Regression Multiple

Introduction

- simple linear regression: On a étudié la relation entre une variable y à expliquer et une variable explicative,
- A présent, on essaye d'expliquer y par plusieurs variables exogènes explicatives indépendantes.

Introduction

- On se propose de voir par la suite la contribution si elle existe de chaque variables.

Multiple Regression

- **Example.** L’étude d’un coût de produit, Y, pour 67 individus, 4 independent variables ont été considérées:
    - X_1: Taille moyenne de l'encours de prêt durant l'année,
    - X_2 : Nombre moyen de prêts en cours,
    - X_3 : Nombre total de nouvelles demandes de prêt traitées, and
    - X_4 : Indice d'échelle de salaire.
- Le model sera alors:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 X_3 + \beta_4 x_4 + \varepsilon$$

Formal Statement of the Model

- Modèle de régression General

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$$

  - β_0, β_1, …, β_k : paramètres
  - X_1, X_2, …,X_k variables connues observées
  - ε , the error terms are independent N(o, σ^2)

Estimating the parameters of the model

- La valeur des paramètres est inconnue, on les estime β_i.
- Comme dans la regression simple, on utilise MCO pour calculer fitted values (y ajustée ou estimée)

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_k x_k$$

- Principe de la MCO: min la somme des carrés des résidus

Estimating the parameters of the model

- Min:

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Estimation des parametères du modèle

- L'estimation des β_i indique la variation de y en cas de variation d'une unite de X_i le reste des variables est supposé constant
- Les paramètres β_i souvent appelés partial regression coefficients

Estimating the parameters of the model

- La variance concernant fitted model est

$$s^2 = \frac{1}{n-k-1}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

K: nombre des variables Xi

$$s = \sqrt{s^2}$$

Écart type:the regression standard error

Estimating the parameters of the model

- Dans le modèle σ^2 and σ mesure la variabilité des résponses.
- Il est naturel d'estimater σ^2 par s^2 and σ par s.

Analysis of Variance Table

- The basic idea of the regression ANOVA table are the same in simple and multiple regression.
- The sum of squares decomposition and the associated degrees of freedom are:

$$\sum (y_i - \bar{y})^2 = \sum (\hat{y}_i - \bar{y})^2 + \sum (y_i - \hat{y}_i)^2$$

$$SST \quad = \quad SSR \quad + \quad SSE$$

- *df*:

$$n - 1 = k + (n - k - 1)$$

Table de l’Analyse de la Variance

Source	Sum of Squares	df	Mean Square	F-test
Regression	SSR (VE)	k	MSR= SSR/k	MSR/MSE
Error	SSE(VR ou SCR)	n-k-1	MSE= SSE/n-k-1	
Total	SST	n-1		

F-test (significativité globale) for the overall fit of the model

- To test the statistical significance of the regression relation between the response variable y and the set of variables $x_1, \ldots, x_k$, (significativité globale):

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$$
$$H_a : \text{not all } \beta_i \, (i = 1, \cdots k) \text{ equal zero}$$

- On utilise le test de Fisher, on calculi F obs:

F (observée) = $$F = \frac{MSR}{MSE}$$

F-test for the overall fit of the model

- The decision rule at significance level α is:
  - Reject H_0 if

    $$F > F(\alpha; k, n-k-1)$$

  - La valeur critique F(α, k, n-k-1) est déterminée par la table de Fisher
- L' existence d'une relation globale ne signifie pas qu'on pourrait passer à la prédiction.
- Noter lorsque k=1, thisce test est réduit à the F-test dans une simple linear regression

Interval estimation of β_i

- For our regression model, we have:

$$\frac{b_i - \beta_i}{s(b_i)} \quad \text{has a t-distribution with n-k-1 degrees of freedom}$$

- Therefore, an interval estimate for β_i with 1- α confidence coefficient is:

$$b_i \pm t(\frac{\alpha}{2}; n-k-1)s(b_i)$$

Where

$$s(b_i) = \sqrt{\frac{MSE}{\sum (x-\bar{x})^2}}$$

Significance tests for β_i

- To test:

$$H_0 : \beta_i = 0$$

$$H_a : \beta_i \neq 0$$

- On utilise le test de student :

$$t = \frac{b_i}{s(b_i)}$$

- Reject H_0 if

$$t > t(\frac{\alpha}{2}; n-k-1) \quad or$$

$$t < -t(\frac{\alpha}{2}; n-k-1)$$

Multiple regression model

- Souvent on a plusieurs variables expliquatives, notre objectif est de les utilizer pour la prediction de y.

