Chapter 10 계량경제

시계열 데이터에는 경향(trend), 계절성(seasonality), 주기(cycle), 불규칙성(irregular) 성분이 있다. 주기를 자기상관(autocorrelation)으로 정의하기도 한다. 자기상관은 시점 t 와 $(t-1)$ 간의 상관 관계를 의미하며 한동안(일정 기간) 증가하거나 감소하는 경우 양의 자기상관이 존재한다고 하고 시점마다 증감이 반복되는 경우 이를 음의 자기상관이라 한다. 양의 자기상관이 일반적이다. 하루 100개 제품을 생산하는 생산 라인의 경우 기계에 문제가 생기면 한동안 100개 미만을 생산하기도 하고 수요가 많아지면 밤샘 작업을 통하여 초과 생산하는 경우 일별 생산량 데이터는 양의 자기상관을 가진다. 냉난방 기술자가 빌딩 온도를 매일 내렸다 올렸다(그의 습관) 하는 경우 실내 온도는 음의 자기상관을 갖는다.

시간 도표로부터 시계열 자료의 특성을 알 수 있다.

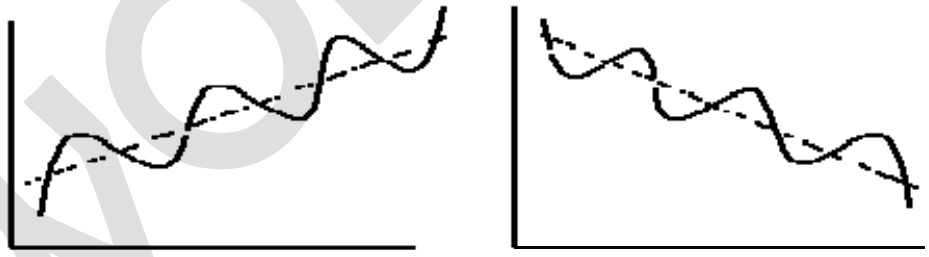
□경향(Trend): $\{Y_t; t=1,2,..,T\}$ 데이터가 증가(감소)하는 경향이 있는지 혹은 안정적인지 알 수 있다. 직선의 기울기가 있는가?

□주기(cycle): 일정한 주기(진폭)마다 유사한 변동이 반복된다.

□계절성(seasonality): 주별, 월별, 분기별, 년별 유사 패턴이 반복된다.

□불규칙성(irregular): 일정한 패턴을 따르지 않는다.

$$Y_t = Trend + Cycle + Seasonality + Irregular$$



시계열(time series) 데이터는 관측치가 시간적 순서를 가지게 된다. 일정 시점에 조사된 데이터는 횡단(cross-sectional) 자료라 한다. ○○전자 주가, △△기업 월별 매출액, 소매물가지수, 실업률, 환율 등이 시계열 자료이다.

시계열 데이터 분석의 목적을 살펴보면 다음과 같다.

□가장 중요한 목적은 미래 값을 **예측**하는 것이다. 향후 일주일간 주가 예측, 다음 달 매출액 예측 등

□시계열 데이터의 특성을 파악한다. 경향(trend), 주기(cycle), 계절성(seasonality), 불규

칙성(irregular) 등

주가, 소매물가지수 등 변수 하나에만 관심을 갖는 경우 이를 일변량(univariate) 시계열 데이터 분석이라 한다. 일변량 데이터의 예측치(forecasting value)는 자신들의 과거의 값들에만 의존한다. 다음은 주가(Stock price)를 예측하기 위한 $AR(p)$ 모형이다.

$$S_t = \alpha + \beta_1 S_{t-1} + \beta_1 S_{t-1} + ... + \beta_p S_{t-p}$$

시계열 데이터에 대한 회귀분석을 계량경제(econometrics)라 한다. 예를 들어 이자율(Interest), 인플레이션(Inflation)이 환율(Exchange rate)에 영향을 미치는 요인이라 하자. 다음과 같은 선형 회귀 모형을 생각할 수 있다. 종속변수의 미래 값을 예측하기 위하여 설명변수의 이전(t-1) 값이 필요하다. 여기서는 시차(time lag)를 (t-1)만 생각했는데 (t-2), (t-3), ...을 고려해야 한다.

$$ER_t = \alpha + \beta_1 IR_{t-1} + \beta_2 IF_{t-1}$$

여기서는 시계열 데이터에 대한 회귀분석 모형을 다루고자 한다. 시계열 데이터에 대한 다중회귀 모형은 다음과 같이 쓸 수 있다. p 는 설명 변수의 개수이고 t 는 관측 시점을 의미한다.

$$Y_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + ... + \beta_p X_{pt} + e_t,\ t = 1,2,..,n,\ e_t \sim iidN(0, \sigma^2)$$

시계열 데이터의 관측 시점간의 거리는 같다고 정의한다.(equally spaced time points) 일별, 월별, 연별 데이터가 수집된다. 주가의 경우 토, 일요일 자료는 없으나 금(t), 월($t+1$), ,.. 이런 식으로 일별 자료로 간주한다.

예제 자료

1951년-1953년 주별 데이터 이다.($n = 30$)

Quantity(아이스크림 소비량) Price(아이스크림 가격)

Income(주별 소득) Temp(주별 평균 온도)

A	B	C	D	E
date	IC	price	income	temp
1	0.386	0.270	78	41
2	0.374	0.282	79	56
3	0.393	0.277	81	63
4	0.425	0.280	80	68
5	0.406	0.272	76	69
6	0.344	0.262	78	65
7	0.327	0.275	82	61

```
data icecream;
    input date quantity price income temp;
    datalines;
1    0.386    0.270    78   41
2    0.374    0.282    79   56
```

10.1 모형 및 추정

우선 예제를 위하여 소득과 온도만이 아이스크림 소비량에 영향을 미친다고 하자.

