Chapter 17

Multiple Regression

Nonindependence of the Error Variable

If we were to observe the auction price of cars every week for, say, a year, that would constitute ***a time series***.

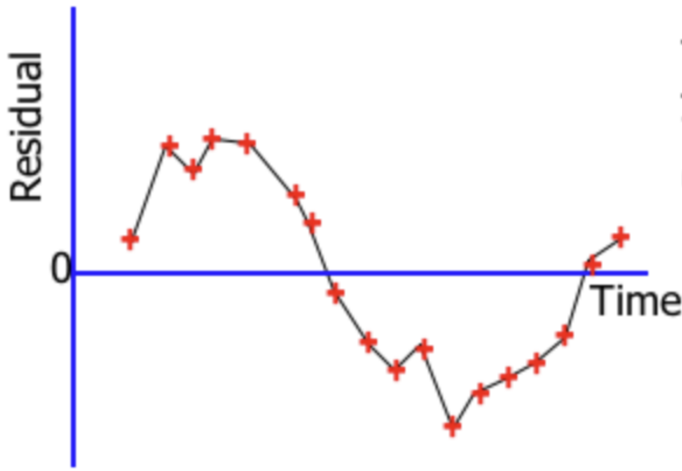
When the data are time series, the errors often are ***correlated***.

Error terms that are correlated over time are said to be ***autocorrelated*** or ***serially correlated***.

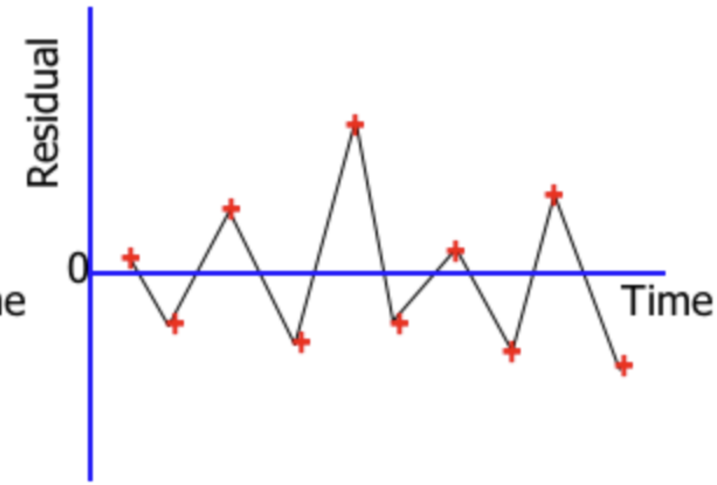
We can often detect autocorrelation by ***graphing the residuals against the time periods***. If a pattern emerges, it is likely that the independence requirement is violated.

Nonindependence of the Error Variable

Patterns in the appearance of the residuals over time indicates that autocorrelation exists:



Note the runs of positive residuals, replaced by runs of negative residuals



Note the oscillating behavior of the residuals around zero.

Negative autocorreleted

Regression Diagnostics – Time Series

- The ***Durbin-Watson test*** allows us to determine whether there is evidence of ***first-order autocorrelation*** — a condition in which a relationship exists between ***consecutive residuals***, i.e. e_{i-1} and e_i (i is the time period). The statistic for this test is defined as:

$$\boxed{d = \frac{\sum_{i=2}^{n}(e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2}}$$

- d has a range of values: $0 \le d \le 4$.

Durbin–Watson

0 d 4

Small values of **d** (**$d < 2$**) indicate a ***positive*** first-order autocorrelation.

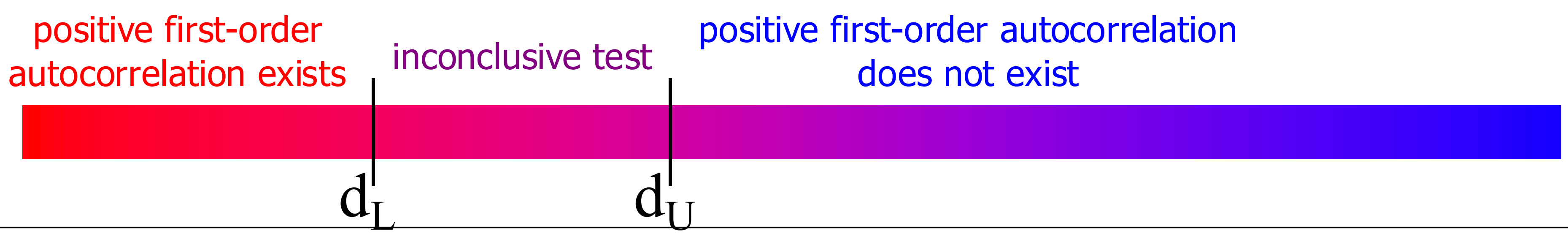
Large values of **d** (**$d > 2$**) imply a ***negative*** first-order autocorrelation.

