L19 Lack of fit test

1. Lack of fit test

   (1) Lack of fit test
   In the one-way ANOVA model $y = \mu(x_1, ..., x_k) + \epsilon$ the tests on

   $$H_0: \mu(x_1, ..., x_k) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k \quad \text{versus } H_a: \mu(x_1, ..., x_k) \neq \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$$
   $$\text{or } H_0: \mu(x_1, .., x_k) = \beta_1 x_1 + \cdots + \beta_k x_k \quad \text{versus } H_a: \mu(x_1, .., x_k) \neq \beta_1 x_1 + \cdots + \beta_k x_k$$

   are the lack of fit tests since H_a says there is lack of fit for regression model

   $$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \epsilon. \quad \text{or} \quad y = \beta_1 x_1 + \cdots + \beta_k x_k + \epsilon.$$

   (2) Test statistic
   Because both original and reduced models are regression models, the LRT statistic is
   Test Statistic: $F = \frac{(SSEr - SSE)/(q-p)}{MSE}$.
   Here SSE is from the full model, the ANOVA model, and is now SSPE
   SSE_r is from the reduced modek, the regression model, and is denoted as SSE
   $SSE_r - SSE$ is $SSE - SSPE$, called SSLF
   DF of SSLF=(DF of SSE)-(DF of SSPE)=(n-p)-(n-q)=q-p
   Thus $F = \frac{MSLF}{MSPE}$ where $MSLF = SSLF/(q-p)$ and $MSPE = SSPE/(n-q)$.

   (3) Decision rule
   Reject H_0 if $F > F_\alpha(q-p, n-q)$. p-value: $P(F(q-p, n-q) > F_{ob})$.

2. Extended ANOVA table with decomposition of SSE

   (1) For model with intercept
   For $y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \epsilon$, $p = k+1$

Source	SS	DF	MS	F	P
Model	SSM	$p-1=k$	MSM	MSM/MSE	$P(F(p-1, n-p) > F_{ob})$
Error	SSE	$n-p$	MSE		
LackFit	SSLF	$q-p$	MSLF	MSLF/MSPE	$P(F(q-p, n-q) > F_{ob})$
PureError	SSPE	$n-q$	MSPE		
C.Total	SSTO	$n-1$			

   (2) For model without intercept
   For model $y = \beta_1 x_1 + \cdots + \beta_k x_k + \epsilon$, $p = k$

Source	SS	DF	MS	F	P
Model	SSM	$p=k$	MSM	MSM/MSE	$P(F(p, n-p) > F_{ob})$
Error	SSE	$n-p$	MSE		
LackFit	SSLF	$q-p$	MSLF	MSLF/MSPE	$P(F(q-p, n-q) > F_{ob})$
PureError	SSPE	$n-q$	MSPE		
U.Total	SSTO	n			

3. Implementation of lack of fit test

   (1) Use extended ANOVA table as tool

Ex: For ANOVA model $y = \mu(x) + \epsilon$ with data

x	1	1	0	0	-1	1
y	1	2	3	4	5	0

we obtained SSPE=2.5 with DF=3. So MSLE= $0.8\dot{3}$. For $y = \beta_0 + \beta_1 x + \epsilon$

Source	SS	DF	MS	F	P
Model	14.7	1	14.7	21	0.0102
Error	2.8	4	0.7		
LackFit	0.3	1	0.3	0.36	0.5908
PureError	2.5	3	$0.8\dot{3}$		
C.Total	17.5	5			

For $y = \beta x + \epsilon$

Source	SS	DF	MS	F	P
Model	1	1	1	0.09	0.77
Error	54	5	10.8		
LackFit	51.5	2	25.75	30.09	0.01
PureError	2.5	3	$0.8\dot{3}$		
U.Total	55	6			

Thus

$H_0 : \mu(x) = \beta x$ versus $H_a : \mu(x) \neq \beta x$
Test Statistics: $F = \frac{MSLF}{MSPE}$
p-value: $P(F(q-1, n-q) > F_{ob})$
$F_{ob} = 30.9$, p-value: $P(F(2,3) > 30.9) = 0.01$
Reject H_0. Model $y = \beta x + \epsilon$ lacks the fit.