Multiple regression model Building

- Parfois l'effet d'une variable depend également de l'effet d'une autre non prévu dans le modèle.
- C'est l'effet de l'interaction.

Multiple regression model Building

- La manière la plus simple de les intégrer est de construire une variable égale au produit des deux,.
- Comment détecter un bon modèle?

Meilleure Regression equation.

- Certaines variables peuvent être écartées:
  - Ne sont pas fondamentales pour le problème
  - Peuvent entrainer une augmentation d el’erreur
  - Peuvent générer une autre variable indépendante au modèle.

Exemple

Données des series temporelles: Correlation

- Dans le modèle de régression on a suppose que les erreurs ε_i sont independants.
- En business, beaucoup d' applications utilisent les " time series data".
- L'hypothèse de l'absence d'autocorrelation des erruers peut être contestable.

Problems of Serial Correlation

- Si les erreurs sont autocorrélées, l'usage de la MCO peut avoir des conséquences
  - MCO va biaiser la variance des "error terms"
  - tests t and F distribution ne sont pas adéquats
  - La quantité ajustée ou fitted values pourrait être biaisée elle aussi d'où une très mauvaise prediction,

Corrélation de premier ordre

- Utilisons une regression simple pour bien illustrer le problème :

$$\varepsilon_t = \rho\,\varepsilon_{t-1} + \nu_t$$

$$y_t = \beta_0 + \beta_1 x_t + \varepsilon_t$$

  - ε_t = error at time t
  - ρ = the parameter that measures correlation between adjacent error terms
  - ν_t normally distributed error terms with mean zero and variance σ^2_ν (hypothèse habituelle)

Durbin-Watson Test for Serial Correlation

- Recall the first-order serial correlation model

$$y_t = \beta_0 + \beta_1 x_t + \varepsilon_t$$

$$\varepsilon_t = \rho \varepsilon_{t-1} + \nu_t$$

- The hypothesis to be tested are:

$$H_0 : \rho = 0$$

$$H_a : \rho > 0$$

- The alternative hypothesis is $\rho > 0$ since in business and economic time series tend to show positive correlation.

Durbin-Watson Test for Serial Correlation

- The Durbin-Watson statistic is defined as

$$DW = \frac{\sum_{t=2}^{n}(e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2}$$

- Where

$$e_t = y_t - \hat{y}_t = \text{the residual for time period t}$$

$$e_{t-1} = y_{t-1} - \hat{y}_{t-1} = \text{the residual for time period t -1}$$

Durbin-Watson Test for Serial Correlation

- The auto correlation coefficient ρ can be estimated by the lag 1 residual autocorrelation $r_1(e)$

$$r_1(e) = \frac{\sum_{t=2}^{n} e_t e_{t-1}}{\sum_{t=1}^{n} e_t^2}$$

- And it can be shown that

$$DW = 2(1 - r_1(e))$$

Durbin-Watson Test for Serial Correlation

- Since $-1 < r_1(e) < 1$ then $0 < DW < 4$
- If $r_1(e) = 0$, then $DW = 2$ (there is no correlation.)
- If $r_1(e) > 0$, then $DW < 2$ (positive correlation)
- If $r_1(e) < 0$, Then $DW > 2$ (negative correlation)

Durbin-Watson Test for Serial Correlation

- Decision rule:
  - If DW > U, Do not reject H_0.
  - If DW < L, Reject H_0
  - If L ≤ DW ≤ U, the test is inconclusive.
- The critical Upper (U) an Lower (L) bound can be found in Durbin-Watson table of your text book.
- To use this table you need to know The significance level (α) The number of independent parameters in the model (k), and the sample size (n).

