Multiple Regression
Part 1: Introduction

STAT 705: Regression and Analysis of Variance

Multiple Linear Regression

- Simple Linear Regression
  - One predictor variable X
  - Probably too simplistic (imprecise)
  - Inadequate description of the behavior of Y
- Multiple Linear Regression
  - Multiple predictor variables: X_1, X_2, X_3, …X_p
  - Can model curved relationships between Y and X's
  - Can attain much more precise inference
  - Can accommodate non-numeric (qualitative) predictors
  - Can incorporate interactions between predictors

Multiple Regression With Two Predictors

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i \qquad i = 1, \ldots n$$

Where

Y_i = observed value of the response variable for subject i

X_{1i} = value of first predictor variable recorded on subject i

X_{2i} = value of second predictor variable recorded in subject i

β_0 = the intercept

β_1 = the slope on X_1

β_2 = the slope on X_2

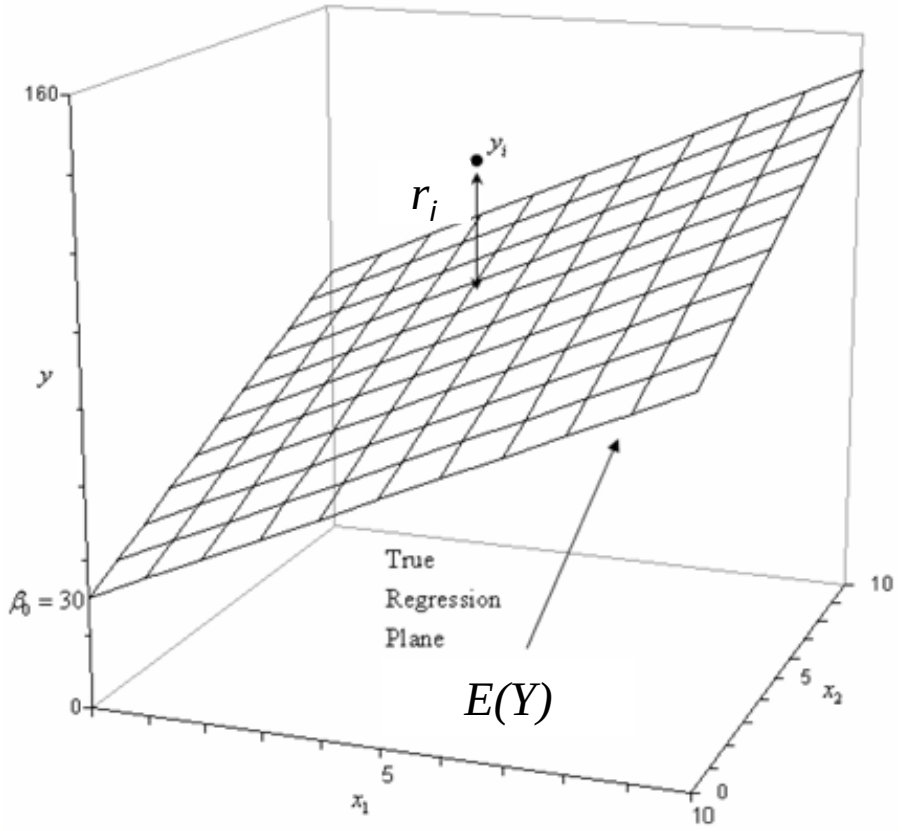
As before, these do NOT depend on *i*

ε_i = the residual corresponding to the i^{th} subject

and $\varepsilon_i \sim \text{NIID}(0, \sigma^2)$

Multiple Regression with Two Predictors

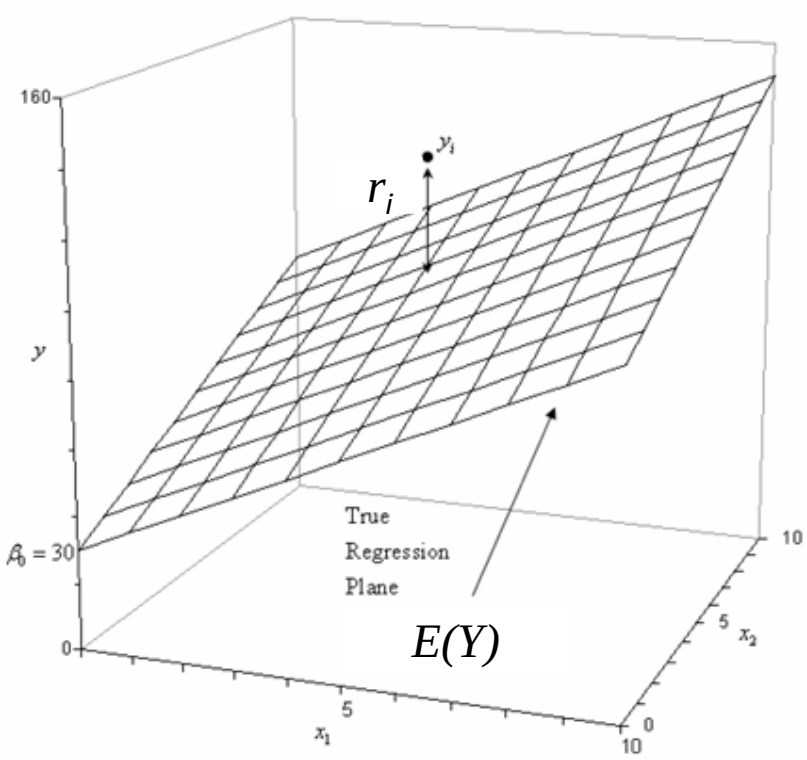
$$Y_i = \beta_o + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i \qquad i = 1, \ldots n$$



Instead of a regression LINE, we have a two dimensional PLANE. A point on the plane indicates the expected response $E(Y)$ for given values of X_1 AND X_2

Least Squares Estimation

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i \qquad i = 1, \ldots n$$



Least Squares estimation of β's will be based on minimizing the squares of the vertical distances between the observations and the plane, i.e., the sum of squared residuals (r_i's)

