ORF 245 Fundamentals of Statistics
Chapter 14
Least Squares Regression

Robert Vanderbei

Fall 2014

Slides last edited on December 12, 2014

Least Squares

(Recalling two slides from Chapter 4)

Suppose we know from some underlying fundamental principle (say physics for example) that some parameter y is related linearly to another parameter x:

$$y = \alpha + \beta x$$

but we don't know α and β. We'd like to do experiments to determine them. A probabilistic model of the experiment has X and Y as random variables. Let's imagine we do the experiment over and over many times and have a good sense of the joint distribution of X and Y. We want to pick α and β so as to minimize

$$\mathbb{E}(Y - \alpha - \beta X)^2$$

Again, we take derivatives and set them to zero. This time we have two derivatives:

$$\frac{\partial}{\partial \alpha}\mathbb{E}(Y - \alpha - \beta X)^2 = \mathbb{E}\left(\frac{\partial}{\partial \alpha}(Y - \alpha - \beta X)^2\right) = -2\mathbb{E}(Y - \alpha - \beta X) = -2(\mu_Y - \alpha - \beta \mu_X) = 0$$

and

$$\frac{\partial}{\partial \beta}\mathbb{E}(Y - \alpha - \beta X)^2 = \mathbb{E}\left(\frac{\partial}{\partial \beta}(Y - \alpha - \beta X)^2\right) = -2\mathbb{E}\left((Y - \alpha - \beta X)X\right) = -2\left(\mathbb{E}(XY) - \alpha\mathbb{E}(X) - \beta\mathbb{E}(X^2)\right) = 0$$

Least Squares – Continued

(Recalling two slides from Chapter 4)

We get two linear equations in two unknowns

$$\alpha + \beta\mu_X = \mu_Y$$

$$\alpha\mu_X + \beta\mathbb{E}(X^2) = \mathbb{E}(XY)$$

Multiplying the first equation by μ_X and subtracting it from the second equation, we get

$$\beta\mathbb{E}(X^2) - \beta\mu_X^2 = \mathbb{E}(XY) - \mu_X\mu_Y$$

This equation simplifies to

$$\beta\sigma_X^2 = \sigma_{XY}$$

and so

$$\beta = \frac{\sigma_{XY}}{\sigma_X^2} = \rho\frac{\sigma_Y}{\sigma_X}$$

Finally, substituting this expression into the first equation, we get

$$\alpha = \mu_Y - \rho\frac{\sigma_Y}{\sigma_X}\mu_X$$

As before, suppose that it is known (or believed) that there is a simple linear relation between two measurable quantities x and y:

$$y = \alpha + \beta x$$

Suppose further that x can be set with arbitrary precision but that y is measured with some error. A statistical model for such as situation involves random variables:

$$Y = \alpha + \beta x + \varepsilon$$

Here, ε and (therefore) Y are random variables.

Our aim is to derive estimates for α and β based on a sampling of size n.

We assume that x can be varied as we like in the sampling process. So, we have n samples like this:

$$Y_i = \alpha + \beta x_i + \varepsilon_i, \qquad i = 1, 2, \ldots, n$$

We wish to choose the α and β that give the "best fit" to the sample. In other words, we choose them so as to minimize

$$\sum_i \varepsilon_i^2 = \sum_i (Y_i - \alpha - \beta x_i)^2$$

Minimizing the Sum of Squares

To minimize, we take the derivatives with respect to α and β and set them to zero:

$$\begin{aligned}\frac{\partial}{\partial\alpha}\sum_i (Y_i-\alpha-\beta x_i)^2 &= \sum_i 2\,(Y_i-\alpha-\beta x_i)\,(-1) = 0\\ \frac{\partial}{\partial\beta}\sum_i (Y_i-\alpha-\beta x_i)^2 &= \sum_i 2\,(Y_i-\alpha-\beta x_i)\,(-x_i) = 0\end{aligned}$$

Dividing both sides by $-2n$, these equations simplify to

$$\begin{aligned}\overline{Y}-\alpha-\beta\overline{x} &= 0\\ \overline{xY}-\alpha\overline{x}-\beta\overline{x^2} &= 0\end{aligned}$$

where

$$\overline{x}=\frac{1}{n}\sum_i x_i, \qquad \overline{Y}=\frac{1}{n}\sum_i Y_i, \qquad \overline{xY}=\frac{1}{n}\sum_i x_iY_i, \qquad \overline{x^2}=\frac{1}{n}\sum_i x_i^2$$

This is two equations in two unknowns. Multiplying the first equation by $\overline{x}$ and subtracting that from the second equation, we can solve for β and, knowing β, we can use the first equation to solve for α:

$$\hat{\beta}=\frac{\overline{xY}-\overline{x}\,\overline{Y}}{\overline{x^2}-\overline{x}^2}, \qquad \hat{\alpha}=\overline{Y}-\hat{\beta}\overline{x}$$

Mean of $\hat{\alpha}$ and $\hat{\beta}$

Recalling that $\hat{\beta} = \frac{\overline{xY} - \overline{x}\,\overline{Y}}{\overline{x^2} - \overline{x}^2}$ and $\hat{\alpha} = \overline{Y} - \hat{\beta}\overline{x}$

