Time Series Fundamentals

OLS Regression and Time Series Data

Or “Don’t go all Durbin Watson on me”

What the heck – let's just run a regression on this time series data

1. Well, we've learned about two interesting time series processes – moving averages and autogression…but let's go back to regression for a minute and see why time series analysis is useful…
2. It might be tempting to use regression on this time series data – remember we did this in the last slide deck to sniff for a single mean, although I warned you there was a problem. Why not just use our old friend multivariate regression on time series?

What the heck – let's just run a regression on this time series data

1. What happens if the error terms are correlated?
   1. The estimates of the regression coefficients are still unbiased – that's good
   2. The estimates of the standard errors of the coefficients are biased – that's bad. In fact they are seriously biased downward.
2. So what does biased downward mean? It means that we are likely to reject a true null hypothesis – that is, we stand a serious risk of

What the heck – let's just run a regression on this time series data

1. So how do we tell if the error terms are correlated – that is, not independent?

2. One way is to look at the residual plot of the errors from the regression model

3. What is a residual plot? Remember that a regression model produces predictions of each data point in the data set. A residual is simply the difference (e.g. error) between the actual data point and the prediction of that data point by the regression model

What the heck – let’s just run a regression on this time series data

1. Here is some sales data along with some ad expenditures in thousands of dollars and also population trends for our sales territory.

2. Note that it is time series data – that is the data is collected for each year of sales

3. Let’s run a regression of

year	sales_units	ad_expend	population
1	3083	75	825000
2	3149	78	830445
3	3218	80	838750
4	3239	82	842940
5	3295	84	846315
6	3374	88	852240
7	3475	93	860760
8	3569	97	865925
9	3597	99	871640
10	3725	104	877745
11	3794	109	886520
12	3959	115	894500
13	4043	120	900400
14	4194	127	904005
15	4318	135	908525
16	4493	144	912160
17	4683	153	917630
18	4850	161	922220
19	5005	170	925910
20	5235	182	929610

What the heck – let's just run a regression on this time series data

1. Here is the regression from GRETL. Note the standard error for the ad expenditures is approximately .143 and that the relationship between ad expenditure and sales is significant p <

Model 4: OLS, using observations 1960-1979 (T = 20)
Dependent variable: sales_units

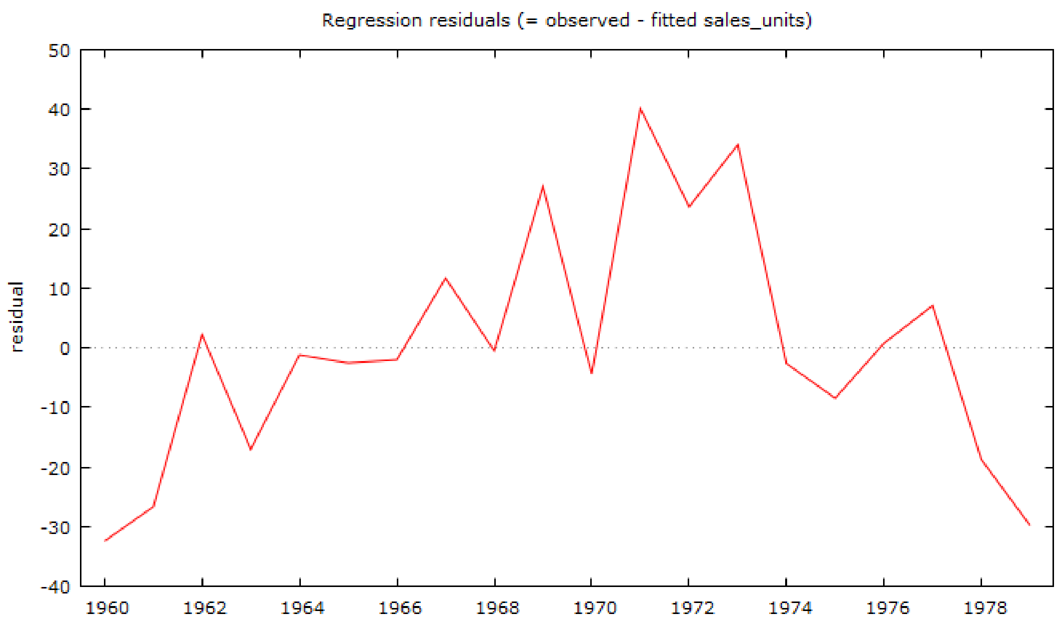
	Coefficient	*Std. Error*	*t-ratio*	*p-value*	
const	1608.83	17.0882	94.1485	<0.0001	***
ad_expend	20.0877	0.143336	140.1441	<0.0001	***

