Chapter 1: Introduction

Statistical Approach to a Scientific Problem

- Ask a question
- Collect data
- Initial, exploratory data analysis
- Answer the question (Inferential statistics)

Ask a question

- Describe something (What is happening?)
- Make predictions (What will happen?)
- Causal inference (What will happen if I ...?)
- Others

Collect data

- Is the data relevant?
- Is there measurement error?
- Is there missing data?
- Is the data a sample?
  - What is the population?
  - Random sample?
- Is the data from an experiment?
  - What was the treatment?
  - How was the treatment allocated? (random?)

Exploratory data analysis

- Organize data
- Display data graphically
- Summarize data
- Be alert for the unexpected

Inferential Statistics

- Estimate parameters
- Make predictions
- Test hypotheses
- What did we learn?
- What is still uncertain / what may have gone wrong?

Regression Analysis

Build a model to "explain" the relationship between a single variable Y and other variables $X_1, \cdots, X_p$

- Y: response variable, output, dependent variable
- X: predictor variable, input, independent variable
  - $p = 1$: simple regression
  - $p > 1$: multiple regression

Goals of Regression Analysis

- Describe data
- Make predictions
- Causal inference

Types of Variables

- Qualitative, categorical: can't say one is bigger than another
- Quantitative, numerical
  - Discrete counts
  - Continuous measures
- In between (ordinal)

What We Will Cover

- X: continuous, discrete or categorical
- Y is a continuous variable
- Y is a binary variable
- Y is a discrete count

Emphases of the Course

- Practice using linear regression models
- Learn what methods are available, and their limitations
- Many examples, less mathematical theory
- More intuition, less derivation of formulas
- Will still learn mathematical foundations behind practical tools

Quick Introduction to R

Pima Data Example: Exploratory Data Analysis

```
## Load the library
> library(faraway)
## Read in the data
> data(pima)
> pima
    pregnant glucose diastolic triceps insulin  bmi ...
1          6     148        72      35       0 33.6 ...
2          1      85        66      29       0 26.6 ...
3          8     183        64       0       0 23.3 ...
...
767        1     126        60       0       0 30.1 ...
768        1      93        70      31       0 30.4 ...
```

```
> help(pima)
 The dataset contains the following variables
     'pregnant' Number of times pregnant
     'glucose' Plasma glucose concentration at 2 hours
              in an oral glucose tolerance test
     'diastolic' Diastolic blood pressure (mm Hg)
     'triceps' Triceps skin fold thickness (mm)
     'insulin' 2-Hour serum insulin (mu U/ml)
     'bmi' Body mass index (weight in kg/(height in m)^2)
     'diabetes' Diabetes pedigree function
     'age' Age (years)
     'test' test whether the patient shows signs of
          diabetes (coded 0 if negative, 1 if positive)
```

```
## Dimension of the data
> dim(pima)
[1] 768   9

## Numerical Summaries
> summary(pima)
    pregnant        glucose        diastolic        triceps
 Min.   : 0.0   Min.   :  0   Min.   :  0   Min.   : 0
 1st Qu.: 1.0   1st Qu.: 99   1st Qu.: 62   1st Qu.: 0
 Median : 3.0   Median :117   Median : 72   Median :23
 Mean   : 3.9   Mean   :121   Mean   : 69   Mean   :21
 3rd Qu.: 6.0   3rd Qu.:140   3rd Qu.: 80   3rd Qu.:32
 Max.   :17.0   Max.   :199   Max.   :122   Max.   :99
```

```
    insulin          bmi           diabetes          age
 Min.   :  0   Min.   : 0.0   Min.   :0.08   Min.   :21
 1st Qu.:  0   1st Qu.:27.3   1st Qu.:0.24   1st Qu.:24
 Median : 31   Median :32.0   Median :0.37   Median :29
 Mean   : 80   Mean   :32.0   Mean   :0.47   Mean   :33
 3rd Qu.:127   3rd Qu.:36.6   3rd Qu.:0.63   3rd Qu.:41
 Max.   :846   Max.   :67.1   Max.   :2.42   Max.   :81
      test
 Min.   :0.000
 1st Qu.:0.000
 Median :0.000
 Mean   :0.349
 3rd Qu.:1.000
 Max.   :1.000
```

```
## Missing Values
> sort(pima$diastolic)
  [1]   0   0   0   0   0   0   0   0   0   0   0   0
 [13]   0   0   0   0   0   0   0   0   0   0   0   0
 [25]   0   0   0   0   0   0   0   0   0   0   0  24
 [37]  30  30  38  40  44  44  44  44 ...
> pima$diastolic[pima$diastolic == 0] = NA
> pima$glucose[pima$glucose == 0] = NA
> pima$triceps[pima$triceps == 0] = NA
> pima$insulin[pima$insulin == 0] = NA
> pima$bmi[pima$bmi == 0] =NA
```

