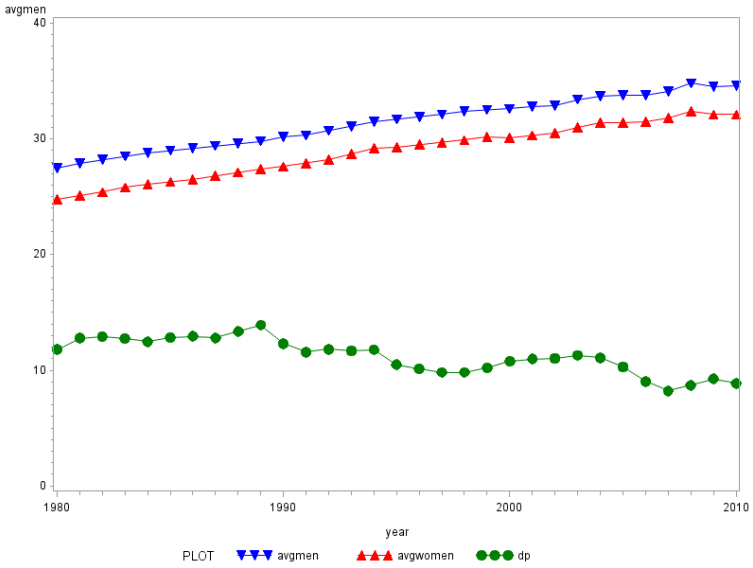
Solution to exercise 4.3
We initially present the SAS program used in the computations.

```
/*SAS-program Exercise 4.3. Data are taken from www.statistikbanken.dk*/
data divorce;
input year avgmen avgwomen dp;
cards;
1980 27.5 24.8 11.79
1981 27.9 25.1 12.78
1982 28.2 25.4 12.9
1983 28.5 25.8 12.73
1984 28.8 26.1 12.47
1985 29 26.3 12.82
1986 29.2 26.5 12.94
1987 29.4 26.8 12.8
1988 29.6 27.1 13.35
1989 29.8 27.4 13.91
1990 30.2 27.6 12.31
1991 30.3 27.9 11.57
1992 30.7 28.2 11.81
1993 31.1 28.7 11.68
1994 31.5 29.2 11.78
1995 31.7 29.3 10.48
1996 31.9 29.5 10.12
1997 32.1 29.7 9.82
1998 32.4 29.9 9.82
1999 32.5 30.2 10.2
2000 32.6 30.1 10.79
2001 32.8 30.3 10.98
2002 32.9 30.5 11.03
2003 33.4 31 11.29
2004 33.7 31.4 11.08
2005 33.8 31.4 10.28
2006 33.8 31.5 9.02
2007 34.1 31.8 8.21
2008 34.8 32.4 8.71
2009 34.5 32.1 9.25
2010 34.6 32.1 8.86
;
run;


Title 'Model Men';
proc reg data = divorce;
model dp = avgmen/r influence;
run;
```

```
Title ’Model Women’;
proc reg data = divorce;
model dp = avgwomen/r influence;
run;


Title 'Plot of divorce data';
proc sgplot data=divorce;
series x=year y=avgmen /lineattrs=(color=blue) markers
markerattrs=(symbol=trianglefilled) MARKERFILLATTRS=(color=blue);
series x=year y=avgwomen /lineattrs=(color=red) markers
markerattrs=(symbol=trianglefilled) MARKERFILLATTRS=(color=red);
series x=year y=dp /lineattrs=(color=green) markers
markerattrs=(symbol=trianglefilled) MARKERFILLATTRS=(color=green);
run;

Title 'Regression on avg. men';
proc reg data = divorce plots(label)=all;
model dp = avgmen/r influence dwprob;
output out=divout residual=dpresid;
run;
proc print data=divout;
run;


proc sgplot data=divout;
series x=year y=dp /lineattrs=(color=green) markers
markerattrs=(symbol=trianglefilled) MARKERFILLATTRS=(color=green);
series x=year y=dpresid /lineattrs=(color=violet) markers
markerattrs=(symbol=trianglefilled) MARKERFILLATTRS=(color=violet);
run;

Title 'Plot autocorrelation function';
proc timeseries data=divout plots=(corr);
var dpresid;
run;

Title 'Regression on avg. women';
proc reg data = divorce;
model dp = avgwomen/r influence dwprob;
run;

Title 'Regression on both averages';
proc reg data = divorce;
model dp = avgwomen avgmen/r influence dwprob;
run;
```