Durbin–Watson (one-tail test)

- To test for ***positive first-order autocorrelation***:

- If $d < d_L$, we conclude that there is enough evidence to show that positive first-order autocorrelation exists.
- If $d > d_U$, we conclude that there is **not** enough evidence to show that positive first-order autocorrelation exists.
- And if $d_L \leq d \leq d_U$, the test is inconclusive.

d_L, d_U from table 11, appendix B

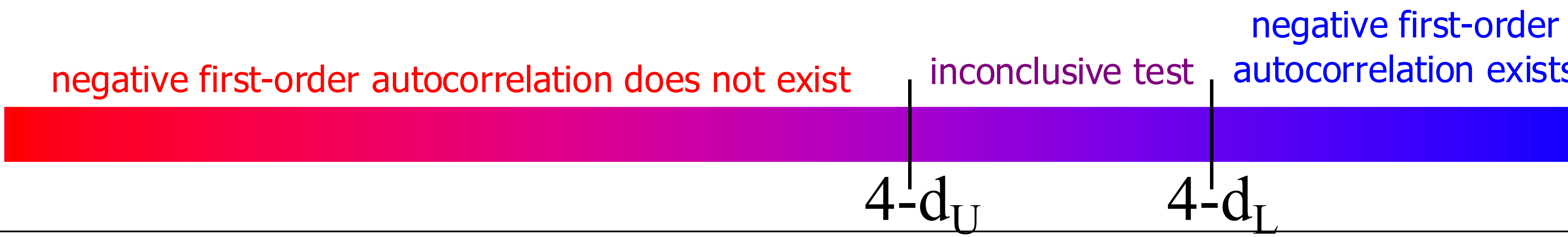


Durbin–Watson (one-tail test)

- To test for ***negative first-order autocorrelation***:

- If $d > 4 - d_L$, we conclude that there is enough evidence to show that negative first-order autocorrelation exists.
- If $d < 4 - d_U$, we conclude that there is not enough evidence to show that negative first-order autocorrelation exists.
- And if $4 - d_U \leq d \leq 4 - d_L$, the test is inconclusive.

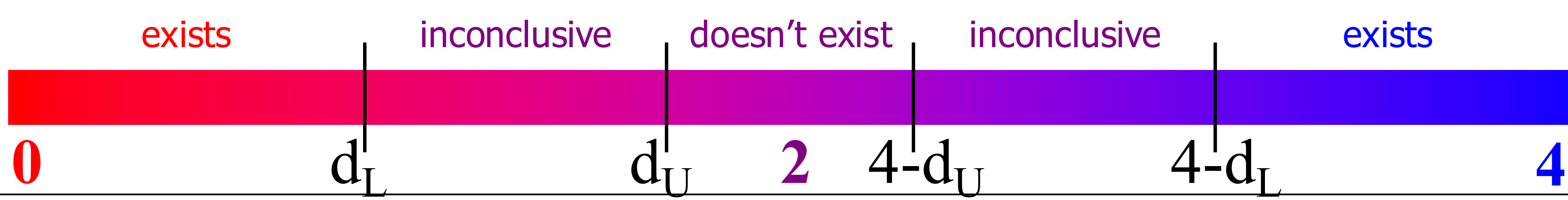
d_L, d_U from table 11, appendix B



Durbin–Watson (two-tail test)

- To test for ***first-order autocorrelation***:

- If $d < d_L$ or $d > 4 - d_L$, first-order autocorrelation **exists**.
- If d falls between d_L and d_U ***or*** between $4 - d_U$ and $4 - d_U$, the test is inconclusive.
- If d falls between d_U and $4 - d_U$ there is no evidence of first order autocorrelation.