We first compute the expected value of $\hat{\beta}$:

$$\begin{aligned}
\mathbb{E}(\hat{\beta}) &= \frac{\mathbb{E}(\overline{xY}) - \overline{x}\,\mathbb{E}(\overline{Y})}{\overline{x^2} - \overline{x}^2} \\
&= \frac{\mathbb{E}(\frac{1}{n}\sum_i x_i Y_i) - \overline{x}\,\mathbb{E}(\frac{1}{n}\sum_i Y_i)}{\overline{x^2} - \overline{x}^2} \\
&= \frac{\frac{1}{n}\sum_i x_i \mathbb{E}(Y_i) - \overline{x}\,\frac{1}{n}\sum_i \mathbb{E}(Y_i)}{\overline{x^2} - \overline{x}^2} \\
&= \frac{\frac{1}{n}\sum_i x_i(\alpha + \beta x_i) - \overline{x}\,\frac{1}{n}\sum_i(\alpha + \beta x_i)}{\overline{x^2} - \overline{x}^2} \\
&= \beta
\end{aligned}$$

Next, we compute the expected value of $\hat{\alpha}$:

$$\begin{aligned}
\mathbb{E}(\hat{\alpha}) &= \mathbb{E}(\overline{Y}) - \mathbb{E}(\hat{\beta})\,\overline{x} \\
&= \mathbb{E}(\frac{1}{n}\sum_i Y_i) - \beta\,\overline{x} \\
&= \frac{1}{n}\sum_i \mathbb{E}(Y_i) - \beta\,\overline{x} \\
&= \frac{1}{n}\sum_i(\alpha + \beta x_i) - \beta\,\overline{x} \\
&= \alpha
\end{aligned}$$

The estimators are *unbiased!*

Variance of $\hat{\alpha}$ and $\hat{\beta}$

Again, recall that $\hat{\beta} = \frac{\overline{xY} - \overline{x}\,\overline{Y}}{\overline{x^2} - \overline{x}^2}$ and $\hat{\alpha} = \overline{Y} - \hat{\beta}\overline{x}$

To compute the variance of $\hat{\alpha}$ and $\hat{\beta}$, we need these computations:

$$\overline{xY} - \overline{x}\,\overline{Y} = \frac{1}{n}\sum_i x_i Y_i - \overline{x}\frac{1}{n}\sum_i Y_i = \frac{1}{n}\sum_i (x_i - \overline{x})\, Y_i$$

$$\text{var}\left(\overline{xY} - \overline{x}\,\overline{Y}\right) = \text{var}\left(\frac{1}{n}\sum_i (x_i - \overline{x})\, Y_i\right) = \frac{1}{n^2}\sum_i (x_i - \overline{x})^2\,\sigma^2$$

$$\hat{\alpha} = \overline{Y} - \frac{\overline{xY} - \overline{x}\,\overline{Y}}{\overline{x^2} - \overline{x}^2}\overline{x} = \frac{\overline{x^2}\,\overline{Y} - \overline{x}\,\overline{xY}}{\overline{x^2} - \overline{x}^2} = \frac{\frac{1}{n}\sum_i \left(\overline{x^2} - \overline{x}\,x_i\right) Y_i}{\overline{x^2} - \overline{x}^2}$$

Using these formulas, we get

$$\text{var}(\hat{\beta}) = \frac{\text{var}(\overline{xY} - \overline{x}\overline{Y})}{\left(\overline{x^2} - \overline{x}^2\right)^2} = \frac{\sigma^2}{n}\frac{\frac{1}{n}\sum_i (x_i - \overline{x})^2}{\left(\overline{x^2} - \overline{x}^2\right)^2} = \frac{\sigma^2}{n}\frac{1}{\overline{x^2} - \overline{x}^2}$$

and

$$\text{var}(\hat{\alpha}) = \frac{\frac{1}{n^2}\sum_i \left(\overline{x^2} - \overline{x}\,x_i\right)^2\sigma^2}{\left(\overline{x^2} - \overline{x}^2\right)^2} = \frac{\sigma^2}{n}\frac{\frac{1}{n}\sum_i \left(\overline{x^2} - \overline{x}\,x_i\right)^2}{\left(\overline{x^2} - \overline{x}^2\right)^2} = \cdots = \frac{\sigma^2}{n}\frac{\overline{x^2}}{\overline{x^2} - \overline{x}^2}$$

Confidence Intervals for α and β

Subtracting the mean and dividing by the standard deviation, we see that the random variables

$$\frac{\hat{\alpha}-\alpha}{\sqrt{\text{var}(\hat{\alpha})}} = \frac{\hat{\alpha}-\alpha}{\frac{\sigma}{\sqrt{n}}\frac{\sqrt{\overline{x^2}}}{\sqrt{\overline{x^2}-\overline{x}^2}}} \quad \text{and} \quad \frac{\hat{\beta}-\beta}{\sqrt{\text{var}(\hat{\beta})}} = \frac{\hat{\beta}-\beta}{\frac{\sigma}{\sqrt{n}}\frac{1}{\sqrt{\overline{x^2}-\overline{x}^2}}}$$

have *mean zero* and *variance one*. If the ε_i's are *Normally* distributed (with mean 0 and variance σ^2), then the above random variables are standard Normal random variables.

