Topic: Dummy Variable

(21) Bars of Soaps are scored for their appearance in a manufacturing operation. These scores are on a 1–10 scale, and the higher the score the better. The difference between operator performance and the speed of the manufacturing line is believed to measurably effect the quality of the appearance. The following data were collected on this problem

Operator	Line Speed	Appearance (Sum for 30 Bars)
1	150	255
1	175	246
1	200	249
2	150	260
2	175	223
2	200	231
3	150	265
3	175	247
3	200	256

a. Using dummy variables, fit a multiple regression model to these data.

22 ~~TOOP~~ Topic - Dummy variable

20. An experimenter suggests the following dummy variable scheme to separate possible level differences among six groups. Is it a workable one?

Z_0	Z_1	Z_2	Z_3	Z_4	Z_5
1	1	−1	−1	−1	−1
1	−1	2	−1	−1	−1
1	−1	−1	3	−1	−1
1	−1	−1	−1	4	−1
1	−1	−1	−1	−1	5
1	−1	−1	−1	−1	−1

Answer. For two ~~cate~~ groups, we use two dummy variables. $(X_0, Z) = (1, 0)$ for ~~group~~ group A

$(1\ 1)$ for group B

If the matrix $X = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}$ has a non zero determinant the setup will work. For the Turkey data, the corresponding matrix is $\begin{pmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}$ and the determinant is 1.

These six column vectors are clearly linearly independent, so the system will work.

(23) (22) Topic: Dummy variable

(22) Here is another six-group scheme. Will it work?

Z_1	Z_2	Z_3	Z_4	Z_5
1	1	1	1	1
2	3	3	2	1
3	2	3	3	2
3	3	2	3	3
2	3	3	2	3
1	2	3	3	1

Answer: Add the $Z_0 = (1\ 1\ 1\ 1\ 1\ 1)'$ vector to all the others. The resulting six vectors are clearly independent, so the system will work.

Topic: Dummy variable

(24) (23) (23) An experimenter says he feels the need to fit two straight lines to ten equally spaced points, the first five of which he believes are on one line, and the second five on another line. He proposes to use dummy system A, below. The statistician on the project suggests system B. Who is right?

System A				System B			
X_0	X_1	X_2	X_3	X_0	X_1	X_2	X_3
1	1	0	−1	1	1	0	0
1	2	0	−1	1	2	0	0
1	3	0	−1	1	3	0	0
1	4	0	−1	1	4	0	0
1	5	0	−1	1	5	0	0
1	0	0	0	1	5	1	1
1	0	1	0	1	5	2	1
1	0	2	0	1	5	3	1
1	0	3	0	1	5	4	1
1	0	4	0	1	5	5	1

Answer: Both system A & B are ~~ca~~ OK.

(24) : (25) Topic: Dummy variables

(4)

"Look at these data," a friend moans. "I don't know whether to fit two straight lines, one straight line, or what." You look at his notes and see that he has two sets of (X, Y) data, given below, which both cover the same X-range. How do you resolve his dilemma? Describe, and give model details, and "things he needs to do."

Set A: X	Y	Set B: X	Y
8	5.3	9	5.1
0	0.9	7	4.4
12	7.1	8	5.2
2	2.4	6	3.8

Solutions: Two sets of data, two dummy variables;

$$(Z_0, Z) = (1, 0) \text{ for set A}$$
$$= (1, 1) \text{ for set B.}$$

We can fit the model ~~$Y = \beta_0 + \beta_1 X + \alpha_0 Z$~~

$$Y = \beta_0 + \beta_1 X + \alpha_0 Z + \alpha_1 XZ + \epsilon.$$ The fitted equation is

$$\hat{Y} = 1.142 + 0.506\,X - 0.0418\,Z - 0.036\,XZ.$$

We can test if a single line is sufficient by testing $H_0: \alpha_0 = \alpha_1 = 0$ ag. H_1: H_0 is not true.

The extra sum of squares $F = \dfrac{0.1818/2}{0.3272/4} = 1.11$

So a single straight line seems appropriate.

