Chapter 4
Multiple Regression Analysis (Part 2)

Terry Dielman

Applied Regression Analysis:
A Second Course in Business and Economic Statistics, fourth edition

4.4 Comparing Two Regression Models

So far we have looked at two types of hypothesis tests. One was about the overall fit:

$$H_0: \beta_1 = \beta_2 = \ldots = \beta_K = 0$$

The other was about individual terms:

$$H_0: \beta_j = 0$$

$$H_a: \beta_j \neq 0$$

4.4.1 Full and Reduced Model Using Separate Regressions

- Suppose we wanted to test a subset of the *x* variables for significance as a group.
- We could do this by comparing two models.
- The first (Full Model) has *K* variables in it.
- The second (Reduced Model) contains only the *L* variables that are NOT in our group.

The Two Models

For convenience, let's assume the group is the last (*K*-*L*) variables. The Full Model is:

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_L x_L + \beta_{L+1} x_{L+1} + \cdots + \beta_K x_K + e$$

The Reduced Model is just:

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_L x_L + e$$

The Partial F Test

We test the group for significance with another F test. The hypothesis is:

H_0: $\beta_{L+1} = \beta_{L+2} = \ldots = \beta_K = 0$

H_a: At least one $\beta \neq 0$

The test is performed by seeing how much SSE changes between models.

The Partial F Statistic

Let SSE_F and SSE_R denote the SSE in the full and reduced models.

$$F = \frac{(SSE_R - SSE_F) / (K - L)}{SSE_F / (n\text{-}K\text{-}1)}$$

The statistic has $(K\text{-}L)$ numerator and $(n\text{-}K\text{-}1)$ denominator d.f.

The "Group"

- In many problems the group of variables has a natural definition.
- In later chapters we look at groups that provide curvature, measure location and model seasonal variation.
- Here we are just going to look at the effect of adding two new variables.

Example 4.4 Meddicorp (yet again)

In addition to the variables for advertising and bonuses paid, we now consider variables for market share and competition.

x_3 = Meddicorp market share in each area

x_4 = largest competitor's sales in each area

The New Regression Model

The regression equation is

SALES = - 594 + 2.51 ADV + 1.91 BONUS + 2.65 MKTSHR - 0.121 COMPET

Predictor	Coef	SE Coef	T	P
Constant	-593.5	259.2	-2.29	0.033
ADV	2.5131	0.3143	8.00	0.000
BONUS	1.9059	0.7424	2.57	0.018
MKTSHR	2.651	4.636	0.57	0.574
COMPET	-0.1207	0.3718	-0.32	0.749

S = 93.77 R-Sq = 85.9% R-Sq(adj) = 83.1%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	4	1073119	268280	30.51	0.000
Residual Error	20	175855	8793		
Total	24	1248974			

```
PROC  REG;
  MODEL SALES = ADV  BONUS  MKTSHR  COMPET;
RUN:
```

SAS 시스템

The REG Procedure
Model: MODEL1
Dependent Variable: SALES

Number of Observations Read	25
Number of Observations Used	25

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	1073119	268280	30.51	<.0001
Error	20	175855	8792.75990		
Corrected Total	24	1248974			

Root MSE	93.76972	R-Square	0.8592
Dependent Mean	1269.02000	Adj R-Sq	0.8310
Coeff Var	7.38914		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-593.53745	259.19585	-2.29	0.0330
ADV	1	2.51314	0.31428	8.00	<.0001
BONUS	1	1.90595	0.74239	2.57	0.0184
MKTSHR	1	2.65101	4.63566	0.57	0.5738
COMPET	1	-0.12073	0.37181	-0.32	0.7488

Did We Gain Anything?

- The old model had $R^2 = 85.5\%$ so we gained only .4%.
- The t ratios for the two new variables are .57 and -.32.
- It does not look like we have an improvement, but we really need the F test to be sure.

