SERIAL CORRELATION

Feb 23 2015

Charlotte Wickham  stat512.cwick.co.nz
Overview of regression

A model for the mean:

\[ \mu\{Y \mid X_1, \ldots, X_p\} = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p \]

+ assumptions:

There is a Normally distributed subpopulation at each combination of explanatory variables values.

The means of the subpopulations fall on the line/surface defined above ( \( \mu\{Y \mid X_1, \ldots, X_p\} \) )

The subpopulation standard deviations are all equal to \( \sigma \)

The selection of an observation from one subpopulation is independent of the selection of any other observation.

The deviation of an observation from the mean, is independent of the deviation from the mean for any other observation.

Ch. 15 & 16
Serial Correlation

The multiple regression tools rely on the observations being independent (after accounting for the effects of the explanatory variables).

Often when measurements are made at adjacent points in time or space the observations are correlated.
Clear cutting (stripping the land of all vegetation) is one method of logging Douglas Fir.

Water quality in streams is adversely affected by clear cutting.

An alternative is patch cutting.

Observe two watersheds, one from patch-cut and one undisturbed.

Measure water quality by nitrates.

Is the mean nitrate level higher for the patch cut watershed?
Your turn

Ignoring of the appropriateness of regression, how would you answer the question of interest?

Is the mean nitrate level higher for the patch cut watershed compared to undisturbed watershed?

Two sample t-test

\[ \mu_{\text{Nitrate}_1} - 3 = \beta_0 + \beta_1 \cdot PC \quad \beta_1 = 0 \]
Nitrates (NO3-N) in runoff from patch-cut and undisturbed watersheds, for five years after logging.

Graph showing nitrate levels (ppm) over weeks post-logging.
After transformation, and subtracting the sample average from each.
Both are centered around 0.
Notice the “runs” of observations above or below the mean.
Serial correlation
a.k.a autocorrelation

Positive serial correlation: an observation on one side of the mean tends to be followed by another observation on the same side of the mean.

Negative serial correlation: an observation on one side of the mean tends to be followed by another observation on the opposite side of the mean.
The “runs” make averages of subsamples much more variable about the mean than for uncorrelated series.

The observations also exhibit less variability than expected without correlation.

The usual SE on the average formula, $\frac{s}{\sqrt{n}}$, will overestimate the precision when there is positive autocorrelation.
Two solutions

1. Adjust standard errors to be more appropriate.

2. Filter variables to remove correlation.

Either way you need to estimate the extent of the correlation (and make an assumption about its structure).

More advanced methods explicitly model the correlation.
We see a positive correlation!

the transformed nitrate concentration from the previous week
Estimating serial correlation

essentially, the sample correlation between current and previous residual

\[ r_1 = \frac{c_1}{c_0} \]

\[ c_1 = \frac{1}{n-1} \sum_{t=2}^{n} \text{res}_t \times \text{res}_{t-1} \]

\[ c_0 = \frac{1}{n-1} \sum_{t=1}^{n} (\text{res}_t)^2 \]

covariance of current and previous residual

variance of residuals

Or in R:

```r
with(case1501, acf(Nocut))$acf[,1]
```

\[ \begin{array}{c}
[1,] 1.0000000000 \\
[2,] 0.7440345412 \\
[3,] 0.6220669301 \\
[4,] 0.4931940407
\end{array} \]

\[ r_1 \] 1st serial correlation coefficient

\[ r_2 \] 2nd serial correlation coefficient

correlation between now & two observations ago
1. An adjusted SE on the sample average

\[ SE_{\bar{Y}} = \sqrt{\frac{1 + r_1}{1 - r_1}} \frac{s}{\sqrt{n}} \]

where \( r_1 \) is the first serial correlation coefficient.

Appropriate under the autoregressive model of order 1 (\textbf{AR}(1)):

- The series is measured at equally spaced times
- Let \( v \) be the long run series mean, then

\[ \mu \{ Y_t - v \mid \text{past history} \} = \alpha ( Y_{t-1} - v ) \]

where \( \alpha \) is the first order autocorrelation coefficient.
A two sample comparison

Do the usual two sample procedure, but adjust the standard error:

\[
\bar{Y}_C - \bar{Y}_U = 2.016 - 1.905 = 0.111
\]

\[
\text{SE}_{\bar{Y}_C - \bar{Y}_U} = \sqrt{\frac{1 + r_1}{1 - r_1}} \times sp \sqrt{\frac{1}{n_c} + \frac{1}{n_u}}
\]

\[
= \sqrt{\frac{1 + 0.644}{1 - 0.644}} 
\times 1.183 \sqrt{\frac{1}{88} + \frac{1}{88}} = 0.383
\]

pooled = assume the same for both watersheds, and use both sets of data to estimate them
2. Filter variables to remove correlation.