We don't know σ^2. It can be shown that

$$S^2 = \frac{1}{n-2}\sum_i \left(Y_i - \hat{\alpha} - \hat{\beta}x_i\right)^2$$

is an *unbiased* estimate of σ^2 and the corresponding normalized rv's are approximately normally distributed (t-distribution with $n-2$ degrees of freedom if n is small).

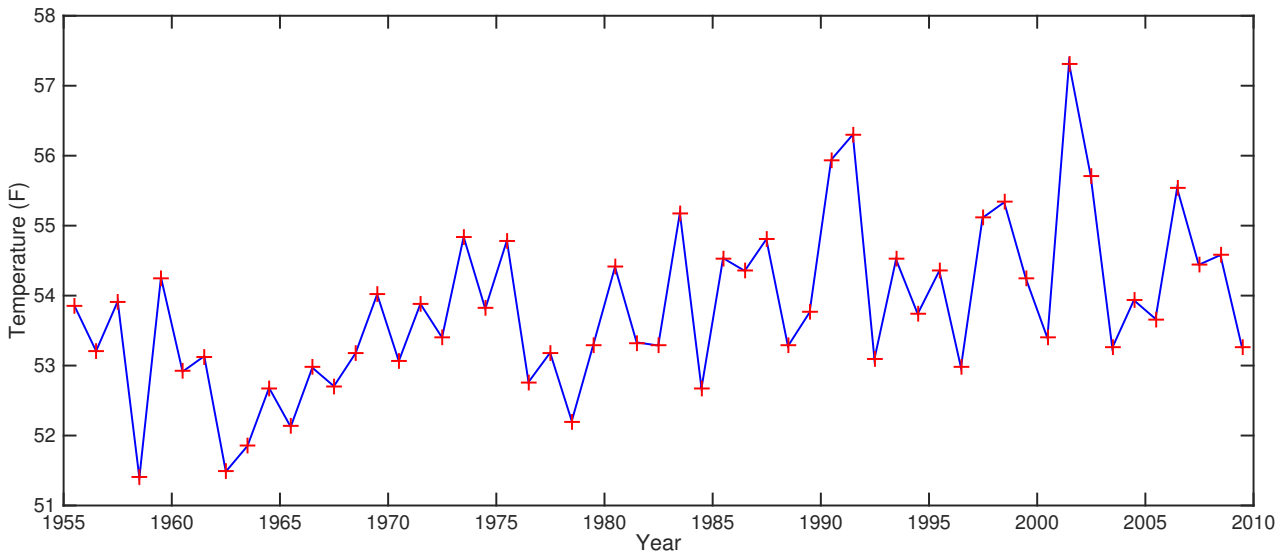
Putting this all together, we get the following 95% *confidence intervals* for $\hat{\alpha}$ and $\hat{\beta}$:

$$\alpha \in \hat{\alpha} \pm 1.96\,\frac{S}{\sqrt{n}}\,\frac{\sqrt{\overline{x^2}}}{\sqrt{\overline{x^2}-\overline{x}^2}} \qquad \beta \in \hat{\beta} \pm 1.96\,\frac{S}{\sqrt{n}}\,\frac{1}{\sqrt{\overline{x^2}-\overline{x}^2}}$$

If n is small (less, say, that 30) then replace 1.96 with $t_{n-2}(0.025)$.

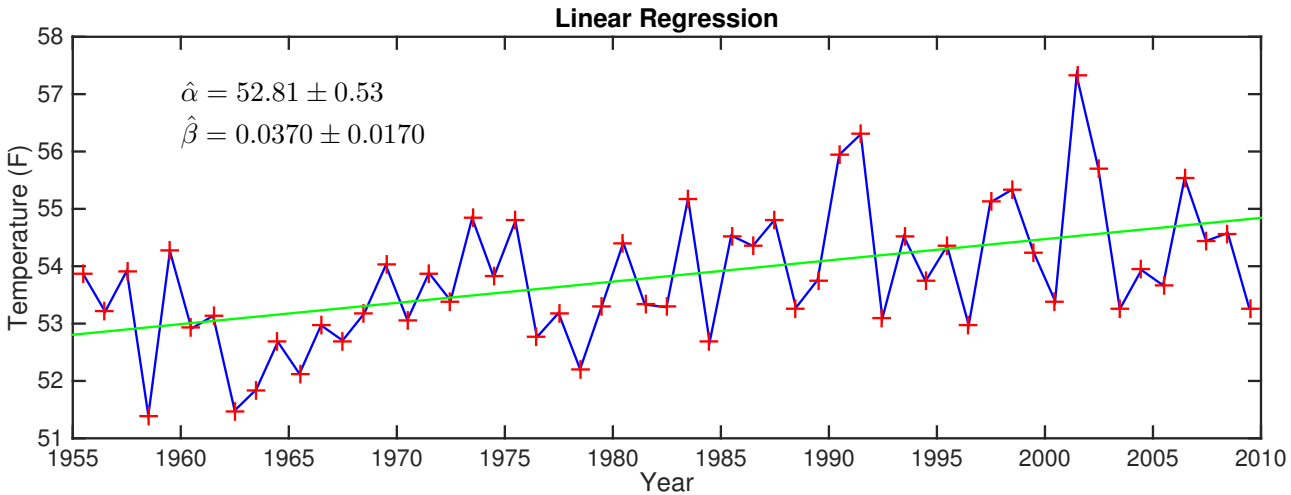
Back to Temp vs. Time at McGuire AFB

Recall from Chapter 10 our plot of one-year averages:



This is $n = 55$ data points where "x" is the year (with, say, year 0 being 1955) and "y" is average temperature for year x.