r_i represents the difference between an observation Y_i and its expected value $E(Y_i)$ on the plane.

Example

- In a long-term study on obesity, researchers are interested in the relationship between body fat and morphological measurements
- Body fat is expensive to measure accurately
- Could there be other, less expensive, measures that would reliably indicate body fat?
- Random sample of 20 healthy females, 2534 years of age

Triceps Skinfold Thickness (mm)	Midarm Circumference (cm)	Body Fat
19.5	29.1	11.9
24.7	28.2	22.8
30.7	37	18.7
29.8	31.1	20.1
19.1	30.9	12.9
25.6	23.7	21.7
31.4	27.6	27.1
27.9	30.6	25.4
22.1	23.2	21.3
25.5	24.8	19.3
31.1	30	25.4
30.4	28.3	27.2
18.7	23	11.7
19.7	28.6	17.8
14.6	21.3	12.8
29.5	30.1	23.9
27.7	25.7	22.6
30.2	24.6	25.4
22.7	27.1	14.8
25.2	27.5	21.1

Example: Model Specification

$$Y_i = \beta_o + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i \qquad i = 1, \ldots n$$

Y_i is the observed body fat content on the i^{th} woman

X_{1i} is the observed triceps skinfold thickness (mm) on the i^{th} woman

X_{2i} is the observed midarm circumference (cm) on the i^{th} woman

β_0, β_1 and β_2 are the (partial) regression coefficients

ε_i are the residuals or left-over noise corresponding to the i^{th} woman

$$\varepsilon_i \sim \text{NIID}\left(0, \sigma^2\right)$$

Components of the Model

- A "model specification" includes
  - The equation, including all subscripts on the variables
  - A short definition for each variable in the equation, including the subscripts
  - A list of all the parameters in the equation (these are the values we will be estimating)
  - The assumptions of the model

Least Squares Estimation

For two predictors : $Y_i = \beta_o + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i \qquad i = 1, \ldots n$

