Least Squares Error Decomposition

C. Durso

Least Squares Line

Theorem

Given n pairs of numbers, $\{(x_1, y_1), (x_2, y_2), \ldots (x_n, y_n)\}$ and the model $y \approx mx + b$, the parameters m and b that minimize the sum of the squares of errors, $\Sigma_{i=1}^{n}\big(y_i - (mx_i + b)\big)^2$, are

- $m = \frac{\Sigma_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\Sigma_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\frac{1}{n}\Sigma_{i=1}^{n} y_i x_i - \bar{y}\bar{x}}{\frac{1}{n}\Sigma_{i=1}^{n} x_i^2 - \bar{x}^2}$

- $b = \bar{y} - m\bar{x}$, m as above

Mean y and Predicted y

Definitions

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

$\hat{y} = mx_i + b$

Square Differences

Definitions

$SSY = \Sigma(y_i - \bar{y_i})^2$, total variation in y

$SSE = \Sigma(y_i - \widehat{y_i})^2$, error sum of squares

$SSR = \Sigma(\widehat{y_i} - \bar{y})^2$, regression sum of squares

Sum of Square Differences

Theorem

$SSY = SSE + SSR$

Proof of $SSY = SSE + SSR$

$$SSY = \Sigma(y_i - \hat{y}_i + \hat{y}_i - \bar{y})^2$$

Proof of $SSY = SSE + SSR$

$$SSY = \Sigma(y_i - \hat{y}_i + \hat{y}_i - \bar{y})^2$$

$$= \Sigma(y_i - \hat{y}_i)^2 + \Sigma(\hat{y}_i - \bar{y})^2 + 2\Sigma(y_i - \hat{y}_i)(\hat{y}_i - \bar{y})$$

Proof of $SSY = SSE + SSR$

$$SSY = \Sigma(y_i - \hat{y}_i + \hat{y}_i - \bar{y})^2$$

$$= \Sigma(y_i - \hat{y}_i)^2 + \Sigma(\hat{y}_i - \bar{y})^2 + 2\Sigma(y_i - \hat{y}_i)(\hat{y}_i - \bar{y})$$

$$= SSE + SSR + 2\Sigma(y_i - \hat{y}_i)(\hat{y}_i - \bar{y})$$

Proof of $SSY = SSE + SSR$

$$SSY = \Sigma(y_i - \hat{y}_i + \hat{y}_i - \bar{y})^2$$

$$= \Sigma(y_i - \hat{y}_i)^2 \; + \Sigma(\hat{y}_i - \bar{y})^2 \; + 2\Sigma(y_i - \hat{y}_i)(\hat{y}_i - \bar{y})$$

$$= SSE + SSR + 2\Sigma(y_i - \hat{y}_i)(\hat{y}_i - \bar{y})$$

Done if $\Sigma(y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) = 0$.

Proof of $\Sigma(y_i - \hat{y})(\hat{y} - \bar{y}) = 0$

$\Sigma(y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) = \Sigma\big(y_i - (mx_i + b)\big)\big((mx_i + b) - \bar{y}\big)$.

Proof of $\Sigma(y_i - \hat{y})(\hat{y} - \bar{y}) = 0$

$\Sigma(y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) = \Sigma\big(y_i - (mx_i + b)\big)\big((mx_i + b) - \bar{y}\big)$.

$= \Sigma\big(y_i - (mx_i + \bar{y} - m\bar{x})\big)(mx_i + \bar{y} - m\bar{x} - \bar{y})$

because $b = \bar{y} - m\bar{x}$

Proof of $\Sigma(y_i - \hat{y})(\hat{y} - \bar{y}) = 0$

$\Sigma(y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) = \Sigma\big(y_i - (mx_i + b)\big)\big((mx_i + b) - \bar{y}\big)$.