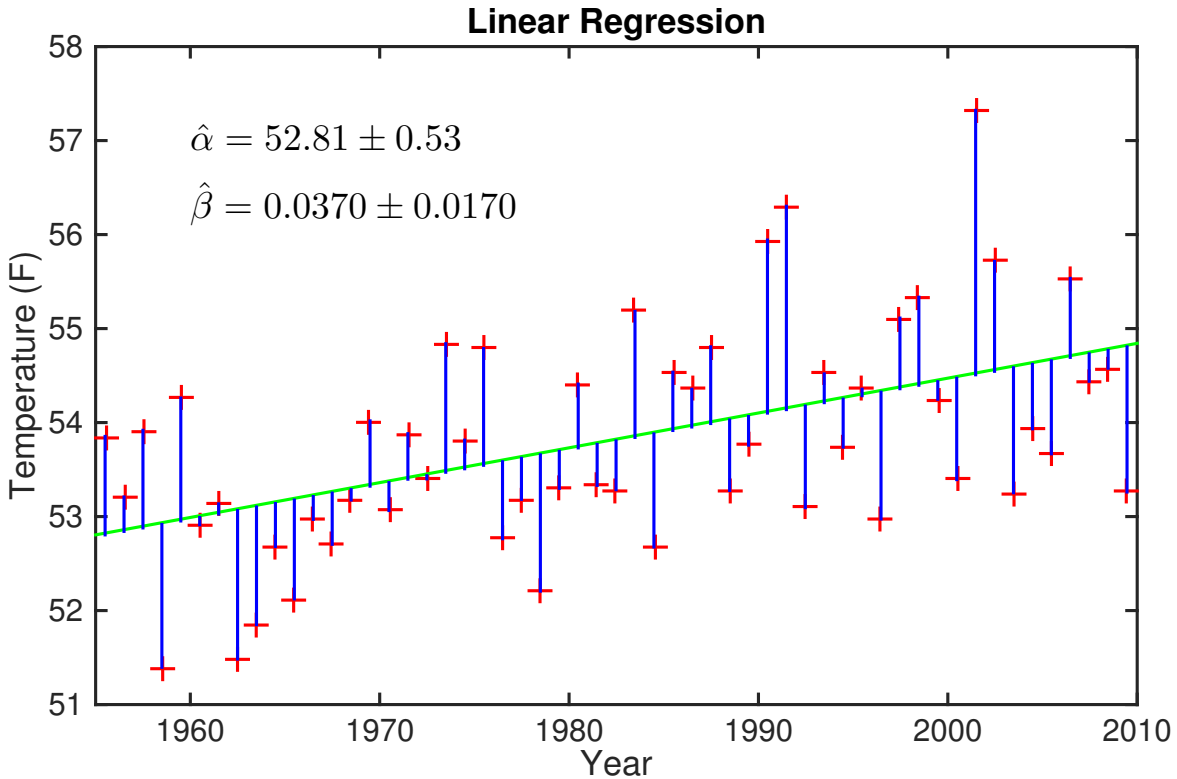
Temp vs. Time – Regression Fit




```
T = McGuireAFB(:,2);
[n m] = size(T);
T = T(1: 365*floor(n/365));
T = reshape(T, 365, floor(n/365));
y = mean(T)';
[n m] = size(y);
x = (0:(n-1))';
xbar = sum(x)/n;
ybar = sum(y)/n;
xybar = sum(x.*y)/n;
x2bar = sum(x.*x)/n;
beta = (xybar - xbar*ybar)/(x2bar - xbar^2)
alpha = ybar - beta*xbar
plot( 1955+x, y,'b-', 1955+x, y,'r+', ...
  [1955 2010], [alpha alpha+beta*(2010-1955)], 'g-');
```