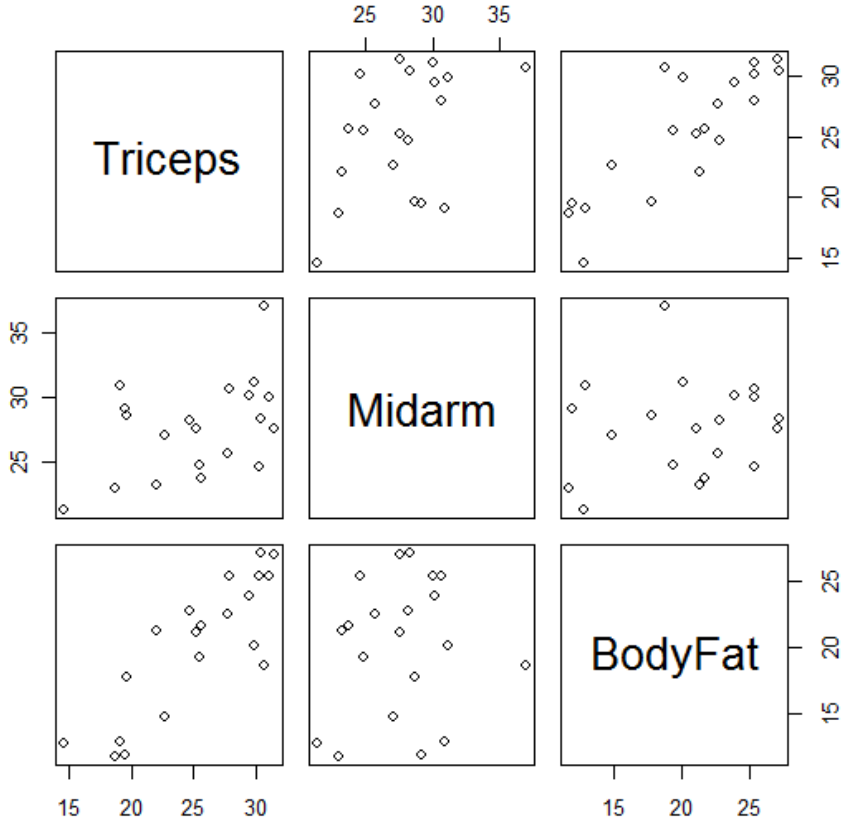
For p predictors : $Y_i = \beta_o + \sum_{k=1}^{p} \beta_k X_{ik} + \varepsilon_i \qquad i = 1, \ldots n$

- Least Squares Criterion:

  minimize Q = $\sum_{i=1}^{n} r_i^2 = \sum_{i=1}^{n} \left\{ Y_i - \left(\beta_0 + \sum_{k=1}^{p} \beta_k X_{ik} \right) \right\}^2$

