# LECTURE 4E – TIME SERIES REGRESSION

Lecture 4E-1

# Regression Assumptions - Revisited

- Population Model:
  - $Y = \boldsymbol{\alpha} + \boldsymbol{\beta_1} X_1 + \boldsymbol{\beta_2} X_2 + \boldsymbol{\beta_3} X_3 + \boldsymbol{\epsilon}$ with $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \boldsymbol{\sigma}_\varepsilon^2)$.
- (Sample) Regression Model or Prediction Model
  - $\hat{y} = \hat{\alpha} + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3$ with residual (also called error or noise) terms $\boldsymbol{e} = (y - \hat{y})$
- In order for our conclusions about the regression model (that we fit) to be valid, we need to check four main assumptions:
  - *Linearity/Nonlinearity between predictors and dependent variable*
  - *Normality of Residuals*
  - *Homoscedasticity* (constant variance of residuals across X) or the opposite, *heteroscedasticity*
  - *Statistical Independence of residuals* (relevant in time-series data)
- In addition, other problems (have to be considered and fixed) such as:
  - *Multicollinearity* of predictors (excessive correlations among predictors)
  - *Missing Data*, especially in large secondary datasets in analytics)
- Failure to address these problems could result in invalid conclusions from the regression analysis.

http://people.duke.edu/~rnau/testing.htm

# Autocorrelation

- When the error terms are dependent on each other (as often happens with time series data where data is collected across time), on measure of the dependence is *autocorrelation*.
- **Autocorrelation** (also called **serial correlation**) is a linear relationship among residuals of the model across time.
- We will use a simple model examining the relationship between Microsoft's marketing and advertising expenditures and its revenues to illustrate autocorrelation.

| | Obs | Year | Quarter | Revenues | Marketing | Summer | Fall | Winter |
|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 1987 | 1 | 60.02 | 13.44 | 0 | 0 | 1 |
| 2 | 2 | 1987 | 2 | 71.62 | 16.80 | 0 | 0 | 0 |
| 3 | 3 | 1987 | 3 | 85.66 | 18.36 | 1 | 0 | 0 |
| 4 | 4 | 1987 | 4 | 86.68 | 22.54 | 0 | 1 | 0 |
| 5 | 5 | 1988 | 1 | 88.74 | 23.26 | 0 | 0 | 1 |
| 6 | 6 | 1988 | 2 | 132.73 | 31.48 | 0 | 0 | 0 |

**REVENUES** Microsoft's real quarterly revenues, in millions of dollars1
**MARKETING** Microsoft's real quarterly expenditures on marketing and
advertising, in millions of dollars
**SUMMER** =1 when REVENUES and MARKETING are from the third quarter (July-September), 0 otherwise
**FALL** = 1 when REVENUES and MARKETING are from the fourth quarter
(Oct.-Dec.), 0 otherwise

# Regression Model (AutoCorr.R)

- $\widehat{\text{Revenues}} = \hat{\alpha} + \hat{\beta}_{mkt}\text{Marketing}$
- We run this model in **R**
- The model appears to be an excellent fit, but always have to check for autocorrelations in data that are a series of observation over a time period.
- In our data set, Obs can be used as a time variable because the Obs 1 represents Time period 1, Obs2 represents Time Period 2 etc. on a quarterly basis. That is, Obs 1 is the first quarter of 1987 and Obs 2 the second quarter of 1987 and so on.
- To check for autocorrelations visually, we can plot the residuals of from the model against time.

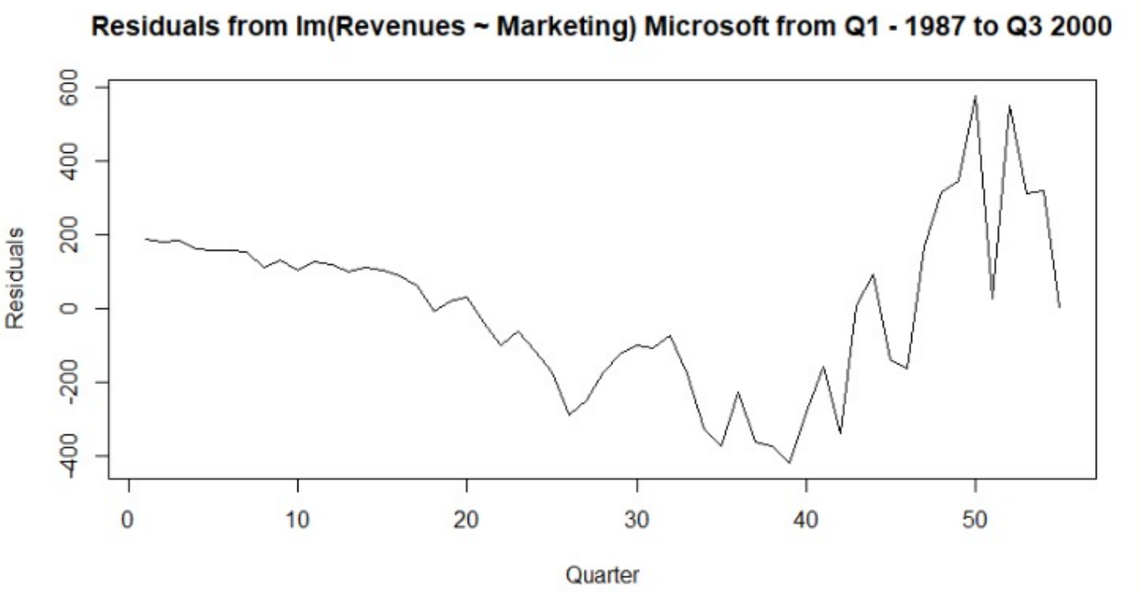
```
> mod1 <- lm(Revenues ~ Marketing, data=df)
> summary(mod1)

Call:
lm(formula = Revenues ~ Marketing, data = df)

Residuals:
    Min      1Q  Median      3Q     Max 
-421.33 -160.95   17.77  141.17  577.79 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -203.6130    49.8504  -4.084  0.00015 ***
Marketing      5.7596     0.1754  32.839  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 227 on 53 degrees of freedom
Multiple R-squared:  0.9532,	Adjusted R-squared:  0.9523 
F-statistic:  1078 on 1 and 53 DF,  p-value: < 2.2e-16
```

```
> # Read csv file as a DataFrame
> #
> setwd("C:\\Users\\sarathy\\Documents\\2019-Teaching\\Fall2019\\Fall2019-MSIS5503\\MSIS-5503-Data")
> df <- read.table('Microsoft.csv',
+                  header = TRUE, sep = ',')
> print(head(df))
  Obs Year Quarter Revenues Marketing Summer Fall Winter
1   1 1987       1    60.02     13.44      0    0      1
2   2 1987       2    71.62     16.80      0    0      0
3   3 1987       3    85.66     18.36      1    0      0
4   4 1987       4    86.68     22.54      0    1      0
5   5 1988       1    88.74     23.26      0    0      1
6   6 1988       2   132.73     31.48      0    0      0
```

