STA 302 / 1001

Note Title 11/24/2011

unambiguous, accurate, deep

Reminder: Andriy in RW 107/109 Fri Nov 25 1-3 pm

Course Evaluations : Thursday December 1

Partial F-tests Test a subset of β's (excluding β_0) are 0 simultaneously.

For these tests, can use "Sequential Sums of Squares" or "Type I SS" to calculate difference in SS needed in numerator of test stat. for partial F-test

Gives the additional contribution to SS Reg. each variable gives over and above the variables listed above it. Depends on order in which variables are named in model statement.

Want variables whose coefficients you are testing to be listed last.

In SAS: add SS1 option to model statement.

__Output:__

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Type I SS
Intercept	1	37.84583	3.62869	10.43	<.0001	75634
time	1	12.15833	2.74303	4.43	0.0004	886.95042
i150	1	29.35000	4.75107	6.18	<.0001	1409.73075
i300	1	20.22500	4.75107	4.26	0.0005	654.36800
i450	1	15.97500	4.75107	3.36	0.0037	538.68000
i600	1	6.12500	4.75107	1.29	0.2146	75.61500
i750	1	1.60000	4.75107	0.34	0.7404	5.12000

SSReg for model: $Y = \beta_0 + \beta_1 time + e$

SSReg for model: $Y = \beta_0 + \beta_1 time + \beta_2 i150 + e = 886.95 + 1409.73$

$886 + 1409 + 654 + 538 + 76 + 5 \doteq SSReg$ for full model

Partial F-test test statistic

$$F_{obs} = \frac{\left(RSS_{red} - RSS_{full}\right) / \#\text{ of } \beta\text{'s being tested}}{MSE_{full}}$$

$= SSReg_{full} - SSReg_{reduced}$

For our example, $SSReg_{full} - SSReg_{reduced}$

$= 1409 + 654 + 538 + 76 + 5$

$= 345091 - 767.47$

Suppose $H_0: \beta_{i750} = 0$ vs $H_a: \beta_{i750} \neq 0$

assumes all other variables are in model.

t-test: test stat = .34 P-value = .74

Partial F-test: test stat = $\dfrac{5.12/1}{45.15}$ → small ⇒ P-value will be big

$= (.34)^2$

Under H_0, this test stat $\sim F(1, 17)$

↖ df error, full

New question: Does the way light intensity affects mean of # of flowers/plant depend on timing?

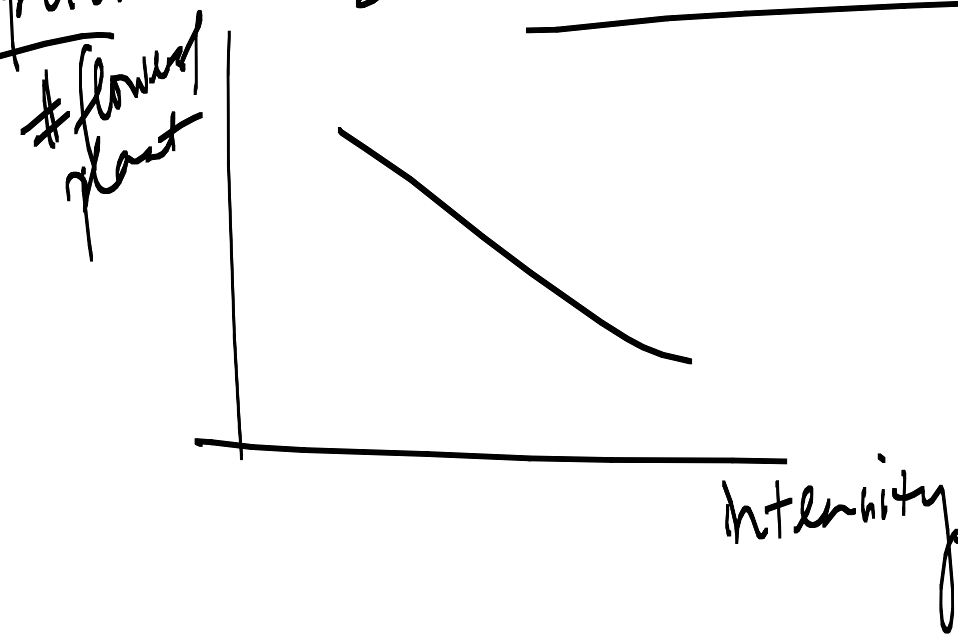
Now: model intensity as quantitative variable