(25) (26) Topic: Dummy variables

An experimenter has two sets of data, of (X, Y) type, and wishes to fit a quadratic equation to each set. She also wishes (later) to test if the two quadratic fits might be identical in "location" and "curvature" but have different intercept values. Explain how you would set this up for her.

Answer: She should fit the six-parameter model

$$Y = \beta_0 + \beta_1 X + \beta_{11} X^2 + Z(\alpha_0 + \alpha_1 X + \alpha_{11} X^2) + \epsilon$$

and for testing ~~not~~ for two "parallel" quadratics

$H_0: \alpha_1 = \alpha_{11} = 0$ is appropriate.

(27) (26) Topic: Dummy variables

You have two sets of data involving values of X and Y, but you are unsure whether to fit the data separately or together. You consider and fit the six-parameter model

$$Y = \beta_0 + \beta_1 X + \beta_{11} X^2 + Z(\alpha_0 + \alpha_1 X + \alpha_{11} X^2) + \epsilon,$$

where Z is a dummy variable whose value is -1 for "set A" and 1 for "set B."

a. What hypothesis would you test to answer the question: "Will a single quadratic model fit all the data?"
b. What hypothesis would you test to answer the question: "Will a single straight line model fit all the data?"
c. How would you obtain *separate* quadratic fits to the two data sets?
d. If a data point in set A and a data point in set B had the same X-value, would those two points be "repeat points" in the fit of the full model written out above?

Answer:

(a) $\alpha_0 = \alpha_1 = \alpha_{11} = 0$

(b) $\beta_{11} = \alpha_0 = \alpha_1 = \alpha_{11} = 0$

(c) By fitting the model as given & setting $Z = 0$ for set A & $Z = 1$ for set B.

(d) No, their Z values would be different.

28 ~~27~~ Topic: polynomial Regression.

D. A finished product is known to lose weight after it is produced. The following data demonstrate this drop in weight.

Time After Production, t	Weight Difference (in $\frac{1}{16}$ oz), Y
0	0.21
0.5	−1.46
1.0	−3.04
1.5	−3.21
2.0	−5.04
2.5	−5.37
3.0	−6.03
3.5	−7.21
4.0	−7.46
4.5	−7.96

Requirements

1. Using orthogonal polynomials, develop a second-order fitted equation that ~~represent~~ represents the loss in weight as a function of time after production.

Answer: $\hat{Y} = -0.0037 - 2.8008\,t + 0.2314\,t^2$

29 ~~28~~ Topic: polynomial Regression.

E. Nine equally spaced levels of a dyestuff were applied to apparently identical pieces ~~of clot~~ of cloth. The color ratings awarded, in order of increasing dyestuff levels, were

$Y = 11,\ 12,\ 10,\ 12,\ 11,\ 14,\ 16,\ 22,\ 28.$

Find a suitable polynomial relationship between Y and the level of dyestuff using ~~orthogo-nal~~ orthogonal polynomials.

Answer: $\hat{\alpha}_0 = 15.111$, $\hat{\alpha}_1 = 1.8667$, $\hat{\alpha}_2 = 0.165945$

$\hat{\alpha}_3 = 0.072727$

The analysis of variance is shown below

Source	df	SS
α_0	1	2055.11
α_1	1	209.07
α_2	1	76.33
α_3	1	5.24
Residual	5	4.36
Total	9	2350.00

The cubic term is not significant. A suitable model is

$$\hat{Y} = \hat{\alpha}_0 + \hat{\alpha}_1 P_1(x) + \hat{\alpha}_2 P_2(x)$$

$$= 11.792 + 1.8667x + 0.4978x^2$$

This model accounts for $R^2 = .9678$ of the total variation about the mean.

Topic: Generalized linear models

Suppose we have n observations of variables

$X_1, X_2, \ldots, X_K, Y$, where the X's are

predictors and Y is a response variable. If the

Y's are binomial ratios $Y_i = \frac{r_i}{m}$, say, where

m is constant, what types of analyses are

feasible?

