

Lab 1. MVN and Regression

Math 586. Due March 2

Conditioning of MVN

Suppose we would like to solve the problem posed at the start of Lecture 5: given the joint distribution of three MVN variables $\mathbf{Y} = (Y_1, Y_2, Y_3)'$ and knowing the values of Y_2, Y_3 , predict Y_1 . We have learned that the best unbiased estimate was in fact linear, and obtained the formulas for the predictive mean and predictive variance.

Example 1. Let's look at the daily temperature time series at Sevilleta, for about 4 years. First, we will ignore the trend (see Exercise 2 below). We can estimate the **autocovariance matrix** for \mathbf{Y} as follows

```
Y = load('./data/st1.txt');
plot(Y)
T = length(Y);
scatter(Y(1:T-1), Y(2:T))           % reveals autocorrelation
Sig = cov( [Y(1:T-2) Y(2:T-1) Y(3:T)] ) % produces covariance matrix
```

This gives us the estimated

$$\Sigma = \begin{bmatrix} 357.9982 & 299.5226 & 240.9003 \\ 299.5226 & 356.4208 & 298.2395 \\ 240.9003 & 298.2395 & 355.3958 \end{bmatrix}$$

For the ease of discussion let's assume that the means of Y 's equal 0 (no trend). To obtain the best estimate of Y_t based on Y_{t-1}, Y_{t-2} , we have to find a_1, a_2 that minimize MSE for

$$\hat{Y}_t = a_1 Y_{t-1} + a_2 Y_{t-2} \quad (1)$$

This is achieved when (see p.4, Lecture 5)

$$\mathbf{a} = \Sigma_{12} \Sigma_{22}^{-1},$$

with the MSE indicating quality of our forecast

$$\text{MSE} = \text{Var}(Y_t | Y_{t-1}, Y_{t-2}) = \text{Var}(Y_t) - \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21}$$

Note that if we were only forecasting $Y_t = 0$ (value equal to its mean, which is the safe bet in the absence of any data), our MSE would equal to the variance of Y_t , that is about 358. Also note $\Sigma_{21} = \Sigma'_{12}$.

Equation 1 is called autoregression equation.

```

Sig12 = Sig(1, 2:3);
Sig22 = Sig(2:3,2:3);
a = Sig12 * Sig22^(-1)
MSE = Sig(1,1) - Sig12 * Sig22^(-1) * Sig12'

```

To obtain the final forecast for Y_t , use the equation (1).

Multiple regression

Let's consider the example of trend estimation for median grain size (Lecture 6, p.4). The data can be found at <http://infohost.nmt.edu/~olegm/586/data/Reg1.txt>

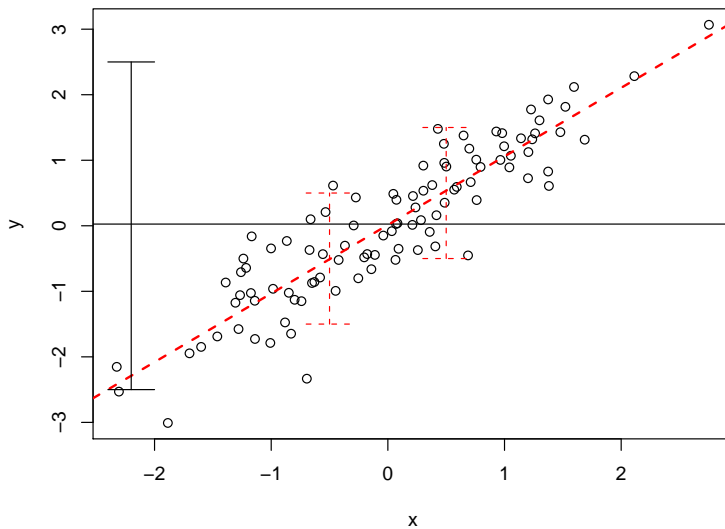
```

data = load('./data/Reg1.txt');
Y = data(:,3);
n = length(Y);
X = [ones(n,1) data(:,1:2)];
betahat = (X' * X)^(-1) * X' * Y;
Yhat = X * betahat;    % predicted values

