

# Unit 4: Review of Regression

Jeffrey Woo

Department of Statistics, University of Virginia

Spring 2016

# Readings for Unit 4

Textbook chapter 2.2.

# Last Unit

- 1 Stationarity
- 2 Autocovariance and Autocorrelation of Stationary Time Series
- 3 Estimating the ACF

# This Unit

- 1 Parameter Estimation
- 2 Model Selection
- 3 Diagnostics

# Motivation

In time series analysis, we frequently would prefer to analyze a stationary process. This allows us to better estimate autocorrelation and other quantities of interest. In addition, ARMA processes provide a rich framework for analyzing stationary processes. A strong trend, however, may \_\_\_\_\_ the behavior of the stationary process. It may, therefore, be necessary to \_\_\_\_\_ a trend; one way to do that is via regression.

# 1 Linear Regression Basics

## 2 Parameter Estimation

## 3 Model Selection

## 4 Diagnostics

## 5 Worked Example

# Linear Regression Basics

The basic data type for regression consists of a list of pairs of numbers,  $(x_1, z_1), \dots, (x_n, z_n)$ , where the  $x_i$  are thought of as the dependent variables and  $z_i$  are thought of as the independent variables. The simple linear regression model would then be

$$x_t = \beta_1 + \beta_2 z_t + w_t$$

for  $t = 1, \dots, n$  where  $w_t, t = 1, \dots, n$  are zero-mean iid normal random variables with variance  $\sigma_w^2$ .

# Linear Regression Basics

A very simple case is where  $z_t = t$ . This would correspond to a basic linear trend. So, the model would look like

$$x_t = \beta_1 + \beta_2 t + w_t.$$



# Linear Regression Basics

We can extend this to multiple independent series with a model

$$x_t = \beta_1 z_{t1} + \dots + \beta_q z_{tq} + w_t.$$

Using vector notation, the linear regression model can be written as

$$x_t = \boldsymbol{\beta}' \mathbf{z}_t + w_t \tag{1}$$

where  $\boldsymbol{\beta}' = (\beta_1, \beta_2, \dots, \beta_q)$  and  $\mathbf{z}_t = (z_{t1}, z_{t2}, \dots, z_{tq})'$ .

1 Linear Regression Basics

2 Parameter Estimation

3 Model Selection

4 Diagnostics

5 Worked Example

# Parameter Estimation

Estimating the parameter vector  $\beta$  is done by minimizing the residual sum of squares

$$\text{RSS} = \sum_{t=1}^n (x_t - \beta' z_t)^2 \quad (2)$$

with respect to  $\beta_1, \beta_2, \dots, \beta_q$ . Let the matrix  $\mathbf{Z} = (z_1, z_2, \dots, z_n)'$  be the  $n \times q$  matrix of  $n$  samples of the independent variables, and  $\mathbf{x} = (x_1, x_2, \dots, x_n)'$  the vector of dependent variables. It turns out that

$$\hat{\beta} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{x} \quad (3)$$

The estimators  $\hat{\beta}$  are unbiased, and are called \_\_\_\_\_.

# Parameter Estimation

An unbiased estimator for the variance  $\sigma_w^2$  is

$$s_w^2 = \frac{\text{RSS}}{n - q}. \quad (4)$$

# Other Terminology

\_\_\_\_\_:

$$\hat{x}_t = \hat{\beta}' \mathbf{z}_t. \quad (5)$$

\_\_\_\_\_:

$$e_i = x_t - \hat{x}_t. \quad (6)$$

# Inference

Assuming Gaussian errors, we can build confidence intervals using statistics such as

$$\frac{\hat{\beta}_i - \beta_i}{\text{standard error}(\hat{\beta}_i)}$$

which have a t-distribution with  $n - q$  d.f, and  $s_w^2$  is distributed proportionally to a  $\chi_{n-q}^2$ .

1 Linear Regression Basics

2 Parameter Estimation

3 Model Selection

4 Diagnostics

5 Worked Example

# Model Selection: Nested Models

Often times, we want to compare various competing models or select a subset of independent variables. Consider a model that only has a subset  $q_1 < q$  independent variables

$$\mathbf{z}_{1t} = (z_{t1}, z_{t2}, \dots, z_{tq_1})',$$

$$x_t = \beta_1' \mathbf{z}_{1t} + w_t. \quad (7)$$

This is called the \_\_\_\_\_ model, and is compared with the \_\_\_\_\_ model, specified in (1) which has all  $q$  independent variables. Models (1) and (7) are called \_\_\_\_\_ models since all the terms in the reduced model occurs in the full model.



