TIME SERIES

USING DUMMY VARIABLES TO MODEL QUALITATIVE IINDEPENDENT VARIABLES

Example 1: A realtor wants to predict the Selling Price of a home as a function of the Area (sqr ft), Age (years), and whether the house has a pool or not. She has collected the data given in the file HomePrices.xlsx (data is shown on the next slide). Fit an MLR equation for the Selling Price as a function of the predictors.

AREA	AGE	POOL	SELLING-PRICE
2600	13	YES	233300
3400	11	YES	303350
3400	8	NO	299050
2400	1	NO	214350
2200	12	YES	202200
2400	15	NO	228000
2700	7	NO	238700
3400	8	YES	301300
2500	8	NO	233550
2500	13	NO	227050
3100	7	NO	264700
2100	6	NO	186850
3200	2	YES	287700
3200	14	NO	263900
3300	15	YES	294500
2300	3	YES	213550
2700	10	YES	245250
2900	2	NO	242700
3100	4	YES	279150
2800	15	YES	241250
3200	6	YES	305850
3400	6	YES	307850

Create a dummy variable for the variable POOL with

1 if house has a pool, and
0 if house has no pool.

AREA	AGE	POOL	SELLING-PRICE
2600	13	1	233300
3400	11	1	303350
3400	8	0	299050
2400	1	0	214350
2200	12	1	202200
2400	15	0	228000
2700	7	0	238700
3400	8	1	301300
2500	8	0	233550
2500	13	0	227050
3100	7	0	264700
2100	6	0	186850
3200	2	1	287700
3200	14	0	263900
3300	15	1	294500
2300	3	1	213550
2700	10	1	245250
2900	2	0	242700
3100	4	1	279150
2800	15	1	241250
3200	6	1	305850
3400	6	1	307850

Regression Analysis: SELLING-PRICE versus AREA, AGE, POOL

The regression equation is
SELLING-PRICE = 22726 + 80.4 AREA - 178 AGE + 8300 POOL

Predictor	Coef	SECoef	T	P	VIF	
Constant	22726	13368	1.7	0.106		AGE not significant , so we will drop AGE.
AREA	80.4	4.6	17.57	0	1.1	
AGE	-177.8	416.6	-0.43	0.675	1	
POOL	8300	3896	2.13	0.047	1.1	

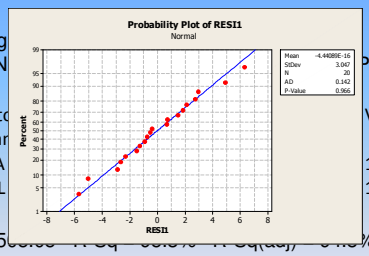
S = 8694.25 R-Sq = 95.3% R-Sq(adj) = 94.6%

Regression Analysis: SELLING-PRICE versus AREA, POOL

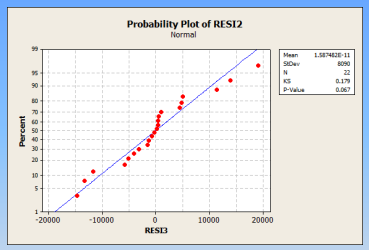
The reg
SELLIN POOL

Predict VIF
Constar
AREA 1.1
POOL 1.1

S = 85



Model is good – check residuals for normality.



Residuals from the final model are normally distributed.

The final regression equation is
SELLING-PRICE = 20979 + 80.5 AREA + 8201 POOL

which means that house selling price/sq ft is $80.5 and a pool adds $8201 to the selling price of a house.

INTERPRETATION OF MODEL WITH DUMMY VARIABLES AND INTERACTION TERM WITH X

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2$$

where

$$x_1 = x,\ x_2 = DS,\ x_1 x_2 = x \times DS$$

If $\beta_{12} \neq 0$, then when $DS = 0$ (Mutual),

$y = \beta_0 + \beta_1 x$ (line for Mutual data)

and when $DS = 1$ (Stock),

$$y = \beta_0 + \beta_1 x + \beta_{12} x = \beta_0 + (\beta_1 + \beta_{12}) x$$

In other words, the slope depends on the level of the categorical variable if there is interaction between x and the dummy variable.

What does the above ANOVA table tell us?

