Non-independence due to Time Correlation
(Chapter 14)

• When we model the mean structure with ordinary least squares, the mean structure explains the general trends in the data with respect to our dependent variable and the independent variables.

The leftover *noise* or errors are assumed to have no pattern (we have diagnostic plots to check this).

For one thing, the errors are assumed to be independent.

• Suppose observations have been collected over time, and observations taken closer in time are more alike than observations taken further apart in time.
This is a time-correlation situation. And we can see the correlation in the errors by plotting the residuals against time.

- **Example**: Time as independent variable

The following scatterplot shows a positive linear trend in $Y$ with respect to time for $\text{Time} = 1, 2, 3, \ldots, 50$.

Let’s look at the ordinary least squares fit...
There is a pattern in the residuals suggesting residuals near to each other are similar (positively correlated).

If a residual is positive, there’s a good chance it’s neighboring residual is also positive.

A lag plot of the residuals gives us information on this...

residual: $e_i$
previous residual in time: $e_{i-1}$
Plotting each residual against the previous residual:

So, there is positive correlation in the lag residuals.

The assumption of independence is violated with respect to the assumption of OLS.

- We can instead move away from OLS, and incorporate this correlation into our modeling...
Autocorrelation

• Autoregressive model: model a series in terms of its own past behavior.

• The first-order autoregressive model, AR(1)

\[ Y_t = \beta_0 + \beta_1 x_t + \epsilon_t \quad \text{for} \quad t = 1, \ldots, T \]

with \( \epsilon_t = \rho \epsilon_{t-1} + u_t \)

and \( u_t \sim N(0, \sigma^2) \)

\( |\rho| < 1 \) is the autocorrelation parameter, it tells how strongly the sequential observations are correlated

The \( t^{th} \) and the \( (t - j)^{th} \) are also correlated, but not as strongly:

\( \text{corr}(\epsilon_t, \epsilon_{t-j}) = \rho^j \)
A simulation of AR(1) data from $n = 50$ uniformly spaced time points with a positive linear trend (with $\beta_1 = 2$) can bring insight into the AR(1) process:

```r
## Generate x-values:
> n=50
> time=1:n

## Assign parameters:
> sigma=3
> rho=.95
> beta=2

## Get start point for time series
## and allocate space for vectors:
> y=rep(0,n)
> e=rep(0,n)
> e[1]=rnorm(1,0,sigma)

## Use AR(1) process to sequentially generate y-values:
> for (i in 2:n){
    e[i]=rho*e[i-1]+rnorm(1,0,sigma)
    y[i]=beta*time[i]+e[i]
}
```

*The data for the plots on the previous pages were made from this code.*
• There is also a test for time-correlated errors called the Durbin-Watson test. It actually looks for AR(1) errors, and uses $H_0 : \rho = 0$ vs. $H_A : \rho \neq 0$.

The test statistic:

$$d = \frac{\sum_{t=2}^{n}(e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2}$$

A small $d$ indicates positive autocorrelation. And $d = 2$ suggests no positive autocorrelation.

• Testing the simulated AR(1) data:

```r
> library(car)
> durbin.watson(lm.out)

<table>
<thead>
<tr>
<th>lag</th>
<th>Autocorrelation</th>
<th>D-W Statistic</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.852054</td>
<td>0.2674768</td>
<td>0</td>
</tr>
</tbody>
</table>

Alternative hypothesis: rho != 0

⇒ Reject $H_0$, there is positive correlation.
The mean structure in the AR(1) is the same as OLS, but we model the errors differently.

\[ Y = X\beta + \epsilon \quad \text{with} \quad \epsilon \sim N_n(0, \Sigma) \]

and \( V(Y) = \Sigma \)

\[
\begin{bmatrix}
\kappa & \kappa\rho & \kappa\rho^2 & \cdots & \kappa\rho^{n-2} & \kappa\rho^{n-1} \\
\kappa\rho & \kappa & \kappa\rho & \cdots & \kappa\rho^{n-3} & \kappa\rho^{n-2} \\
\vdots & & & & & \\
\kappa\rho^{n-2} & \kappa\rho^{n-3} & \kappa\rho^{n-4} & \cdots & \kappa & \kappa\rho \\
\kappa\rho^{n-1} & \kappa\rho^{n-2} & \kappa\rho^{n-3} & \cdots & \kappa\rho & \kappa
\end{bmatrix}_{n \times n}
\]

\( \text{Var}(\epsilon_t) = \kappa = \frac{\sigma^2}{1-\rho^2} \)

And we again have a Generalized Least Squares estimation for \( \hat{\beta} \).

\[ \hat{\beta} = (X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}Y \]

* Notice the similarity to the OLS form, but now with the \( \Sigma^{-1} \).
• **Example:** Daily value of stock

  – The dataset *soccho* for this example shows the value of 1 unit of CREF Social Choice stock fund on each day of a year starting on 10/21/99.
  – We’re interested in fitting a linear model over time.
  – But, the independence assumption for OLS is probably violated.
  – We can use the Durbin-Watson test to determine whether this is true.
  – If so, we will fit an AR(1) model to the data.

