Time Series

Lesson 3

Grant Foster

Time Series Process

Ultimately boils down to a *joint probability function* for x at all moments of time t.

If x continuous, pdf = probability density function.
If x discrete, pmf = probability mass function.

Deterministic: $x(t) = f(t)$ so $p(x) = \delta(x - f(t))$,

$$p(x_1, x_2, ..., x_n) = \prod_{j=1}^{n} \delta(x_j - f(t_j)).$$

Regression

We often examine data which we believe shows a deterministic signal, and we wish to characterize that signal. In fact we often have reason to believe we know the *form* of the signal, but we have to use the data to estimate the *parameters* of our model. This process can be called *regression*.

Least Squares Regression

There are many roads to regression, but by far the best-known and most common method is *least-squares regression.*

Suppose, for instance, that we believe the data follow a straight line, but also include random noise. In this case the observed values x_n will be given by

$$x_n = \beta_o + \beta_1 t_n + \varepsilon_n.$$

$\beta_o =$ intercept, $\quad \beta_1 =$ slope, $\quad \varepsilon_n =$ noise.

Least Squares Regression

For the moment, assume that the noise is zero-mean white noise

$$\langle \varepsilon_n \rangle = 0,$$

and that the variance of the noise is given by σ^2 so

$$\langle \varepsilon_j \varepsilon_k \rangle = \sigma^2 \delta_{jk},$$

where δ_{jk} is the Kronecker delta,

$$\delta_{jk} = \begin{cases} 1 & j = k \\ \\ 0 & \text{else} \end{cases}$$

Least Squares Regression

(Later ... we'll consider the effect if the noise follows some other process.)

Least Squares Regression

For any given set of parameters β_o and β_1, we have a *model* of the behavior of the data. The model, of course, enables us to compute what the data values would be in the absence of noise

$$y_n = \beta_o + \beta_1 t_n.$$

We can take the difference between the observed values x_n and the values from a particular model y_n as the definition of the *residuals*

$$R_n = x_n - y_n = x_n - \beta_o - \beta_1 t_n.$$

Least Squares Regression

Now we can take the sum of the squares of all the residuals as a measure of the "total error" of the model

$$E = \sum_{n=1}^{N} (R_n)^2 = \sum_{n=1}^{N} (x_n - \beta_o - \beta_1 t_n)^2$$

The method of least squares selects the parameter values which give the smallest total error, or *sum of squared residuals* (SSR).

Hence the name "least squares."

Least Squares Regression

How do we find those parameter values? We simply find the values for which the partial derivative of the SSR with respect to each and every parameter is equal to zero.

Least Squares Regression

For the intercept parameter β_o we have

$$\frac{\partial E}{\partial \beta_o} = -2\sum_{n=1}^{N}(x_n - \beta_o - \beta_1 t_n) = 0.$$

For the slope parameter β_1 we have

$$\frac{\partial E}{\partial \beta_1} = -2\sum_{n=1}^{N} t_n(x_n - \beta_o - \beta_1 t_n) = 0.$$

These are two equations in two unknowns (β_o and β_1), enabling us to determine the parameters.

Least Squares Regression

We can write them as

$$\sum_{n=1}^{N} x_n = \sum_{n=1}^{N} \beta_o + \sum_{n=1}^{N} \beta_1 t_n = N\beta_o + \beta_1 \sum_{n=1}^{n} t_n,$$

and

$$\sum_{n=1}^{N} t_n x_n = \sum_{n=1}^{N} \beta_o t_n + \sum_{n=1}^{N} \beta_1 (t_n)^2 = \beta_o \sum_{n=1}^{N} t_n + \beta_1 \sum_{n=1}^{n} (t_n)^2.$$

These equations are *linear* in the parameters β_o and β_1, so this process is called *linear least squares*.

Least Squares Regression

For conceptual simplicity, I'll divide these equations by N and define the average data value

$$\bar{x} = \frac{1}{N} \sum_{n-1}^{N} x_n,$$

and average time

$$\bar{t} = \frac{1}{N} \sum_{n=1}^{N} t_n.$$

Least Squares Regression

Then the equations become

$$\bar{x} = \beta_o + \beta_1 \bar{t},$$

and

$$\frac{1}{N} \sum_{n=1}^{N} t_n x_n = \beta_o \bar{t} + \frac{\beta_1}{N} \sum_{n=1}^{n} (t_n)^2.$$

Least Squares Regression

Or,

$$\langle x \rangle = \beta_o + \beta_1 \langle t \rangle,$$

and

$$\langle tx \rangle = \beta_o \langle t \rangle + \beta_1 \langle t^2 \rangle.$$

Keep in mind that since we assume that t_n and x_n are actual data rather than just an abstract *process*, the angle brackets denote *average values* rather than *expected values*.