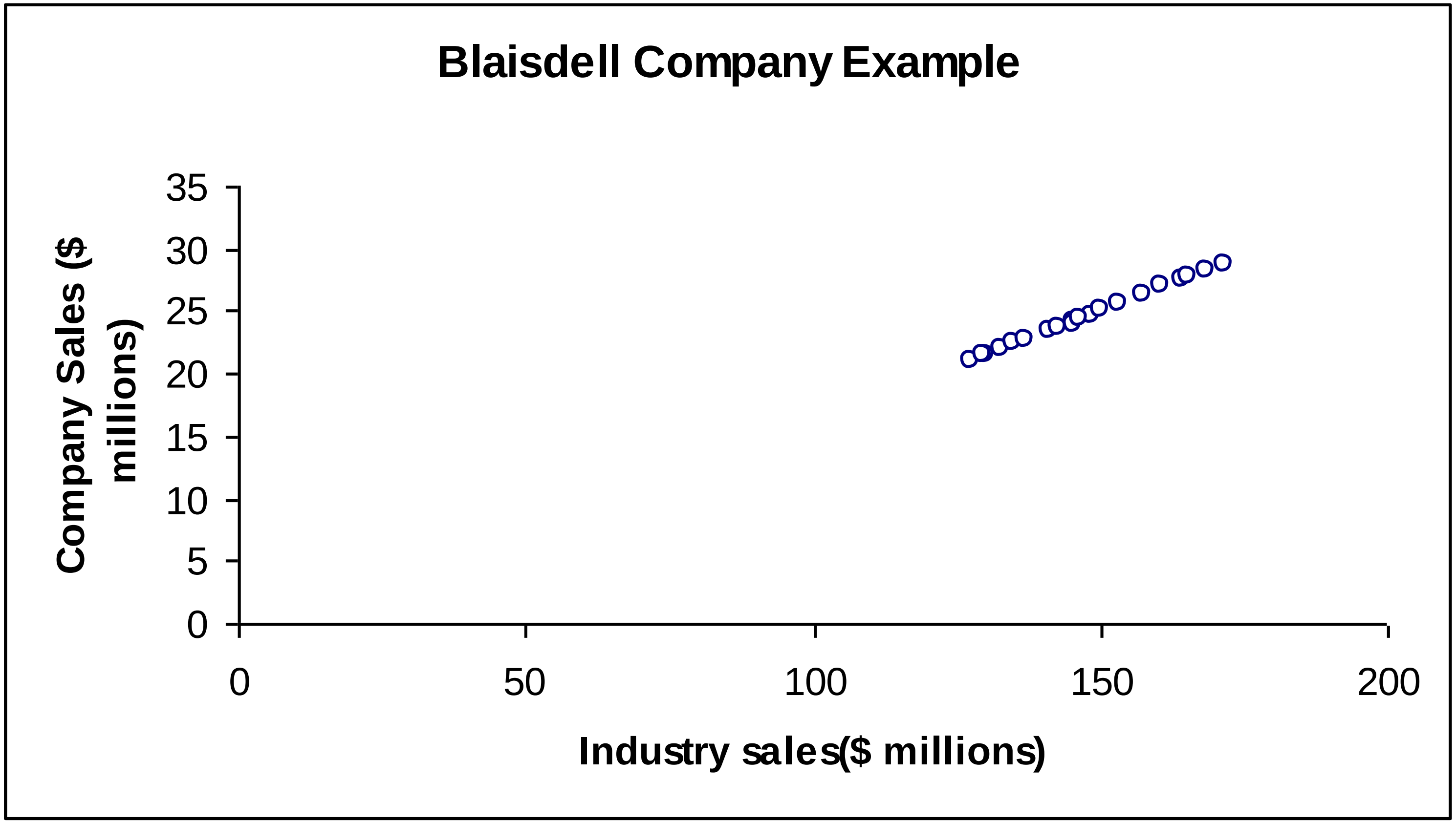
Example

- The Blaisdell Company wished to predict its sales by using industry sales as a predictor variable. The following table gives seasonally adjusted quarterly data on company sales and industry sales for the period 1983-1987.