Mean dependent var	3914.900	S.D. dependent var	662.9769
Sum squared resid	7646.729	S.E. of regression	20.61112
R-squared	0.999084	Adjusted R-squared	0.999033
F(1, 18)	19640.36	P-value(F)	8.39e-29
Log-likelihood	−87.84178	Akaike criterion	179.6836
Schwarz criterion	181.6750	Hannan-Quinn	180.0723
rho	0.380272	Durbin-Watson	1.074299

What the heck – let's just run a regression on this time series data

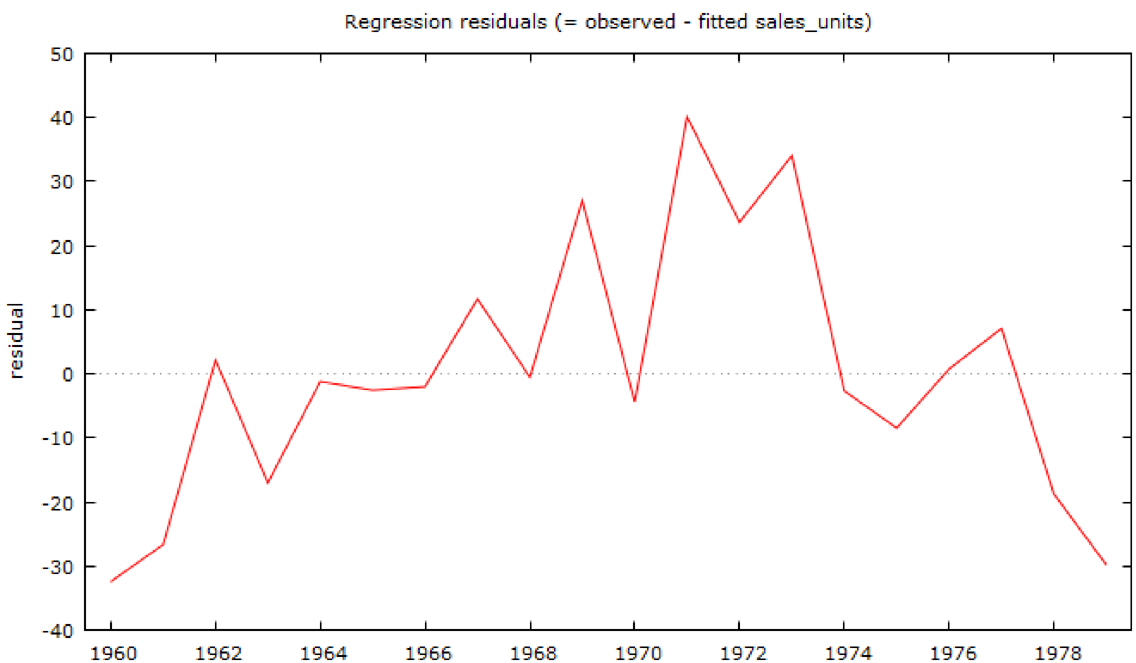
1. If we plot the residuals – that is the errors from the model – against time, what do we see?

2. If the errors are uncorrelated we should see them "flip flop" back and forth across the zero line as ?



What the heck – let's just run a regression on this time series data

1. Looks like we may have correlated error terms and as a result of that, a deflated standard error for the regression coefficient – the fact that ad expenditures was statistically significant and related to sales could be a Type I error



The Mysterious Durbin Watson Statistic

1. Is there some other way besides the Mark I eyeball test to tell if there is serial correlation in the residuals?

2. Yes there i W

$$d = \frac{\sum (\hat{e}_t - \hat{e}_{t-1})^2}{\sum \hat{e}_t^{\,2}}$$

Differences between consecutive errors

Durbin Watson statistic

Error term for each prediction

The Mysterious Durbin Watson Statistic

1. Let's look at the null and alternative hypotheses for the Durbin Watson test. The symbol ϕ stands for the serial correlation coefficient between errors (in this case consecutive errors). Remember correlation coefficients? If a correlation coefficient is zero then there is no correlation – that's good for us in this case – the error terms are not correlated.

