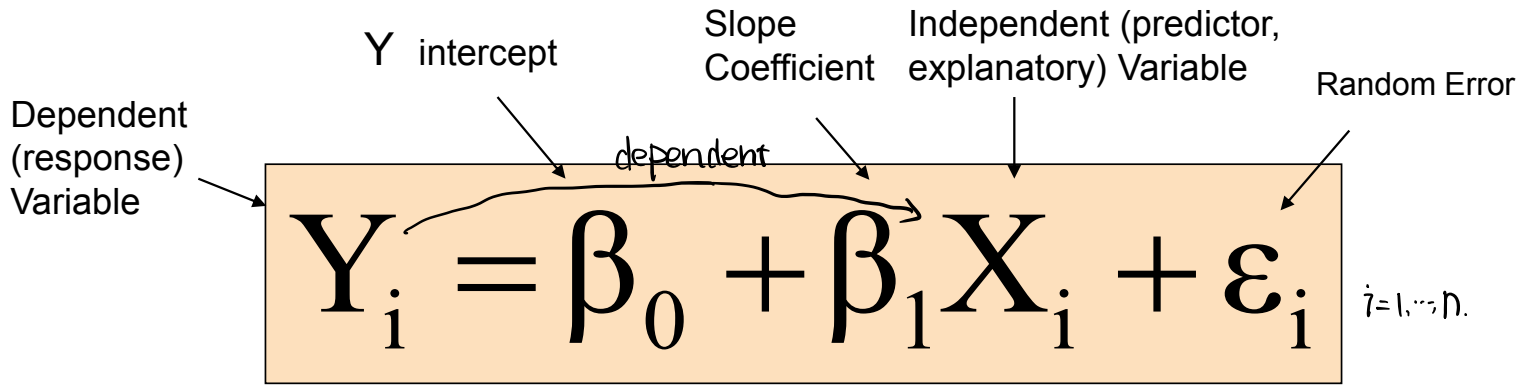
Ch2. Simple Linear Regression

- **Relationship between 2 variables**
- **The regression model**
- **Assumptions**
- **Estimation and method of least squares**
- **Inferences concerning β_1 and β_0**
- **Estimation of the mean of the response variable for a given level of X**
- **Prediction of new observation**
- **Analysis of variance approach to regression analysis**
- **Measures of linear association between X and Y**

Simple Linear Regression Model



Assumptions:

- $E(\varepsilon_i) = 0$
- Variance $(\varepsilon_i) = \sigma^2$
- Covariance $(\varepsilon_i, \varepsilon_j) = 0$

$$\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

In matrix notation?

$$\underset{n \times 1}{\underset{\sim}{Y}} = \underset{n \times 2}{\underset{\sim}{X}} \; \underset{2 \times 1}{\underset{\sim}{\beta}} + \underset{n \times 1}{\underset{\sim}{\varepsilon}}$$

Simple Linear Regression Equation

The simple linear regression equation provides an estimate of the population regression line

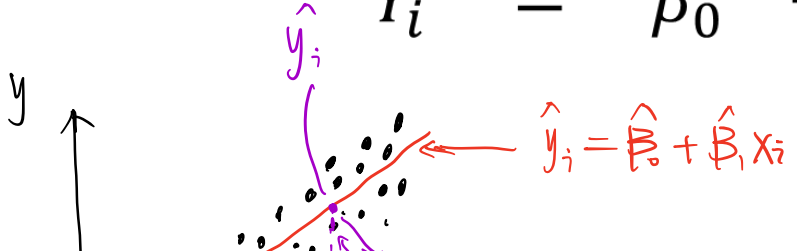
Estimated (or predicted) Y value for observation i

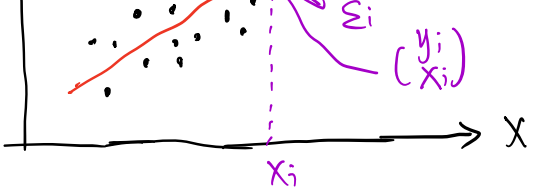
Estimate of the regression intercept

Estimate of the regression slope

Value of *X* for observation *i*

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$





Interpretation of the Slope and the Intercept

- $\hat{\beta}_0$ is the estimated average value of Y when the value of X is zero
- $\hat{\beta}_1$ is the estimated change in the average value of Y as a result of a one-unit change in X

Example

- A real estate agent wishes to examine the relationship between the selling price of a home and its size (measured in square feet)

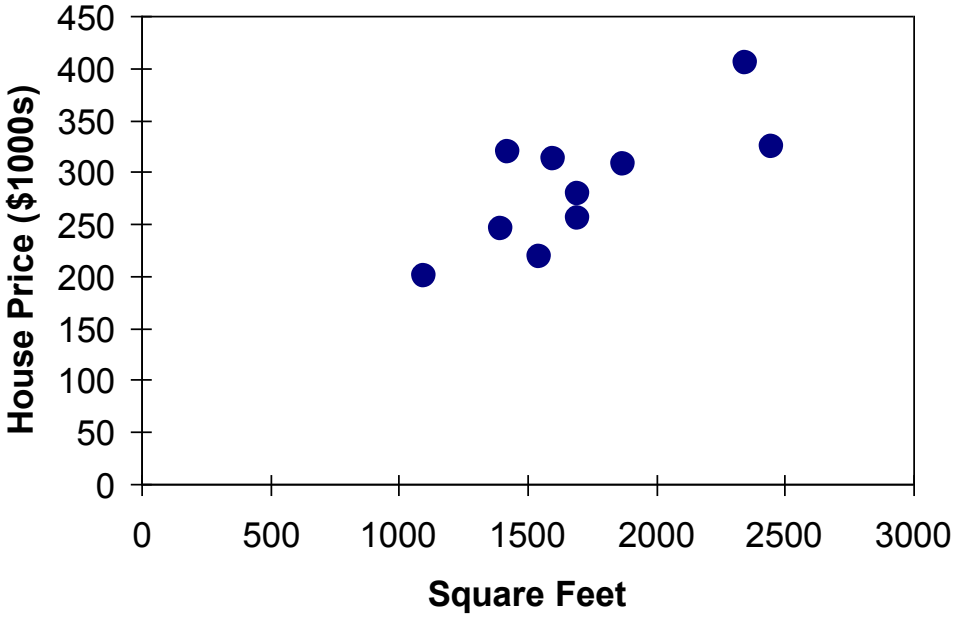
- A random sample of 10 houses is selected
  - Dependent variable (*Y*) = house price in $1000s
  - Independent variable (*X*) = square feet