n: sample size

Critical Values for the Durbin-Watson Statistic, $\alpha=0.5$

k is the number of independent variables

	k=1		k=2		k=3		k=4		k=5	
n	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U
15	1.08	1.36	.95	1.54	.82	1.75	.69	1.97	.56	2.21
16	1.10	1.37	.98	1.54	.86	1.73	.74	1.93	.62	2.15
17	1.13	1.38	1.02	1.54	.90	1.71	.78	1.90	.67	2.10
18	1.16	1.39	1.05	1.53	.93	1.69	.82	1.87	.71	2.06
19	1.18	1.40	1.08	1.53	.97	1.68	.86	1.85	.75	2.02
20	1.20	1.41	1.10	1.54	1.00	1.68	.90	1.83	.79	1.99
21	1.22	1.42	1.13	1.54	1.03	1.67	.93	1.81	.83	1.96
22	1.24	1.43	1.15	1.54	1.05	1.66	.96	1.80	.86	1.94
23	1.26	1.44	1.17	1.54	1.08	1.66	.99	1.79	.90	1.92
24	1.27	1.45	1.19	1.55	1.10	1.66	1.01	1.78	.93	1.90
25	1.29	1.45	1.21	1.55	1.12	1.66	1.04	1.77	.95	1.89
26	1.30	1.46	1.22	1.55	1.14	1.65	1.06	1.76	.98	1.88
27	1.32	1.47	1.24	1.56	1.16	1.65	1.08	1.76	1.01	1.86
28	1.33	1.48	1.26	1.56	1.18	1.65	1.10	1.75	1.03	1.85
29	1.34	1.48	1.27	1.56	1.20	1.65	1.12	1.74	1.05	1.84
30	1.35	1.49	1.28	1.57	1.21	1.65	1.14	1.74	1.07	1.83

Example 17.3

Can we create a model that will predict lift ticket sales at a ski hill based on two weather parameters?

Variables:

y - lift ticket sales during Christmas week,

x_1 - total snowfall (inches), and

x_2 - average temperature (degrees Fahrenheit)

Our ski hill manager collected 20 years of data. Xm17-03

Example 17.3

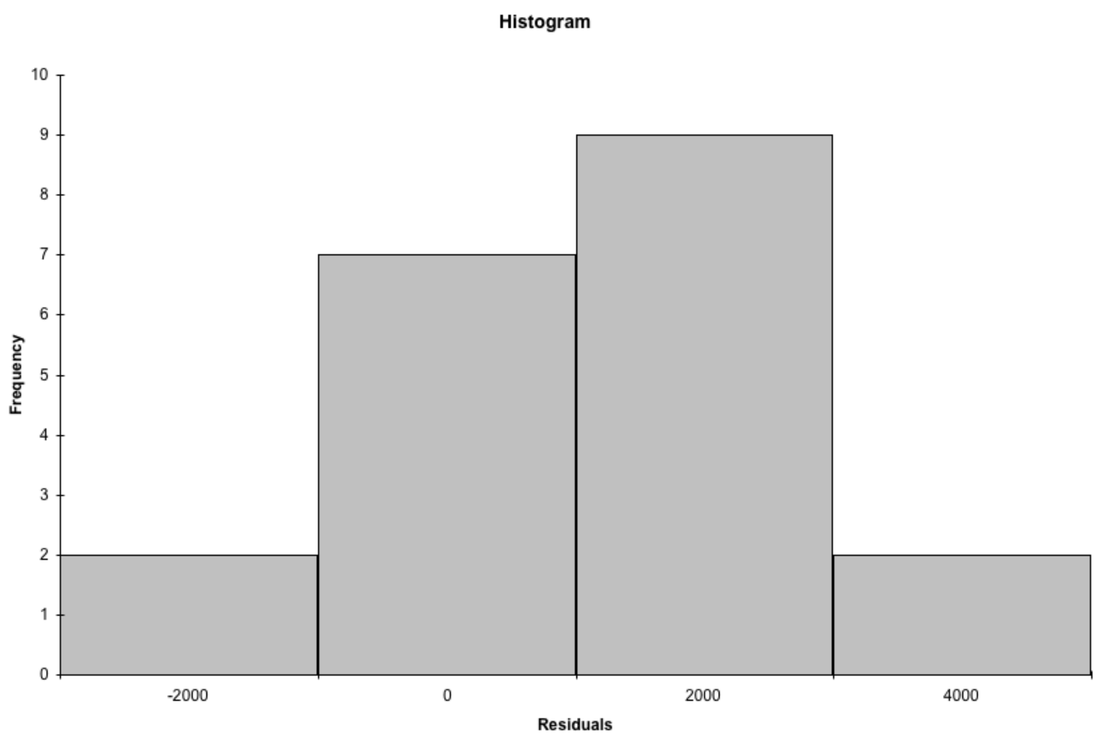
	A	B	C	D	E	F
1	SUMMARY OUTPUT					
2						
3	*Regression Statistics*					
4	Multiple R	0.3465				
5	R Square	0.1200				
6	Adjusted R Square	0.0165				
7	Standard Error	1712				
8	Observations	20				
9						
10	ANOVA					
11		*df*	*SS*	*MS*	*F*	*Significance F*
12	Regression	2	6793798.248	3396899	1.16	0.3373
13	Residual	17	49807213.95	2929836		
14	Total	19	56601012.2			
15						
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	
17	Intercept	8308.0	903.7	9.19	0.0000	
18	Snowfall	74.59	51.57	1.45	0.1663	
19	Temperature	-8.75	19.70	-0.44	0.6625	