The Formal Test

Numerator df = *(K-L) = 4-2 = 2*
Denominator df = *(n-K-1) = 20*

At a 5% level, $F_{2,20} = 3.49$

$H_0: \beta_{MKTSHR} = \beta_{COMPET} = 0$
$H_a:$ At least one is $\neq 0$

Reject H_0 if $F > 3.49$

Things We Need

Full Model: *(K = 4)*

$SSE_F = 175855$

(n-K-1) = 20

Reduced Model: *(L = 2)*

SSE_R

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	2	1067797	533899	64.83	0.000
Residual Error	22	181176	8235		
Total	24	1248974			

Computations

$$F = \frac{(SSE_R - SSE_F) / (K - L)}{SSE_F / (n-K-1)}$$

$$= \frac{(181176 - 175855) / (4 - 2)}{175855 / (25-4-1)}$$

$$= \frac{5321/2}{8793} = .3026$$

4.4.2 Full and Reduced Model Comparisons Using Conditional Sums of Squares

- In the standard ANOVA table, SSR shows the amount of variation explained by all variables together.
- Alternate forms of the table break SSR down into components.
- For example, Minitab shows sequential SSR which shows how much SSR increases as each new term is added.

Sequential SSR for Meddicorp

```
S = 93.77       R-Sq = 85.9%     R-Sq(adj) = 83.1%

Analysis of Variance

Source            DF          SS          MS          F          P
Regression         4     1073119      268280      30.51      0.000
Residual Error    20      175855        8793
Total             24     1248974

Source       DF      Seq SS
ADV           1     1012408
BONUS         1       55389
MKTSHR        1        4394
COMPET        1         927
```

```
PROC  REG;
  MODEL SALES = ADV  BONUS  MKTSHR  COMPET / SS1 ;
RUN:
```

Number of Observations Read	25
Number of Observations Used	25

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	1073119	268280	30.51	<.0001
Error	20	175855	8792.75990		
Corrected Total	24	1248974			

Root MSE	93.76972	R-Square	0.8592
Dependent Mean	1269.02000	Adj R-Sq	0.8310
Coeff Var	7.38914		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Type I SS
Intercept	1	-593.53745	259.19585	-2.29	0.0330	40260294
ADV	1	2.51314	0.31428	8.00	<.0001	1012408
BONUS	1	1.90595	0.74239	2.57	0.0184	55389
MKTSHR	1	2.65101	4.63566	0.57	0.5738	4394.15366
COMPET	1	-0.12073	0.37181	-0.32	0.7488	927.06773

```
PROC  REG;
  MODEL SALES = ADV  BONUS  / SS1 ;
RUN:
```

Number of Observations Read	25
Number of Observations Used	25

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	1067797	533899	64.83	<.0001
Error	22	181176	8235.29179		
Corrected Total	24	1248974			

Root MSE	90.74851	R-Square	0.8549
Dependent Mean	1269.02000	Adj R-Sq	0.8418
Coeff Var	7.15107		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Type I SS
Intercept	1	-516.44428	189.87570	-2.72	0.0125	40260294
ADV	1	2.47318	0.27531	8.98	<.0001	1012408
BONUS	1	1.85618	0.71573	2.59	0.0166	55389

Meaning What?

1. If ADV was added to the model first, SSR would rise from 0 to 1012408.
2. Addition of BONUS would yield a nice increase of 55389.
3. If MKTSHR entered third, SSR would rise a paltry 4394.
4. Finally, if COMPET came in last, SSR would barely budge by 927.

Implications

- This is another way of showing that once you account for advertising and bonuses paid, you do not get much more from the last two variables.
- The last two sequential SSR values add up to 5321, which was the same as the (SSE_R – SSE_F) quantity computed in the partial F test.
- Given that, it is not surprising to learn that the partial F test can be stated in terms of sequential sums of squares.

4.5 Prediction With a Multiple Regression Equation

As in simple regression, we will look at two types of computations:

1. Estimating the mean *y* that can occur at a set of *x* values.
2. Predicting an individual value of *y* that can occur at a set of x values.

4.5.1 Estimating the Conditional Mean of *y* Given $x_1, x_2, \ldots, x_K$

This is our estimate of the point on our regression surface that occurs at a specific set of *x* values.

For two *x* variables, we are estimating:

$$\mu_{y|x_1,x_2} = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

Computations

The point estimate is straightforward, just plug in the *x* values.

$$\hat{y}_m = b_0 + b_1 x_1 + b_2 x_2$$

The difficult part is computing a standard error to use in a confidence interval. Thankfully, most computer programs can do that.

4.5.2 Predicting an Individual Value of y Given $x_1, x_2, \ldots, x_K$

Now the quantity we are trying to estimate is:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + e_i$$

Our interval will have to account for the extra term (e_i) in the equation, thus will be wider than the interval for the mean.