Some selected output

Plot and print-out of the data



Obs	year	avgmen	avgwomen	dp
1	1980	27.5	24.8	11.79
2	1981	27.9	25.1	12.78
3	1982	28.2	25.4	12.90
4	1983	28.5	25.8	12.73
5	1984	28.8	26.1	12.47
6	1985	29.0	26.3	12.82
7	1986	29.2	26.5	12.94
8	1987	29.4	26.8	12.80
9	1988	29.6	27.1	13.35
10	1989	29.8	27.4	13.91
11	1990	30.2	27.6	12.31
12	1991	30.3	27.9	11.57
13	1992	30.7	28.2	11.81
14	1993	31.1	28.7	11.68
15	1994	31.5	29.2	11.78
16	1995	31.7	29.3	10.48
17	1996	31.9	29.5	10.12
18	1997	32.1	29.7	9.82
19	1998	32.4	29.9	9.82
20	1999	32.5	30.2	10.20
21	2000	32.6	30.1	10.79
22	2001	32.8	30.3	10.98
23	2002	32.9	30.5	11.03
24	2003	33.4	31.0	11.29
25	2004	33.7	31.4	11.08
26	2005	33.8	31.4	10.28
27	2006	33.8	31.5	9.02
28	2007	34.1	31.8	8.21
29	2008	34.8	32.4	8.71
30	2009	34.5	32.1	9.25
31	2010	34.6	32.1	8.86

1. The general linear model for the divorce rate written as a function of average age of the male is

$$\begin{bmatrix} DP_{1980} \\ DP_{1981} \\ \vdots \\ DP_{2010} \end{bmatrix} = \begin{bmatrix} 1 & avgmen_{1980} \\ 1 & avgmen_{1981} \\ \vdots & \vdots \\ 1 & avgmen_{2010} \end{bmatrix} \begin{bmatrix} \alpha_{men} \\ \beta_{men} \end{bmatrix} + \begin{bmatrix} \varepsilon_{1980} \\ \varepsilon_{1981} \\ \vdots \\ \varepsilon_{2010} \end{bmatrix}$$

or shortly

$$\boldsymbol{DP} = \boldsymbol{x}_{men}\boldsymbol{\theta}_{men} + \boldsymbol{\varepsilon}$$

Similarly for women

$$\boldsymbol{DP} = \boldsymbol{x}_{women}\boldsymbol{\theta}_{women} + \boldsymbol{\delta}$$

We furthermore assume that $\boldsymbol{\varepsilon} \sim N_{31}(\mathbf{0}, \sigma^2_{men}\boldsymbol{I})$ and $\boldsymbol{\delta} \sim N_{31}(\mathbf{0}, \sigma^2_{women}\boldsymbol{I})$.

2. The parameters estimated using model statements like

model dp = avgmen;

will automatically include an intercept and thus give estimates for both parameters for men and similarly for women.

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	50.21905	50.21905	77.76	<.0001
Error	29	18.72969	0.64585		
Corr. Total	30	68.94874			

Variable	DF	Par. Est-imate	Std. Error	t Val-ue	Pr > |t|
Intercept	1	29.70795	2.10246	14.13	<.0001
avgmen	1	-0.58910	0.06681	-8.82	<.0001

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	49.66591	49.66591	74.69	<.0001
Error	29	19.28283	0.66493		
Corr. Total	30	68.94874			

Variable	DF	Par.Est-imate	Std. Error	t Val-ue	Pr > |t|
Intercept	1	27.25241	1.86172	14.64	<.0001
avgwomen	1	-0.55490	0.06421	-8.64	<.0001

