36-617: Applied Linear Regression

Generalized Least Squares

Brian Junker
132E Baker Hall
brian@stat.cmu.edu

Announcements

- Project 01: I am grading them this week!
- Project 02 will come out in a week or two.
  - The schedule may be a bit compressed compared to project 01
- Quiz on Sheather Chapter 9 today!
- HW08 Due Wednesday this week
- HW09 Due *Next* Wednesday
- On Wednesday I will begin talking about hierarchical mixed effects models (a.k.a. multilevel models [MLM], hierarchical linear models [HLM], linear mixed effects regression [LMER], etc…)
  - Please start Sheather 10.1 (not 10.2) for Wednesday’s lecture.
- There is a brief description of 36-663 (Hierarch. Models) in the week09 folder.

Outline

- Review ML -> OLS
- What happens to the theory when $\epsilon \sim N(0, \Sigma)$
- Estimating Σ
- Applications:
  - WLS – unequal sample sizes
  - Time series correlation: AR(1), etc.

Review: ML/LS Estimates

$$\boxed{y = X\beta + \epsilon,\ \epsilon \sim N(0, \sigma^2 I)}$$

- $\hat{\beta} = (X^T X)^{-1} X^T y$
- $\hat{y} = X\hat{\beta} = X(X^T X)^{-1} X^T y = Hy$
- The "residual SD" is the square root of

$$\hat{\sigma}^2 = \tfrac{1}{n-k} \sum_{i=1}^{n} (y_i - X_i\hat{\beta})^2 = \tfrac{1}{n-k} (y - X\hat{\beta})^T (y - X\hat{\beta})$$

- Basic distribution properties on the next slide…

Review: $\hat{\beta}$, H, $\hat{y}$ & $\hat{e}$ for ML/LS

$$Y \sim N(\mu, \Sigma) \Rightarrow AY \sim N(A\mu, A\Sigma A^T)$$

$$\boxed{y \sim N(X\beta, \sigma^2 I), \quad \hat{\beta} = (X^TX)^{-1}X^Ty}$$

$$\Rightarrow E[\hat{\beta}] = \beta, \quad \mathrm{Var}(\hat{\beta}) = (X^TX)^{-1}\sigma^2$$

$$\boxed{\hat{\beta} \sim N(\beta, (X^TX)^{-1}\sigma^2)}$$

$$\boxed{H = X(X^TX)^{-1}X^T}$$

$$\Rightarrow E[\hat{y}] = E[Hy] = X\beta, \quad \mathrm{Var}(\hat{y}) = \mathrm{Var}(Hy) = H\sigma^2$$

$$\boxed{\hat{y} \sim N(X\beta, H\sigma^2)}$$

$$E[\hat{e}] = E[(I-H)y] = 0,$$

$$\mathrm{Var}(\hat{e}) = \mathrm{Var}((I-H)y) = (I-H)\sigma^2$$

$$\boxed{\hat{e} \sim N(0, (I-H)\sigma^2)}$$

Generalized Least Squares

- Suppose instead of $y = X\beta + \epsilon,\ \epsilon \sim N(0, \sigma^2 I)$, we have $y = X\beta + \epsilon,\ \epsilon \sim N(0, \Sigma)$
- Then $y \sim N(X\beta, \Sigma)$, which by our earlier definition (week 4!) means that

$$\Sigma^{-1/2}(y - X\beta) \sim N(0, I)$$

- More precisely, we will let $\Sigma = \sigma^2 W$, where W is symmetric & positive definite, so there exists a lower-triangular matrix S such that[1] $SS^T = W$ and hence

$$S^{-1}(y - X\beta) \sim N(0, \sigma^2 I)$$

[1]This is called a "Cholesky decomposition".

