STAT 626: Outline of Lecture ~~6~~ 7
Regression in the Time Series Contexts (§2.2,)

1. **Review of Stationarity, Preview of TS Models**

2. **A Quick Review of Multiple Regression**

$$x = Z\beta + w,$$

   LSE of β :

$$\widehat{\beta} = (Z'Z)^{-1}Z'x.$$

3. ~~**Tests, CIs and Variable Selection**~~

4. ~~**AIC (Akaike Information Criterion)**~~

5. ~~**Example 2.2: Pollution, Temperature and Mortality (PTM) Data**~~

6. ~~**Example 2.3: Regression with Lagged Variables**~~

7. **Examples 2.8, 2.9: Trig. Regression and Periodogram**

Review of Stationarity, Preview of TS Models

1. **Linear Processes:** $x_t = \mu + \sum_{j=-\infty}^{+\infty} \psi_j w_{t-j}$ is stationary with the *autocovariance function*

$$\gamma(h) = \sum_{j=-\infty}^{\infty} \psi_{j+h}\psi_j.$$

2. **MA(q) Models**: $x_t = w_t + \theta_1 w_{t-1} + \ldots + \theta_q w_{t-q}, \quad \theta_q \neq 0$, is stationary, its autocovariance is zero at lags greater than q.

3. **Autoregressive Models of order p or AR(p) Models:**

$$x_t = \phi_1 x_{t-1} + \ldots + \phi_p x_{t-p} + w_t, \quad \phi_p \neq 0.$$

QUESTION: Is a time series $\{x_t\}$ defined via an AR(p) model always stationary? If so, what is its autocovariance function?
To get a feel for the answer consider the AR(1):

$$x_t = \phi x_{t-1} + w_t,$$

what happens when $\phi = 1$?

4. **The Backshift Operator B:** $Bx_t = x_{t-1}$.

5. **MA(q) and B:**

$$x_t = w_t + \theta_1 w_{t-1} + \ldots + \theta_q w_{t-q} = (1 + \theta_1 B + \ldots + \theta_q B^q) w_t = \theta(B) w_t.$$

6. **AR(p) and B:**

$$x_t - \phi_1 x_{t-1} - \ldots - \phi_p x_{t-p} = w_t, \quad (1 - \phi_1 B - \ldots - \phi_p B^p) x_t = \phi(B) x_t = w_t.$$

7. **The ROOTS of the polynomial equation**

$$\phi(B) = 0,$$

hold the key to the question of stationarity of the solutions of AR models.

Regression and Forecasting:

PROBLEM (POPULATION INFORMATION): Given the value of a random variable X, **find β to minimize the mean-square error (MSE) of predicting Y by $\widehat{Y} = \beta X$**:

$$\text{MSE}(\beta) = E(Y - \beta X)^2.$$

SOLUTION: The minimizer satisfies the *normal equation*:

$$\text{Var}(X) \;\; \widehat{\beta} = \text{Cov}(X, Y)$$

or

$$\widehat{\beta} = \frac{\text{Cov}(X, Y)}{\text{Var}(X)},$$

and

$$\text{MSE}(\widehat{\beta}) = E(Y - \widehat{Y})^2 = (1 - \rho^2)\text{Var}(Y).$$

Time Series Prediction: Given the time series data $x_1, \ldots, x_n$ from a zero-mean stationary process $\{x_t\}$ with **known** autocovariance function,$\gamma(h)$ find the forecast value of the process at the next time point, x_{n+1}. More precisely, find $\phi_{n1}, \ldots, \phi_n$ to minimize the MSE of forecasting:

$$E(x_{n+1} - \phi_{n1}x_n - \ldots - \phi_{nn}x_1)^2.$$

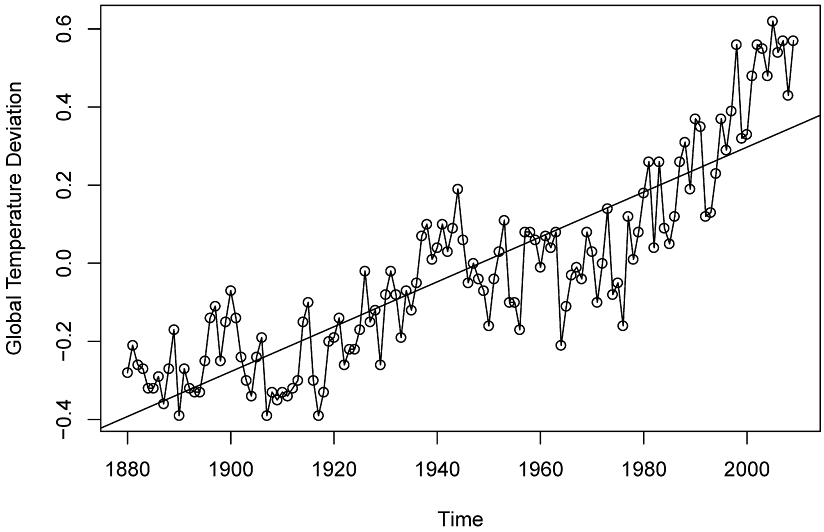

Fig. 2.1. Global temperature deviations shown in Figure 1.2 with fitted linear trend line.

```
summary(fit <- lm(gtemp~time(gtemp)))  # regress gtemp on time
plot(gtemp, type="o", ylab="Global Temperature Deviation")
abline(fit)  # add regression line to the plot
```

The linear model described by (2.1) above can be conveniently written in a more general notation by defining the column vectors $\boldsymbol{z}_t = (z_{t1}, z_{t2}, \dots, z_{tq})'$ and $\boldsymbol{\beta} = (\beta_1, \beta_2, \dots, \beta_q)'$, where $'$ denotes transpose, so (2.1) can be written in the alternate form

$$x_t = \boldsymbol{\beta}'\boldsymbol{z}_t + w_t. \tag{2.2}$$

where $w_t \sim \text{iid N}(0, \sigma_w^2)$. It is natural to consider estimating the unknown coefficient vector $\boldsymbol{\beta}$ by minimizing the error sum of squares

$$Q = \sum_{t=1}^{n} w_t^2 = \sum_{t=1}^{n} (x_t - \boldsymbol{\beta}'\boldsymbol{z}_t)^2, \tag{2.3}$$

with respect to $\beta_1, \beta_2, \dots, \beta_q$. Minimizing Q yields the ordinary least squares estimator of $\boldsymbol{\beta}$. This minimization can be accomplished by differentiating (2.3) with respect to the vector $\boldsymbol{\beta}$ or by using the properties of projections. In the notation above, this procedure gives the normal equations

$$\left(\sum_{t=1}^{n} \boldsymbol{z}_t\boldsymbol{z}_t'\right) \widehat{\boldsymbol{\beta}} = \sum_{t=1}^{n} \boldsymbol{z}_t x_t. \tag{2.4}$$

The notation can be simplified by defining $Z = [\boldsymbol{z}_1 \,|\, \boldsymbol{z}_2 \,|\, \cdots \,|\, \boldsymbol{z}_n]'$ as the $n \times q$ matrix composed of the n samples of the input variables, the observed $n \times 1$ vector $\boldsymbol{x} = (x_1, x_2, \dots, x_n)'$ and the $n \times 1$ vector of errors

$\boldsymbol{w} = (w_1, w_2, \ldots, w_n)'$. In this case, model (2.2) may be written as

$$\boldsymbol{x} = Z\boldsymbol{\beta} + \boldsymbol{w}. \tag{2.5}$$

The normal equations, (2.4), can now be written as

$$(Z'Z)\,\widehat{\boldsymbol{\beta}} = Z'\boldsymbol{x} \tag{2.6}$$

and the solution

$$\widehat{\boldsymbol{\beta}} = (Z'Z)^{-1}Z'\boldsymbol{x} \tag{2.7}$$

when the matrix $Z'Z$ is nonsingular. The minimized error sum of squares (2.3), denoted SSE, can be written as

$$\begin{aligned} SSE &= \sum_{t=1}^{n}(x_t - \widehat{\boldsymbol{\beta}}'\boldsymbol{z}_t)^2 \\ &= (\boldsymbol{x} - Z\widehat{\boldsymbol{\beta}})'(\boldsymbol{x} - Z\widehat{\boldsymbol{\beta}}) \\ &= \boldsymbol{x}'\boldsymbol{x} - \widehat{\boldsymbol{\beta}}'Z'\boldsymbol{x} \\ &= \boldsymbol{x}'\boldsymbol{x} - \boldsymbol{x}'Z(Z'Z)^{-1}Z'\boldsymbol{x}, \end{aligned} \tag{2.8}$$

to give some useful versions for later reference. The ordinary least squares estimators are unbiased, i.e., $E(\widehat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$, and have the smallest variance within the class of linear unbiased estimators.

