MM 225 – AI and Data Science

Day 24: Supervised Learning : Regression Analysis-2

Instructors: Hina Gokhale, MP Gururajan, N. Vishwanathan

1 OCTOBER 2024

Least Squares Estimators of β_0, β_1

- Let $(Y_i, x_i) : i = 1, 2, \ldots, n$ be the data.
- These can be expressed as $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i \; : \; i = 1, 2, \ldots, n$
- Want to find β_0, β_1 by minimizing the squared error between values of Y_i and its estimator $\beta_0 + \beta_1 x_i$
- Let us denote estimated value of β_0, β_1 as $\widehat{\beta_0}$ and $\widehat{\beta_1}$ respectively
- Want find β_0 and β_1 that would minimize sum of squares of deviations from the observations from the regression line:
- $SS = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_i)^2$

$$\frac{\partial SS}{\partial \beta_0} = -2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i) = 0$$

$$\frac{\partial SS}{\partial \beta_1} = -2\sum_{i=1}^{n} x_i(y_i - \beta_0 - \beta_1 x_i) = 0$$

Simplify these two equation leads to

$$\left.\begin{aligned} n\beta_0 + \beta_1 \sum_{i=1}^{n} x_i &= \sum_{i=1}^{n} y_i \\ \beta_0 \sum_{i=1}^{n} x_i + \beta_1 \sum_{i=1}^{n} x_i^2 &= \sum_{i=1}^{n} y_i x_i \end{aligned}\right\} \text{Least Squares Normal Equations}$$

Further simplification we will get

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

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

Notations

Let estimated value of Y_i and ϵ_i be denoted by $\hat{Y}_i$ and e_i for i =1, 2,…,n

$\widehat{Y_i} = \widehat{\beta_0} + \widehat{\beta_1} x_i$ and $e_i = Y_i - \widehat{Y_i}$

Sum of Squares of residuals = $SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2$

$$Sxx = \sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}$$

$$Sxy = \sum_{i=1}^{n} (x_i - \bar{x})(Y_i - \bar{Y}) = \sum_{i=1}^{n} x_i Y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} Y_i\right)}{n}$$

Properties of Estimated Residuals

- $SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2$
- It can be shown that E(SSE) = (n-2) σ^2
- Hence SSE/(n-2) is an unbiased estimator of σ^2 and it is denoted by

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2}{n-2}$$

- SSE can be simplified as

$$SSE = SST - \widehat{\beta_1} * Sxy$$

Where, $SST = Total\ corrected\ sum\ of\ squares = \sum_{i=1}^{n} (Y_i - \bar{Y}_i)^2$

Properties of Estimated Residuals

- *For i = 1, 2, …, n*
- $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ and $E(\epsilon_i) = 0$ and $Var(\epsilon_i) = \sigma^2$
- It implies that

$$E(Y_i) = \beta_0 + \beta_1 x_i$$

$$Var(Y_i) = \sigma^2$$

$$E\left(\widehat{\beta_1}\right) = \frac{\sum(x_i - \bar{x})E(Y_i)}{\sum x_i^2 - n\bar{x}^2} = \beta_1 \; and \; Var\left(\widehat{\beta_1}\right) = \frac{\sigma^2}{Sxx}$$

$$E\left(\widehat{\beta_0}\right) = \sum_{i=1}^{n} \frac{E(Y_i)}{n} - \bar{x}E(B) = \beta_0 \; and \; Var\left(\widehat{\beta_0}\right) = \sigma^2\left[\frac{1}{n} + \frac{\bar{x}^2}{Sxx}\right]$$

Properties of LS Estimators

A is an unbiased estimator of β_0

B is an unbiased estimator of β_1

It can be shown that Cov(A, B) = $-\frac{\sigma^2 \bar{x}}{Sxx}$

Standard Errors of estimator of intercept and slope are respectively

$$SE(\widehat{\beta_0}) = \sqrt{\sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{Sxx}\right]}$$

$$SE(\widehat{\beta_1}) = \sqrt{\frac{\sigma^2}{Sxx}}$$

Estimated std error for $\widehat{\beta_0}$ and $\widehat{\beta_1}$ can be obtained by replacing σ^2 by its unbiased estimate $\hat{\sigma}^2$

Testing Hypothesis on regression parameters

$Y = \beta_0 + \beta_1 x + \epsilon$ and β_0 and β_1 are estimated as $\widehat{\beta_0}$ and $\widehat{\beta_1}$.

How do we know that statistically this relationship is significant?

If β_1 = 0 then this implies that Y is not dependent on x!