3. We start with males. We are considering the new observation DP_{2011} and the prediction becomes

$$U = \widehat{DP}_{2011} = [1 \quad avgmen_{2011}] \begin{bmatrix} \hat{\alpha}_{men} \\ \hat{\beta}_{men} \end{bmatrix} = 9.21$$

Furthermore, we in general have $V(U) = \sigma^2 z^T (x^T \Sigma^{-1} x) z = \sigma^2 c$. We insert the numbers

$$c = [1 \; 34.8] \begin{bmatrix} 6.8442 & -0.2170 \\ -0.2170 & 0.0069 \end{bmatrix} \begin{bmatrix} 1 \\ 34.8 \end{bmatrix} = 0.1123$$

With $\hat{\sigma}^2 = 0.64585$ we get the estimated standard deviation of $\widehat{DP}_{2011}$

$$\hat{\sigma}\sqrt{c} = \sqrt{0.64585}\sqrt{0.1123} = 0.2693$$

Since $t(29)_{0.975} = 2.045$, we get the following confidence interval for the mean of the predicted variable

$$\widehat{DP}_{2011} \pm 2.045 \times 0.2693 = 9.21 \pm 0.5517 = [8.66, 9.76]$$

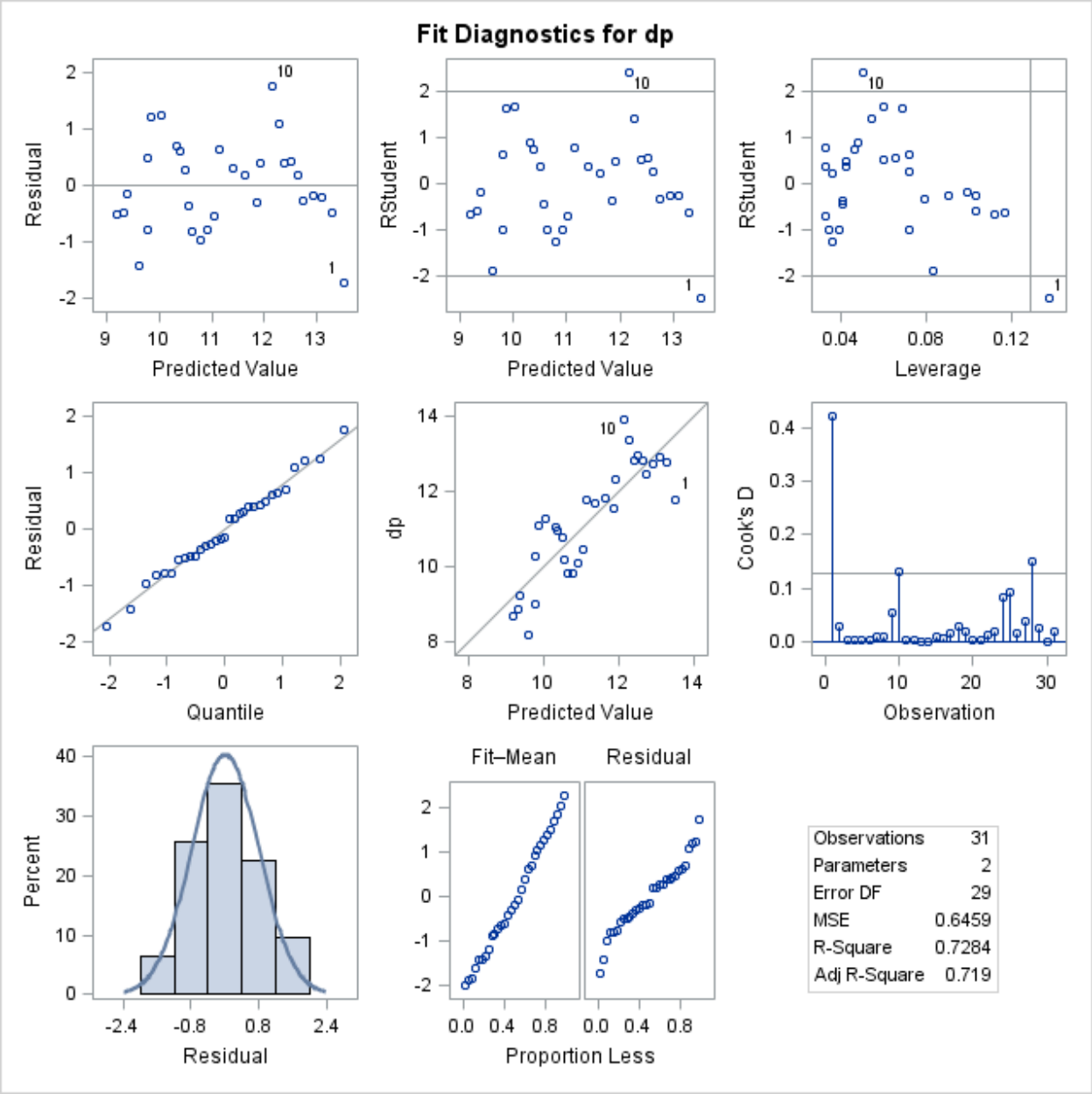
The prediction interval for the observation itself becomes