Least Squares Regression

We can write these equations in matrix form, as

$$\begin{bmatrix} \langle x \rangle \\ \\ \langle tx \rangle \end{bmatrix} = \begin{bmatrix} 1 & \langle t \rangle \\ \\ \langle t \rangle & \langle t^2 \rangle \end{bmatrix} \begin{bmatrix} \beta_o \\ \\ \beta_1 \end{bmatrix}.$$

The equations can be solved for the coefficients β_n by multiplying both sides by the inverse of the matrix. This gives

$$\begin{bmatrix} \beta_o \\ \\ \beta_1 \end{bmatrix} = \frac{1}{\langle t^2 \rangle - \langle t \rangle^2} \begin{bmatrix} \langle t^2 \rangle & -\langle t \rangle \\ \\ -\langle t \rangle & 1 \end{bmatrix} \begin{bmatrix} \langle x \rangle \\ \\ \langle tx \rangle \end{bmatrix}.$$

Least Squares Regression

Knowledge of the coefficients β_0 and β_1 tells us the straight line which "best" fits the data in the least-squares sense. This process is called *linear regression*.

Least Squares Regression

One should be careful about terminology.

In the name "linear regression" the word "linear" refers to the fact that the model is a straight line.

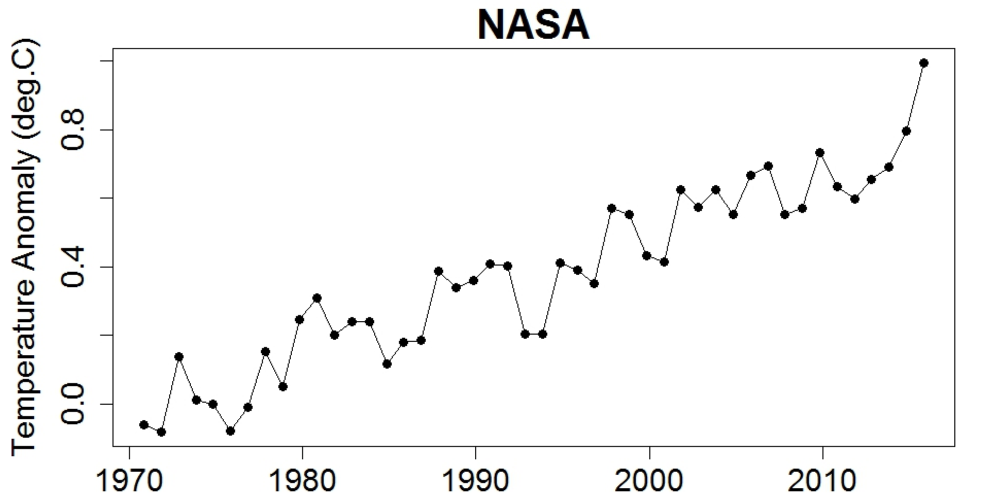
BUT in the name "linear least squares" the word "linear" refers to the fact that the model is linear in its regression coefficients (whatever those coefficients might refer to).

