Chapter 1 Simple Linear Regression (part 4)

1 Analysis of Variance (ANOVA) approach to regression analysis

Recall the model again

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad i = 1, ..., n$$

The observations can be written as

obs	Y	X
1	Y_1	X_1
2	Y_2	X_2
$\vdots$	$\vdots$	$\vdots$
n	Y_n	X_n

The deviation of each Y_i from the mean $\bar{Y}$,

$$Y_i - \bar{Y}$$

The fitted $\hat{Y}_i = b_0 + b_1 X_i, i = 1, ..., n$ are from the regression and determined by X_i. Their mean is

$$\bar{\hat{Y}} = \frac{1}{n}\sum_{i=1}^{n} Y_i = \bar{Y}$$

Thus the deviation of $\hat{Y}_i$ from its mean is

$$\hat{Y}_i - \bar{Y}$$

The residuals $e_i = Y_i - \hat{Y}_i$, with mean is

$$\bar{e} = 0 \qquad (why?)$$

Thus the deviation of e_i from its mean is

$$e_i = Y_i - \hat{Y}_i$$

Write

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

obs	deviation of Y_i	deviation of $\hat{Y}_i = b_0 + b_1 X_i$	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:

$$\begin{aligned}
\sum_{i=1}^n (Y_i - \bar{Y})^2 &= \sum_{i=1}^n (\hat{Y}_i - \bar{Y} + Y_i - \hat{Y}_i)^2 \\
&= \sum_{i=1}^n \{(\hat{Y}_i - \bar{Y})^2 + (Y_i - \hat{Y}_i)^2 + 2(\hat{Y}_i - \bar{Y})(Y_i - \hat{Y}_i)\} \\
&= SSR + SSE + 2\sum_{i=1}^n (\hat{Y}_i - \bar{Y})(Y_i - \hat{Y}_i) \\
&= SSR + SSE + 2\sum_{i=1}^n (\hat{Y}_i - \bar{Y})e_i \\
&= SSR + SSE + 2\sum_{i=1}^n (b_0 + b_1 X_i - \bar{Y})e_i \\
&= SSR + SSE + 2b_0 \sum_{i=1}^n e_i + 2b_1 \sum_{i=1}^n X_i e_i - 2\bar{Y}\sum_{i=1}^n e_i \\
&= SSR + SSE
\end{aligned}$$

It is also easy to check

$$SSR = \sum_{i=1}^n (b_0 + b_1 X_i - b_0 - b_1 \bar{X})^2 = b_1^2 \sum_{i=1}^n (X_i - \bar{X})^2 \quad (1)$$

Breakdown of the degree of freedom

The degrees of freedom for SST is $n-1$: noticing that

$$Y_1 - \bar{Y}, ..., Y_n - \bar{Y}$$

have one constraint $\sum_{i=1}^{n}(Y_i - \bar{Y}) = 0$

The degrees of freedom for SSR is 1: noticing that

$$\hat{Y}_i = b_0 + b_1 X_i$$

(see Figure 1)

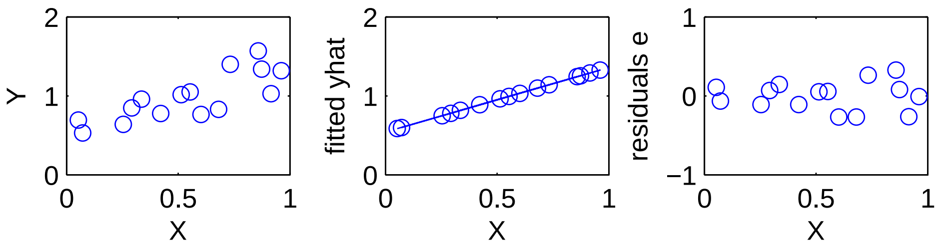

Figure 1: A figure shows the degree of freedom

The degrees of freedom for SSE is $n-2$: noticing that

$$e_1, ..., e_n$$

have TWO constraints $\sum_{i=1}^{n} e_i = 0$ and $\sum_{i=1}^{n} X_i e_i = 0$ (i.e., the normal equation).

