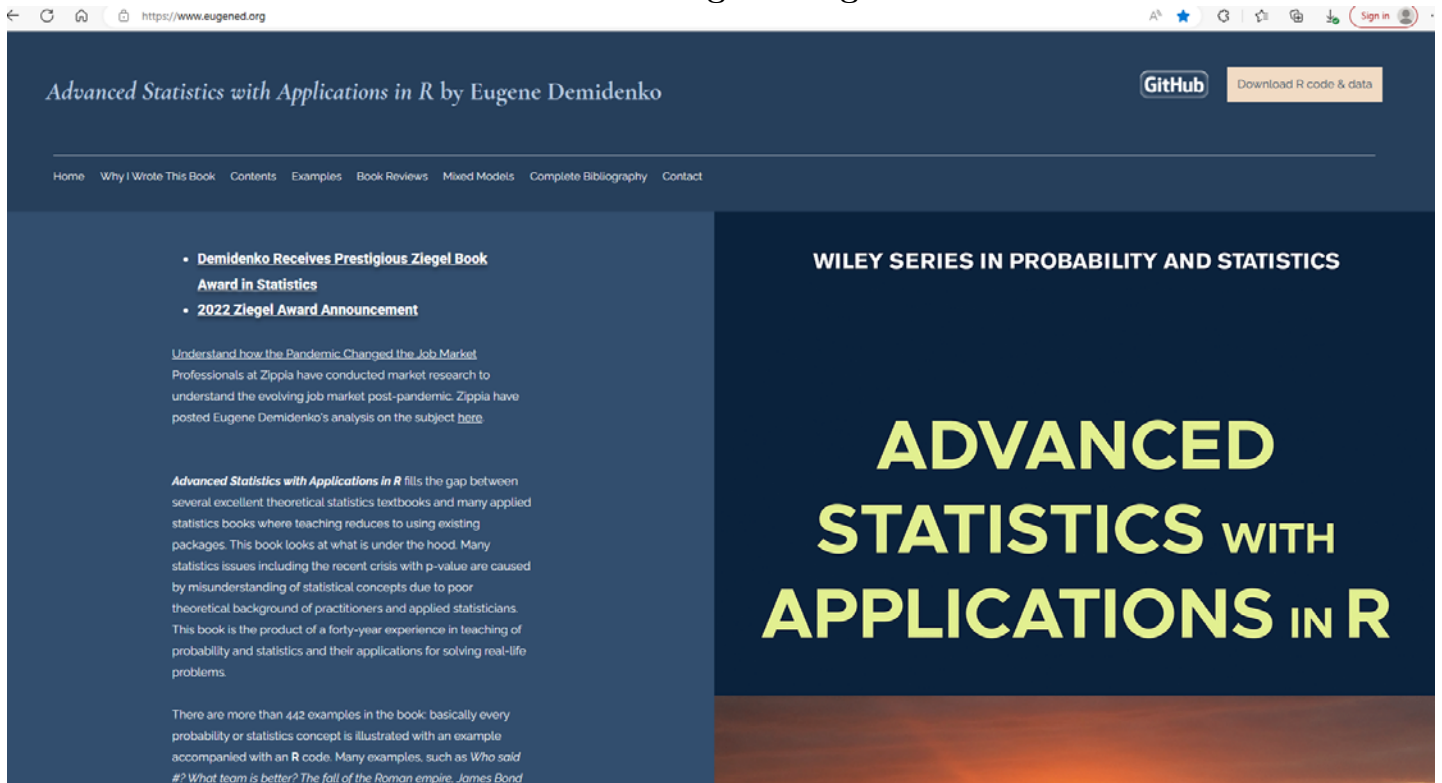


Probability, Statistics, Computer Science, and Data Science

Read the preface: "Why I wrote this book"

www.eugened.org

The image is a screenshot of a web browser displaying the homepage of the book "Advanced Statistics with Applications in R" by Eugene Demidenko. The browser's address bar shows the URL "https://www.eugened.org". The website has a dark blue header with the book title and author's name on the left, and a "GitHub" logo and a "Download R code & data" button on the right. Below the header is a navigation menu with links: Home, Why I Wrote This Book, Contents, Examples, Book Reviews, Mixed Models, Complete Bibliography, and Contact. The main content area is split into two columns. The left column contains a list of bullet points: "Demidenko Receives Prestigious Ziegel Book Award in Statistics" and "2022 Ziegel Award Announcement". Below these is a link "Understand how the Pandemic Changed the Job Market" followed by a paragraph about Zippla's market research. Further down is a paragraph about the book's purpose, stating it fills the gap between theoretical and applied statistics. The right column features the text "WILEY SERIES IN PROBABILITY AND STATISTICS" at the top, followed by the book title "ADVANCED STATISTICS WITH APPLICATIONS IN R" in large, bold, yellow-green letters. At the bottom of the page, there is a paragraph stating that there are more than 442 examples in the book, each with an R code example, such as "Who said #? What team is better? The fall of the Roman empire. James Bond".