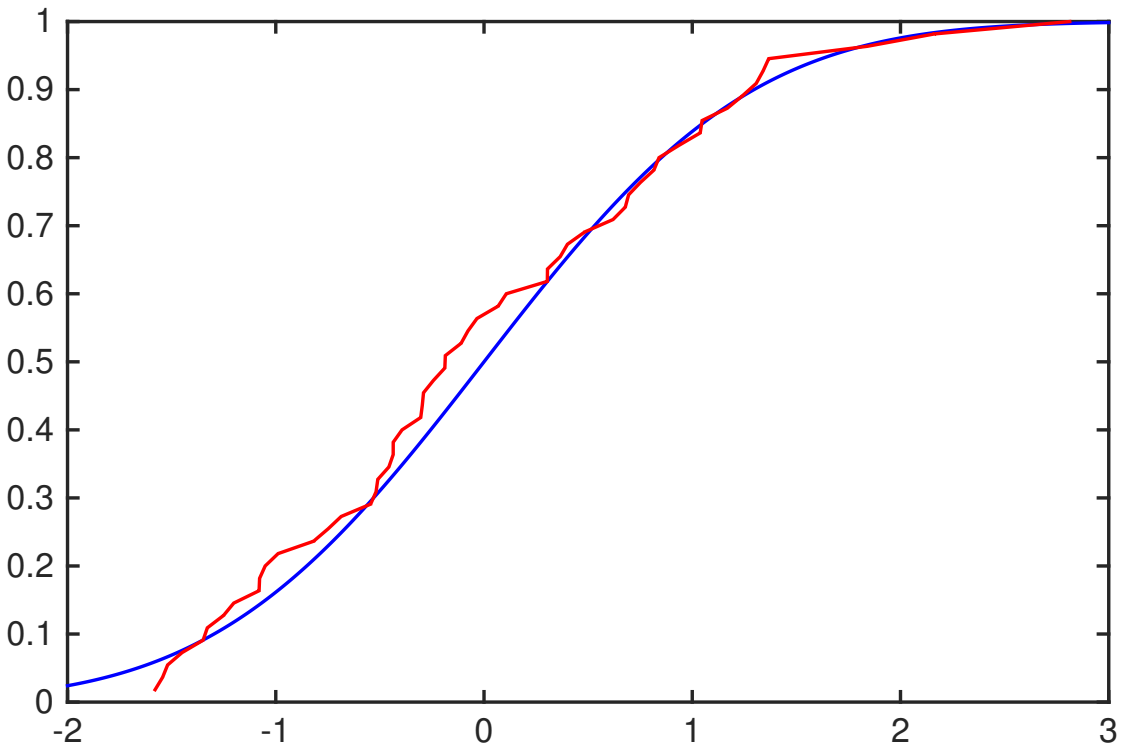
```
% Code for confidence interval...

eps = y - alpha - beta*x;
S = sqrt(sum(eps.^2)/(n-2));
alpha_width = 1.96 * (S/sqrt(n)) ...
    * (sqrt(x2bar))/(sqrt(x2bar-xbar^2));
beta_width = 1.96 * ...
    (S/sqrt(n)) /(sqrt(x2bar-xbar^2));
```

Temp vs. Time – Epsilons

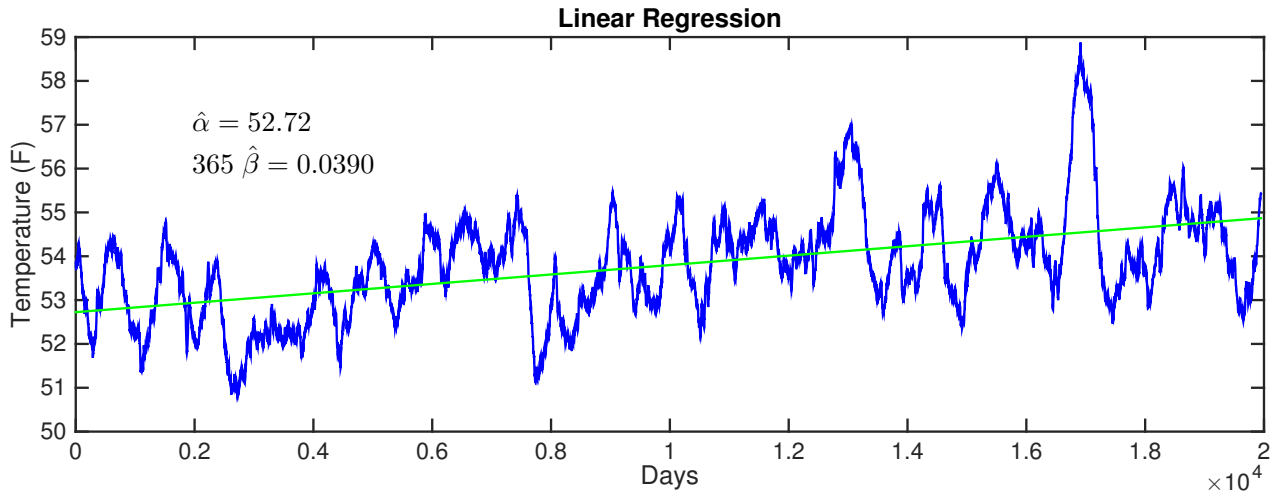


Temp vs. Time – Regression Fit – Goodness of Fit

Here we plot the empirical cdf for the ε_i's on top of the cdf of a normal distribution with the same mean and variance:



Temp vs. Time – Regression Fit – Rolling Average




```
T = McGuireAFB(:,2);
window = ones(365,1)/365;
y = conv(T,window,'valid');
[n m] = size(y);
x = (0:(n-1))';
xbar = sum(x)/n;
ybar = sum(y)/n;
xybar = sum(x.*y)/n;
x2bar = sum(x.*x)/n;
betahat = (xybar - xbar*ybar)/(x2bar - xbar^2)
alphahat = ybar - betahat*xbar
plot(x,y,'b-', [x(1) x(end)], [alphahat alphahat+betahat*365*(2010-1955)], 'g');
```

Note: Confidence interval not computed because ε_i's are not independent.