# Residual Plot to check for independence of Residuals

- We will plot the residuals against their position in time.
- If the residuals (from one observation/period) to the next were independent, and therefore uncorrelated, we should see no particular pattern, and they should be spread uniformly above and below the zero line within a rectangle defined by the value boundaries of the X (time) and Y axes.
- Instead, we see a definite pattern over time, suggesting the existence of **autocorrelation** (or also called **serial correlation**).
- We also see that the magnitude of residuals gets larger over time (heteroscedasticity). So, we will convert Revenues to lnRevenues using a log transformation.
  - lnRevenues <- log(Revenues)
- We re-do the regression and residuals for lnRevenues

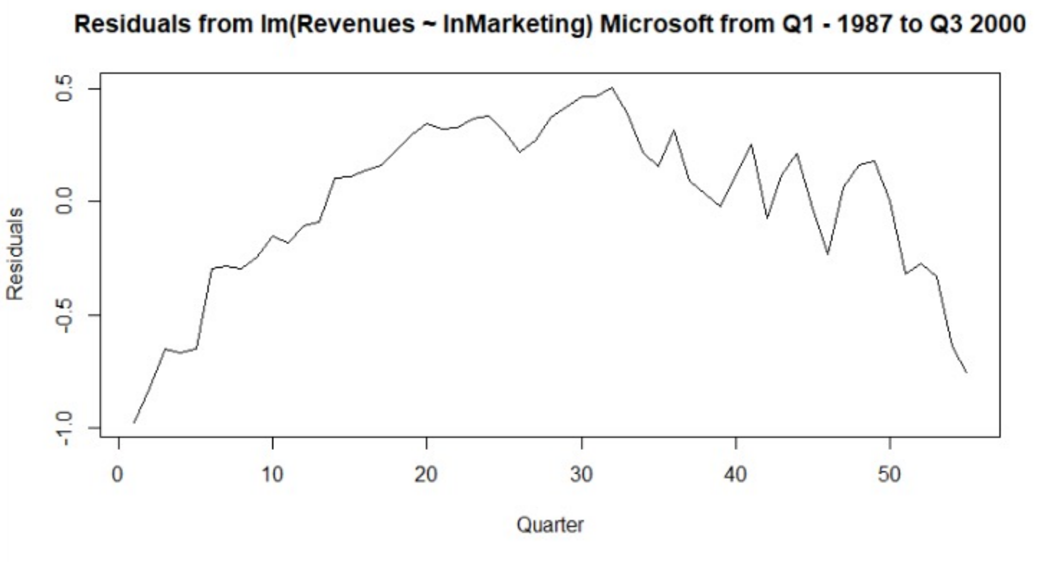
```
> mod1 <- lm(Revenues ~ Marketing, data=df)
> summary(mod1)

Call:
lm(formula = Revenues ~ Marketing, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-421.33 -160.95   17.77  141.17  577.79

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -203.6130    49.8504  -4.084  0.00015 ***
Marketing      5.7596     0.1754  32.839  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 227 on 53 degrees of freedom
Multiple R-squared:  0.9532,    Adjusted R-squared:  0.9523
F-statistic:  1078 on 1 and 53 DF,  p-value: < 2.2e-16

> resid1 <- residuals(mod1)
> obs <- df$Obs
> #
> plot (obs, resid1,
+       main= "Residuals from lm(Revenues ~ Marketing) Microsoft from Q1 - 1987 to Q3 2000",
+       xlab = "Quarter",
+       ylab = "Residuals",
+       type = "l")
```



# Modified Regression Model

- $\widehat{\ln Revenues} = \hat{\alpha} + \hat{\beta}_{mkt} Marketing$
- The residuals look better but continue to show a pattern, rather than being uniformly distributed around 0 across the X-axis (time).
- The residuals are clearly correlated with each other across time, displaying **autocorrelation**.

```
> lnRevenues <- log(df$Revenues)
> #
> mod2 <- lm(lnRevenues ~ Marketing, data=df)
> summary(mod2)

Call:
lm(formula = lnRevenues ~ Marketing, data = df)

Residuals:
    Min      1Q  Median      3Q     Max 
-0.9767 -0.2395  0.1084  0.2792  0.5056 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 4.9852310  0.0808894   61.63   <2e-16 ***
Marketing   0.0064078  0.0002846   22.52   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3683 on 53 degrees of freedom
Multiple R-squared:  0.9054,	Adjusted R-squared:  0.9036 
F-statistic:   507 on 1 and 53 DF,  p-value: < 2.2e-16

> resid2 <- residuals(mod2)
> plot (obs, resid2,
+       main= "Residuals from lm(Revenues ~ lnMarketing) Microsoft from Q1 - 1987 to Q3 2000",
+       xlab = "Quarter",
+       ylab = "Residuals",
+       type = "l")
```



# Autocorrelation

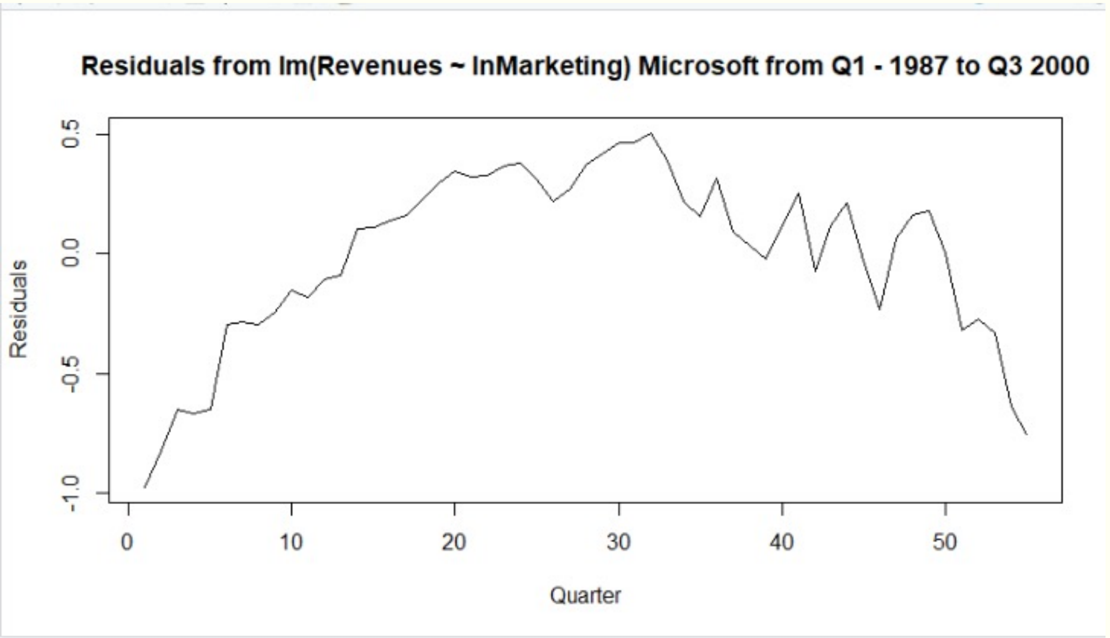
- Autocorrelation measures the *linear dependence between residuals* of successive (immediate or later) observations.