Example

Year	Quarter	t	CompSale	InduSale
1983	1	1	20.96	127.3
	2	2	21.4	130
	3	3	21.96	132.7
	4	4	21.52	129.4
1984	1	5	22.39	135
	2	6	22.76	137.1
	3	7	23.48	141.2
	4	8	23.66	142.8
1985	1	9	24.1	145.5
	2	10	24.01	145.3
	3	11	24.54	148.3
	4	12	24.3	146.4
1986	1	13	25	150.2
	2	14	25.64	153.1
	3	15	26.36	157.3
	4	16	26.98	160.7
1987	1	17	27.52	164.2
	2	18	27.78	165.6
	3	19	28.24	168.7
	4	20	28.78	171.7

Example

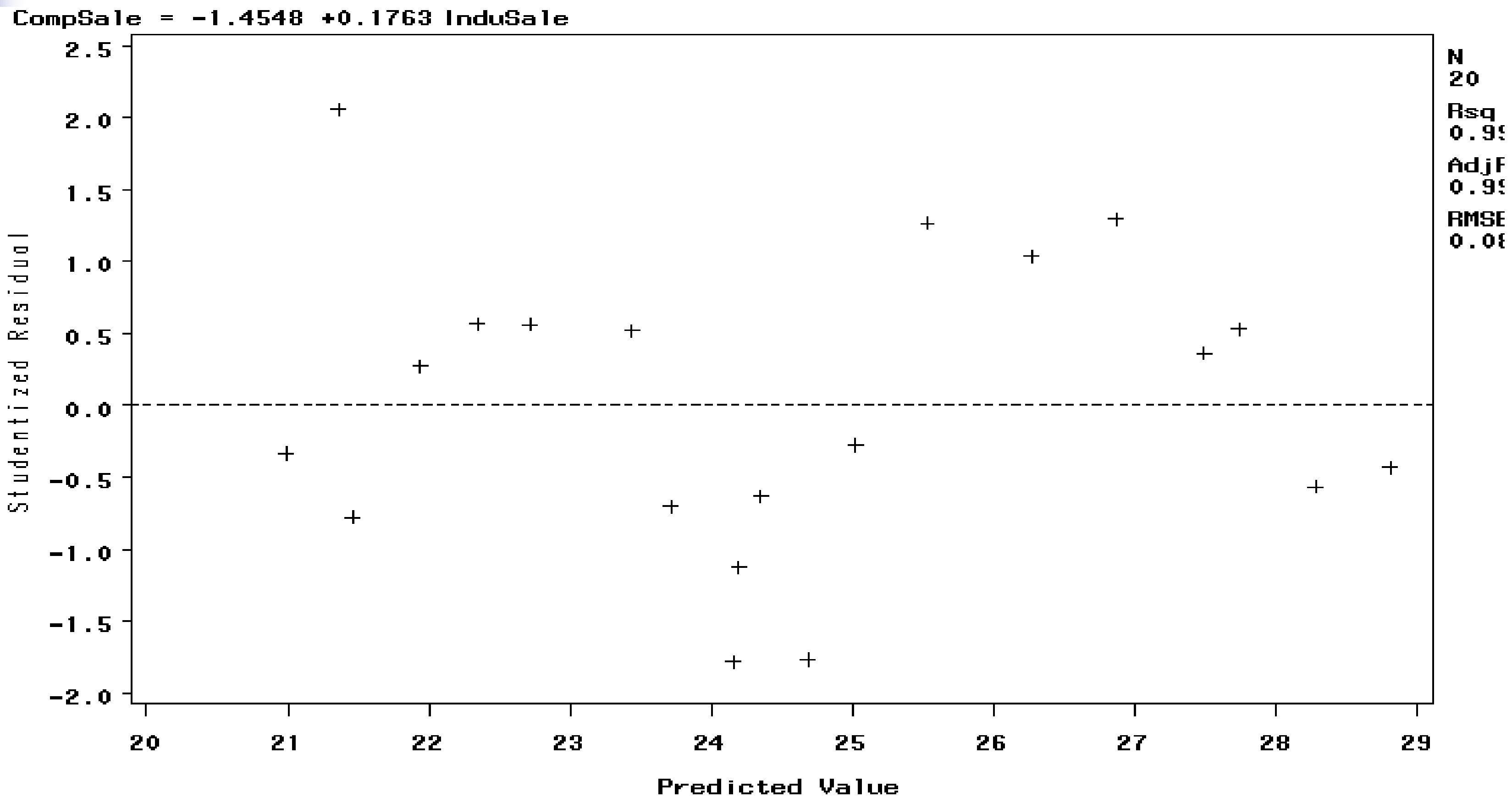


Blaisdell Company Example
Company Sales ($ millions)
35
30
25
20
15
10
5
0
0
50
100
150
200
Industry sales($ millions)

Example

- The scatter plot suggests that a linear regression model is appropriate.
- Least squares method was used to fit a regression line to the data.
- The residuals were plotted against the fitted values.
- The plot shows that the residuals are consistently above or below the fitted value for extended periods.

Example



CompSale = -1.4548 +0.1763 InduSale
2.5
2.0
1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5
-2.0
Studentized Residual
20
21
22
23
24
25
26
27
28
29
Predicted Value
N
20
Rsq
AdjR
RMSE

Example

- To confirm this graphic diagnosis we will use the Durbin-Watson test for:

$$H_0 : \rho = 0$$
$$H_a : \rho > 0$$

- The test statistic is:

$$DW = \frac{\sum_{t=2}^{n} (e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2}$$

Example

Year	Quarter	t	Company sales(y)	Industry sales(x)	e_t	e_t-e_{t-1}	(e_t-e_{t-1})^2	e_t^2