Mean (of) Squares

$$MSR = SSR/1 \quad \text{called } \textbf{regression mean square}$$
$$MSE = SSE/(n-2) \quad \text{called } \textbf{error mean square}$$

Analysis of variance (ANOVA) table Based on the break-down, we write it as a table

Source of variation	**SS**	**df**	**MS**	**F-value**	$P(> F)$
Regression	SSR $= \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$	1	MSR $= \frac{SSR}{1}$	$F^* = \frac{MSR}{MSE}$	p-value
Error	SSE $=\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$	n-2	MSE $= \frac{SSE}{n-2}$		
Total	SST $= \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$	n-1			

R command for the calculation

```
anova(object, ...)
```

where "object" is the output of a regression.

Expected Mean Squares

$$E(MSE) = \sigma^2$$

and

$$E(MSR) = \sigma^2 + \beta_1^2 \sum_{i=1}^{n} (X_i - \bar{X})^2$$

[Proof: the first equation was proved (where?). By (1), we have

$$\begin{aligned} E(MSR) &= E(b_1)^2 \sum_{i=1}^{n} (X_i - \bar{X})^2 = [Var(b_1) + (Eb_1)^2] \sum_{i=1}^{n} (X_i - \bar{X})^2 \\ &= [\frac{\sigma^2}{\sum_{i=1}^{n} (X_i - \bar{X})^2} + \beta_1^2] \sum_{i=1}^{n} (X_i - \bar{X})^2 = \sigma^2 + \beta_1^2 \sum_{i=1}^{n} (X_i - \bar{X})^2 \end{aligned}$$

]

2 F-test of $H_0 : \beta_1 = 0$

Consider the hypothesis test

$$H_0 : \beta_1 = 0, \quad H_a : \beta_1 \neq 0.$$

Note that $\hat{Y}_i = b_0 + b_1 X_i$ and

$$SSR = b_1^2 \sum_{i=1}^{n} (X_i - \bar{X})^2$$

If $b_1 = 0$ then $SSR = 0$ (why). Thus we can test $\beta_1 = 0$ based on SSR. i.e. under H_0, SSR or MSR should be "small".

We consider the F-statistic

$$F = \frac{MSR}{MSE} = \frac{SSR/1}{SSE/(n-2)}.$$

Under H_0,

$$F \sim F(1, n-2)$$

For a given significant level α, our criterion is

$$\text{If } F^* \le F(1-\alpha, 1, n-2) \text{ (i.e. indeed small), accept } H_0$$
$$\text{If } F^* > F(1-\alpha, 1, n-2) \text{(i.e. not small), reject } H_0$$

where $F(1-\alpha, 1, n-2)$ is the $(1-\alpha)$ quantile of the F distribution.

We can also do the test based on the p-value $= P(F > F^*)$,

$$\text{If p-value} \ge \alpha, \text{ accept } H_0$$
$$\text{If p-value} < \alpha, \text{ reject } H_0$$

Example 2.1 For the example above (with $n = 25$, in part 3), we fit a model

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

(By **(R code)**), we have the following output

Analysis of Variance Table

Response: Y						
	Df	Sum Sq	Mean Sq	F value	$Pr(>F)$	
X	1	252378	252378	105.88	4.449e-10	***
Residuals	23	54825	2384			

Suppose we need to test $H_0 : \beta_1 = 0$ with significant level 0.01, based on the calculation, the p-value is 4.449×10^{-10} <0.01, we should reject H_0.

Equivalence of F-test and t-test We have two methods to test $H_0 : \beta_1 = 0$ versus $H_1 : \beta_1 \neq 0$. Recall $SSR = b_1^2 \sum_{i=1}^n (X_i - \bar{X})^2$. Thus

$$F^* = \frac{SSR/1}{SSE/(n-2)} = \frac{b_1^2 \sum_{i=1}^n (X_i - \bar{X})^2}{MSE}$$

But since $s^2(b_1) = MSE / \sum_{i=1}^n (X_i - \bar{X})^2$ (where?), we have under H_0,

$$F^* = \frac{b_1^2}{s^2(b_1)} = \Big(\frac{b_1}{s(b_1)}\Big)^2 = (t^*)^2.$$

Thus

$$F^* > F(1-\alpha, 1, n-2) \Longleftrightarrow (t^*)^2 > (t(1-\alpha/2, n-2))^2 \Longleftrightarrow |t^*| > t(1-\alpha/2, n-2).$$

and

$$F^* \le F(1-\alpha, 1, n-2) \Longleftrightarrow (t^*)^2 \le (t(1-\alpha/2, n-2))^2 \Longleftrightarrow |t^*| \le t(1-\alpha/2, n-2).$$

