Multiple Regression and ANOVA (Ch. 9.2)

Yifan Zhu

Iowa State University

Outline

Multiple Regression and ANOVA

- **Analysis of variance (ANOVA)**: the use of sums of squares to construct a test statistic for comparing nested models.
- **Nested models**: a pair of models such that one contains all the parameters of the other.
  - Examples:
    - Full model: $Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i$ with the reduced model: $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$.
    - Full model: $Y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \varepsilon_i$ with the reduced model: $Y_i = \beta_0 + \varepsilon_i$

Sums of Squares

- **Total sum of squares (SST)**: the total amount of variation in the response.

$$SST = \sum_i (y_i - \overline{y})^2$$

- **Regression sum of squares (SSR)**: the amount of variation in response explained by the model.

$$SSR = \sum_i (\widehat{y}_i - \overline{y})^2$$

- **Error sum of squares (SSE)**: the amount of variation in the response *not* explained by the model.

$$SSE = \sum_i (y_i - \widehat{y}_i)^2$$

Properties of Sums of Squares

- They add up:

$$SST = SSR + SSE$$

- We can use them to calculate R^2:

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

- We can calculate the **mean squared error (MSE)**:

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

which satisfies:

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

$MSE = s_{LF}^2$ for simple linear regression and s_{SF}^2 for multiple regression.

- The **regression mean square (MSR)** is:

$$MSR = \frac{1}{p-1} SSR$$

Inference: deciding between nested models

- Suppose I have the full model:

$$Y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \cdots + \beta_{p-1} x_{p-1,i} + \varepsilon_i$$

- And an intercept-only reduced model:

$$Y_i = \beta_0 + \varepsilon_i$$

- I want to do a hypothesis test to decide if the full model works better than the reduced model.
  - Does the full model explain significantly more variation in the response than the reduced model?
  - This is a job for the sums of squares.

The hypothesis test: intercept-only model vs. full model

1.
   - $H_0 : \beta_1 = \beta_2 = \cdots = \beta_{p-1} = 0$
   - H_a : not all of the β_i's $= 0$ $(i = 1, 2, \ldots, p-1)$
2. α is some sensible value (< 0.1).
3. The test statistic is:

$$F = \frac{SSR/(p-1)}{SSE/(n-p)} = \frac{MSR}{MSE} \sim F_{p-1,\ n-p}$$

   Assume:
   - H_0 is true.
   - The full model is valid with the ε_i's iid N(0,σ^2)
4. Reject H_0 if observed F $> F_{p-1,n-p,1-\alpha}$. Or use the p-value: $P(F_{p-1,n-p} > observedF)$; reject H_0 when p-value is small.

Example: stack loss

1. Consider a chemical plant that makes nitric acid from ammonia.
2. We want to predict stack loss (y, 10 times the % ammonia that escapes from the absorption column) using:
   - x_1: air flow, the rate of operation of the plant
   - x_2, inlet temperature of the cooling water
   - x_3: (% circulating acid - 50%)$\times$10

Example: stack loss

i, Observation Number	x_{1i}, Air Flow	x_{2i}, Cooling Water Inlet Temperature	x_{3i}, Acid Concentration	y_i, Stack Loss
1	80	27	88	37
2	62	22	87	18
3	62	23	87	18
4	62	24	93	19
5	62	24	93	20
6	58	23	87	15
7	58	18	80	14
8	58	18	89	14
9	58	17	88	13
10	58	18	82	11
11	58	19	93	12
12	50	18	89	8
13	50	18	86	7
14	50	19	72	8
15	50	19	79	8
16	50	20	80	9
17	56	20	82	15

Example: stack loss

- Given:
  - $n = 17$
  - y: stack loss of nitrogen from the chemical plant.
  - x_1: air flow, the rate of operation of the plant
  - x_2, inlet temperature of the cooling water
  - x_3: (% circulating acid - 50%)$\times 10$
- We'll test the full model:

$$Y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \beta_3 x_{3,i} + \varepsilon_i$$

against the reduced model:

$$Y_i = \beta_0 + \varepsilon_i$$

at $\alpha = 0.05$.

Example: stack loss

1.
    - $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$
    - Not all of the β_i's are 0, $i = 1, 2, 3$.
2. $\alpha = 0.05$
3. The test statistic is:

$$F = \frac{SSR/(p-1)}{SSE/(n-p)} = \frac{MSR}{MSE} \sim F_{p-1,\ n-p}$$

Assume:

- H_0 is true.
- The full model is valid with the ε_i's iid N$(0,\sigma^2)$

Reject H_0 if $F > F_{p-1,\ n-p,\ 1-\alpha} = F_{4-1,\ 17-4,\ 1-0.05} = F_{3,13,0.95} = 3.41$.