1983	1	1	20.96	127.3	-0.02605			0.000679
	2	2	21.4	130	-0.06202	-0.03596	0.001293	0.003846
	3	3	21.96	132.7	0.022021	0.084036	0.007062	0.000485
	4	4	21.52	129.4	0.163754	0.141733	0.020088	0.026815
1984	1	5	22.39	135	0.04657	-0.11718	0.013732	0.002169
	2	6	22.76	137.1	0.046377	-0.00019	3.76E-08	0.002151
	3	7	23.48	141.2	0.043617	-0.00276	7.61E-06	0.001902
	4	8	23.66	142.8	-0.05844	-0.10205	0.010415	0.003415
1985	1	9	24.1	145.5	-0.0944	-0.03596	0.001293	0.008911
	2	10	24.01	145.3	-0.14914	-0.05474	0.002997	0.022243
	3	11	24.54	148.3	-0.14799	0.001152	1.33E-06	0.021901
	4	12	24.3	146.4	-0.05305	0.094937	0.009013	0.002815
1986	1	13	25	150.2	-0.02293	0.030125	0.000908	0.000526
	2	14	25.64	153.1	0.105852	0.12878	0.016584	0.011205
	3	15	26.36	157.3	0.085464	-0.02039	0.000416	0.007304
	4	16	26.98	160.7	0.106102	0.020638	0.000426	0.011258
1987	1	17	27.52	164.2	0.029112	-0.07699	0.005927	0.000848
	2	18	27.78	165.6	0.042316	0.013204	0.000174	0.001791
	3	19	28.24	168.7	-0.04416	-0.08648	0.007478	0.00195
	4	20	28.78	171.7	-0.03301	0.011152	0.000124	0.00109
							0.097941	0.133302

Example

$$DW = \frac{.09794}{.13330} = .735$$

- Using Durbin Watson table of your text book, for k = 1, and n=20, and using α = .01 we find U = 1.15, and L = .95
- Since DW = .735 falls below L = .95 , we reject the null hypothesis, namely, that the error terms are positively autocorrelated.

