Chapter 2 Multiple Regression (Part 2)

1 Analysis of Variance in multiple linear regression

Recall the model again

$$Y_i = \underbrace{\beta_0 + \beta_1 X_{i1} + ... + \beta_p X_{ip}}_{\text{predictable}} + \underbrace{\varepsilon_i}_{\text{unpredictable}}, \quad i = 1, ..., n$$

For the fitted model $\hat{Y}_i = b_0 + b_1 X_{i1} + ... + b_p X_{ip}$,

$$Y_i = \hat{Y}_i + e_i \quad i = 1, ..., n$$

$$\underbrace{Y_i - \bar{Y}}_{\text{Total deviation}} = \underbrace{\hat{Y}_i - \bar{Y}}_{\substack{\text{Deviation} \\ \text{due the regression}}} + \underbrace{e_i}_{\substack{\text{Deviation} \\ \text{due to the error}}}$$

obs	deviation of Y_i	deviation of $\hat{Y}_i = b_0 + b_1 X_{i1} + ... + b_p X_{ip}$	deviation of $e_i = Y_i - \hat{Y}_i$
1	$Y_1 - \bar{Y}$	$\hat{Y}_1 - \bar{Y}$	$e_1 - \bar{e} = e_1$
2	$Y_2 - \bar{Y}$	$\hat{Y}_2 - \bar{Y}$	$e_2 - \bar{e} = e_2$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	$Y_n - \bar{Y}$	$\hat{Y}_n - \bar{Y}$	$e_n - \bar{e} = e_n$
Sum of squares	$\sum_{i=1}^n (Y_i - \bar{Y})^2$ Total Sum of squares (SST)	$\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2$ Sum of squares due to regression (SSR)	$\sum_{i=1}^n e_i^2$ Sum of squares of error/residuals (SSE)

We have

$$\underbrace{\sum_{i=1}^n (Y_i - \bar{Y})^2}_{\text{SST}} = \underbrace{\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2}_{\text{SSR}} + \underbrace{\sum_{i=1}^n e_i^2}_{\text{SSE}}$$

[Proof:

$$\begin{aligned}
\sum_{i=1}^{n}(Y_i - \bar{Y})^2 &= \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y} + Y_i - \hat{Y}_i)^2 \\
&= \sum_{i=1}^{n}\{(\hat{Y}_i - \bar{Y})^2 + (Y_i - \hat{Y}_i)^2 + 2(\hat{Y}_i - \bar{Y})(Y_i - \hat{Y}_i)\} \\
&= SSR + SSE + 2\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})(Y_i - \hat{Y}_i) \\
&= SSR + SSE + 2\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})e_i \\
&= SSR + SSE
\end{aligned}$$

where $\sum_{i=1}^{n} \hat{Y}_i e_i = 0$ and $\sum_{i=1}^{n} e_i = 0$ are used, which follow from the Normal equations.]

- $$SST = \sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \mathbf{Y}^{'}\mathbf{Y} - \frac{\mathbf{1}}{\mathbf{n}}\mathbf{Y}^{'}\mathbf{J}\mathbf{Y} = \mathbf{Y}^{'}(\mathbf{I} - \frac{\mathbf{1}}{\mathbf{n}}\mathbf{J})\mathbf{Y}$$

  Degree of freedom? **n-1** (with n being the number of observations)

- $$SSE = \sum_{i=1}^{n} e_i^2 = \mathbf{e}^{'}\mathbf{e} = (\mathbf{Y} - \mathbf{X}\mathbf{b})^{'}(\mathbf{Y} - \mathbf{X}\mathbf{b}) = \mathbf{Y}^{'}(\mathbf{I} - \mathbf{H})\mathbf{Y}$$

  Degree of freedom? **n-p-1** (with p+1 being the number of coefficients)

- Let $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{X}$ and and $\mathbf{J} = \mathbf{1}\mathbf{1}'/\mathbf{n}$. Note that

  $$\hat{\mathbf{Y}} = \mathbf{H}\mathbf{Y}$$

  and by the fact $\sum_{i=1}^{n} e_i = 0$ (see the normal equations),

  $$\bar{\hat{Y}} = \bar{Y} = \mathbf{1}'\mathbf{Y}/n.$$

  Thus

  $$\begin{aligned}
  SSR &= (\hat{\mathbf{Y}} - \bar{Y})' * (\hat{\mathbf{Y}} - \bar{Y}) = \mathbf{Y}'(\mathbf{H} - \mathbf{J}/n)'(\mathbf{H} - \mathbf{J}/n)'\mathbf{Y} \\
  &= \mathbf{Y}'(\mathbf{H} - \mathbf{J}/n)\mathbf{Y}.
  \end{aligned}$$

  Degree of freedom? **p** (the number of variables).