$$\widehat{DP}_{2011} \pm 2.045 \times \hat{\sigma}\sqrt{1 + c} = 9.21 \pm 1.7333 = [7.48, 10.94]$$

If the predictions are based on females the confidence and the prediction intervals become

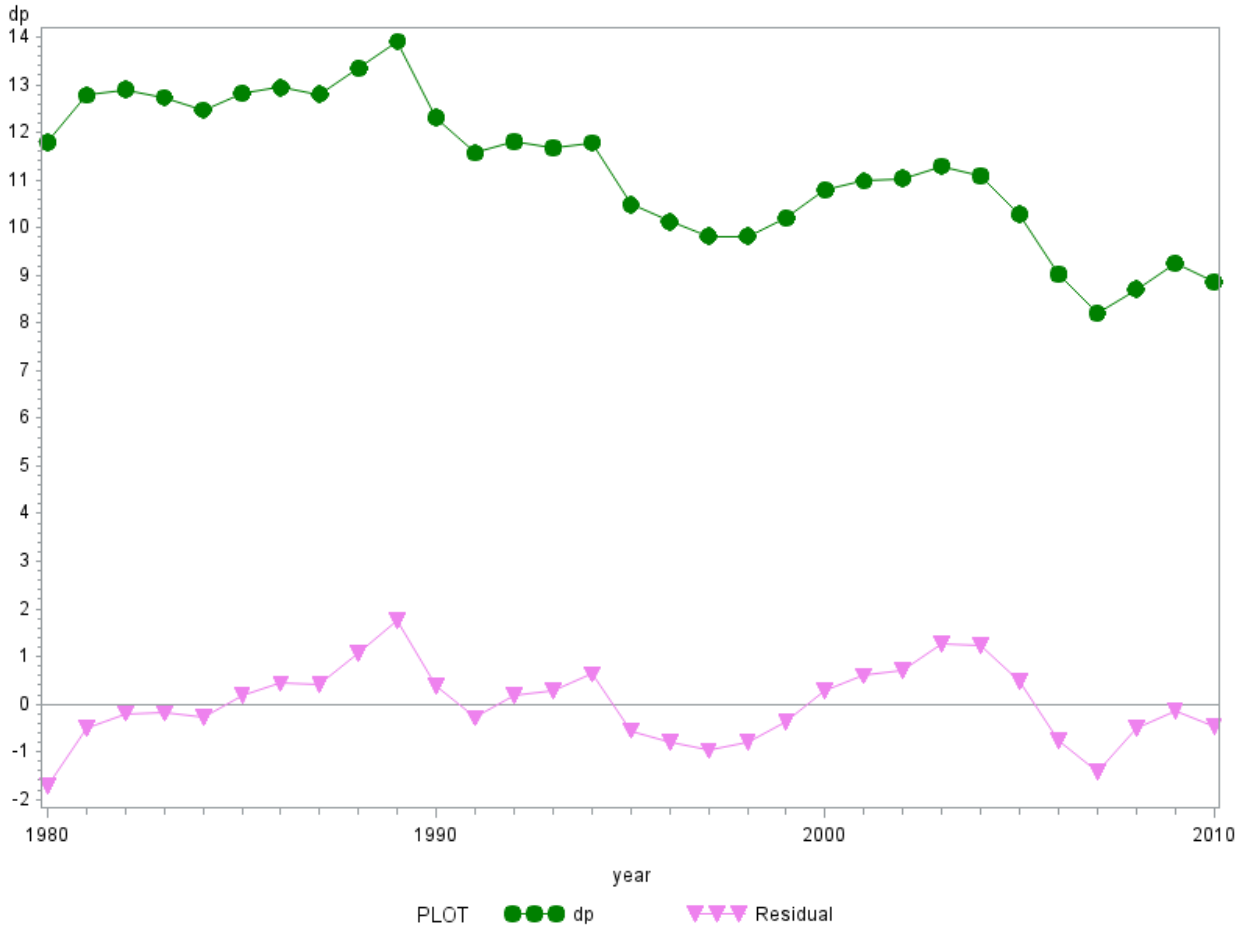
$$[8.73, 9.81] \quad \& \quad [7.52, 11.03]$$