- The presence of autocorrelation does not mean that the values of an independent variable are correlated over time, or with each other, as occurs with multicollinearity.

- Instead, it is really about the *relationship between X and Y (say, Revenues and Marketing) over time.*

- The most common type of autocorrelation, **first-order autocorrelation**, is present when an observed error tends to be influenced by the observed error that **immediately precedes** it in the previous time period.

- We call this *first-order* autocorrelation because only *one time period* separates the two correlated error term observations.



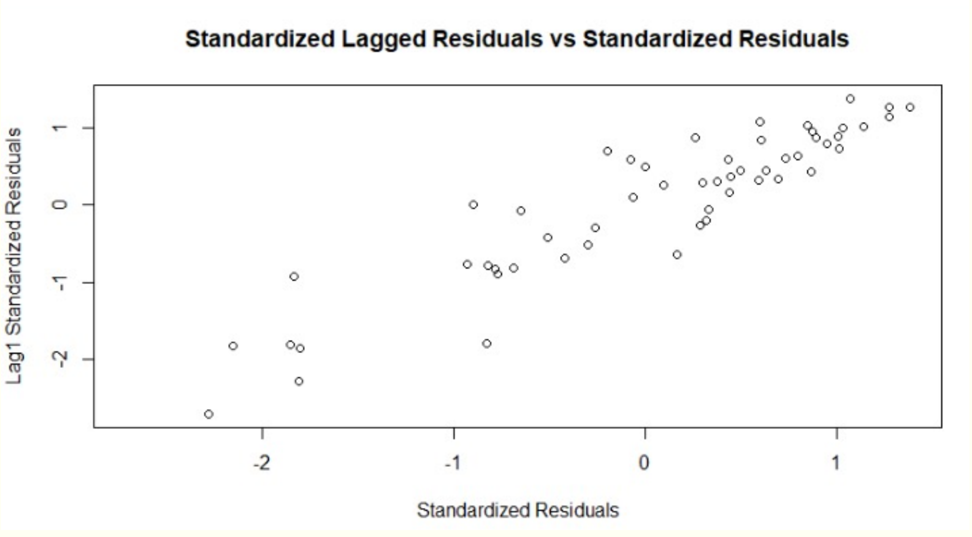
# Autocorrelation

- Since autocorrelation is a measure of linear relationship we have the following simple regression *among the residuals*:
  - $e_t = \boldsymbol{\rho} * e_{t-1} + u_t$
  - Where $e_t$ = residual from time period t and $e_{t-1}$ = residual from time period (t-1) i.e., previous time period
- If we standardized the residuals, we know that $\boldsymbol{\rho}$ is just the correlation between $e_t$ and $e_{t-1}$. Assuming the residuals are standardized, the value of $\boldsymbol{\rho}$ must fall between -1 and 1 since it measures correlation.
- If $\boldsymbol{\rho}$ were to **exactly** equal 1 or -1, the effect of one error on the next would not die out over time. For this reason, $\boldsymbol{\rho}$ must be greater than -1 and less than 1.

# Understanding Autocorrelation

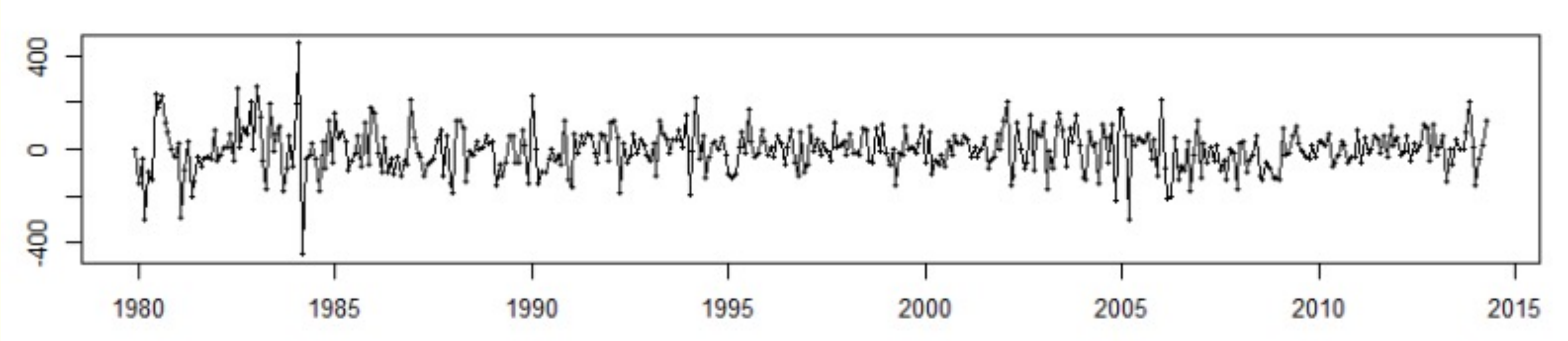
- We create the standardized residuals for our model, first.
- We create the lag1 residuals; i.e., we lag the residuals by 1 period so that the residual in time period 2, becomes the residual for time 1 in the ne set of residuals. Clearly, there will be one missing value.
- In **R**, we create lag1 standardized residuals using the head() function.
- To calculate the correlation, we have to account for the missing lagged observation in the first period. Using the *na.omit()* function.
- Here is what the originals and lagged residuals look like
- The first-order autocorrelation is the correlation between these two columns of data.
- A scatterplot shows a definite linear trend between the original and lagged residuals.

```
> st_resid2 <- rstandard(mod2)
> st_resid2_l1 <- c(NA, head(st_resid2, -1))
> M <- cbind(st_resid2, st_resid2_l1)
> print(head(M))
   st_resid2 st_resid2_l1
1 -2.7132728           NA
2 -2.2812072    -2.713273
3 -1.8115111    -2.281207
4 -1.8520245    -1.811511
5 -1.7994899    -1.852024
6 -0.8277391    -1.799490
> print(cor(na.omit(M)))
             st_resid2 st_resid2_l1
st_resid2    1.0000000    0.9188692
st_resid2_l1 0.9188692    1.0000000
> #
> plot(st_resid2, st_resid2_l1,
+      main= "Standardized Lagged Residuals vs Standardized Residuals",
+      xlab = "Standardized Residuals",
+      ylab = "Lag1 Standardized Residuals")
```