$= \Sigma\big(y_i - (mx_i + \bar{y} - m\bar{x})\big)(mx_i + \bar{y} - m\bar{x} - \bar{y})$

because $b = \bar{y} - m\bar{x}$

$= \Sigma\big(y_i - \bar{y} - m(x_i - \bar{x})\big)(mx_i - m\bar{x})$

Proof of $\Sigma(y_i - \hat{y})(\hat{y} - \bar{y}) = 0$

$$\Sigma(y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) = \Sigma\big(y_i - (mx_i + b)\big)\big((mx_i + b) - \bar{y}\big).$$

$$= \Sigma\big(y_i - (mx_i + \bar{y} - m\bar{x})\big)(mx_i + \bar{y} - m\bar{x} - \bar{y})$$

because $b = \bar{y} - m\bar{x}$

$$= \Sigma\big(y_i - \bar{y} - m(x_i - \bar{x})\big)(mx_i - m\bar{x})$$

$$= m[(y_i - \bar{y})(x_i - \bar{x}) - m(x_i - \bar{x})^2]$$

Proof of $\Sigma(y_i - \hat{y})(\hat{y} - \bar{y}) = 0$

$$\Sigma(y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) = \Sigma\big(y_i - (mx_i + b)\big)\big((mx_i + b) - \bar{y}\big).$$

$$= \Sigma\big(y_i - (mx_i + \bar{y} - m\bar{x})\big)(mx_i + \bar{y} - m\bar{x} - \bar{y})$$

because $b = \bar{y} - m\bar{x}$

$$= \Sigma\big(y_i - \bar{y} - m(x_i - \bar{x})\big)(mx_i - m\bar{x})$$

$$= m[(y_i - \bar{y})(x_i - \bar{x}) - m(x_i - \bar{x})^2]$$

$$m\left[\sum(y_i - \bar{y})(x_i - \bar{x}) - \left(\frac{\Sigma(y_i - \bar{y})(x_i - \bar{x})}{\Sigma(x_i - \bar{x})^2}\right)\Sigma(x_i - \bar{x})^2\right] = 0$$

R^2

Definition

$$R^2 = \frac{SSY - SSE}{SSY} = \frac{SSR}{SSY}$$

- R^2 is the percent of the variability in y accounted for by the variability in $\hat{y}$.
- R^2 close to 1 shows strong linear relation between x and y.

$R^2 = cor(x, y)^2$

$$R^2 = \frac{\Sigma(\hat{y}_i - \bar{y})^2}{\Sigma(y_i - \bar{y})^2}$$

$$= \frac{\Sigma(mx_i + (\bar{y} - m\bar{x}) - \bar{y})^2}{\Sigma(y_i - \bar{y})^2}$$

$$= \frac{m^2\Sigma(x_i - \bar{x})^2}{\Sigma(y_i - \bar{y})^2}$$

$$= \frac{\left(\Sigma(y_i - \bar{y})(x_i - \bar{x})\right)^2 \Sigma(x_i - \bar{x})^2}{(\Sigma(x_i - \bar{x})^2)^2 \Sigma(y_i - \bar{y})^2}$$

$$= \frac{\left(\Sigma(y_i - \bar{y})(x_i - \bar{x})\right)^2}{\Sigma(x_i - \bar{x})^2 \Sigma(y_i - \bar{y})^2} = \left(\frac{\Sigma(y_i - \bar{y})(x_i - \bar{x})}{\sqrt{\Sigma(x_i - \bar{x})^2 \Sigma(y_i - \bar{y})^2}}\right)^2$$

Linear Model With Normal Error

C. Durso

Normal Errors

- Model: $y_i = mx_i + b + \varepsilon_i$ where the ε_i's are independent, identically distributed sample from $normal(0, \sigma^2)$.
- Find maximum likelihood estimators $\widehat{m}$, $\hat{b}$ for m, b (new notation).

Maximum Likelihood Estimators

Maximize $\log\left(\prod (2\pi\sigma^2)^{-1/2} \exp\left(-\frac{\left(y_i-(mx_i+b)\right)^2}{2\sigma^2}\right)\right) = c -$

$\frac{n}{2}\log(\sigma^2) - \sum \frac{\left(y_i-(mx_i+b)\right)^2}{2\sigma^2} = L(m,b,\sigma^2)$