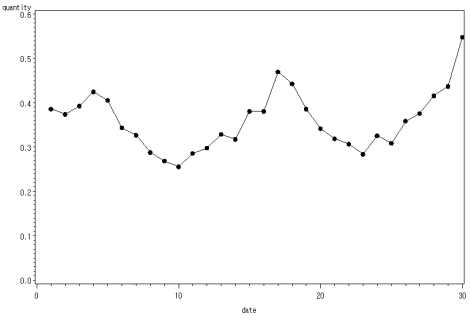
$$Q_t = \beta_0 + \beta_1 \times I_t + \beta_2 \times T_t + e_t, \quad e_t \sim iidN(0, \sigma^2)$$

오차항은 (1)정규성(normality) (2)등분산성(homoscadicity) (3)독립성(independency)을 가정한다. 횡단 자료에서는 독립성을 검정하지 않으나 시계열 데이터에서 중요하다.

10.1.1 Time Plot

시계열 데이터의 구조를 파악하기 위하여 시간도표를 먼저 그린다. 그러나 이것으로 설명변수의 상관 관계를 알 수 있거나 어떤 정보를 얻을 수 있는 것은 아니어서 회귀분석에서 유용한 도구는 아니다.

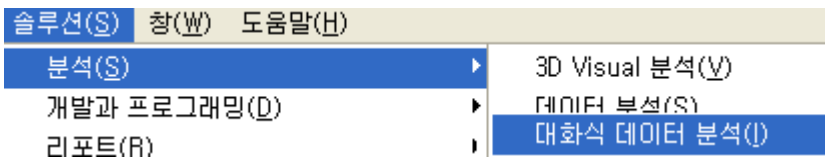
```
proc gplot data=icecream;
    axis1 order=0 to 0.6 by 0.1;
    symbol i=join v=dot;
    plot quantity*date/vaxis=axis1;
run;
```



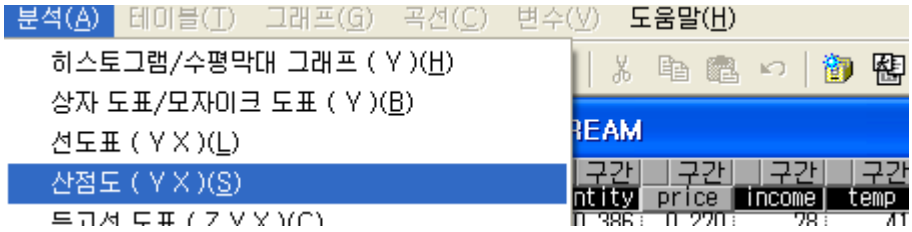
시간도표에 의하면 계절성과 경향이 존재하는 것 같다.

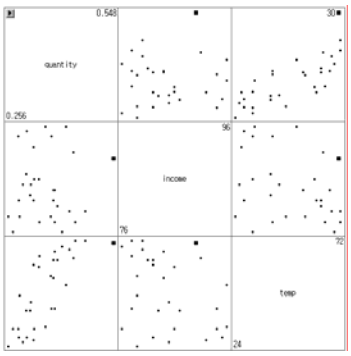
10.1.2 산점도 행렬

종속변수와 설명변수간의 선형 관계 존재 여부, 설명변수간의 다중공선성 존재 여부 등을 미리 파악하기 위하여 산점도 행렬을 그린다.



CRTL을 누른 후 변수 3개를 선택하고 분석 메뉴에서 산점도를 선택한다.





온도는 소비량에 양의 영향을 미친다. 소득은 글쎄? 소득과 온도간에는 상관 관계가 없어 보인다. 마지막 30번째 관측치가 이상치처럼 보인다.

10.1.3 회귀계수 유의성 검정

회귀 모형이 적합한지 알아보려면 분산분석에 의한 F-검정을 실시하고 회귀 계수의 유의성은 t-검정을 하면 된다. 이 모두를 한꺼번에 하려면

```
proc reg data=icecream;
    model quantity=income temp/selection=stepwise;
run;
```

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	0.08812	0.04406	31.81	<.0001
Error	27	0.03740	0.00139		
Corrected Total	29	0.12552			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	-0.11320	0.10828	0.00151	1.09	0.3051
income	0.00353	0.00117	0.01261	9.10	0.0055
temp	0.00354	0.00044496	0.08784	63.41	<.0001

소득과 온도 모두 양의 영향을 미친다. 어느 설명변수의 영향력이 더 큰지 알아보려면 표준화 회귀계수(standardized beta co-efficient)를 구한다.

```
proc reg data=one;
    model quantity=income temp/stb;
run;
```

온도의 영향력이 더 크다.

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate
Intercept	1	-0.11320	0.10828	-1.05	0.3051	0
income	1	0.00353	0.00117	3.02	0.0055	0.33512
temp	1	0.00354	0.00044496	7.96	<.0001	0.88444

10.1 4 다중공선성 확인

산점도 행렬에 의해 다중공선성(Multicollinearity) 문제가 발생하지 않을 것이라는 것을 짐작했지만 검정통계량에 의해 확인하자.

```
proc reg data=icecream;
    model quantity=income temp/vif collin;
run;
```

VIF($=1/(1-R_k^2)$)가 10이하이므로 문제가 없다.