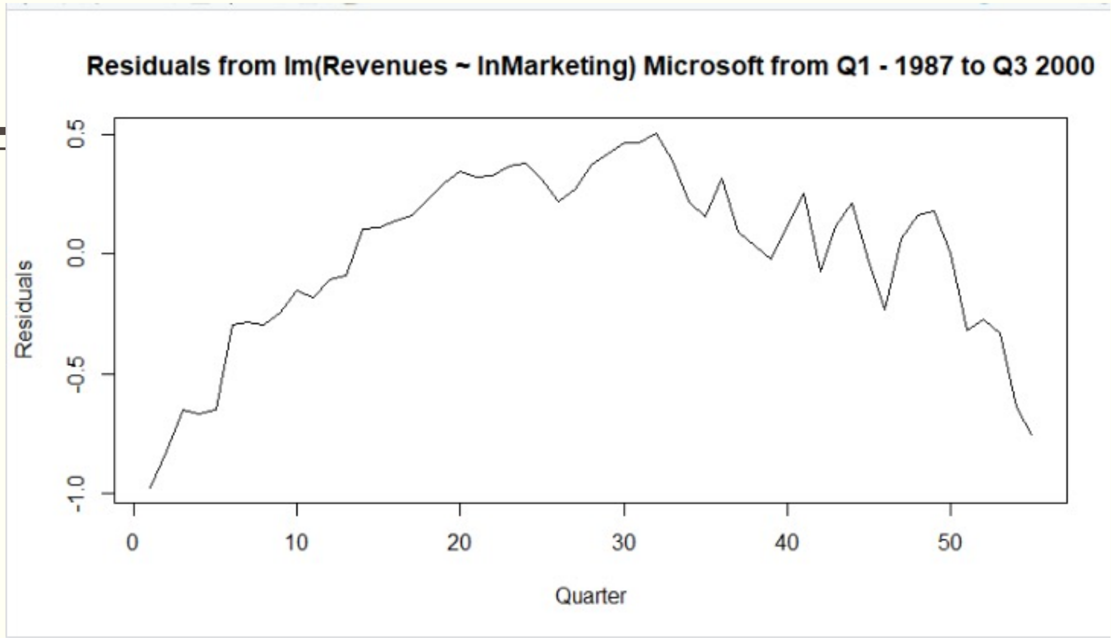
# Zero Autocorrelation ("White Noise")

- If **ρ** is zero, then one error has nothing to do with the next error, so there is no autocorrelation.
- The residuals we see **will not** follow a pattern over time.
- However, it is often hard to tell visually. There are other advanced functions that can be used to confirm lack of autocorrelation, that are beyond the scope of this course.



# Positive Autocorrelation

- If **ρ** is positive, the residuals tend (for the most part) to have the same sign from one period to the next.
  - If $e_{t-1}$ is positive, then $e_t$ tends to be positive; if $e_{t-1}$ is negative, $e_t$ tends to be negative.
- A positive **ρ** indicates positive autocorrelation, also called *positive serial correlation*. In our example, the error term observations exhibit **positive first-order autocorrelation**. But a scatterplot is not always the best way to see this.




```
> st_resid2 <- rstandard(mod2)
> st_resid2_l1 <- c(NA, head(st_resid2, -1))
> M <- cbind(st_resid2, st_resid2_l1)
> print(head(M))
   st_resid2 st_resid2_l1
1 -2.7132728           NA
2 -2.2812072    -2.713273
3 -1.8115111    -2.281207
4 -1.8520245    -1.811511
5 -1.7994899    -1.852024
6 -0.8277391    -1.799490
> print(cor(na.omit(M)))
             st_resid2 st_resid2_l1
st_resid2    1.0000000    0.9188692
st_resid2_l1 0.9188692    1.0000000
> #
> plot(st_resid2, st_resid2_l1,
+      main= "Standardized Lagged Residuals vs Standardized Residuals",
+      xlab = "Standardized Residuals",
+      ylab = "Lag1 Standardized Residuals")
```

# Negative Autocorrelation

- If **ρ** is negative, the errors tend to alternate signs, indicating **negative first-order autocorrelation** or negative serial correlation. In this situation, a positive observed error term is usually followed by a negative one, which is usually followed by a positive one, and so on. Negative first-order autocorrelation is less common than positive autocorrelation.

# Effects of Autocorrelation

- The presence of autocorrelation indicates that *an important predictor (namely, time) is missing* from the model.

- While the estimates of the slope coefficients remain unbiased, the standard errors of estimates of predictors in the model tend to be smaller than they should be.

- Therefore, the *t-statistic tends to inflate and the overall model F-test as well as the tests of coefficients tend to become significant*, even if they may not be.

- Therefore, even though Marketing appears to be a significant predictor of lnRevenues in our model (p-value near 0), we cannot trust this conclusion from our model if we find that significant autocorrelation is present among the residuals.

- To test for significance of autocorrelation amongst the residuals, we use the **Durbin-Watson** statistic and the Durbin-Watson test.

```
> lnRevenues <- log(df$Revenues)
> #
> mod2 <- lm(lnRevenues ~ Marketing, data=df)
> summary(mod2)

Call:
lm(formula = lnRevenues ~ Marketing, data = df)

Residuals:
    Min      1Q  Median      3Q     Max 
-0.9767 -0.2395  0.1084  0.2792  0.5056 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 4.9852310  0.0808894   61.63   <2e-16 ***
Marketing   0.0064078  0.0002846   22.52   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3683 on 53 degrees of freedom
Multiple R-squared:  0.9054,    Adjusted R-squared:  0.9036 
F-statistic:   507 on 1 and 53 DF,  p-value: < 2.2e-16
```

# DURBIN WATSON TEST FOR SIGNIFICANT AUTOCORRELATION

Lecture 4E-2

# Test for First-order Autocorrelation – The Durbin Watson Statistic (AutoCorr.R)

- The Durbin-Watson statistic provides a test **a test for first-order autocorrelation**.

- The Durbin-Watson statistic is:
  - $$\frac{(\hat{e}_2 - \hat{e}_1)^2 + (\hat{e}_3 - \hat{e}_2)^2 + \ \dots + (\hat{e}_n - \hat{e}_{n-1})^2}{(\hat{e}_1{}^2 + \hat{e}_2{}^2 + \hat{e}_3{}^2 + \cdots + \hat{e}_n{}^2)}$$

- In **R**, you need to install package “lmtest” and then use the and then use the dwtest() function.

- The Durbin-Watson statistic is ***approximately*** equal to (2 - 2 **ρ**).

- In our example, the first-order autocorrelation was 0.9189 and the D-W statistic for our model is 0.1447.

