STA 303 / 1002

Note Title 1/30/2012

Pygmalion Example: Two-way Analysis of Variance

Does mean treatment effect differ with company?

pygmalion vs control

H_0: $\beta_{11} = \cdots = \beta_{19} = 0$ (coef. of interaction terms)

H_a: at least one not = 0

Partial F-test: Given in Type III SS

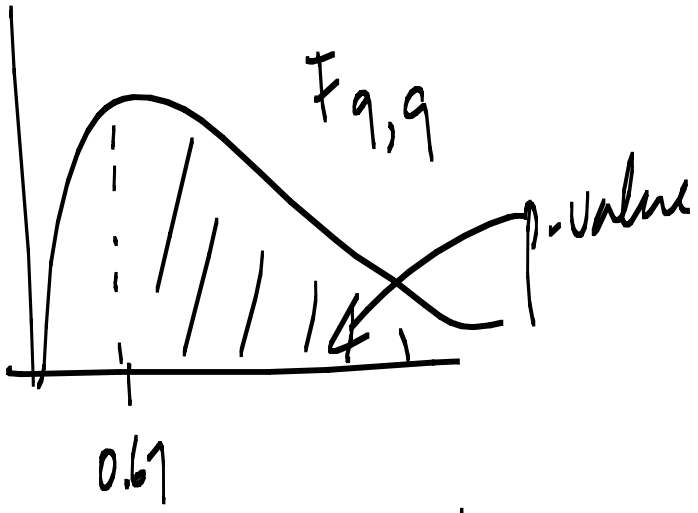
Type III SS = 311.46 = difference in SSReg between full and reduced models

(reduced model doesn't have 9 interaction terms)

MSE for full model = 51.89

Test stat $\frac{311.46/9}{51.89} = 0.67$

If H_0 is true this is an observation from a F distribution with 9, 9 ← df for error for full model



p-value is large

Data are consistent with 0 coefficients for interaction terms that treatment effect differs with company.

No evidence

Next step: Since interaction is not significant, fit additive model so that we can assess "main effects" (Are there differences among companies? or treatments?)

Model $Y_i = \beta_0 + \beta_1 I_{PYG,i} + \beta_2 I_{comp1,i} + \beta_3 I_{comp2,i} + \cdots + \beta_{10} I_{comp9,i} + \epsilon_i$

Company	$E(Y_i)$ Treatment: Pygmalion	$E(Y_i)$ Control	Treatment effect: Pyg − Control
2	$\beta_0 + \beta_1 + \beta_2$	$\beta_0 + \beta_2$	β_1
3	$\beta_0 + \beta_1 + \beta_3$	$\beta_0 + \beta_3$	β_1
4			
5			
⋮	⋮	⋮	⋮
10	$\beta_0 + \beta_1$	β_0	β_1

Test: Is there a difference in mean score between

Pygmalion and control groups?

$H_0: \beta_1 = 0$ vs $H_a: \beta_1 \neq 0$

Test stat: $F_{obs} = 7.84$

This is an observation from a F distribution with 1, 18

p-value = .0119

We have evidence of a difference in mean score between pygmalion and control platoons (over and above differences among companies)

Differences among companies?

$H_0: \beta_2 = \dots = \beta_{10} = 0$ vs H_a: at least one not 0

Test stat. $F_{obs} = 1.75$, obsn from $F(9, 18)$ under H_0

p-value = 0.1484

No evidence of differences among companies

Means statement output

	N	Mean	Std
Pygmalion	10	78.7	2.57
Control	19	71.63	7.24

71.63 → $\frac{\sum Y_i \text{ (ctrl)}}{n_{ctrl}}$

7.24 → s.d. of 19 control observations

LSmeans Statment output

Pygmalion	78.7	Values estimated from model.
Control	71.48	

Estimate of $\hat{E}(Y_i)$ from model:

Company	Pygmalion	Control
1	72.62 - 4.33 + 7.22	72.62 - 4.33
2	72.62 + 1.13 + 7.22	72.62 + 1.13
3		
⋮		
10	72.62 + 7.22	72.62