If the errors w_t are normally distributed, $\widehat{\boldsymbol{\beta}}$ is also the maximum likelihood estimator for $\boldsymbol{\beta}$ and is normally distributed with

$$\operatorname{cov}(\widehat{\boldsymbol{\beta}}) = \sigma_w^2\left(\sum_{t=1}^{n}\boldsymbol{z}_t\boldsymbol{z}_t'\right)^{-1} = \sigma_w^2(Z'Z)^{-1} = \sigma_w^2 C, \tag{2.9}$$

where

$$C = (Z'Z)^{-1} \tag{2.10}$$

is a convenient notation for later equations. An unbiased estimator for the variance σ_w^2 is

$$s_w^2 = MSE = \frac{SSE}{n-q}, \tag{2.11}$$

where MSE denotes the *mean squared error*, which is contrasted with the maximum likelihood estimator $\widehat{\sigma}_w^2 = SSE/n$. Under the normal assumption, s_w^2 is distributed proportionally to a chi-squared random variable with $n-q$ degrees of freedom, denoted by χ^2_{n-q}, and independently of $\widehat{\beta}$. It follows that

$$t_{n-q} = \frac{(\widehat{\beta}_i - \beta_i)}{s_w\sqrt{c_{ii}}} \tag{2.12}$$

has the t-distribution with $n-q$ degrees of freedom; c_{ii} denotes the i-th diagonal element of C, as defined in (2.10).

Table 2.1. Analysis of Variance for Regression

Source	df	Sum of Squares	Mean Square
$z_{t,r+1},\dots,z_{t,q}$	$q-r$	$SSR = SSE_r - SSE$	$MSR = SSR/(q-r)$
Error	$n-q$	SSE	$MSE = SSE/(n-q)$
Total	$n-r$	SSE_r	

Various competing models are of interest to isolate or select the best subset of independent variables. Suppose a proposed model specifies that only a subset $r < q$ independent variables, say, $\boldsymbol{z}_{t:r} = (z_{t1}, z_{t2}, \dots, z_{tr})'$ is influencing the dependent variable x_t. The reduced model is

$$\boldsymbol{x} = Z_r \boldsymbol{\beta}_r + \boldsymbol{w} \tag{2.13}$$

where $\boldsymbol{\beta}_r = (\beta_1, \beta_2, \dots, \beta_r)'$ is a subset of coefficients of the original q variables and $Z_r = [\boldsymbol{z}_{1:r} \mid \cdots \mid \boldsymbol{z}_{n:r}]'$ is the $n \times r$ matrix of inputs. The null hypothesis in this case is H_0: $\beta_{r+1} = \cdots = \beta_q = 0$. We can test the reduced model (2.13) against the full model (2.2) by comparing the error sums of squares under the two models using the F-statistic

$$F_{q-r,n-q} = \frac{(SSE_r - SSE)/(q-r)}{SSE/(n-q)}, \tag{2.14}$$

which has the central F-distribution with $q-r$ and $n-q$ degrees of freedom when (2.13) is the correct model. Note that SSE_r is the error sum of squares under the reduced model (2.13) and it can be computed by replacing Z with Z_r in (2.8). The statistic, which follows from applying the likelihood ratio criterion, has the improvement per number of parameters added in the numerator compared with the error sum of squares under the full model in the denominator. The information involved in the test procedure is often summarized in an Analysis of Variance (ANOVA) table as given in Table 2.1 for this particular case. The difference in the numerator is often called the regression sum of squares

