# Unit 4: Review of Regression

Taylor R. Brown PhD

Department of Statistics, University of Virginia

Spring 2020

# Readings for Unit 4

Textbook chapter 2.1.

# 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 **obscure** the behavior of the stationary process. It may, therefore, be necessary to **remove** a trend; one way to do that is via regression.

# Linear Regression Basics

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

$$x_t = \beta_0 + \beta_1 z_t + w_t$$

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

# Linear Regression Basics

We can extend this to multiple predictors with a model

$$x_t = \beta_0 + \beta_1 z_{t1} + \ldots + \beta_q z_{tq} + w_t.$$

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

$$x_t = \boldsymbol{\beta'} \boldsymbol{z_t} + w_t \tag{1}$$

where $\boldsymbol{\beta'} = (\beta_0, \beta_1, \cdots, \beta_q)$ and $\boldsymbol{z_t} = (1, z_{t1}, z_{t2}, \cdots, z_{tq})'$.

# Parameter Estimation

Estimating the parameter vector $\boldsymbol{\beta}$ is done by minimizing the error sum of squares

$$Q = \sum_{t=1}^{n} (x_t - \boldsymbol{\beta}' \boldsymbol{z_t})^2 \tag{2}$$

with respect to $\beta_0, \beta_1, \cdots, \beta_q$. Let the matrix $\boldsymbol{Z} = (1, z_1, z_2, \cdots, z_n)'$ be the $n \times (q+1)$ matrix of $n$ samples of the predictor variables, and $\boldsymbol{x} = (x_1, x_2, \cdots, x_n)'$ the vector of response variables. It turns out that

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

The estimators $\hat{\boldsymbol{\beta}}$ are unbiased, and are called **ordinary least squares estimators**.

## Parameter Estimation

The minimized error sum of squares, denoted by $SSE$, is

$$SSE = \sum_{t=1}^{n} (x_t - \hat{\beta}' \boldsymbol{z_t})^2 \tag{4}$$

# Parameter Estimation

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

$$s_w^2 = MSE = \frac{SSE}{n-(q+1)}. \tag{5}$$

# Other Terminology

**Fitted values**:

$$\hat{x}_t = \hat{\beta}' \boldsymbol{z_t}. \tag{6}$$

**Residuals**:

$$e_i = x_t - \hat{x}_t. \tag{7}$$

# Inference

Assuming independent 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 $n-(q+1)$ d.f, and $s_w^2$ is distributed proportionally to a $\chi^2_{n-(q+1)}$.

# Model Selection: Nested Models

Often times, we want to compare various competing models or select a subset of predictors. Consider a model that only has a subset $r < q$ predictors $\boldsymbol{z_{t,1:r}} = (z_{t1}, z_{t2}, \cdots, z_{tr})'$,

$$x_t = \boldsymbol{\beta_r'} \boldsymbol{z_{t,1:r}} + w_t. \tag{8}$$

(8) is called the **reduced** model, and is compared with the **full** model, as specified in (1), which has all $q$ predictor variables. Models (1) and (8) are called **nested** models since all the terms in the reduced model occur in the full model.

# Model Selection: Nested Models

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

$$F_{q-r,n-q-1} = \frac{SSE_r - SSE}{SSE} \frac{n-q-1}{q-r}, \tag{9}$$

where $SSE_r$ denotes the $SSE$ of the reduced model.

# Model Selection: Non-Nested Models

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

$$\text{AIC} = \log \hat{\sigma_k}^2 + \frac{n + 2k}{n}, \tag{10}$$

where $\hat{\sigma_k}^2 = \frac{SSE(k)}{n}$ and $SSE(k)$ is the $SSE$ for a model with $k$ regression coefficients. For model selection, we would like to **minimize** the AIC.

# Coefficient of Determination

The coefficient of determination, $R^2$, is a popular measure of model fit. For example, for the full model,

$$R^2 = \frac{SSE_0 - SSE}{SSE_0}, \tag{11}$$

where $SSE_0$ is the total sum of squares. $R^2$ is interpreted as the proportion of the variance in the response variable that can be explained by our model.