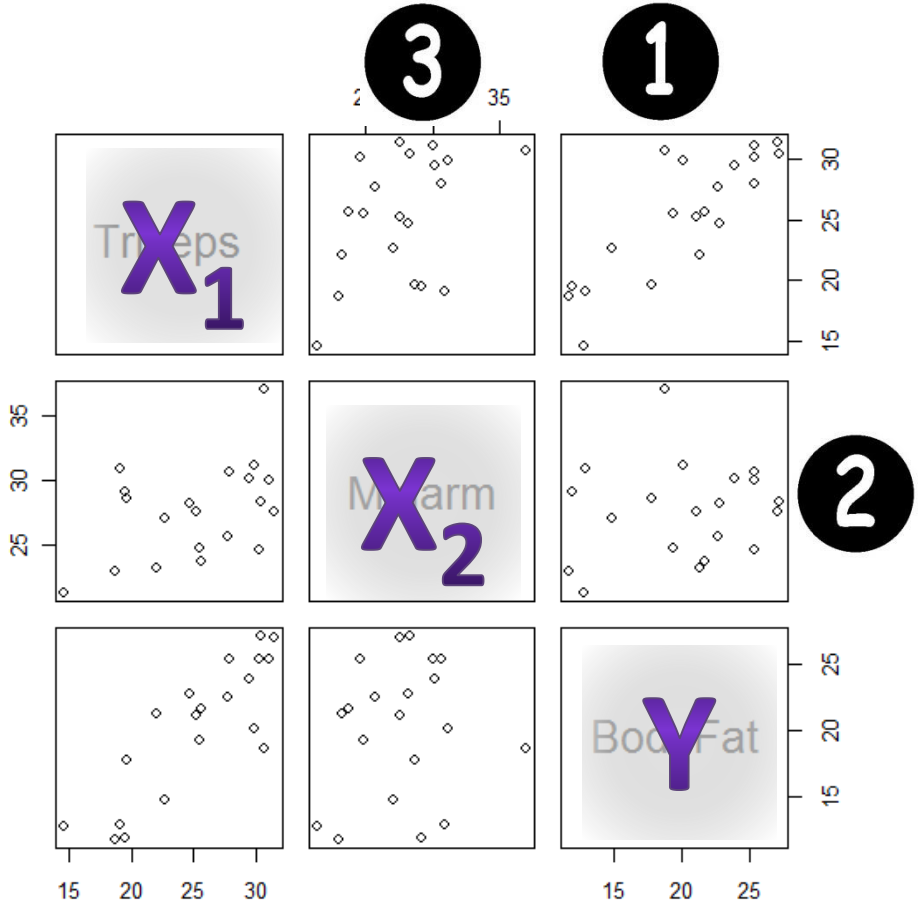
- Find values of all the β's that minimize Q
  - Get the partial derivatives of Q with respect to each β
  - Set each derivative equal to zero and solve the simultaneous equations for the β's
- Does this sound familiar?

Scatterplot Matrix



- Begin the analysis with a scatterplot matrix
- Each variable is plotted against the other variables
- Graphs in upper right are mirror images of those in lower left
- Often, these plots reveal patterns (e.g. curves) that are not readily apparent by looking only at correlation
- More predictors ⇨ more graphs

Scatterplot Matrix, Continued



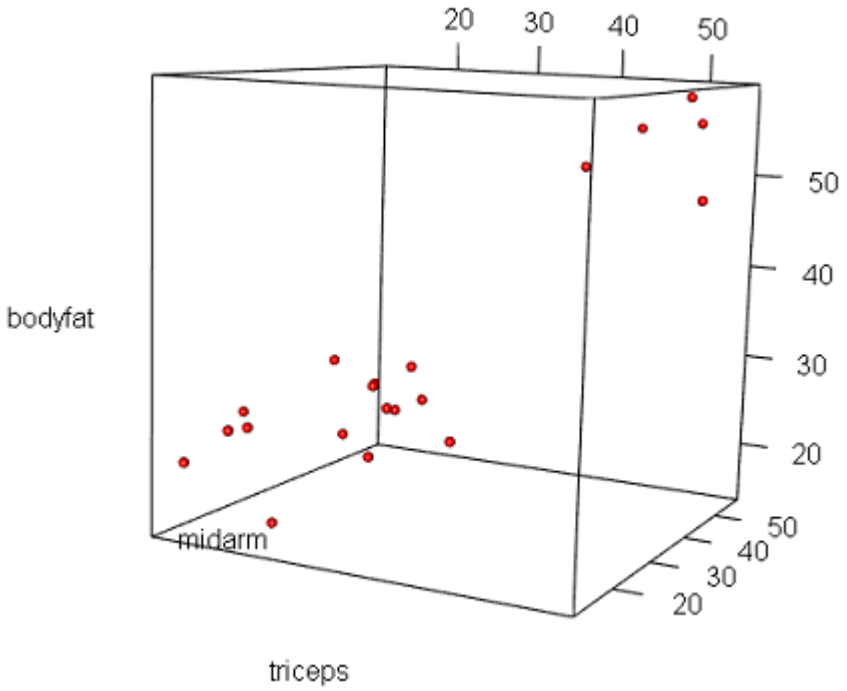
Replace each variable with its generic designation, X_1, X_2 or Y