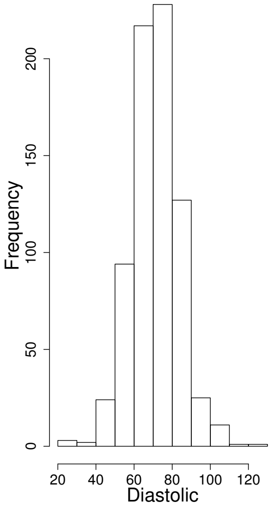
```
## Categorical Variable
> pima$test = factor(pima$test)
> summary(pima$test)
  0   1
500 268
> levels(pima$test) = c("negative", "positive")
> summary(pima$test)
negative positive
     500      268
```

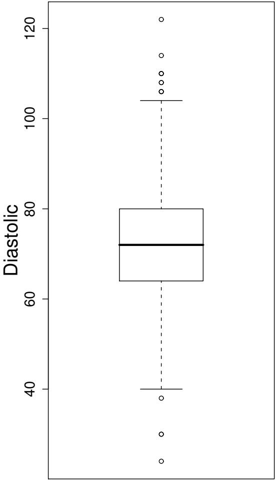
```
## New Summary
> summary(pima)
    pregnant        glucose        diastolic        triceps
 Min.   : 0.0   Min.   : 44   Min.   : 24   Min.   :  7
 1st Qu.: 1.0   1st Qu.: 99   1st Qu.: 64   1st Qu.: 22
 Median : 3.0   Median :117   Median : 72   Median : 29
 Mean   : 3.8   Mean   :122   Mean   : 72   Mean   : 29
 3rd Qu.: 6.0   3rd Qu.:141   3rd Qu.: 80   3rd Qu.: 36
 Max.   :17.0   Max.   :199   Max.   :122   Max.   : 99
                NA's   :  5   NA's   : 35   NA's   :227
```

```
   insulin          bmi             diabetes             age
 Min.   : 14   Min.   :18.2   Min.   :0.08   Min.   :21
 1st Qu.: 76   1st Qu.:27.5   1st Qu.:0.24   1st Qu.:24
 Median :125   Median :32.3   Median :0.37   Median :29
 Mean   :156   Mean   :32.5   Mean   :0.47   Mean   :33
 3rd Qu.:190   3rd Qu.:36.6   3rd Qu.:0.63   3rd Qu.:41
 Max.   :846   Max.   :67.1   Max.   :2.42   Max.   :81
 NA's   :374   NA's   :11.0
       test
 negative:500
 positive:268
```