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	-0.11320	0.10828	-1.05	0.3051	0
income	1	0.00353	0.00117	3.02	0.0055	1.11786
temp	1	0.00354	0.00044496	7.96	<.0001	1.11786

상태 지수(condition index $=\sqrt{\lambda_{\max}/\lambda_k}$)가 10 이상이면 문제가 발생한다. 3번째 경우 문제가 발생하는 것처럼 보이나 절편과 소득에 의한 문제이므로 설명변수간에는 다중공선성 문제가 발생하지 않는다.

Collinearity Diagnostics

Number	Eigenvalue	Condition Index	Proportion of Variation: Intercept	income	temp
1	2.92519	1.00000	0.00045572	0.00053896	0.00970
2	0.07266	6.34500	0.00643	0.01362	0.80522
3	0.00215	36.86474	0.99311	0.98584	0.18508

10.1.5 이상치 혹은 영향치 진단

RSTUDENT(표준화 제외잔차, ± 2), HAT($X'(X'X)^{-1}X$, $2(p+1)/n$, 영향치 진단), DFFITS(예측차이, $2\sqrt{(p+1)/n}$), DFBETAS(회귀계수 차이, 1) 등을 살펴 이상치나 영향치를

진단한다. 첫번째, 30번째 관측치가 이상치로 보인다. 30번째 관측치는 온도에 의해 이상치가 되었다. 산점도 행렬 결과와 일치한다.

```
proc reg data=icecream;
    model quantity=income temp/influence;
run;
```

Obs	Residual	RStudent	Hat Diag H	DFFITS	DFBETAS Intercept	DFBETAS income	DFBETAS temp
1	0.0786	2.4144	0.0988	0.7995	0.6689	-0.6077	-0.4176
2	0.009887	0.2695	0.0617	0.0691	0.0397	-0.0415	0.0070
3	-0.002977	-0.0810	0.0615	-0.0208	-0.0033	0.0050	-0.0108
29	0.002178	0.0615	0.1288	0.0237	-0.0181	0.0171	0.0161
30	0.0919	3.0911	0.1596	1.3469	-0.9622	0.8590	1.0687

10.1.6 독립성 검정

오차의 독립성은 Durbin and Watson(1951) 통계량의 의해 검정한다. DW 통계량은 오차 자기상관 존재여부를 판단한다. $e_t = \rho e_{t-1} + \varepsilon_t$, $\varepsilon_t \sim iidN(0, \sigma_*^2)$ 자기상관이 존재한다는 것은 회귀계수 ρ 가 0이 아니라는 것이다. 다음은 자기상관을 검정하는 DW 검정통계량이다.

$$DW = \sum_{i=2}^{n}(e_i - e_{i-1})^2 / \sum_{i=1}^{n} e_i^2$$

만약 자기상관이 존재하지 않으면 DW는 2에 근사한다.(why? DW 검정통계량에 $e_t = \rho e_{t-1} + \varepsilon_t$ 을 넣고 $\rho = 0$ 으로 해 보자) 임계치 D_L 과 D_U 는 자료의 개수와 설명변수의 개수 p 에 의존하며 표가 따로 주어진다. 만약 $D_L \le DW \le D_U$ 이면 귀무가설 채택한다. 그렇지 않으면 귀무가설 기각한다.

```
proc reg data=icecream;
    model quantity=income temp/dw;
run;
```

Dependent Variable: quantity

Durbin-Watson D	1.003
Number of Observations	30
1st Order Autocorrelation	0.303

SAS도 DW 검정통계량에 대한 유의확률이 주어지지 않으므로 표를 찾아야 하는 번거로움이 있다. 오차의 자기상관계수($Corr(e_t, e_{t-1})$) r 과 $DW \approx 2(1-r)$ 의 관계가 있으므로 오차(잔차, 오차의 추정치)의 자기상관계수를 이용하여 독립성을 검정할 수 있다. DW 통계량 표는 강의노트에서 다운받기 바란다. 다음은 그 일부분이다. p 는 설명변수의 개수이다.

$(0, D_L)$	(D_L, D_U)	$(D_U, 4-D_U)$	$(4-D_U, 4-D_L)$	$(4-D_L, D_L)$
귀무가설 기각 양의 자기 상관	미결정 H_0 기각도 채택도 하지 않음	귀무가설 채택 자기상관 없음	미결정 H_0 기각도 채택도 하지 않음	귀무가설 기각 음의 자기 상관

데이터($n=30, p=2$) DW 통계량은 1.003이었는데 DW-통계표에서 $(D_L=1.28, D_U=1.57)$ 이므로 오차의 자기상관이 존재한다. $(0, D_L)$ 사이에 있으므로 양의 자기상관이 존재한다.