```

R^2 computation

As mentioned in Lecture 6, one of the most important characteristics of performance of regression is, along with MSE, the “coefficient of determination” R^2 . It equals the “percent of variation in Y explained by its linear relationship with predictors X_1, \dots, X_p ”. To understand the concept, think about simple linear regression and compare two error bars:



By definition,

$$R^2 = \frac{\text{SSReg}}{\text{SST}} = \frac{\text{SST} - \text{SSE}}{\text{SST}} = \frac{\sum_{i=1}^N (Y_i - \bar{Y})^2 - \sum_{i=1}^N (Y_i - \hat{Y}_i)^2}{\sum_{i=1}^N (Y_i - \bar{Y})^2}$$

SSE is the sum of squared residuals and SST plays the role in estimating the variance of Y .

Smaller SSE leads to higher R^2 . The following simple code will find R^2 :

```
resid = Y - Yhat;  
SSE = sum(resid.^2)  
SST = sum((Y-mean(Y)).^2);  
Rsquared = (SST - SSE)/SST
```

We can use R^2 to compare the performance of different models. For example, does including quadratic terms into regression make a big difference? Later, we will learn a more precise technique to answer this question.

Exercises:

1. Consider the best prediction of the next day's average temperature based on Example 1.

- (a) Suppose now you are only using the previous day's data: find coefficient \tilde{a}_1 such that

$$\hat{Y}_t = \tilde{a}_1 Y_{t-1}.$$

Compare the resulting MSE with the one obtained from 2-day prediction $\hat{Y}_t = a_1 Y_{t-1} + a_2 Y_{t-2}$.

- (b) Now, suppose that you have data available on Y_{t-1} and Y_{t+1} , and you'd like to restore a missing observation at the day t :

$$\hat{Y}_t = a_1^* Y_{t-1} + a_2^* Y_{t+1}$$

Find the best coefficients a_1^* , a_2^* and compare the resulting MSE with the one obtained from 2-day prediction $\hat{Y}_t = a_1 Y_{t-1} + a_2 Y_{t-2}$.

2. Now let's fit the seasonal trend of temperature data we previously ignored. Since the temperatures appear to follow a sinusoidal trend, stack the second column of \mathbf{X} -matrix with $\sin\left(\frac{2\pi t}{365}\right)$ and the third column with $\cos\left(\frac{2\pi t}{365}\right)$. (We don't expect any linear trend, but we can later

test for it. Use 365.25 if you're concerned about leap years.)
 This leads to the model

$$\hat{Y}_t = \beta_0 + \beta_1 \sin\left(\frac{2\pi t}{365}\right) + \beta_2 \cos\left(\frac{2\pi t}{365}\right)$$

- (a) Run the *linear* regression with the matrix \mathbf{X} described above. What are the estimates for $\boldsymbol{\beta}$? Obtain a scatterplot with trend line overlaid.
- (b) using the identity $\sin(A - B) = \sin(A)\cos(B) - \cos(A)\sin(B)$, find the *amplitude* β_1^* and the *phase* ϕ_0 in the *non-linear* regression

$$\hat{Y}_t = \beta_0 + \beta_1^* \sin\left(\frac{2\pi(t - \phi_0)}{365}\right)$$

Based on ϕ_0 , which day of the year is typically the hottest?

- (c) Is there still structure left in the residuals? Investigate and display the appropriate plot(s).
 - (d) Recalculate the autocovariance matrix for the residuals. How did it change compared to $\boldsymbol{\Sigma}$ on page 1?
3. Fit the quadratic trend to the grain size data. [Hint: append columns to \mathbf{X} that contain x- and y- coordinates squared and their cross-product.] Does the inclusion of quadratic terms lead to a significant improvement of the model?
 4. High Plains (Kansas) aquifer data at
http://infohost.nmt.edu/~olegm/586/data/hi_plain.txt

The variables are ID, County, Well, Longitude, Latitude, Easting (miles), Northing (miles), Land Surface Elevation (ft), Water Table Elevation (ft), Water Depth (ft).

Consider predictors $X_1 = \text{Easting}$, $X_2 = \text{Northing}$, and response variable $Y = \text{Water Table Elevation}$ (9th column).

Estimate the linear trend. Does there appear to be a quadratic trend?