Sample Data for House Price Model

House Price in $1000s (*Y*)	Square Feet (*X*)
245	1400
312	1600
279	1700
308	1875
199	1100
219	1550
405	2350
324	2450
319	1425
255	1700

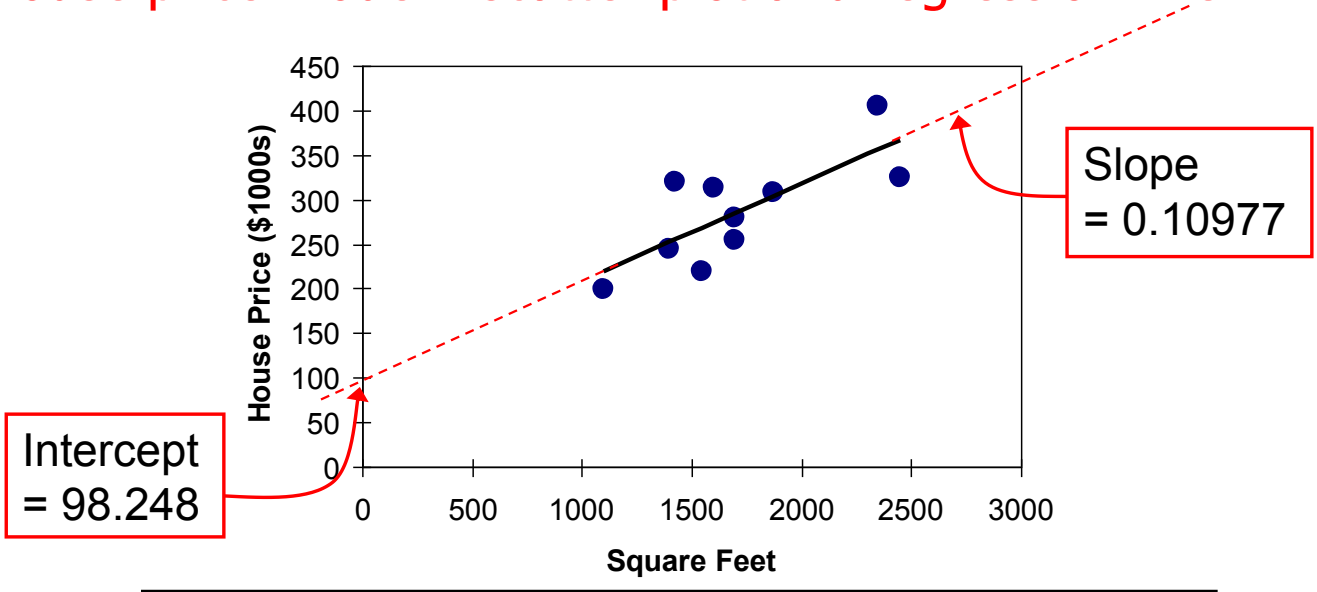
Graphical Presentation

House price model: scatter plot



Graphical Presentation

House price model: scatter plot and regression line



$$\widehat{\text{house price}} = 98.24833 + 0.10977(\text{square feet})$$

Interpretation of the Intercept, $\hat{\beta}_0$

$$\widehat{\text{house price}} = 98.24833 + 0.10977(\text{square feet})$$

- $\hat{\beta}_0$ is the estimated average value of Y when the value of X is zero (if $X = 0$ is in the range of observed X values)
- Here, no houses had 0 square feet, so $\hat{\beta}_0 = 98.24833$ just indicates that, for houses within the range of sizes observed, \$98,248.33 is the portion of the house price not explained by square feet

Interpretation of the Slope Coefficient, $\hat{\beta}_1$

解释

$$\widehat{\text{house price}} = 98.24833 + 0.10977(\text{square feet})$$

- $\hat{\beta}_1$ measures the estimated change in the average value of Y as a result of a one-unit change in X
  - Here, $\hat{\beta}_1$ = .10977 tells us that the average value of a house increases by .10977($1000) = $109.77, on average, for each additional one square foot of size

