Trends:

Goodness-of-Fit and Residual Analysis

Week III: Video 10

STAT 485/685, Fall 2020, SFU

Sonja Isberg

Review: Trends in Time Series

Suppose our process of interest, $\{Y_t\}$, has some mean function μ_t (which may or may not be a function of t).

We can separate out the mean from the rest of the process by writing:

$$Y_t = \mu_t + X_t$$

where $\{X_t\}$ is the "de-trended" version of the process, i.e. $E(X_t) = 0$.

Review: Trends in Time Series

Suppose our process of interest, $\{Y_t\}$, has some mean function μ_t (which may or may not be a function of t).

We can separate out the mean from the rest of the process by writing:

$$Y_t = \mu_t + X_t$$

where $\{X_t\}$ is the "de-trended" version of the process, i.e. $E(X_t) = 0$.

We have learned about three different types of trends, and how to fit them:

- Constant trend: $\mu_t = \mu$ for all t
- Linear trend: $\mu_t = \beta_0 + \beta_1 t$
- Cyclical/seasonal trend: e.g., $\mu_t = \mu_{t-12}$ for all t
  - Two different models: *seasonal means model* and *cosine trend model*

We estimate the mean at time t by plugging in the $\hat{\beta}$'s to obtain $\hat{\mu}_t$.

Video 10 Learning Objectives

By the end of this video, we should be able to:

- Assess the goodness-of-fit of a time series trend model, using the measures of R^2 and adjusted R^2
- Understand which assumptions are made in R when standard errors and p-values are given for a trend model
- Define the residuals of a trend model
- Interpret several types of residual plots, to assess the adequacy of the model assumptions
- Define the sample autocorrelation function, and interpret its plot to judge whether a process appears to be white noise

Some Things To Be Aware Of

R output usually provides estimates of the standard deviations of the $\hat{\beta}$'s, but relies heavily on some assumptions on $\{X_t\}$ which are usually not true:

- Standard error values assume that $\{X_t\}$ is a white noise process (i.e., iid random variables).
- t-statistics and p-values also assume that the X_t's must be approximately Normal.

So we have to be very careful when interpreting R output!

Measures of Goodness-of-Fit

After fitting the trend, X_t is the remaining unobserved stochastic component. We can estimate each X_t using the **residual**:

$$\hat{X}_t = Y_t - \hat{\mu}_t$$

The **residual standard deviation / residual standard error** is one measure of the goodness-of-fit of a model:

$$s = \sqrt{\frac{1}{n-p}\sum_{t=1}^{n}(Y_t - \hat{\mu}_t)^2},$$

where p is the number of parameters estimated in μ_t. A smaller value of s implies a better fit.

Measures of Goodness-of-Fit

After fitting the trend, X_t is the remaining unobserved stochastic component. We can estimate each X_t using the **residual**:

$$\hat{X}_t = Y_t - \hat{\mu}_t$$

The **residual standard deviation / residual standard error** is one measure of the goodness-of-fit of a model:

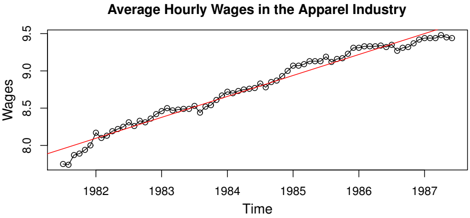
$$s = \sqrt{\frac{1}{n-p}\sum_{t=1}^{n}(Y_t - \hat{\mu}_t)^2},$$

where p is the number of parameters estimated in μ_t. A smaller value of s implies a better fit.

Another measure of goodness-of-fit: R^2 (**coefficient of determination / "Multiple R-squared"**). It is the fraction of the variation in the series that is explained by the estimated trend. So: $0 \leq R^2 \leq 1$.

Adjusted R-squared is like R^2, but slightly adjusted to account for the number of estimated parameters.

R Example



R Code:

```
data(wages)
model.lin <- lm(wages~time(wages))
summary(model.lin)
```

Output:
```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -5.490e+02  1.115e+01  -49.24   <2e-16 ***
time(wages)  2.811e-01  5.618e-03   50.03   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.08257 on 70 degrees of freedom
Multiple R-squared:  0.9728,    Adjusted R-squared:  0.9724
F-statistic:  2503 on 1 and 70 DF,  p-value: < 2.2e-16
```

Standard Errors and Hypothesis Tests

The p-values in the R output are for testing the null hypothesis that each of the corresponding parameters is zero.

However (!):

- Standard error values assume that $\{X_t\}$ is a white noise process (i.e., iid random variables).
- t-statistics and p-values also assume that the X_t's must be approximately Normal.

Unless we have reason to believe that the above assumptions are true, we should not be making conclusions from the standard errors and p-values in the table.

