R.R. If $t > t_{\alpha/2;\, n-(K+1)}$
or $t < -t_{\alpha/2;\, n-(K+1)}$

Fdrop (or Fpart) $\Rightarrow$ R.R. We reject if Fdrop or Fpart) $>$ $F_{\alpha(1;n-(K+1)}$

$t^2_{n-(K+1)} = F_{part}(1; n-(K+1)) = F_{drop}(1; n-(K+1))$

Fpart
- is based on breaking down the SSR into components i.e. we determine the contribution to SSR made by variable x_j after all other variables have been included

i.e. SS(due to x_j after all other x's in the model) - SSR(x_j | all other x_i's)
"given happened already"
$= SSR(\text{due all x's}) - SSR(\text{all x's w/out } x_j) =$
$= SSR(x_1, x_2 \ldots x_{j-1}, x_j, x_{j+1} \cdots x_K) - SSR(x_1, x_2 \ldots x_{j-1}, x_{j+1}, \ldots x_K)$
$= SSR_F - SSR_r$

$r^2 = \frac{SSR}{TSS}$ = determines how much of the tot. variation in the data can be explained by the regr. line and hence how much is due to error

↑ def. of determination

- if r^2 is = 0 (or close to) = when none of the x's in the model have strong linear relationship w/ y
- if $r^2 = 1$ (or close to) it does not necessarily imply a gd model as there may be only few x's that are linearly

related to y (and hence the rest of the x's do not contribute)
- also SSR increases as # of x's increases, hence $r^2 = \frac{SSR}{TSS}$ will increase also (it does not reflect the goodness of fit)
- better measure is so called r^2 adjusted which will drop down if unimportant x's are entered into the model.

$$r^2_{adj} = 1 - \frac{MSE}{TSS/n-1} = 1 - \frac{SSE/n-(k+1)}{TSS/n-1}$$

i.e. - If 96 vs 70 there is an issue

Dummy variables (Indicator variables)
- are used to represent qualitative (categorical) variables such as sex, marital status, geographic region, mother tongue)

the # of dummy variables needed to represent such a variable is always one less than # of categories the variable in question has.

i.e: sex = {m, f} ⇒ 2 categories ⇒ 1 dummy $X = \begin{cases} 0 \text{ if } m \\ 1 \text{ if } f \end{cases}$

coffee cup size {s, m, l} ⇒ 3 categories ⇒ 2 dum. $X_1 = \begin{cases} 1 \text{ if } s \\ 0 \text{ otherwise} \end{cases}$ $X_2 = \begin{cases} 1 \text{ if } m \\ 0 \text{ otherwise} \end{cases}$

if s $X_1 = 1$ $X_2 = 0$ if m $X_1 = 0$ $X_2 = 1$ if l $X_1 = 0$ $X_2 = 0$

(Ex) want to examine the relationship btwn university salary (y), the # of yrs of experience (x_1) and the sex of professor (x_2)

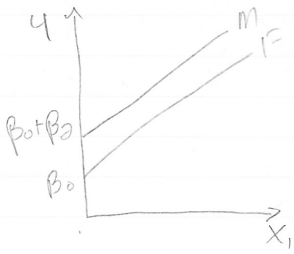
We might suspect a linear relationship btwn y & x_1 for both males & females i.e. we run SLR model $y = \beta_0 + \beta_1 x_1 + \varepsilon$ 2 times (once for males & once for F)

- get 2 lines & can compare them
- But we want x_2 (sex of prof) in the same model w/ x_1

i.e. $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$ where $x_2 = \begin{cases} 0, \text{if F} \\ 1, \text{if M} \end{cases}$

if prof M: $y = \beta_0 + \beta_1 x_1 + \beta_2(1) + \varepsilon$
or $y = \beta_0 + \beta_2 + \beta_1 x_1 + \varepsilon \ldots$ (line M)

if prof F: $y = \beta_0 + \beta_1 x_1 + \beta_2(0) + \varepsilon$
or $y = \beta_0 + \beta_1 x_1 + \varepsilon \ldots$ (line F)

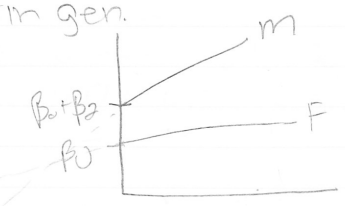


To test that starting sal.'s are the same
$H_0: \beta_2 = 0$
$H_a: \beta_2 \neq 0$

β_0 = y inter of F
β_1 = slope of F & M
$\beta_2 = (\beta_0 + \beta_2) - \beta_0$ difference in y intercepts btwn M & F's

- Different y intercepts, but same is. (parallel) slopes "Diff starting sal.'s but same incs"

- In gen.