Example 2: (Interaction with a Dummy Variable)

A convenience store purchases novelty items each month from two vendors, A and B. Each vendor gives a discount (%) that is related to the total amount of purchases in that past. Data on monthly purchase (PURCHASE) and discount (%) for 2 years from the two vendors is given in the file Vendor.xlsx (data is shown on the next slide). The store owner wants to fit a regression model to purchases as a function of DISCOUNT and VENDOR.

PURCHASE	DISCOUNT	VENDOR	PURCHASE	DISCOUNT	VENDOR
500.55	4.14	A	1044.09	7.54	B
350.99	4.2	A	703.17	5.21	B
549.53	4.91	A	639.85	5.15	B
325.27	4.16	A	954.42	6.9	B
816.26	6.96	A	1102.34	7.93	B
604.7	5.32	A	850.74	6.73	B
528.26	5.29	A	722.2	5.89	B
461.12	4.05	A	847.64	7.13	B
712.16	5.28	A	779.18	6.1	B
367.09	4.32	A	876.5	7.93	B
			893.25	6.86	B
			992.35	7.73	B
			816.14	6.17	B
			997.14	7.54	B

We first create DUMMY variable column DVENDOR (0 for A, 1 for B). We also create an interaction column
DiscxVendor = DISCOUNT * DVENDORT
which allows for a regression equation with different intercepts and slopes for the vendors A and B (see next slide).

PURCHASE	DISCOUNT	VENDOR	DVENDOR	DiscxDVendor
500.55	4.14	A	0	0
350.99	4.2	A	0	0
549.53	4.91	A	0	0
325.27	4.16	A	0	0
816.26	6.96	A	0	0
604.7	5.32	A	0	0
528.26	5.29	A	0	0
461.12	4.05	A	0	0
712.16	5.28	A	0	0
367.09	4.32	A	0	0
1044.09	7.54	B	1	7.54
703.17	5.21	B	1	5.21
639.85	5.15	B	1	5.15
954.42	6.9	B	1	6.9
1102.34	7.93	B	1	7.93
850.74	6.73	B	1	6.73
722.2	5.89	B	1	5.89
847.64	7.13	B	1	7.13
779.18	6.1	B	1	6.1
876.5	7.93	B	1	7.93
893.25	6.86	B	1	6.86
992.35	7.73	B	1	7.73
816.14	6.17	B	1	6.17
997.14	7.54	B	1	7.54

Regression Analysis: PURCHASE versus DISCOUNT, DVENDOR, DiscxDVendor

The regression equation is

PURCHASE = - 229 + 154 DISCOUNT + 233 DVENDOR - 26.0 DiscxDVendor

VIFs are very high, and both DVENDOR and DiscxDVendor are insignificant; drop DVENDOR, DiscxDVendor one at a time

Predictor	Coef	SECoef	T	P	VIF
Constant	-229.2	126.8	-1.81	0.086	
DISCOUNT	154.39	25.68	6.01	0	5.5
DVENDOR	232.7	188.9	1.23	0.232	42.8
DiscxDVendor	-26.03	32.85	-0.79	0.437	61.9

S = 69.7440 R-Sq = 91.8% R-Sq(adj) = 90.5%

Predictor	Coef	SECoef	T	P	VIF
Constant	-151.84	80.21	-1.89	0.072	
DISCOUNT	138.48	15.87	8.73	0	2.1
DVENDOR	86.81	41.68	2.08	0.05	2.1

S = 69.1233 R-Sq = 91.5% R-Sq(adj) = 90.7%

Predictor	Coef	SECoef	T	P	VIF
Constant	-124.28	95.12	-1.31	0.205	
DISCOUNT	133.47	19.5	6.84	0	3.1
DiscxDVendor	13.429	7.403	1.81	0.084	3.1

S = 70.5993 R-Sq = 91.2% R-Sq(adj) = 90.3%

Clearly the 1st model PURCHASE = -151.84 + 138.48 DISCOUNT + 86.81 DVENDOR (R^2 = 91.5) is better than the 2nd model on this slide.

PARTIAL F-TEST FOR TESTING PORTION OF A REGRESSION MODEL

Consider the MLR model

$$y = \beta_0 + \beta_1 x_1 + ... + \beta_g x_g + \beta_{g+1} x_{g+1} + ... + \beta_k x_k + e$$

To test the hypothesis that none of the variables $x_{g+1},...,x_k$ affect y, we can test

$H_0 : \beta_{g+1} = ... = \beta_k x_k = 0$ vs.