Predictions using Regression Analysis

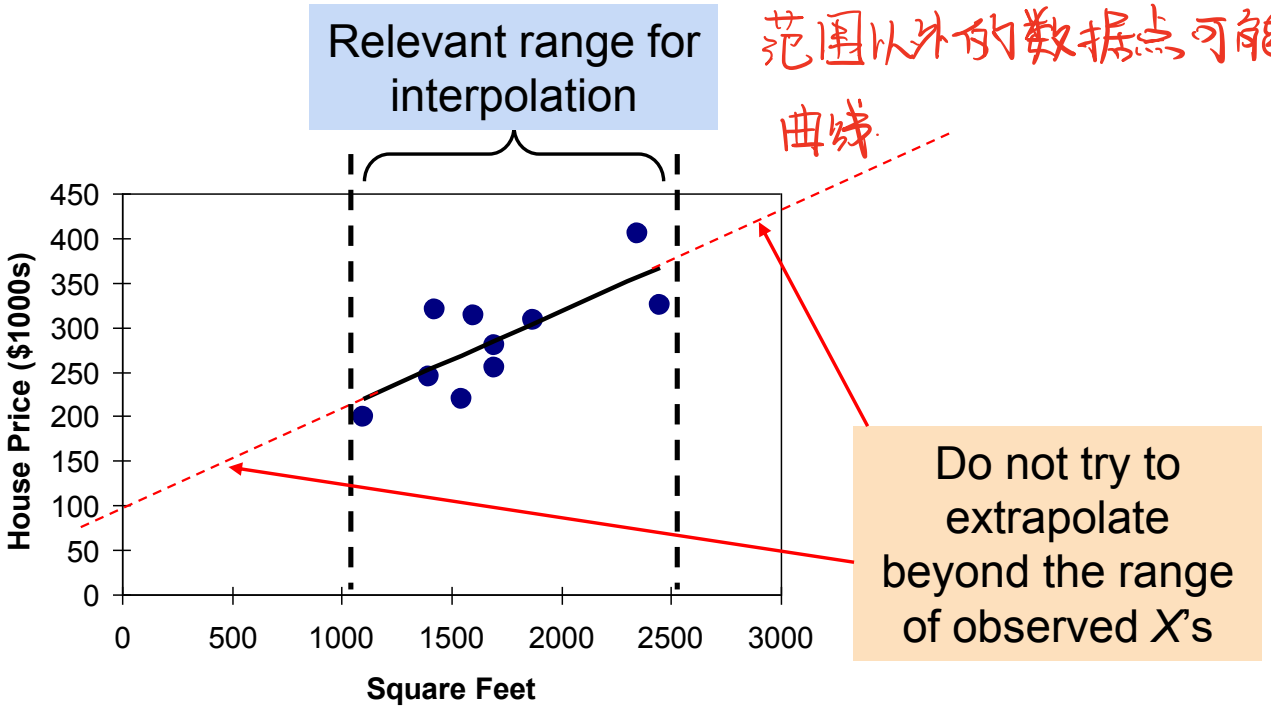
Predict the price for a house with 2000 square feet:

$$\widehat{\text{house price}} = 98.25 + 0.1098\,(\text{sq.ft.})$$
$$= 98.25 + 0.1098\,(2000)$$
$$= 317.85$$

The predicted price for a house with 2000 square feet is 317.85($1,000s) = $317,850

Interpolation vs. Extrapolation

- When using a regression model for prediction, only predict within the relevant range of data



Estimation (Method of Least Squares)

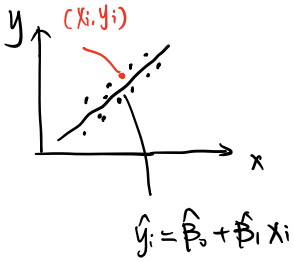
- $\hat{\beta}_0$ and $\hat{\beta}_1$ are obtained by finding the values of β_0 and β_1 that minimize the sum of the squared differences between Y and $\hat{Y}$

Remark 2.1

$\Rightarrow$ estimate β_0, β_1

Remark 2.1: (1) how to estimate β_0, β_1

(2) how good is the estimation



Data: $\mathcal{D} = \{\begin{pmatrix} y_i \\ x_i \end{pmatrix}, i = 1, \cdots, n\}$

$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$

We want to minize $\sum_{i=1}^n dis\,((x_i, y_i), (x_i, \hat{y}_i))$

$\Rightarrow \begin{cases} L_2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 & \text{— better. 平方后曲线光滑性质更好.} \\ L_1 = \sum_{i=1}^n |y_i - \hat{y}_i| & \text{— 连续但不一定光滑} \end{cases}$

$* \min_{\beta_0, \beta_1} \boxed{\sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2} \to$ SS (Sum of squares) $*$ $\begin{cases} \frac{\partial SS}{\partial \beta_0} = 0 = -2\sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i) \\ \frac{\partial SS}{\partial \beta_1} = 0 = -2\sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i) x_i \end{cases}$

(LSE) least square error

$(*) \Rightarrow 0 = -2\left(\sum_{i=1}^n y_i - n\beta_0 - \beta_1 \sum_{i=1}^n x_i\right)$

$= -2n(\bar{y} - \beta_0 - \beta_1 \bar{x})$

$\Rightarrow \bar{y} = \beta_0 + \beta_1 \bar{x}$

$\Rightarrow \beta_0 = \bar{y} - \beta_1 \bar{x}$

$(**) \Rightarrow 0 = -2(\sum_{i=1}^n x_i y_i - \beta_0 \sum_{i=1}^n x_i - \beta_1 \sum_{i=1}^n x_i^2)$

$= -2(\sum_{i=1}^n x_i y_i - n\beta_0 \bar{x} - \beta_1 \sum_{i=1}^n x_i^2)$

$\beta_0 = \bar{y} - \beta_1 \bar{x} \longrightarrow = -2(\sum_{i=1}^n x_i y_i - n(\bar{y} - \beta_1 \bar{x})\bar{x} - \beta_1 \sum_{i=1}^n x_i^2)$

$= -2\left[\sum_{i=1}^n x_i y_i - n\bar{x}\bar{y} + n\beta_1 \bar{x}^2 - \beta_1 \sum_{i=1}^n x_i^2\right]$