- $\begin{cases} \frac{\partial}{\partial m} L(m,b,\sigma^2) = \frac{1}{\sigma^2}\sum\left(y_i - (mx_i+b)\right)x_i = 0 \\ \frac{\partial}{\partial b} L(m,b,\sigma^2) = \frac{1}{\sigma^2}\sum\left(y_i - (mx_i+b)\right) = 0 \end{cases}$ same maximizing values as least squares
- $\frac{\partial}{\partial \sigma^2} L(m,b,\sigma^2) = -\frac{1}{2}\left(\frac{n}{\sigma^2} - \sum\left(y_i - (mx_i+b)\right)^2 \frac{1}{(\sigma^2)^2}\right) = 0,$

  $\sigma^2 = \sum \frac{\left(y_i-(\hat{m}x_i+\hat{b})\right)^2}{n}$

Unbiased Error Variance

Definition

$$s^2 = {}^{SSE}\!/_{(n-2)}$$

$\sigma^2 \approx s^2 = {}^{SSE}\!/_{(n-2)}$ unbiased estimate

Inference for Linear Regression $\hat{m}$ and $\hat{b}$

C. Durso

Normal Errors

- Model: $y_i = mx_i + b + \varepsilon_i$ where the ε_i's are independent, identically distributed sample from $normal(0, \sigma^2)$.
- Find maximum likelihood estimators $\widehat{m}$, $\hat{b}$ for m, b (new notation).

Applications

Given the model assumptions:

- Is linear model significantly better than just predicting $y_i = \bar{y}$?
- What slopes are plausible?
- What intercepts are plausible?

Inference for Regression
Normal (Gaussian) Errors

- Is linear association significant?
- Null hypothesis: true $m = 0$
- Test: $pf\left({}^{SSR}/_{s^2}, 1, n-2\right), s^2 = \frac{SSE}{n-2}$
- Equivalent, generalizable test statistic:

$$\frac{\frac{SSY - SSE}{(n-1)-(n-2)}}{\frac{SSE}{n-2}}$$

- Numerator will be small if regression is useless.

Expected Value for $\widehat{m}$

$\widehat{m}$ distributed as $M = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\Sigma(X_i - \bar{X})^2}$

$$E[Y_i] = E[mX_i + b + \varepsilon_i] = mX_i + b$$

$$E[\bar{Y}] = E[m\bar{X} + b + \bar{\varepsilon}] = m\bar{X} + b$$

$$E[M] = \frac{\sum(X_i - \bar{X})\big(mX_i + b - (m\bar{X} + b\,)\big)}{\Sigma(X_i - \bar{X})^2}$$

$$E[M] = \frac{m\sum(X_i - \bar{X})^2}{\Sigma(X_i - \bar{X})^2} = m$$

Variance for $\widehat{m}$

$\widehat{m}$ distributed as $M = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\Sigma(X_i - \bar{X})^2}$

Numerator is $\Sigma(X_i - \bar{X})Y_i - \Sigma(X_i - \bar{X})\bar{Y} = \Sigma(X_i - \bar{X})Y_i$

$Var[Y_i] = Var[\varepsilon_i] = \sigma^2$

$Var[M] = \frac{\Sigma(X_i - \bar{X})^2 \sigma^2}{(\Sigma(X_i - \bar{X})^2)^2} = \frac{\sigma^2}{\Sigma(X_i - \bar{X})^2}$, estimate by $\frac{s^2}{\Sigma(X_i - \bar{X})^2}$

Inference for $\widehat{m}$

$$\frac{\widehat{m}-m}{\sqrt{\frac{s^2}{\Sigma\left(X_i-\bar{X}\right)^2}}} \sim \text{Student's t with } n-2 \text{ degrees of freedom}$$

Observations

- Y_i, Y_j independent for $i \neq j$
- $M, \bar{Y}$ independent:
  - $Cov\left[\Sigma(X_i - \bar{X})Y_i, Y_j\right] = \sigma^2\left(X_j - \bar{X}\right)$
  - $Cov[\Sigma(X_i - \bar{X})Y_i, \bar{Y}] = \Sigma\frac{\sigma^2}{n}\left(X_j - \bar{X}\right) = 0$
  - Jointly normally distributed random variables with covariance equal to 0 are independent.