In terms of Table 2.1, it is conventional to write the F-statistic (2.14) as the ratio of the two mean squares, obtaining

$$F_{q-r,n-q} = \frac{MSR}{MSE}, \tag{2.15}$$

where MSR, the *mean squared regression*, is the numerator of (2.14). A special case of interest is $r = 1$ and $z_{t1} \equiv 1$, when the model in (2.13) becomes

$$x_t = \beta_1 + w_t,$$

and we may measure the proportion of variation accounted for by the other variables using

Example 2.8 Using Regression to Discover a Signal in Noise

In Example 1.12, we generated $n = 500$ observations from the model

$$x_t = A\cos(2\pi\omega t + \phi) + w_t, \tag{2.38}$$

where $\omega = 1/50$, $A = 2$, $\phi = .6\pi$, and $\sigma_w = 5$; the data are shown on the bottom panel of Figure 1.11 on page 16. At this point we assume the frequency of oscillation $\omega = 1/50$ is known, but A and ϕ are unknown parameters. In this case the parameters appear in (2.38) in a nonlinear way, so we use a trigonometric identity[4] and write

$$A\cos(2\pi\omega t + \phi) = \beta_1\cos(2\pi\omega t) + \beta_2\sin(2\pi\omega t),$$

where $\beta_1 = A\cos(\phi)$ and $\beta_2 = -A\sin(\phi)$. Now the model (2.38) can be written in the usual linear regression form given by (no intercept term is needed here)

$$x_t = \beta_1\cos(2\pi t/50) + \beta_2\sin(2\pi t/50) + w_t. \tag{2.39}$$

Using linear regression on the generated data, the fitted model is

$$\widehat{x}_t = -.71_{(.30)}\cos(2\pi t/50) - 2.55_{(.30)}\sin(2\pi t/50) \tag{2.40}$$

with $\widehat{\sigma}_w = 4.68$, where the values in parentheses are the standard errors. We note the actual values of the coefficients for this example are $\beta_1 = 2\cos(.6\pi) = -.62$ and $\beta_2 = -2\sin(.6\pi) = -1.90$. Because the parameter estimates are significant and close to the actual values, it is clear that we are able to detect the signal in the noise using regression, even though the signal appears to be obscured by the noise in the bottom panel of Figure 1.11. Figure 2.9 shows data generated by (2.38) with the fitted line, (2.40), superimposed.

To reproduce the analysis and Figure 2.9 in R, use the following commands:

```
set.seed(1000)  # so you can reproduce these results
x = 2*cos(2*pi*1:500/50 + .6*pi) + rnorm(500,0,5)
z1 = cos(2*pi*1:500/50);  z2 = sin(2*pi*1:500/50)
summary(fit <- lm(x~0+z1+z2))  # zero to exclude the intercept
plot.ts(x, lty="dashed")
lines(fitted(fit), lwd=2)
```

Example 2.9 Using the Periodogram to Discover a Signal in Noise

The analysis in Example 2.8 may seem like cheating because we assumed we knew the value of the frequency parameter ω. If we do not know ω, we could try to fit the model (2.38) using nonlinear regression with ω as a parameter. Another method is to try various values of ω in a systematic way. Using the

[4] $\cos(\alpha \pm \beta) = \cos(\alpha)\cos(\beta) \mp \sin(\alpha)\sin(\beta)$.

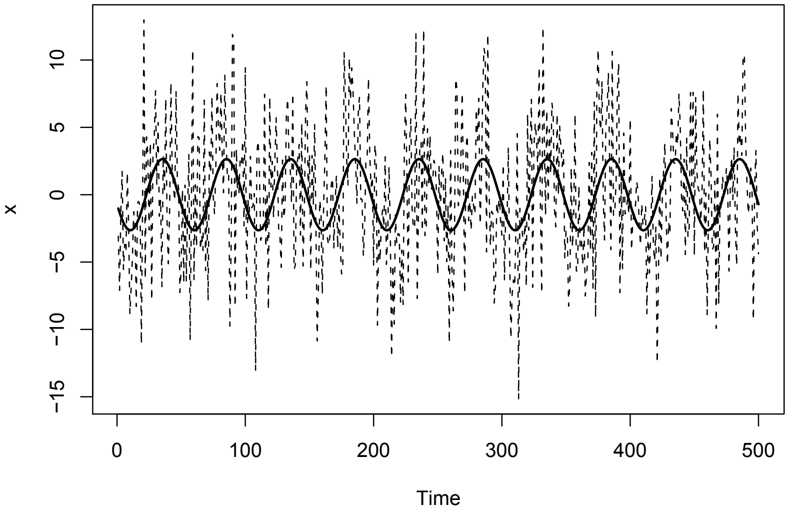

Fig. 2.9. Data generated by (2.38) [dashed line] with the fitted [solid] line, (2.40), superimposed.