Therefore, it is of interest to test the hypothesis

$$H_0\colon\ \beta_1 = 0$$

In general it would be of interest to test the hypothesis that

$$H_0\colon\ \beta_1 = \beta_{1,0}$$

To statistically test the hypothesis an additional assumption needs to be made:

$\boldsymbol{\epsilon \sim N(o, \sigma^2)}$

Distribution of the regression estimators

$\boldsymbol{\epsilon \sim N\left(o, \sigma^2\right)}$

Hence $Y \sim N(\beta_0 + \beta_1 x, \boldsymbol{\sigma}^2)$

Estimator $\widehat{\beta_1} = \frac{\sum(x_i - \bar{x})Y_i}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}$ is a linear combination of independent RV Y_i

Hence $\widehat{\beta_1} \sim N\left(\beta_1, \frac{\sigma^2}{Sxx}\right)$

Similarly $\widehat{\beta_0} \sim N\left(\beta_0, \sigma^2\left[\frac{1}{n} + \frac{\bar{x}^2}{Sxx}\right]\right)$

And $\frac{(n-2)\hat{\sigma}^2}{\sigma^2} \sim \chi_{n-2}^2$

The test statistic for testing $\beta_1 = \beta_{1,0}$

Unbiased estimator of β_1 is B and estimated SE(B) = $\sqrt{\frac{\hat{\sigma}^2}{Sxx}}$

Hence the test statistic for testing $\beta_1 = \beta_{1,0}$ is

$$T = \frac{\widehat{\beta_1} - \beta_{1,0}}{\sqrt{\frac{\hat{\sigma}^2}{Sxx}}}$$

Hence when H_0 is true: T$\sim t(n-2)$

Critical region for testing $\beta_1 = \beta_{1,0}$

Alternative Hypothesis	Critical region for given α
$\boldsymbol{\beta_1 \neq \beta_{1,0}}$: Two sided alternative	$\left\{\frac{\widehat{\beta_1} - \boldsymbol{\beta_{1,0}}}{\sqrt{\frac{\widehat{\boldsymbol{\sigma}}^2}{\boldsymbol{Sxx}}}} < \boldsymbol{t}_{\alpha/2}(\boldsymbol{n}-2)\right\} \cup \left\{\frac{\widehat{\beta_1} - \boldsymbol{\beta_{1,0}}}{\sqrt{\frac{\widehat{\boldsymbol{\sigma}}^2}{\boldsymbol{Sxx}}}} > \boldsymbol{t}_{1-\alpha/2}(\boldsymbol{n}-2)\right\}$
$\boldsymbol{\beta_1 < \beta_{1,0}}$: one sided alternative	$\left\{\frac{\widehat{\beta_1} - \boldsymbol{\beta_{1,0}}}{\sqrt{\frac{\widehat{\boldsymbol{\sigma}}^2}{\boldsymbol{Sxx}}}} < \boldsymbol{t}_{\alpha}(\boldsymbol{n}-2)\right\}$
$\boldsymbol{\beta_1 > \beta_{1,0}}$: one sided alternative	$\left\{\frac{\widehat{\beta_1} - \boldsymbol{\beta_{1,0}}}{\sqrt{\frac{\widehat{\boldsymbol{\sigma}}^2}{\boldsymbol{Sxx}}}} > \boldsymbol{t}_{1-\alpha}(\boldsymbol{n}-2)\right\}$

The test statistic for testing $\beta_0 = \beta_{0,0}$

Unbiased estimator of β_0 is A and estimated SE(B) = $\sqrt{\hat{\sigma}^2\left[\frac{1}{n}+\frac{\bar{x}^2}{Sxx}\right]}$

Hence the test statistic for testing $\beta_0 = \beta_{0,0}$ is

$$T = \frac{\widehat{\beta_0} - \beta_{0,0}}{\sqrt{\hat{\sigma}^2\left[\frac{1}{n}+\frac{\bar{x}^2}{Sxx}\right]}}$$

