THE UNIVERSITY OF TEXAS AT AUSTIN
McCOMBS SCHOOL OF BUSINESS

# Residuals and autocorrelation 2

## Lecture 6

### STA 371G

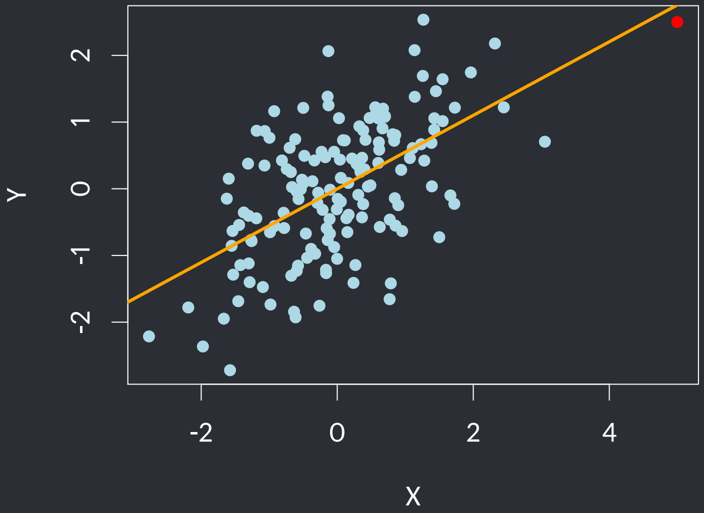
1. Unusual observations

2. Autocorrelation

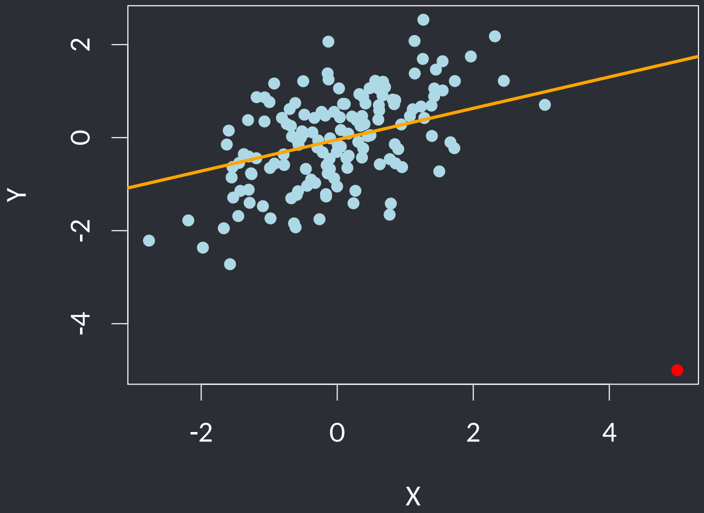
# What a single case can do

Even a single case can wreak havoc on the regression line. Let's add one outlier, at $X = 5$, and see what happens with different $Y$ values.