regression results of §2.2, we can show the estimated regression coefficients in Example 2.8 take on the special form given by

$$\widehat{\beta}_1 = \frac{\sum_{t=1}^n x_t \cos(2\pi t/50)}{\sum_{t=1}^n \cos^2(2\pi t/50)} = \frac{2}{n}\sum_{t=1}^n x_t \cos(2\pi t/50); \tag{2.41}$$

$$\widehat{\beta}_2 = \frac{\sum_{t=1}^n x_t \sin(2\pi t/50)}{\sum_{t=1}^n \sin^2(2\pi t/50)} = \frac{2}{n}\sum_{t=1}^n x_t \sin(2\pi t/50). \tag{2.42}$$

This suggests looking at all possible regression parameter estimates,[5] say

$$\widehat{\beta}_1(j/n) = \frac{2}{n}\sum_{t=1}^n x_t \cos(2\pi t\, j/n); \tag{2.43}$$

$$\widehat{\beta}_2(j/n) = \frac{2}{n}\sum_{t=1}^n x_t \sin(2\pi t\, j/n), \tag{2.44}$$

where, $n = 500$ and $j = 1, \ldots, \frac{n}{2} - 1$, and inspecting the results for large values. For the endpoints, $j = 0$ and $j = n/2$, we have $\widehat{\beta}_1(0) = n^{-1}\sum_{t=1}^n x_t$ and $\widehat{\beta}_1(\frac{1}{2}) = n^{-1}\sum_{t=1}^n (-1)^t x_t$, and $\widehat{\beta}_2(0) = \widehat{\beta}_2(\frac{1}{2}) = 0$.

For this particular example, the values calculated in (2.41) and (2.42) are $\widehat{\beta}_1(10/500)$ and $\widehat{\beta}_2(10/500)$. By doing this, we have regressed a series, x_t, of

[5] In the notation of §2.2, the estimates are of the form $\sum_{t=1}^n x_t z_t \,/\, \sum_{t=1}^n z_t^2$ where $z_t = \cos(2\pi t j/n)$ or $z_t = \sin(2\pi t j/n)$. In this setup, unless $j = 0$ or $j = n/2$ if n is even, $\sum_{t=1}^n z_t^2 = n/2$; see Problem 2.10.

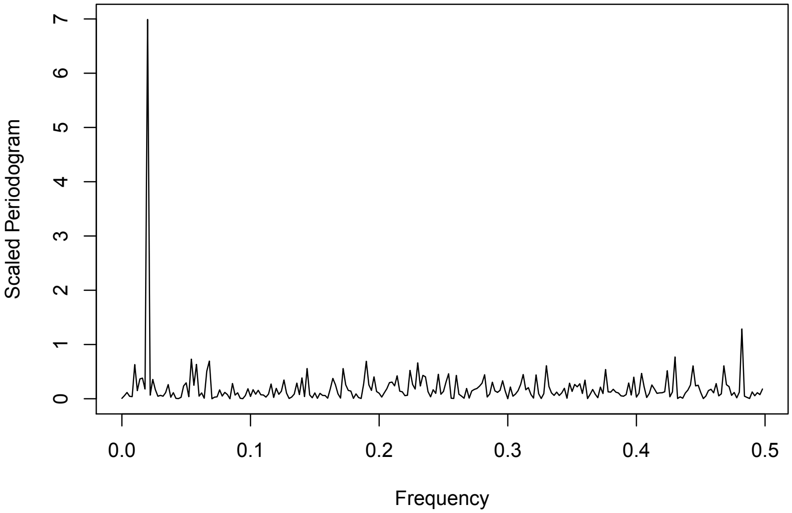

Fig. 2.10. The scaled periodogram, (2.45), of the 500 observations generated by (2.38); the data are displayed in Figures 1.11 and 2.9.

length n using n regression parameters, so that we will have a perfect fit. The point, however, is that if the data contain any cyclic behavior we are likely to catch it by performing these saturated regressions.