(you can check in the statistical table $F(1-\alpha, 1, n-2) = (t(1-\alpha/2, n-2))^2$) Therefore, the test results based on F and t statistics are the same. (But ONLY for simple linear regression model)

3 General linear test approach

To test whether $H_0 : \beta_1 = 0$, we can do it by comparing two models

$$\text{Full model} : Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

and

$$\text{Reduced model} : Y_i = \beta_0 + \varepsilon_i$$

Denote the SSR of the FULL and REDUCED models by $SSR(F)$ and $SSR(R)$ respectively (and SSE(R), SSR(F)). We have immediately

$$SSR(F) \geq SSR(R)$$

or

$$SSE(F) \leq SSE(R).$$

A question: when does the equality hold?

Note that if $H_0 : \beta_1 = 0$ holds, then

$$\frac{SSE(R) - SSE(F)}{SSE(F)} \text{ should be small}$$

Considering the degree of freedoms, define

$$F = \frac{(SSE(R) - SSE(F))/(df_R - df_F)}{SSE(F)/df_F} \text{ should be small}$$

where df_R and df_F indicate the degrees of freedom of $SSE(R)$ and $SSE(F)$ respectively. Under $H_0 : \beta_1 = 0$, it is proved that

$$F \sim F(df_R - df_F, df_F)$$

Suppose we get the F value as F^*, then

$$\text{If } F^* \leq F(1-\alpha, df_R - df_F, df_F), \text{ accept } H_0$$
$$\text{If } F^* > F(1-\alpha, df_R - df_F, df_F), \text{ reject } H_0$$

Similarly, based on the p-value $= P(F > F^*)$,

$$\text{If p-value} \geq \alpha, \text{ accept } H_0$$
$$\text{If p-value} < \alpha, \text{ reject } H_0$$

4 Descriptive measures of linear association between X and Y

It follows from

$$SST = SSR + SSE$$

that

$$1 = \frac{SSR}{SST} + \frac{SSE}{SST}$$

where

- $\frac{SSR}{SST}$ is the proportion of Total sum of squares that can be explained/predicted by the predictor X
- $\frac{SSE}{SST}$ is the proportion of Total sum of squares that caused by the random effect.

A "good" model should have large

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

R^2 is called $R-$**square**, or **coefficient of determination**

Some facts about R^2 for simple linear regression model

1. $0 \leq R^2 \leq 1$.
2. if $R^2 = 0$, then $b_1 = 0$ (because $SSR = b_1^2 \sum_{i=1}^n (X_i - \bar{X})^2$)
3. if $R^2 = 1$, then $Y_i = b_0 + b_1 X_i$ (why?)
4. the correlation coefficient between

   $$r_{X,Y} = \pm\sqrt{R^2}$$

   [Proof:

   $$R^2 = \frac{SSR}{SST} = \frac{b_1^2 \sum_{i=1}^n (X_i - \bar{X})^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2} = r_{XY}^2$$

5. R^2 only indicates the fitness in the observed range/scope. We need to be careful if we make prediction outside the range.
6. R^2 only indicates the "linear relationships". $R^2 = 0$ does not mean X and Y have no nonlinear association.

5 Considerations in Applying regression analysis

1. In prediction a new case, we need to ensure the model is applicable to the new case.

2. Sometimes we need to predict X, and thus predict Y. As a consequence, the prediction accuracy also depends on the prediction of X

3. The range of X for the model. If a new case X is far from the range, in the prediction, we need be careful

4. $\beta_1 \neq 0$ only indicates the correlation relationship, but not a cause-and-effect relation (causality).

5. Even if $\beta_1 = 0$ can be concluded, we cannot say Y has no relationship/association with X. We can only say there is no LINEAR relationship/association between X and Y.

6 Write an estimated model

$$\begin{array}{cccc} \hat{Y} & = & b_0 & + \quad b_1 \ X \\ \text{(S.E.)} & & (s(b_0)) & (s(b_1)) \end{array}$$

$$\hat{\sigma}^2(\text{or MSE}) = ..., \quad R^2 = ...,$$
$$\text{F-statistic} = ... \text{ (and others)}$$

Other formats of writing a fitted model can be found in Part 3 of the lecture notes.