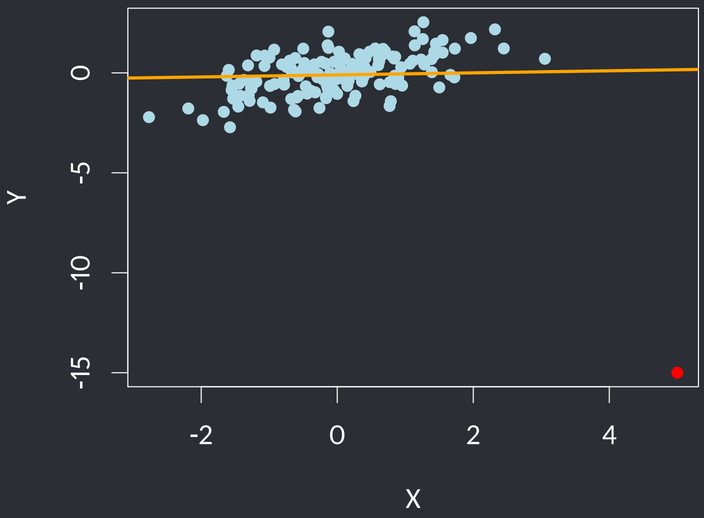
# What a single case can do



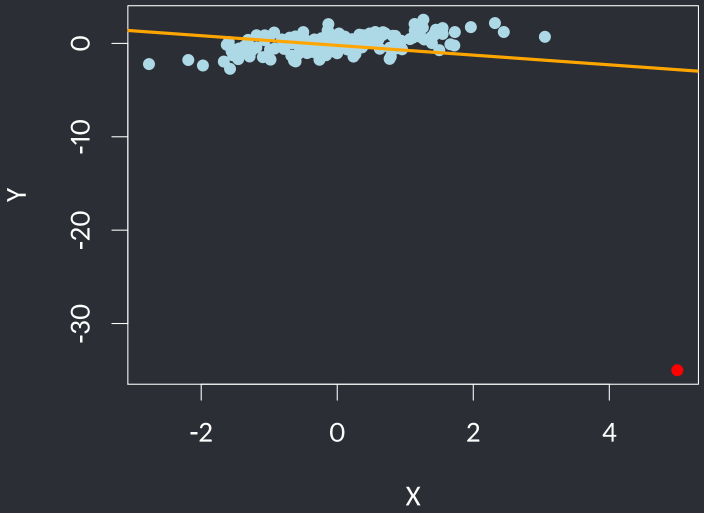
# What a single case can do



# What a single case can do



# What a single case can do




BLACKMAIL

## Regression is like blackmail

Blackmail:

- Compromising information gives a blackmailer leverage—the *potential* to have a big impact
- Once the blackmailer uses the information, that gives them influence

# Regression is like blackmail

Blackmail:

- Compromising information gives a blackmailer leverage—the *potential* to have a big impact
- Once the blackmailer uses the information, that gives them influence

Regression:

- When a point has a very unusual $X$ value (i.e., far from $\bar{X}$), it has leverage—the *potential* to have a big impact on the regression line
- When that point *also* has a $Y$ value that is out of line with the general trend, it will pull the regression line towards it—giving it influence

# How do I know what points are influential?

- Cook's distance can be a useful metric for finding influential points

# How do I know what points are influential?

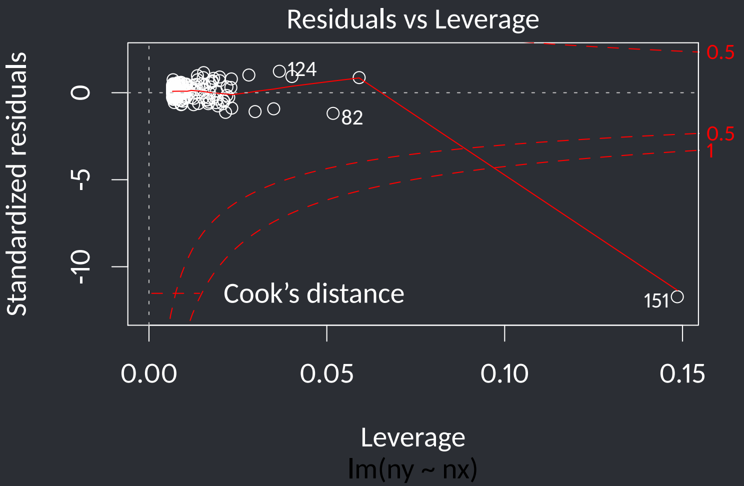
- Cook's distance can be a useful metric for finding influential points
- Larger values of Cook's distance indicate more influential points, but there is no firm cutoff

# How do I know what points are influential?

- Cook's distance can be a useful metric for finding influential points
- Larger values of Cook's distance indicate more influential points, but there is no firm cutoff
- Use Cook's distance to help you find points that might be influential, and then run the regression both with and without the point to judge for yourself

# Using the Cook’s distance plot in R

```
> plot(model, which=5)
```



1. Unusual observations

2. Autocorrelation

# What is autocorrelation?

- The first assumption of regression is that the errors are independent.

# What is autocorrelation?

- The first assumption of regression is that the errors are independent.
- In many time series data sets, this isn't the case—what happens in one time period is often correlated with those time periods right before or after.

# What is autocorrelation?

- The first assumption of regression is that the errors are independent.
- In many time series data sets, this isn't the case—what happens in one time period is often correlated with those time periods right before or after.
- Autocorrelation is when we can consistently predict the value of a variable at a particular time based on other times.

# How can autocorrelation be detected?

- Sometimes, it's clear that there is likely to be autocorrelation.
- Any time that the value of a time series builds on the previous stage (e.g., daily stock price, annual revenue) autocorrelation is a likely danger.
- But it's worth checking for autocorrelation in any time series!

# What about when it's not obvious?

The Durbin-Watson test lets us test the null hypothesis that the errors in a regression come from a population where successive errors are uncorrelated.

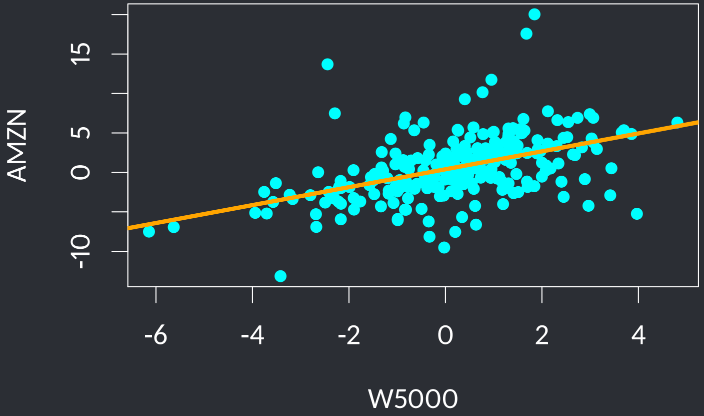
# What about when it's not obvious?