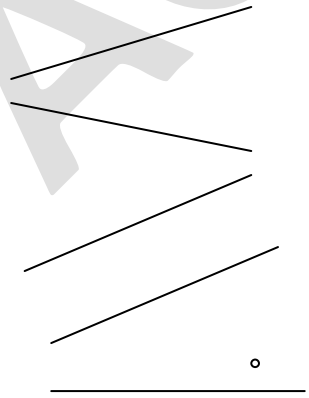
n	$p=1$		$p=2$		$p=3$		$p=4$	
	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U
29	1.34	1.48	1.27	1.56	1.20	1.65	1.12	1.74
30	1.35	1.49	1.28	1.57	1.21	1.65	1.14	1.74

(유의수준 5%)

10.1.7 잔차분석

비록 독립성 가정이 무너졌지만 회귀분석의 최종 단계인 잔차분석을 실시해보자. 잔차분석에서는 등분산성, 정규성과 이상치 판단 및 잔차의 패턴 분석을 실시한다.

```
goption reset=all;
proc gplot data=out1;
    symbol v=circle;
    plot res*phat;
run;

proc univariate data=out1 normal;
    var res;
run;
```



이분산 ▶ WLS

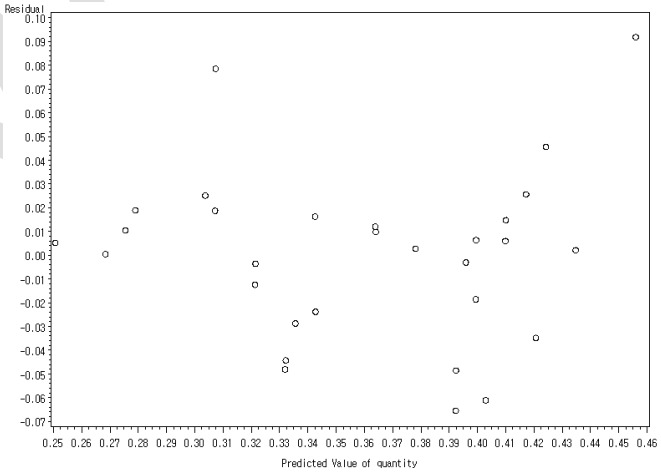
주요한 설명변수?

이상치

정규성 검정

검정		----통계량----		-------p-값-------	
Shapiro-Wilk	W	0.94798	Pr < W	0.1492	
Kolmogorov-Smirnov	D	0.136895	Pr > D	>0.1500	▶ 정규성 OK

잔차가 사이클을 갖는 패턴이 있다. 오차의 자기상관이 존재하기 때문일 것이다.



이제 무엇을 해야 하나?

10.2 해결책

오차항이 자기상관을 갖는 경우 간단한 해결책으로 1차 차분(first differencing) 방법을 생각할 수 있을 이용하는 것이다. $\nabla Y_t = Y_t - Y_{t-1}$ 그러나 이 방법은 설명변수까지 차분해야 하는 문제가 발생하므로 회귀모형에서는 적절하지 않다. (why?)

오차의 자기상관이란 설명변수들에 의해 설명되지 못한 오차들이 서로 영향을 주고 있다는 것이다. 지난 주의 소득이 영향을 주지 않을까? 새로운 모형을 다음과 같이 제안할 수 있다.

$$Q_t = \beta_0 + \beta_1 \times I_{t-1} + \beta_2 \times Temp_t + e_t, \quad e_t \sim iidN(0, \sigma^2)$$

아니 왜 이번 주 소득은 없는가? 만약 금주 소득과 지난 주 소득을 동시에 넣으면 다중공선성 문제가 발생하는 것은 당연하다. 의심이 납니까? 한번 해보기 바랍니다. (VIF, COLLIN 옵션)

```
data icecream1;
    set icecream;
    lag_in=lag(income);
run;
proc reg data= icecream1;
    model quantity=lag_in temp/vif collin dw;
    plot student.*predicted./vref=-2 2;
    output out=out1 r=res p=phat;
run;
```

LAG 함수는 $(t-1)$ 관측치

VIF, COLLIN, DW 옵션을 동시에 사용하는 것은 좋지 않으나 우리는 이미 이전 결과가 있으므로 예측할 수 있다.

```
data out1;
    set out1;
    lagres=lag(res);
run;
```

```
proc corr data=out1;
    var res lagres;
run;
```

```
proc arima data=out1;
    identify var=res;
    estimate p=1 noconstant;
run;
```

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate	Variance Inflation
Intercept	1	-0.20332	0.08768	-2.32	0.0285	0	0
lag_in	1	0.00458	0.00096686	4.73	<.0001	0.42975	1.05491
temp	1	0.00356	0.00036436	9.76	<.0001	0.88649	1.05491

Collinearity Diagnostics

Number	Eigenvalue	Condition Index	Proportion of Variation: Intercept	lag_in	temp
1	2.92570	1.00000	0.00050590	0.00057782	0.01041
2	0.07194	6.37712	0.00774	0.01403	0.87469
3	0.00236	35.22070	0.99175	0.98539	0.11489

VIF는 10보다 높지 않으니 문제가 없다. 상태지수(condition index)가 10이(어떤 이는 30을 기준으로 본다) 넘는 곳이 있다. 그러나 절편과 LOGIN 변수가 문제를 일으키므로(변동비율이 높은 변수간, 일반적으로 0.8 이상) 다중공선성 문제는 아니다.

오차의 독립성 검정

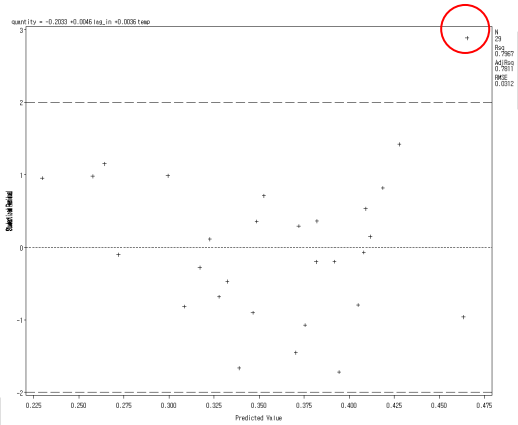
Dependent Variable: quantity

Durbin-Watson D	1.229
Number of Observations	29
1st Order Autocorrelation	0.243

오차의 자기상관이 존재하는지 결론을 내릴 수 없는 것으로 판단된다. 그러므로 오차의 독립성이 가정된다고 하자??? 일단은 (최종적 해결은 10.4절 참고) 이상치 진단과 잔차분석을 실시하자.

