1. How can you tell if there is serial correlation?
2. AR to model serial correlation.
3. Ignoring serial correlation.
4. GLS.
5. Projects.

1) Identifying serial correlation.

Plot Y_t versus Y_{t-1}. See p307 bottom.
Better yet, look at the correlogram. See p308.
Often the correlogram starts out high and gradually decays.
For iid observations Y_i with mean 0 and variance σ^2, ρ_k = 0, for $k \neq 0$.
$r_k \sim -1/n$, +/- about $2 / \sqrt{n}$, for $k \neq 0$.
So for confidence bounds, you can use $-1/n \pm 1.96/\sqrt{n}$.
Correlogram values outside the confidence bounds indicate statistically significant serial correlations.

2) AR to model serially correlated errors.
Suppose you have Y and 4 explanatory variables, recorded in time.
We would like a model of the form
$Y_t = \beta_0 + \beta_1 X_{1,t} + \ldots + \beta_4 X_{4,t} + \varepsilon_t$, where ε_t are not iid.
Instead, ε_t are autoregressive (AR).
AR(1): $\varepsilon_t = \alpha_1 \varepsilon_{t-1} + \delta_t$, where δ_t are iid, or more generally,

AR(p): $\varepsilon_t = \alpha_1 \varepsilon_{t-1} + \alpha_2 \varepsilon_{t-2} + \ldots \alpha_p \varepsilon_{t-p} + \delta_t$.

Motivation: ε_t depends on ε_{t-1}, ε_{t-2}, etc. (continuity, esp if α_1 is close to 1 .
For an AR(1) process, the autocorrelation goes quickly down to zero. The autocorrelation $\rho_k = |\alpha_1|^k$ for $k > 0$.
AR(0): ε_t are iid. For AR(0), the autocorrelation ρ_k is zero at any nonzero time lag.

```
par(mfrow=c(2,3))
d = rnorm(100)
plot(1:100,d,xlab="time",ylab="y",main="iid noise")
## y = d[2:100] + 0.8*d[1:99]
y = rep(0,100)
y[1] = d[1]
for(i in 2:100){
y[i] = 0.9*y[i-1] + d[i]
}
plot(1:100,y,xlab="time",ylab="y",main="AR(1) noise")
## y2 = d[3:100] + 0.8*d[2:99] + 0.4[1:98]
y2 = rep(0,100)
y2[1:2] = d[1:2]
for(i in 3:100){
y2[i] = 0.6*y2[i-1] + .2*y2[i-2] + d[i]
}

plot(1:100,y2,xlab="time",ylab="y",main="AR(2)
noise")
acf(d)
acf(y)
acf(y2)
```

3) Ignoring serial correlation.
When serial correlation is present, using OLS to estimate β, the estimates will be unbiased, but with incorrect estimates of the variance, so CIs and tests will not be valid. See p311.

4) GLS.
When serial correlation is present, to get optimal estimates of β, you can use generalized least squares, GLS. If you know the covariance Σ between the observations, or can estimate it using the correlogram, then you can estimate β using GLS. See the top of p313.
If ∑ = S S^T, then you can define generalized residuals S^{-1} e, as described on p317. These should be uncorrelated. See p319.
GLS can be implemented using the gls() function in R.

5) Projects.
The project is due by email to me (frederic@stat.ucla.edu) before 11pm on Sat, June 5.
a) Find a dataset on something you know more than the average person about, or are extremely interested in. It could be on the web or you can collect it yourself. Your topic may be non-academic, and should be based on a genuine interest of yours, such as one of your hobbies or extra-curricular activities.

b) Your response variable, Y, should be non-negative-integer-valued.
c) You should also have at least 2 explanatory variables, and at least 20 observations (n).
d) Analyze your data using OLS, Poisson regression, kernel regression, and one other method (logistic regression, least trimmed squares, GLS, m-estimation, or LASSO).
e) Your written report should be 3-5 pages of written text, in a reasonable font with reasonable spacing, followed by as many figures and tables as you'd like at the end, in an appendix.
f) Do not include the figures in your text. Instead, just have all the figures at the end.
g) There is no need to explain in your text what the different methods you use are. Instead, focus on interpreting your results. Write your paper to an audience that knows statistics but does not know about your application.
h) Your report should contain an introduction (1/2 to 1 page) explaining why your data are interesting or important, a results section (2-3 pages), in which you comment on each of your figures, tables, and results, and a conclusion (1/2 to 1 page), in which you summarize your main findings and describe problems with your dataset and analysis.
i) In the results section, be sure to explain the main interesting features of your figures. Any figures that you do not comment on in your text should not be part of your report at all.
j) Do not submit R code.

k) In your conclusion, you are encouraged to speculate in interpreting your results and, in particular, to speculate on how any problems with your data collection or analysis may have influenced your results. You can speculate about what other methods might be better, what models might be better, or what variables might be more appropriate if you were going to study this further in the future.

l) For each of the methods you use, show your residuals and analyze goodness-of-fit as appropriate.

m) Show your fitted models. Point out the most interesting things in terms of the data. Point out particular outliers or interesting observations.