- in this case, we say that x_1 & x_2 interact and therefore interaction term (ie. cross-product term) needs to be added to the model

cross product i.e. interaction term

i.e. $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \varepsilon$ ← "interaction model"

Here we have different y intercepts and slopes

If a prof is a male: $y = \beta_0 + \beta_1 x_1 + \beta_2(1) + \beta_3 x_1(1) + \varepsilon$
or $y = (\beta_0 + \beta_2) + (\beta_1 + \beta_3)x_1 + \varepsilon \quad \ldots (m)$

- If our prof F: $y = \beta_0 + \beta_1 x_1 + \beta_2(0) + \beta_3 x_1(0) + \varepsilon$
or $y = \beta_0 + \beta_1 x_1 + \varepsilon \quad \ldots (F)$

β_0 = y-interc. for F
$\beta_2 = (\beta_0 + \beta_2) - \beta_0$ = difference in y-intercepts btwn m + F
β_1 = slope of F
$\beta_3 = (\beta_1 + \beta_3) - \beta_3$ = diff. in slopes for m + F

※ Interaction models

- if the relationship btwn y + x_1 for example, is different for diff. values of x_2 then there is an interaction effect btwn x_1 + x_2 which should be accntd for in the model by adding the cross product (i.e. interaction term) $x_1 x_2$

i.e. $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \varepsilon$

NOTE: When using dummy variable(s) you often need to include terms representing the possible interactions of the dummy variable(s) w/ all the other non-dummy variables

- Comparing two lines for m's + F's

testing if the two lines have the same y-intercept

(one par. use t or F)

$H_0: \beta_2 = 0$; α
$H_a: \beta_2 \neq 0$

~~t-test~~ $= t = \frac{\hat{\beta}_2}{\sqrt{Var_{22} \cdot MSE}}$ Rej. if $t > t_{\alpha/2; n-4}$ or $t < -t_{\alpha/2; n-4}$

$F_{part} = \frac{[SSR_F - SSR_R] / [df_{SSR_F} - df_{SSR_R}]}{MSE_F}$

Full: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$
R: $Y = \beta_0 + \beta_1 X_1 + \beta_3 X_1 X_2 + \varepsilon$

$\frac{[SSR_F - SSR_R] / [3 - 2]}{SSE_F / n-4}$ (3 - 2 = # of β's) $\Rightarrow$ R.R. if $F_{part} > F_\alpha(1, n-4)$

- test if the 2 lines have the same slopes or test for parallelism

$H_0: \beta_3 = 0$; α / t test $= \frac{\hat{\beta}_3}{\sqrt{V_{33} \cdot MSE}} \Rightarrow$ R.R. if $t > t_{\alpha/2; n-4}$ or $t < -t_{\alpha/2; n-4}$
$H_a: \beta_3 \neq 0$

$F_{drop} = \frac{[SSE_R - SSE_F] / [df_{SSE_R} - df_{SSE_F}]}{MSE_F}$

Full: same
R: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$

$= \frac{[SSE_R - SSE_F] / [n-3-(n-4)]}{SSE_F / n-4}$ R.R. if $F_{drop} > F_\alpha(1; n-4)$

- test if the 2 lines are coincident (i.e. same y intercept and the same slope

(+ because cont. var.'s)

$H_0: \beta_2 = \beta_3 = 0$

H_a: @ least one of the β's $\neq 0$; α

$$F_{part} = \frac{[SSR_F - SSR_r] - [df_{SSR_F} - df_{SSR}]}{mSE_F}$$

E: same

R: $y = \beta_0 + \beta_1 x_1 + \varepsilon$ R.R. $F_{part.} > F_{\alpha(2;n-4)}$

$$\frac{[SSR_F - SSR_r]/(3-1)}{SSE_F/n-4}$$

$$F_{dwp} = \frac{[SSE_r - SSE_F] / [df_{SSE_r} - df_{SSE_F}]}{mSE_F} \quad \text{or equiv } \frac{(n-2-(n-4))}{(n-4)}$$

R.R. if $F_{dwp} > F_{\alpha(2;n-4)}$

A3Q2 An experimenter wished to compare the potencies of 3 different drug products. To do this 12 test tubes were inoculated w/ a culture of the virus under study and incubated for 2 days @ 35°C. Four dosage levels (0.2, 0.4, 0.8 + 1.6 mg/tube) were to be used from each of the 3 different drug products A, B, + C, w/ only one dose-drug combo. for each of the 12 test tube cultures