**Question:** When should $R^2$ be used / not used?

# Assumptions for Linear Regression

The assumptions for linear regression are

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

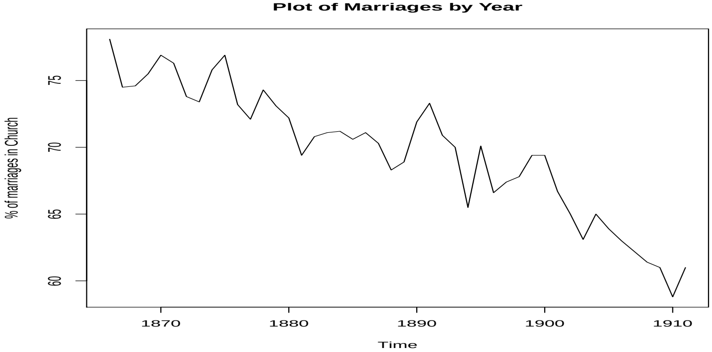
# Diagnostics

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.

# Marriages in Church of England

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



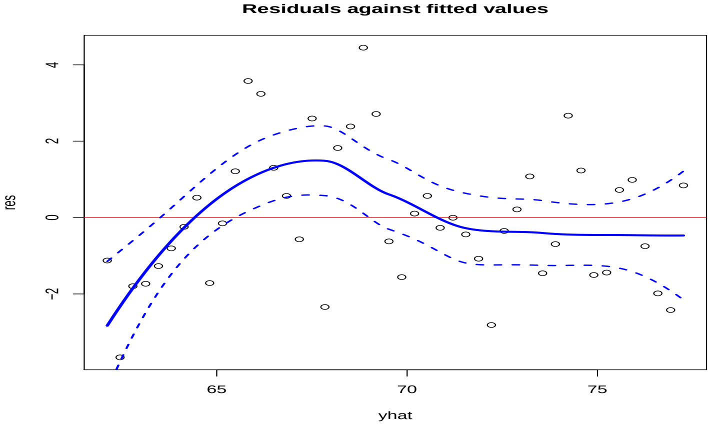

We choose to fit a simple linear regression because of the apparent decreasing trend.