Prediction in Minitab

Here we predict sales for a territory with 500 units of advertising and 250 units of bonus

```
Predicted Values for New Observations

New Obs     Fit     SE Fit      95.0% CI           95.0% PI
1        1184.2       25.2   (1131.8, 1236.6)  ( 988.8, 1379.5)

Values of Predictors for New Observations

New Obs       ADV     BONUS
1             500       250
```

Interpretations

We are 95% sure that the *average* sales in territories with $50,000 advertising and $25,000 of bonuses will be between $1,131,800 and $1,236,600.

We are 95% sure that any *individual* territory with this level of advertising and bonuses will have between $988,800 and $1,379,500 of sales

4.6 Multicollinearity: A Potential Problem in Multiple Regression

- In multiple regression, we like the *x* variables to be highly correlated with *y* because this implies good prediction ability.
- If the *x* variables are highly correlated among themselves, however, much of this prediction ability is redundant.
- Sometimes this redundancy is so severe that it causes some instability in the coefficient estimation. When that happens we say *multicollinearity* has occurred.

4.6.1 Consequences of Multicollinearity

1. The standard errors of the b_j are larger than they should be. This could cause all the *t* statistics to be near 0 even though the *F* is large.

2. It is hard to get good estimates of the β_j. The b_j may have the wrong sign. They may have large changes in value if another variable is dropped from or added to the regression.

4.6.2 Detecting Multicollinearity

Several methods appear in the literature. Some of these are:

1. Examining pairwise correlations
2. Seeing large *F* but small *t* ratios
3. Computing Variance Inflation Factors

Examining Pairwise Correlations

- If it is only a *collinearity* problem, you can detect it by examining the correlations for pairs of *x* values.
- How large the correlation needs to be before it suggests a problem is debatable. One rule of thumb is .5, another is the maximum correlation between *y* and the various *x* values.
- The major limitation of this is that it will not help if there is a linear relationship involving several *x* values, for example,

  $x_1 = 2x_2 - .07x_3$ + a small random error

Large *F*, Small *t*

- With a significant *F* statistic you would expect to see at least one significant predictor, but that may not happen if all the variables are fighting each other for significance.
- This method of detection may not work if there are, say, six good predictors but the multicollinearity only involves four of them.
- This method also may not help identify what variables are involved.

Variance Inflation Factors

- This is probably the most reliable method for detection because it both shows the problem exists and what variables are involved.
- We can compute a VIF for each variable. A high VIF is an indication that the variable's standard error is "inflated" by its relationship to the other *x* variables.

Auxiliary Regressions

Suppose we regressed each *x* value, in turn, on all of the other *x* variables.

Let R_j^2 denote the model's R^2 we get when x_j was the "temporary *y*".

The variable's VIF is $VIF_j = \frac{1}{1 - R_j^2}$

VIF_j and R_j^2

If x_j was totally uncorrelated with the other x variables, its VIF would be 1.

This table shows some other values.

R_j^2	VIF_j
0%	1
50%	2
80%	5
90%	10
99%	100

Auxiliary Regressions: A Lot of Work?

- If there were a large number of *x* variables in the model, obtaining the auxiliaries would be tedious.
- Most statistics package will compute the VIF statistics for you and report them with the coefficient output.
- You can then do the auxiliary regressions, if needed, for the variables with high VIF.

Using VIFs

- A general rule is that any VIF > 10 is a problem.
- Another is that if the average VIF is considerably larger than 1, SSE may be inflated.
- The average VIF indicates how many times larger SSE is due to multicollinearity than if the predictors were uncorrelated.
- Freund and Wilson suggest comparing the VIF to $1/(1-R^2)$ for the main model. If the VIF are less than this, multicollinearity is not a problem.

Our Example

Pairwise correlations

Correlations: SALES, ADV, BONUS, MKTSHR, COMPET

	SALES	ADV	BONUS	MKTSHR
ADV	0.900			
BONUS	0.568	0.419		
MKTSHR	0.023	-0.020	-0.085	
COMPET	0.377	0.452	0.229	-0.287

The maximum correlation among the x variables is .452 so if multicollinearity exists it is well hidden.