Residuals

The **residual** corresponding to the t^{th} observation:

$$\hat{X}_t = Y_t - \hat{\mu}_t$$

If we wish to use the standard errors in the table, we should check whether $\{X_t\}$ appears to be a white noise process.

If we wish to use the t-statistics and p-values, we should check whether $\{X_t\}$ is a *normal* white noise process.

We can do this by plotting the behaviour of the residuals.

<u>Note:</u> We usually plot **standardized residuals** or **studentized residuals**. These values standardize the residual in some way, to make it easier to interpret.

Residuals vs. Time

Plot of residuals vs. time will tell us what the process $\{X_t\}$ might look like.

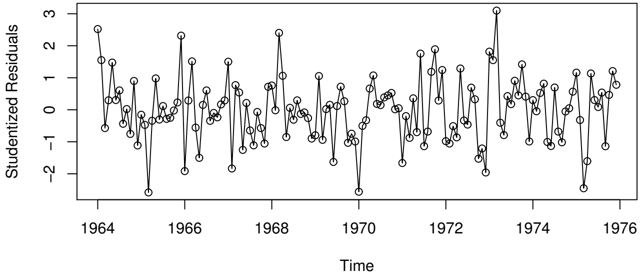
If $\{X_t\}$ is white noise, and the trend is adequately modelled, then we should see a random scatter around 0.

Residuals vs. Time

Plot of residuals vs. time will tell us what the process $\{X_t\}$ might look like.

If $\{X_t\}$ is white noise, and the trend is adequately modelled, then we should see a random scatter around 0.

```
data(tempdub)
month. <- season(tempdub)
model.seasonal <- lm(tempdub~month.-1)
plot(y=rstudent(model.seasonal1), x=as.vector(time(tempdub)), type='o',
     xlab='Time', ylab='Studentized Residuals')
```



Residuals vs. Time (cont'd)

If the data are seasonal, we should also investigate any patterns relating to different months of the year.

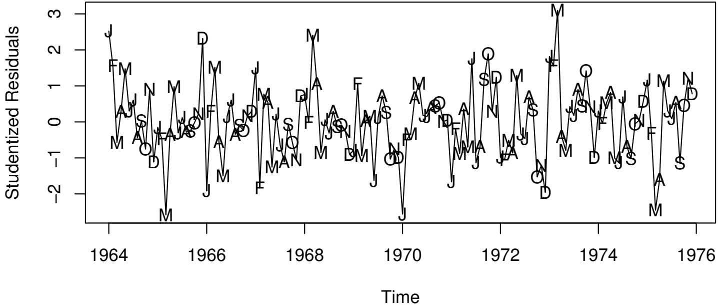
Plot residuals vs. time, with labels for the different months:

Residuals vs. Time (cont'd)

If the data are seasonal, we should also investigate any patterns relating to different months of the year.

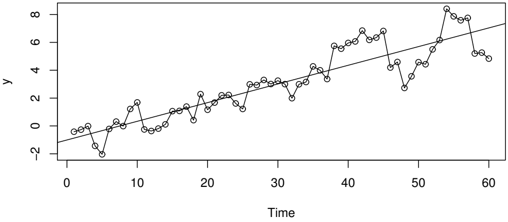
Plot residuals vs. time, with labels for the different months:

```
plot(y=rstudent(model.seasonal1), x=as.vector(time(tempdub)), type='o',
     xlab='Time', ylab='Studentized Residuals',
     pch=as.vector(season(tempdub)))
```



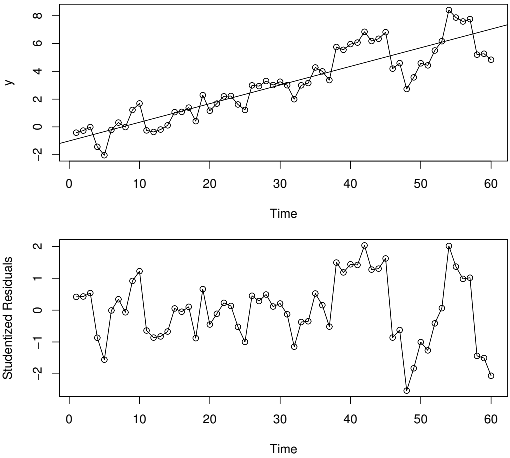
Residuals vs. Time (cont'd)

Another example: Fitting a linear trend to random walk data:



Residuals vs. Time (cont'd)

Another example: Fitting a linear trend to random walk data:



Here, the observations "hang together" too much to be white noise, and there is more variability at larger t.

Residuals vs. Trend Estimate

Plot of residuals vs. $\hat{\mu}_t$ will tell us if the trend fits the data well. If the chosen trend model is adequate, we should see a random scatter about 0.