(2) Use SAS to display the extended ANOVA table

```
data a;
   infile "D:\ex.txt";
   input y x1 x2 @@;
proc reg;
   model y=x1 x2/lackfit;
   run;
proc reg;
   model y=x1 x2/noint lackfit;
   run;
```

displays the extended ANOVA tables.

L20 Type I SS

1. Type I SS for regression with intercept

   (1) Definition
   SAS: proc reg; model y=x1 x2 x3/ss1; run;
   attaches Type I SS in parameter table to each β_i, $i = 0, 1, 2, 3$.

Model with	Test on	Type I SS
β_0	$H_0: \beta_0 = 0$	$SSI_0 = SSE(\emptyset) - SSE(\beta_0)$
β_1, β_0	$H_0: \beta_1 = 0$	$SSI_1 = SSE(\beta_0) - SSE(\beta_1, \beta_0)$
β_2, β_1, β_0	$H_0: \beta_2 = 0$	$SSI_2 = SSE(\beta_1, \beta_0) - SSE(\beta_2, \beta_1, \beta_0)$
β_3, β_2, β_1, β_0	$H_0: \beta_3 = 0$	$SSI_3 = SSE(\beta_2, \beta_1, \beta_0) - SSE(\beta_3, \beta_2\beta_1, \beta_0)$

   (2) Usage
   Numerator of test statistic F

Test on	Test statistic	
$H_0: \beta_0 = 0$	$F = \frac{SSI_0}{MSE(\beta_0)}$	$= \frac{SSI_0}{SSE(\beta_0)/(n-1)}$
$H_0: \beta_1 = 0$	$F = \frac{SSI_1}{MSE(\beta_0)}$	$= \frac{SSI_1}{SSE(\beta_1,\beta_0)/(n-2)}$
$H_0: \beta_2 = 0$	$F = \frac{SSI_2}{MSE(\beta_2,\beta_1,\beta_0)}$	$= \frac{SSI_0}{SSE(\beta_2,\beta_1,\beta_0)/(n-3)}$
$H_0: \beta_3 = 0$	$F = \frac{SSI_3}{MSE(\beta_3,\beta_2,\beta_1,\beta_0)}$	$= \frac{SSI_0}{SSE(\beta_3,\beta_2,\beta_1,\beta_0)/(n-4)}$

   (3) SSE in sequential models
   The denominator of F depends on the $SSE = SSE(\beta_k, ..., \beta_0)$ in the sequential models.
   $SSE(\beta_0) = \sum(y_i - \overline{y})^2 = SSTO$ in the full model listed in ANOVA table.
   $\sum_{i=1}^{k} SSI_i = SSE(\beta_0) - SSE(\beta_k, ..., \beta_0)$. So $SSE(\beta_k, ..., \beta_0) = SSTO - \sum_{i=1}^{k} SSI_i$.

Ex1: With data in Table94.txt for model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$

```
proc reg; model y=x1 x2/noprint; test x2=0 run;
```
produced

Source	DF	MS	F	$Pr > F$
Numerator	1	1.47716	0.12	0.7345
Denominator	9	12.06648		

for test on $H_0: \beta_0 = 0$ vs $H_a: \beta_2 \neq 0$.

```
proc reg; model y=x1 x2 x3/ss1; run;
```
produced

$SSTO = 117.53684$ with $DF = 11$

$SS1_0 = 1232.84168$, $SSI_1 = 7.46139$, $SSI_2 = 1.47716$, $SSI_3 = 0.43322$.

Thus $SSE(\beta_0, \beta_1, \beta_2) = SST0 - (SSI_1 + SSI_2) = 108.59829$ with $DF = 12 - 3 = 9$.

So $MSE(\beta_0, \beta_1, \beta_2) = 12.06648$ Hence $F = \frac{SSI_2}{MSE(\beta_2,\beta_1,\beta_0)} = \frac{1.47716}{12.06648} = 0.1224$

2. Type I SS for regression without intercept

   (1) Definition
   SAS: proc reg; model y=x1 x2 x3/noint ss1; run;
   attaches Type I SS in parameter table to each β_i, $i = 1, 2, 3$.