$= -2\left[\sum_{i=1}^n x_i y_i - n\bar{x}\bar{y} + \beta_1 [n\bar{x}^2 - \sum_{i=1}^n x_i^2]\right]$

$\Rightarrow \hat{\beta}_1 = \frac{\sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^n x_i^2 - n\bar{x}^2} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})(x_i - \bar{x})} \overset{d}{=} \frac{S_{xy}}{S_{xx}}$

Estimation (Method of Least Squares)

LSE (estimation by least square)

- $\hat{\beta}_1 = \dfrac{\sum_{i=1}^{n}(X_i-\overline{X})(Y_i-\overline{Y})}{\sum_{i=1}^{n}(X_i-\overline{X})^2} \overset{\text{d}}{=} \dfrac{S_{xy}}{S_{xx}}$
- $\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}$

note that

$S_{XY} = \sum_{i=1}^{n}(X_i-\overline{X})(Y_i-\overline{Y}) = \sum_{i=1}^{n} X_i Y_i - n\overline{X}\,\overline{Y}$

$S_{XX} = \sum_{i=1}^{n}(X_i-\overline{X})^2 = \sum_{i=1}^{n} X_i^2 - n\overline{X}^2$

$S_{YY} = \sum_{i=1}^{n}(Y_i-\overline{Y})^2 = \sum_{i=1}^{n} Y_i^2 - n\overline{Y}^2$

Relationship between slope (b_1) and sample correlation (r)

- $\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = r \cdot \sqrt{\frac{S_{yy}}{S_{xx}}}$

Similarity:

* $H_0: r = 0$
  $H_0: \beta_1 = 0$ → Same conclusion
* signs of $\hat{\beta}_1$ and r are the same

相关系数

- $r = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}}$

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

Difference

* meaning of r and β_1 are different
* Interpretation of $\hat{\beta}_1$

Estimation of error terms variance σ^2

- The estimator of σ^2 is

$$S^2 = MSE = \frac{SSE}{n-2} = \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n-2} = \frac{\sum_{i=1}^{n} e_i^2}{n-2}$$

- S^2 is an unbiased estimator of σ^2

严格的证明留到期中后

目前可以通过自由度来理解

Remark 2.2 对于自由度的理解

Recall: $y_1, \cdots, y_n \sim N(\mu, \sigma^2)$.

$S^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}$ unbiased estimator of σ^2

$n-1$: degree of freedom

$\hat{\sigma}^2 = S^2 \leftarrow$ unbiased estimation of σ^2.

$\sum y_i^2$ — degree of freedom $= n$ (d.f)

$\sum (y_i - \bar{y})^2$ — d.f $= n-1$

↳ $\sum (y_i - \bar{y}) = 0$. constrained equation (约束条件)

推广

$\Rightarrow y_i \sim (\mu_i, \sigma^2)$ $\mu_i \neq \beta_0 + \beta_1 x_i$, $\hat{y}_i = \hat{\mu}_i = \hat{\beta}_0 + \hat{\beta} x_i$

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

$\begin{cases} \frac{\partial SS}{\partial \beta_0} = 0 \\ \frac{\partial SS}{\partial \beta_1} = 0 \end{cases}$ 两个限制方程

degree of freedom need -2

Estimation (Method of Maximum Likelihood)

- $\hat{\beta}_1 = \dfrac{\sum_{i=1}^{n}(X_i-\overline{X})(Y_i-\overline{Y})}{\sum_{i=1}^{n}(X_i-\overline{X})^2}$

Normal Error Model

Remark 2.3

- $\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}$

- The estimator of σ^2 is $\dfrac{SSE}{n} = \dfrac{n-2}{n} S^2$

Under the assumption.

- MLE of β_0 = LSE of β_0 (unbiased)
- MLE of β_1 = LSE of β_1 (unbiased)
- MLE of σ^2 (< unbiased estimator of σ^2) asymptotically unbiased

一起估计. LSE of σ^2 是分开估计的

Remark 2.3: 使用MLE来估计 $\hat{\beta}_0$、$\hat{\beta}_1$.

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

$y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$

$$P(y_i) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma^2}\right\}$$

Likelihood: $L = L(\beta_0, \beta_1, \sigma^2) = \prod_{i=1}^{n} P(y_i | \beta_0, \beta_1, \sigma^2)$

log-likelihood: $l = \sum_{i=1}^{n} \log P(y_i | \beta_0, \beta_1, \sigma^2) = l(\underset{\sim}{\theta})$ $\theta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \sigma^2 \end{pmatrix}$

MLE of θ: $\hat{\theta} = \arg\max_{\theta} l(\theta) \Rightarrow$ (Assignment 1)

$$\left.\begin{aligned} \hat{\beta}_0 &= \frac{S_{xy}}{S_{xx}} \\ \hat{\beta}_1 &= \bar{y} - \hat{\beta}_1 \bar{x} \end{aligned}\right\} \text{MLE} = \text{LSE}$$

$$\hat{\sigma}^2 = \frac{\sum (y_i - \hat{y}_i)^2}{n} = \frac{n-2}{n} S^2$$

$$E(\hat{\sigma}^2) = \frac{n-2}{n}\sigma^2 \xrightarrow{\text{as } n \to \infty} \sigma^2.$$

MLE of σ^2 is an approximated unbiased estimation.

continue

接上：证明无偏估计：

统计中最困难的是建模

("If" is correct)

big "if"

1°. $*\hat{\beta}_1$, want to prove $E(\hat{\beta}_1)=\beta_1$

If $y_i=\beta_0+\beta_1 x_i+\varepsilon_i$, $\varepsilon_i\sim N(0,\sigma^2)$ 假定模型正确的情况下.