Generalized Least Squares, cont'd…

- Since $S^{-1}(y - X\beta) \sim N(0, \sigma^2 I)$, we know

$$\underbrace{S^{-1}y}_{y^*} \quad \sim \quad N(\underbrace{S^{-1}X\beta}_{X^*\beta}, \sigma^2 I)$$

- So $y = X\beta + \epsilon,\ \epsilon \sim N(0, \Sigma)$ is equivalent to

$$y^* = X^*\beta + \epsilon^*,\ \epsilon^* \sim N(0, \sigma^2 I)$$

with solution

$$\begin{aligned}\hat{\beta}^* &= ({X^*}^T X^*)^{-1} {X^*}^T y^* \\ &= (X^T (S^{-1})^T (S^{-1}) X)^{-1} X^T (S^{-1})^T (S^{-1}) y \\ &= (X^T W^{-1} X)^{-1} X^T W^{-1} y\end{aligned}$$

- $\hat{\beta}^* = \mathrm{argmin}_\beta \mathrm{RSS}^*,\ \mathrm{RSS}^* = (y^* - X^*\beta)^T (y^* - X^*\beta) = (y - X\beta)^T W^{-1} (y - X\beta)$, and $\hat{\sigma}^2 = \mathrm{RSS}^*/(n - df)$

(where we use the facts that $(A^{-1})^T = (A^T)^{-1}$ and $(AB)^{-1} = (B^{-1})(A^{-1})$ so that $(S^{-1})^T(S^{-1}) = (S^T)^{-1}(S^{-1}) = (SS^T)^{-1} = \Sigma^{-1}$)

$\hat{\beta}^*$, H^*, $\hat{y}^*$ & $\hat{e}^*$ under GLS

Under $y = X\beta + \epsilon$, $\epsilon \sim N(0, \Sigma)$, with $\Sigma = \sigma^2 W$

$$
\begin{aligned}
y &\sim N(X\beta, \Sigma), \quad \text{i.e. } y^* \sim N(X^*\beta, \sigma^2 I) \\
& \quad (y^* = S^{-1}y, \; X^* = S^{-1}X)
\end{aligned}
$$

We get

$$
\begin{aligned}
\hat{\beta}^* &\sim N(\beta, \sigma^2({X^*}^T X^*)^{-1}) \; = \; N(\beta, \sigma^2(X^T W^{-1} X)^{-1}) \\
H^* &= S^{-1}X(X^T W^{-1} X)^{-1} X^T (S^{-1})^T \\
\hat{y}^* &= X^* \hat{\beta}^* \; \sim \; N(X^*\beta, \sigma^2 H^*) \\
\hat{e}^* &= y^* - \hat{y}^* \; \sim \; N(0, \sigma^2(I - H^*))
\end{aligned}
$$

Not exactly what we want, if we want to predict y and not y*

To predict y from GLS estimates…

- Rather than $\hat{y}^* = X^*\hat{\beta}^*$, we could use $\hat{y} = X\hat{\beta}^*$
- Using the results from the previous slide, we get

$$
\begin{aligned}
E[\hat{y}] &= E[X\hat{\beta}^*] = XE[\hat{\beta}^*] = X\beta \\
\text{Var}(\hat{y}) &= \text{Var}(X\hat{\beta}^*) = X\text{Var}(\hat{\beta}^*)X^T \\
&= \sigma^2 X(X^T W^{-1} X)^{-1} X^T = \sigma^2 S H^* S^T
\end{aligned}
$$

So, after some calculation,

$$
\begin{aligned}
\hat{y} &\sim N(X\beta, \sigma^2 S H^* S^T) \\
\hat{e} &= y - \hat{y} \sim N(0, \sigma^2 S(I - H^*)S^T)
\end{aligned}
$$

Aside: A recommendation…

- Base casewise diagnostic plots on $y^*, \hat{y}^* = X^*\hat{\beta}^*$ and $\hat{e}^* = y^* - \hat{y}^*$ from the model

$$y^* = X^*\beta + \epsilon^*,\ \epsilon \sim N(0, \sigma^2 I)$$

(where $y^* = S^{-1}y$ & $X^* = S^{-1}X$)

- For prediction, better off using $y, \hat{y} = X\hat{\beta}^*$ and $\hat{e} = y - \hat{y}$ from the original model

$$y = X\beta + \epsilon,\ \epsilon \sim N(0, \Sigma)$$

Estimating Σ …

- For $y_i = X_i\beta + \epsilon_i,\ i = 1, \ldots, n,$

$$\Sigma = \mathsf{Var}\begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \cdots & \sigma_n^2 \end{bmatrix}$$

where $\sigma_i^2 = \mathsf{Var}\left(\epsilon_i\right)$ and $\sigma_{ij} = \mathsf{Cov}\left(\epsilon_i, \epsilon_j\right)$