Both the coefficient of determination and the p-value of the F-test indicate the model is poor…

Neither variable is linearly related to ticket sale…

Example 17.3

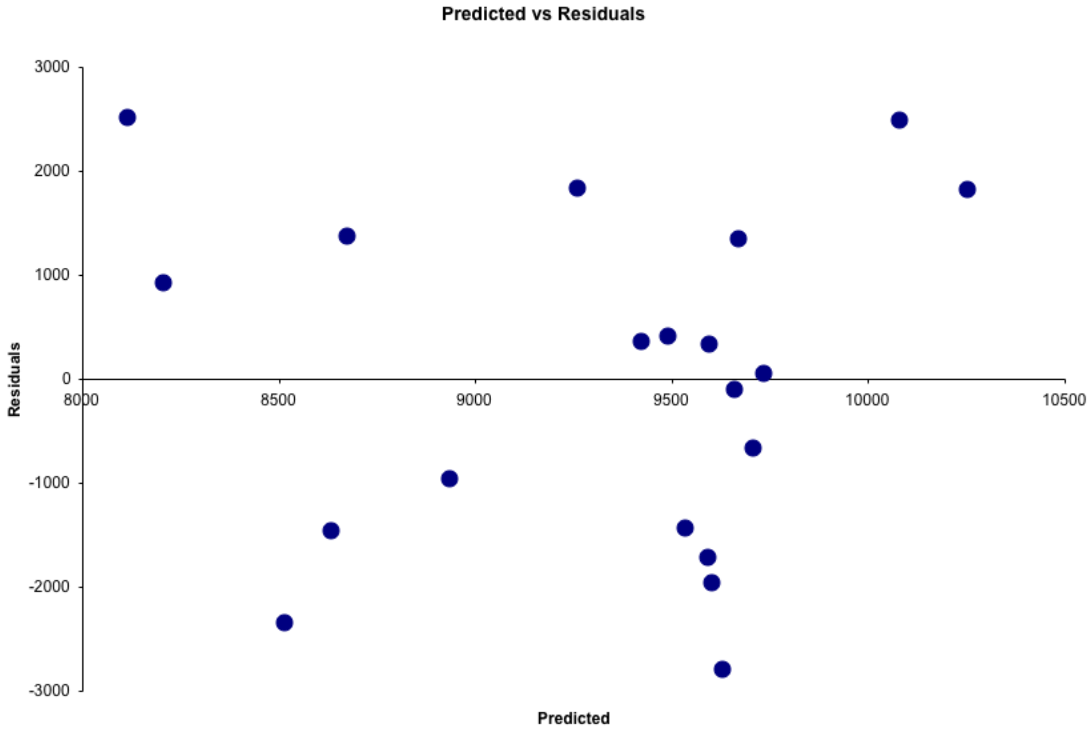
- The histogram of residuals…



- reveals the errors may be normally distributed…

Example 17.3

- In the plot of residuals versus predicted values (testing for heteroscedasticity) — the error variance appears to be constant…



Example 17.3 Durbin-Watson

- Apply the Durbin-Watson Statistic from Data Analysis Plus to the entire ***list of residuals***.

