STAC67: Regression Analysis

Lecture 11

Sohee Kang

Feb. 24, 2021

Outline of the Lecture

- Write the CI and PI in terms of matrix notation
- Re-write the Sum of Squares in ANOVA with quadratic forms
- Prove that MSE is the unbiased estimator of σ^2
- Introduce the adjusted R^2

- The prediction of a new observation

$$Y_0 = \underset{\sim}{x}_0'\underset{\sim}{\beta} + \epsilon_0$$

for a vector of values of independent variables $\underset{\sim}{x}_0$ is $\begin{bmatrix} 1 \\ x_1 \\ x_2 \\ \vdots \\ x_p \end{bmatrix}$

$$\widehat{Y}_{pred_0} = \underset{\sim}{x}_0'\underset{\sim}{\widehat{\beta}}$$

- The mean and variance of $\widehat{Y}_{pred_0}$ are

$E(\hat{Y}_{pred}) = X_0' B$ $\quad$ $Var(\hat{Y}_{pred}) = (\underbrace{1}_{\text{accomadating for } \varepsilon_0} + X_0(X'X)^{-1}X_0')\sigma^2$

Moreover, when $\underset{\sim}{\epsilon} \sim N(\underset{\sim}{0}, \sigma^2 \boldsymbol{I})$ then

$$\hat{Y}_{pred} \sim N(X_0'B,\ (1 + X_0(X'X)^{-1}X_0')\sigma^2)$$

Properties of a prediction $\widehat{Y}_{pred_0}$

- The variance of a prediction for a new observation is larger than the variance of the estimator of the mean response even though the point estimate is the same. That is, for a vector of values of predictor variables $\underset{\sim}{x}_0$
  - Prediction:

$$\widehat{Y}_{pred_0} = \underset{\sim}{x}_0' \underset{\sim}{\widehat{\beta}} \quad \text{with variance} \quad [1 + \underset{\sim}{x}_0'(\boldsymbol{X}'\boldsymbol{X})^{-1}\underset{\sim}{x}_0]\sigma^2$$

  - Mean response:

$$\widehat{Y}_0 = \underset{\sim}{x}_0' \underset{\sim}{\widehat{\beta}} \quad \text{with variance} \quad \underset{\sim}{x}_0'(\boldsymbol{X}'\boldsymbol{X})^{-1}\underset{\sim}{x}_0\sigma^2$$

$\sim N(x_0' \hat{\beta} \ , \ x_0 (X'X)^{-1} x_0' \sigma^2)$

Estimates and precision: summary

Consider the model $\underset{\sim}{Y} = \boldsymbol{X}\underset{\sim}{\beta} + \epsilon$, with $E(\underset{\sim}{\epsilon}) = \underset{\sim}{0}$, $Var(\underset{\sim}{\epsilon}) = \boldsymbol{I}\sigma^2$

Quantity	Estimator	Variance of the estimator
$\underset{\sim}{\beta}$	$\hat{B}$	$(X'X)^{-1}\sigma^2$
$E(\underset{\sim}{Y})$	$\hat{Y} = HY$	$\sigma^2 H$
$\underset{\sim}{\epsilon}$	$e = (I-H)Y$	$(I-H)\sigma^2$
Y_0	$X_0'\hat{B}$	$(1 + X_0'(X'X)^{-1}X_0)\sigma^2$

Analysis of Variance and Quadratic Forms

- Quadratic forms of $\underset{\sim}{Y}$: $\underset{\sim}{Y}'\boldsymbol{A}\underset{\sim}{Y}$, where $\boldsymbol{A}$ is a symmetric matrix of coefficients called defining matrix
- Next section: Study the properties of residual, regression and total sum of squares and sum of squares used in inference
- They are all quadratic forms of $\underset{\sim}{Y}$

Partitioning of total sum of squares

- We know that

$$\underset{\sim}{Y} = \underset{\sim}{\widehat{Y}} + \underset{\sim}{e}$$

- We will generalize the partitioning of the total sum of squares that we had for simple linear regression, i.e.

$$SST = SSR + SSE$$

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

to multiple linear regression.