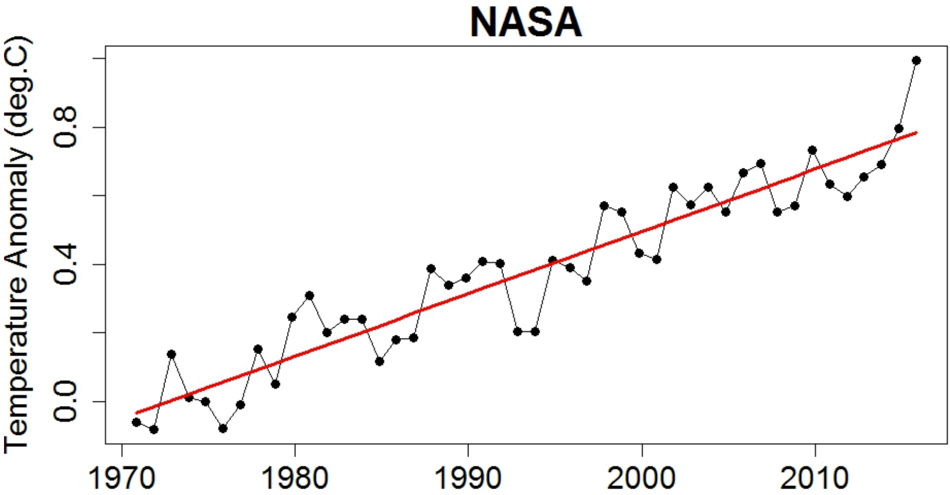
Least Squares Regression

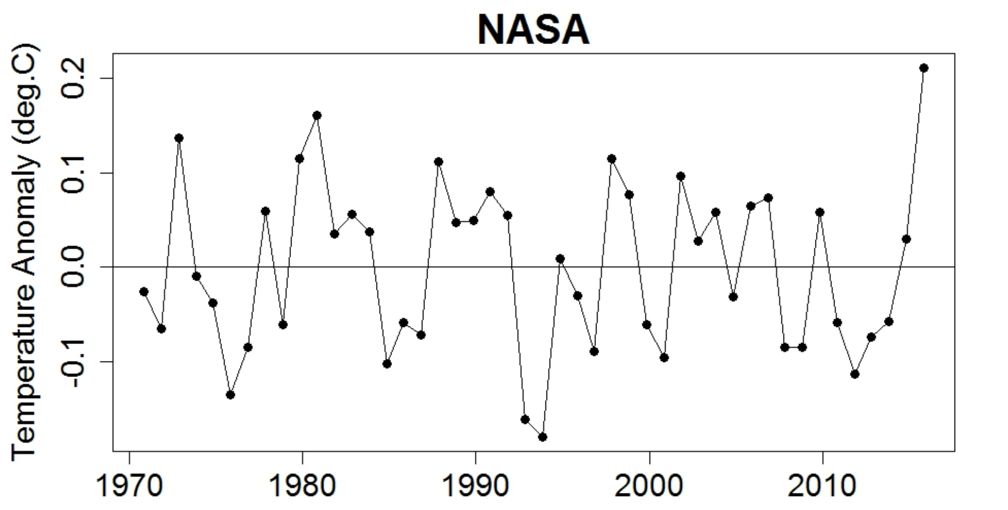
Some researchers make the mistake of saying they've applied "nonlinear" least squares when they've actually used linear least squares, but the model has terms which are nonlinear in their arguments. But they're linear in the *regression coefficients* so it's linear least squares.