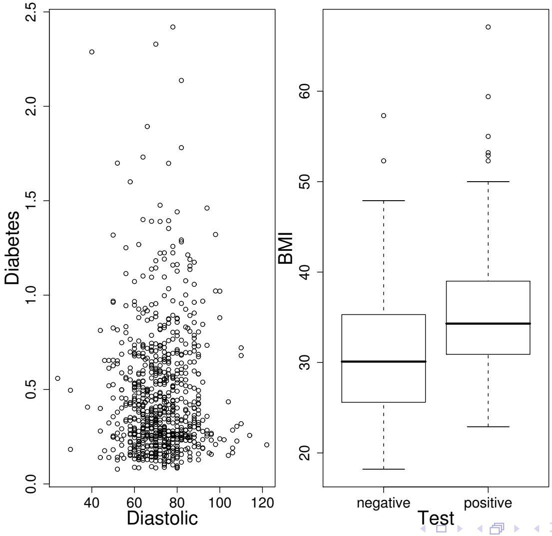
```
## Individual summary functions
> mean(pima$diastolic, na.rm=T)
[1] 72.40518
> median(pima$diastolic, na.rm=T)
[1] 72
> range(pima$diastolic, na.rm=T)
[1]  24 122
> quantile(pima$diastolic, na.rm=T)
  0%  25%  50%  75% 100%
  24   64   72   80  122
## Other functions:  var(), sd()
```

```
## Graphical Summaries:  single variable
> hist(pima$diastolic)
> boxplot(pima$diastolic)
```





```
## Graphical Summaries:  two variables
> plot(pima$diastolic, pima$diabetes)
> plot(pima$test, pima$bmi)
```



```
## Selecting Subsets of the Data
## The second row
> pima[2,]
  pregnant glucose diastolic triceps insulin
2        1      85        66      29      NA
       bmi diabetes age     test
2     26.6    0.351  31 negative
## The third column
> pima[,3]
  [1]  72  66  64  66  40  74  50  NA  70 ...
## The (2,3) element
> pima[2,3]
[1] 66
```

```
## The first, second and fourth row
> pima[c(1,2,4), ]
  pregnant glucose diastolic triceps insulin ...
1        6     148        72      35      NA ...
2        1      85        66      29      NA ...
4        1      89        66      23      94 ...
## The third through sixth rows
> pima[3:6, ]
  pregnant glucose diastolic triceps insulin ...
3        8     183        64      NA      NA ...
4        1      89        66      23      94 ...
5        0     137        40      35     168 ...
6        5     116        74      NA      NA ...
```

```
## "Everything but"
> pima[, -c(1,2)]
    diastolic triceps insulin  bmi diabetes age     test
1          72      35      NA 33.6    0.627  50 positive
2          66      29      NA 26.6    0.351  31 negative
3          64      NA      NA 23.3    0.672  32 positive
...

## Cases which have pregnant greater than 14
> pima[pima$pregnant > 14, ]
    pregnant glucose diastolic triceps insulin ...
89        15     136        70      32     110 ...
160       17     163        72      41     114 ...
```

```
## Help
> help(boxplot)
> ?boxplot
> help('*')
> help.start()
```

Chapter 2: Estimation

Regression Analysis

- y: response, output
- $x = (x_1, x_2, \ldots, x_p)$: predictors, input
- Goal: model the relationship between y and $x_1, \ldots, x_p$

- General form: $y = f(x) + \epsilon$
- $f(\cdot)$: underlying truth. Unknown
- Usually we are given a set of data

$$(x_{11}, \ldots, x_{1p}, y_1), \cdots, (x_{n1}, \ldots, x_{np}, y_n)$$

Galapagos Example

- Interested in how the number of species of tortoise on a Galapagos Island relates to other features of the island
- y: number of species of tortoise
- $x_1, \ldots, x_5$: area of the island, highest elevation of the island, distance from the nearest island, distance from Santa Cruz Island, area of the adjacent island

Galapagos Example

```
## Load the data
> library(faraway)
> data(gala)
## Check out the data
> gala
          Species Endemics    Area Elevation Nearest ...
Baltra         58       23   25.09       346     0.6 ...
Bartolome      31       21    1.24       109     0.6 ...
Caldwell        3        3    0.21       114     2.8 ...
Champion       25        9    0.10        46     1.9 ...
Coamano         2        1    0.05        77     1.9 ...
...
```

Linear Regression Analysis

- There is no way to estimate $f(\cdot)$ directly given a finite number of samples.
- We put some restrictions/structure on $f(\cdot)$.
- Assume
$$f(x) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$
where β_j's are unknown parameters and β_0 is the intercept.
- Estimation of $f(\cdot)$ is reduced to estimation of β_j's

What Does "Linear" Mean?