정규성 검정

검정		----통계량----	-------p-값-------
Shapiro-Wilk	W	0.973656	Pr < W 0.6621
Kolmogorov-Smirnov	D	0.065995	Pr > D >0.1500



이상치 하나를 제외하고 최종 회귀 모형을 구하자.

```
proc reg data=icecream1;
    model quantity=lag_in temp/stb;
    reweight obs.=30;
run;
```

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate
Intercept	1	-0.12247	0.07740	-1.58	0.1261	0
lag_in	1	0.00378	0.00084549	4.47	0.0001	0.41481
temp	1	0.00321	0.00032228	9.97	<.0001	0.92558

$$\hat{Q}_t = -0.12247 + 0.00378 I_{t-1} + 0.00321 Temp_t$$

최종 회귀 모형: $(p = 0.001)$ $(p < 0.001)$, 지난 주 소득과 온도는 아이스크림 소비량에 양의 영향을 미치고 온도의 영향력이 소득에 비해 더 높음을 알 수 있다. 또한 온도가 소득에 비해 아이스크림 소비량에 영향을 많이 미침을 알 수 있다. (표준화 회귀계수)

이 모형의 문제는 예측에 있다. 다음 주 아이스크림 소비량(Q_{t+1})을 예측하기 위하여 다음 중 평균 온도에 대한 예측치($Temp_{t+1}$)가 있어야 한다. 온도에 대한 예측은 어느 정도 신뢰할 수 있지만 그러나 경제 변수인 경우에는??? 그래서 설명변수로는 과거치를 넣으려는 시도가 많다.

10.3 변환 데이터 회귀 모형

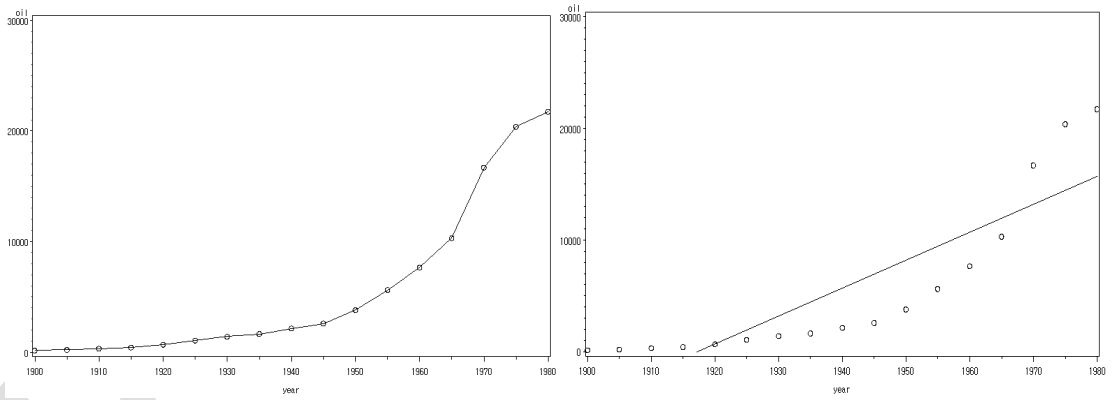
인구와 같이 꾸준히 증가하거나(요즈음은 다소 주춤하지만) 감소하는 시계열 데이터의 경우 시간을 설명 변수로 사용할 수 있을 것이다. $Y_t = \alpha + \beta T + e_t$

OIL.txt 데이터는 1900부터 1980년($n=17$, 매 5년 단위) 전세계 오일 생산량(OIL, 단위 백만 배럴)을 조사한 자료이다. 우선 Time plot을 그려보자.

참고YEAR 사용할 때 1900년을 그대로 사용하는 것과 $t=1$로 사용하는 것의 차이가 있는가? 차이는 없다. 시작이 1인가 1990인가 밖에 차이는 없다. 위의 모형의 경우

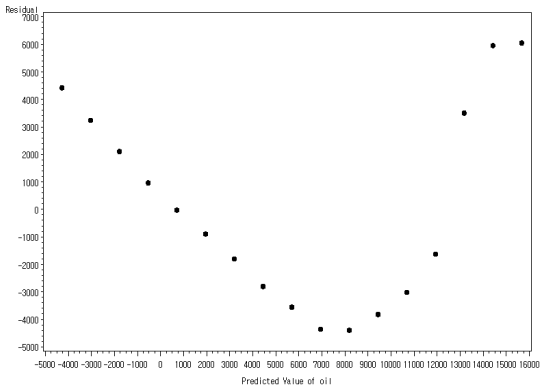
```
proc gplot data=oil;
    symbol i=join v=circle;
    plot oil*year;
run;
```

```
proc gplot data=oil;
    symbol i=rl v=circle;
    plot oil*year;
run;
```