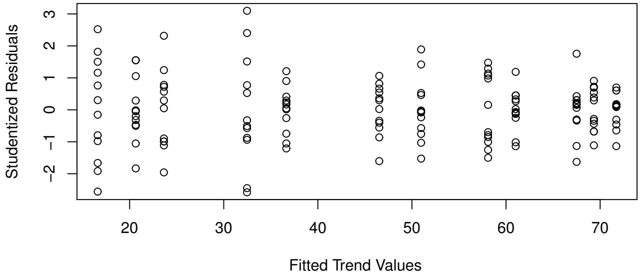
Any patterns in the plot may indicate that a different trend model should be used: Are smaller residuals associated with smaller $\hat{\mu}_t$'s? Is there more variation on one end than at the other?

Residuals vs. Trend Estimate

Plot of residuals vs. $\hat{\mu}_t$ will tell us if the trend fits the data well. If the chosen trend model is adequate, we should see a random scatter about 0.

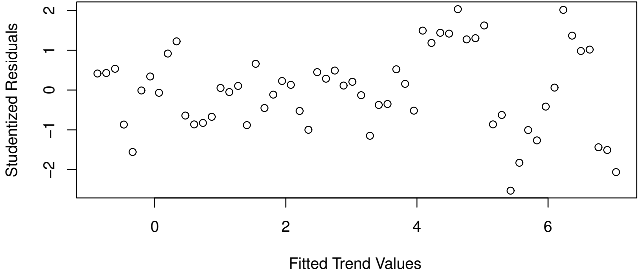
Any patterns in the plot may indicate that a different trend model should be used: Are smaller residuals associated with smaller $\hat{\mu}_t$'s? Is there more variation on one end than at the other?

```
plot(y=rstudent(model.seasonal1), x=as.vector(fitted(model.seasonal1)),
     xlab='Fitted Trend Values', ylab='Studentized Residuals')
```



Residuals vs. Trend Estimate (cont'd)

Another example: Fitting a linear trend to random walk data:



We see more variability at larger $\hat{\mu}_t$-values.

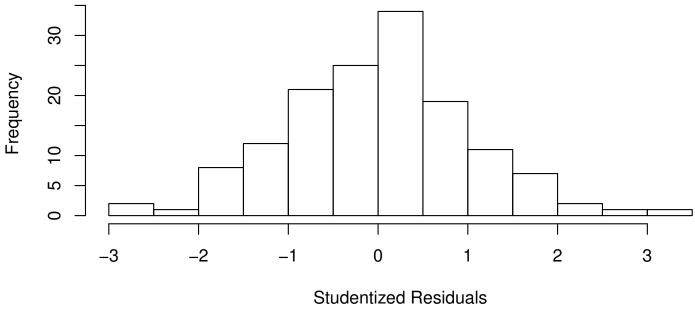
Histogram of Residuals

Histogram of residuals will tell us whether or not the process $\{X_t\}$ appears to be normal.

Histogram of Residuals

Histogram of residuals will tell us whether or not the process $\{X_t\}$ appears to be normal.

```
hist(rstudent(model.seasonal1), xlab='Studentized Residuals')
```



Does it look symmetric and bell-shaped?

Q-Q Plot of Residuals

Q-Q plot of residuals will also tell us whether or not $\{X_t\}$ appears to be normal.

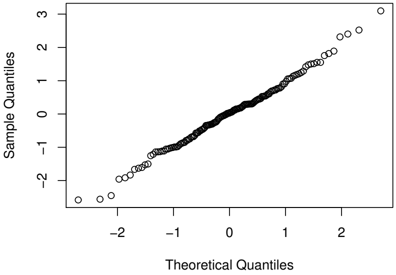
A Q-Q plot displays the quantiles of residuals vs. the theoretical quantiles of a normal distribution. If the residuals are approximately normal, we should see a straight line.

Q-Q Plot of Residuals

Q-Q plot of residuals will also tell us whether or not $\{X_t\}$ appears to be normal.

A Q-Q plot displays the quantiles of residuals vs. the theoretical quantiles of a normal distribution. If the residuals are approximately normal, we should see a straight line.

```
qqnorm(rstudent(model.seasonal1))
```



More formal way of testing for normality: Shapiro-Wilk test.

Testing for Independence

We might also wish to test for independence in $\{X_t\}$. This can be done using the **runs test**.

Null hypothesis: The X_t's are independent.