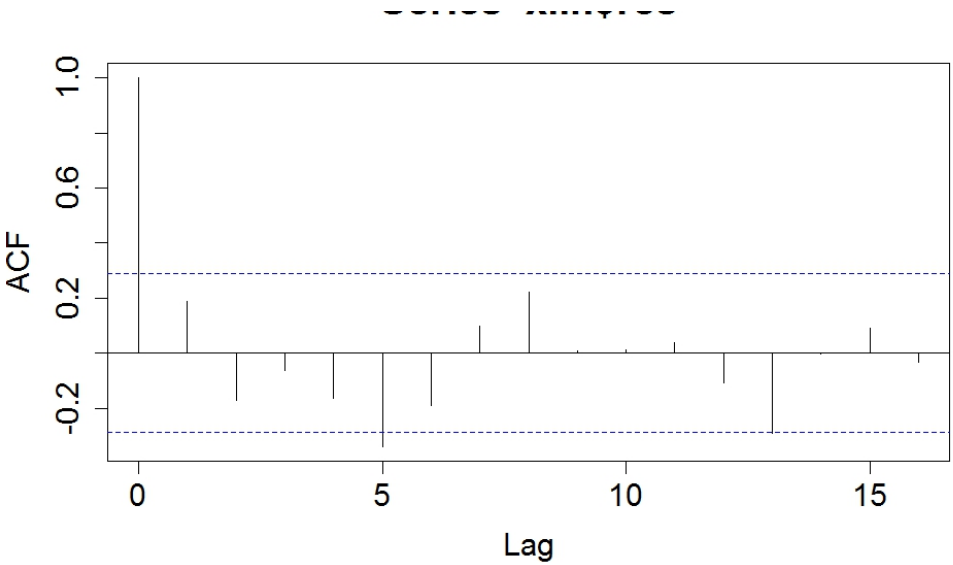
Least Squares Regression

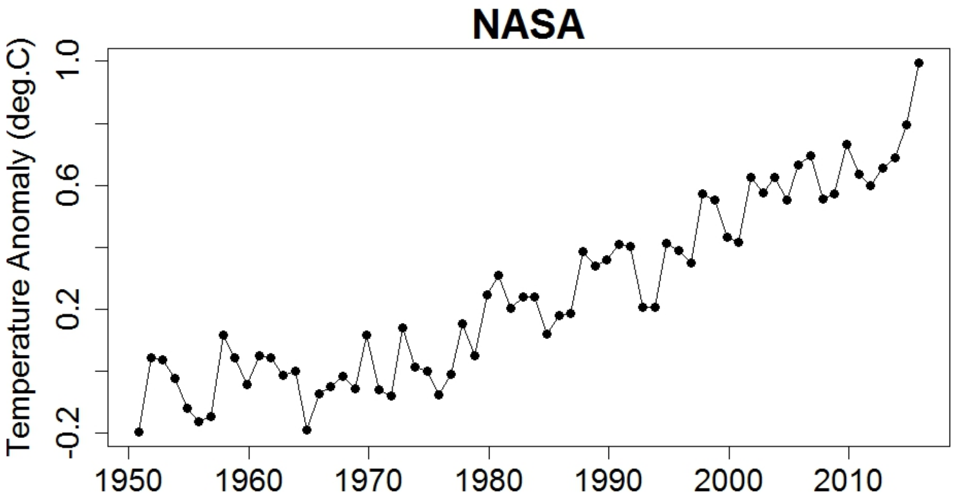
Example: Global Temperature

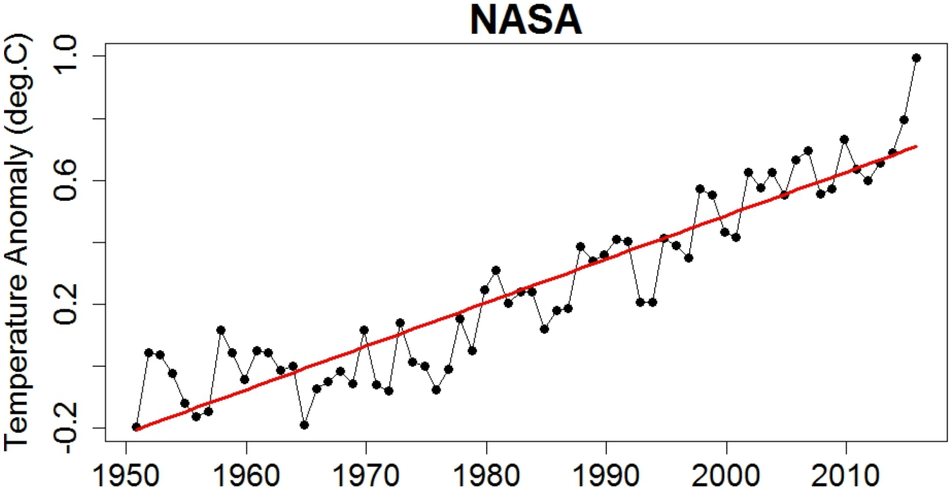

NASA
Temperature Anomaly (deg.C)
0.0
0.4
0.8
1970
1980
1990
2000
2010

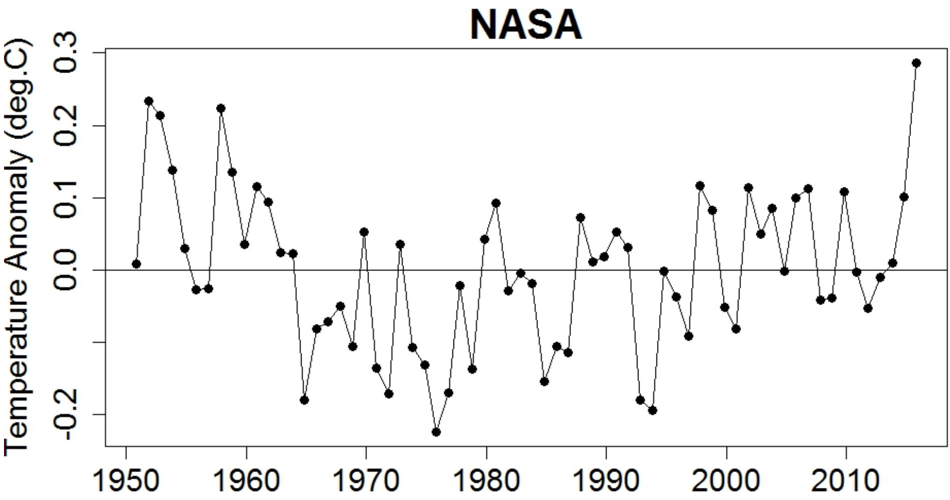

NASA
Temperature Anomaly (deg.C)
0.0
0.4
0.8
1970
1980
1990
2000
2010

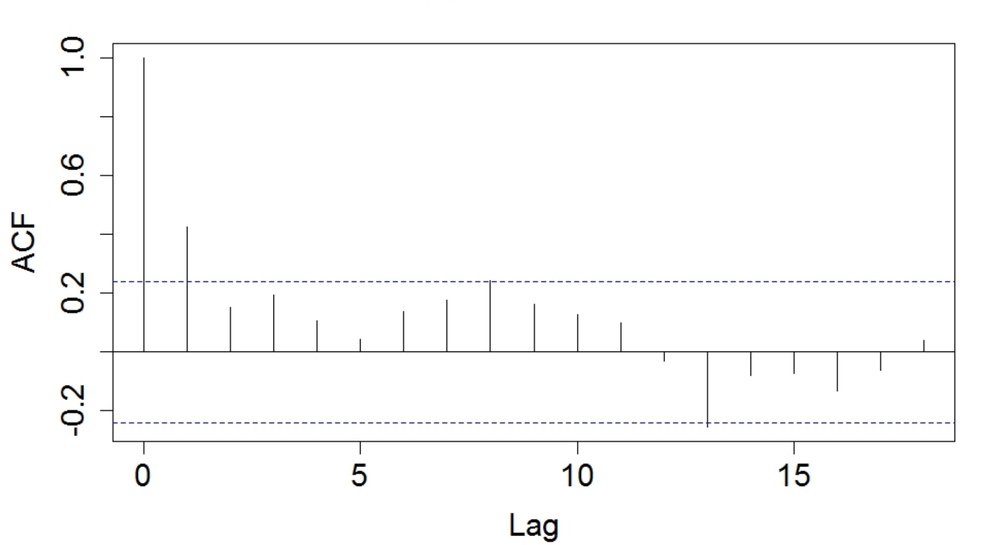

NASA
Temperature Anomaly (deg.C)