A linear model is linear in parameters, not linear in predictors. Formally, a function g is linear in β if

$$g(a \cdot \beta + a^* \cdot \beta^*) = a \cdot g(\beta) + a^* \cdot g(\beta^*)$$

where $a, a^* \in \mathbb{R}, \beta, \beta^* \in \mathbb{R}^p$.

Example: $f(x) = \beta_0 + \beta_1 e^{x_1} + \beta_2 \ln(x_2) + \beta_3 x_1 x_3$ is a linear model, but $f(x) = \beta_0 + \beta_1 x_1^{\beta_2}$ is not.

Transformation

$f(x) = \beta_0 x_1^{\beta_1}$ is not a linear model. However, notice that

$$\ln f(x) = \ln \beta_0 + \beta_1 \ln x_1$$

Hence if we let $f^*(x) = \ln f(x), \beta_0^* = \ln \beta_0, \beta_1^* = \beta_1$, we have

$$f^*(x) = \beta_0^* + \beta_1^* \ln x_1$$

which is a linear model.

Implications

- Linear models are less restrictive than you might think
- They can be made very flexible by transformation of the response and the predictors.
- Linear models are not necessarily straight lines (for example, $y = ax^2 + bx + c$).

Simple Linear Regression

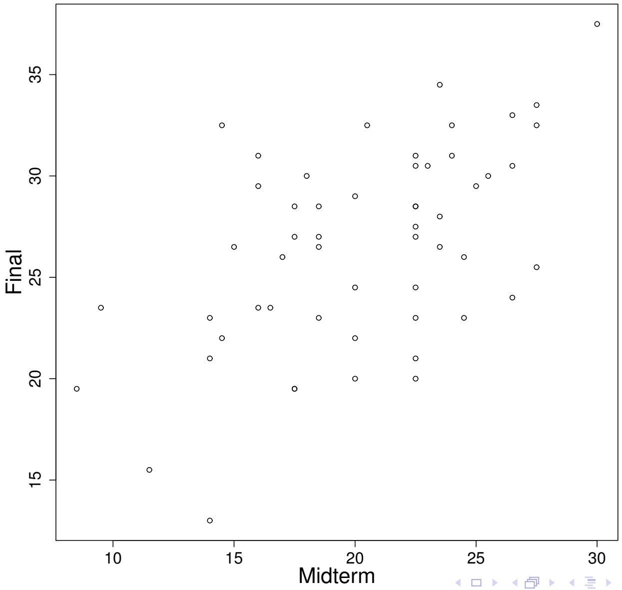
- $p = 1$, only one predictor variable
- The model is:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, \ldots, n$$

Example

- Scores from previous Stats 500
- y: final score
- x: midterm score
- $y = \beta_0 + \beta_1 x + \epsilon$

Stats 500 Data

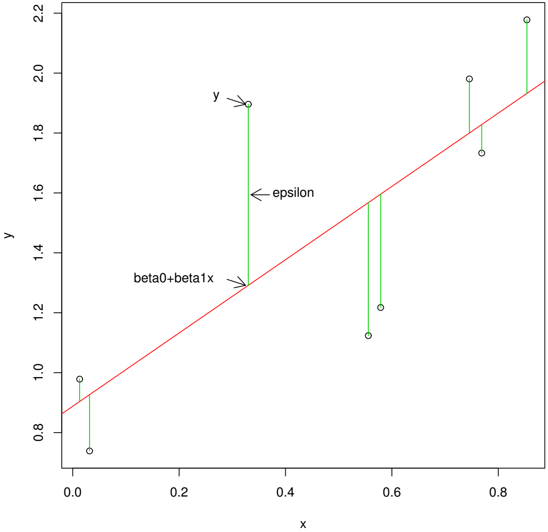


Simple Linear Regression Ctd

- Goal: given $(y_i, x_i),\ i = 1, \ldots, n$, estimate β_0, β_1
- ϵ_i is the error term; can always assume $E\epsilon = 0$.
- Minimize errors – how do we define that?
- One criterion is **least squares**:

$$\min_{\beta_0, \beta_1} \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

Least Squares Estimate



Estimating β_0, β_1

Differentiate the criterion with respect to β_0, β_1 and set the derivatives equal to 0, we get:

$$\begin{aligned}
\frac{\partial}{\partial \beta_0} &= (-2)\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i) = 0 \\
\frac{\partial}{\partial \beta_1} &= (-2)\sum_{i=1}^{n} x_i(y_i - \beta_0 - \beta_1 x_i) = 0
\end{aligned}$$

Solving for β_0 and β_1, we have:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$$

"Hat" notation is used for estimates.