2. If the error terms are correlated – they are likely positively correlated and so that would be bad – resulting in underestimated standard errors for our regression coefficient for ad expenditures

$$H_{null}: \ \phi = 0$$

$$H_{alt}: \ \phi > 0$$

The Mysterious Durbin Watson Statistic

1. What is the Durbin Watson statistic look like for our regression?

2. According to the analysis, our Durbin Watson statistic = 1.074. Is that bad mojo? How do I tell?

Model 4: OLS, using observations 1960-1979 (T = 20)
Dependent variable: sales_units

	Coefficient	*Std. Error*	*t-ratio*	*p-value*	
const	1608.83	17.0882	94.1485	<0.0001	***
ad_expend	20.0877	0.143336	140.1441	<0.0001	***

Mean dependent var	3914.900	S.D. dependent var	662.9769
Sum squared resid	7646.729	S.E. of regression	20.61112
R-squared	0.999084	Adjusted R-squared	0.999033
F(1, 18)	19640.36	P-value(F)	8.39e-29
Log-likelihood	−87.84178	Akaike criterion	179.6836
Schwarz criterion	181.6750	Hannan-Quinn	180.0723
rho	0.380272	Durbin-Watson	1.074299

The Mysterious Durbin Watson Statistic

1. You have to use a Durbin Watson table like this one from Savin and White

Table A-2

Models with an intercept (from Savin and White)

Durbin-Watson Statistic: 5 Per Cent Significance Points of dL and dU

	k'*=1		k'=2		k'=3		k'=4		k'=5		k'=6		k'=7		k'=8		k'=9		k'=10	
n	dL	dU	dL	dU	dL	dU	dL	dU	dL	dU	dL	dU	dL	dU	dL	dU	dL	dU	dL	dU
6	0.610	1.400	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----
7	0.700	1.356	0.467	1.896	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----
8	0.763	1.332	0.559	1.777	0.367	2.287	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----
9	0.824	1.320	0.629	1.699	0.455	2.128	0.296	2.588	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----
10	0.879	1.320	0.697	1.641	0.525	2.016	0.376	2.414	0.243	2.822	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----
11	0.927	1.324	0.758	1.604	0.595	1.928	0.444	2.283	0.315	2.645	0.203	3.004	-----	-----	-----	-----	-----	-----	-----	-----
12	0.971	1.331	0.812	1.579	0.658	1.864	0.512	2.177	0.380	2.506	0.268	2.832	0.171	3.149	-----	-----	-----	-----	-----	-----
13	1.010	1.340	0.861	1.562	0.715	1.816	0.574	2.094	0.444	2.390	0.328	2.692	0.230	2.985	0.147	3.266	-----	-----	-----	-----
14	1.045	1.350	0.905	1.551	0.767	1.779	0.632	2.030	0.505	2.296	0.389	2.572	0.286	2.848	0.200	3.111	0.127	3.360	-----	-----
15	1.077	1.361	0.946	1.543	0.814	1.750	0.685	1.977	0.562	2.220	0.447	2.471	0.343	2.727	0.251	2.979	0.175	3.216	0.111	3.438
16	1.106	1.371	0.982	1.539	0.857	1.728	0.734	1.935	0.615	2.157	0.502	2.388	0.398	2.624	0.304	2.860	0.222	3.090	0.155	3.304
17	1.133	1.381	1.015	1.536	0.897	1.710	0.779	1.900	0.664	2.104	0.554	2.318	0.451	2.537	0.356	2.757	0.272	2.975	0.198	3.184
18	1.158	1.391	1.046	1.535	0.933	1.696	0.820	1.872	0.710	2.060	0.603	2.258	0.502	2.461	0.407	2.668	0.321	2.873	0.244	3.073
19	1.180	1.401	1.074	1.536	0.967	1.685	0.859	1.848	0.752	2.023	0.649	2.206	0.549	2.396	0.456	2.589	0.369	2.783	0.290	2.974
20	1.201	1.411	1.100	1.537	0.998	1.676	0.894	1.828	0.792	1.991	0.691	2.162	0.595	2.339	0.502	2.521	0.416	2.704	0.336	2.885
21	1.221	1.420	1.125	1.538	1.026	1.669	0.927	1.812	0.829	1.964	0.731	2.124	0.637	2.290	0.546	2.461	0.461	2.633	0.380	2.806
22	1.239	1.429	1.147	1.541	1.053	1.664	0.958	1.797	0.863	1.940	0.769	2.090	0.677	2.246	0.588	2.407	0.504	2.571	0.424	2.735