- Want to estimate *n(n+1)/2* parameters with n observations… need constraints… ***Applications!***

Application: Weighted Least Squares (WLS)

- In many situations we know Σ is diagonal, and we know the structure of Σ, up to a constant multiple ... *For example:*
  - The y_i's are averages of n_i observations each, so that $Var(y_i) = \sigma^2/n_i$; or…
  - $Var(y_i)$ is proportional to the k^{th} predictor: $Var(y_i) = \sigma^2 x_{ki}$; or…
  - Etc…
- In cases like this,

$$\Sigma = \begin{bmatrix} \mathrm{Var}(y_1) & 0 & \cdots & 0 \\ 0 & \mathrm{Var}(y_2) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \mathrm{Var}(y_n) \end{bmatrix} = \sigma^2 \begin{bmatrix} w_1 & 0 & \cdots & 0 \\ 0 & w_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & w_n \end{bmatrix} = \sigma^2 W$$

where $w_i = 1/n_i$, or $w_i = x_{ki}$, etc., and we have just 1 parameter to estimate!

WLS Example[1]

Following are data from an experiment to study the interaction of certain kinds of elementary particles on collision with proton targets. The experiment was designed to test certain theories about the nature of the strong interaction. The cross-section (`crossx`) variable is believed to be linearly related to the inverse of the energy (`energy` - has already been inverted). At each level of the momentum, a very large number of observations were taken so that it was possible to accurately estimate the standard deviation of the response (`sd`).

	momentum	energy	crossx	sd
1	4	0.345	367	17
2	6	0.287	311	9
3	8	0.251	295	9
4	10	0.225	268	7
5	12	0.207	253	7
6	15	0.186	239	6
7	20	0.161	220	6
8	30	0.132	213	6
9	75	0.084	193	5
10	150	0.060	192	5

[1]From http://www.biostat.jhsph.edu/~iruczins/teaching/jf/ch5.pdf

Fitting the WLS model

```
> strongx <-
+ read.table(stdin(),header=T)
0:   momentum energy crossx sd
1: 1          4   0.345     367 17
2: 2          6   0.287     311  9
3: 3          8   0.251     295  9
4: 4         10   0.225     268  7
5: 5         12   0.207     253  7
6: 6         15   0.186     239  6
7: 7         20   0.161     220  6
8: 8         30   0.132     213  6
9: 9         75   0.084     193  5
10: 10      150   0.060     192  5
11:
> summary(wls.1 <- lm(crossx ~
+ energy, data=strongx,
+ weights=sd^(-2)))
```

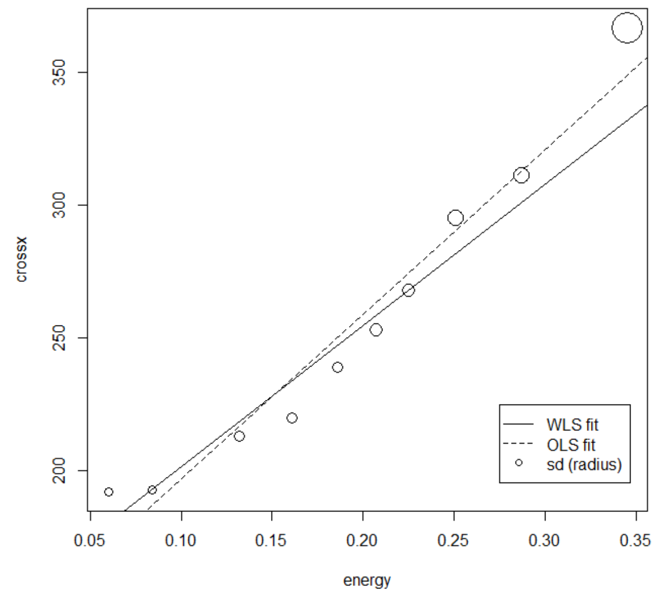
We give lm() the diagonal elements of W^{-1}, without the unknown residual variance σ^2

```
Call:
lm(formula = crossx ~ energy, data = strongx,
weights = sd^(-2))

Weighted Residuals:
    Min      1Q  Median      3Q     Max
-2.3230 -0.8842  0.0000  1.3900  2.3353

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  148.473      8.079   18.38 7.91e-08
energy       530.835     47.550   11.16 3.71e-06
---

Residual standard error: 1.657 on 8 degrees of
freedom
Multiple R-squared:  0.9397,
Adjusted R-squared:  0.9321

F-statistic: 124.6 on 1 and 8 DF,
p-value: 3.71e-06
```