Temp vs. Time – Multiple Linear Regression

Let's take into account the *seasonal oscillations*:

$$y_i = \beta_1 + \beta_2 x_i + \beta_3 \cos(2\pi x_i/365.25) + \beta_4 \sin(2\pi x_i/365.25) + \varepsilon_i \qquad i = 1, 2, \ldots, n$$

Again, we look for values of the parameters that minimize the sum of the squares of the ε_i's.

This time, however, we have to set four derivatives equal to zero. That's four equations in four unknowns. Not fun. Fortunately, Matlab provides a tool precisely for this purpose. We put the temperatures in a column vector y,

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix},$$

and we put the numbers that multiply the β_j's into a matrix X with n rows and 4 columns

$$X = \begin{bmatrix} 1 & x_1 & \cos(2\pi x_1/365.25) & \sin(2\pi x_1/365.25) \\ 1 & x_2 & \cos(2\pi x_2/365.25) & \sin(2\pi x_2/365.25) \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_n & \cos(2\pi x_n/365.25) & \sin(2\pi x_n/365.25) \end{bmatrix}.$$

Multiple Linear Regression – Continued

It may look complicated but y and X are just a vector and a matrix of numbers. Here's a few entries from y and X:

$$y = \begin{bmatrix} 37.6 \\ 43.2 \\ 40.0 \\ 42.2 \\ 30.0 \\ 36.3 \\ \vdots \\ 81.2 \\ 82.0 \\ 75.2 \\ 70.5 \end{bmatrix} \qquad X = \begin{bmatrix} 1 & 1 & 0.9999 & 0.0172 \\ 1 & 2 & 0.9994 & 0.0344 \\ 1 & 3 & 0.9987 & 0.0516 \\ 1 & 4 & 0.9976 & 0.0688 \\ 1 & 5 & 0.9963 & 0.0859 \\ 1 & 6 & 0.9947 & 0.1030 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 20311 & -0.7765 & -0.6301 \\ 1 & 20312 & -0.7656 & -0.6433 \\ 1 & 20313 & -0.7544 & -0.6564 \\ 1 & 20314 & -0.7430 & -0.6693 \end{bmatrix}$$

And, our statistical model can be written in this matrix notation:

$$y = X\beta + \varepsilon$$

where

$$\beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{bmatrix}$$

and ε is a long vector of "errors".

Multiple Linear Regression – Matlab

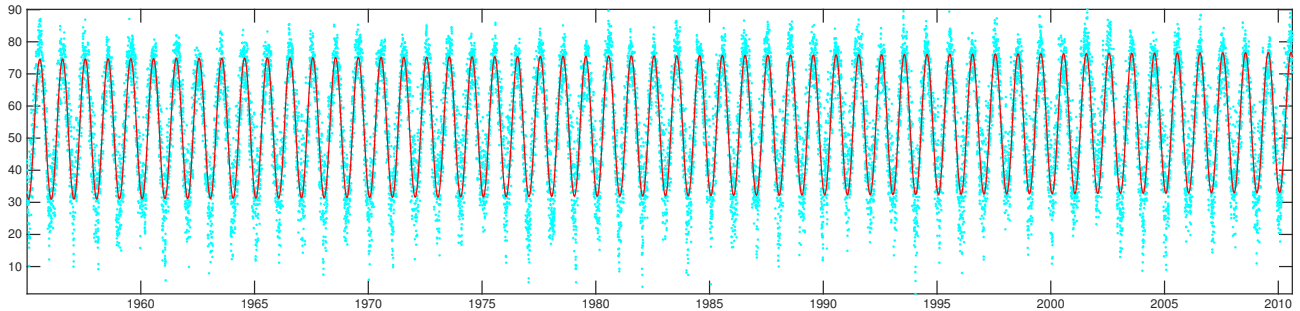
Linear least squares is so important that Matlab has made it especially easy to compute the vector β that minimizes the sum of the squares of the ε's:

$$\hat{\beta} = X \backslash y$$

Here's the full Matlab code used to make the plot on the following page:

```
load -ascii data/McGuireAFB.dat;
y = McGuireAFB(:,2);
[n m] = size(y);

% least squares regression

dt = (1:n)';
X = [ ones(n,1) dt cos(2*pi*dt/365.25) sin(2*pi*dt/365.25) ];
betahat = X\y
betahat(2)*356.25*100

plot( 1955+dt/365.25, y, 'b+', 1955+dt/365.25, X*betahat, 'r-');
```

Multiple Linear Regression – Output



The numerical answers are:

$$\hat{\beta}_1 = 52.7400, \qquad \hat{\beta}_2 = 0.00010618, \qquad \hat{\beta}_3 = -20.2127, \qquad \hat{\beta}_4 = -8.1828$$

The linear trend is encoded in $\hat{\beta}_2$.