Another interpretation

Letting $r = Cor(x, y)$, $s_y = SD(y)$, $s_x = SD(x)$, can rewrite the line equation (simple algebra) as

$$\frac{y - \bar{y}}{s_y} = r\frac{x - \bar{x}}{s_x},$$

or, if x and y are standardized first (mean 0, sd 1), simply

$$y = rx.$$

Two regression lines

- Suppose x and y have both been standardized.
- Regress y on x: $y = rx$
- Regress x on y: $x = ry$

Regression effect: predictions always "regress" towards the mean

- Regression effect is usually uninteresting
- Example: husband's and wife's education

Multiple Linear Regression

Model: $y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_i$
$\#$ predictors $=$
$\#$ parameters $=$
Assume $E(\epsilon_i) = 0, \quad i = 1, \ldots, n$

Matrix Notation

Let

$$y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \quad X = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ \vdots & \vdots & x_{ij} & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix}$$

$$\beta = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_p \end{pmatrix}, \quad \epsilon = \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

Then we can write the model for the data as:

$$y_{n\times 1} = X_{n\times(p+1)}\beta_{(p+1)\times 1} + \epsilon_{n\times 1}$$

This is the same model in more compact notation.

Estimating β

- Observe y and X (n samples)
- Want to minimize errors
- Least squares criterion:

$$
\begin{aligned}
\min_{\beta} \sum_{i=1}^{n} \epsilon_i^2 &= \epsilon^T \epsilon \\
&= (y - X\beta)^T (y - X\beta) \\
&= y^T y - 2y^T X\beta + \beta^T X^T X\beta
\end{aligned}
$$

Differentiating the criterion with respect to β and setting the derivative equal to 0:

- The **normal equation**:

$$X^T X\hat{\beta} = X^T y$$

- Solve for β:

$$\hat{\beta} = \left(X^T X\right)^{-1} X^T y$$

- X full rank $\Leftrightarrow$ $X^T X$ invertible

Fitted Model

- Fitted values: $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_p x_{ip}$
- Fitted model: $\hat{f}(x) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_p x_p$
- Residuals: $\hat{\epsilon}_i = y_i - \hat{y}_i$
- Residual sum of squares (RSS): $\sum_{i=1}^{n} \hat{\epsilon}_i^2$

Hat Matrix

- $\hat{y} = X\hat{\beta} = X\left(X^T X\right)^{-1} X^T y = Hy$, where

$$H = X\left(X^T X\right)^{-1} X^T$$

  is called the "Hat" matrix.
- Fitted values: $\hat{y} = Hy$
- Residuals: $\hat{\epsilon} = y - \hat{y} = (I - H)y$
- H is a projection matrix.

Projection Matrix

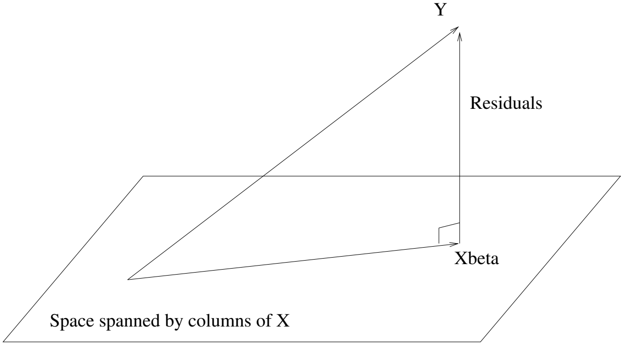
Definition: H is a projection matrix if

- $H^T = H$ (H is symmetric).
- $HH = H$ (H is idempotent).