$$\hat{\beta}_1=\frac{\sum_{i=1}^n (x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^n (x_i-\bar{x})^2}=\sum_{i=1}^n \frac{(x_i-\bar{x})}{S_{xx}}y_i-\bar{y}\sum_{i=1}^n \frac{(x_i-\bar{x})}{S_{xx}}=0$$

$$k_i\triangleq\frac{x_i-\bar{x}}{S_{xx}},\ \sum k_i=0$$

$$=\sum_{i=1}^n k_i y_i.$$

thus. $E(\hat{\beta}_1)=\sum_{i=1}^n k_i E y_i=\sum_{i=1}^n k_i(\beta_0+\beta x_i)=\beta_0\sum k_i+\beta_1\sum k_i x_i$

($\sum k_i=0$, $\sum k_i x_i=1$)

$$=\beta_1$$

$$\sum k_i x_i=\sum k_i x_i-\sum k_i\bar{x}$$
$$=\sum k_i(x_i-\bar{x})$$
$$=\sum\frac{(x_i-\bar{x})(x_i-\bar{x})}{S_{xx}}=1$$

thus. $Var(\hat{\beta}_1)=\sum_{i=1}^n k_i^2 Var(y_i)=\sigma^2\sum_{i=1}^n\frac{(x_i-\bar{x})^2}{S_{xx}^2}$

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

$\Rightarrow$ · y_i are independent.

· If ε_i iid $N(0,\sigma^2)$, $\hat{\beta}_1\sim N(\beta_1,\frac{\sigma^2}{S_{xx}})$

Continue

2°. 证明 $\hat{\beta}_0$ 为无偏估计

$$E\hat{\beta}_0 = E\bar{y} - \bar{x}E\hat{\beta}_1$$
$$= \beta_0 + \beta_1\bar{x} - \bar{x}\beta_1$$
$$= \beta_0$$

$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$

$E\bar{y} = \frac{1}{n}\sum_{i=1}^{n}(\beta_0 + \beta_1 x_i) = \beta_0 + \beta_1\bar{x}$

$$Var(\hat{\beta}_0) = Var(\bar{y} - \bar{x}\hat{\beta}_1) = Var(\bar{y}) + \bar{x}^2 Var(\hat{\beta}_1) - 2\bar{x}\,Cov(\bar{y}, \hat{\beta}_1)$$
$$= \frac{\sigma^2}{n} + \frac{\bar{x}^2}{S_{xx}}\sigma^2$$
$$= \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)\sigma^2$$

(其中 $Cov(\bar{y}, \hat{\beta}_1) = 0$)

证: $Cov(\bar{y}, \hat{\beta}_1) = 0$.

$$\bar{y} = \left(\frac{1}{n}, \cdots, \frac{1}{n}\right)\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = A^T y$$

$$\hat{\beta}_1 = (k_1, \cdots, k_n)\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = B^T y.$$

$$Cov(\bar{y}, \hat{\beta}_1) = Cov(A^T y, B^T y)$$
$$= A^T Cov(y, y) B$$
$$= \sigma^2 A^T B = \sigma^2 \sum_{i=1}^{n} \frac{k_i}{n} = 0$$

$\Rightarrow$ If $\varepsilon_i \sim N(0, \sigma^2)$, i.i.d.

$\hat{\beta}_0 \sim N\left(\beta_0, \left(\frac{1}{n} + \frac{\bar{x}}{S_{xx}}\right)\sigma^2\right)$

3°. 证明 $\hat{\sigma}^2_{MLE}$ 是近似 unbiased

$$\hat{\sigma}^2_{MLE} = \frac{SSE}{n} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y}^2)$$
$$= \frac{n-2}{n}\boxed{\sum_{i=1}^{n}\frac{(y_i - \hat{y}_i)^2}{n-2}} = S^2$$

$\Rightarrow E\hat{\sigma}^2_{MLE} = \frac{n-2}{n}\sigma^2 \to \sigma^2$ as $n \to \infty$

More on inference

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

Assumptions:

- X_i are known constants, $i = 1, \ldots n$
- $\varepsilon_i \overset{ind}{\sim} N(0, \sigma^2)$

Therefore, $Y_i \overset{ind}{\sim} N(\beta_0 + \beta_1 X_i, \sigma^2)$

1° N.T. prove: $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$

$E(\hat{\beta}_1) = \beta_1$

Distribution of $\hat{\beta}_1$

- $$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2}$$

$$= \sum_{i=1}^{n} k_i Y_i$$

where $$k_i = \frac{(X_i - \overline{X})}{\sum_{i=1}^{n}(X_i - \overline{X})^2}$$

$E\hat{\beta}_1 = \sum_{i=1}^{n} k_i E y_i = \sum_{i=1}^{n} k_i(\beta_0 + \beta_1 x) = \beta_0 \Sigma k_i + \beta_1 \Sigma k_i x_i = \beta_1$

Distribution of $\hat{\beta}_1$

Hence, if $\varepsilon_i \sim N(0, \sigma^2)$

(1). $\hat{\beta}_1 \sim N(\beta_1, \sigma^2/S_{xx})$ where $S_{xx} = \sum_{i=1}^{n} (X_i - \bar{X})^2$

(2). $\frac{(n-2)S^2}{\sigma^2} \sim \chi^2(n-2)$

(3). $\hat{\beta}_1$ and S^2 are independent

$$\Rightarrow \frac{\hat{\beta}_1 - \beta_1}{\sigma / S_{xx}^{\frac{1}{2}}} \Big/ \left(\frac{(n-2)S^2}{\sigma^2} \Big/ n-2 \right)^{\frac{1}{2}} \sim t_{n-2}$$

$$t = \frac{\hat{\beta}_1 - \beta_1}{s / S_{xx}^{\frac{1}{2}}} \sim t_{n-2}$$

Testing (Two-sided test of β_1)