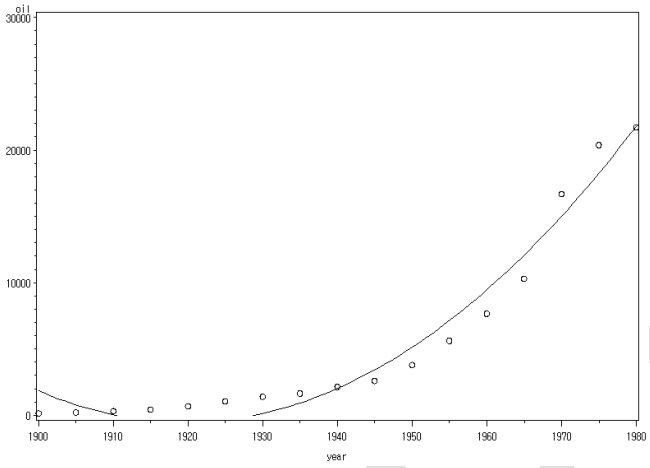
잔차분석의 잔차와 $\hat{Y}$의 산점도가 아래 형태를 띠면 해결책은 무엇인가?

```
goptions reset=all;
proc gplot data=out1;
    symbol v=dot;
    plot res*yhat;
run;
```



설명변수의 제곱 항을 설명변수로 고려하면 될 것이다.

```
proc gplot data=oil;
    symbol i=rq v=circle;
    plot oil*year;
run;
```



다음과 같이 설명변수의 제곱 항을 넣고 실시한 회귀분석 절차이다. 제곱 항을 넣었으므로 다중공선성(year와 year2) 문제가 발생한다. 좋은 해결 방법은 아니다.

```
data oil1;
    set oil;
        year2=year**2;
    run;
```

```
proc reg data=oil1;
    model oil=year year2/dw vif collin;
    output out=out1 p=yhat r=res;
run;
```

```
goptions reset=all;
proc gplot data=out1;
    symbol v=dot;
    plot res*yhat;
run;
```

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	22517697	2417891	9.31	<.0001	0
year	1	-23461	2492.95036	-9.41	<.0001	31694
year2	1	6.11104	0.64250	9.51	<.0001	31694

Collinearity Diagnostics

Number	Eigenvalue	Condition Index	Proportion of Variation		
			Intercept	year	year2
1	2.99968	1.00000	2.236156E-9	5.58914E-10	2.233589E-9
2	0.00031860	97.03197	0.00003159	2.89311E-13	0.00003154
3	3.352592E-9	29912	0.99997	1.00000	0.99997

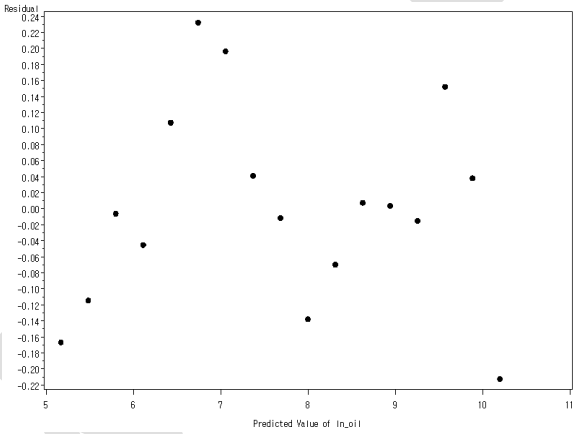
모형을 다음과 같이 생각해 보자. $O_t = \alpha\beta^{Year} e_t$ (지수 성장 모형) 양변에 자연 로그를 취하면 $\ln Q_t = \ln\alpha + \ln(\beta)\times Year + \ln(e_t)$ <=> $Q_t^* = \alpha^* + \beta^* \times Year + e_t^*$ (선형 회귀 모형)

```
data oil2;
    set oil;
        ln_oil=log(oil);
    run;

proc reg data=oil2;
    model ln_oil=year/dw;
    output out=out1 p=yhat r=res;
run;

data out2;
    set out1;
        lag_res=lag(res);
run;

proc corr data=out2;
    var res lag_res;
run;

proc arima data=out1;
    identify var=res;
    estimate p=1 noconstant;
run;

goptions reset=all;
proc gplot data=out1;
    symbol v=dot;
    plot res*yhat;
run;
```

피어슨 상관 계수
H0: Rho=0 검정에 대한 Prob > |r|
관측치 개수

	res	lag_res
res	1.00000	0.47715
Residual		0.0616
	17	16

Parameter	Estimate	Standard Error	t Value	Approx Pr > |t|
AR1,1	0.50322	0.24684	2.04	0.0584



Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-114.24977	2.44416	-46.74	<.0001
year	1	0.06285	0.00126	49.89	<.0001

최종 회귀 모형은 다음과 같다. $\ln(\hat{Oil}_t) = -114.25 + 0.06285 * Year_t$

이제 이것을 이용하여 1985년 1990년 오일 생산량을 예측해 보자. 데이터 마지막 라인에 예측하고자 하는 연도를 넣고 오일은 결측치로 처리한다.

```
1980    21732
1985 .
1990 .

data oil2;
    set oil;
        ln_oil=log(oil);
    run;

proc reg data=oil2;
    model ln_oil=year/p cli;
run;
```

출력 결과 마지막 두 행은 1985년, 1990년 오일 소비량 예측치와 신뢰구간이다.

Obs	Dep Var ln_oil	Predicted Value	Std Error Mean Predict	95% CL Predict		Residual
1	5.0039	5.1707	0.0591	4.8717	5.4697	-0.1667
2	5.3706	5.4850	0.0538	5.1905	5.7794	-0.1143
16	9.9228	9.8847	0.0538	9.5902	10.1791	0.0381
17	9.9865	10.1989	0.0591	9.8999	10.4979	-0.2124
18	.	10.5132	0.0645	10.2091	10.8173	.
19	.	10.8274	0.0701	10.5178	11.1371	.