LSMEANS FOR TREATMENT are averages over 10 companies

Lsmeans for Pygmalion is average of 10 #'s in Pygmalion column

" " Control " " " 10 " " Control column

Here LSMEANS for Control is not equal to Mean of the 19 Control observations because there is not an equal number of control observations per company.
(Company 3 only had 1 control platoon, the rest of the companies had 2)

Note: Because of this lack of balance, can't use the Type III SS to assess main effects for a model that has an interaction
(SAS uses LSMEANS in its calculation of Type III SS so it is no longer equivalent to the Partial F-test)

for main effects when model has interaction term

Check model assumptions

From plots - no outliers
- perhaps decreasing variance
- normality OK

Independent observations? OK if we assume platoons weren't interacting

→ p-values may only be approximate

Conclusion: Evidence of a difference in mean score between Pygmalion and control platoons ($p = .01$)
(will consider this only weak evidence since I have some concern about the variance estimate)
On average, Pygmalion platoons scored higher (mean = 78.7) than control platoons (mean 71.6)

Estimation of Regression Parameters

Model: $Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip} + \varepsilon_i$

Least squares find $\hat{\beta}_0, \ldots, \hat{\beta}_p$ to minimize

$$RSS = \sum \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \cdots \hat{\beta}_p x_{ip} \right)^2$$

Because one assumption is $\varepsilon_i \sim N(0, \sigma^2)$

then $Y_i \mid X_{i1} = x_{i1}, \ldots, X_{ip} = x_{ip}$

$$\sim N\left(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}, \sigma^2\right)$$

So Joint density function of

$$Y_1 \mid X_{11} = x_{11}, \ldots, X_{1p} = x_{1p}, \quad \ldots, \quad Y_n \mid X_{n1} = x_{n1}, \ldots, X_{np} = x_{np}$$

(Since Y_i's uncorrelated because ε_i's are uncorrelated)

is $\prod_{i=1}^{n} f(y_i \mid x_i)$

$$= \left(\frac{1}{\sqrt{2\pi}\,\sigma}\right)^n \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_{i1} - \dots - \beta_p x_{ip})^2\right\}$$

Find β's that Minimize RSS is equivalent to finding the β's that maximize the joint density function of the y's conditional on the Xs

Least squares is equivalent to maximum likelihood estimation under the assumption of normally distributed error.

New example

Donner Party Example 1846 - covered wagon party left Illinois for California

- got stuck in snow in Sierra Nevada mountains in October
- rescued in April 1847

Data: for adults ($\geq$ 15 years old)

- age
- sex
- survived or not

Want to model odds of survival based on age and sex.

Y_i is a binary variable (survived, died)

Model: (Binary) Logistic Regression

$$Y_i \mid X_i = \begin{cases} 1 & \text{if response in category of interest} \\ 0 & \text{otherwise} \end{cases}$$

$$Y_i \mid X_i \sim \text{Bernoulli}(\pi_i)$$

$$E(Y_i \mid X_i) = \pi_i$$

$$\text{Var}(Y_i \mid X_i) = \pi_i(1-\pi_i)$$

Logistic regression model is an example of a Generalized Linear Model

Generalized Linear Models

Have response: Y.

+ set of explanatory variables $X_1, \ldots, X_p$

Could have $Y|X \sim$ Bernoulli
or $\sim$ Binomial
or $\sim$ Normal (regular regression)
or $\sim$ Poisson
or $\sim$ Gamma

Want to model $E(Y)$ as function of $X_1, \ldots, X_p$

often the function is $f(X) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$
"linear" $= X\beta$

Key idea of Generalized Linear Models

LINK FUNCTION, function g s.t.

$$g(E(Y)) = X\beta$$

Some choices for the link function

① Identity link: $g(E(Y)) = E(Y)$

Model: $E(Y) = X\beta$

This is regression (STA 302)

Usual distribution:

$$Y|X \sim \text{Normal}$$