Remedial Measures for Serial Correlation

- Addition of one or more independent variables to the regression model.
  - One major cause of autocorrelated error terms is the omission from the model of one or more key variables that have time-ordered effects on the dependent variable.
- Use transformed variables.
  - The regression model is specified in terms of changes rather than levels.

Extensions of the Multiple Regression Model

- In some situations, nonlinear terms may be needed as independent variables in a regression analysis.
  - Business or economic logic may suggest that non-linearity is expected.
  - A graphic display of the data may be helpful in determining whether non-linearity is present.
- One common economic cause for non-linearity is diminishing returns.
  - Fore example, the effect of advertising on sales may diminish as increased advertising is used.

Extensions of the Multiple Regression Model

- Some common forms of nonlinear functions are :

$$Y = \beta_0 + \beta_1(X) + \beta_2(X^2)$$

$$Y = \beta_0 + \beta_1(X) + \beta_2(X^2) + \beta_3(X^3)$$

$$Y = \beta_0 + \beta_1(1/X)$$

$$Y = e^{\beta_0} X^{\beta_1}$$

Extensions of the Multiple Regression Model

- To illustrate the use and interpretation of a non-linear term, we return to the problem of developing a forecasting model for private housing starts (PHS).
- So far we have looked at the following model

$$PHS = \beta_0 + \beta_1(MR) + \beta_2(Q2) + \beta_3(Q3) + \beta_4(Q4)$$

  - Where MR is the mortgage rate and Q2, Q3, and Q4 are indicators variables for quarters 2, 3, and 4.

Example: Private Housing Start

- First we add real disposable personal income per capita (DPI) as an independent variable. Our new model for this data set is:

$$PHS = \beta_0 + \beta_1(MR) + \beta_2(Q2) + \beta_3(Q3) + \beta_4(Q4) + \beta_5(DPI)$$

- Regression results for this model are shown in the next slide.

Example: Private Housing Start

SUMMARY OUTPUT						
Regression Statistics						
Multiple R	0.943791346					
R Square	0.890742104					
Adjusted R Square	0.874187878					
Standard Error	19.05542121					
Observations	39					
ANOVA						
	df	*SS*	*MS*	*F*	*Significance F*	
Regression	5	97690.01942	19538	53.80753	6.51194E-15	
Residual	33	11982.59955	363.1091			
Total	38	109672.619				
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	-31.06403714	105.1938477	-0.2953	0.769613	-245.0826992	182.9546249
MR	-20.1992545	4.124906847	-4.8969	2.5E-05	-28.59144723	-11.80706176
Q2	97.03478074	8.900711541	10.90191	1.78E-12	78.9261326	115.1434289
Q3	75.40017073	8.827185877	8.541813	7.17E-10	57.44111179	93.35922967
Q4	20.35306822	8.83373887	2.304015	0.027657	2.380677107	38.32545934
DPI	0.022407799	0.004356973	5.142974	1.21E-05	0.013543464	0.031272134

Example: Private Housing Start

- The prediction model is

$$P\hat{HS} = -31.06 - 20.19(MR) + 97.03(Q2) + 75.40(Q3) + 20.35(Q4) + 0.02(DPI)$$