0.2
0.1
0.0
-0.1
1970
1980
1990
2000
2010


ACF
1.0
0.6
0.2
-0.2
0
5
10
15
Lag

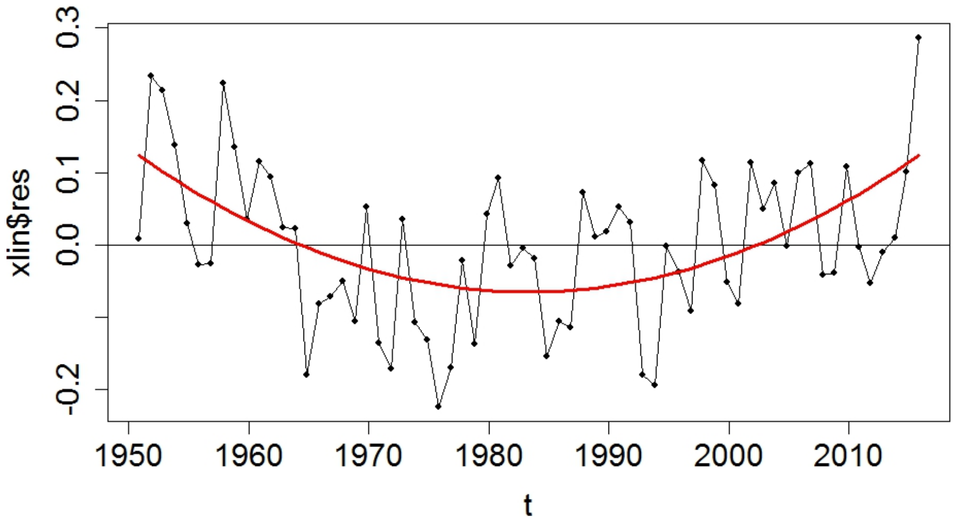

NASA
Temperature Anomaly (deg.C)
-0.2
0.2
0.6
1.0
1950
1960
1970
1980
1990
2000
2010


NASA
Temperature Anomaly (deg.C)
-0.2
0.2
0.6
1.0
1950
1960
1970
1980
1990
2000
2010

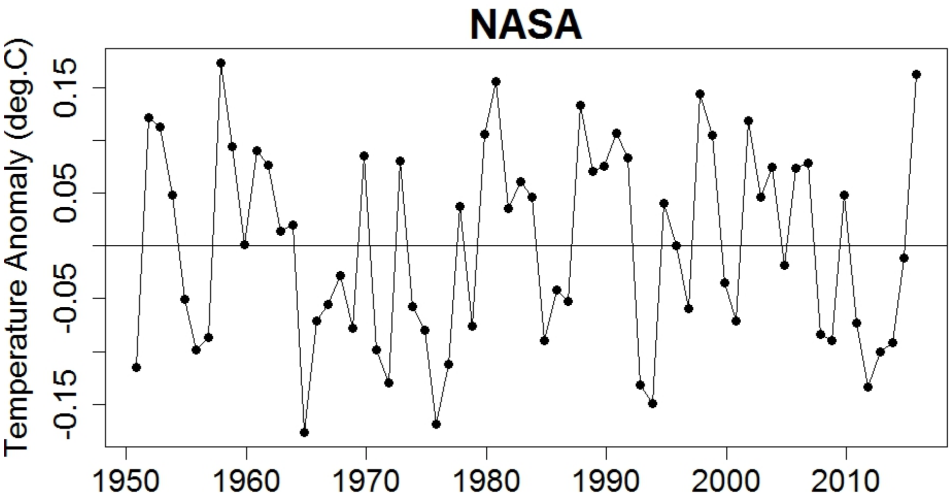

NASA
Temperature Anomaly (deg.C)
0.3
0.2
0.1
0.0
-0.2
1950
1960
1970
1980
1990
2000
2010