Does $X\left(X^TX\right)^{-1}X^T$ satisfy these two conditions?

The projection matrix H projects $y_{n\times 1}$ onto the column space of $X_{n\times(p+1)}$, which leads to the vector space interpretation of least squares estimate.

Vector Space Interpretation



$\min_\beta (y - X\beta)^T(y - X\beta)$ minimizes the Euclidean distance between y and the linear space spanned by the columns of X.

Properties of $\hat{\beta}$

- Unbiased: $E(\hat{\beta}) = \beta$. Check:

- $\mathrm{Var}(\hat{\beta}) =$? Assume $\mathrm{Var}(\epsilon) = \sigma^2 I$, then

$$\mathrm{Var}(\hat{\beta}) = \sigma^2 (X^T X)^{-1}$$

Estimating variance

- σ^2 can also be estimated:

$$\hat{\sigma}^2 = \frac{\sum_i (y_i - \hat{y}_i)^2}{n - (p+1)},$$

  where $n - (p+1)$ is the degrees of freedom.
- Unbiased: $E(\hat{\sigma}^2) = \sigma^2$

Galapagos Example

```
## Get the X matrix
> dim(gala)
[1] 30  7
> n = dim(gala)[1]
> p = dim(gala)[2] - 2
> x = cbind(1, as.matrix(gala[, 3:7]))
> ## Compute the inverse of (X^T X)
> xtx = t(x) %*% x
> xtxi = solve(xtx)
> beta = xtxi %*% t(x) %*% gala[,1]
```

```
> beta
                 [,1]
           7.068220709
Area      -0.023938338
Elevation  0.319464761
Nearest    0.009143961
Scruz     -0.240524230
Adjacent  -0.074804832
> ## Residual sum of squares
> rss = sum((gala[,1] - x %*% beta)^2)
> sigma2 = rss / (n - (p+1))
> sigma = sqrt(sigma2)
> sigma
[1] 60.97519
```

```
> ## Use the lm() function
> temp = lm(Species ~ Area + Elevation + Nearest
            + Scruz + Adjacent, data=gala)
> summary(temp)
```

```
Call:
lm(formula = Species ~ Area + Elevation + Nearest +
Scruz +  Adjacent,  data = gala)
Residuals:
     Min        1Q   Median        3Q       Max
-111.679  -34.898   -7.862   33.460  182.584
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.068221  19.154198   0.369 0.715351
Area        -0.023938   0.022422  -1.068 0.296318
Elevation    0.319465   0.053663   5.953 3.82e-06 ***
Nearest      0.009144   1.054136   0.009 0.993151
Scruz       -0.240524   0.215402  -1.117 0.275208
Adjacent    -0.074805   0.017700  -4.226 0.000297 ***
---
Signif. codes:   0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1

Residual standard error: 60.98 on 24 degrees of freedom
Multiple R-Squared: 0.7658,     Adjusted R-squared: 0.7171
F-statistic:  15.7 on 5 and 24 DF,  p-value: 6.838e-07
```

Goodness of Fit

- Need a measure of how well the model fits with the data
- Residual sum of squares (RSS): $\sum_i (y_i - \hat{y}_i)^2$
  Seems reasonable, but what about units?

Coefficient of determination (R^2)

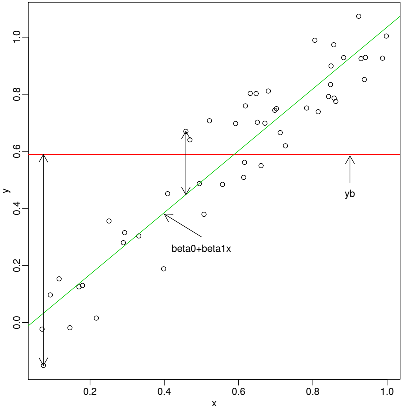
$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$$

- Alternative expression:

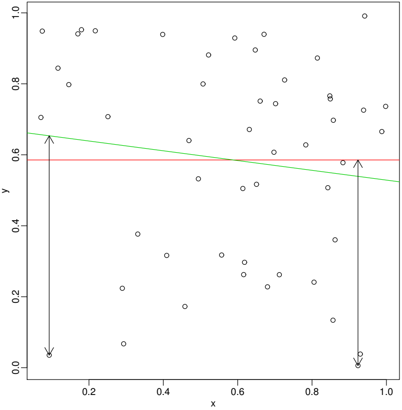
$$R^2 = \frac{\sum_i (\hat{y}_i - \bar{y})^2}{\sum_i (y_i - \bar{y})^2}$$

- $0 \leq R^2 \leq 1$. (Why?)
- R^2 "close" to 1 indicates good fit.