VIFs in Minitab

The regression equation is

SALES = - 594 + 2.51 ADV + 1.91 BONUS + 2.65 MKTSHR - .121 COMPET

Predictor	Coef	SE Coef	T	P	VIF
Constant	-593.5	259.2	-2.29	0.033	
ADV	2.5131	0.3143	8.00	0.000	1.5
BONUS	1.9059	0.7424	2.57	0.018	1.2
MKTSHR	2.651	4.636	0.57	0.574	1.1
COMPET	-0.1207	0.3718	-0.32	0.749	1.4

S = 93.77 R-Sq = 85.9% R-Sq(adj) = 83.1%

No Problem!

```
PROC  REG;
  MODEL SALES = ADV  BONUS  MKTSHR  COMPET / VIF ;
RUN:
```

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	-593.53745	259.19585	-2.29	0.0330	0
ADV	1	2.51314	0.31428	8.00	<.0001	1.47989
BONUS	1	1.90595	0.74239	2.57	0.0184	1.22186
MKTSHR	1	2.65101	4.63566	0.57	0.5738	1.11406
COMPET	1	-0.12073	0.37181	-0.32	0.7488	1.39368

VIF > 10 ??

4.6.3 Correction for Multicollinearity

- One solution would be to leave out one or more of the redundant predictors.
  >> Which one ? >> PCA (multivariate)

- Another would be to use the variables differently. If x_1 and x_2 are collinear, you might try using x_1 and the ratio x_2/x_1 instead.

- Finally, there are specialized statistical procedures (Ridge Reg.) that can be used in place of ordinary least squares.

Another Examples for Correlated Predictors

- Lagged Independent Variables

$$y_t = \beta_0 + \beta_1 x_t + \beta_2 x_{t-1} + \beta_3 x_{t-2} + e_t$$

- Lagged Dependent & Independent Variables

$$y_t = \beta_0 + \beta_1 y_{t-1} + \beta_2 x_t + \beta_3 x_{t-1} + e_t$$

4.7 Lagged Variables as Explanatory Variables in Time-Series Regression

- When using time series data in a regression, the relationship between y and x may be *concurrent* or x may serve as a *leading indicator*.
- In the latter, a past value of x appears as a predictor, either with or without the current value of x.
- An example would be the relationship between housing starts as y and interest rates as x. When rates drop, it is several months before housing starts increase.

Lagged Variables

The effect of advertising on sales is often cumulative so it would not be surprising see it modeled as:

$$y_t = \beta_0 + \beta_1 x_t + \beta_2 x_{t-1} + \beta_3 x_{t-2} + e_t$$

Here x_t is advertising in the current month and the lagged variables x_{t-1} and x_{t-2} represent advertising in the two previous months.

Potential Pitfalls

- If several lags of the same variable are used, it could cause multicollinearity if x_t was highly *autocorrelated* (correlated with its own past values).

- Lagging causes lost data. If x_{t-2} is included in the model, the first time it can be computed is at time period $t = 3$. We lose any information in the first two observations.

```
Data b  ;  set  a ;
  Lag1_adv = Lag1(ADV);


PROC  REG data=b ;
  MODEL SALES = ADV  LAG1_ADV   BONUS  / VIF ;
RUN:
```

Root MSE	87.14041	R-Square	0.8681
Dependent Mean	1281.75000	Adj R-Sq	0.8484
Coeff Var	6.79855		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	-620.04988	203.81216	-3.04	0.0064	0
ADV	1	2.73182	0.29925	9.13	<.0001	1.31199
lag1_adv	1	-0.30505	0.27939	-1.09	0.2879	1.35178
BONUS	1	2.29084	0.72635	3.15	0.0050	1.20548

Lagged *y* Values

- Sometimes a past value of *y* is used as a predictor as well. A relationship of this type might be:

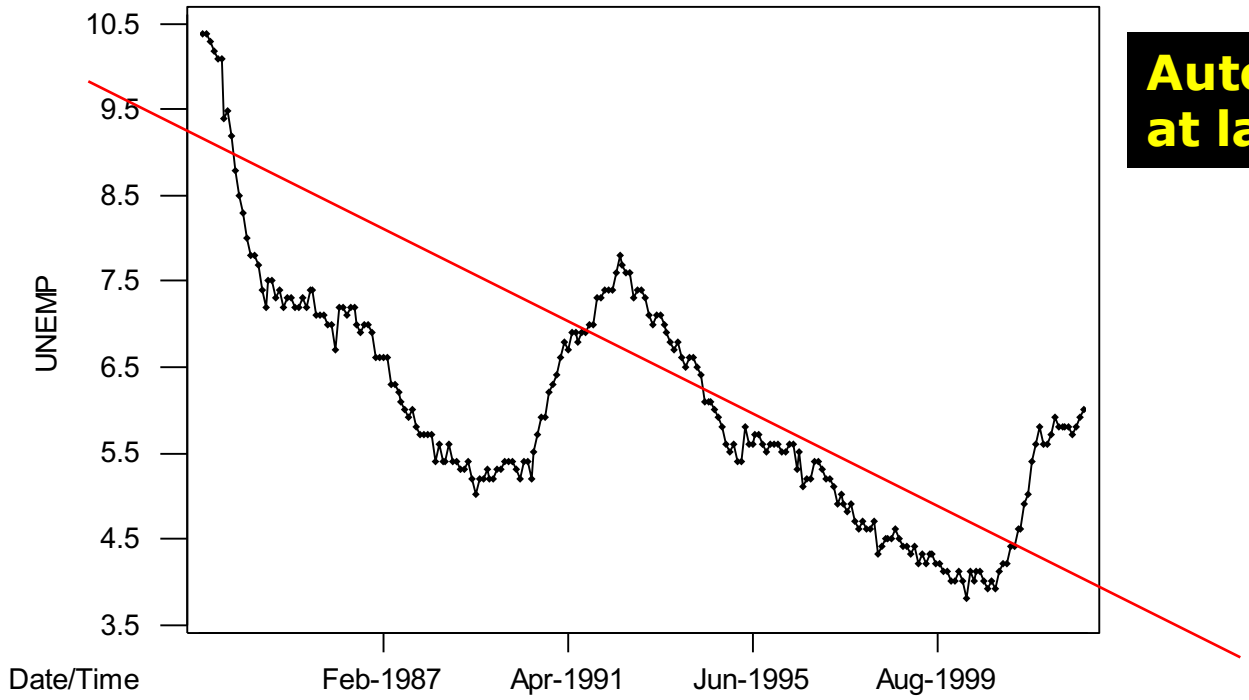
$$y_t = \beta_0 + \beta_1 y_{t-1} + \beta_2 x_t + \beta_3 x_{t-1} + e_t$$

- This implies that this month's sales y_t are related to by two months of advertising expense x_t and x_{t-1} plus last month's sales y_{t-1}.

Example 4.6 Unemployment Rate

- The file UNEMP4 contains the national unemployment rates (seasonally-adjusted) from January 1983 through December 2002.
- On the next few slides are a time series plot of the data and regression models employing first and second lags of the rates.