Total sum of squares

- $SST = \sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2$ in matrix notation:

Exercise

Show that $SST = \underset{\sim}{Y}'\underset{\sim}{Y} - \frac{1}{n}\underset{\sim}{Y}'\boldsymbol{J}\underset{\sim}{Y}$, where $\boldsymbol{J}$ is the $n \times n$ square matrix with all elements equal to 1.

$$SST = \sum Y_i^2 - n\bar{Y}^2$$
$$= \sum Y_i^2 - n\left(\frac{1}{n}\sum Y_i\right)^2$$
$$= Y'Y - \frac{1}{n}Y'JY$$

- SST is a quadratic form of $\underset{\sim}{Y}$ because

$$Y'\underbrace{\left(I - \frac{1}{n}J\right)}_{\text{is symmetric}}Y$$

- The defining matrix associated is

$$I - \frac{1}{n}J$$

Residual sum of squares

- $SSE = \sum_{i=1}^{n} e_i^2$ in matrix notation:

Exercise

Show that $SSE = \underset{\sim}{Y}'\underset{\sim}{Y} - \hat{\underset{\sim}{\beta}}'\boldsymbol{X}'\underset{\sim}{Y}$

$$= \sum e_i^2$$

$$= e'e$$

$$= [Y - X\hat{\beta}]'[Y - X\hat{\beta}]$$

$$= Y'Y - \underbrace{Y'X\hat{\beta} - \hat{\beta}'X'Y + \hat{\beta}'X'X\hat{\beta}}_{Y'XB \text{ (show by } B' = (X'X)^{-1}X'Y)} = Y'Y - \hat{\beta}'X'Y$$

- SSE is a quadratic forms of $\underset{\sim}{Y}$ because

$$SSE = \underset{\sim}{Y}'(\boldsymbol{I} - \boldsymbol{H})\underset{\sim}{Y}$$

- The defining matrix is $\boldsymbol{I} - \boldsymbol{H}$.

Regression sum of squares

- $SSR = \sum_{i=1}^{n}(\widehat{Y}_i - \overline{Y})^2$ in matrix notation:

Exercise

Show that $SSR = \underset{\sim}{\widehat{\beta}}'\boldsymbol{X}'\underset{\sim}{Y} - \frac{1}{n}\underset{\sim}{Y}'\boldsymbol{J}\underset{\sim}{Y}$

$= \sum \hat{Y}_i^2 - n\bar{Y}^2$

$= \hat{Y}'\hat{Y} - \frac{1}{n}(\sum Y_i)^2$

$= \hat{\beta}'X'X\hat{\beta} - \frac{1}{n}Y'JY$

$= \hat{\beta}'X'\cancel{X(X'X)^{-1}}X'Y - \frac{1}{n}Y'JY$

$= \hat{\beta}'X'Y - \frac{1}{n}Y'JY$

Regression sum of squares

- SSR is a quadratic forms of $\underset{\sim}{Y}$ because
  $Y'X(X'X)^{-1}X'Y - \frac{1}{n}Y'JY = Y'(H - \frac{1}{n}J)Y$
- The defining matrix associated is: $H - \frac{1}{n}J$

Exercise

Check that

$$SST = SSR + SSE$$

$$Y'(I - \frac{1}{n}J)Y = Y'(H - \frac{1}{n}J)Y + Y'(I - H)Y$$

$$= Y'(I - \frac{1}{n}J)Y$$

Degrees of freedom

- For now: the number of values in the calculation of a statistic that can freely vary.
- SST has $n-1$ degrees of freedom
- SSE has $n-(p+1)$ degrees of freedom
- SSR has $(p+1)-1$ degrees of freedom.