We see

1. A fairly strong linear association between X_1 and Y
2. A much weaker (or no) linear association between X_2 and Y
3. A moderate linear association between X_1 and X_2

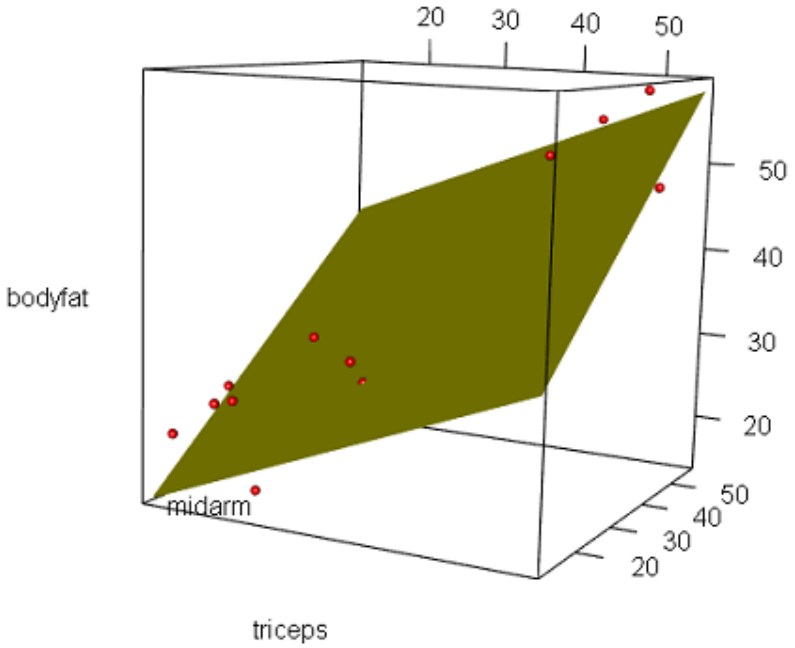
3D Scatterplot: Body Fat Data

When there are exactly two predictors, we can plot all the data in three dimensions.

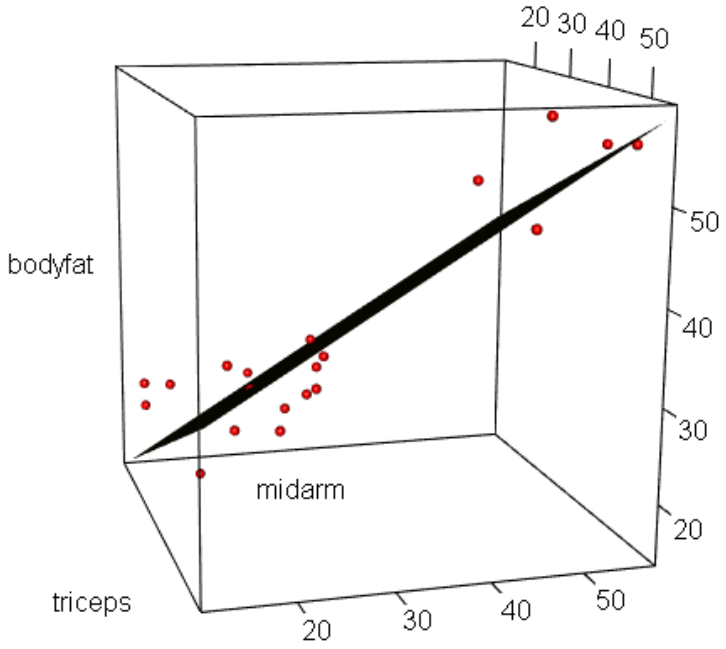


- Vertical axis is the response (body fat)
- Two horizontal axes are the two predictors -- triceps and midarm
- The Least Squares procedure fits a surface (a plane) through these points