The Durbin-Watson test lets us test the null hypothesis that the errors in a regression come from a population where successive errors are uncorrelated.

If we **reject** the null hypothesis, then there is evidence of autocorrelation, and the independence assumption is violated.

# Example

Consider the stock market data from Lecture 4, when we regressed a company's weekly return on the weekly return of a market index (Wilshire 5000). Is autocorrelation present?



# Example

```
> library(lmtest)
> model <- lm(AMZN ~ W5000, data=stock.market)
> dwtest(model)


Durbin-Watson test

data:  model
DW = 1.9759, p-value = 0.4249
alternative hypothesis: true autocorrelation is greater than 0
```

Here, $p = 0.42 > 0.05$, so we fail to reject the null hypothesis: there is no evidence of (first-order) autocorrelation.

# Example

```
> library(lmtest)
> model <- lm(AMZN ~ W5000, data=stock.market)
> dwtest(model)


Durbin-Watson test

data:  model
DW = 1.9759, p-value = 0.4249
alternative hypothesis: true autocorrelation is greater than 0
```

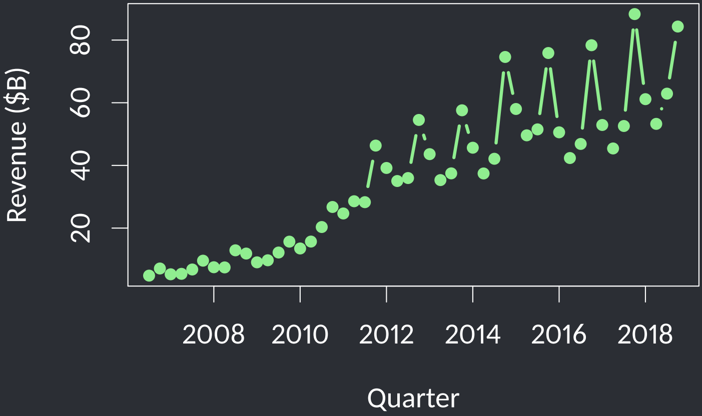
Here, $p = 0.42 > 0.05$, so we fail to reject the null hypothesis: there is no evidence of (first-order) autocorrelation. (Surprised?)

## First-order?!

- Durbin-Watson only lets us test for first-order autocorrelation; that is, correlation between the error at time $t$ (today) and the error at time $t - 1$ (yesterday, last month, last year, etc).
- Sometimes the error at time $t$ is correlated not with time $t - 1$ but time $t - 2$ or $t - 3$, etc.

# Example

Let's look at Apple's quarterly revenue since 2006:




What do you think the peaks correspond to?

## The autocorrelation function

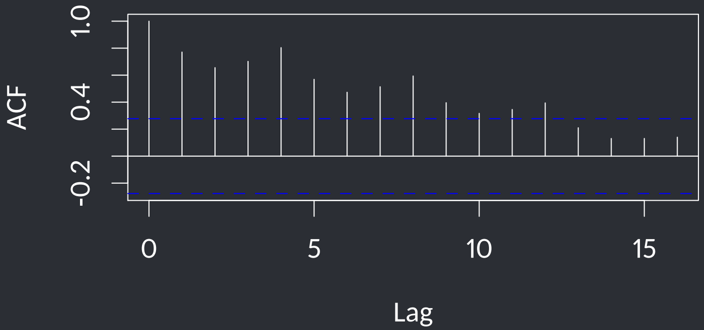
Another approach to suss out autocorrelation is to calculate all possible correlations with the time series and itself, “lagged” back by different time steps:

| | Apple revenue ($B) | Lag 1 | Lag 2 | Lag 3 |
|---|---|---|---|---|
| 2006 Q3 | 4.837 | NA | NA | NA |
| 2006 Q4 | 7.115 | 4.837 | NA | NA |
| 2007 Q1 | 5.264 | 7.115 | 4.837 | NA |
| 2007 Q2 | 5.410 | 5.264 | 7.115 | 4.837 |
| 2007 Q3 | 6.789 | 5.410 | 5.264 | 7.115 |
| 2007 Q4 | 9.608 | 6.789 | 5.410 | 5.264 |

# The autocorrelation function

The autocorrelations are highest for lag 1 and lag 4 (i.e., one quarter and one year ago):