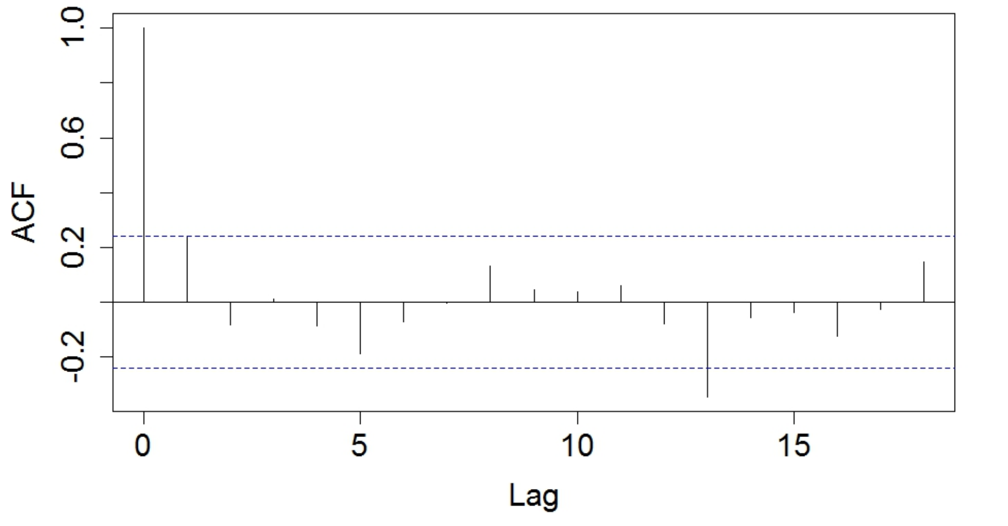


ACF
1.0
0.6
0.2
-0.2
0
5
10
15
Lag

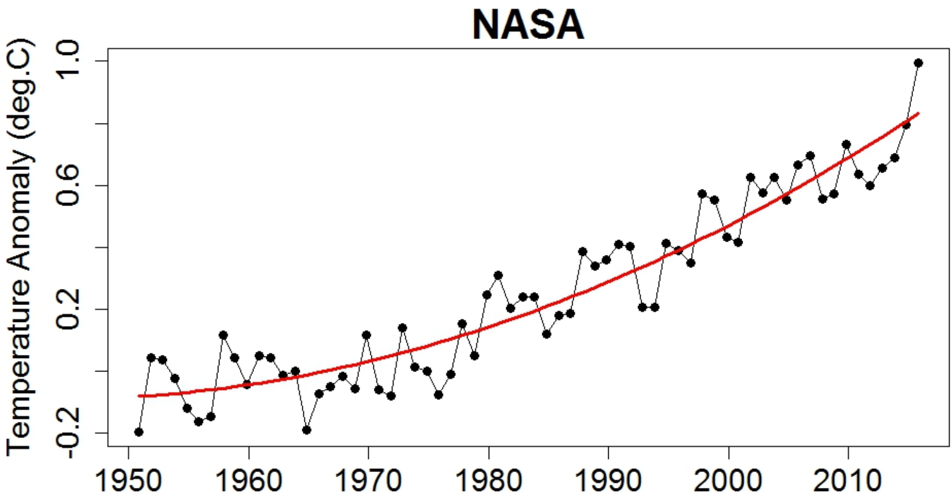


xlin$res
0.3
0.2
0.1
0.0
-0.2
1950
1960
1970
1980
1990
2000
2010
t


NASA
Temperature Anomaly (deg.C)
0.15
0.05
-0.05
-0.15
1950
1960
1970
1980
1990
2000
2010



ACF
1.0
0.6
0.2
-0.2
0
5
10
15
Lag


NASA
Temperature Anomaly (deg.C)
-0.2
0.2
0.6
1.0
1950
1960
1970
1980
1990
2000
2010

Basis for Least-Squares Regression

Why define the total error of a model by the sum of the squared residuals?

Basis for Least-Squares Regression

Why define the total error of a model by the sum of the squared residuals?

Suppose the data actually equal our model (straight line), plus zero-mean i.i.d. Gaussian white noise. Then if the model is correct, the residuals are zero-mean Gaussian iid noise so the pdf for any single residual R_j is the normal distribution

$$P(\varepsilon_j) = \frac{e^{-\frac{1}{2}R_j^2/\sigma^2}}{\sigma\sqrt{2\pi}}.$$

Basis for Least-Squares Regression

Since the noise values are independent, the joint pdf for all of them taken together is the product of their individual probability densities

$$L(R_1, R_2, ...) = \prod_{j=1}^{N} \left[\frac{e^{-\frac{1}{2} R_j^2/\sigma^2}}{\sigma\sqrt{2\pi}} \right] = \frac{e^{-\frac{1}{2}(R_1^2 + R_2^2 + ...)/\sigma^2}}{\sigma^N (2\pi)^{\frac{1}{2}N}}.$$

Basis for Least-Squares Regression

It is often also useful to define the *log-likelihood function*, which is just the logarithm of that

$$\Lambda(R_1, R_2, ...) = \ln(L(R_1, R_2, ...))$$

$$= -\frac{1}{2\sigma^2}\sum_{j=1}^{N} R_j^2 - N\ln(\sigma) - \frac{1}{2}N\ln(2\pi).$$

Basis for Least-Squares Regression

Now consider, what set of parameters (what regression fit) would give the greatest likelihood of the observed data, i.e., give the greatest value of the likelihood function? Maximizing the likelihood function is equivalent to maximizing the log-likelihood function, and maximizing the log-likelihood is equivalent to minimizing the *negative* of the log-likelihood. Therefore we really want to minimize the quantity

$$-\Lambda(R_1, R_2, ...) = \frac{1}{2\sigma^2} \sum_{j=1}^{N} R_j^2 + N \ln(\sigma) + \frac{1}{2} N \ln(2\pi).$$

Basis for Least-Squares Regression

It turns out that the regression parameters which minimize this are those which minimize the sum in the first term only, i.e., those which minimize

$$\sum_{j=1}^{N} R_j^2.$$

But this is just the sum of the squared residuals. Therefore when the data follow our model plus zero-mean i.i.d. Gaussian noise, least-squares gives the *maximum-likelihood* solution.