	A	B	C	D
22				
23	RESIDUAL OUTPUT			
24				
25	*Observation*	*Predicted Tickets*	*Residuals*	*Standard Residuals*
26	1	9629	-2793.99	-1.726
27	2	9593	-1723.23	-1.064
28	3	8515	-2342.03	-1.447
29	4	8936	-956.95	-0.591
30	[illegible]	9603	-1963.73	-1.213
31	[illegible]	8632	-1465.27	-0.905
32	[illegible]	9533	-1439.07	-0.889
33	8	9489	414.07	0.256
34	9	9423	364.91	0.225
35	10	9660	-102.82	-0.064
36	11	9734	49.96	0.031
37	12	10252	1823.37	1.126
38	13	8208	920.10	0.568
39	14	9707	-660.40	-0.408
40	15	8115	2515.57	1.554
41	16	10081	2482.26	1.533
42	17	9669	1343.06	0.830
43	18	8673	1368.40	0.845
44	19	9594	334.65	0.207
45	20	9260	1831.16	1.131
46				

Sheet2 DURBIN4 Sheet3 Sheet1

Ready Sum=0

	A	B
1	**Durbin-Watson Statistic**	
2		
3	d = 0.5931	
4		

Sheet2

Ready

Example 17.3

To test for positive first-order autocorrelation with $\alpha = .05$, we find in Table 8(a) in Appendix B

$d_L = 1.10$ and $d_U = 1.54$

The null and alternative hypotheses are

H_0 : There is no first-order autocorrelation.

H_1 : There is positive first-order autocorrelation.

The rejection region is $d < d_L = 1.10$. Since $d = .59$, we reject the null hypothesis and conclude that there is enough evidence to infer that positive first-order autocorrelation exists.

Example 17.3

Autocorrelation usually indicates that the model needs to include an independent variable that has a time-ordered effect on the dependent variable.

The simplest such independent variable represents the time periods. We included a third independent variable that records the number of years since the year the data were gathered. Thus, $x_3 = 1, 2, \ldots, 20$. The new model is

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$$

Example 17.3

	A	B	C	D	E	F
1	SUMMARY OUTPUT					
2						
3	*Regression Statistics*					
4	Multiple R	0.8608				
5	R Square	0.7410				
6	Adjusted R Square	0.6924				
7	Standard Error	957				
8	Observations	20				
9						
10	ANOVA					
11		*df*	*SS*	*MS*	*F*	*Significance F*
12	Regression	3	41940217	13980072	15.26	0.0001
13	Residual	16	14660795	916300		
14	Total	19	56601012			
15						
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	
17	Intercept	5965.6	631.3	9.45	0.0000	
18	Snowfall	70.18	28.85	2.43	0.0271	
19	Temperature	-9.23	11.02	-0.84	0.4145	
20	Time	229.97	37.13	6.19	0.0000	

The fit of the model is high,

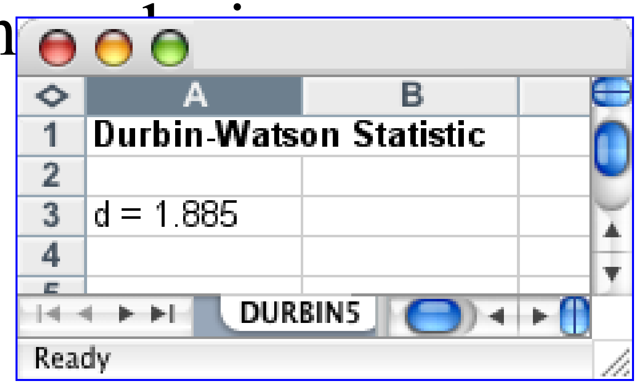
The model is valid…

our new variable

Snowfall and time are linearly related to ticket sales; temperature is not…

Example 17.3

- If we re-run the Durbin-Watson statistic against the residuals from our Regression



- we can conclude that there is not enough evidence to infer the presence of first-order autocorrelation. *(Determining d_L and d_U is left as an exercise for the reader…)*

- Hence, we have improved out model dramatically!

Example 17.3

Notice that the model is improved dramatically.

The F-test tells us that the model is valid. The t-tests tell us that both the amount of snowfall and time are significantly linearly related to the number of lift tickets.

This information could prove useful in advertising for the resort. For example, if there has been a recent snowfall, the resort could emphasize that in its advertising.

If no new snow has fallen, it may emphasize their snow-making facilities.