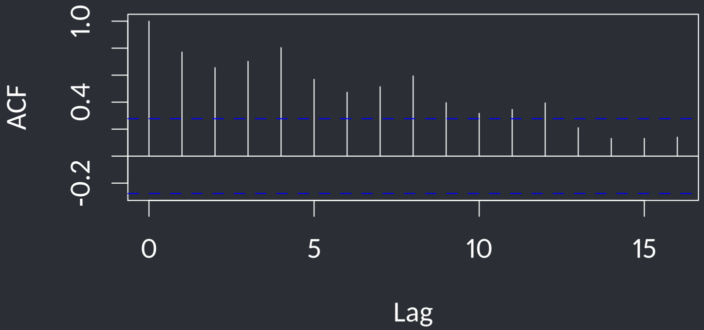
```
> acf(apple$Revenue.Billions)
```



# The autocorrelation function

The autocorrelations are highest for lag 1 and lag 4 (i.e., one quarter and one year ago):

```
> acf(apple$Revenue.Billions)
```



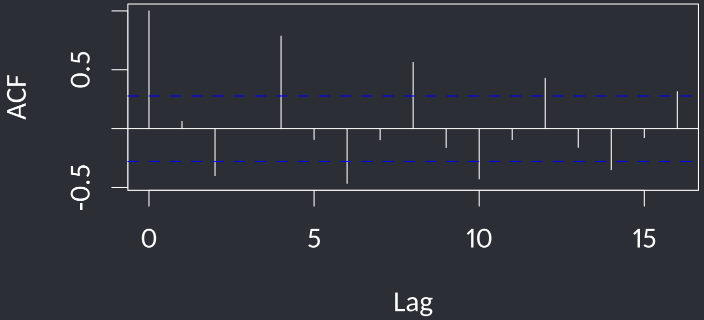
# The autocorrelation function, applied to residuals

What we really need to test is the autocorrelation of the *residuals*, not of the revenue itself. Durbin-Watson suggests there is no first-order autocorrelation:

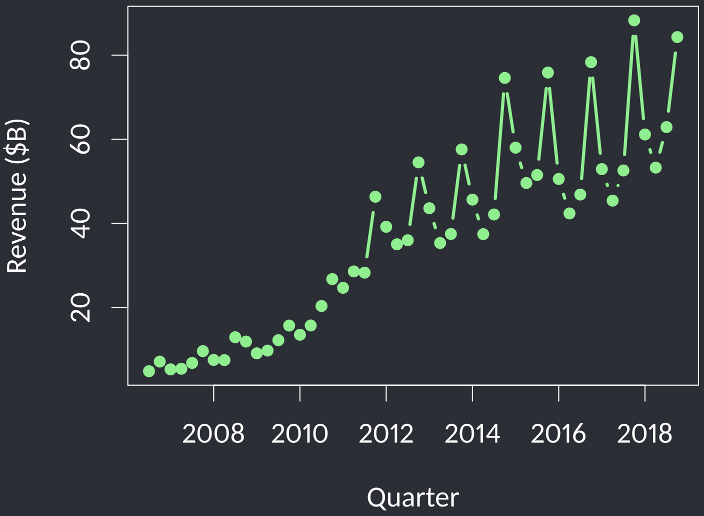
```
> model <- lm(Revenue.Billions ~ Time, data=apple)
> dwtest(model)


Durbin-Watson test

data:  model
DW = 1.8392, p-value = 0.2347
alternative hypothesis: true autocorrelation is greater than 0
```

But there is clearly second- and fourth-order autocorrelation!

```
> model <- lm(Revenue.Billions ~ Time, data=apple)
> acf(residuals(model))
```




How do we interpret these–what accounts for this pattern?


Revenue ($B)
80
60
40
20
2008
2010
2012
2014
2016
2018
Quarter

## How to handle autocorrelation

- When time-series data is used, it’s important to check for autocorrelation.

## How to handle autocorrelation

- When time-series data is used, it's important to check for autocorrelation.
- If present, the independence assumption is violated, and we shouldn't trust the $p$-values and confidence intervals that come out of the regression.

# How to handle autocorrelation

- When time-series data is used, it's important to check for autocorrelation.
- If present, the independence assumption is violated, and we shouldn't trust the $p$-values and confidence intervals that come out of the regression.
- There may be a way to transform the data to remove the autocorrelation: for example, stock prices have strong autocorrelation, but the percentage changes from day to day (or week to week, etc.) do not.

# A final note on regression assumptions

- The purpose of most regression assumptions is ensuring that the *p*-values and confidence intervals are accurate.

## A final note on regression assumptions

- The purpose of most regression assumptions is ensuring that the *p*-values and confidence intervals are accurate.
- But nothing prevents you from building a regression even when the assumptions are violated!

## A final note on regression assumptions

- The purpose of most regression assumptions is ensuring that the *p*-values and confidence intervals are accurate.
- But nothing prevents you from building a regression even when the assumptions are violated!
- Unless the linearity assumption is violated, the regression equation may still be useful for making predictions.