Intuition



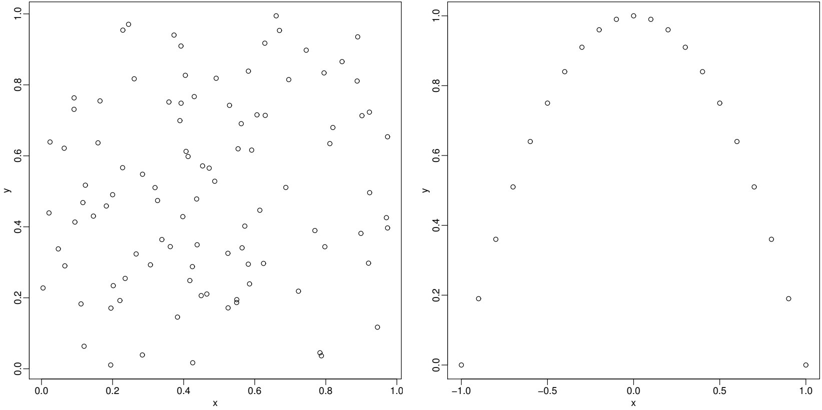
$R^2 = 0.89$



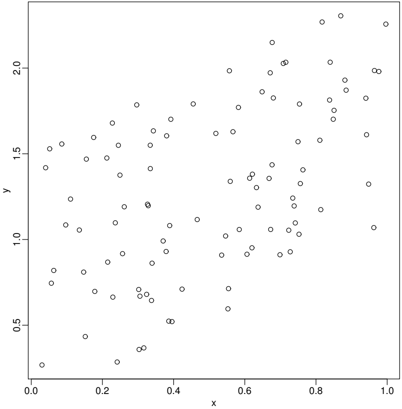
$R^2 = 0$

Remarks on R^2

- R^2 near 0 could be

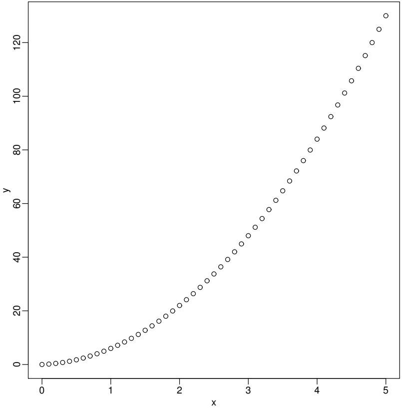


- Small R^2 does not mean that y and X are not linearly related (can have slight trend with high variance).



$R^2 = 0.2.$

- R^2 close to 1 does not mean the linear model is correct.



$R^2 = 0.9.$

The Gauss-Markov Theorem

- Why use the least squares estimate $\hat{\beta}$?
- Theorem: Suppose $y = X\beta + \epsilon$, X is full rank, $E(\epsilon) = 0$ and $\mathrm{Var}(\epsilon) = \sigma^2 I$. Consider $\psi = c^T\beta$. Then among all unbiased linear estimates of ψ, $\hat{\psi} = c^T\hat{\beta}$ has the minimum variance and is unique.
- Example: Let $c^T = (1, x_1, \ldots, x_p)$, then $\psi = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$.
- Best Linear Unbiased Estimate (BLUE)

What Can Go Wrong?

- $X^T X$ could be singular (happens if predictors are linearly dependent or if $p > n$)
- Assumed $\mathrm{Var}(\epsilon) = \sigma^2 I$
- Best only among linear, unbiased estimates