The Mysterious Durbin Watson Statistic

1. The n along side the rows is the sample size. In our case we have 20 cases (20 years of data) so pick n=20.

2. k = number of IVs which in our case is one (ad expenditures)

Table A-2

Models with an intercept (from Savin and White)

Durbin-Watson Statistic: 5 Per Cent Significance Points of dL and dU

	k'*=1		k'=2		k'=3		k'=4		k'=5		k'=6		k'=7		k'=8		k'=9		k'=10	
n	dL	dU	dL	dU	dL	dU	dL	dU	dL	dU	dL	dU	dL	dU	dL	dU	dL	dU	dL	dU
6	0.610	1.400	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----
7	0.700	1.356	0.467	1.896	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----
8	0.763	1.332	0.559	1.777	0.367	2.287	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----
9	0.824	1.320	0.629	1.699	0.455	2.128	0.296	2.588	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----
10	0.879	1.320	0.697	1.641	0.525	2.016	0.376	2.414	0.243	2.822	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----
11	0.927	1.324	0.758	1.604	0.595	1.928	0.444	2.283	0.315	2.645	0.203	3.004	-----	-----	-----	-----	-----	-----	-----	-----
12	0.971	1.331	0.812	1.579	0.658	1.864	0.512	2.177	0.380	2.506	0.268	2.832	0.171	3.149	-----	-----	-----	-----	-----	-----
13	1.010	1.340	0.861	1.562	0.715	1.816	0.574	2.094	0.444	2.390	0.328	2.692	0.230	2.985	0.147	3.266	-----	-----	-----	-----
14	1.045	1.350	0.905	1.551	0.767	1.779	0.632	2.030	0.505	2.296	0.389	2.572	0.286	2.848	0.200	3.111	0.127	3.360	-----	-----
15	1.077	1.361	0.946	1.543	0.814	1.750	0.685	1.977	0.562	2.220	0.447	2.471	0.343	2.727	0.251	2.979	0.175	3.216	0.111	3.438
16	1.106	1.371	0.982	1.539	0.857	1.728	0.734	1.935	0.615	2.157	0.502	2.388	0.398	2.624	0.304	2.860	0.222	3.090	0.155	3.304
17	1.133	1.381	1.015	1.536	0.897	1.710	0.779	1.900	0.664	2.104	0.554	2.318	0.451	2.537	0.356	2.757	0.272	2.975	0.198	3.184
18	1.158	1.391	1.046	1.535	0.933	1.696	0.820	1.872	0.710	2.060	0.603	2.258	0.502	2.461	0.407	2.668	0.321	2.873	0.244	3.073
19	1.180	1.401	1.074	1.536	0.967	1.685	0.859	1.848	0.752	2.023	0.649	2.206	0.549	2.396	0.456	2.589	0.369	2.783	0.290	2.974
20	1.201	1.411	1.100	1.537	0.998	1.676	0.894	1.828	0.792	1.991	0.691	2.162	0.595	2.339	0.502	2.521	0.416	2.704	0.336	2.885
21	1.221	1.420	1.125	1.538	1.026	1.669	0.927	1.812	0.829	1.964	0.731	2.124	0.637	2.290	0.546	2.461	0.461	2.633	0.380	2.806
22	1.239	1.429	1.147	1.541	1.053	1.664	0.958	1.797	0.863	1.940	0.769	2.090	0.677	2.246	0.588	2.407	0.504	2.571	0.424	2.735

The Mysterious Durbin Watson Statistic

1. Let's look closer at this. Notice that there is a column marked dL (Durbin lower limit) and another column marked dU (Durbin upper limit).