"how well the drug work"

a) Write a general lin. model relating the response (y) to the independent var.'s: "dose" and "drug product" Make x_1 = ln dose Identify the parameters in the model

- here drug product - (A, B or C) ← categorical variable ∴ we need 2 dummy variables:

- $X_2 = \begin{cases} 1; \text{if B} \\ 0; \text{otherwise} \end{cases}$, $X_3 = \begin{cases} 1; \text{if C} \\ 0; \text{otherwise} \end{cases}$

MLR $y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_1 X_2 + \beta_5 X_1 X_3 + \varepsilon$

<u>if we have drug A</u>

$X_2 = 0$
$X_3 = 0$

$y = \beta_0 + \beta_1 X_1 + \beta_2(0) + \beta_3(0) + \beta_4 X_1(0) + \beta_5 X_1(0) + \varepsilon$
$y = \beta_0 + \beta_1 X_1 + \varepsilon \ldots\ldots (A)$

<u>If drug B</u>

$X_2 = 1$
$X_3 = 0$

$y = \beta_0 + \beta_1 X_1 + \beta_2(1) + \beta_3(0) + \beta_4 X_1(1) + \beta_5 X_1(0) + \varepsilon$
$y = \beta_0 + \beta_1 X_1 + \beta_2 + \beta_4 X_1 + \varepsilon$
$y = (\beta_0 + \beta_2) + (\beta_1 + \beta_4) X_1 + \varepsilon \ldots (B)$

<u>if drug C</u>

$X_2 = 0$
$X_3 = 1$

$y = \beta_0 + \beta_1 X_1 + \beta_2(0) + \beta_3(1) + \beta_4 X_1(0) + \beta_5 X_1(1) + \varepsilon$
$y = (\beta_0 + \beta_3) + (\beta_1 + \beta_5) X_1 + \varepsilon \ldots (C)$

A is the base line

β_0 = y intercept for line drug A

$\beta_2 = (\beta_0 + \beta_2) - \beta_0$ = diff. in y-intercs. for the lines A + B

$\beta_3 = (\beta_0 + \beta_3) - \beta_0$ = diff in y-intercs. for the lines A + C

β_1 = slope of line A
$\beta_4 = (\beta_1 + \beta_4) - \beta_1$ = diff. in slopes btwn A + B
$\beta_5 = (\beta_1 + \beta_5) - \beta_1$ = diff. in slopes btwn A + C

b.) It would be reasonible to assume that the 3 lines have a common y-intec. I.e. test if their y intec.'s are =.

$H_0: \beta_2 = \beta_3 = 0$ $\quad n = 12$

$H_a:$ @ least one of $\beta \neq 0$; α

2 partial t-test

test stat. $F_{part} = \dfrac{[SSR_F - SSR_R]/[df_{SSR_F} - df_{SSR_R}]}{MSE_F} =$

reduced: $y = \beta_0 + \beta_1 x_1 + \beta_4 x_1 x_2 + \beta_5 x_1 x_3 + \varepsilon$

$= \dfrac{[SSR_F - SSR_R]/(5-3)}{SSE_F/n-6}$

R.R. Rej H_0 if $F_{part} > F_\alpha(2;6)$

or

$F_{drop} = \dfrac{[SSE_R - SSE_F]/[df_{SSE_R} - df_{SSE_F}]}{MSE_F}$

$= \dfrac{[SSE_R - SSE_F]/(n-4)-(n-6)}{SSE_F/n-6}$

R.R. Rej H_0 if $F_{drop} > F_\alpha(2;6)$

c.) A3.Q2 Test whether the 3 drug lines are parallel (i.e. the same slopes).

in SAS $\ln x$ does not exist
use $\log(x)$

In Real life	In SAS
$\ln X$	$\log(x)$
$\log_{10} X$	$\log 10(x)$
$\log_2 X$	$\log 2(x)$

Not on exam

Multicollinearity (not in textbook)

① If independent (i.e. explanatory x's) variables are uncorrelated, then $r^2_{y,x_1x_2\dots x_K} = r^2_{y,x_1} + r^2_{y,x_2} + \dots + r^2_{y,x_K}$

Are x's independent?