H_1 : at least one of the above betas $\neq 0$

which is equivalent to comparing the FULL MODEL

$y = \beta_0 + \beta_1 x_1 + ... + \beta_g x_g + \beta_{g+1} x_{g+1} + ... + \beta_k x_k + e$ to

$y = \beta_0 + \beta_{g+1} x_{g+1} + ... + \beta_k x_k + e$ (REDUCED MODEL)

To compare the FULL model to the REDUCED model, compute

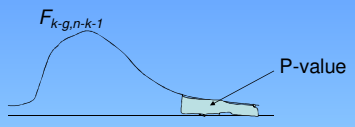
$SSE_C = SSE$ for the Complete model

$SSE_R = SSE$ for the Reduced model

$SSE_R - SSE_C$ = drop in SSE attributable to $x_{g+1},..., x_k$

Then compute the PARTIAL F-STATISTIC

$$F = \frac{(SSE_R - SSE_C)/(k-g)}{SSE_C/(n-k-1)} \sim F_{k-g,n-k-1}$$



The null hypothesis is rejected if P < α.

EXAMPLE 3: Y=demand, X4 = PriceDif, X3 – AdvExp (data file Demand.xlsx)
To fit model relating Y to X4, X3, $X3^2$, X4X3, DB, DC to the following 30 data points.

Period	Price	IndPrice	PriceDif	AdvExp	Demand	Ad_Campaign
1	3.85	3.80	-0.05	5.50	7.38	B
2	3.75	4.00	0.25	6.75	8.51	B
3	3.70	4.30	0.60	7.25	9.52	B
4	3.70	3.70	0.00	5.50	7.50	A
5	3.60	3.85	0.25	7.00	9.33	C
6	3.60	3.80	0.20	6.50	8.28	A
7	3.60	3.75	0.15	6.75	8.75	C
8	3.80	3.85	0.05	5.25	7.87	C
9	3.80	3.65	-0.15	5.25	7.10	B
10	3.85	4.00	0.15	6.00	8.00	C
11	3.90	4.10	0.20	6.50	7.89	A
12	3.90	4.00	0.10	6.25	8.15	C
13	3.70	4.10	0.40	7.00	9.10	C
14	3.75	4.20	0.45	6.90	8.86	A
15	3.75	4.10	0.35	6.80	8.90	B

Example 3, continued

Period	Price	IndPrice	PriceDif	AdvExp	Demand	Ad_Campaign
16	3.80	4.10	0.30	6.80	8.87	B
17	3.70	4.20	0.50	7.10	9.26	B
18	3.80	4.30	0.50	7.00	9.00	A
19	3.70	4.10	0.40	6.80	8.75	B
20	3.80	3.75	-0.05	6.50	7.95	B
21	3.80	3.75	-0.05	6.25	7.65	C
22	3.75	3.65	-0.10	6.00	7.27	A
23	3.70	3.90	0.20	6.50	8.00	A
24	3.55	3.65	0.10	7.00	8.50	A
25	3.60	4.10	0.50	6.80	8.75	A
26	3.65	4.25	0.60	6.80	9.21	B
27	3.70	3.65	-0.05	6.50	8.27	C
28	3.75	3.75	0.00	5.75	7.67	B
29	3.80	3.85	0.05	5.80	7.93	C
30	3.70	4.25	0.55	6.80	9.26	C

Model to be fitted to this data is:

$$y = \beta_0 + \beta_4 x_4 + \beta_3 x_3 + \beta_{33} x_3^2 + \beta_{43} x_4 x_3 + \beta_{DB} DB + \beta_{DC} DC$$

To understand the meanings of the parameters, set $x_4 = d, x_3 = a$:

Since we have 3 periods, we need 2 dummy variables, DB, DC

Period A: DB = 0, DC = 0

Period B: DB = 1, DC = 0

Period C: DB = 0, DC = 1

$$\mu_{[d,a,A]} = \beta_0 + \beta_4 d + \beta_3 a + \beta_{33} a^2 + \beta_{43} da$$

$$\mu_{[d,a,B]} = \beta_0 + \beta_4 d + \beta_3 a + \beta_{33} a^2 + \beta_{43} da + \beta_{DB}$$

$$\mu_{[d,a,C]} = \beta_0 + \beta_4 d + \beta_3 a + \beta_{33} a^2 + \beta_{43} da + \beta_{DC}$$

$$\mu_{[d,a,B]} - \mu_{[d,a,A]} = \beta_{DB}$$

$$\mu_{[d,a,C]} - \mu_{[d,a,A]} = \beta_{DC}$$

$$\mu_{[d,a,C]} - \mu_{[d,a,B]} = \beta_{DC} - \beta_{DB}$$

$H_0 : \mu_A = \mu_B = \mu_C$, $H_1 : H_0$ is false

so above H_0 is equivalent to testing

$\beta_{DB} = 0, \beta_{DC} = 0$

Regression Analysis: y=Demand versus x4=PriceDif, x3=AdvExp,

The regression equation is
y=Demand = 25.6 + 9.06 x4=PriceDif - 6.54 x3=AdvExp + 0.584 x3_sqr - 1.16 x3x4
+ 0.214 DB + 0.382 DC

Predictor	Coef	SE Coef	T	P
Constant	25.613	4.794	5.34	0.000
x4=PriceDif	9.059	3.032	2.99	0.007
x3=AdvExp	-6.538	1.581	-4.13	0.000
x3_sqr	0.5844	0.1299	4.50	0.000
x3x4	-1.1565	0.4557	-2.54	0.018
DB	0.21369	0.06215	3.44	0.002
DC	0.38178	0.06125	6.23	0.000

S = 0.130811 R-Sq = 97.1% R-Sq(adj) = 96.3%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	6	13.0650	2.1775	127.25	0.000
Error	23	0.3936	0.0171		
Total	29	13.4586			

Source	DF	Seq SS
x4=PriceDif	1	10.6527
x3=AdvExp	1	1.2722
x3_sqr	1	0.2604
x3x4	1	0.2089
DB	1	0.0061
DC	1	0.6648

Continue with Ex. 3– to test the null hypothesis that there is no effect of the different ad campaigns on y.

$H_0 : \mu_B = \mu_M = \mu_T$, $H_1 : H_0$ is false

Also,

$\beta_{DM} = \mu_M - \mu_B$, and $\beta_{DT} = \mu_T - \mu_B$

so above H_0 is equivalent to testing $\beta_{DM} = 0, \beta_{DT} = 0$

Regression Analysis: Reduced Model (NO DB or DC)

The regression equation is
y=Demand = 29.1 + 11.1 x4=PriceDif - 7.61 x3=AdvExp + 0.671 x3_sqr - 1.48 x3x4

Predictor	Coef	SE Coef	T	P
Constant	29.113	7.483	3.89	0.001
x4=PriceDif	11.134	4.446	2.50	0.019
x3=AdvExp	-7.608	2.469	-3.08	0.005
x3_sqr	0.6712	0.2027	3.31	0.003
x3x4	-1.4777	0.6672	-2.21	0.036

S = 0.206339 R-Sq = 92.1% R-Sq(adj) = 90.8%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	4	12.3942	3.0985	72.78	0.000
Residual Error	25	1.0644	0.0426		
Total	29	13.4586			

Source	DF	Seq SS
x4=PriceDif	1	10.6527
x3=AdvExp	1	1.2722
x3_sqr	1	0.2604
x3x4	1	0.2089

Full Model:

$y = \beta_0 + \beta_4 x_4 + \beta_3 x_3 + \beta_{33} x_3^2 + \beta_{43} x_4 x_3 + \beta_{DB} DB + \beta_{DC} DC$

Reduced Model:

$y = \beta_0 + \beta_4 x_4 + \beta_3 x_3 + \beta_{33} x_3^2 + \beta_{43} x_4 x_3$

$SSE_C = 0.3936$, df = 23
$SSE_R = 1.0644$, $SSE_R - SSE_C = 0.1677$
$F_{calc} = 23 \times .1677/.3936 = 9.8$
$F_{4,23,.05} = 2.80$

Reject null hypothesis that $H_0 : \beta_{DB} = 0, \beta_{DC} = 0$

and conclude that DB and DC terms do have significant effects on y.