Example: stack loss

4. In JMP, fit the full model and look at the **ANOVA table**:

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	3	795.83449	265.278	169.0432
Error	13	20.40080	1.569	**Prob > F**
C. Total	16	816.23529		<.0001*

by reading directly from the table, we can see:

- $p - 1 = 3$, $n - p = 13$, $n - 1 = 16$
- $SSR = 795.83, SSE = 20.4, SST = 816.24$
- $MSR = SSR/(p - 1) = 795.83/3 = 265.28$
- $MSE = SSE/(n - p) = 20.4/13 = 1.57$
- $observedF = MSR/MSE = 265.78/1.57 = 169.04$
- Prob>F gives the p-value, $P(F_{3,13} > observedF) < 0.0001$.

5. With $F = 169.04 > 3.41$, we reject H_0 and conclude H_a.
6. There is overwhelming evidence that at least one of air flow, inlet temperature, and % circulating acid is important in explaining the variation in stack loss.

What if I want to compare different nested models?

1. - $H_0 : \beta_{l_1} = \beta_{l_2} = \cdots = \beta_{l_k} = 0$
   - H_a : not all of $\beta_{l_1}, \beta_{l_2}, \cdots, \beta_{l_k}$ are 0.
   - (For example, $H_0 : \beta_2 = \beta_3 = 0$ vs $H_a :$ either β_2 or $\beta_3 \neq 0$ or both. The model is $Y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \beta_3 x_{i,3} + \beta_4 x_{i,4} + \varepsilon_i$, and $k = 2$)
2. α is some sensible value.
3. The test statistic is:

$$F = \frac{(SSR_f - SSR_r)/k}{SSE_f/(n-p)} \sim F_{k,\ n-p}$$

- SSR_r is for the reduced model and SSR_f is for the full model.
- Of course, we assume H_0 is true and the full model is valid with the ε_i's iid $N(0, \sigma^2)$.

What if I want to compare different nested models?

4. We can construct a combined ANOVA table:

Source	SS	df	MS	F
Reg (full)	SSR_f	$p-1$		
Reg (reduced)	SSR_r	$p-k-1$		
Reg (full | red)	$SSR_f - SSR_r$	k	$\frac{SSR_f - SSR_r}{k}$	$\frac{MSR_{f\|r}}{MSE_f}$
Error	SSE_f	$n-p$	$\frac{SSE_f}{n-p}$	
Total	SST	$n-1$		

5. Reject H_0 if observed F $> F_{p-1,n-p,1-\alpha}$. Or use the p-value: $P(F_{p-1,n-p} > observedF)$; reject H_0 when p-value is small.

Example: stack loss

1.
   - $H_0 : \beta_2 = \beta_3 = 0$
   - H_a : either $\beta_2 \neq 0$ or $\beta_3 \neq 0$
2. $\alpha = 0.05$
3. The test statistic is:

$$F = \frac{(SSR_f - SSR_r)/k}{SSE_f/(n-p)} = \frac{(SSR_f - SSR_r)/2}{SSE_f/(17-4)}$$
$$= \frac{(SSR_f - SSR_r)/2}{SSE_f/13}$$

- Assume H_0 is true and the full model is valid with the ε_i's iid $N(0, \sigma^2)$.
- Then, $F \sim F_{k,\ n-p} = F_{2,13}$.
- I will reject H_0 if $F > F_{2,13,0.95} = 3.81$.

Example: stack loss

4. Look at the ANOVA tables in JMP for both the full model ($Y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \beta_3 x_{3,i} + \varepsilon_i$):

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	3	795.83449	265.278	169.0432
Error	13	20.40080	1.569	**Prob > F**
C. Total	16	816.23529		<.0001*

and the reduced model ($Y_i = \beta_0 + \beta_1 x_{1,i} + \varepsilon_i$):

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	775.48219	775.482	285.4318
Error	15	40.75311	2.717	**Prob > F**
C. Total	16	816.23529		<.0001*

Example: stack loss

I construct a different ANOVA table for this test:

Source	SS	df	MS	F
Reg (full)	795.83	4		
Reg (reduced)	775.48	2		
Reg (full | red)	20.35	2	10.18	6.48
Error	20.4	13	1.57	
Total	SST	16		

5. With $observedF = 6.48 > 3.81$, I reject H_0 and conclude H_a.
6. There is enough evidence to conclude that at least one of inlet temperature and % circulating acid is associated with stack loss.

Example: stack loss

- Attempt to eliminate inlet temperature (x_2) from the model at $\alpha = 0.05$. Here is the ANOVA table for the full model:

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	3	795.83449	265.278	169.0432
Error	13	20.40080	1.569	**Prob > F**
C. Total	16	816.23529		<.0001*

and for the reduced model:

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	2	776.84496	388.422	138.0520
Error	14	39.39033	2.814	**Prob > F**
C. Total	16	816.23529		<.0001*

Example: stack loss

1. $H_0 : \beta_2 = 0,\ H_a : \beta_2 \neq 0$
2. $\alpha = 0.05$
3. The test statistic is:

$$F = \frac{(SSR_f - SSR_r)/k}{SSE_f/(n-p)} = \frac{SSR_f - SSR_r}{SSE_f/(17-4)}$$
$$= \frac{SSR_f - SSR_r}{SSE_f/13}$$

- Assume H_0 is true and the full model is valid with the ε_i's iid $N(0, \sigma^2)$.
- Then, $F \sim F_{k,\ n-p} = F_{1,13}$.
- I will reject H_0 if $F > F_{1,13,0.95} = 4.67$.