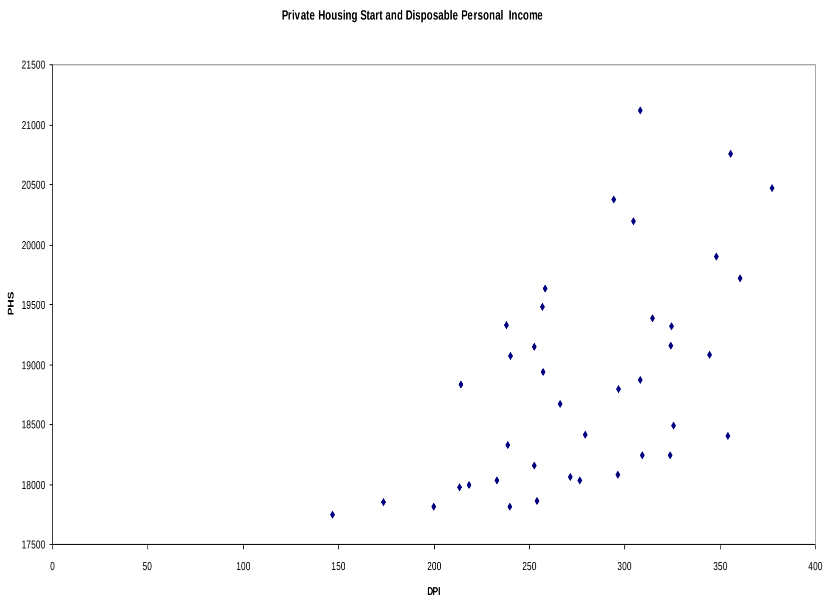
- In comparison with the previous model, we see that the R-squared has improved.It has changed from 78% to 89%.
- The standard error of the estimate has decreased from 26.49 for the previous model to 19.05 for the new model.

Example: Private Housing Start

- The value of the DW test has changed from 0.88 for the previous model to 0.78 for the new model.
- At 5% level the critical value for DW test, from Durbin-Watson table, for k = 5, and n = 39 is L= 1.22, and U = 1.79.
- Since The value of the DW test is smaller than L=1.22, we reject the null hypothesis H_0: $\rho = 0$
- This implies that there is serial correlation in both models, the assumption of the independence of the error terms is not valid.

Example: Private Housing Start

- The Plot of PHS against DPI shows a curve linear relation.
- Next we introduce a nonlinear term into the regression.
- The square of disposable personal income per capita (DPI^2) is included in the regression model.



Example: Private Housing Start

- We also add the dependent variable, lagged one quarter, as an independent variable in order to help reduce serial correlation.
- The third model that we fit to our data set is:

$$PHS = \beta_0 + \beta_1(MR) + \beta_2(Q2) + \beta_3(Q3) + \beta_4(Q4) + \beta_5(DPI) + \beta_6(DPI^2) + \beta_7(LPHS)$$

- Regression results for this model are shown in the next slide.

Example: Private Housing Start

SUMMARY OUTPUT						
Regression Statistics						
Multiple R	0.97778626					
R Square	0.956065971					
Adjusted R Square	0.946145384					
Standard Error	12.46719572					
Observations	39					
ANOVA						
	df	*SS*	*MS*	*F*	*Significance F*	
Regression	7	104854.2589	14979.17985	96.37191	3.07085E-19	
Residual	31	4818.360042	155.4309691			
Total	38	109672.619				
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	716.5926532	1017.664989	0.704153784	0.486593	-1358.949934	2792.13524
MR	-13.65521724	3.093504134	-4.414158396	0.000114	-19.96446404	-7.345970448
Q2	106.9813297	6.069780998	17.62523718	1.04E-17	94.60192287	119.3607366
Q3	27.72122303	9.111432565	3.042465916	0.004748	9.138323433	46.30412262
Q4	-13.37855186	7.653050858	-1.748133144	0.09034	-28.98706069	2.22995698
DPI	-0.060399279	0.104412354	-0.578468704	0.567127	-0.273349798	0.15255124
DPI SQUARED	0.000335974	0.000536397	0.626354647	0.535668	-0.000758014	0.001429963
LPHS	0.655786939	0.097265424	6.742241114	1.51E-07	0.457412689	0.854161189

Example: Private Housing Start

- The inclusion of DPI^2 and Lagged PHS has increased the R-squared to 96%
- The standard error of the estimate has decreased to 12.45
- The value of the DW test has increased to 2.32 which is greater than U = 1.79 which rule out positive serial correlation.
- You see that the third model worked best for this data set.
- The following slide gives the data set.