2. In this case dL = 1.201 and dU = 1.411

	k'*=1		k'=2	
n	dL	dU	dL	dU
6	0.610	1.400	-----	-----
7	0.700	1.356	0.467	1.896
8	0.763	1.332	0.559	1.777
9	0.824	1.320	0.629	1.699
10	0.879	1.320	0.697	1.641
11	0.927	1.324	0.758	1.604
12	0.971	1.331	0.812	1.579
13	1.010	1.340	0.861	1.562
14	1.045	1.350	0.905	1.551
15	1.077	1.361	0.946	1.543
16	1.106	1.371	0.982	1.539
17	1.133	1.381	1.015	1.536
18	1.158	1.391	1.046	1.535
19	1.180	1.401	1.074	1.536
20	1.201	1.411	1.100	1.537
21	1.221	1.420	1.125	1.538
22	1.239	1.429	1.147	1.541
23	1.257	1.437	1.168	1.543

The Mysterious Durbin Watson Statistic

1. Here is the hypothesis testing rule for Durbin Watson

2. If $d < dL$ then reject H_{null}: $\phi = 0$

   If $d > dU$ do not reject H_{null}: $\phi = 0$

If $dL \leq d \leq dU$ then the test is inconclusive

	k'*=1		k'=2	
n	dL	dU	dL	dU
6	0.610	1.400	-----	-----
7	0.700	1.356	0.467	1.896
8	0.763	1.332	0.559	1.777
9	0.824	1.320	0.629	1.699
10	0.879	1.320	0.697	1.641
11	0.927	1.324	0.758	1.604
12	0.971	1.331	0.812	1.579
13	1.010	1.340	0.861	1.562
14	1.045	1.350	0.905	1.551
15	1.077	1.361	0.946	1.543
16	1.106	1.371	0.982	1.539
17	1.133	1.381	1.015	1.536
18	1.158	1.391	1.046	1.535
19	1.180	1.401	1.074	1.536
20	1.201	1.411	1.100	1.537
21	1.221	1.420	1.125	1.538
22	1.239	1.429	1.147	1.541
23	1.257	1.437	1.168	1.543

The Mysterious Durbin Watson Statistic

1. Here is our regression analysis. Notice that our Durbin Watson from the regression is = 1.074

2. Remember from the chart that dL = 1.201 and dU = 1.411

Model 4: OLS, using observations 1960-1979 (T = 20)
Dependent variable: sales_units

	Coefficient	*Std. Error*	*t-ratio*	*p-value*	
const	1608.83	17.0882	94.1485	<0.0001	***
ad_expend	20.0877	0.143336	140.1441	<0.0001	***

Mean dependent var	3914.900	S.D. dependent var	662.9769
Sum squared resid	7646.729	S.E. of regression	20.61112
R-squared	0.999084	Adjusted R-squared	0.999033
F(1, 18)	19640.36	P-value(F)	8.39e-29
Log-likelihood	−87.84178	Akaike criterion	179.6836
Schwarz criterion	181.6750	Hannan-Quinn	180.0723
rho	0.380272	Durbin-Watson	1.074299

The Mysterious Durbin Watson Statistic

1. If d from the data (d=1.074) is < dL then reject H_{null}: $\phi = 0$

$$1.074 < 1.201$$

$$d < dL$$

so reject H_{null}

Model 4: OLS, using observations 1960-1979 (T = 20)
Dependent variable: sales_units

	Coefficient	Std. Error	t-ratio	p-value	
const	1608.83	17.0882	94.1485	<0.0001	***
ad_expend	20.0877	0.143336	140.1441	<0.0001	***

Mean dependent var	3914.900	S.D. dependent var	662.9769
Sum squared resid	7646.729	S.E. of regression	20.61112
R-squared	0.999084	Adjusted R-squared	0.999033
F(1, 18)	19640.36	P-value(F)	8.39e-29
Log-likelihood	−87.84178	Akaike criterion	179.6836
Schwarz criterion	181.6750	Hannan-Quinn	180.0723
rho	0.380272	Durbin-Watson	1.074299

The Mysterious Durbin Watson Statistic

1. Oh, that's bad – we have serially correlated error terms and so we violated a major regression assumption. Our significance test for the regression coefficient of ad expenditures is mess up – the standard error is wrong and too small. Should we just quit and go to Hills and Dales?

