - \hat{\boldsymbol{\beta}}^T\mathbf{X}^T\mathbf{y}}{n-p}$

4. Alternative formulation (for models with an intercept)
   $\mathbb{E}(Y_i|\mathbf{X}) = \gamma + \beta_1(x_{i1} - \bar{x}_1) + \beta_2(x_{i2} - \bar{x}_2) + \cdots + \beta_q(x_{iq} - \bar{x}_q)$
   $\mathbb{E}(\mathbf{Y}|\mathbf{X}) = \gamma\mathbf{1}_n + \dot{\mathbf{X}}\dot{\boldsymbol{\beta}}$, where $\dot{\mathbf{X}}_{ij} = x_{ij} - \bar{x}_j$, $\dot{\boldsymbol{\beta}} = \begin{pmatrix}\beta_1 & \cdots & \beta_q\end{pmatrix}^T$, $\gamma = \beta_0 + \beta_1\bar{x}_1 + \cdots + \beta_q\bar{x}_q$
   Least squares *unbiased* estimators: $\hat{\gamma} = \bar{y}$, $\hat{\dot{\boldsymbol{\beta}}} = \left(\dot{\mathbf{X}}_T\dot{\mathbf{X}}\right)^{-1}\dot{\mathbf{X}}^T\mathbf{y}$
   $\mathrm{Var}\left(\hat{\dot{\boldsymbol{\beta}}}\middle|\mathbf{X}\right) = \sigma^2\left(\dot{\mathbf{X}}^T\dot{\mathbf{X}}\right)^{-1}$, $\mathrm{Var}\left(\hat{\gamma}|\mathbf{X}\right) = n^{-1}\sigma^2$, $\mathrm{cov}\left(\hat{\dot{\boldsymbol{\beta}}}, \hat{\gamma}\middle|\mathbf{X}\right) = \mathbf{0}$

5. Distributional results:
   - $\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1})$
   - regression (model) SS $\mathbf{Y}^T\mathbf{P_X}\mathbf{Y} \sim \sigma^2\chi^2(q, \sigma^{-2}\dot{\boldsymbol{\beta}}^T\dot{\mathbf{X}}^T\dot{\mathbf{X}}\dot{\boldsymbol{\beta}})$
   - RSS $\mathbf{Y}^T(\mathbf{I}_n - \mathbf{P_X})\mathbf{Y} \sim \sigma^2\chi^2(n - q - 1, 0)$
   - RSS and regression SS are independent
   - $\dfrac{\mathbf{c}^T\hat{\boldsymbol{\beta}} - \mathbf{c}^T\boldsymbol{\beta}}{\sigma\sqrt{\mathbf{c}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{c}}} \sim N(0,1)$ and
     $\dfrac{\mathbf{c}^T\hat{\boldsymbol{\beta}} - \mathbf{c}^T\boldsymbol{\beta}}{\hat{\sigma}\sqrt{\mathbf{c}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{c}}} \sim t(n-p)$ (test hypotheses about linear funcs of the parameters)

6. 95% Confidence interval: $\mathbf{c}^T\hat{\boldsymbol{\beta}} \pm t_{0.025}\hat{\sigma}\sqrt{\mathbf{c}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{c}}$
   CI for future response: $\mathbf{x}_*^T\hat{\boldsymbol{\beta}} \pm t_{0.025}\hat{\sigma}\sqrt{\mathbf{x}_*^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_*}$
   Prediction interval: $\mathbf{x}_*^T\hat{\boldsymbol{\beta}} \pm t_{0.025}\hat{\sigma}\sqrt{1 + \mathbf{x}_*^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_*}$

7. To test $\boldsymbol{\beta} = \mathbf{0}$,
   $F = \frac{\text{regression SS}}{\text{RSS}} \sim F(p, n-p)$ (chk simple linear regression, but with different SS)

8. To test a more general linear hypothesis about the coefficients of the model, $\mathbf{C}\boldsymbol{\beta} = \mathbf{d}$,
   Extra SS $= \left(\mathbf{C}\hat{\boldsymbol{\beta}} - \mathbf{d}\right)^T\left(\mathbf{C}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{C}^T\right)^{-1}\left(\mathbf{C}\hat{\boldsymbol{\beta}} - \mathbf{d}\right)$
   $$F = \frac{(\text{ESS for } H_0)/c}{\text{RMS}} = \frac{(\text{RSS under } H_0 - \text{RSS under full model})/c}{\text{RMS}} \sim F(k, n-p)$$