Hence when H_0 is true: T$\sim t(n-2)$

Critical region for testing $\beta_0 = \beta_{0,0}$

Alternative Hypothesis	Critical region for given α
$\boldsymbol{\beta_0 \neq \beta_{0,0}}$: Two sided alternative	$\left\{ \dfrac{\widehat{\beta_0} - \boldsymbol{\beta_{0,0}}}{\sqrt{\boldsymbol{\hat{\sigma}^2}\left[\dfrac{\mathbf{1}}{\boldsymbol{n}} + \dfrac{\boldsymbol{\bar{x}^2}}{\boldsymbol{Sxx}}\right]}} < \boldsymbol{t_{\alpha/2}(n-2)} \right\} \cup \left\{ \dfrac{\widehat{\beta_0} - \boldsymbol{\beta_{0,0}}}{\sqrt{\boldsymbol{\hat{\sigma}^2}\left[\dfrac{\mathbf{1}}{\boldsymbol{n}} + \dfrac{\boldsymbol{\bar{x}^2}}{\boldsymbol{Sxx}}\right]}} > \boldsymbol{t_{1-\alpha/2}(n-2)} \right\}$
$\boldsymbol{\beta_0 < \beta_{0,0}}$: one sided alternative	$\left\{ \dfrac{\widehat{\beta_0} - \boldsymbol{\beta_{0,0}}}{\sqrt{\boldsymbol{\hat{\sigma}^2}\left[\dfrac{\mathbf{1}}{\boldsymbol{n}} + \dfrac{\boldsymbol{\bar{x}^2}}{\boldsymbol{Sxx}}\right]}} < \boldsymbol{t_{\alpha}(n-2)} \right\}$
$\boldsymbol{\beta_0 > \beta_{0,0}}$: one sided alternative	$\left\{ \dfrac{\widehat{\beta_0} - \boldsymbol{\beta_{0,0}}}{\sqrt{\boldsymbol{\hat{\sigma}^2}\left[\dfrac{\mathbf{1}}{\boldsymbol{n}} + \dfrac{\boldsymbol{\bar{x}^2}}{\boldsymbol{Sxx}}\right]}} > \boldsymbol{t_{1-\alpha}(n-2)} \right\}$

Prediction

Suppose new value Y_{n+1} is to be predicted when $x = x_{n+1}$

Then the point estimator Y_{n+1} can be given by $\hat{Y}_{n+1} = \hat{\beta}_0 + \hat{\beta}_1 x_{n+1}$

Error in prediction $e_p = Y_{n+1} - \hat{Y}_{n+1}$

Note that

$$E(e_p) = E(Y_{n+1} - \hat{Y}_{n+1}) = 0$$

$$Var\left(\hat{Y}_{n+1} = \hat{\beta}_0 + \hat{\beta}_1 x_{n+1}\right) = \sigma^2 \left[\frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{Sxx}\right]$$

$$Var(Y_{n+1}) = \sigma^2$$

Also note that Y_{n+1} refers to the future observation while $\hat{Y}_{n+1}$ is estimated from the model developed. Hence, Y_{n+1} and $\hat{Y}_{n+1}$ are independent.

Therefore: $Var\left(e_p\right) = \sigma^2 \left[1 + \frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{Sxx}\right]$

Prediction Interval

Thus we have

$$Y - \hat{\beta}_0 - \hat{\beta}_1 x_{n+1} \sim N\left(0, \sigma^2\left[1 + \frac{1}{n} + \frac{(\bar{x} - x_{n+1})^2}{Sxx}\right]\right)$$

And hence $\dfrac{Y_{n+1} - \hat{Y}_{n+1}}{\sqrt{\hat{\sigma}^2\left[1+\frac{1}{n}+\frac{(\bar{x}-x_{n+1})^2}{Sxx}\right]}} \sim t(n-2)$

Therefore prediction interval at 100(1-α)% confidence level is

$$\hat{y}_{n+1} - t_{n-2,\alpha/2}\sqrt{\hat{\sigma}^2\left[1 + \frac{1}{n} + \frac{(\bar{x} - x_{n+1})^2}{Sxx}\right]} \le Y_{n+1}$$

$$\le \hat{y}_{n+1} + t_{n-2,\alpha/2}\sqrt{\hat{\sigma}^2\left[1 + \frac{1}{n} + \frac{(\bar{x} - x_{n+1})^2}{Sxx}\right]}$$

Summary

Statistical properties of estimated errors = residuals

Statistical properties of least squares estimators

Testing of Hypothesis for regression coefficients

Prediction and Prediction Interval

Thank you….

Difference between mean value prediction and Prediction of future value of Y

Suppose we are interested in predicting Y for given x_{n+1} say $Y(x_{n+1})$

Difference between mean response $\beta_0 + \beta_1 x_0$ and $Y(x_{n+1})$

- Example: let x_0 be temperature and Y be response to an experiment carried out at temperature x_0, then
  - When several experiments are carried out at a given x_0, then expected value would be mean value of $\beta_0+\beta_1 x_0$
  - However, if only one experiment is carried out Y will be only one response….Present case relates to this possibility