Model with	Test on	Type I SS
β_1	$H_0: \beta_1 = 0$	$SSI_1 = SSE(\emptyset) - SSE(\beta_1)$
β_2, β_1	$H_0: \beta_2 = 0$	$SSI_2 = SSE(\beta_1) - SSE(\beta_2, \beta_1)$
β_3, β_2, β_1	$H_0: \beta_3 = 0$	$SSI_3 = SSE(\beta_2, \beta_1) - SSE(\beta_3, \beta_2\beta_1)$

   (2) Usage
   Numerator of test statistic F

Test on	Test statistic	
$H_0: \beta_1 = 0$	$F = \frac{SSI_1}{MSE(\beta_1)}$	$= \frac{SSI_1}{SSE(\beta_1)/(n-1)}$
$H_0: \beta_2 = 0$	$F = \frac{SSI_2}{MSE(\beta_2,\beta_1)}$	$= \frac{SSI_0}{SSE(\beta_2,\beta_1)/(n-2)}$
$H_0: \beta_3 = 0$	$F = \frac{SSI_3}{MSE(\beta_3,\beta_2,\beta_1)}$	$= \frac{SSI_0}{SSE(\beta_3,\beta_2,\beta_1)/(n-3)}$

(3) SSE in sequential models
The denominator of F depends on the $SSE = SSE(\beta_k, ..., \beta_1)$ in the sequential models.
Note that $SSE(\emptyset) = \sum y_i^2 = SSTO$ in the full model listed in ANOVA table.
$\sum_{i=1}^k SSI_i = SSE(\emptyset) - SSE(\beta_k, ..., \beta_1)$. So $SSE(\beta_k, ..., \beta_1) = SSTO - \sum_{i=1}^k SSI_i$.

Ex2: With data in Table94.txt for model $y = \beta_1 x_1 + \beta_2 x_2 + \epsilon$

```
proc reg; model y=x1 x2/noint noprint; test x2=0 run;
```

produced

Source	DF	MS	F	$Pr > F$
Numerator	1	119.58182	1.76	0.2144
Denominator	10	68.03391		

for test on $H_0 : \beta_0 = 0$ vs $H_a : \beta_2 \neq 0$.

```
proc reg; model y=x1 x2 x3/noint ss1; run;
```

produced
$SSTO = 1350.37852$ with $DF = 12$
$SSI_1 = 550.45760$, $SSI_2 = 119.58182$, $SSI_3 = 334.93396$.
Thus $SSE(\beta_1, \beta_2) = SST0 - (SSI_1 + SSI_2) = 680.3391$ with $DF = 12 - 2 = 10$.
So $MSE(\beta_1, \beta_2) = 68.03391$ Hence $F = \frac{SSI_2}{MSE(\beta_2,\beta_1)} = \frac{119.58182}{68.03391} = 1.7577$

3. Confidence region for β

(1) $\frac{(\widehat{\beta}-\beta)'(X'X)(\widehat{\beta}-\beta)/p}{MSE} \sim F(p, n-p)$.

Proof.

$H_0 : A\beta = b$ vs $H_a : A\beta \neq b$
Test statistic: $F = \frac{(A\widehat{\beta}-b)'[A(X'X)^{-1}A']^{-1}/q}{MSE}$
Reject H_0 if $F > F_\alpha(q, n-p)$

is an α-level test since $P(F > F_\alpha(q, n-p)|H_0) = \alpha$. Hence $F \overset{H_0}{\sim} F(q, n-p)$, i.e.,

$$\frac{(A\widehat{\beta} - A\beta)'[A(X'X)^{-1}A']^{-1}(A\widehat{\beta} - A\beta)/q}{MSE} \sim F(q, n-p).$$

Let $A = I_p$. Then $\frac{(\widehat{\beta}-\beta)'(X'X)(\widehat{\beta}-\beta)/p}{MSE} \sim F(p, n-p)$.

(2) Confidence region for β.
Let $CR = \left\{\beta : \frac{(\widehat{\beta}-\beta)'(X'X)(\widehat{\beta}-\beta)/p}{MSE} < F_\alpha(p, n-p)\right\}$. Then CR is a confidence region for β with confidence coefficient $1 - \alpha$ since $P(CR) = \alpha$.

Comment: CR is an ellipsoid in R^p.