STA 302 / 1001

Note Title 11/17/2011

mean, bigmove, accurate, dep

Assignment 3 - Q1 - don't worry if you don't have scatterplots of all of the pairs of explanatory variables

- new code on assignment web site.

Assignment 2 - slight adjustment made to marks adjusted for TA differences

Rainfall example:

Does the relationship of rainfall with yield depend on year?

New model:

$$\text{Yield} = \beta_0 + \beta_1 \text{rain} + \beta_2 \text{rain}^2 + \beta_3 \text{year} + \beta_4 \text{rain} * \text{year} + e$$

rain*year is an "interaction" term

Two explanatory variables are said to interact if the effect that one of them has on the response depends on the value of the other.

What we can learn from the rain*year interaction term: Does rain - yield relationship change with year?

Fitted model:

$$\hat{Yield} = -1909.5 + 158.8\,rain - .186\,rain^2 + 1.008\,year - .08064\,rain*year$$

From t-tests, conclude that we have evidence that coefficient of rain*year is stat. sig. different from 0 given the other terms are in the model.

Also this model has smaller MSE and large Adj R^2 than additive model

Conclude that adding the interaction term is worthwhile

Year goes up by 10 years.

Estimated change in mean of yield is

$= 1.008(10) - .08064\,\text{rain}(10)$

So change over time in mean of yield is adjusted for rain

Why? Irrigation lessening rain effect

effect of year is greater for years with little rain

Should we routinely add interaction terms?

No

Why not? Consider assignment 3 - 14 explanatory variables

$\binom{14}{2} = 91$ possible 2-way interaction terms

(not including 3-way or higher order interaction terms)

When to add interaction terms?

- research question that has to do with an interaction

Standard practice - If an interaction term is in the model, also include the individual terms for the predictor variables, even if their coef. are not stat. sig. different from 0

Avoids this logical inconsistency: The effect of X_2 on Y depends on X_1 but there is no effect of X_1

Model that also includes $rain^2 * year$ interaction

Model: $Yield = \beta_0 + \beta_1 rain + \beta_2 rain^2 + \beta_3 year + \beta_4 rain * year + \beta_5 rain^2 * year + \varepsilon$

F-test: $H_0: \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$

H_a: at least one of $\beta_1, \ldots, \beta_5$ not 0

$p < .0001$

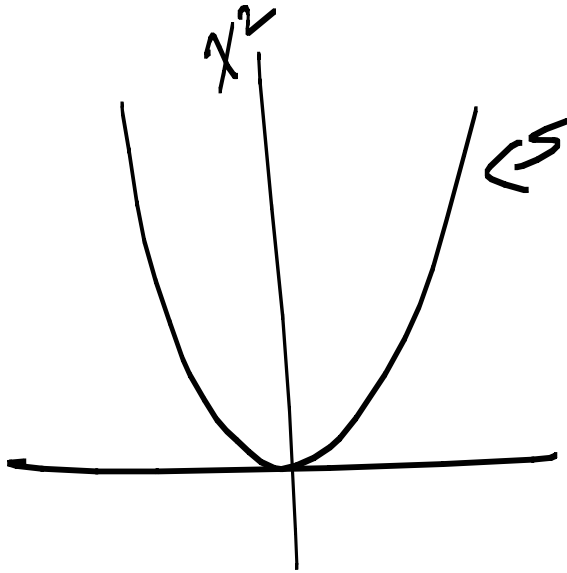
Conclude that we have strong evidence that at least one of coefficients is not 0

But p-values for 5 tests, $H_0: \beta_j = 0$ vs $H_a: \beta_j \neq 0$ $j = 1, \ldots, 5$

have p-values ranging from .417 to .876

Do we conclude that all of $\beta_1, \ldots, \beta_7$ are 0?

No The t-tests are for the effect of one explanatory variable given that the others are in the model.



← approximately straight for large values of n.

Mauna Loa CFCs example

- plot of standardized residuals versus time indicates a quadratic model may be appropriate
- added time^2 to model
  ⇒ SAS: "Model is not full-rank"

Why? For large x, x and x^2 are close to linearly related

$\Rightarrow$ columns of X are linearly related

$\Rightarrow$ can't invert $X'X$

$\Rightarrow$ no unique solution for least squares estimates

Solution - centre the explanatory variable

use $(x - \bar{x})$ and $(x - \bar{x})^2$

Allows us to fit a quadratic model.

But the plots of the residuals vs $(x - \bar{x})$ still show a pattern

Problem: autocorrelation → observations close together in time tend to be more closely correlated than observations further apart in time

Solution: Time Series model

Multiple Regression example 2 Meadowfoam – plant

Effects of light on meadowfoam flowering

- a randomized experiment

6 light intensities: 150, 300, 450, 600, 750, 900 $\mu mol/m^2/s$

2 timings { early (coded 1)
late (coded 0)

$\Rightarrow$ 12 treatments

$n = 24$, 2 replications of each treatment

Response: average # of flowers on 10 plants in one pot

Questions of interest: What is the effect of timing?
What is the effect of light intensity?
on number of flowers/plant

Define indicator variable $time = \begin{cases} 1 & \text{if timing early} \\ 0 & \text{if timing late} \end{cases}$

Will also treat intensity as a categorical variable,

Why? - I can do this because intensity has a small number of values with multiple observations for each.

- May be useful for learning which

intensity is "best". (highest value of y on average)

- Doesn't impose a particular form of relationship (eg. linear or quadratic) on intensity-y relationship.

Allows

Define 6 new indicator variables

$$i150 = \begin{cases} 1 & \text{if intensity is 150} \\ 0 & \text{otherwise} \end{cases}$$

$$i300 = \begin{cases} 1 & \text{if intensity is 300} \\ 0 & \text{o/w} \end{cases}$$

$$i900 = \begin{cases} 1 & \text{if intensity is 900} \\ 0 & \text{o/w} \end{cases}$$

If try to fit linear regression model with indicator variable for timing + 6 indicator variables for intensity

SAS: "Model is not full-rank"

8x42
X
r1
Limit

Columns for $i150 + i300 + \cdots + i900$ gives a column of 1's

$\Rightarrow$ linear dependence among columns

Using all 6 indicator variables is redundant:
if 5 are 0, you know the 6th is 1

In general, for categorical variable with K categories, need K-1 indicator variables

So, fit a model with only 5 of the 6 indicator variables for intensity

$$y = \beta_0 + \beta_1 i150 + \beta_2 i300 + \beta_3 i450 + \beta_4 i600 + \beta_5 i750 + \beta_6 time + \epsilon$$

Fitted model:

$$\hat{y} = 37.8 + 29.4\, i150 + 20.2\, i300 + 16.0\, i450 + 6.1\, i600 + 1.6\, i750 + 12.2\, time$$

When light intensity is 150 $\mu mol/m^2/s$ and time is early, what is estimate of mean # of flower/plant?

Answer: $37.8 + 29.4 + 12.2$

What does the intercept estimate?

mean # of flowers/plant when intensity is 900 and timing is late