Estimate of the residual SD, σ

Comparing with OLS…

```
> summary(ols.1 <- lm(crossx ~ energy,
data=strongx))
[...]
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   135.00      10.08    13.4 9.21e-07
energy        619.71      47.68    13.0 1.16e-06
[...]
Residual standard error: 12.69 on 8 degrees
of freedom
Multiple R-squared:  0.9548,
Adjusted R-squared:  0.9491
F-statistic: 168.9 on 1 and 8 DF,
p-value: 1.165e-06
> with(strongx,
+ plot(energy, crossx, cex=sd/4))
> abline(wls.1); abline(ols.1, lty=2)
> legend(0.275,225,legend =
+ c("WLS fit","OLS fit","sd (radius)"),
+ lty=c(1,2,NA), pch=c(NA,NA,1))
```



- OLS seems to follow the data better, ***<u>but…</u>***
- WLS weights observations with lower variance more, in minimizing RSS* =

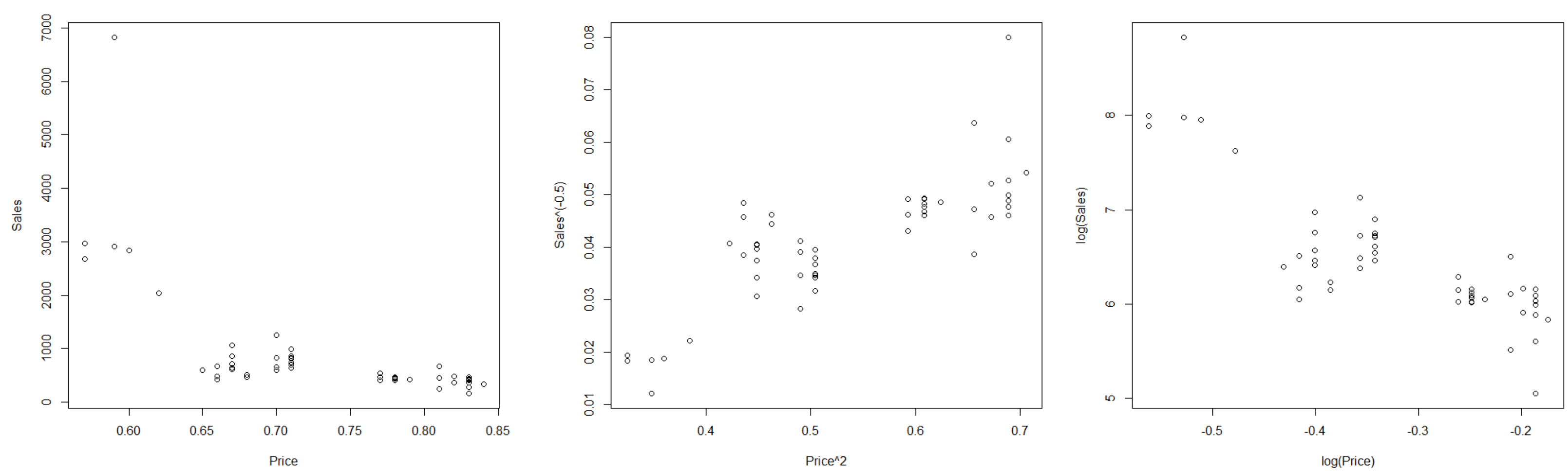
$$(y - X\beta)^T W^{-1}(y - X\beta)$$
$$= \sum_i 1/w_i(y_i - X_i\beta)^2$$

Application: serial correlation

- If the order in which the data came is important then it is worth checking to see if any typical time series models for Σ apply.
  - `confoods2.txt` contains weekly sales data for 52 weeks, for a canned food product (Sheather, Ch 3 & Ch 9). The goal is to understand how `Price` and `Promotion` (0/1 dummy) affect `Sales`
  - Because the data come sequentially in time, and customers' behavior in one week is *unlikely* to be independent of their behavior the next week, it is worth considering serial correlation in the data.