R Code:

```
runs(rstudent(model.seasonal1))
```

Sample Autocorrelation Function

Recall: Correlation between any two random variables X and Y:

$$Corr(X, Y) = \frac{Cov(X, Y)}{\sqrt{Var(X)Var(Y)}} = \frac{E[(X - E(X))(Y - E(Y))]}{\sqrt{Var(X)Var(Y)}}$$

Sample Autocorrelation Function

<u>Recall:</u> Correlation between any two random variables X and Y:

$$Corr(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}} = \frac{E[(X-E(X))(Y-E(Y))]}{\sqrt{Var(X)Var(Y)}}$$

This is usually estimated by the sample correlation coefficient:

$$\widehat{Corr}(X,Y) = \frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i-\bar{X})^2\sum_{i=1}^{n}(Y_i-\bar{Y})^2}}$$

Sample Autocorrelation Function

<u>Recall:</u> Correlation between any two random variables X and Y:

$$Corr(X, Y) = \frac{Cov(X, Y)}{\sqrt{Var(X)Var(Y)}} = \frac{E[(X - E(X))(Y - E(Y))]}{\sqrt{Var(X)Var(Y)}}$$

This is usually estimated by the sample correlation coefficient:

$$\widehat{Corr}(X, Y) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

For a time series $\{Y_1, Y_2, \dots, Y_n\}$, we can estimate its autocorrelation function ρ_k using the **sample autocorrelation function** (**sample ACF**):

$$r_k = \widehat{Corr}(Y_t, Y_{t-k}) = \frac{\sum_{t=k+1}^{n}(Y_t - \bar{Y})(Y_{t-k} - \bar{Y})}{\sum_{t=1}^{n}(Y_t - \bar{Y})^2}$$

Note: We assume stationarity in this definition.

Sample Autocorrelation Function of the Residuals

The sample ACF is a very useful tool! We'll be seeing plots of r_k vs. lag k a lot in this course.

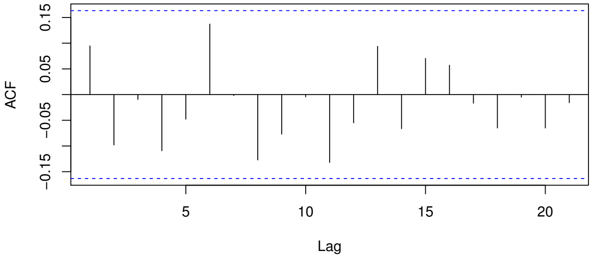
Right now, we are interested in examining the sample ACF of the residuals $\hat{X}_t$, in order to see if there may be dependence in $\{X_t\}$:

Sample Autocorrelation Function of the Residuals

The sample ACF is a very useful tool! We'll be seeing plots of r_k vs. lag k a lot in this course.

Right now, we are interested in examining the sample ACF of the residuals $\hat{X}_t$, in order to see if there may be dependence in $\{X_t\}$:

```
acf(rstudent(model.seasonal1))
```

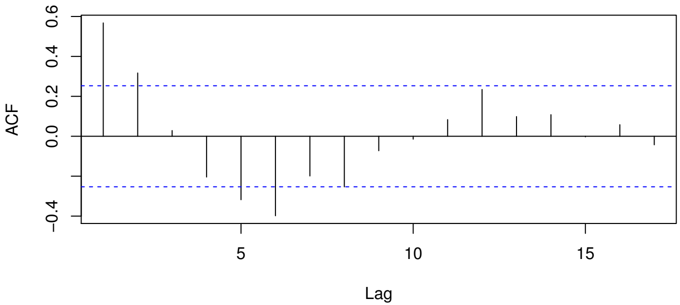


Any values outside of the dashed lines would suggest that that $\rho_k \neq 0$.

The plot above suggests that $\{X_t\}$ may be white noise, since none of the ρ_k's are significantly different from 0.

Sample Autocorrelation Function of the Residuals (cont'd)

Another example: Fitting a linear trend to random walk data:



This confirms the smoothness we saw in the plot of residuals vs. time. The residuals are correlated!

In particular: $\rho_1 > 0$ and $\rho_2 > 0$. (Similarly, $\rho_5 < 0$ and $\rho_6 < 0$.)

So, we have reason to believe that the stochastic part of the random walk (once the linear trend is removed) is not a white noise process.

Final Comments

That's all for now!

In this video, we've learned about a few different methods for assessing a trend model's goodness of fit. We've seen several different residual plots, and learned how to interpret them to assess model assumptions.

Finally, we learned about the all-important sample ACF, and how it can help us assess model assumptions as well.

Next Week in STAT 485/685: Some more practice with these concepts, and review of Ch. 1-3.

References

[1] Cryer, J. D., & Chan, K. S. (2008). *Time series analysis: with applications in R.* Springer Science and Business Media.

[2] Chan, K. S., & Ripley, B. (2020). TSA: Time Series Analysis. R package version 1.2.1.