Answer: we might consider fitting a linear

model function

$$f(Y_i) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_K X_{iK} + \epsilon_i$$

where $f(Y_i) = \ln\left(\frac{Y_i}{1-Y_i}\right)$

Topic: Non-Linear Estimation

31

Estimate the parameter θ in the nonlinear model

$$Y = e^{-\theta t} + \epsilon$$

from the following observations:

t	Y
1	0.80
4	0.45
16	0.04

Construct an approximate 95% confidence interval for θ.

Answer: Rao..

Answer: $\hat{\theta} = 0.20345$ & 95% Confidence interval for θ is: $0.179 \leq \theta \leq 0.231$.

32. Topic: Non-linear Estimation

B. Estimate the parameter θ in the nonlinear model

$$Y = e^{-\theta t} + \epsilon$$

from the following observations:

t	Y
0.5	0.96, 0.91
1	0.86, 0.79
2	0.63, 0.62
4	0.48, 0.42
8	0.17, 0.21
16	0.03, 0.05

Construct an approximate 95% confidence interval for θ.

Answer: $\hat{\theta} = 0.20691$ & 95% Confidence interval for θ: $0.190 \leq \theta \leq 0.225$.

Topic: Non-linear Estimation.

33. TRUE / False Question

(a) The model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_2 X_3 \ln X_1 + \epsilon$ is a linear model

(b) The model $Y = \beta_0 + \beta_1 X + \beta_2 (\beta_3)^X + \epsilon$ is a linear model

(10)

(c) The model $Y = \theta + \alpha X_1 X_3 + \beta X_1 + \alpha\beta X_2 + \epsilon$ with parameters (θ, α, β) is non-linear

(d) The model $Y = \beta_0 + \beta_1 (X_1 - X_2) + \beta_2 (X_1 - X_2)^2 + \epsilon$ is a non-linear model

Answer

(a) TRUE (b) FALSE (c) TRUE (d) FALSE.

(34) Topic : SELECTING THE BEST Regression Equation

~~3~~. The demand for a consumer product is affected by many factors. In one study, measurement on the relative urbanization, educational level, and relative income of nine geographic areas were made in an attempt to determine their effect on the product usage. The data collected were as follows:

Area Number	Relative Urbanization X_1	Educational Level X_2	Relative Income X_3	Product Usage Y
1	42.2	11.2	31.9	167.1
2	48.6	10.6	13.2	174.4
3	42.6	10.6	28.7	160.8
4	39.0	10.4	26.1	162.0
5	34.7	9.3	30.1	140.8
6	44.5	10.8	8.5	174.6
7	39.1	10.7	24.3	163.7
8	40.1	10.0	18.6	174.5
9	45.9	12.0	20.4	185.7
Means	41.86	10.62	22.42	167.07
Standard deviations	s_1 4.1765	s_2 0.7463	s_3 7.9279	s_4 12.6452

The correlation matrix is

	X_1	X_2	X_3	Y
X_1	1	0.684	−0.616	0.802
X_2	0.684	1	−0.172	0.770
X_3	−0.616	−0.172	1	−0.629
Y	0.802	0.770	−0.629	1

Requirements

a. Use the stepwise procedure to determine a fitted first-order model using $F = 2{\cdot}00$ for both entering and rejecting variables.

b. write out the analysis of variance table and comment on the adequacy of the final fitted equation after examining residuals.

Ans: (a) The stepwise procedure enters X_1 ($F = 12.60$), enters X_2 ($F = 2.04$), enters X_3 ($F = 3.62$). Now however, X_1 has weakened with partial $F = 0.06$. X_1 is rejected and both X_2 and X_3 remain. The final equation is

$$\hat{Y} = 63.021 + 11.517 X_2 - 0.816 X_3 .$$

(b)

ANOVA

Source of variation	df	SS	MS	F
Regression	2	1079.12	539.56	16.181
Residual	6	200.08	33.34	
Total	8	1279.2		

Regression model is significant. Residual plots reveal no problem.