# Model Selection: Nested Models

With nested models, we compare the RSS of both models using the  $F$  statistic

$$F_{q-q_1, n-q} = \frac{\text{RSS}_1 - \text{RSS}}{\text{RSS}} \frac{n-q}{q-q_1}. \quad (8)$$

# Model Selection: Non-Nested Models

When comparing non-nested models, we can use the Akaike's Information Criterion (AIC)

$$\text{AIC} = \log \hat{\sigma}_k^2 + \frac{n + 2k}{n}, \quad (9)$$

where  $\hat{\sigma}_k^2 = \frac{\text{RSS}_k}{n}$  and  $\text{RSS}_k$  is the residual sum of squares for a model with  $k$  regression coefficients. For model selection, we would like to \_\_\_\_\_ the AIC.

- 1 Linear Regression Basics
- 2 Parameter Estimation
- 3 Model Selection
- 4 Diagnostics**
- 5 Worked Example

# Assumptions for Linear Regression

The assumptions for linear regression are

- There exist a linear relationship between the dependent and independent variables.
- $E(w_i) = 0$ .
- $\text{Var}(w_i) = \sigma_w^2$  is constant and finite.
- $w_i$ 's are uncorrelated.
- $w_i$  are iid normal.

# Diagnostics

Use the following to check regression assumptions are satisfied:

- Residual plot: to check if right regression equation used, variance of errors is constant, mean of errors is zero.
- ACF plot: to determine correlation.
- Normal probability plot: to check for normality.

1 Linear Regression Basics

2 Parameter Estimation

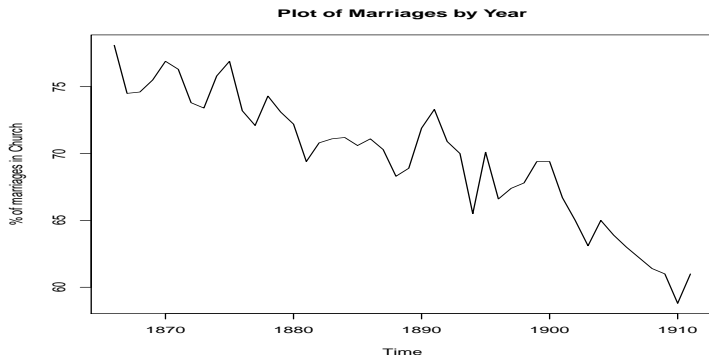
3 Model Selection

4 Diagnostics

5 Worked Example

# Marriages in Church of England

In this example, we go back to the data regarding number of marriages in the Church of England.



We choose to fit a simple linear regression because of the apparent decreasing trend.

# Marriages in Church of England

```
> timefit<-lm(marriages~time)
> summary(timefit)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	704.78168	37.93471	18.58	<2e-16 ***
time	-0.33629	0.02009	-16.74	<2e-16 ***

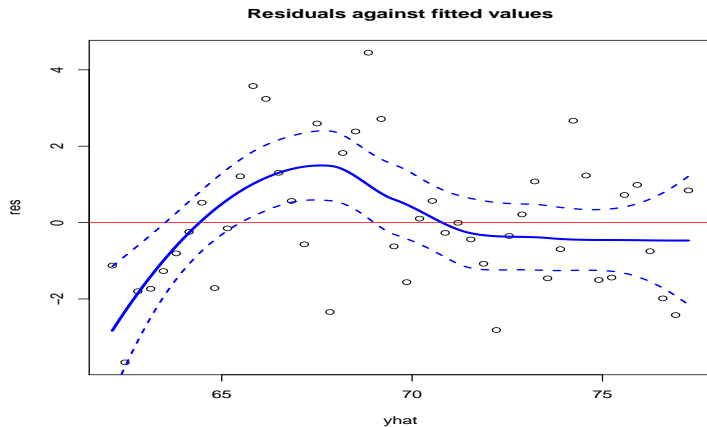
Residual standard error: 1.809 on 44 degrees of freedom  
Multiple R-squared: 0.8643,      Adjusted R-squared: 0.8612  
F-statistic: 280.3 on 1 and 44 DF,   p-value: < 2.2e-16