# 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.

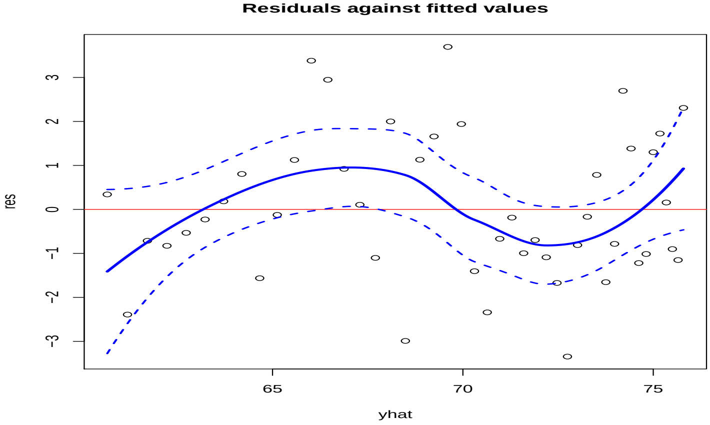
## Marriages in Church of England

```
> 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.

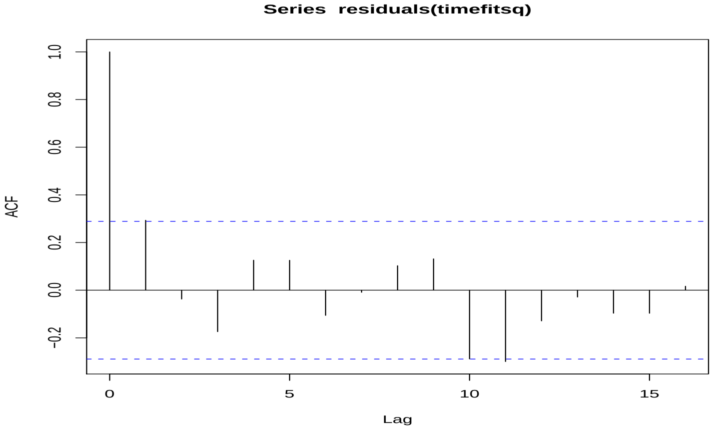
# Marriages in Church of England

Check residual plot.



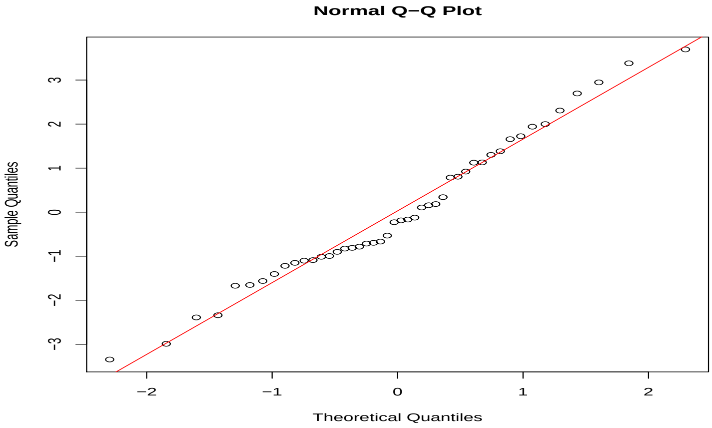
# Marriages in Church of England

Check ACF plot.



# Marriages in Church of England

Check QQ plot.



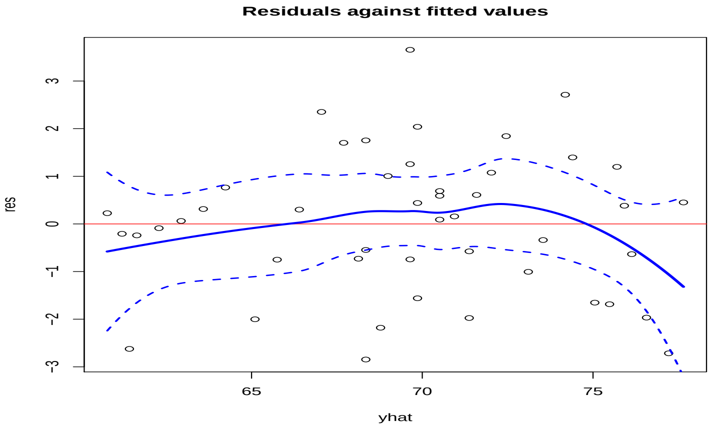
# Marriages in Church of England and Mortality Rate

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?

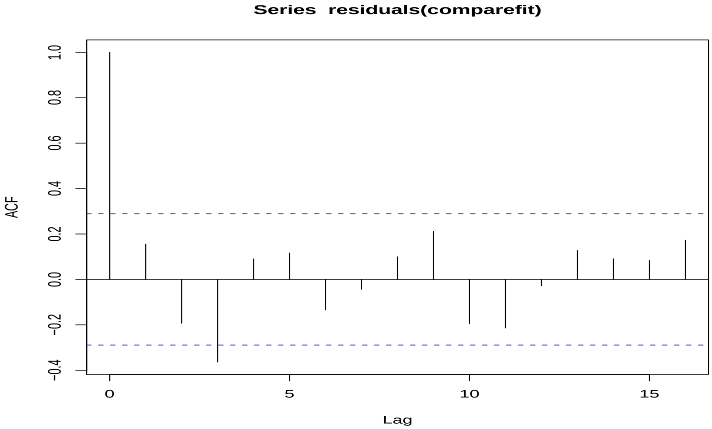
# Marriages in Church of England and Mortality Rate

```
> 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
```

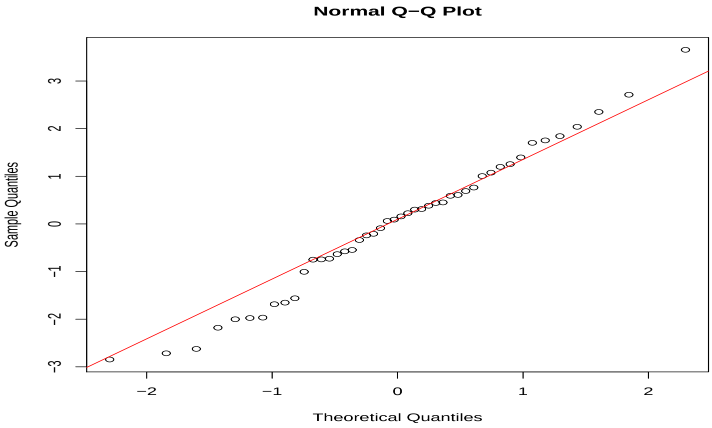
# Marriages in Church of England and Mortality Rate



# 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 predictor 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
```