Y: # of flowers/plant

For models, 3 options

Option 1: Same lines for each timing

"COINCIDENT"



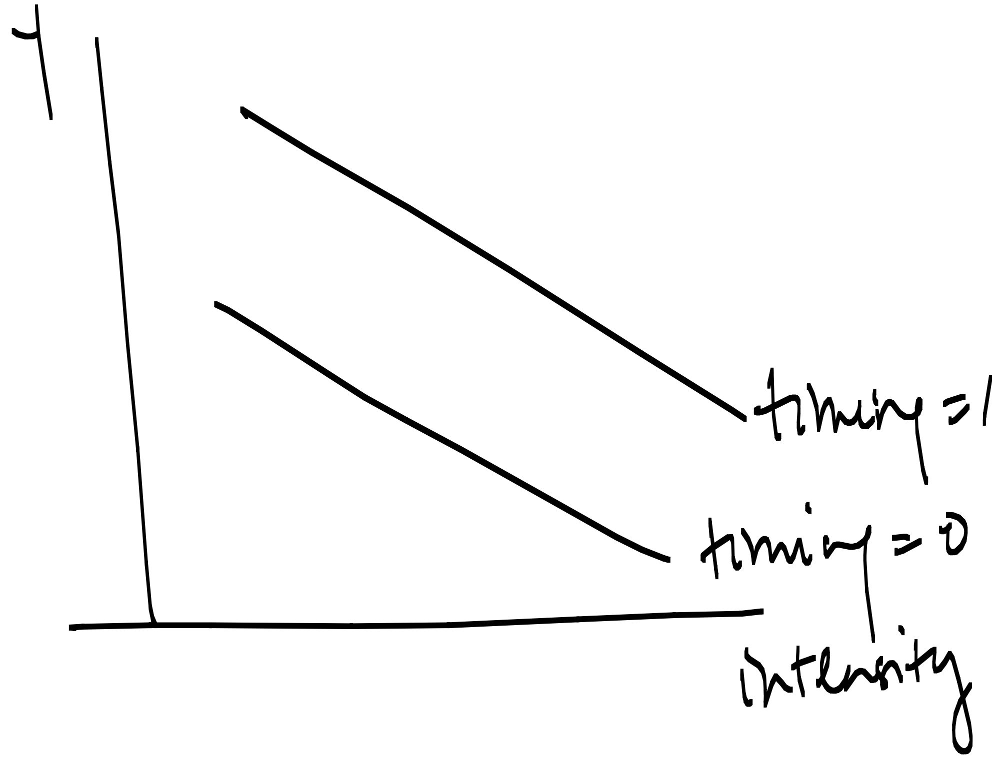
Model:

$Y = \beta_0 + \beta_1 \text{ intensity} + e$

- not a good idea here

From plot, we get higher values of Y for timing = 1

Option 2 **Parallel lines**



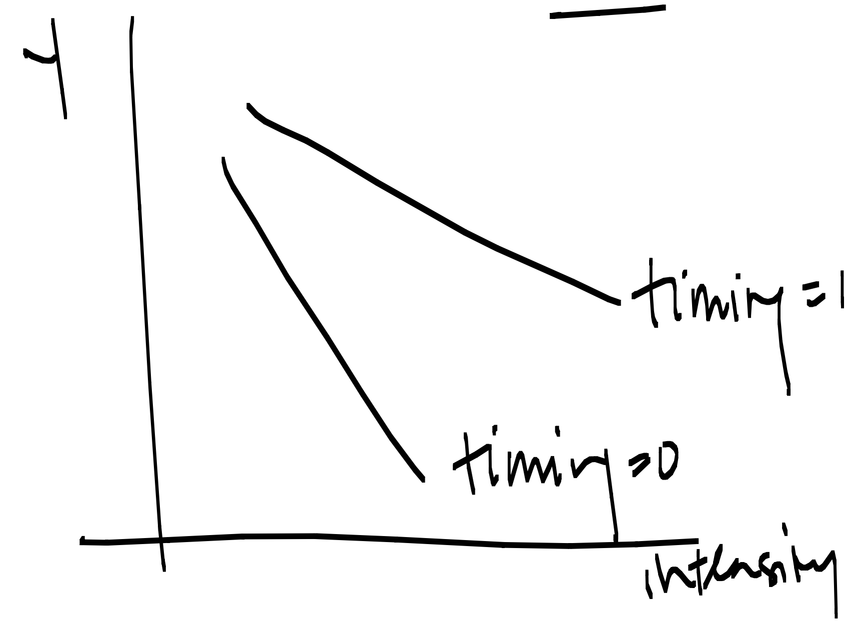
Model:

$$Y = \beta_0 + \beta_1 \text{timing} + \beta_2 \text{intensity} + e$$

When timing = 0: Model: $Y = \beta_0 + \beta_2 \text{intensity} + e$

when timing = 1: Model: $Y = (\beta_0 + \beta_1) + \beta_2 \text{intensity} + e$

Option 3 Separate lines for each timing
Unrelated regressions



Model: $Y = \beta_0 + \beta_1\,\text{timing} + \beta_2\,\text{intensity} + \beta_3\,\text{timing} * \text{intensity} + e$

If timing $= 0$, Model: $Y = \beta_0 + \beta_2\,\text{intensity} + e$

If timing $= 1$, Model: $Y = (\beta_0 + \beta_1) + (\beta_2 + \beta_3)\,\text{intensity} + e$

Test whether resulting change in y when intensity changes is same for early and late timings:

$H_0: \beta_3 = 0$

Is intensity - # of flowers/plant relationship the same for early and late timings? Do the have the same line?

$H_0: \beta_1 = \beta_3 = 0$ (tests both same slopes + same intercept)

Option 3: another example of an interaction term
Now - between categorical variable and a quantitative variable

- Result: lines with different slopes

Called: "<u>Analysis of Covariance</u>".

Conclusions

From unrelated regressions model:

There is no evidence that effect of light intensity on # of flowers/plant differs with timing ($p = .91$)

If interactions, difficult (impossible?) to talk about effects of the individual predictor variables because they depend on value of other.

Since coef. of interaction is not stat. sig. different from 0, remove it so that we can talk about individual effects of timing and intensity.

There is strong evidence $(p < 0.001)$ that light intensity affects # of flowers/plant over and above timing.

For same timing, increasing light intensity by 100 $\mu mol/m^2/s$ decrease # of flowers/plant on average by 4

95% CI for this decrease

$$100\left[-.04047 \pm \underbrace{t_{21,.025}}_{2.080}(.00513)\right]$$

$$= (-5.1, -3.0) \qquad 21 = df_{error}$$

There is strong evidence ($p = .0001$) that timing affects # of flowers/plant over and above light intensity.

For same intensity, for early timing get on average 12.2 more flowers/plant.

95% CI for this increase

$12.2 \pm 2.080\,(2.63)$

Could have fit 2 separate regression lines by splitting data into the 12 early observations and 12 late observations

Advantages of using Analysis of Covariance (ANCOVA):

- have tests for equal slopes & intercepts
- have higher df_{error} ⇒ power increased, narrower CIs
- get est of error variance based on 24 observations rather than 12 (better est.)

<u>Disadvantage</u>: implicit assumption that both groups have same error variance

New <u>Example</u>: House <u>prices</u> in Chicago (a long time ago)

Y - selling price (in \$1000's)

X's - bdr - # of bedrooms

flr - floor space in ft^2

fp - # of fireplaces

rms - # of rooms

st - indicator variable for storm windows

lot - frontage (in feet)

bth - # of baths

gar - # of garage parking spaces

Comments from pairwise scatterplots:

- no obvious non linear Y-X relationships; no obvious need for transformations
- perhaps an influential point (house with 8 bedrooms)
- all X's seem to have some relationship with Y
- also relationships between some of X's (in particular, bdr & rms)

From correlations: (mostly focussed on p-values / signs rather than values)

- all X's stat sig with Y (appropriate test for sig?? NO)

- many pairs of X s that are linearly related