- The test shows that there is significant autocorrelation, at $\alpha = 0.05$, because the p-value is almost 0.

```
> library(lmtest)
> dwtest(mod2, alternative = c("greater"))

        Durbin-Watson test

data:  mod2
DW = 0.14474, p-value < 2.2e-16
alternative hypothesis: true autocorrelation is greater than 0
```

```
> print(cor(na.omit(M)))
             st_resid2 st_resid2_l1
st_resid2    1.0000000    0.9188692
st_resid2_l1 0.9188692    1.0000000
```

# Test for First-order Autocorrelation – The Durbin Watson Statistic

- Since, the Durbin-Watson statistic is ***approximately*** equal to (2 - 2 $\boldsymbol{\rho}$).
  - If there is no autocorrelation, then $\boldsymbol{\rho}$ = 0. The Durbin-Watson statistic = 2.
  - The worst possible case of positive first-order autocorrelation occurs when $\boldsymbol{\rho}$ is very close to 1. If $\boldsymbol{\rho}$ = 1, the Durbin-Watson statistic = 0.
    - This means the closer the Durbin-Watson statistic is to zero, the more likely serious first-order positive autocorrelation exists.
  - For *negative first-order autocorrelation*, the worst possible case occurs when $\boldsymbol{\rho}$ is close to -1. If $\boldsymbol{\rho}$ = -1, the Durbin-Watson statistic = 4.
    - This means that when the Durbin-Watson statistic is closer to 4, the chances of first-order negative first-order autocorrelation increase.

- In summary, the Durbin-Watson statistic varies from 0 to 4:
  - values closer to 0 indicate *positive* first-order autocorrelation;
  - values close to 2 indicate *no* first-order autocorrelation; and
  - values closer to 4 indicate *negative* first-order autocorrelation.

# Test for First-order Autocorrelation – The Durbin Watson Statistic

- Most hypothesis tests use a critical value to separate the regions where the null hypothesis is rejected or not rejected.
- The Durbin-Watson statistic has **three regions** with two critical values referred to as $\mathbf{d_U}$ ("d-upper") and $\mathbf{d_L}$ ("d-lower") and are given in Tables.
- The decision rules for the Durbin-Watson test for **positive** first-order autocorrelation are as follows:
    - If the Durbin-Watson statistic is less than $\mathbf{d_L}$, **reject** the null hypothesis of no first-order autocorrelation; assume positive first-order autocorrelation.
    - If the Durbin-Watson statistic is greater than $\mathbf{d_u}$, **do not reject** the null hypothesis of no first-order autocorrelation; assume no first-order autocorrelation.
    - If the Durbin-Watson statistic lies between $\mathbf{d_L}$ and $\mathbf{d_u}$ (or exactly equal to either $\mathbf{d_L}$ or $\mathbf{d_u}$), the test is **inconclusive** regarding first-order autocorrelation.
- Use the number of independent variables in your regression to find the correct column for $\mathbf{d_U}$ and $\mathbf{d_L}$ in the table, and the sample size n to find the correct row.

**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 | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| n | 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 |
| 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 |
| 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 |
| 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 |
| 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 |
| 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 |
| 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 |
| 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 |
| 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 |
| 23 | 1.257 | 1.437 | 1.168 | 1.543 | 1.078 | 1.660 | 0.986 | 1.785 | 0.895 | 1.920 | 0.804 | 2.061 | 0.715 | 2.208 | 0.628 | 2.360 |
| 24 | 1.273 | 1.446 | 1.188 | 1.546 | 1.101 | 1.656 | 1.013 | 1.775 | 0.925 | 1.902 | 0.837 | 2.035 | 0.750 | 2.174 | 0.666 | 2.318 |
| 25 | 1.288 | 1.454 | 1.206 | 1.550 | 1.123 | 1.654 | 1.038 | 1.767 | 0.953 | 1.886 | 0.868 | 2.013 | 0.784 | 2.144 | 0.702 | 2.280 |
| 26 | 1.302 | 1.461 | 1.224 | 1.553 | 1.143 | 1.652 | 1.062 | 1.759 | 0.979 | 1.873 | 0.897 | 1.992 | 0.816 | 2.117 | 0.735 | 2.246 |
| 27 | 1.316 | 1.469 | 1.240 | 1.556 | 1.162 | 1.651 | 1.084 | 1.753 | 1.004 | 1.861 | 0.925 | 1.974 | 0.845 | 2.093 | 0.767 | 2.216 |
| 28 | 1.328 | 1.476 | 1.255 | 1.560 | 1.181 | 1.650 | 1.104 | 1.747 | 1.028 | 1.850 | 0.951 | 1.959 | 0.874 | 2.071 | 0.798 | 2.188 |
| 29 | 1.341 | 1.483 | 1.270 | 1.563 | 1.198 | 1.650 | 1.124 | 1.743 | 1.050 | 1.841 | 0.975 | 1.944 | 0.900 | 2.052 | 0.826 | 2.164 |
| 30 | 1.352 | 1.489 | 1.284 | 1.567 | 1.214 | 1.650 | 1.143 | 1.739 | 1.071 | 1.833 | 0.998 | 1.931 | 0.926 | 2.034 | 0.854 | 2.141 |
| 31 | 1.363 | 1.496 | 1.297 | 1.570 | 1.229 | 1.650 | 1.160 | 1.735 | 1.090 | 1.825 | 1.020 | 1.920 | 0.950 | 2.018 | 0.879 | 2.120 |
| 32 | 1.373 | 1.502 | 1.309 | 1.574 | 1.244 | 1.650 | 1.177 | 1.732 | 1.109 | 1.819 | 1.041 | 1.909 | 0.972 | 2.004 | 0.904 | 2.102 |
| 33 | 1.383 | 1.508 | 1.321 | 1.577 | 1.258 | 1.651 | 1.193 | 1.730 | 1.127 | 1.813 | 1.061 | 1.900 | 0.994 | 1.991 | 0.927 | 2.085 |
| 34 | 1.393 | 1.514 | 1.333 | 1.580 | 1.271 | 1.652 | 1.208 | 1.728 | 1.144 | 1.808 | 1.079 | 1.891 | 1.015 | 1.978 | 0.950 | 2.069 |
| 35 | 1.402 | 1.519 | 1.343 | 1.584 | 1.283 | 1.653 | 1.222 | 1.726 | 1.160 | 1.803 | 1.097 | 1.884 | 1.034 | 1.967 | 0.971 | 2.054 |
| 36 | 1.411 | 1.525 | 1.354 | 1.587 | 1.295 | 1.654 | 1.236 | 1.724 | 1.175 | 1.799 | 1.114 | 1.876 | 1.053 | 1.957 | 0.991 | 2.041 |
| 37 | 1.419 | 1.530 | 1.364 | 1.590 | 1.307 | 1.655 | 1.249 | 1.723 | 1.190 | 1.795 | 1.131 | 1.870 | 1.071 | 1.948 | 1.011 | 2.029 |
| 38 | 1.427 | 1.535 | 1.373 | 1.594 | 1.318 | 1.656 | 1.261 | 1.722 | 1.204 | 1.792 | 1.146 | 1.864 | 1.088 | 1.939 | 1.029 | 2.017 |
| 39 | 1.435 | 1.540 | 1.382 | 1.597 | 1.328 | 1.658 | 1.273 | 1.722 | 1.218 | 1.789 | 1.161 | 1.859 | 1.104 | 1.932 | 1.047 | 2.007 |
| 40 | 1.442 | 1.544 | 1.391 | 1.600 | 1.338 | 1.659 | 1.285 | 1.721 | 1.230 | 1.786 | 1.175 | 1.854 | 1.120 | 1.924 | 1.064 | 1.997 |
| 45 | 1.475 | 1.566 | 1.430 | 1.615 | 1.383 | 1.666 | 1.336 | 1.720 | 1.287 | 1.776 | 1.238 | 1.835 | 1.189 | 1.895 | 1.139 | 1.958 |
| 50 | 1.503 | 1.585 | 1.462 | 1.628 | 1.421 | 1.674 | 1.378 | 1.721 | 1.335 | 1.771 | 1.291 | 1.822 | 1.246 | 1.875 | 1.201 | 1.930 |
| 55 | 1.528 | 1.601 | 1.490 | 1.641 | 1.452 | 1.681 | 1.414 | 1.724 | 1.374 | 1.768 | 1.334 | 1.814 | 1.294 | 1.861 | 1.253 | 1.909 |
| 60 | 1.549 | 1.616 | 1.514 | 1.652 | 1.480 | 1.689 | 1.444 | 1.727 | 1.408 | 1.767 | 1.372 | 1.808 | 1.335 | 1.850 | 1.298 | 1.894 |