The value shown above is in degrees Fahrenheit per day.

To convert to degrees per century, we need to multiply by 100×365.25.

We get

$$\text{linear trend} = 100 \times 365.25 \times \hat{\beta}_2 = 3.8781^{\circ}\text{F/century}$$

Confidence Intervals via Bootstrap

It's possible, but tedious, to derive formulas for the variances (and covariances) of the $\hat{\beta}_i$'s in the climate model.

An alternate method for producing confidence intervals is called *Bootstrap*.

The idea is as follows.

When n is large (say, $n = 20{,}314$, as in the climate model), then we can use the computed regression coefficients to produce a large empirical pool of ε's:

$$\hat{\varepsilon}_i = y_i - \left(\hat{\beta}_1 + \hat{\beta}_2 x_i + \hat{\beta}_3 \cos(2\pi x_i/365.25) + \hat{\beta}_4 \sin(2\pi x_i/365.25)\right)$$

In other words,

$$\hat{\varepsilon} = y - X\hat{\beta}$$

By sampling (with replacement) from this pool of $\hat{\varepsilon}_i$'s, we can artificially generate new data sets that are statistically similar to the original one:

$$\tilde{y} = X\hat{\beta} + \tilde{\varepsilon}$$

(here, $\tilde{\varepsilon}_i$ is one of the $\hat{\varepsilon}_j$'s drawn at random from the pool of such epsilons). Using this new data set, we can compute new values for the estimators:

$$\tilde{\beta} = X \backslash \tilde{y}$$

Confidence Intervals via Bootstrap – Continued

We can do this over and over again to generate a large sample of new estimators. The expected value of these new estimators is unchanged from the original. But, they won't be all the same and therefore we can use them to compute an empirical standard deviation and use that to compute confidence intervals.

The Matlab code to do this bootstrap is rather straightforward:

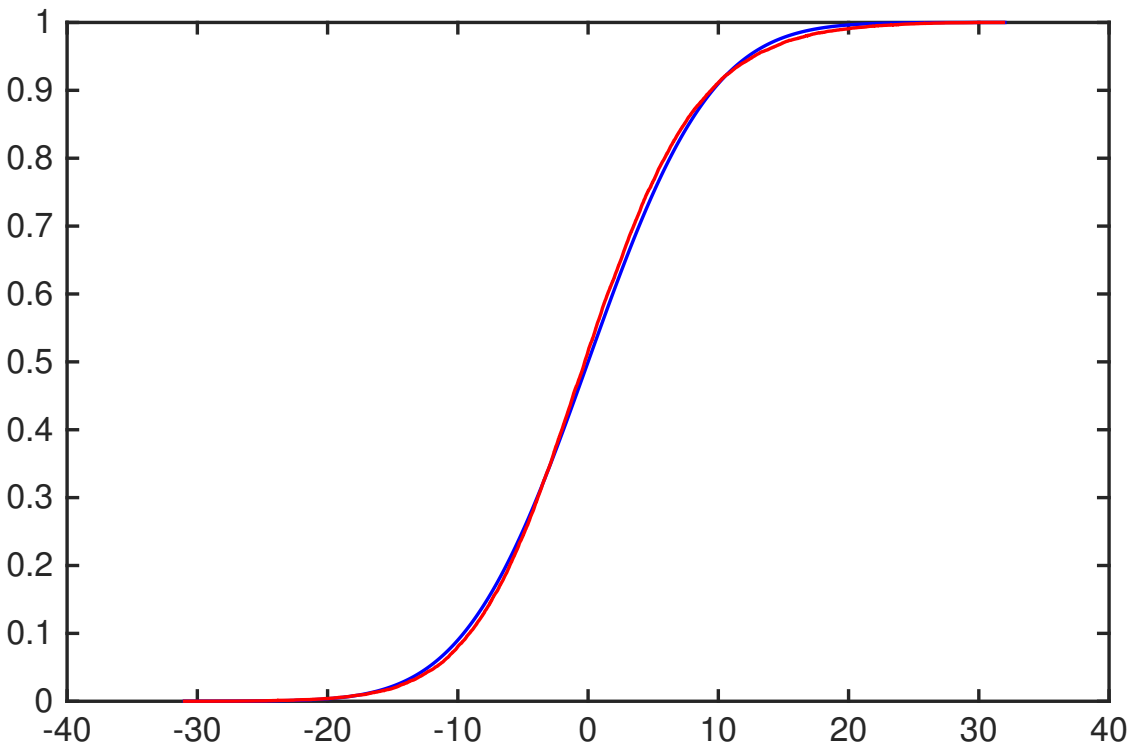
```
eps = y - X*beta;
beta_new = zeros(4,100);
indices = randi(n,n);  % an nxn matrix of random numbers between 1 and n
for i=1:100
    y_new = X*beta + eps(indices(:,i));
    beta_new(:,i) = X\y_new;
end
betahat = mean(beta_new');
betahatstd = std(beta_new');
36525 * betahat(2)
36525 * 1.96*betahatstd(2)
```

And, the answer is...

$$100 \times 365.25 \times \beta_2 = 3.8662 \pm 0.6635^{\circ}\text{F/century}$$

Goodness of Fit

As before, we plot the empirical cdf for the ε_i's on top of the cdf of a normal distribution with the same mean and variance:



A Few Final Remarks on Climate Data

The climate data for McGuire AFB can be downloaded from here...

http://www.princeton.edu/~rvdb/ampl/nlmodels/LocalWarming/McGuireAFB/data/McGuireAFB.txt

Here is a published paper that describes the data and its analysis in some detail...

http://www.princeton.edu/~rvdb/tex/LocalWarming/LocalWarmingSIREVrev.pdf

Had Linear Algebra Been A Prereq...