그러므로 19985년 오일 생산량은 $e^{10.5132} = 36798$ 이다.

10.4 계절성

$Q_t = \beta_0 + \beta_1 \times P_t + \beta_2 \times Temp_t + e_t$ 모형의 오차항은 독립성을 만족하지 못한다. 해결책으로 설명변수 과거치를 설명변수화을 사용해도 독립성 불만족 문제를 해결하지 못한다. (다른 방법은 없는가?

아이스크림 Time plot을 살펴보면 주기(cycle)가 13번째마다 반복됨(seasonality)을 알 수 있다. 그래서 설명변수로 13주전 아이스크림 소비량을 고려하였고 아이스크림이 증가하는 경향(trend)이 있으므로 시간(주별 데이터이므로 주)을 설명변수로 추가하였다. LAG13은 13 시점 전의 관측치를 의미하는 것으로 Q_{t-13} 이다.

```
data icecream1;
    set icecream;
    lag_q=lag13(quantity);
run;

proc reg data=icecream1;
    model quantity=week lag_q temp/ vif collin dw;
    reweight obs.=30;
run;
```

위 모형은 $Q_t = \beta_0 + \beta_1 \times W_t + \beta_2 \times Q_{t-13} + \beta_3 \times Temp_t + e_t$ 이다.

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	-0.04343	0.04950	-0.88	0.3976	0
week	1	0.00469	0.00111	4.21	0.0012	1.20961
lag_q	1	0.65492	0.14414	4.54	0.0007	2.46048
temp	1	0.00172	0.00048076	3.58	0.0038	2.44617

Collinearity Diagnostics

Number	Eigenvalue	Condition Index	Proportion of Variation: Intercept	week	lag_q	temp
1	3.88382	1.00000	0.00058368	0.00229	0.00057813	0.00223
2	0.09539	6.38097	0.00303	0.17943	0.00322	0.17089
3	0.01600	15.57995	0.13896	0.57123	0.13188	0.57063
4	0.00479	28.47140	0.85743	0.24705	0.86432	0.25625

Durbin-Watson D	2.234
Number of Observations	16
1st Order Autocorrelation	-0.224

(오차의 독립성 성립)

최종회귀모형: $Q_t = -0.043 + 0.0047 W_t + 0.65 Q_{t-13} + 0.002 Temp_t$

EXAMPLE

1978년~1985년까지 4개 가정용품 출하액과 미국인 내구재 소비 지출액(DUR)을 분기마다 조사한 자료이다. 🖫APPLIANCE.txt FRIG(냉장고) 출하액에 미치는 영향으로 DUR, 분기, 시간 변수를 고려하고 회귀분석을 실시하시오.

QTR: Quarter, from 1st quarter 1978 to 4th quarter 1985

DISH: Unit factory shipments of dishwashers (thousands)

DISP: Unit factory shipments of disposers (thousands)

FRIG: Unit factory shipments of refrigerators (thousands)

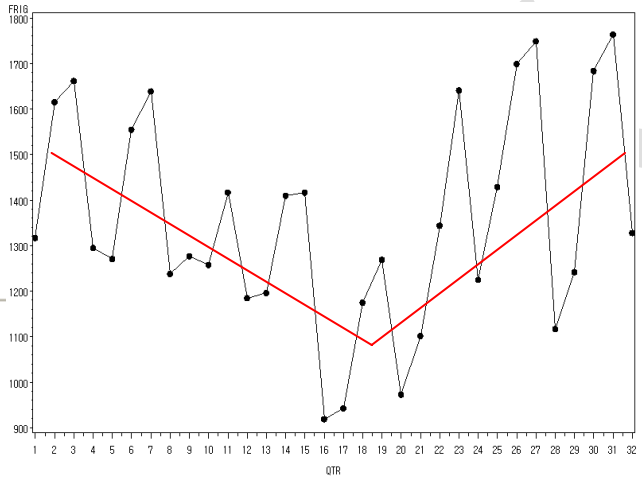
WASH: Unit factory shipments of washers (thousands)

DUR: U.S. durable goods expenditures (billions of 1982 dollars)

우선 냉장고 출하액에 산점도를 그려보자.

```
data appliance;
  input QTR DISH DISP FRIG WASH DUR;
  cards;
1       841     798     1317    1271    252.6
2       957     837     1615    1295    272.4
```

```
goptions reset=all;
proc gplot data=appliance;
    axis1 order=1 to 32 ;
    symbol i=join v=dot;
    plot frig*qtr/haxis=axis1;
run;
```



분기별 데이터이므로 분기에 따른 차이가 있을 것이므로 분기(I, II, III, IV)를 지시변수(indicator variable 혹은 가변수 dummy)로 고려하자. 분기가 4분기이므로 지시변수의 개수는 3개이어야 한다.(D1, D2, D3라 하자.) 그리고 감소하다가 감소하는 경향이 있으므로 시간에 대한 2차 함수를 넣으면 될 것이다. 즉 최초의 모형은 다음과 같다.

$$F_t = \beta_0 + \beta_1 \times Q_t + \beta_2 \times Q_t^2 + \beta_3 \times D1_t + \beta_4 \times D2_t + \beta_5 \times D3_t + \beta_6 \times DUR_t + e_t$$

우선 지시변수와 제곱 변수를 만들자.

```
data appliance;
  input QTR DISH DISP FRIG WASH DUR;
  if qtr=(int(qtr/4)*4) then do;
        d1=0; d2=0; d3=0;
  end;
  if qtr=(int(qtr/4)*4+1) then do;
        d1=1; d2=0; d3=0;
  end;
  if qtr=(int(qtr/4)*4+2) then do;
        d1=0; d2=1; d3=0;
  end;
  if qtr=(int(qtr/4)*4+3) then do;
        d1=0; d2=0; d3=1;
  end;
  qtr2=qtr*qtr;
  cards;
1       841     798     1317    1271
2       957     837     1615    1295
```

데이터 만드는 곳에서 변수 변환이 가능하다.