Time Series Plot



Autocorrelation is .97 at lag 1 and .94 at lag 2

$$y_t = \beta_0 + \beta_1 t + \epsilon_t$$

$$y_t = \beta_0 + \beta_1 y_{t-1} + \epsilon_t$$

$$y_t = \beta_0 + \beta_1 y_{t-1} + \beta_2 y_{t-2} + \epsilon_t$$

Trend Model

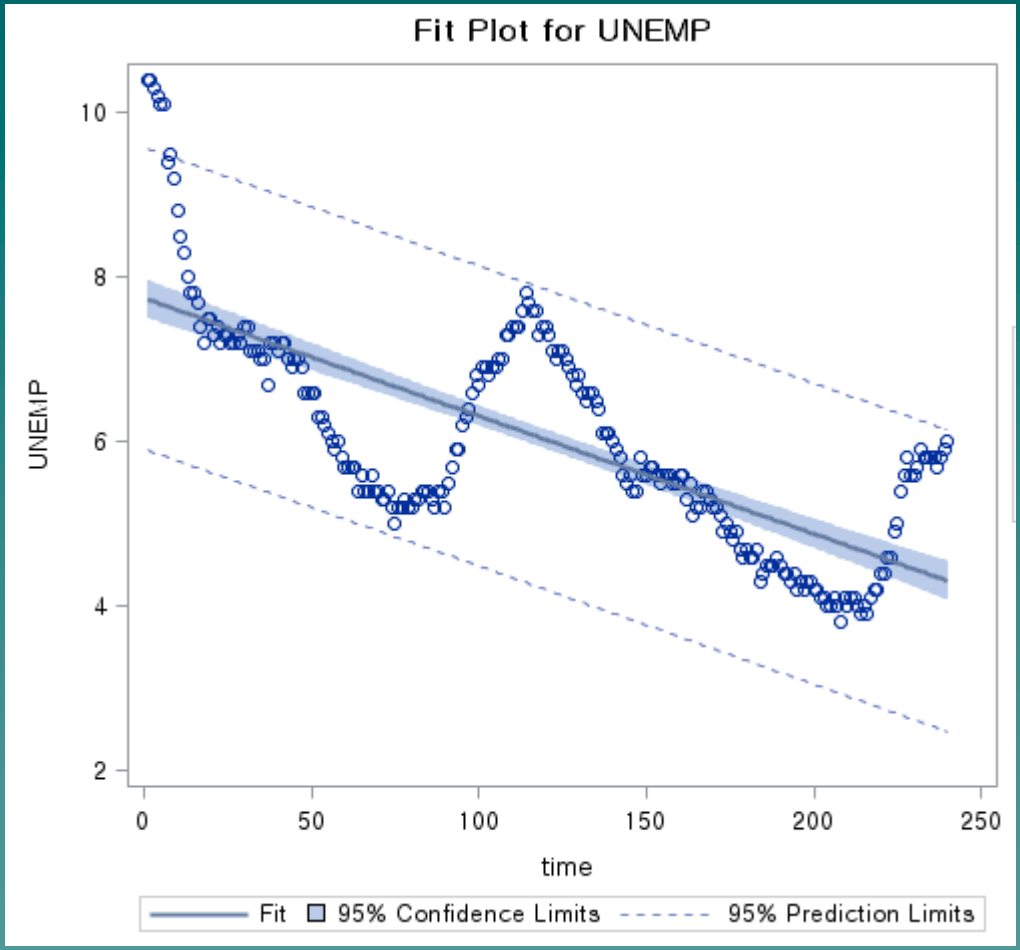
$$y_t = \beta_0 + \beta_1 t + \epsilon_t$$

```
PROC REG;
  MODEL UNEMP = TIME ;
RUN:
```

Root MSE	0.92400	R-Square	0.5378
Dependent Mean	6.02000	Adj R-Sq	0.5358
Coeff Var	15.34889		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	7.74628	0.11966	64.73	<.0001
time	1	-0.01433	0.00086090	-16.64	<.0001



- Not a good fit ! (R2 = 0.54)
- Errors are not independent !

Regression With First Lag

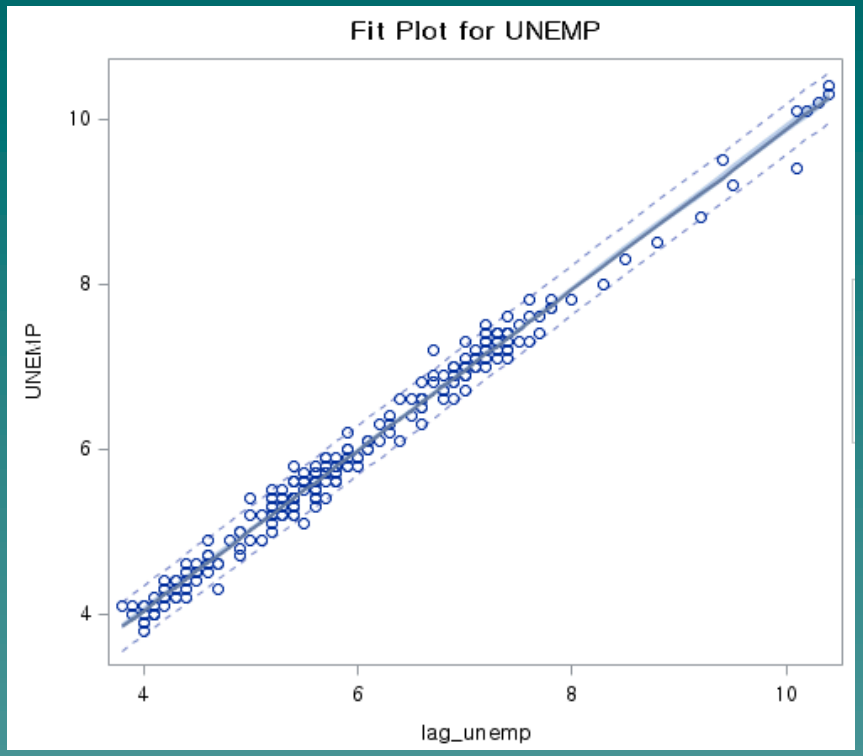
$$y_t = \beta_0 + \beta_1 y_{t-1} + \epsilon_t$$