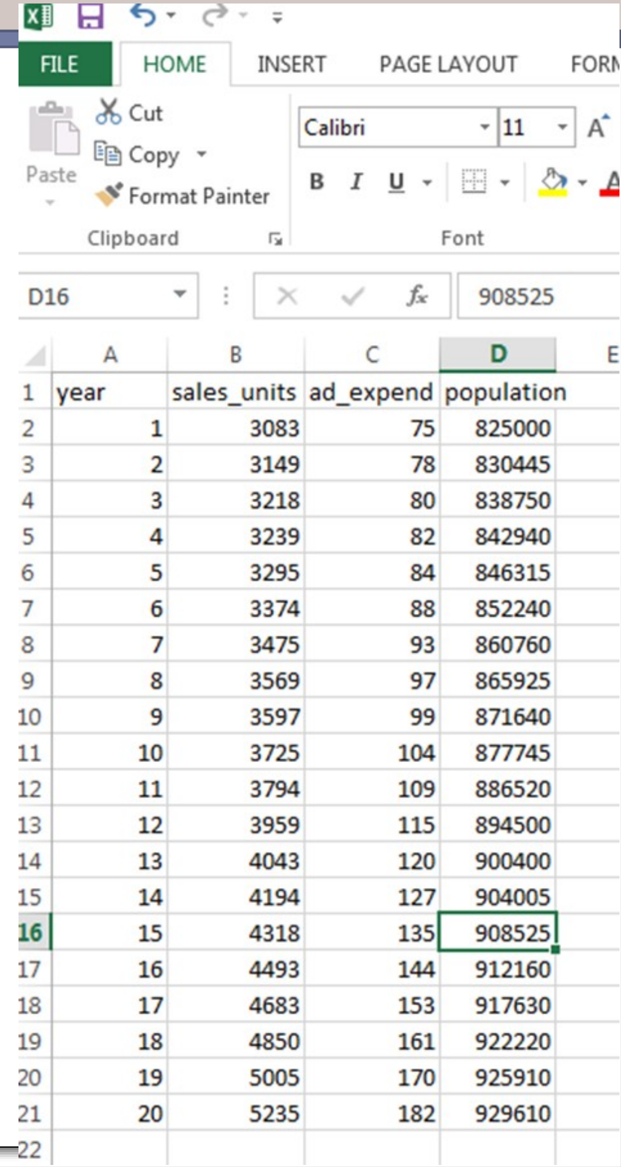
2. No we can fix this. One way to fix this is to include another important variable in the equation that we might have forgotten to include. That independent variable is time dependent.

The Mysterious Durbin Watson Statistic

1. Let's look at our data again…

2. We also have the variable that shows population in the sales region. That's likely to be time dependent and also an important variable that is likely to influence sales. Let's jam that variable into the model as well so our new model looks like:

$$Sales = \alpha + b_1 \text{ ad\_expend} + b_2 \text{ pop}$$



	A	B	C	D
1	year	sales_units	ad_expend	population
2	1	3083	75	825000
3	2	3149	78	830445
4	3	3218	80	838750
5	4	3239	82	842940
6	5	3295	84	846315
7	6	3374	88	852240
8	7	3475	93	860760
9	8	3569	97	865925
10	9	3597	99	871640
11	10	3725	104	877745
12	11	3794	109	886520
13	12	3959	115	894500
14	13	4043	120	900400
15	14	4194	127	904005
16	15	4318	135	908525
17	16	4493	144	912160
18	17	4683	153	917630
19	18	4850	161	922220
20	19	5005	170	925910
21	20	5235	182	929610
22				

The Mysterious Durbin Watson Statistic

1. Here is our new regression analysis from GRETL. Let's try our hypothesis test on the Durbin Watson again

2. If d > dU do not reject H_{null}: $\phi = 0$ (remember k=2 now)

3.056 > 1.527 (with k=2) so cannot reject H_{null}

d > dU

Model 5: OLS, using observations 1960-1979 (T = 20)
Dependent variable: sales_units

	Coefficient	Std. Error	t-ratio	p-value	
const	313.321	217.344	1.4416	0.1676	
ad_expend	18.4215	0.291553	63.1841	<0.0001	***
population	0.00168826	0.000282935	5.9670	<0.0001	***