①INT 함수는 정수 값 얻는다.

②1분기이면 D1=1, D2=0, D3=0 이런 식으로 분기가 구별된다. 이 변수는 절편에 추가된다.

③QTR2는 QTR의 제곱

우선 STEPWISE 옵션을 이용하여 유의한 변수를 찾자. SAS에서 SLS=0.15가 default이므로 D1 변수도 채택되었다. 유의수준을 0.05로 하려면 SLS=0.05로 하시오.

```
proc reg data=appliance;
    model frig=qtr qtr2 dur d1 d2 d3/selection=stepwise;
run;
```

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	-632.60150	198.71842	72158	10.13	0.0038
qtr2	-0.60323	0.09620	279955	39.32	<.0001
DUR	7.15873	0.80604	561641	78.88	<.0001
d1	64.23923	42.47595	16286	2.29	0.1425
d2	320.88703	42.31870	409395	57.50	<.0001
d3	392.75338	42.29582	613970	86.23	<.0001

0.05에서 유의한 변수만을 가지고 다중공선성(QTR과 QTR2에 의해 다중공선성 문제를 발생하나 마침 QTR은 제외되었다.) 문제 진단과 독립성 검정을 실시해 보자.

```
proc reg data=appliance;
    model frig=qtr qtr2 dur d2 d3/dw vif collin;
run;
```

다중공선성 문제는 없다.

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	-606.68022	202.64235	-2.99	0.0058	0
qtr2	1	-0.61516	0.09814	-6.27	<.0001	4.07491
DUR	1	7.19608	0.82465	8.73	<.0001	4.08728
d2	1	288.66925	37.42733	7.71	<.0001	1.12668
d3	1	360.66694	37.45191	9.63	<.0001	1.12816

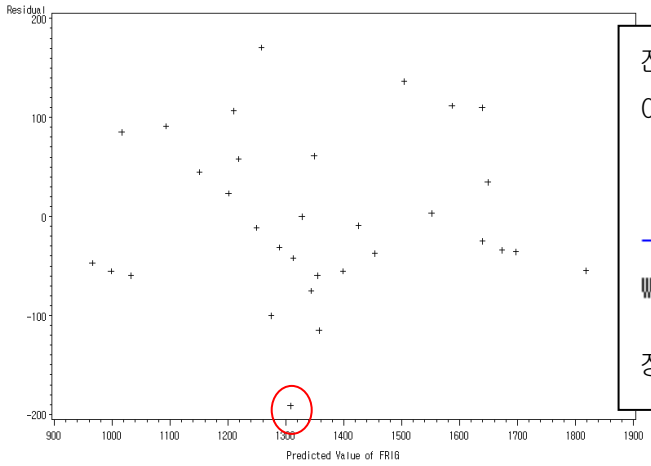
Collinearity Diagnostics

Condition Index	Proportion of Variation: Intercept	qtr2	DUR	d2	d3
1.00000	0.00049180	0.00739	0.00038369	0.01943	0.02071
1.80778	0.00000116	0.00006109	5.74356E-8	0.34587	0.31849
2.60241	0.00002212	0.06945	0.00016097	0.42686	0.49108
3.65835	0.00681	0.19385	0.00222	0.20460	0.16915
36.82812	0.99267	0.72924	0.99724	0.00324	0.00057107

DW 통계량 표가 있으므로 이것을 이용하자. ($p=4, n=32$)이므로 5%에서 (1.18, 1.73)가 채택 역이다. 그러므로 독립성이 만족한다.

Dependent Variable: FRIG

Durbin-Watson D	1.518
Number of Observations	32
1st Order Autocorrelation	0.213



잔차분석 실시 결과 이상치가 한 개 이상 존재하였다.

정규성 검정

----통계량----		-------p-값-------	
W	0.968664	Pr < W	0.4634

정규성을 만족한다.

관측치와 예측치를 한 그래프에 그려 추정된 회귀모형에 의해 얼마나 잘 예측되는지 시각적으로 알아보자. 이상치는 제외하지 않았다.

```
proc reg data=appliance;
    model frig=qtr2 dur d2 d3;
    reweight obs.=28;
    reweight obs.=29;
    reweight obs.=23;
    reweight obs.=25;
    reweight obs.=27;
    reweight obs.=26;
    output out=out1 p=yhat r=res  student=rs;
run;
```

STUDENT는 표준화 제외 잔차로 2 이상이면 이상치이다. 이상치인 관측치를 REWEIGHT 사용하여 하나씩 제외한다. 값을 보려면 PRINT 사용

```
proc print data=out1;
run;
```

하면 된다.

```
goptions reset=all;
proc gplot data=out1;
    symbol i=join v=dot;
    axis1 order=1 to 32 ;
    plot (frig yhat)*qtr/haxis=axis1 overlay;
run;
quit;
```