$\left(r^2 = \frac{SSR}{TSS}\right)$ $\quad SSR_{(x_1x_2\dots x_K)} = SSR(x_1) + SSR(x_2) + \dots + SSR(x_K)$

$= SSR_{(x_j \mid \text{all other } x\text{'s})} = SSR_1 - SSR_r = SSR_{(x_1x_2\dots x_K)} - SSR_{(x_1x_2\dots x_K)}$

$= SSR(x_j)$

- If x's are related this is not true

② If the independent (i.e. explanatory) variables are correlated then we have so-called multicollinearity

- If occurs when some of the explanatory variables are related to each other i.e. when they are correlated among themselves
- For example:

$x_3 = 2x_1 + 3x_2 - 5x_4$ ← x_3 is linearly related to x_1 x_2 & x_4

- x_3 does not bring anything new to the model, but can't be disregarded

$\Rightarrow$ • $SSR(x_j \mid \underset{\text{in the model}}{\text{all other } x\text{'s}}) = 0$ (or very small #)

but if $SSR(x_j)$ - it might be a very lrg # as x_j is linearly related to other x's

• X^TX is not invertible (since 1 row is a linear combo. of some other row) and hence there are infinitely many sol.'s to $\beta = (X^TX)^{-1}(X^TY)$

- not rejecting $H_0: \beta_j = 0$ means only that X_j has no additional contribution (linear relationship) w/ Y when the other X's to which it is related are in the model.

Indications of Multicollinearity in a Regression

1. Large correlation coefficient btwn pairs of X's
2. $SSR(X_j \mid \text{all other } X\text{'s present}) \ll SSR(X_j)$ + much, much smaller than
3. Adding or deleting an X variable that is highly related to some of the other X's in the model will cause a huge chng in the val.'s of the β of those X's to which (X_j) is related.
4. Some of the regression coefficients ($\hat{\beta}_j$'s) have opposite signs to what you would expect from theory or common sense.
5. Wide confidence intervals for β_j's, ie. lrg sampling variances (v_{jj} MSE) for β_j's
6. Even though F test $= \frac{MSR}{MSE}$ for testing whether $\beta_1 = \beta_2 = \beta_k = 0$ (i.e for testing for a significant linear relationship btwn Y + @ least one of the X's) will indicate a high significant linear relationship; if however, we test for individual contributions of X's to Y ie. (testing $H_0: \beta_j = 0$ $(j = 1 \ldots K)$, the results are not significant
7. Large (variance inflation factors) VIF's, where $VIF = \frac{1}{1 - R_j^2}$, where R_j^2 is coefficient of determination for the regression of X_j on all other X's.

Not on Exam

i.e. $\hat{x}_j = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \dots + \hat{\beta}_{j-1} x_{j-1} + \hat{\beta}_{j+1} x_{j+1} + \dots + \hat{\beta}_K x_K$

If x_j is highly related to all other x's then R_j^2 is high, then $1 - R_j^2$ is very small

$\Rightarrow \frac{1}{1-R_j^2}$ is very lrg

~~Rule of thumb~~

VIF > 5 - moderate multicollinearity in the data.

VIF > 10 - serious multicollinearity in the data

in SAS
```
proc reg corr;
model y = x1 ... xK / VIF;
run;
```

(Ex) Full: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$ Collinearity?

we get $R = \begin{matrix} \\ x_1 \\ x_2 \\ x_3 \end{matrix} \begin{matrix} \begin{matrix} x_1 & x_2 & x_3 \end{matrix} \\ \begin{bmatrix} 1 & 0.9463 & .982 \\ & 1 & 0.8761 \\ & & 1 \end{bmatrix} \end{matrix}$ ex. correlation btwn $x_1, x_2 = .9463$

high pairwise correlation coefficients

ie $x_1 x_2 = 0.9463$

$x_1 x_3 = .982$

$x_2 x_3 = .8761$

- Start w/ testing x_1 (highest correlation)

- $SSR(x_1 | x_2 x_3) \ll SSR(x_1)$ ← ? true?

Reduced model - $y = \beta_0 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$

$SSR(x_2, x_3) = 322,134,328$

Full: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$ → OVER

$SSR(x_1, x_2, x_3)$ 325,360,308

Not on exam

$y = \beta_0 + \beta_1 x_1 + \varepsilon$

$SSR(x_1) = 314,544,161$

$SSR(x_1 \mid x_2, x_3) = SSR_f - SSR_r = SSR(x_1, x_2, x_3) - SSR(x_2, x_3)$
$= 3,325,980 \ll 314,544,161 \; (SSR(x_1))$ ✓

confirmation x_1 is creating problems ↑

full: $\hat{\beta}_1 = 0.854$, $\hat{\beta}_2 = 44.375$, $\hat{\beta}_3 = 82.285$ ← lrg jump
reduced: $\hat{\beta}_2 = 71.976$, $\hat{\beta}_3 = 117.454$

- full - $H_0: \beta_1 = \beta_2 = \beta_3 = 0$ $\quad F_{(3,17)} = 45.093$, pval = .0001
highly significant linear relationship btwn y and @ least one of the x's
$H_0: \beta_j = 0$

To find out which x:

(x_1) $j = 1$ $\quad t_{(17)} = 1.176 \rightarrow$ pval = 0.2558 } non-significant
(x_2) $j = 2$ $\quad t_{(17)} = 1.411 \rightarrow$ pval = .1762 } linear relationship
(x_3) $j = 3$ $\quad t_{(17)} = 1.3978 \rightarrow$ pval = 0.1805 } btwn y + any of the x's (i.e. don't reject H_0)

- large variance inflation factors
VIF's 12.88, 9.71 + 5.795 > 10 serious multic.