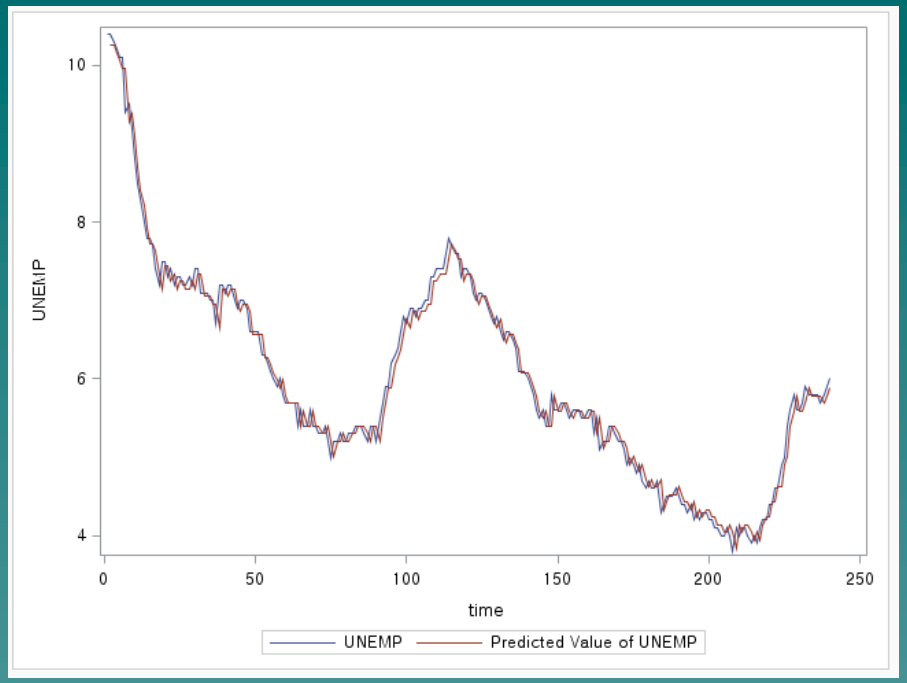
```
DATA  B;  SET  A;
  LAG_UNEMP = LAG1(UNEMP);

PROC  REG  DATA=B;
  MODEL  UNEMP  =  LAG_UNEMP  ;
RUN:
```

Root MSE	0.15153	R-Square	0.9871
Dependent Mean	6.00167	Adj R-Sq	0.9870
Coeff Var	2.52478		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	0.15319	0.04460	3.44	0.0007
lag_unemp	1	0.97149	0.00723	134.43	<.0001





- Good fit ! (R2 = 0.987)

Regression With First Lag

```
The regression equation is
UNEMP = 0.153 + 0.971 Unemp1

239 cases used 1 cases contain missing values

Predictor         Coef      SE Coef           T         P
Constant       0.15319      0.04460        3.44     0.001
Unemp1        0.971495     0.007227      134.43     0.000

S = 0.1515       R-Sq = 98.7%     R-Sq(adj) = 98.7%

Analysis of Variance

Source              DF           SS           MS          F         P
Regression           1       414.92       414.92   18070.47     0.000
Residual Error     237         5.44         0.02
Total              238       420.36
```

High R^2 because of autocorrelation

Regression With Two Lags

$$y_t = \beta_0 + \beta_1 y_{t-1} + \beta_2 y_{t-2} + \epsilon_t$$

```
The regression equation is
UNEMP = 0.168 + 0.890 Unemp1 + 0.0784 Unemp2

238 cases used 2 cases contain missing values

Predictor         Coef      SE Coef           T         P       VIF
Constant       0.16764      0.04565        3.67     0.000
Unemp1         0.89032      0.06497       13.70     0.000      77.4
Unemp2         0.07842      0.06353        1.23     0.218      77.4

S = 0.1514       R-Sq = 98.7%     R-Sq(adj) = 98.6%

Analysis of Variance

Source             DF          SS          MS          F         P
Regression          2      395.55      197.77    8630.30     0.000
Residual Error    235        5.39        0.02
Total             237      400.93
```

Comments

- It does not appear that the second lag term is needed. Its *t* statistic is 1.23.
- Because we got $R^2 = 98.7\%$ from the model with just one term, there was not much variation left for the second lag term to explain.
- Note that the second model also had a lot of multicollinearity.