Aside: Transformations

- The Box-Cox method suggests replacing Sales with $(\text{Sales})^{-1/2}$ and replacing Price with $(\text{Price})^2$.
- However, this is harder to explain to consultee or collaborator, so we also try log transformations:



(Review: Interpreting log transform)

- Since $\log(1+x) \approx x$, if we subtract the two regression relations

$$\begin{array}{rcl} y_1 & = & \beta_0 + \beta_1 \log(x) + \epsilon_1 \\ y_2 & = & \beta_0 + \beta_1 \log(x + \Delta x) + \epsilon_2 \\ \hline \Delta y = y_2 - y_1 & = & \beta_1 \log\left(1 + \frac{\Delta x}{x}\right) + \text{(error)} \\ \text{so } \Delta y & \approx & \beta_1 \frac{\Delta x}{x} + \text{(error)} \end{array} \qquad (1)$$

Hence β_1 is the approximate* change in y induced by a relative change $\Delta x/x$ in x. E.g. if the relative change is $\Delta x/x = 0.01$, then y changes by approximately $\beta_1 \cdot (0.01)$.

- If we replace y with $\log(y)$, then equation (1) becomes

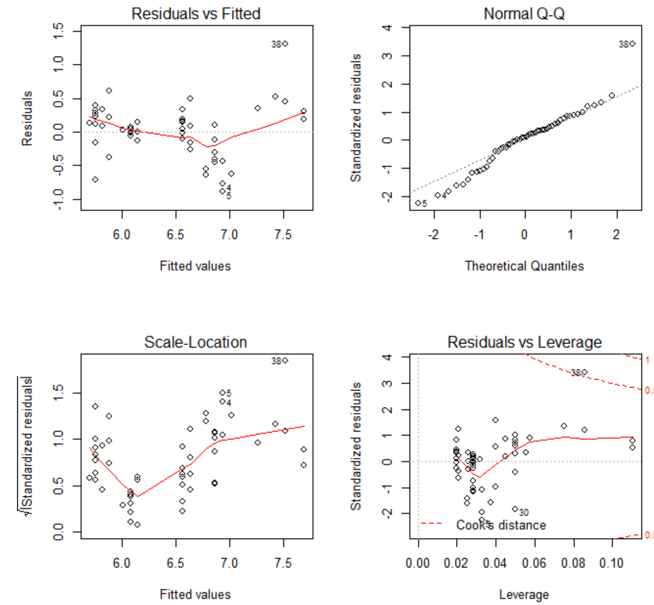
$$\begin{array}{rcl} \frac{\Delta y}{y} & \approx & \beta_1 \frac{\Delta x}{x} + \text{(error)} \\ \frac{\Delta y}{y} \cdot 100\% & \approx & \beta_1 \frac{\Delta x}{x} \cdot 100\% + \text{(rescaled error)} \end{array}$$

If we set $\Delta x/x = 0.01$ again, we see that, since $\Delta x/x \cdot 100\% = 1\%$, β_1 is now approximately* the percent change in y for a 1% change in x.