```
> AIC(timefit)
[1] 189.0146
```



# Marriages in Church of England

Check residual plot.



Curvature present. Let's add a square term for time.

# Marriages in Church of England

```
> timesq<-time^2
> timefitsq<-lm(marriages~time+timesq)
> anova(timefitsq)
Analysis of Variance Table
```

Response: marriages

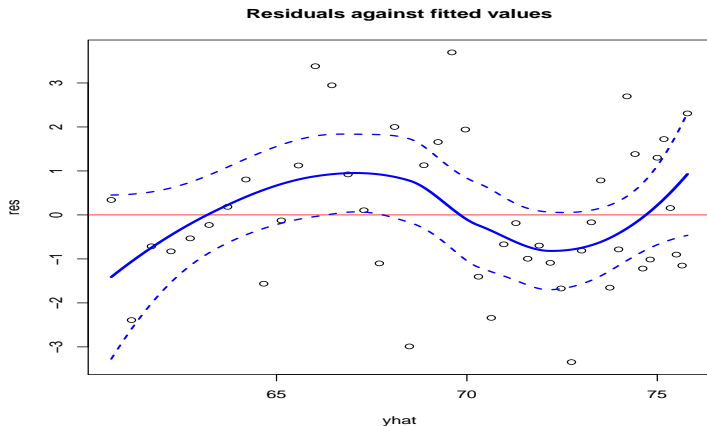
	Df	Sum Sq	Mean Sq	F value	Pr(>F)
time	1	916.90	916.90	324.6874	< 2.2e-16 ***
timesq	1	22.50	22.50	7.9682	0.007182 **
Residuals	43	121.43	2.82		

```
> AIC(timefitsq)
[1] 183.1945
```

P-value for timesq is significant. AIC has gone down, indicating the fit of the model has improved.

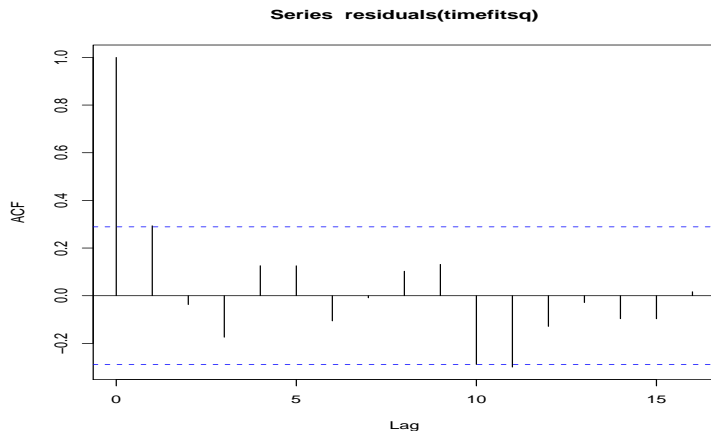
# Marriages in Church of England

Check residual plot.



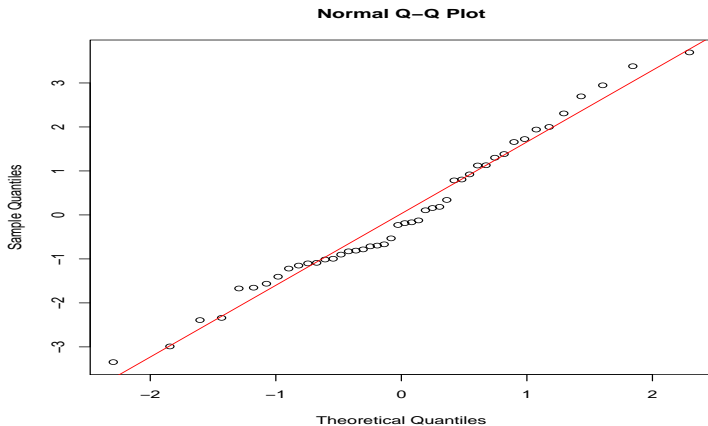
# Marriages in Church of England

Check ACF plot.



# Marriages in Church of England

Check QQ plot.



# Marriages in Church of England and Mortality Rate

Another possibility is that we wish to compare two time series via regression. We can treat one series as fixed, and the other series as simply a linearly transformed, perturbed version of that series. For example, is the percentage of marriages in the Church of England linearly related to the mortality rate in England?

# Marriages in Church of England and Mortality Rate