(35) Topic Regression model with Auto Correlated errors

Q. The following 24 residuals from a straight line fit are equally spaced in time and are given in time sequential order. Is there any evidence of lag-1 serial correlation, do you think? (Use a two-sided test at level $\alpha = 0.05$.)

8, −5, 7, 1, −3, −6, 1, −2, 10, 1, −1, 8, −6, 1, −6, −8, 10, −6, 9, −3, 3, −5, 1, −9

Answer: The Durbin-Watson ~~st~~ test statistic $d = 2.67$,

So $4-d = 1.33$. This is the upper end of the range $(d_L, d_U) = (1.16, 1.33)$ for $n=24$, $K=1$, in the 2.5% table. So it is not significant, and there is no evidence of lag-1 serial correlation.

36 Topic:
Regression model with Autocorrelated errors

1~~4~~3 Consider the simple linear regression model $y_t = \beta_0 + \beta_1 x + \varepsilon_t$, where the errors are generated by the second-order autoregressive process

$$\varepsilon_t = \rho_1 \varepsilon_{t-1} + \rho_2 \varepsilon_{t-2} + a_t$$

Discuss how the Cochrane–Orcutt iterative procedure could be used in this situation. What transformations would be used on the variables y_t and x_t? How would you estimate the parameters ρ_1 and ρ_2?

Answer: Consider the transformation

$$y_t' = y_t - \rho_1 y_{t-1} - \rho_2 y_{t-2} \quad \&$$

$$x_t' = x_t - \rho_1 x_{t-1} - \rho_2 x_{t-2}.$$

Remaining part is left to the reader.

Topic: Generalized linear models

(37) The pareto ~~distribution~~ probability density function is

$$f(u,\theta) = \theta u^{-(1+\theta)}, \quad u \geq 0$$

show that the pareto is a member of exponential family.

Ans

$$f(u,\theta) = \exp\{-(1+\theta)\ln u + \ln\theta\}$$

(38) Topic: Model Adequacy checking

Why do we plot the residuals $e_i = Y_i - \hat{Y}_i$ against the $\hat{Y}_i$ and not against the Y_i, for the usual linear model?

Answer: Because the e's and Y's are usually correlated but the e's and the $\hat{Y}$'s are not.

Topic: Simple linear regression

39

~~3.26~~ Consider the simple linear regression model

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

where the intercept β_0 is known.

a. Find the least-squares estimator of β_1 for this model. ~~Does this answer seem reasonable~~?

b. What is the variance of the slope ($\hat{\beta}_1$) for the least-squares estimator found in part a?

c. Find a $100(1 - \alpha)$ percent confidence interval for β_1. Is this interval narrower than the estimator for the case where both slope and intercept are unknown?

Answer 1

(a) $$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (y_i - \beta_0) x_i}{\sum x_i^2}$$

(b) $$V(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^{n} x_i^2}$$

(c) $$\hat{\beta}_1 - t_{\alpha/2, n-1} \sqrt{\frac{MS_{Res}}{\sum x_i^2}} \leq \beta_1 \leq \hat{\beta}_1 + t_{\alpha/2, n-1} \sqrt{\frac{MS_{Res}}{\sum x_i^2}}$$

Yes.

Topic

(40) Measurement Errors and Calibration problem

A mechanical engineer is calibrating a thermocouple. He has chosen 16 levels of temperatures evenly spaced over the interval 100 - 400°C. The ~~actual~~ actual temperature x and the observed reading on the thermocouple y are shown in the Table below. Suppose a new observation on temperature of $y_0 = 200°C$ is obtained using the thermocouple. Find a point estimate of the actual temperature, from the calibration line.

Answer: The straight line model is

$$\hat{y} = -6.67 + 0.953x$$

$$\hat{x}_0 = \frac{y - \hat{\beta}_0}{\hat{\beta}_1} = \frac{200 - (-6.67)}{0.953}$$

$$= 216.86^\circ C$$