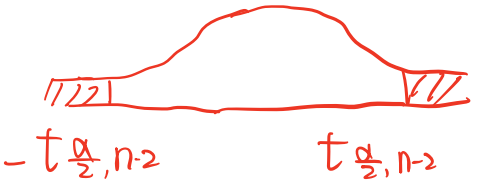
H_0: $\beta_1 = 0$ (no linear relationship)
H_1: $\beta_1 \neq 0$ (linear relationship does exist between X and Y)

Test statistics:

$$t = \frac{\hat{\beta}_1}{s/S_{xx}^{\frac{1}{2}}} \overset{H_0}{\sim} t_{n-2}$$

Decision rule:

Rejection field: if $|t| > t_{\frac{\alpha}{2},n-2}$



or p-value: $= Pr(t_{n-2} > |t|)$

Testing (Two-sided test of β_1)

exercise

$$H_0: \beta_1 = k$$
$$H_1: \beta_1 \neq k$$

Test statistics:

Decision rule:

Distribution of $\hat{\beta}_0$

$$\hat{\beta}_0 = \overline{Y} - b_1 \overline{X}$$

Hence,

$$\hat{\beta}_0 \sim \mathrm{N}(\beta_0, \sigma^2 \left[\frac{1}{n} + \frac{\overline{X}^2}{\sum_{i=1}^{n} (X_i - \overline{X})^2} \right])$$

Estimation of the mean of the response variable for a given level of X

Example

Suppose you want to develop a model to predict selling price of homes based on assessed value. A sample of 30 recently sold single-family houses in a small city is selected to study the relationship between selling price (Y, in $000) and assessed value (X, in $000). The houses in the city had been reassessed at full value 1 year prior to the study.

Estimate of the **average** selling price for houses with an assessed value of $70,000.

Remark 2.4:

model: $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad i = 1, \cdots, n.$

$D = \{(y_i, x_i), i = 1, \cdots, n\}$ $\quad \hat{\beta}_1 = S_{xy}/S_{xx}$

$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$

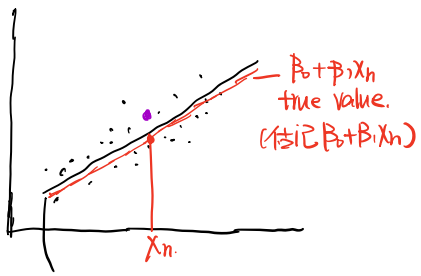
$x_h = \$70k$. N.T. estimate the average selling price

$y_h = \beta_0 + \beta_1 x_h \quad \hat{y}_h = \hat{\beta}_0 + \hat{\beta}_1 x_h$

How good is $\hat{y}_h$ = ?

1° $E(\hat{y}_h) = y_h$

2° $Var(\hat{y}_h) = ?$



$\hat{\beta}_0 + \hat{\beta}_1 x_h$ (fitted)

how to estimate: $Ey_h = \beta_0 + \beta_1 x_h \triangleq \theta_h$.

Estimation (Prediction): $\hat{\theta}_h \triangleq \hat{\beta}_0 + \hat{\beta}_1 x_h$

$E(\hat{\theta}_h) = E(\hat{\beta}_0) + E(\hat{\beta}_1) x_h \quad \Rightarrow E(\hat{\theta}_h - \theta_h) = 0$

$= \beta_0 + \beta_1 x_h = \theta_h$

then Consider the Variance of $\hat{\theta}_h - \theta_h$

$Var(\hat{\theta}_h - \theta_h) = Var(\bar{y} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 x_h - \beta_0 - \beta_1 x_h)$

$= Var(\bar{y} + \hat{\beta}_1 (x_h - \bar{x}))$

$= Var(\bar{y}) + (x_h - \bar{x})^2 Var(\hat{\beta}_1) + 2(x_h - \bar{x}) \underbrace{Cov(\bar{y}, \hat{\beta}_1)}_{0}$

$= \frac{b^2}{n} + \frac{(x_h - \bar{x})^2 b^2}{S_{xx}}$

$= \left(\frac{1}{n} + \frac{(x_h - \bar{x})^2}{S_{xx}}\right) b^2$

$\Rightarrow \dfrac{\hat{\theta}_h - \theta_h}{\left[\frac{1}{n} + \frac{(x_h - \bar{x})}{S_{xx}}\right]^{\frac{1}{2}} b} \Big/ \left(\frac{(n-2)S^2}{b^2} / n-2\right) \sim t_{n-2}$

$\Rightarrow \hat{\theta}_h - \theta_h \sim N(0, (\frac{1}{n} + (x_h - \bar{x})^2 / S_{xx}) b^2)$

$\Rightarrow \dfrac{\hat{\theta}_h - \theta_h}{\left[\frac{1}{n} + \frac{(x_h - \bar{x})^2}{S_{xx}}\right]^{\frac{1}{2}} b} \Big/ \left(\frac{(n-2)S^2}{b^2} / n-2\right) \sim t_{n-2}$

$N(0,1) \sim$

$\Rightarrow \dfrac{\hat{\theta}_h - \theta_h}{S\left[\frac{1}{n} + \frac{(x_h - \bar{x})^2}{S_{xx}}\right]^{\frac{1}{2}}} \sim t_{n-2}$