# Testing for Positive First-order Autocorrelation

- For positive first-order autocorrelation the hypotheses are:
  - $H_0$: $\rho = 0$ and $H_a$: $\rho > 0$.
- In our case, the Durbin-Watson statistic was 0.1447, our sample $\rho$ was 0.9189, indicating *strong positive first-order autocorrelation*.
- For our example, from **Tables**:
  - k = 1 and n=55,
  - At $\alpha = 0.05$, $d_L = 1.528$ and $d_U = 1.601$.
  - The calculated statistic is: 0.1447
  - In our case, 0.1447 < dL (=1.528) so we reject the null hypothesis of no first-order positive autocorrelation at a 5% significance level.
- From the Durbin-Watson test in **R**, we see that the p-value is close to 0, indicating that the **null hypothesis of zero autocorrelation is rejected**.
- We can assume that there is *significant positive first-order autocorrelation*.
- Therefore, because of the significant autocorrelation, we cannot safely conclude that the coefficient for MARKETING is statistically significant even though the t-test in reports it is significant at 1%.
- This t-statistic for the slope of Marketing could be inflated by the significant autocorrelation.

```
> mod2 <- lm(lnRevenues ~ Marketing, data=df)
> summary(mod2)

Call:
lm(formula = lnRevenues ~ Marketing, data = df)

Residuals:
    Min      1Q  Median      3Q     Max 
-0.9767 -0.2395  0.1084  0.2792  0.5056 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 4.9852310  0.0808894   61.63   <2e-16 ***
Marketing   0.0064078  0.0002846   22.52   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3683 on 53 degrees of freedom
Multiple R-squared:  0.9054,	Adjusted R-squared:  0.9036 
F-statistic:   507 on 1 and 53 DF,  p-value: < 2.2e-16
```

```
# install.packages("lmtest")
library(lmtest)
dwtest(mod2, alternative = c("greater"))
```

```
	Durbin-Watson test

data:  mod2
DW = 0.14474, p-value < 2.2e-16
alternative hypothesis: true autocorrelation is greater than 0
```

# Testing for Negative First-order Autocorrelation

- The decision rules for a Durbin-Watson negative autocorrelation test are different from those for positive first-order autocorrelation.
- The null and alternative hypotheses are $\mathbf{H_0}$: $\boldsymbol{\rho} = 0$ and $\mathbf{H_a}$: $\boldsymbol{\rho} < 0$.
- When the Durbin-Watson statistic comes out greater than 2, negative first-order autocorrelation may be present.
- The decision rules for a Durbin-Watson test of negative first-order autocorrelation:
  - If the Durbin-Watson statistic is greater than 4 - $\mathbf{d_L}$, reject the null hypothesis of no first-order autocorrelation; assume negative first-order autocorrelation.
  - If the Durbin-Watson statistic is less than 4 - $\mathbf{d_U}$, do not reject the null hypothesis of no first-order autocorrelation; assume no first-order autocorrelation.
  - If the Durbin-Watson statistic lies between 4 - $\mathbf{d_L}$ and 4 - $\mathbf{d_U}$ (or exactly equal to either 4 - $\mathbf{d_L}$ or 4 - $\mathbf{d_U}$ ), the test is ***inconclusive*** regarding negative first-order autocorrelation.

# Using Time as a Predictor in OLS Regression Models

- In the usual OLS regression models, when significant autocorrelation is present, one approach is to include time as a predictor
- Suppose that the observed series is $y_t$, for t=1,2,…,n.
  - For a linear trend, use t (the time index) as a predictor variable in a regression.
  - For a quadratic trend, we might consider using both t and $t^2$.
  - For quarterly data, with possible seasonal (quarterly) effects, we can define indicator variables such as $S_j$ = 1 if observation is in quarter j of a year and 0 otherwise. There are 3 such indicators.
- Model with linear and quadratic trends and seasonal factor
  - $\hat{y}_t = \hat{\propto} + \hat{\beta}_1 t + \hat{\beta}_2 t^2 + \hat{\beta}_{s1} S_1 + \hat{\beta}_{s2} S_2 + \hat{\beta}_{s3} S_3$

| | Obs | Year | Quarter | Revenues | Marketing | Summer | Fall | Winter |
|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 1987 | 1 | 60.02 | 13.44 | 0 | 0 | 1 |
| 2 | 2 | 1987 | 2 | 71.62 | 16.80 | 0 | 0 | 0 |
| 3 | 3 | 1987 | 3 | 85.66 | 18.36 | 1 | 0 | 0 |
| 4 | 4 | 1987 | 4 | 86.68 | 22.54 | 0 | 1 | 0 |
| 5 | 5 | 1988 | 1 | 88.74 | 23.26 | 0 | 0 | 1 |
| 6 | 6 | 1988 | 2 | 132.73 | 31.48 | 0 | 0 | 0 |

- **Notes**:
  - We can include other X predictors (like MARKETING in our example)
  - $\hat{\beta}_1$ models a linear trend in $\hat{y}_t$ with time, adding $\hat{\beta}_2$ models a quadratic trend in $\hat{y}_t$ with time
  - $\hat{\beta}_{s1}$, $\hat{\beta}_{s2}$, and $\hat{\beta}_{s3}$ model seasonal components as dummy variables.