The course is over. What follows is a short advertisement for ORF 405.

Had Linear Algebra Been A Prereq...

Regression model: $y = X\beta + \varepsilon$.
Find β that minimizes sum of squared errors:

$$\hat{\beta} \;=\; \text{argmin}_\beta \|y - X\beta\|^2 \;=\; \text{argmin}_\beta (y - X\beta)^T (y - X\beta) \;=\; \text{argmin}_\beta f(\beta)$$

where

$$f(\beta) = y^T y - \beta^T X^T y - y^T X\beta + \beta^T X^T X\beta.$$

Take the gradient with respect to β and equate to zero:

$$X^T X\hat{\beta} = X^T y \qquad \Longrightarrow \qquad \hat{\beta} = (X^T X)^{-1} X^T y$$

The estimator is unbiased:

$$\mathbb{E}(\hat{\beta}) \;=\; (X^T X)^{-1} X^T \mathbb{E}(y) \;=\; (X^T X)^{-1} X^T X\beta = \;\beta$$

The covariances among all of the coefficients is also easy to compute:

$$\begin{aligned}
\text{Cov}(\hat{\beta}) \;&=\; \mathbb{E}(\hat{\beta}\hat{\beta}^T) - \beta\beta^T \;=\; \mathbb{E}\left((X^T X)^{-1} X^T y y^T X (X^T X)^{-1}\right) - \beta\beta^T \\
&=\; \mathbb{E}\left((X^T X)^{-1} X^T (X\beta + \varepsilon)(X\beta + \varepsilon)^T X (X^T X)^{-1}\right) - \beta\beta^T \\
&=\; \sigma^2 (X^T X)^{-1}
\end{aligned}$$

The 2-Dimensional Case

Suppose that

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} \qquad \text{and} \qquad X = \begin{bmatrix} e & x \end{bmatrix}$$

where e denotes a column vector of all ones and x is a column vector containing the x_i's. In this case,

$$X^T X = \begin{bmatrix} e^T \\ x^T \end{bmatrix} \begin{bmatrix} e & x \end{bmatrix} = \begin{bmatrix} n & e^T x \\ e^T x & x^T x \end{bmatrix}$$

Hence

$$\begin{aligned} \sigma^2 (X^T X)^{-1} &= \sigma^2 \begin{bmatrix} x^T x & -e^T x \\ -e^T x & n \end{bmatrix} \frac{1}{n x^T x - (e^T x)^2} \\ &= \sigma^2 \begin{bmatrix} n\overline{x^2} & -n\overline{x} \\ -n\overline{x} & n \end{bmatrix} \frac{1}{n^2 \overline{x^2} - (n\overline{x})^2} \\ &= \frac{\sigma^2}{n} \begin{bmatrix} \overline{x^2} & -\overline{x} \\ -\overline{x} & 1 \end{bmatrix} \frac{1}{\overline{x^2} - (\overline{x})^2} \end{aligned}$$

The entries on the diagonal are consistent with the formulas we derived before.

Back To The k-Dimensional Case

Recall:

$$y = X\beta + \varepsilon, \qquad \mathbb{E}(\varepsilon) = 0$$
$$\hat{\beta} = (X^T X)^{-1} X^T y, \qquad \mathbb{E}(\hat{\beta}) = \beta$$

Let

$$\hat{\varepsilon} \;=\; y - X\hat{\beta} \;=\; y - X(X^T X)^{-1} X^T y \;=\; Py \;=\; P(X\beta + \varepsilon)$$

where

$$P \;=\; I - X(X^T X)^{-1} X^T \;=\; UU^T$$

where U is a $n \times (n-k)$ orthonormal matrix ($U^T U = I$). Note that $PX = 0$ and $P^T P = P$. We compute...

$$\begin{aligned}\mathbb{E}\left(\|\hat{\varepsilon}\|^2\right) \;&=\; \mathbb{E}\left(P(X\beta + \varepsilon)\right)^T \left(P(X\beta + \varepsilon)\right) \;=\; \beta^T X^T P X \beta + \mathbb{E}\varepsilon^T P^T P \varepsilon \\ &=\; \mathbb{E}\varepsilon^T P \varepsilon \;=\; \mathbb{E}\varepsilon^T U U^T \varepsilon \;=\; \cdots \;=\; (n-k)\sigma^2\end{aligned}$$

Therefore, $S_n = \frac{1}{n-k}\|\hat{\varepsilon}\|^2 = \frac{1}{n-k}\|y - X\hat{\beta}\|^2$ is an *unbiased estimator of* σ^2.