$\Rightarrow$ CI of $\beta_0 + \beta_1 x_h = \theta_h = Ey_h$ is $\hat{y}_h \pm t_{\frac{\alpha}{2}, n-2} \cdot S\left[\frac{1}{n} + \frac{(x_h - \bar{x})^2}{S_{xx}}\right]^{\frac{1}{2}}$

$(\hat{y}_h = \hat{\theta}_h)$

Now, consider

Estimation: $y_h = \beta_0 + \beta_1 x_h + \varepsilon_h \quad \rightarrow E y_h = \beta_0 + \beta_1 x_h$

(Prediction) $\hat{y}_h = \hat{\beta}_0 + \hat{\beta}_1 x_h$ (average)

↓ uncertainity ↑

Prediction of y_h: new individual $\quad Var(\hat{\theta}_h) + \sigma^2$

$\Rightarrow$ Prediction Interval (PI) at x_h: $\hat{y}_h \pm t_{\frac{\alpha}{2}, n-2} \cdot S\left[1 + \frac{1}{n} + \frac{(x_h - \bar{x})}{S_{xx}}\right]^{\frac{1}{2}}$ ☆

Prediction: $y_{h(new)} = \beta_0 + \beta_1 x_h + \varepsilon_h$

$\hat{y}_{h(new)} = \hat{\beta}_0 + \hat{\beta}_1 x_h$

$\Rightarrow \dfrac{y_{h(new)} - \hat{\beta}_0 - \hat{\beta}_1 x_h}{S\left[1 + \frac{1}{n} + \frac{(x_h - \bar{x})^2}{S_{xx}}\right]^{\frac{1}{2}}} \sim t_{n-2}$

estimate (predict) $y_{h(new)} = \beta_0 + \beta_1 x_h + \varepsilon_h$

$\hat{y}_{h(new)} = \hat{\beta}_0 + \hat{\beta}_1 x_h$

$(y_{h(new)} - \hat{\beta}_0 - \hat{\beta}_1 x_h) \sim N(0, Var(\hat{\theta}_h) + \sigma^2) \Rightarrow \dfrac{y_{h(new)} - \hat{\beta}_0 - \hat{\beta}_1 x_h}{S\left[1 + \frac{1}{n} + \frac{(x_h - \bar{x})^2}{S_{xx}}\right]^{\frac{1}{2}}} \sim t_{n-2} \Rightarrow$ ☆

Estimation of the mean of the response variable for a given level of X (Estimate $E(Y_h)$)

Let X_h denote the level of X for which we wish to estimate the mean response [to be estimated by $\hat{Y}_h$]. (Note: Given X_h, the mean response is $E(Y_h) = \beta_0 + \beta_1 X_h$ according to the model)

For estimation, Given X_h

$$\hat{Y}_h = \hat{\beta}_0 + \hat{\beta}_1 X_h$$

Distribution of $E(Y_h) - \hat{Y}_h$:

$$E(Y_h) - \hat{Y}_h \sim N\left(0, \sigma^2 \left[\frac{1}{n} + \frac{(X_h - \overline{X})^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2}\right]\right)$$

special case of predict interval

Confidence Interval for $E(Y_h)$

Two-sided 100(1-α)% C.I. for $E(Y_h)$:

$$\left(\hat{Y}_h - t_{\alpha/2,n-2} s \sqrt{\frac{1}{n} + \frac{(X_h - \overline{X})^2}{\sum_{i=1}^{n} (X_i - \overline{X})^2}}, \quad \hat{Y}_h + t_{\alpha/2,n-2} s \sqrt{\frac{1}{n} + \frac{(X_h - \overline{X})^2}{\sum_{i=1}^{n} (X_i - \overline{X})^2}} \right)$$

Prediction of a new observation $Y_{h(new)}$

Example

Suppose you want to develop a model to predict selling price of homes based on assessed value. A sample of 30 recently sold single-family houses in a small city is selected to study the relationship between selling price (Y, in $000) and assessed value (X, in $000). The houses in the city had been reassessed at full value 1 year prior to the study.

Estimate the selling price of **an individual** house with an assessed value of $70,000.

Prediction of a new observation $Y_{h(new)}$

Prediction of $Y_{h(new)}$ corresponding to a given level X of the predictor variable by $\hat{Y}_h$.

$$\hat{Y}_h = \hat{\beta}_0 + \hat{\beta}_1 X_h$$

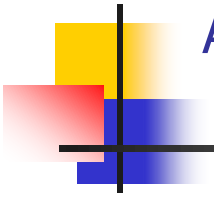
Distribution of $Y_{h(new)} - \hat{Y}_h$:

$$Y_{h(new)} - \hat{Y}_h \sim N\left(0, \sigma^2\left[1 + \frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right]\right)$$

Confidence Interval for $Y_{h(new)}$

Two-sided 100(1-α)% C.I. for $Y_{h(new)}$:

$$\left(\hat{Y}_h - t_{\alpha/2,n-2} s \sqrt{\left[1 + \frac{1}{n} + \frac{(X_h - \overline{X})^2}{\sum_{i=1}^{n} (X_i - \overline{X})^2} \right]}, \quad \hat{Y}_h + t_{\alpha/2,n-2} s \sqrt{\left[1 + \frac{1}{n} + \frac{(X_h - \overline{X})^2}{\sum_{i=1}^{n} (X_i - \overline{X})^2} \right]} \right)$$



Analysis of variance approach to regression analysis

- Partitioning of Total Sum of Squares
- Mean Squares
- Analysis of Variance (ANOVA) Table

Remark 2.5

Remark 2.5

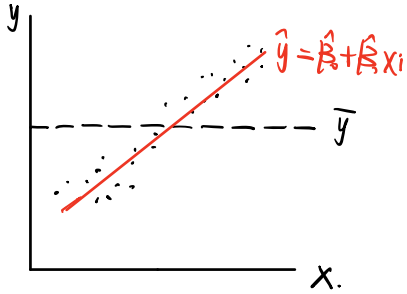
total variation of y_i (without considering the model)

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

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

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

i 개념설명



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

↑ SSE. SSR ← sum of squares due to regression

sum of squares of residuals ↑

↑ explained variation by the model

unexplained variation by the model

$R^2 = \frac{SSR}{SST}$ — % of the variation can be explained by the model.