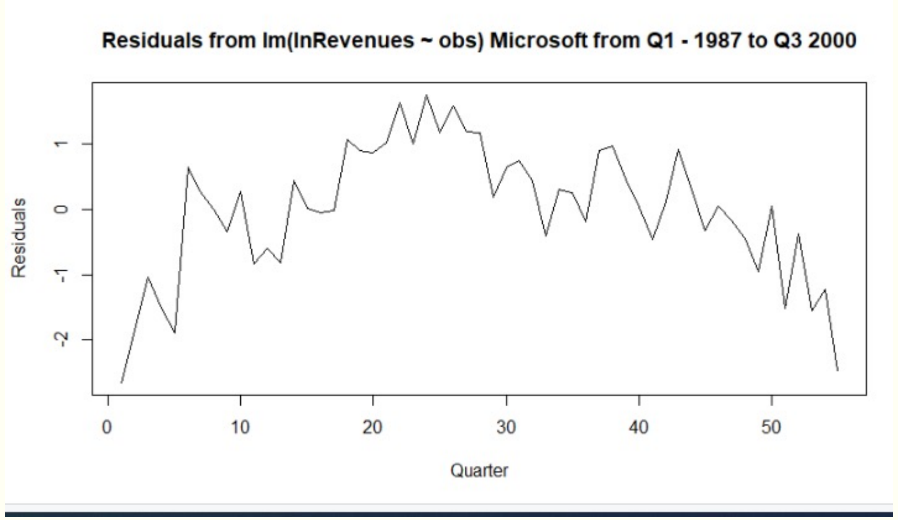
# Using Time as a Predictor in OLS Regression Models

- Going back to our example, we will first try a model with linear trend term.
  - $\widehat{\ln Revenues} = \hat{\alpha} + \hat{\beta}_1 obs$ **(note: obs is our time indicator, t)**
  - $\widehat{\ln Revenues} = -4.36301 + 0.0736t$
- We will look at model fit, the Durbin-Watson statistic, as well as the plot of residuals against time
- The plot shows a quadratic trend with time
- In **R**, the D-W statistic of 0.5377 has a p-value near 0, and still shows significant positive auto-correlation at $\alpha$ = 0.05.

Using Tables, k = 1 and n = 55:

- At $\alpha$ = 0.05, $\mathbf{d_L}$ = 1.528 and $\mathbf{d_U}$ = 1.601.
- The calculated statistic is: 0.5377
- For positive autocorrelation the hypotheses are: H0: $\boldsymbol{\rho}$ = 0 and Ha: $\boldsymbol{\rho}$ > 0.
- In our case 0.5377 < $\mathbf{d_L}$ so we **reject the null hypothesis of no first-order autocorrelation** at $\alpha$ = 0.05.
- We can assume that there is significant positive autocorrelation.

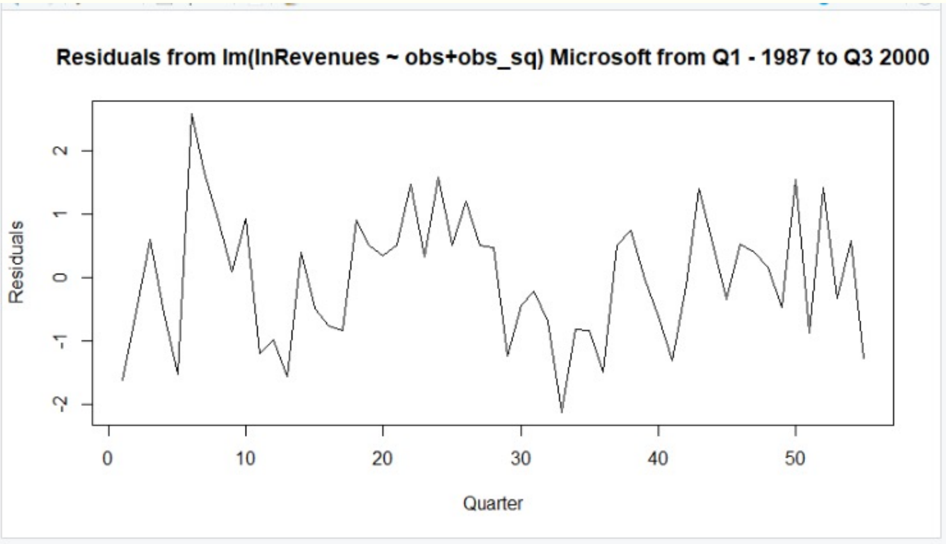
```
> mod3 <- lm(lnRevenues ~ obs)
> summary(mod3)

Call:
lm(formula = lnRevenues ~ obs)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.34190 -0.05929  0.00643  0.10630  0.23197 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 4.363010   0.036472  119.63   <2e-16 ***
obs         0.073567   0.001133   64.92   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1334 on 53 degrees of freedom
Multiple R-squared:  0.9876,    Adjusted R-squared:  0.9873 
F-statistic:  4215 on 1 and 53 DF,  p-value: < 2.2e-16

> st_resid3 <- rstandard(mod3)
> plot (obs, st_resid3,
+        main= "Residuals from lm(lnRevenues ~ obs) Microsoft from Q1 - 1987 to Q3 2000",
+        xlab = "Quarter",
+        ylab = "Residuals",
+        type = "l")
> dwtest(mod3, alternative = c("greater"))

        Durbin-Watson test

data:  mod3
DW = 0.5377, p-value = 1.125e-11
alternative hypothesis: true autocorrelation is greater than 0
```



# Using Time as a Predictor in OLS Regression Models

- We next try a model with linear and quadratic trend terms.
  - $\widehat{\ln Revenues} = \hat{\alpha} + \hat{\beta}_1 obs + \hat{\beta}_2 obs^2$
  - $\widehat{\ln Revenues} = 4.1109 + 0.1t - 0.00047t^2$
- The Durbin-Watson statistic (1.532) looks much better
- The spread of the residuals show no curved pattern of autocorrelations and the model fit is excellent suggesting we have a good model.
- In **R,** the p-value of the DW Test is 0.0195 suggesting that we would reject the null hypothesis of no autocorrelation and conclude that there is still significant autocorrelation at $\alpha = 0.05$.

```
> #
> mod4 <- lm(lnRevenues ~ obs+obs_sq)
> summary(mod4)

Call:
lm(formula = lnRevenues ~ obs + obs_sq)

Residuals:
      Min        1Q    Median        3Q       Max
-0.162740 -0.060743  0.006585  0.042251  0.193853

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.111e+00  3.273e-02  125.59  < 2e-16 ***
obs          1.001e-01  2.697e-03   37.12  < 2e-16 ***
obs_sq      -4.739e-04  4.668e-05  -10.15 6.15e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.07799 on 52 degrees of freedom
Multiple R-squared:  0.9958,    Adjusted R-squared:  0.9957
F-statistic:  6218 on 2 and 52 DF,  p-value: < 2.2e-16

> st_resid4 <- rstandard(mod4)
> plot (obs, st_resid4,
+       main= "Residuals from lm(lnRevenues ~ obs+obs_sq) Microsoft from Q1 - 1987 to Q3 2000",
+       xlab = "Quarter",
+       ylab = "Residuals",
+       type = "l")
> dwtest(mod4, alternative = c("greater"))

        Durbin-Watson test

data:  mod4
DW = 1.5322, p-value = 0.01953
alternative hypothesis: true autocorrelation is greater than 0
```