  [Another Proof:[1]

  $$\hat{\mathbf{Y}} - \bar{Y} = \mathbf{H}\mathbf{Y} - \mathbf{1}'/\mathbf{n}\mathbf{Y} = (\mathbf{H} - \mathbf{J}/\mathbf{n})\mathbf{Y}.$$

[1] please ignore this proof

Write $\mathbf{X} = (\mathbf{1} \vdots \mathbf{X_1})$. Then

$$H(\mathbf{1} \vdots \mathbf{X_1}) = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-\mathbf{1}}\mathbf{X}^\top\mathbf{X} = \mathbf{X} = (\mathbf{1} \vdots \mathbf{X_1})^\top$$

Thus

$$H(\mathbf{1} \vdots \mathbf{X_1}) = \mathbf{1}$$

Similarly, $\mathbf{1}'\mathbf{H} = \mathbf{1}'$. Thus

$$(\mathbf{H} - \mathbf{J}/n)'(\mathbf{H} - \mathbf{J}/n)' = \mathbf{H} - \mathbf{J}/n\mathbf{H} - \mathbf{HJ}/n + \mathbf{J}/n = \mathbf{H} - \mathbf{J}/n$$

]

- It follows that

$$SST = SSR + SSE$$

We further define

$$MSR = \frac{SSR}{p} \qquad \text{called } \textbf{regression mean square}$$

$$MSE = \frac{SSE}{n-p-1} \qquad \text{called } \textbf{error mean square (or mean squared error)}$$

2 ANOVA table

Source of Variation	SS	df	MS	F-statistic
Regression	$SSR = \mathbf{Y}'(\mathbf{H} - \mathbf{J}/n)\mathbf{Y}$	p	$MSR = \frac{SSR}{p}$	MSR/MSE
Error	$SSE = \mathbf{Y}'(\mathbf{I} - \mathbf{H})\mathbf{Y}$	$n-p-1$	$MSE = \frac{SSE}{n-p-1}$	
Total	$SST = \mathbf{Y}'(\mathbf{I} - \mathbf{J}/n)\mathbf{Y}$	$n-1$		

3 F test for regression relation

- $H_0 : \beta_1 = \beta_2 = ... = \beta_p = 0$ versus H_a : not all $\beta_k (k = 1, ..., p)$ equal zero
- Under H_0, the reduced model: $Y_i = \beta_0 + \varepsilon_i$

$$SSE(R) = SST = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$$

degrees of freedom $n-1$

- Full model: $Y_i = \beta_0 + \beta_1 X_{i1} + ... + \beta_p X_{ip} + \varepsilon_i$

$$SSE(F) = SSE = e^{'}e = (\mathbf{Y} - \mathbf{X}b)^{'}(\mathbf{Y} - \mathbf{X}b)$$

  degrees of freedom $n - p - 1$

- F test statistic (also called F-test for the model)

$$F^* = \frac{(SSE(R) - SSE(F))/(df(R) - df(F))}{SSE(F)/df(F)} = \frac{SSR/p}{SSE/(n-p-1)}$$

- If $F^* \leq F(1-\alpha;\ p,\ n-p-1)$, conclude(accept) H_0

  IF $F^* > F(1-\alpha;\ p,\ n-p-1)$, conclude H_a (reject H_0)

4 R^2 and the adjusted R^2

- $SSR = SST - SSE$ is the part of variation explained by regression model

- Thus, define coefficient of multiple determination

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

  which is the **proportion of variation in the response that can be explained by the regression model (or that can be explained by the predictors** $X_1, ..., X_p$ **linearly)**

- $0 \leq R^2 \leq 1$

- with more predictor variables, SSE is smaller and R^2 is larger. To evaluate the contribution of the predictors fair, we define the adjusted R^2:

$$R_a^2 = 1 - \frac{\frac{SSE}{n-p-1}}{\frac{SST}{n-1}} = 1 - (\frac{n-1}{n-p-1})\frac{SSE}{SST}$$

  More discussion will be given later about R_a^2.

- For two models with the same number of predictor variables, R^2 can be used to indicate which model is better.

- If model A include more predictor variables than model B, then the R^2 of A must be equal or greater than that of model B. In that case, it is better to use the adjusted R^2.

5 Dwaine studios example

- Y-sales, X_1- number of persons aged 16 or less, X_2- income
- $n = 21, p = 3$
- SST$= 26,196.21$, SSE$= 2,180.93$, SSR$= 26,196.21 - 2,180.93 = 24,015.28$
- $F^* = \frac{24,015.28/2}{2,180.93/18} = 99.1$

  For $H_0 : \beta_1 = \beta_2 = 0$ with $\alpha = 0.05$, $F(0.95; 2, 18) = 3.55$. because

  $$F^* > F(0.95; 2, 18)$$

  we reject H_0

- $$R^2 = \frac{24,015.28}{26.196.21} = 0.917, \qquad R_a^2 = 0.907$$

Writing a fitted regression model

Coefficients:

	Estimate	Std. Error	t value	$Pr(>\|t\|)$	
(Intercept)	-68.8571	60.0170	-1.147	0.2663	
x1	1.4546	0.2118	6.868	2e-06	***
x2	9.3655	4.0640	2.305	0.0333	*

```
Residual standard error:  11.01 on 18 degrees of freedom
Multiple R-squared:  0.9167, Adjusted R-squared:  0.9075
F-statistic:  99.1 on 2 and 18 DF, p-value:  1.921e-10
```

The fitted model is

$$\begin{array}{ccccccc} \hat{Y} & = & \text{-68.86} & + & 1.45X_1 & + & 9.937X_2 \\ \text{(S.E.)} & & (60.02) & & (0.21) & & (4.06) \end{array}$$

$R^2 = 0.9167$, $\quad R_a^2 = 0.9075$, $\quad$ F-statistic: 99.1 on 2 and 18 DF,