Now, we want to prove that $\sum_{i=1}^{n} (y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) = 0$

* $\hat{y}_i - \bar{y} = \hat{\beta}_0 + \hat{\beta}_1 x_i - \bar{y}$

$= \bar{y} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 x_i - \bar{y}$

$= \hat{\beta}_1 (x_i - \bar{x})$

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

$= \hat{\beta}_1 \sum_{i=1}^{n} (y_i - \hat{y}_i)(x_i - \bar{x})$

$= \hat{\beta}_1 \sum_{i=1}^{n} x_i (y_i - \hat{y}_i) = 0$

$- \bar{x} \hat{\beta}_1 \sum_{i=1}^{n} (y_i - \hat{y}_n) = 0$ } two equations to calculate LSE

$p=2$

Sum of square:	SS	d.f. (degree of freedom)	MS.	F
Regression:	SSR	$p-1$	$SSR/p-1$	$MSR/MSE \sim F_{p-1,n-p}$
ERROR	SSE	$n-p$	$SSE/n-p$	
Total	SST	$n-1$		

$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2$ $\quad SSE = \sum (y_i - \hat{y}_i)^2$. $p=2$, β_0, β_1 $\quad SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$

H_0: the model with no independent variables fits the data as well as the model.

H_1: model fits the data better.

Analysis of variance approach to regression analysis

Total Variation = Unexplained Variation + Explained Variation

$$1 = \frac{\text{Unexplained Variation}}{\text{Total Variation}} + \frac{\text{Explained Variation}}{\text{Total Variation}}$$

$$\frac{\text{Explained Variation}}{\text{Total Variation}} = 1 - \frac{\text{Unexplained Variation}}{\text{Total Variation}}$$

Analysis of variance approach to regression analysis

ANOVA Table

	Sum of Squares (SS)	Degrees of freedom (df)	Mean squares (MS)	*F*
Regression	SSR	1	MSR = SSR/1	MSR/MSE
Error	SSE	n-2	MSE = SSE/(n-2)	
Total	SST	n-1		

H_0: $\beta_1 = 0$ (no linear relationship)
H_1: $\beta_1 \neq 0$ (linear relationship does exist between X and Y)

Test Statistics: $F = \text{MSR/MSE}$
Rejection Rule: reject the null hypothesis if $F > F_{(\alpha, 1, n-2)}$

Measures of linear association between X and Y

- Coefficient of Determination R^2

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

  b) $0 \leq R^2 \leq 1$

- Coefficient of Correlation

  a) $r = \pm\sqrt{R^2}$

  b) A plus or minus sign is attached to this measure according to whether the slope of the fitted regression line is positive or negative.

Ch2. Revision: $y = \beta_0 + \beta_1 x + \varepsilon$ (ε ⇐ random error) $\varepsilon \sim (0, b^2)$

response: y; model: $\beta_0 + \beta_1 x$; independent: x

Inference: estimate β_0, β_1, b^2, etc.

study how good they are.

Prediction: given a new x^*, what y^* would be?

} based on a data set $\{(y_i, x_i): i = 1, \cdots, n\}$

model for data

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad \varepsilon_i \sim N(0, b^2)$$

assumption:

1°. y_i and x_i have a linear relationship.

2° y_i's (or ε_i's) are independent for $i = 1, \cdots, n$.
还有变化来性

3°. $\mathrm{Var}(y_i) = \mathrm{Var}(\varepsilon_i) = b^2$.

4°. $y_i \sim N(\ ,\ )$ 服从正态分布.

~~Inference~~. Inference:

LSE \ MLE. if $\varepsilon_i \sim N(0, b^2)$ t-test. CI.

$\hat{\beta}_0 = ? \quad \sim N(\beta_0, ?)$

$\hat{\beta}_1 = ? \quad \sim N(\beta_1, ?)$ $H_0: \beta_1 = 0$ v.s $H_1: \beta_1 \neq 0$.

$\hat{b}^2 = ? \quad \sim \chi^2$

— How good is the model:

$R^2 = \frac{SSR}{SST}$ —% of the variation can be explained by the model.

— H_0: the model with no independent variables fits the data as well as the model

— H_1: the model with the independent variables fits the data better

Case I: $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$

$\Leftrightarrow H_0: \beta_1 = 0$ v.s $H_1: \beta_1 \neq 0$.

Case II: $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$

$\Leftrightarrow H_0: \beta_1 = 0$ and $\beta_2 = 0$ vs $H_1: \beta_j \neq 0$ for at least one of $\{j = 1, 2\}$

F-test: (overall test for the model ANOVA table.

prediction

— estimate the average of Y_h of X_h. $E(Y_h) = \beta_0 + \beta_1 X_h$.

prediction. $\hat{\beta}_0 + \hat{\beta}_1 X_h$.

C.I. of $E(y_h) = ?$

— estimate (predict) of a new observation $y_{h,new} = \beta_0 + \beta_1 X_h + \varepsilon_h$.

prediction. $\hat{\beta}_0 + \hat{\beta}_1 X_h$.

CI or PI of $y_{h,new}$