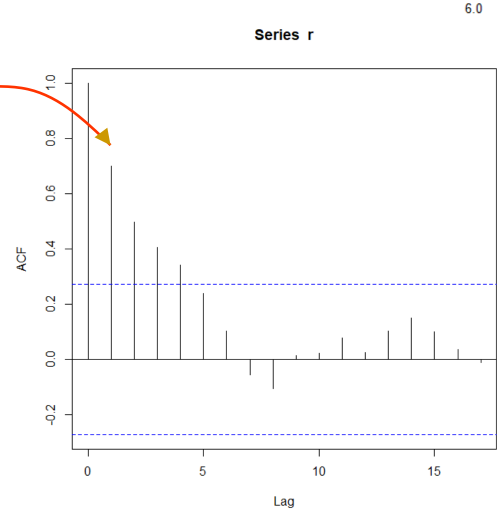
*These "approximate" statements can be made exact by taking expected values everywhere. See "log xform and percent interpretation.pdf" in the same folder as this lecture.

Autocorrelation of residuals…

```
> summary(lm.1 <- lm(log(Sales) ~ log(Price))
[...]
                  Est        SE        t Pr(>|t|)
(Intercept)   4.8029  0.1744   27.53  < 2e-16
log(Price)   -5.1477  0.5098  -10.10 1.16e-13
[...]
Residual standard error: 0.4013 on 50 degrees
of freedom
Multiple R-squared:  0.671,
Adjusted R-squared:  0.6644
F-statistic:   102 on 1 and 50 DF,
p-value: 1.159e-13
```




```
> par(mfrow=c(2,2))
> plot(lm.1)

> r <- resid(lm.1)
> cor(r[-1],r[-length(r)])
[1] 0.717101

> acf(r)
```



- First spike is always 1
- Next spike is lag-1 correlation
- Next is lag-2
- Etc.

AR(1) – Autoregressive order 1 (the simplest autocorrelation model)

- Suppose $\epsilon_i = \rho\epsilon_{i-1} + \nu_i$, $\nu_i \sim N(0, \sigma_\nu^2)$. Then
  - $\sigma_\epsilon^2 = \text{Var}\left(\epsilon_i\right) = \text{Var}\left(\rho\epsilon_{i-1} + \nu_i\right) = \rho^2\sigma_\epsilon^2 + \sigma_\nu^2$, so $\sigma_\epsilon^2 = \frac{\sigma_\nu^2}{1-\rho^2}$
  - $\text{Cov}\left(\epsilon_i, \epsilon_{i-1}\right) = \text{Cov}\left(\rho\epsilon_{i-1} + \nu_i, \epsilon_{i-1}\right) = \rho\sigma_\epsilon^2$
- Thus

$$\text{Cor}\left(\epsilon_i, \epsilon_{i-1}\right) = \frac{\text{Cov}\left(\epsilon_i, \epsilon_{i-1}\right)}{\sqrt{\text{Var}\left(\epsilon_i\right)\text{Var}\left(\epsilon_{i-1}\right)}} = \frac{\rho\sigma_\epsilon^2}{\sigma_\epsilon^2} = \rho$$

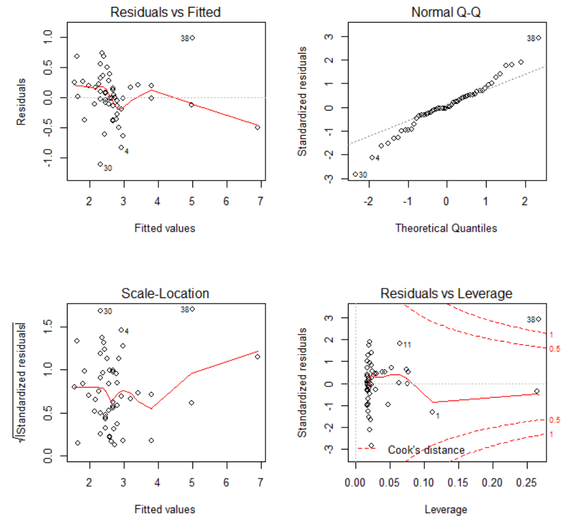
- Similarly, can show $\text{Cor}(\epsilon_i, \epsilon_{i-\ell}) = \rho^{i-\ell}$, $\ell = 0, \ldots, i-1$, and thus