Mean squares

- Mean squares: sum of squares divided by its associated degrees of freedom
- Regression mean squares:

$$MSR = \frac{SSR}{p}$$

- Residual mean squares:

$$MSE = \frac{SSE}{n-(p+1)}$$

Analysis of variance table

- Analysis of variance (ANOVA) table to display the sum of squares and degrees of freedom

Source of variation	Sum of squares	df	Mean squares
Regression	SSR	$(p+1)-1$	$MSR = \frac{SSR}{(p+1)-1}$
Residual	SSE	$n-(p+1)$	$MSE = \frac{SSE}{n-(p+1)}$
Total	SST	$n-1$	

$F = \frac{MSR}{MSE}$

$\sim F((p+1)-1, n-(p+1))$

- The results in the ANOVA table will be used to construct a global test for the regression coefficients.

$H_0: B_1 = B_2 = \cdots = B_p = 0$

$H_1:$ at least one $B_i \neq 0$

Properties of a quadratic form of a random vector

$Var(e) = (I-H)\sigma^2$

$E(e) = 0$

- Consider a quadratic form

$HY = XB$

$$U = \underset{\sim}{Z}'\boldsymbol{A}\underset{\sim}{Z}$$

of a random vector $\underset{\sim}{Z}$ where $\boldsymbol{A}$ is a symmetric matrix (the defining matrix). We have

$$E(\underset{\sim}{Z}'\boldsymbol{A}\underset{\sim}{Z}) = tr[A\,Var(z)] + E(z)'A\,E(z)$$

$$Var(\underset{\sim}{Z}'\boldsymbol{A}\underset{\sim}{Z}) = 2\,tr(A\,Var(z)\,A\,Var(z)) + 4\,E(z)'A\,Var(z)\,A\,E(z)$$

Unbiased estimator of σ^2

Exercise: Show that $E(\underset{\sim}{e}'\underset{\sim}{e}) = (n - p')\sigma^2$.

Hint: use that $tr(\mathbb{H}) = p'$ (without proof) and the quadratic formulation of $\underset{\sim}{e}'\underset{\sim}{e}$.

H

$$E(e'e) = tr(Var(e)) + E(e)'E(e)$$
$$= tr((I-H)\sigma^2) + 0$$
$$= \sigma^2(tr(I) - tr(p)) = \sigma^2(n-p') \quad p' = p+1$$

Exercise:Show that the estimator

$$MSE = \quad s^2 = \frac{\underset{\sim}{e}'\underset{\sim}{e}}{n - p'}$$

is an unbiased estimator of σ^2.

$$E(e'e/(n-p')) = \frac{1}{n-p'} \cdot (n-p')\sigma^2 = \sigma^2$$

Thus is an unbiased estimator.

Coefficient of multiple determination

$$R^2 = \frac{SSR}{SST}$$

- Fraction of the variation in Y explained by the model (i.e. its linear relationship with $X_1, \ldots, X_p$)
- We have $0 \leq R^2 \leq 1$

Exercise Show that

1. $R^2 = 0$ when $\widehat{\beta}_k = 0$ for each $k = 1, \ldots, p$.
2. $R^2 = 1$ when $\widehat{Y}_i = Y_i$ for each $i = 1, \ldots, n$, i.e. when all the observations fall on the fitted regression surface.

Adjusted coefficient of multiple determination}

- Adding more X to the model R^2.
- Adjusted R^2: modified measure that accounts for the number of variables in the model.

- Adjusted coefficient of multiple determination:

$$R^2_{adj} = 1 - \frac{\frac{SSE}{n-p'}}{\frac{SST}{n-1}} = 1 - \left(\frac{n-1}{n-p'}\right)\frac{SSE}{SST}$$

  - R^2_{adj} does not have the same interpretation as R^2.
  - R^2_{adj} may decrease when we add a new variable because

  - R^2_{adj} useful for selecting explanatory variables.