Regression Surface

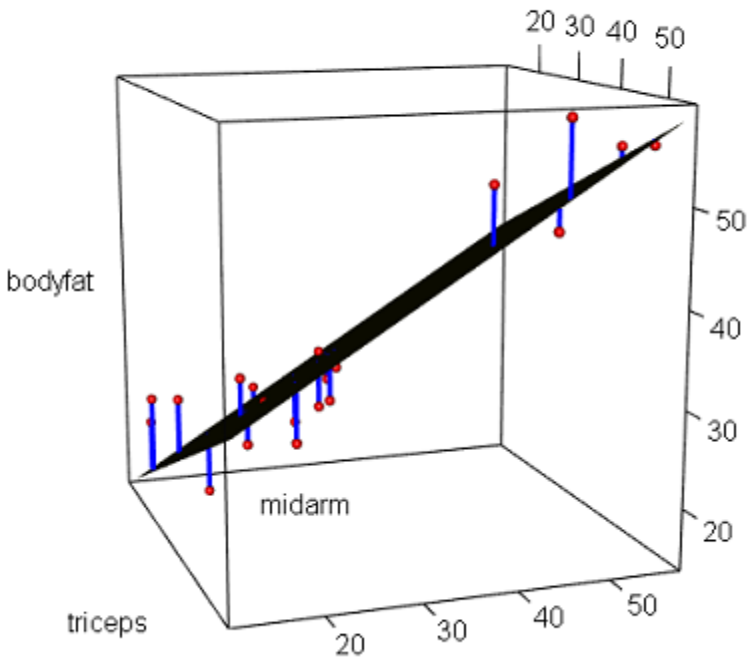


Scatterplot with estimated Least Squares plane



Rotate the graph to see points both above and below the plane

Residuals in 3 Dimensions



- Residuals are the vertical distance from the point to the plane
- The location of the Least Squares plane is guaranteed to minimize the sum of the squared residuals

Least Squares Estimators

- The model $Y_i = \beta_o + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i \quad i = 1, \ldots n$

  has three population parameters: β_0, β_1, and β_2
- The estimated surface $\hat{Y}_i = \hat{\beta}_o + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i}, \quad i = 1, \ldots n$

  has three estimators: $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\beta}_2$
- The estimators
  - are point estimates for the population parameters
  - are calculated from the sample data
  - are random variables
  - have probability distributions

Properties of Least Squares Estimators

- Estimators are unbiased
  - Over repeated sampling, $E(\hat{\beta}_k) = \beta_k$ for each of the β's in the model
- Estimators are NOT independent
  - They are each calculated from the same sample data
  - The dependence among estimators is recorded in a variance-covariance matrix
- For an estimated model with two predictors (plus an intercept), the variance-covariance matrix has 3 rows and 3 columns

Variance-Covariance Matrix

- Is symmetric (upper right entries are mirrored in lower left)
- Diagonal entries are the variances of the sampling distributions for the three estimators
- Off-diagonal entries are the covariances

$$\begin{bmatrix} \text{var}\left(\hat{\beta}_o\right) & \text{cov}\left(\hat{\beta}_o, \hat{\beta}_1\right) & \text{cov}\left(\hat{\beta}_o, \hat{\beta}_2\right) \\ \text{cov}\left(\hat{\beta}_o, \hat{\beta}_1\right) & \text{var}\left(\hat{\beta}_1\right) & \text{cov}\left(\hat{\beta}_1, \hat{\beta}_2\right) \\ \text{cov}\left(\hat{\beta}_o, \hat{\beta}_2\right) & \text{cov}\left(\hat{\beta}_1, \hat{\beta}_2\right) & \text{var}\left(\hat{\beta}_2\right) \end{bmatrix}$$

Importance of Var-Cov Matrix

- Hypothesis tests and confidence/prediction intervals require
  - point estimates (the beta hats) of the parameters (the beta's), and
  - the standard errors of these point estimates
- The standard errors incorporate both the variances and covariances of the estimates
- Ignoring the covariances leads to incorrect standard errors, and therefore faulty inference

What You Should Know

- Be able to write a complete model specification
- Interpret correlation and a scatterplot matrix to decide which predictors are likely to be important in modeling the response
- Explain the Least Squares criterion in three dimensions

- In the next lesson, we will use SAS to fit a model to the body fat data and interpret the output