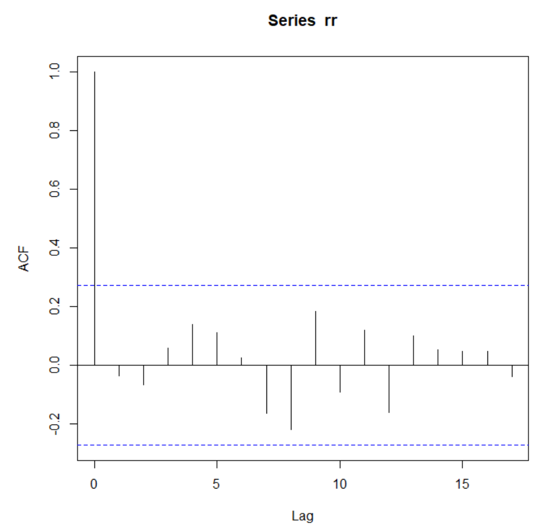
$$\Sigma = \begin{bmatrix} \sigma_\epsilon^2 & \rho\sigma_\epsilon^2 & \cdots & \rho^{n-1}\sigma_\epsilon^2 \\ \rho\sigma_\epsilon^2 & \sigma_\epsilon^2 & \cdots & \rho^{n-2}\sigma_\epsilon^2 \\ \vdots & \vdots & \ddots & \vdots \\ \rho^{n-1}\sigma_\epsilon^2 & \rho^{n-2}\sigma_\epsilon^2 & \cdots & \sigma_\epsilon^2 \end{bmatrix} = \sigma_\epsilon^2 \begin{bmatrix} 1 & \rho & \cdots & \rho^{n-1} \\ \rho & 1 & \cdots & \rho^{n-2} \\ \vdots & \vdots & \ddots & \vdots \\ \rho^{n-1} & \rho^{n-2} & \cdots & 1 \end{bmatrix} = \sigma_\epsilon^2 W$$

Estimation Strategy 1: Plug in estimate of ρ.

```
> rho <- 0.717101 ## estimate of rho
> Sigma <- diag(length(Sales))
> Sigma <- rho^abs(row(Sigma)-col(Sigma))
> S <- chol(Sigma)
> St_inv <- solve(t(S))
> ystar <- St_inv %*% log(Sales)
> X <- model.matrix(lm.1)
> Xstar <- St_inv%*% X
> summary(lm.2 <- lm(ystar ~ Xstar - 1))
[…]
                      Est      SE        t
Xstar(Intercept)    4.5844 0.2090    21.93
Xstarlog(Price)    -5.7976 0.4889   -11.86
[…]
Residual standard error: 0.3949 on 50
degrees of freedom
> par(mfrow=c(2,2))
> plot(lm.2)
> par(mfrow=c(1,1))
> rr <- resid(lm.2)
> acf(rr)
```

Estimate of the residual SD, σ





Estimation Strategy 2: Estimate ρ, β together using maximum likelihood

```
> library(nlme)
> summary(gls.1 <- gls(log(Sales) ~ log(Price), correlation=corAR1(),method="ML"))
AIC     BIC     logLik
  20.02102 27.826 -6.010511
Correlation Structure: AR(1)
 Formula: ~1
 Parameter estimate(s):
      Phi
0.7406252

Coefficients:
                Value Std.Error   t-value p-value
(Intercept)  4.577421 0.2161245  21.17956       0
log(Price)  -5.815621 0.4882088 -11.91216       0

Standardized residuals:
       Min         Q1        Med         Q3        Max
-2.3476869 -0.4815969  0.1580394  0.6130209  2.9496345

Residual standard error: 0.401166
Degrees of freedom: 52 total; 50 residual
```

ML Estimate of ρ

Was the parameter rho needed?

```
> logLik(lm.1)
'log Lik.' -25.2911 (df=3)
> logLik(gls.1)
'log Lik.' -6.010511 (df=4)
> (chisq <-
+ -2*(logLik(lm.1) - logLik(gls.1)))
[1] 38.56117
> pchisq(chisq,df=1,lower.tail=F)
[1] 5.30641e-10
```

Estimate of the residual SD, σ

Summary

- Review ML -> OLS
- What happens to the theory when $\epsilon \sim N(0, \Sigma)$
- Estimating Σ
- Applications:
  - WLS – unequal sample sizes
  - Time series correlation: AR(1), etc.