Remedies for Multicollinearity

1. Drop some of the correlated variables from the model if it is reasonable to do so. (Do not drop a variable that is known from theory to belong in the model)
2. Fit the regression equation using $(x_j - \bar{x}_j)$ instead of x_j (especially if we have polynomial model)

Not on exam!

polynomial models

in general, $y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \dots \beta_K X^K$

- only one variable (high correlation)

or

$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots \beta_K X_K + \varepsilon$

where $X_1 = X$

$X_2 = X^2$

$\vdots$

$X_K = X^K$

instead of doing $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \dots \hat{\beta}_K X_K$ fit

$\hat{y} = \hat{\beta}_0^* + \hat{\beta}_1^*(X_1 - \bar{X}_1) + \hat{\beta}_2^*(X - \bar{X}^2) + \dots \hat{\beta}_K^*(X - \bar{X})^K$

* means small difference in β's

- Which X's should be in model? (not in book)
- model (or variable) selection

- how do we decide which variable should be used in the model?
- how do we select (choose) "best-fitting" model?
  (- select highest, R^2_{adj} if different results)
- to test procedures may give different results

X_1, X_2, X_3, X_4

X_1 - r^2 - MSE - choose 2nd smallest full is 1st
X_2 - r^2 - MSE
X_3 - r^2 "
X_4 - r^2 "
X_1X_2 - r^2 "
X_1X_3 - r^2 "
X_1X_4 - r^2 "
X_2X_3 - r^2 "

full is 1st select highest r^2 2nd

$X_1\ X_2\ X_3$
$X_1\ X_2\ X_4$
$X_2\ X_3\ X_4$
$X_1\ X_2\ X_3\ X_4$ → r^2 → most will always be highest

$r^2 = \frac{SSR}{TSS}$

→ MSE → smallest always

→ use if close to r^2 of full

part of exam

- max R^2 criterion
  - we calculate $r^2 = \frac{SSR}{TSS}$ for each possible model and then select "best-fitting" model, the one w/ the "highest" r^2 (other than the full model) if possible)

NOTE Choose as best-fitting model the one w/ higher r^2_{adj} if deciding btwn two models.

- min. MSE criterion
  - is equivalent to max R^2
  - we calculate MSE for each possible model and then select as the "best-fitting" model the one w/ the smallest MSE (other than the full model if possible)

(Ex) n=20 independent pharmacies. Want to predict the sales based on which variables among the following: X_1 = tot flr. space
X_2 = # of parking spots
X_3 = location of the pharmacy
X_4 = space allocated to prescription dept

# of variables in the model

	R^2	variable
one-var.		
$Y = \beta_0 + \beta_1 X_1 + \varepsilon$	.439	X_4
two-var.		
$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$	.666	X_1, X_4
3-var.		
$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$	.691	X_1, X_3, X_4
Full	.694	X_1, X_2, X_3, X_4

since these are close, we could calc R^2_{adj} and use model w/ highest value.

← 2nd highest "best-fitting model is X_1, X_3, X_4

not on exam

- Mallow's Cp Criterion

- It measures the difference of a fitted regression line from the true model, along w/ random errors
- Theory suggests that the "best-fitting" model is one w/ Cp closest to p (where p is the tot. # of parameters in the model including β_0)

We need to calculate Cp statistic for each one of the possible models where Cp stat. is

$C_p = \frac{SSE_p}{MSE_F} - (n-2p)$, where SSE_p = SSE for the model w/ p parameters

NOTE: When p=1 i.e. no X's is $(Y = \beta_0 + \varepsilon)$

$C_p = \frac{TSS}{MSE_F} - (n-2p)$ → if this is closest p, all X's are no good.

	p	Cp	variable
one var	2 (closest to 2)	1.14	X_4
two var	3 (closest to 3)	2.47	X_3, X_4
3 var	4 ← distance 4-2.96 (closest to 4)	2.96	X_1, X_2, X_4
4 var	5 (closest to 5)	4.04	X_1, X_2, X_3, X_4

- must draw graph to find shortest distance.