4. We just show the results for the model based on males. The panel with the Fit Diagnostic plots is shown below



It is seen that observations number 1 and number 10 has large residuals, but only number 1 also a large leverage. This causes a very large value of Cook's D. Evidently observation no 1 is the most influential!

In the plot below we have shown a time series plot of the residuals, and evidently they do not look random. There seem to be a large correlation between a residual and its time neighbors.



You should take this as a general caution against doing regression on time series, as our assumption of independence is often violated. We can test this as shown below. *It is mostly aimed at those of you who have taken courses on time series analysis.*

We add an analysis on investigating the possible serial correlation in the residuals. This is done b.m.o. the so-called **Durbin-Watson** test statistic. It is defined as

$$D = \frac{\sum_{i=2}^{n}(R_i - R_{i-1})^2}{\sum_{i=1}^{n} {R_i}^2}$$

If there is no autocorrelation between the residuals, D will be close to 2 since we then will have that $V(R_i - R_{i-1}) = 2\sigma^2$. Very large values correspond to a negative correlation between consecutive values of R_i and very small values to a positive correlation. The p-values are computed in **proc reg** if we use the option **dwprob** in the model statement. SAS also provides an estimate of the 1.st order autocorrelation

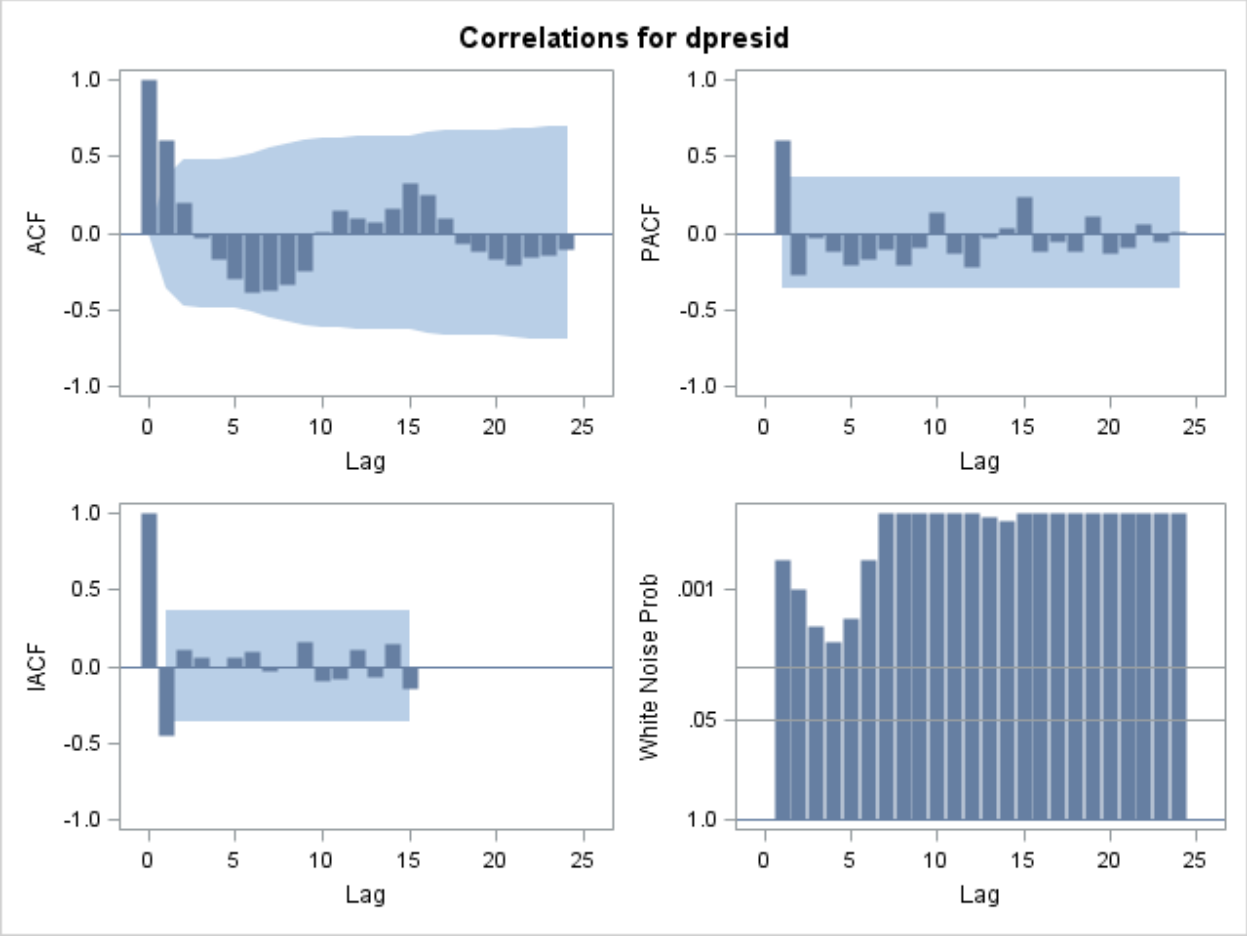
$$\hat{\varphi}_1 = \frac{\sum_{i=2}^{n} R_i R_{i-1}}{\sum_{i=1}^{n} {R_i}^2}$$

The values obtained for males were

Durbin-Watson D	0.620
Pr < DW	<.0001
Pr > DW	1.0000
Number of Observations	31
1st Order Autocorrelation	0.605

Note: Pr<DW is the p-value for testing positive autocorrelation, and Pr>DW is the p-value for testing negative autocorrelation.

We see that the value of D is considerably lower than 2 (p-value <.0001) and we thus conclude that the residuals are positively autocorrelated.