Mean dependent var	3914.900	S.D. dependent var	662.9769
Sum squared resid	2471.155	S.E. of regression	12.05662
R-squared	0.999704	Adjusted R-squared	0.999669
F(2, 17)	28717.12	P-value(F)	1.01e-30
Log-likelihood	−76.54586	Akaike criterion	159.0917
Schwarz criterion	162.0789	Hannan-Quinn	159.6748
rho	−0.533077	Durbin-Watson	3.056839

The Mysterious Durbin Watson Statistic

1. So adding the variable population removed the serial correlation in our error term. So our standard error for the regression coefficient should not be biased downward.

2. That means the standard error for ad_expend here (.291) should be bigger than the original model (.1433) – is it? Yep…

Model with added time dependent variable

Model 5: OLS, using observations 1960-1979 (T = 20)
Dependent variable: sales_units

	Coefficient	*Std. Error*	*t-ratio*	*p-value*	
const	313.321	217.344	1.4416	0.1676	
ad_expend	18.4215	0.291553	63.1841	<0.0001	***
population	0.00168826	0.000282935	5.9670	<0.0001	***

Mean dependent var	3914.900	S.D. dependent var	662.9769
Sum squared resid	2471.155	S.E. of regression	12.05662
R-squared	0.999704	Adjusted R-squared	0.999669
F(2, 17)	28717.12	P-value(F)	1.01e-30
Log-likelihood	−76.54586	Akaike criterion	159.0917
Schwarz criterion	162.0789	Hannan-Quinn	159.6748
rho	−0.533077	Durbin-Watson	3.056839

Original Model

Model 4: OLS, using observations 1960-1979 (T = 20)
Dependent variable: sales_units

	Coefficient	*Std. Error*	*t-ratio*	*p-value*	
const	1608.83	17.0882	94.1485	<0.0001	***
ad_expend	20.0877	0.143336	140.1441	<0.0001	***

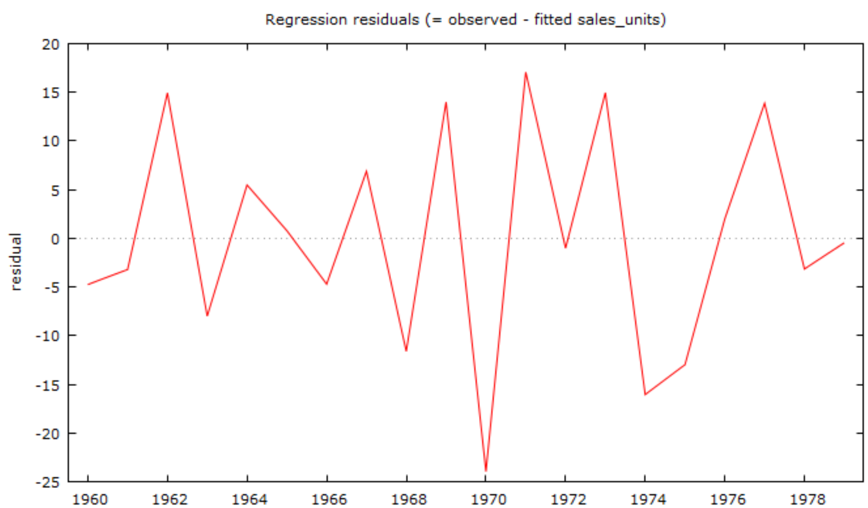
Mean dependent var	3914.900	S.D. dependent var	662.9769
Sum squared resid	7646.729	S.E. of regression	20.61112
R-squared	0.999084	Adjusted R-squared	0.999033
F(1, 18)	19640.36	P-value(F)	8.39e-29
Log-likelihood	−87.84178	Akaike criterion	179.6836
Schwarz criterion	181.6750	Hannan-Quinn	180.0723
rho	0.380272	Durbin-Watson	1.074299

The Mysterious Durbin Watson Statistic

1. That also means that our residual chart for the new model with the time dependent variable of population should look a lot prettier and should “flip flop”– let’s see, does it?

2. Maybe the Durbin Watson statistic is not so mysterious after all?

Model with added time dependent variable



Original Model