```r
> head(soccho)
    account unitval date
1  CREFsoci   88.7151 10/21/99
2  CREFsoci   89.4194 10/22/99
3  CREFsoci   89.4194 10/23/99
4  CREFsoci   89.4194 10/24/99
5  CREFsoci   89.1719 10/25/99
6  CREFsoci   88.7471 10/26/99
```
As there are weekend days in the data set, we will first remove these:

> n=length(soccho$unitval)
> n
[1] 365

## 10/21/99 is a Thursday, get indices for removal:
> a1=(seq(1,365,7)+3)
> a1=a1[-length(a1)]
> a2=(seq(1,365,7)+4)
> a2=a2[-length(a2)]
> a=sort(c(a1,a2))

## Subset data down to weekdays:
> day.values=soccho$unitval[-a]
> day=1:length(day.values)

> length(day.values)
[1] 261

> plot(day,day.values,pch=16)
It’s pretty apparent that there is time-based correlation in the data, but we will fit a regular linear model assuming independence and then test for correlation over time.
```r
> lm.out=lm(day.values~day)
> summary(lm.out)

Coefficients:
               Estimate  Std. Error   t value  Pr(>|t|)  
(Intercept)   92.653292   0.235187   393.95   <2e-16 *** 
day           0.027864   0.001556    17.90   <2e-16 *** 
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

Residual standard error: 1.894 on 259 degrees of freedom
Multiple R-Squared: 0.5531, Adjusted R-squared: 0.5514
F-statistic: 320.6 on 1 and 259 DF,  p-value: < 2.2e-16
```
A plot of the residuals vs. fitted also show the time-based correlation.

```r
> plot(lm.out$fitted.values, lm.out$residuals, pch=16)
> abline(h=0)
```

Residuals that are positive tend to be near other positive residuals, and vice versa for negative residuals.
This is more apparent in a lag plot where we plot a residual vs. its neighboring residual: $e_i$ vs. $e_{i-1}$

> lag.plot(lm.out$residuals, do.lines=FALSE)

There is a positive correlation in the lag residuals (residuals tend to be more like their near neighbors).
We can use the **Durbin-Watson test** to formally test for time dependence (uses the relationship between $e_i$ and $e_{i-1}$).

```r
> library(car)
> durbin.watson(lm.out)

<table>
<thead>
<tr>
<th>lag</th>
<th>Autocorrelation</th>
<th>D-W Statistic</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.9025303</td>
<td>0.1684356</td>
<td>0</td>
</tr>
</tbody>
</table>

Alternative hypothesis: rho != 0

The test strongly rejects the null of independence ($H_0: \rho = 0$).

We will fit a first-order autoregressive model to the data, or an AR(1).
• **Fitting the AR(1) model**

The `gls` function [generalized least squares] in the `nlme` library [non-linear mixed effects] fits regression models with a variety of correlated-error and non-constant error-variance structures.

```r
> library(nlme)

## The ~1 below says the data is in order by time.
> gls.out=gls(day.values~day,
  correlation=corAR1(form = ~1))

> summary(gls.out)
Generalized least squares fit by REML
Model: day.values ~ day
Data: NULL

       AIC      BIC      logLik
Phi 615.644 629.8713  -303.822

Correlation Structure: AR(1)
Formula: ~1
Parameter estimate(s):
  Phi
0.9412842
```
Coefficients:

<table>
<thead>
<tr>
<th>Value</th>
<th>Std. Error</th>
<th>t-value</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>92.12831</td>
<td>1.4044915</td>
<td>65.59549</td>
</tr>
<tr>
<td>day</td>
<td>0.02897</td>
<td>0.0090084</td>
<td>3.21585</td>
</tr>
</tbody>
</table>

Residual standard error: 2.26733
Degrees of freedom: 261 total; 259 residual

*Day* is a significant linear predictor for stock price.

\( \hat{\rho} = 0.9413 \), and sequential observations are strongly correlated.
• Comments:

1. When you have many covariates, you can plot the residuals from the OLS fitted model against time as a time-correlation diagnostic. If there is time-correlation, this plot will show a pattern rather than a random scatter.

2. Including *time* as a predictor does not necessarily remove time-correlated errors. As in the *soccho* example, time was a predictor in the OLS model, which meant there was a general trend over time, but there was still correlation in the errors after *time* was included.