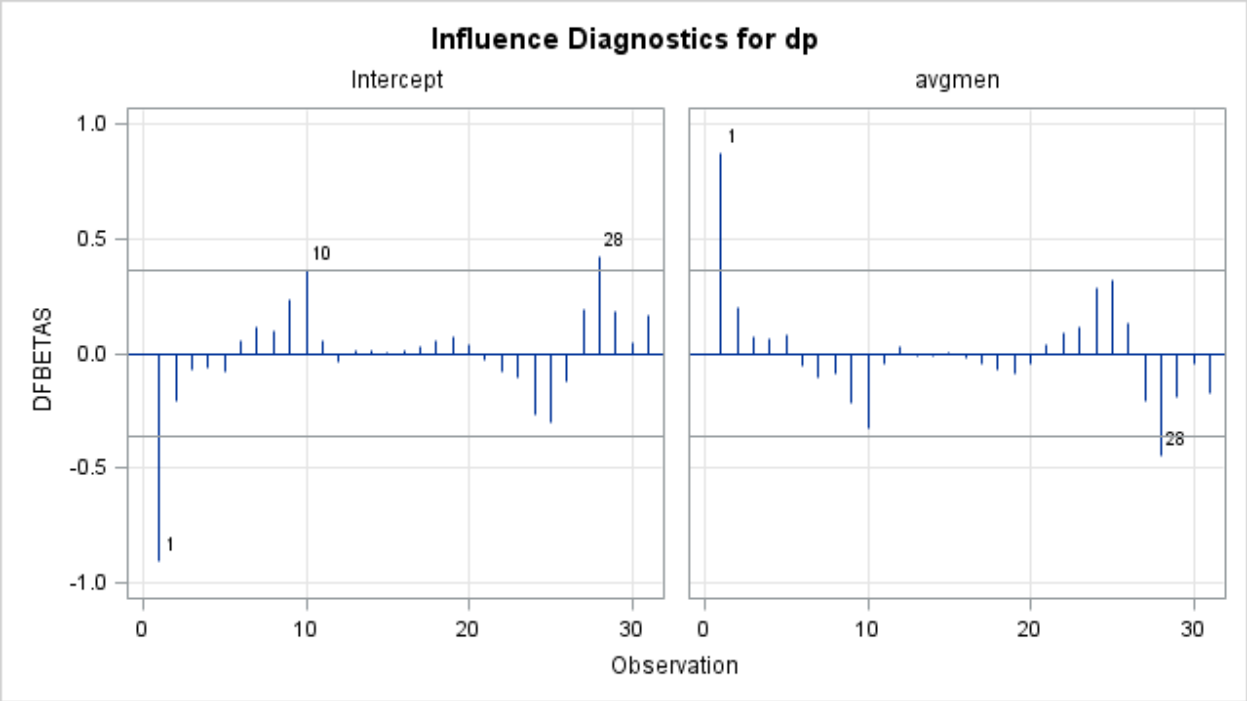


If we want to make a more sophisticated analysis, we may use a time series program

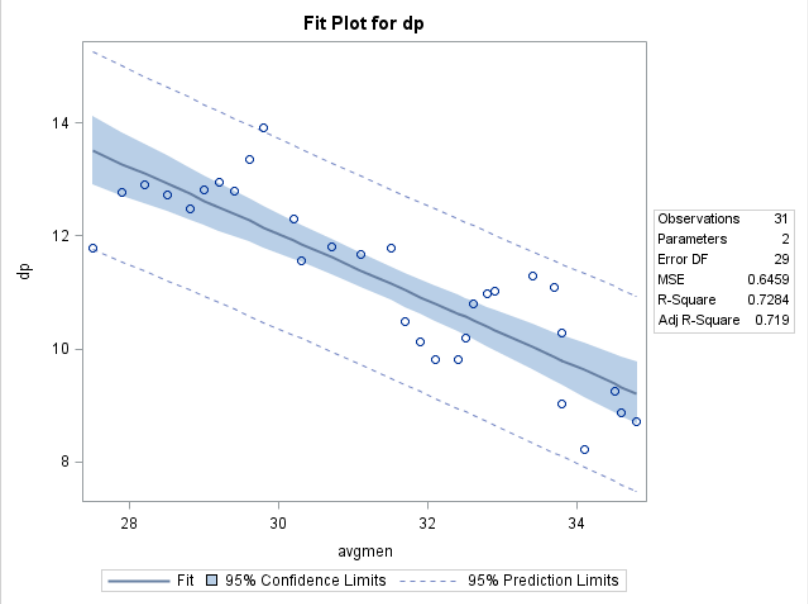
```
proc timeseries data=divout plots=(corr);
var dpresid;
run;
```

and obtain the plot given above. It indicates that also the second order autocorrelation is significantly different from 0, but we shall not go into further details in this direction.

5. The size adjusted cut-off for DFBETAS is $2/\sqrt{n}$ which in our case equals 0.36. From the list or from the plot given below we see that observations 1, 10 and 28 have 'extreme' values of DFBETAS, in good accordance with e.g. the results for Cook's D.



6.



A linear fit of the divorce rate as a function of the average marriage rate gives a reasonable reduction in the variance ($R^2 = 0.7284$). However, the divorce rate shows a clear cyclical behavior that is not captured by the simple linear (affine) model.

■