Basis for Least-Squares Regression

Least-squares regression is the maximum-likelihood solution when the noise is zero-mean, independent, identically distributed Gaussian noise. Since that's the most common assumption about the noise in time series, the least-squares solution applies in a large number of cases.

Its applicability is even wider, because of a result known as the *Gauss-Markov theorem*.

Basis for Least-Squares Regression

For white noise (doesn't have to be i.i.d. and it doesn't have to be Gaussian), then the least-squares solution is "BLUE," meaning *Best Linear Unbiased Estimator*. "Best" means least-variance, i.e., that the uncertainty in our estimated parameters is as small as possible. "Linear" means that the solution is a linear function of the input data. "Unbiased" means that the expected value of the regression fit is equal to the true regression fit. In a wide variety of cases, we shouldn't expect to do better than least-squares regression. Least-squares regression is the workhorse of regression modelling – and with good reason.

Uncertainty of Regression Parameters

Examples: least squares of random noise, to demonstrate that parameter estimates are random variables.

Uncertainty of Regression Parameters

What is the uncertainty of the estimated parameters (β_o and β_1) from linear regression? Start from the fact that the regression parameters are linear in the input data. For the intercept, e.g., there are coefficients $\psi_j^\dagger$ such that

$$\hat{\beta}_o = \sum_{j=1}^{N} \psi_j^\dagger x_j,$$

where $\hat{\beta}_o$ is the *estimated* intercept. We'll learn a **lot more** about those coefficients $\psi_j^\dagger$ later. (call 'em *projection vector*)

Uncertainty of Regression Parameters

Suppose the data actually follow our model, i.e.

$$x_j = \beta_o + \beta_1 t_j + \varepsilon_j,$$

with ε_j white noise, and β_o is the *true* intercept. Then estimated intercept is

$$\hat{\beta}_o = \sum_{j=1}^{N} \psi_j^\dagger (\beta_o + \beta_1 t_j + \varepsilon_j) = \beta_o \sum_{j=1}^{N} \psi_j^\dagger + \beta_1 \sum_{j=1}^{N} \psi_j^\dagger t_j + \sum_{j=1}^{N} \psi_j^\dagger \varepsilon_j.$$

Uncertainty of Regression Parameters

We'll see (later) that $\psi_j^\dagger$ has some very useful properties, including

$$\sum_{j=1}^{N} \psi_j^\dagger = 1,$$

and

$$\sum_{j=1}^{N} \psi_j^\dagger t_j = 0.$$

Uncertainty of Regression Parameters

Because of those properties, the estimated intercept is

$$\hat{\beta}_o = \beta_o + \sum_{j=1}^{N} \psi_j^\dagger \varepsilon_j .$$

$\langle \varepsilon_j \rangle = 0$ (for all j), so expected value of intercept *estimate* is

$$\langle \hat{\beta}_o \rangle = \beta_o + \sum_{j=1}^{N} \psi_j^\dagger \langle \varepsilon_j \rangle = \beta_o ,$$

i.e. $\hat{\beta}_o$ is an *unbiased* estimate. That's good!

Uncertainty of Regression Parameters

What about its uncertainty? Difference from true value is

$$(\hat{\beta}_o - \beta)^2 = (\sum_{j=1}^{N} \psi_j^\dagger \varepsilon_j)^2 = \sum_{j=1}^{N} \sum_{k=1}^{N} \psi_j^\dagger \psi_k^\dagger \varepsilon_j \varepsilon_k.$$

Now we use the fact that ε_j is white noise, so

$$\langle \varepsilon_j \varepsilon_k \rangle = \sigma^2 \delta_{jk}.$$

Since $\delta_{jk} = 1$ when $j \neq k$, the only terms surviving in the sum are those for $j = k$.

Uncertainty of Regression Parameters

Therefore the variance of the intercept estimate is

$$\sigma^2_{(\beta_o)} = \langle(\hat{\beta}_o - \beta_o)^2\rangle = \sigma^2 \sum_{j=1}^{N} (\psi_j^\dagger)^2,$$

and $\sigma_{(\beta_o)}$ is the square root of that.