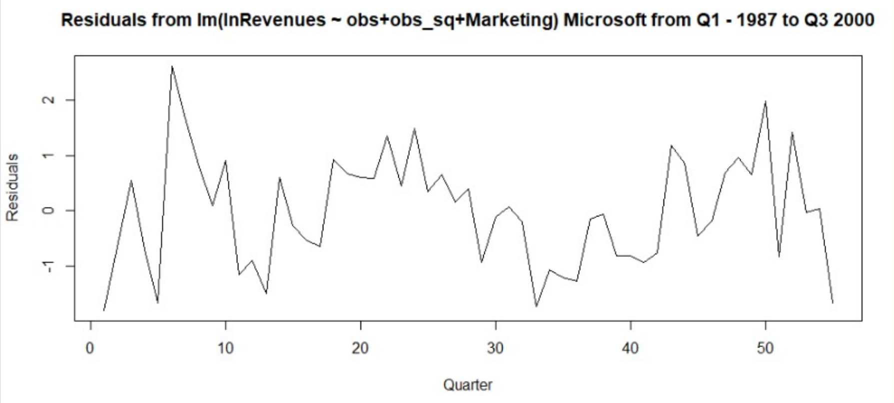

However, from **Tables**, k = 2 and n = 55:

- At $\alpha = 0.05$, $\mathbf{d_L} = 1.490$ and $\mathbf{d_U} = 1.641$.
- The calculated statistic is: 1.5322
- For positive autocorrelation the hypotheses are: H0: $\boldsymbol{\rho} = 0$ and Ha: $\boldsymbol{\rho} > 0$.
- In our case $\mathbf{d_U} > 1.5322 > \mathbf{d_L}$ so the **test of no first-order autocorrelation is inconclusive** at $\alpha = 0.05$.

# Using Time as a Predictor in OLS Regression Models

- We model *ln(revenues)* with linear and quadratic trend terms and with **Marketing** as predictor.
  - $\ln(\widehat{revenues_t}) = \hat{\beta}_1 obs + \hat{\beta}_2 obs^2 + \hat{\beta}_{mkt} marketing$
  - $\mathbf{\ln(\widehat{revenues})} = 4.10 + 0.097t - 0.0007t^2 + 0.0013 marketing$
- Using **R**, the Durbin-Watson statistic is 1.3835 is smaller and the DW Test has a p-value of 0.003227, implying significant autocorrelation at $\alpha = 0.05$.
- The model suggests that marketing is a significant predictor with the time-related variables in the model.
- However, given the larger D_W statistic is smaller compared to the previous model and the significant first-order autocorrelation we may prefer the previous model.

```
> mod5 <- lm(lnRevenues ~ obs+obs_sq+Marketing, data=df)
> summary(mod5)

Call:
lm(formula = lnRevenues ~ obs + obs_sq + Marketing, data = df)

Residuals:
      Min        1Q    Median        3Q       Max
-0.122498 -0.057361 -0.001252  0.047121  0.185079

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.102e+00  3.103e-02 132.212  < 2e-16 ***
obs          9.738e-02  2.730e-03  35.670  < 2e-16 ***
obs_sq      -6.746e-04  8.552e-05  -7.888 2.17e-10 ***
Marketing    1.306e-03  4.770e-04   2.738   0.0085 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.07353 on 51 degrees of freedom
Multiple R-squared:  0.9964,	Adjusted R-squared:  0.9962
F-statistic:  4665 on 3 and 51 DF,  p-value: < 2.2e-16

> st_resid5 <- rstandard(mod5)
> plot (obs, st_resid5,
+       main= "Residuals from lm(lnRevenues ~ obs+obs_sq+Marketing) Micro
+       xlab = "Quarter",
+       ylab = "Residuals",
+       type = "l")
> dwtest(mod5, alternative = c("greater"))

	Durbin-Watson test

data:  mod5
DW = 1.3835, p-value = 0.003227
alternative hypothesis: true autocorrelation is greater than 0
```




Using **Tables**, k = 3 and n = 55:

- At $\alpha = 0.05$, $\mathbf{d_L} = 1.452$ and $\mathbf{d_U} = 1.681$.
- The calculated statistic is: 1.383
- For positive autocorrelation the hypotheses are: H0: $\boldsymbol{\rho} = 0$ and Ha: $\boldsymbol{\rho} > 0$.
- In our case $1.383 < \mathbf{d_L}$ so we **reject the null hypothesis of no first-order autocorrelation** at $\alpha = 0.05$.

# Using Time as a Predictor in OLS Regression Models

- Among the 4 models we tested, Model 3 with only the linear and quadratic terms for time as predictor, resulted in the least auto-correlation.
- The DW Test in **R** suggests continued presence of significant autocorrelation, while the table suggests that the test is inconclusive.
- We could try the seasonal terms $\widehat{y_t} = \widehat{\propto} + \widehat{\beta}_1 t + \widehat{\beta}_2 t^2 + \widehat{\beta}_{s1} S_1 + \widehat{\beta}_{s2} S_2 + \widehat{\beta}_{s3} S_3$ model, but the residuals do not display a clear seasonal pattern; also, it is beyond the scope of this course.
- We can therefore stop with the quadratic model: $\widehat{\ln Revenues} = 4.1109 + 0.1t - 0.00047t^2$ (given the basic modeling approach that we used).
- Note that this is **not** a very "constructive model" because it models revenue purely as a function of time. It may be good for prediction, but does not give insights into variables that determine revenue
- More advanced techniques may be required to obtain a better model.
- In particular there are other time series modeling approaches that may suggest better models.

# Advanced Methods to Account for Autocorrelation

- Other more advanced approaches to Time Series Models:

  1) *Auto-regressive (AR)* Models use lagged versions of the dependent variable as a predictor i.e., in predicting $Y_t$, we use $Y_{t-1}$ and/or $Y_{t-2}$, etc. as predictors. That is, we are regressing Y on a version of itself (auto-regressive).

  2) *Moving average (MA)* Models use a moving average predictor term that consists of a past (lagged) error term multiplied by a coefficient. i.e., in predicting $Y_t$, we use $\theta e_{t-1}$ as a predictor.

  3) Use a *differencing* approach – Use a difference between the predictor at time t and its lagged version to predict the difference between the dependent variable and its lagged version. i.e., for example, use $(X_t - X_{t-1})$ to predict $(Y_t - Y_{t-1})$

  4) Combine 2), 3) and 4) in various ways – the so-called ARIMA models. ARIMA models use Maximum Likelihood Estimation (MLE) and not OLS (typically)