Simple Linear Regression

Dustin Long, PhD

Department of Biostatistics
University of Alabama at Birmingham

September 3, 2019

Notes

Outline:

- Simple Linear Regression

Notes

Outline:

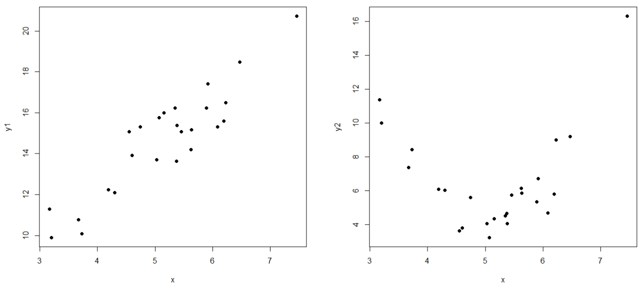
- Assumptions and properties of regression with one variable
- Determination and measures for the line of best fit
- Inference and Interpretations of parameters

Notes

Notes

- When do we use linear regression?
- What type of model do we use?
- What is the best fitting model, and what do we mean by "best fit"?

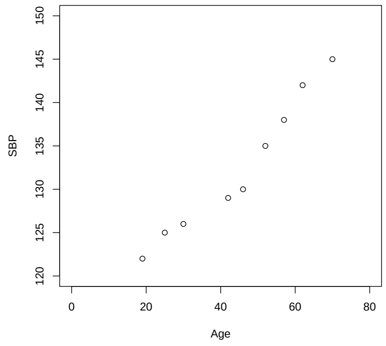
Notes



Notes

Example: SBP and Age

Obs	Age	SBP
1	19	122
2	25	125
3	30	126
4	42	129
5	46	130
6	52	135
7	57	138
8	62	142
9	70	145

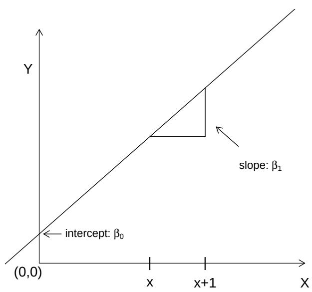


Notes

Mathematical Properties of a Straight Line

- A line is defined by an intercept and slope, i.e., $y = Mx + B$
- Recall properties of straight lines

Notes



Notes

Definition of Linear Regression

- The model used for simple linear regression is:
- $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$
- where Y_i is a dependent (outcome) random variable, X_i is an independent random variable, $i = 1 \ldots n$, β_0 is the population intercept, β_1 is the population slope, and $\epsilon_i \sim N(0, \sigma^2)$
- $\boldsymbol{Y} = \boldsymbol{X}\beta + \boldsymbol{e}$

Notes

Statistical Assumptions of a Straight Line

- Assumption 1: Homoscedasticity, $\sigma^2_x = \sigma^2 \ \ \forall x$
- Assumption 2: Independence, Y values are independent of one another
- Assumption 3: Linearity, $E[Y|X = x] = \mu_{Y|X=x} = \beta_0 + \beta_1 x$ or $Y = \beta_0 + \beta_1 x + E$
- Assumption 4: Existence, $(Y|X = x \sim f(\mu, \sigma^2_x)$, where $\mu, \sigma^2_x < \infty, \ \forall x$
- Assumption 5: Normality, $Y|X = x \sim N(\mu_{Y|X=x}, \sigma^2)$
-

Notes

Determining the Best-fitting Line

- Least Squares Method, OLS Example
- Another example
- Minimum-variance method, Best Linear Unbiased Estimators (BLUE)
- Under Assumptions 1-5, OLS estimates and BLUE are the same
- $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x}$
- $\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$

Notes

Least Squares Estimation

- Least squares estimators are values of β_0 and β_1 that minimize

$$\sum_{i=1}^{N}(Y_i - \beta_0 - \beta_1 X_i)^2$$

- Set partial derivatives equal to 0, solve for β_0 and β_1
- Or can set $\boldsymbol{Y} = \boldsymbol{X}\hat{\beta}$ and solve for $\hat{\beta}$

Notes

- When using hats, all estimated quantities get a hat
- WRONG: $Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$
- CORRECT: $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$
- Regression lines with out hats must have ϵ_i or it is just a line NOT regression
- $\hat{\epsilon}_i = \hat{\boldsymbol{e}}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$ and $\hat{\boldsymbol{e}}$ are all residuals
- $\sum_{i=1}^{n} \hat{\boldsymbol{e}}_i = 0$.

Notes

- *MSE*, mean square error, is the primary estimate of σ^2
- $SSE = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$
- $SSE = \hat{\boldsymbol{e}}'\hat{\boldsymbol{e}}$
- If $SSE = 0$ then the estimated regression line fits the data perfectly
- $MSE = SSE/(n-2)$ ONLY FOR SIMPLE LINEAR REGRESSION
- *SSE* and *MSE* are given directly from software

Notes

Inference about the Slope and Intercept

- Under Assumptions 1-5, $\hat{\beta}_0$ and $\hat{\beta}_1$ are Normally distributed
- To test $H_0 : \beta_1 = \beta_1^{(0)}$, the statistic is
- $T = \frac{\hat{\beta}_1 - \beta_1^{(0)}}{S_{\hat{\beta}_1}}$
- where $S_{\hat{\beta}_1} = \frac{\sqrt{MSE}}{\sqrt{(n-1)\sum_{i=1}^{n}(x_i - \bar{x})^2}}$
- $T \sim T_{n-2}$ under H_0
- C.I. for β_1 is constructed by inverting the previous test statistic
- Test for β_0 exists but rarely used

Notes

Interpretations of Tests

- Most common test for slope, $H_0 : \beta_1 = 0$
- If H_0 not rejected, it does NOT mean the there is no relationship between Y and X, just there is no evidence that the relationship is linear
- If H_0 is rejected, there is at minimum a linear relationship between Y and X but that might not be the entire story

Notes

Inferences about the regression line

- Recall that $\mu_{Y|X=x} = \beta_0 + \beta_1 x$, a particular point in the regression line
- Can test $H_0 : \mu_{Y|X=x} = \mu_{Y|X=x}^{(0)}$
- $S_{\hat{Y}_x} = S_{Y|X}\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{(n-1)S_x^2}}$
- This tests the value of the line at a single point, x
- Inverted C.I. called confidence bands when constructed for all observed values of X

Notes

- For prediction of $\hat{Y}_i$ at a particular value of X_i use
- $S_{\hat{Y}_x} = S_{Y|X}\sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{(n-1)S_x^2}}$
- to create prediction intervals and bands

Notes

ANOVA TABLE

Notes

Questions?

Notes