Example: Private Housing Start

PERIOD	PHS	MR	LPHS	Q2	Q3	Q4	DPI	DPI SQUARED
30-Jun-90	271.3	10.3372	217	1	0	0	18063	1,631,359.85
30-Sep-90	233	10.1033	271.3	0	1	0	18031	1,625,584.81
31-Dec-90	173.6	9.9547	233	0	0	1	17856	1,594,183.68
31-Mar-91	146.7	9.5008	173.6	0	0	0	17748	1,574,957.52
30-Jun-91	254.1	9.5265	146.7	1	0	0	17861	1,595,076.61
30-Sep-91	239.8	9.2755	254.1	0	1	0	17816	1,587,049.28
31-Dec-91	199.8	8.6882	239.8	0	0	1	17811	1,586,158.61
31-Mar-92	218.5	8.7098	199.8	0	0	0	18000	1,620,000.00
30-Jun-92	296.4	8.6782	218.5	1	0	0	18085	1,635,336.13
30-Sep-92	276.4	8.0085	296.4	0	1	0	18036	1,626,486.48
31-Dec-92	238.8	8.2052	276.4	0	0	1	18330	1,679,944.50
31-Mar-93	213.2	7.7332	238.8	0	0	0	17975	1,615,503.13
30-Jun-93	323.7	7.4515	213.2	1	0	0	18247	1,664,765.05
30-Sep-93	309.3	7.0778	323.7	0	1	0	18246	1,664,582.58
31-Dec-93	279.4	7.0537	309.3	0	0	1	18413	1,695,192.85
31-Mar-94	252.6	7.2958	279.4	0	0	0	18154	1,647,838.58
30-Jun-94	354.2	8.4370	252.6	1	0	0	18409	1,694,456.41
30-Sep-94	325.7	8.5882	354.2	0	1	0	18493	1,709,955.25
31-Dec-94	265.9	9.0977	325.7	0	0	1	18667	1,742,284.45
31-Mar-95	214.2	8.8123	265.9	0	0	0	18834	1,773,597.78
30-Jun-95	296.7	7.9470	214.2	1	0	0	18798	1,766,824.02
30-Sep-95	308.2	7.7012	296.7	0	1	0	18871	1,780,573.21
31-Dec-95	257.2	7.3508	308.2	0	0	1	18942	1,793,996.82
31-Mar-96	240	7.2430	257.2	0	0	0	19071	1,818,515.21
30-Jun-96	344.5	8.1050	240	1	0	0	19081	1,820,422.81
30-Sep-96	324	8.1590	344.5	0	1	0	19161	1,835,719.61
31-Dec-96	252.4	7.7102	324	0	0	1	19152	1,833,995.52
31-Mar-97	237.8	7.7905	252.4	0	0	0	19331	1,868,437.81
30-Jun-97	324.5	7.9255	237.8	1	0	0	19315	1,865,346.13
30-Sep-97	314.6	7.4692	324.5	0	1	0	19385	1,878,891.13
31-Dec-97	256.8	7.1980	314.6	0	0	1	19478	1,896,962.42
31-Mar-98	258.4	7.0547	256.8	0	0	0	19632	1,927,077.12
30-Jun-98	360.4	7.0938	258.4	1	0	0	19719	1,944,194.81
30-Sep-98	348	6.8657	360.4	0	1	0	19905	1,980,963.41
31-Dec-98	304.6	6.7633	348	0	0	1	20194	2,038,980.00
31-Mar-99	294.1	6.8805	304.6	0	0	0	20377	2,076,010.87
30-Jun-99	377.1	7.2037	294.1	1	0	0	20472	2,095,440.74
30-Sep-99	355.6	7.7990	377.1	0	1	0	20756	2,153,982.23
31-Dec-99	308.1	7.8338	355.6	0	0	1	21124	2,231,020.37