```
> comparefit<-lm(marriages~mortality)
> summary(comparefit)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	30.0553	1.9439	15.46	<2e-16 ***
mortality	2.1633	0.1054	20.52	<2e-16 ***

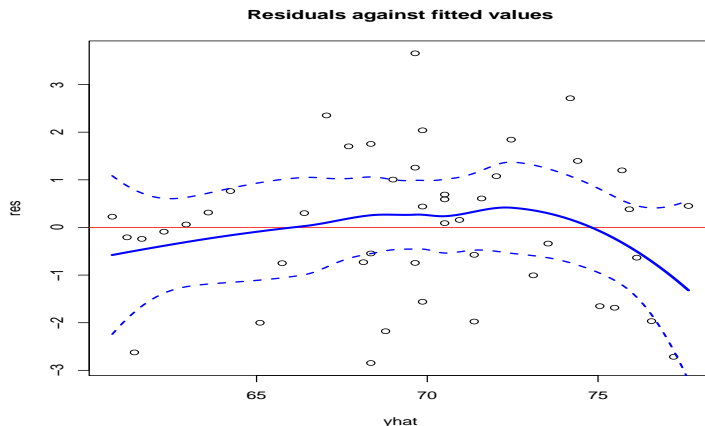
Residual standard error: 1.51 on 44 degrees of freedom

Multiple R-squared: 0.9054,      Adjusted R-squared: 0.9033

F-statistic: 421.3 on 1 and 44 DF,   p-value: < 2.2e-16

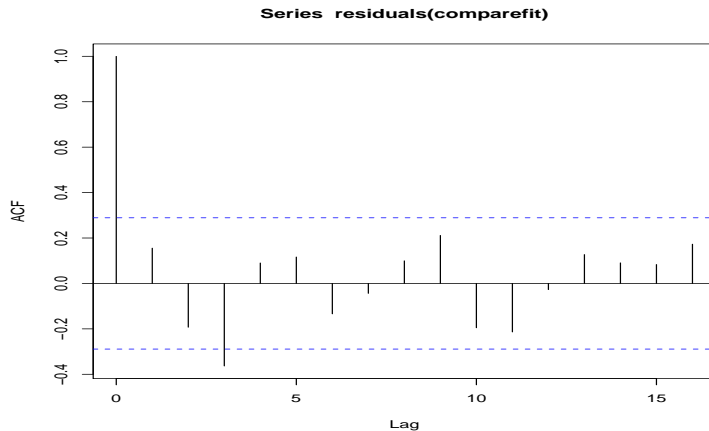
```
> AIC(comparefit)
[1] 172.4094
```

# Marriages in Church of England and Mortality Rate

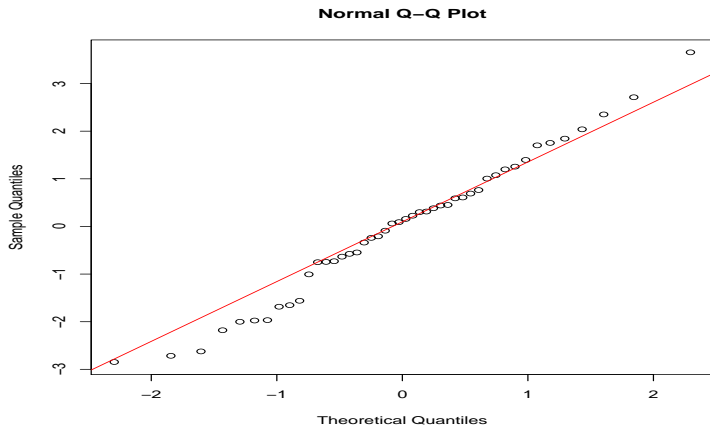




# Marriages in Church of England and Mortality Rate



# Marriages in Church of England and Mortality Rate



# Marriages in Church of England against Mortality Rate and Year

Use both mortality rate and year as independent variables.

```
> comparetimefit<-lm(marriages~mortality+time)
> anova(comparetimefit)
```

Analysis of Variance Table

Response: marriages

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
mortality	1	960.52	960.52	416.149	<2e-16 ***
time	1	1.07	1.07	0.464	0.4994
Residuals	43	99.25	2.31		

```
> AIC(comparetimefit)
[1] 173.9157
```