Example: stack loss

4. I construct a different ANOVA table for this test:

Source	SS	df	MS	F
Reg (full)	795.83	4		
Reg (reduced)	776.84	3		
Reg (full | red)	18.99	1	18.99	12.10
Error	20.4	13	1.57	
Total	SST	16		

5. With $observedF = 12.10 > 4.67$, we reject H_0.
6. There is enough evidence to conclude that stack loss varies with inlet temperature.

Example: stack loss

- The F test for eliminating one parameter is analogous to the t test from before:

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-37.65246	4.732051	-7.96	<.0001*
x1	0.7976856	0.067439	11.83	<.0001*
x2	0.5773405	0.165969	3.48	0.0041*
x3	-0.06706	0.061603	-1.09	0.2961

- The t statistic for $H_0 : \beta_2 = 0$ vs. $H_0 : \beta_2 \neq 0$ is 3.48.
- But $3.48^2 = 12.1$, which is our F statistic from the ANVOA test!
- Fun fact:

$$F_{1,\ \nu} = t_\nu^2$$

The F test statistic and R^2

- If F is the test statistic from a test of $H_0 : \beta_1 = \beta_2 = \cdots = \beta_{p-1} = 0$ vs. H_a : not all of $\beta_1, \beta_2, \ldots, \beta_{p-1}$ are 0, then F can be expressed in terms of the coefficient of determination of the full model:

$$F = \frac{R^2/(p-1)}{(1-R^2)/(n-p)}$$

- For the stack loss example, the full model's $R^2 = 0.975$, and so:

$$F = \frac{0.975/(4-1)}{(1-0.975)/(17-4)} = 169$$

The F test statistic and R^2

Summary of Fit

RSquare	0.975006
RSquare Adj	0.969238
Root Mean Square Error	1.252714
Mean of Response	14.47059
Observations (or Sum Wgts)	17

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	3	795.83449	265.278	169.0432
Error	13	20.40080	1.569	**Prob > F**
C. Total	16	816.23529		<.0001*

The F test statistic and R^2

- For $H_0 : \beta_1 = \beta_2 = \cdots = \beta_{p-1} = 0$ vs. H_a : not all of $\beta_1, \beta_2, \ldots, \beta_{p-1}$,

$$F = \frac{SSR \frac{1}{p-1}}{SSE \frac{1}{n-p}} = \frac{\frac{SSR}{SST} \frac{1}{p-1}}{\frac{SSE}{SST} \frac{1}{n-p}} = \frac{\frac{SSR}{SST} \frac{1}{p-1}}{\frac{SST-SSR}{SST} \frac{1}{n-p}} = \frac{\frac{SSR}{SST} \frac{1}{p-1}}{\left(1 - \frac{SSR}{SST}\right) \frac{1}{n-p}}$$

$$= \frac{R^2 \frac{1}{p-1}}{(1 - R^2) \frac{1}{n-p}}$$

The F test statistic and R^2

- If F is the test statistic from a test of $H_0 : \beta_{l_1} = \beta_{l_2} = \cdots = \beta_{l_k} = 0$ vs. H_a : not all of $\beta_{l_1}, \beta_{l_2}, \ldots, \beta_{l_k}$ are 0, then F can be expressed in terms of the coefficient of determination of the full model (R_f^2) and that of the reduced model (R_r^2):

$$F = \frac{(R_f^2 - R_r^2)/k}{(1 - R_f^2)/(n - p)}$$

- For the stack loss example when we tested $H_0 : \beta_2 = \beta_3 = 0$, $R_f^2 = 0.975$ and $R_r^2 = 0.95$.

$$F = \frac{(0.975 - 0.95)/2}{(1 - 0.975)/(17 - 4)} = 6.50$$

which is close to the test statistic of 6.48 that we calculated before.

The F test statistic and R^2

- When we tested $H_0 : \beta_2 = 0$, R_r^2 was 0.9517, so:

$$F = \frac{(0.975 - 0.9517)/1}{(1 - 0.975)/(17 - 4)} = 12.117$$

which is close to the test statistic of 12.10 that was calculated directly from the ANOVA table.

The F test statistic and R^2

$$F = \frac{(SSR_f - SSR_r)\frac{1}{k}}{SSE_f \frac{1}{n-p}} = \frac{\frac{SSR_f - SSR_r}{SST}\frac{1}{k}}{\frac{SSE_f}{SST}\frac{1}{n-p}} = \frac{\left(\frac{SSR_f}{SST} - \frac{SSR_r}{SST}\right)\frac{1}{k}}{\frac{SST - SSR_f}{SST}\frac{1}{n-p}}$$

$$= \frac{\left(\frac{SSR_f}{SST} - \frac{SSR_r}{SST}\right)\frac{1}{k}}{\left(1 - \frac{SSR_f}{SST}\right)\frac{1}{n-p}} = \frac{\left(R_f^2 - R_r^2\right)\frac{1}{k}}{(1 - R_f^2)\frac{1}{n-p}}$$