Next, note that the regression coefficients $\widehat{\beta}_1(j/n)$ and $\widehat{\beta}_2(j/n)$, for each j, are essentially measuring the correlation of the data with a sinusoid oscillating at j cycles in n time points.[6] Hence, an appropriate measure of the presence of a frequency of oscillation of j cycles in n time points in the data would be

$$P(j/n) = \widehat{\beta}_1^2(j/n) + \widehat{\beta}_2^2(j/n), \tag{2.45}$$

which is basically a measure of squared correlation. The quantity (2.45) is sometimes called the periodogram, but we will call $P(j/n)$ the scaled periodogram and we will investigate its properties in Chapter 4. Figure 2.10 shows the scaled periodogram for the data generated by (2.38), and it easily discovers the periodic component with frequency $\omega = .02 = 10/500$ even though it is difficult to visually notice that component in Figure 1.11 due to the noise.

Finally, we mention that it is not necessary to run a large regression

$$x_t = \sum_{j=0}^{n/2} \beta_1(j/n)\cos(2\pi t j/n) + \beta_2(j/n)\sin(2\pi t j/n) \tag{2.46}$$

to obtain the values of $\beta_1(j/n)$ and $\beta_2(j/n)$ [with $\beta_2(0) = \beta_2(1/2) = 0$] because they can be computed quickly if n (assumed even here) is a highly

[6] Sample correlations are of the form $\sum_t x_t z_t / \left(\sum_t x_t^2 \sum_t z_t^2\right)^{1/2}$.

composite integer. There is no error in (2.46) because there are n observations and n parameters; the regression fit will be perfect. The discrete Fourier transform (DFT) is a complex-valued weighted average of the data given by

$$\begin{aligned} d(j/n) &= n^{-1/2} \sum_{t=1}^{n} x_t \exp(-2\pi i t j/n) \\ &= n^{-1/2} \left(\sum_{t=1}^{n} x_t \cos(2\pi t j/n) - i \sum_{t=1}^{n} x_t \sin(2\pi t j/n) \right) \end{aligned} \tag{2.47}$$

where the frequencies j/n are called the Fourier or fundamental frequencies. Because of a large number of redundancies in the calculation, (2.47) may be computed quickly using the fast Fourier transform (FFT)[7], which is available in many computing packages such as Matlab®, S-PLUS® and R. Note that[8]

$$|d(j/n)|^2 = \frac{1}{n} \left(\sum_{t=1}^{n} x_t \cos(2\pi t j/n) \right)^2 + \frac{1}{n} \left(\sum_{t=1}^{n} x_t \sin(2\pi t j/n) \right)^2 \tag{2.48}$$

and it is this quantity that is called the periodogram; we will write

$$I(j/n) = |d(j/n)|^2.$$

We may calculate the scaled periodogram, (2.45), using the periodogram as

$$P(j/n) = \frac{4}{n} I(j/n). \tag{2.49}$$

We will discuss this approach in more detail and provide examples with data in Chapter 4.

Figure 2.10 can be created in R using the following commands (and the data already generated in x):

```
I = abs(fft(x))^2/500   # the periodogram
P = (4/500)*I[1:250]    # the scaled periodogram
f = 0:249/500           # frequencies
plot(f, P, type="l", xlab="Frequency", ylab="Scaled Periodogram")
```

~~2.4 Smoothing in the Time Series Context~~

~~In §1.4, we introduced the concept of smoothing a time series, and in Example 1.9, we discussed using a moving average to smooth white noise. This method is useful in discovering certain traits in a time series, such as long-term~~

[7] Different packages scale the FFT differently; consult the documentation. R calculates (2.47) without scaling by $n^{-1/2}$.

[8] If $z = a - ib$ is complex, then $|z|^2 = z\overline{z} = (a - ib)(a + ib) = a^2 + b^2$.