If the AR(1) model is adequate and
\[ \mu\{ Y_t | X_t \} = \beta_0 + \beta_1 X_t \]
Then the filtered variables:
\[ V_t = Y_t - \alpha Y_{t-1} \]
\[ U_t = X_t - \alpha X_{t-1} \]
are related by the same slope:
\[ \mu\{ V_t | U_t \} = \beta_0 (1 - \alpha) + \beta_1 U_t \]
with no serial correlation
Use \( r_1 \) as an estimate for \( \alpha \).
Filter response and explanatory.
Then regress filtered variables.
case1502: Global Temperature

The data are the temperatures (in degrees Celsius) averaged for the northern hemisphere over a full year, for years 1880 to 1987. The 108-year average temperature has been subtracted, so each observation is the temperature difference from the series average.
Your turn

Ignoring of the appropriateness of regression, how would you answer the question of interest?

Is the mean temperature increasing?
Is serial correlation a problem?

```r
> fit_slr <- lm(Temp ~ Year, data = case1502)
> summary(fit_slr)$coef

|                      | Estimate | Std. Error | t value | Pr(>|t|) |
|----------------------|----------|------------|---------|----------|
| (Intercept)          | -8.786714263 | 0.6795783683 | -12.92966 | 1.592721e-23 |
| Year                 | 0.004493603 | 0.0003514301 | 12.78662 | 3.281042e-23 |
```

this will be an underestimate

$r_1 = 0.45$

because there is positive serial correlation
2. Use filtering to get SE

If the AR(1) model is adequate and
\[ \mu\{\text{Temp}_t \mid \text{Year}_t\} = \beta_0 + \beta_1 t \]

**Filtered variables:**
\[ V_t = \text{Temp}_t - r_1 \text{Temp}_{t-1} \]
\[ U_t = t - r_1(t - 1) \]

Regress \( V_t \) on \( U_t \)

```r
> case1502$lag_Year <- c(NA, case1502$Year[-nrow(case1502)])
> case1502$lag_Temp <- c(NA, case1502$Temp[-nrow(case1502)])
>
> # regress filtered variables
> fit_filt <- lm(I(Temp - r1*lag_Temp) ~I(Year - r1*lag_Year) , data = case1502)
> summary(fit_filt)$coef

             Estimate  Std. Error   t value     Pr(>|t|)
(Intercept)  -4.92680691  0.6154922045 -8.004662   1.709667e-12
I(Year - r1 * lag_Year) 0.00460353  0.0005809344  7.924355    2.562586e-12
```
1. Use adjustment to get SE

\[ SE_{\beta_1} = \sqrt{\frac{1 + r_1}{1 - r_1}} \]

SE_{\beta_1}^{slr}

\[
\text{summary(fit_slr)$coef[2, 2]}
\]

\[
\text{sqrt((1+r1)/(1-r1)) $* summary(fit_slr)$coef[, 2]}
\]

(Intercept) 1.1068682408 Year 0.0005723943

(Intercept) -8.786714263 0.6795783683 -12.92966 1.592721e-23
Year 0.004493603 0.0003514301 12.78662 3.281042e-23
Examine for serial correlation in the **residuals**. Not the raw response.
The filtering method extends to multiple explanatory variables.
Testing for serial correlation

Large sample test

\[ Z = \frac{r_1}{\sqrt{n}} \]

If there is no serial correlation, \( Z \) has a Normal distribution.

only appropriate when \( n > 100 \)

Runs test

Count how many runs there are and compare to how many we would expect by chance alone with no serial correlation. non-parametric
Is the AR(1) model adequate?

The primary tool is the PACF plot:

```
pacf(residuals(fit_slr))
```
Partial autocorrelation functions for four different types of time series

- **A. White Noise**: no serial correlation
- **B. Autoregression, Order = 3**: "easy" extension of AR(1)
- **C. Moving Average / ARIMA**: complicated
- **D. Non-Stationary / ARIMA**: probably needs a trend removed