Uncertainty of Regression Parameters

There's a *different* "projection vector" $\psi_j^\dagger$ for the *slope* parameter. A similar analysis hows that it too is an unbiased estimate, with variance given by

$$\sigma^2_{(\beta_o)} = \langle(\hat{\beta}_o - \beta_o)^2\rangle = \sigma^2 \sum_{j=1}^{N} (\psi_j^\dagger)^2,$$

(using the *other* $\psi_j^\dagger$, the one for the slope).

Uncertainty of Regression Parameters

The details depend on the quantities $\psi_j^\dagger$, which depend on the times of observation (but not on the data values).

There is an interesting special case: when the mean time is zero, i.e. $\langle t_j \rangle = 0$, we have the case that for the intercept

$$\psi_j^\dagger = \frac{1}{N},$$

Uncertainty of Regression Parameters

In that case, the intercept estimate is

$$\hat{\beta}_o = \sum_{j=1}^{N} \frac{x_j}{N} = \langle x_j \rangle,$$

i.e. the estimated intercept is the average data value (when the average time is zero).

Uncertainty of Regression Parameters

Its variance is then

$$\sigma^2_{(\beta_o)} = \langle(\hat{\beta}_o - \beta_o)^2\rangle = \sigma^2 \sum_{j=1}^{N} \left(\frac{1}{N}\right)^2 = \frac{\sigma^2}{N}.$$

This is the usual expression for the variance of an average, so the estimated intercept is the usual average of x_j and its variance is the usual variance of the average.

Uncertainty of Regression Parameters

We still need an estimate of σ^2 (variance of the white-noise process)! Estimate it as the variance of the residuals, with one exception.

When we estimate the variance of data, we usually use

$$\hat{\sigma}^2_{(x)} = \frac{1}{N-1}\sum_{j=1}^{N}(x_j - \bar{x})^2,$$

and we divide by $N-1$ instead of N, because subtracting the average $\bar{x}$ removes 1 degree of freedom.

Uncertainty of Regression Parameters

For linear regression, removing the linear fit (to generate residuals) removes 2 degrees of freedom (slope and intercept), so we estimate the white-noise variance from the residuals via

$$\hat{\sigma}^2 = \frac{1}{N-2}\sum_{j=1}^{N}(R_j)^2.$$

Note I didn't subtract $\bar{R}$, because the residuals *already* have mean value zero.

Distribution of Regression Parameters

OK, parameters have the given mean (equal to true value) and variance (given by forulae). But what is the *probability distribution*?

Answer: because their deviations are sums of random variables with given coefficients, the *central limit theorem* tells us it is *asymptotically normal*.

Only true asymptotically. Unless: noise is truly iid Gaussian. Then it's truly normal.

Distribution of Regression Parameters

Even when truly normal, the *test statistic* (testing whether it's different from zero)

$$t = \frac{\hat{\beta}}{\hat{\sigma}_{(\beta)}},$$

Isn't normal, because it's the *ratio* of a normal variable ($\hat{\beta}$) to the square root of a chi-square variable ($\hat{\sigma}_{(\beta)}$). That follows the t-distribution.

t is a t-statistic with $N - 2$ degrees of freedom.

Problem 1

Find a time series – one which interests *you*.

Use whatever software you like to use, to fit a linear time trend to those data – one of the form

$$x_j = \beta_o + \beta_1 t_j + \varepsilon_j.$$

Treat the noise ε_j as white noise.

Examine the residuals.

Muse on your results.

Problem 2

Even though we haven't yet studied how (we will in the next lesson), your software can probably fit a more complicated model. Try a quadratic regression of the form

$$x_j = \beta_o + \beta_1 t_j + \beta_2 t_j^2 + \varepsilon_j,$$

and again treat the noise as white noise.

Problem 3

Using the same data, offset the times by one trillion (1,000,000,000,000) so the *new* times are defined by

$$t_{new} = t_{old} + 1000000000000.$$

Repeat the quadratic regression using the new time variable. Discuss the differences introduced by offsetting the times by such a large amount.

Problem 4

Earlier, things simplified when the average time was zero. We can always do that, by offsetting the times to define a new time variable

$$t_{new} = t_{old} - \langle t_{old} \rangle.$$

Why might this be a worthwhile thing to do?

Problem 5

Take your time series from problem 1, re-define time according to problem 4, then perform 5 different regressions:

$$x_j = \beta_o + \beta_1 t_j + \varepsilon_j,$$

$$x_j = \beta_o + \beta_1 t_j + \beta_2 t_j^2 + \varepsilon_j,$$

$$x_j = \beta_o + \beta_1 t_j + \beta_2 t_j^2 + \beta_3 t_j^3 + \varepsilon_j,$$

$$x_j = \beta_o + \beta_1 t_j + \beta_2 t_j^2 + \beta_3 t_j^3 + \beta_4 t_j^4 + \varepsilon_j,$$

$$x_j = \beta_o + \beta_1 t_j + \beta_2 t_j^2 + \beta_3 t_j^3 + \beta_4 t_j^4 + \beta_5 t_j^5 + \varepsilon_j.$$