Expect Value of $\hat{b}$

$\hat{b}$ distributed as $B = \bar{Y} - M\bar{X} = m\bar{X} + b + \frac{1}{n}\Sigma\varepsilon_i - M\bar{X}$

$$E[B] = E\left[(m - M)\bar{X} + \frac{1}{n}\Sigma\varepsilon_i + b\right] = b$$

$\hat{b}$ unbiased

Inference for $\hat{b}$

$\hat{b} = \bar{y} - \hat{m}\bar{x}$

$Var[\bar{Y} - M\bar{X}] = \frac{\sigma^2}{n} + \frac{\bar{X}^2}{\Sigma(X_i - \bar{X})^2}\sigma^2$ by independence

Estimate standard deviation by $s_b = \sqrt{s^2\left(\frac{1}{n} + \frac{\bar{X}^2}{\Sigma(X_i - \bar{X})}\right)}$

$\frac{\hat{b}-b}{s_b} \sim$ Student's t with $n-2$ degrees of freedom

Distributions

Theorem, version 1

Given n observations from $Y_i = mX_i + b + \varepsilon_i$ where $\varepsilon_i \sim normal(0, \sigma^2)$:

- $\hat{m} \sim normal\left(m, \frac{\sigma^2}{\Sigma(X_i - \bar{X})^2}\right)$
- $\hat{b} \sim normal\left(b, \sigma^2\left(\frac{1}{n} + \frac{\bar{X}^2}{\Sigma(X_i - \bar{X})^2}\right)\right)$
  - Estimation of σ by s results in Student's t distributions with $n - 2$ degrees of freedom

Inference for Linear Regression

New Observations

C. Durso

Application

- What is the expected value for a new observation at a particular X_h? How well do we know this?
- What range of values for a new observation at X_h are plausible?

Distributions

Theorem, version 1

Given n observations from $Y_i = mX_i + b + \varepsilon_i$ where $\varepsilon_i \sim normal(0, \sigma^2)$:

- $\hat{m} \sim normal\left(m, \frac{\sigma^2}{\Sigma(X_i - \bar{X})^2}\right)$
- $\hat{b} \sim normal\left(b, \sigma^2\left(\frac{1}{n} + \frac{\bar{X}^2}{\Sigma(X_i - \bar{X})^2}\right)\right)$
    - Estimation of σ by s results in Student's t distributions with $n - 2$ degrees of freedom

Given X_h Mean and Variance of $\hat{Y}_h$

$\hat{Y}_h = \hat{m}X_h + \hat{b}$ (estimated expected value)

$E[MX_h + B] = mX_h + b$

$Var[MX_h + B] = Var[MX_h + \bar{Y} - M\bar{X}] = Var[\bar{Y} + (X_h - \bar{X})M]$: Use independence of $\bar{Y}$ and M:

$$Var[MX_h + B] = \frac{\sigma^2}{n} + \frac{\sigma^2 (X_h - \bar{X})^2}{\Sigma(X_i - \bar{X})^2} \approx s^2 \left(\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\Sigma(X_i - \bar{X})^2}\right)$$

Distribution of $\hat{Y}_h$

Definition

$$s_{Y_h}^2 = s^2\left(\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\Sigma(X_i - \bar{X})^2}\right)$$

$\frac{\hat{Y}_h - E[Y_h]}{s_{Y_h}}$ has a Student's t distribution with $n - 2$ degrees of freedom

Distribution of New Observation $mX_h + b + \varepsilon_h$, Estimated

$$\hat{Y}_h new = \hat{Y}_h + \varepsilon_h$$

$$E[MX_h + B + \varepsilon_h] = mX_h + b$$

$$Var[MX_h + B + \varepsilon_h] = \frac{(n+1)\sigma^2}{n} + \frac{\sigma^2(X_h - \bar{X})^2}{\Sigma(X_i - \bar{X})^2}$$

$\frac{\hat{Y}_h new - E[Y_h]}{\sqrt{s_{Y_h}^2 + s^2}}$ has a Student's t distribution with $n - 2$ degrees of freedom

Distributions

Theorem

Given n observations from $Y_i = mX_i + b + \varepsilon_i$ where $\varepsilon_i \sim normal(0, \sigma^2)$:

- $\hat{m} \sim normal\left(m, \frac{\sigma^2}{\Sigma(X_i - \bar{X})^2}\right)$
- $\hat{b} \sim normal\left(b, \sigma^2\left(\frac{1}{n} + \frac{\bar{X}^2}{\Sigma(X_i - \bar{X})^2}\right)\right)$
- $\hat{Y}_h \sim normal\left(mX_h + b, \sigma^2\left(\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\Sigma(X_i - \bar{X})^2}\right)\right)$
- $\hat{Y}_h new \sim normal\left(mX_h + b, \sigma^2\left(\frac{n+1}{n} + \frac{(X_h - \bar{X})^2}{\Sigma(X_i - \bar{X})^2}\right)\right)$
  - Estimation of σ by s results in Student's t distributions with $n - 2$ degrees of freedom

Regression Diagnostics Overview

C. Durso

Checks After Model Fitting

- Check assumptions before model fit with scatter plot.
- Some assumptions checked after model fit (unlike one- and two-sample parametric tests).

Roles of Model Assumptions

- Model: $\boldsymbol{Y} = m\boldsymbol{X} + b + \boldsymbol{\varepsilon}$
- Least squares best fit line: no assumptions
- Least squares = maximum likelihood: $\varepsilon_1, \varepsilon_2, \ldots \varepsilon_n \; iid \; normal(0, \sigma^2)$
- Unbiased $\hat{m}$ and $\hat{b}$: $\varepsilon_1, \varepsilon_2, \ldots \varepsilon_n$ have mean 0, linear model is correct
- Student's t confidence intervals: $\varepsilon_1, \varepsilon_2, \ldots \varepsilon_n \; iid \; normal(0, \sigma^2)$ and linear model is correct

Check Assumptions

- Linearity of relationship between $\boldsymbol{X}$ and $\boldsymbol{Y}$ (plotting data, plotting standardized residuals against predictions)
- $\varepsilon_1, \varepsilon_2, \dots \varepsilon_n$ iid with mean 0 (plotting standardized residuals against predictions)
- $\varepsilon_1, \varepsilon_2, \dots \varepsilon_n$ $iid\ normal$ (usual normality tests, qq plotting…)

Robustness Issues

- Coefficients in linear regression may be sensitive to individual (x_i, y_i).
- Use *leverage* and *Cook's distance* to examine sensitivity.

Leverage

C. Durso

Diagnostic for Influential Observations

Definition

The leverage of the i^{th} case in regression with a single explanatory variable X equals $\frac{1}{n} + \frac{(X_i - \bar{X})^2}{\sum (X_j - \bar{X})^2}$

Measure of how far X_i is from mean $\bar{X}$. High leverage point is influential if it lies off the regression line.

Motivation

- Shows sensitivity of regression line to y_i.
- Regression line passes through $(\bar{x}, \bar{y})$.
- For x_i far from $\bar{x}$, small change in $\widehat{m}$ gives large change in $\hat{y}_i$.

Derivation

Leverage is $\frac{\partial \hat{y}_i}{\partial y_i}$:

$$\frac{\partial\left[\hat{m}x_i + \hat{b}\right]}{\partial y_i} = \frac{\partial[\hat{m}x_i + \bar{y} - \hat{m}\bar{x}]}{\partial y_i}$$

But $\frac{\partial \hat{m}}{\partial y_i} = \frac{\partial\left[\frac{\sum(x_j - \bar{x})(y_j - \bar{y})}{\sum(x_j - \bar{x})^2}\right]}{\partial y_i} = \left[\frac{(x_i - \bar{x}) - \frac{1}{n}\sum(x_j - \bar{x})}{\sum(x_j - \bar{x})^2}\right]$

$= \frac{(x_i - \bar{x})}{\sum(x_j - \bar{x})^2}$ because the summation cancels.

Derivation

Conclude

$$\frac{\partial[\hat{m}x_i+\bar{y}-\hat{m}\bar{x}]}{\partial y_i}=\frac{\partial[\bar{y}+\hat{m}(x_i-\bar{x})\,]}{\partial y_i}=\frac{1}{n}+\frac{(x_i-\bar{x})^2}{\sum\left(x_j-\bar{x}\right)^2}$$