Advanced Statistics with Applications in R by Eugene Demidenko

GitHub Download R code & data

Home Why I Wrote This Book Contents Examples Book Reviews Mixed Models Complete Bibliography Contact

- Demidenko Receives Prestigious Ziegel Book Award in Statistics
- 2022 Ziegel Award Announcement

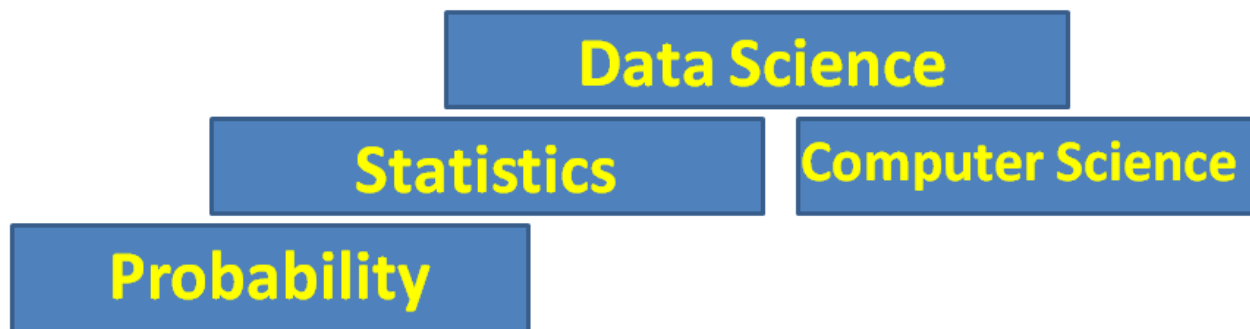
[Understand how the Pandemic Changed the Job Market](#)
Professionals at Zippla have conducted market research to understand the evolving job market post-pandemic. Zippla have posted Eugene Demidenko's analysis on the subject [here](#).

Advanced Statistics with Applications in R fills the gap between several excellent theoretical statistics textbooks and many applied statistics books where teaching reduces to using existing packages. This book looks at what is under the hood. Many statistics issues including the recent crisis with p-value are caused by misunderstanding of statistical concepts due to poor theoretical background of practitioners and applied statisticians. This book is the product of a forty-year experience in teaching of probability and statistics and their applications for solving real-life problems.

There are more than 442 examples in the book: basically every probability or statistics concept is illustrated with an example accompanied with an R code. Many examples, such as *Who said #? What team is better? The fall of the Roman empire. James Bond*

WILEY SERIES IN PROBABILITY AND STATISTICS

ADVANCED STATISTICS WITH APPLICATIONS IN R



Where is MATH 70?

- **Probability theory** determines the chance of a complex event given the probability of elementary events/distribution. Example: X is a Poisson distributed random variable with a known rate parameter. What is the chance that $X > 5$? The hallmark results: the law of large numbers and central limit theorem (gives birth to Gaussian distribution).
- **Statistics** is the inverse probability: how to reconstruct parameter distribution given the data/sample, or more generally, how to infer properties of the general population given as a sample? Statistics is concerned with precision, standard error, estimation of parameters, and statistical hypothesis testing. Example: given sample X_1, \dots, X_n from a Poisson distribution, we reconstruct/estimate the rate as $\hat{\lambda} = \bar{X}$. The hallmark results: unbiased estimators with minimum variance, maximum likelihood estimation yields the most precise (minimum variance) estimators when $n \rightarrow \infty$. Statistics determines optimal procedures/algorithms by minimizing the variance as a criterion.
- **Computer science**, or more specifically **machine learning**, concerns with developing **algorithms** for prediction within the sample treated as the general population (sample=general population). The concept of sample vs general population is tackled with computer experiments by splitting the data into training vs testing subsamples. Computer science solved problems cannot be solved by traditional statistical means such as penalized regression $\min_{\beta, \lambda} [\|\mathbf{y} - \mathbf{X}\beta\|^2 + \lambda \|\beta\|^2]$, or the bandwidth kernel density estimation. To date, the concept of the uncertainty estimation (standard error of prediction) is absent. The hallmark results: cross-validation, regression trees, CNN (convolutional neural network) and deep learning. There is no optimal machine learning algorithm – it's very much depends on the concrete application. Computer science versus Statistics = Engineering versus Physics.
- **Data science** is a combination of statistics and machine learning. Make decision by the means of the data analysis. The old-fashion “expert/boss” decision making is replaced with data-driven and evidence based. Data scientist use the data to answer vital questions such as: what is the optimal stock portfolio, what is the probability that the unemployment rate in the country next year will be greater than 10%, should a new pharmaceutical drug be approved by FDA (Food and Drug Administration), etc. ?

Week 1. Matrix algebra and calculus

Section 10.2, slope model of Section 6.7

Matrix inverse, trace, idempotent matrix

Properties:

1. $(\mathbf{AB})' = \mathbf{B}'\mathbf{A}'$
2. $(\alpha\mathbf{A} + \beta\mathbf{B})' = \alpha\mathbf{A}' + \beta\mathbf{B}'$
3. $(\mathbf{A}^{-1})' = (\mathbf{A}')^{-1}$
4. $(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$

\mathbf{A}' is the transposition of \mathbf{A} , sometimes \mathbf{A}^T is used.

R uses \mathbf{t} for transposition, e.g. $\mathbf{t}(\mathbf{A})$

Definition 1 A square matrix is symmetric if $\mathbf{A}' = \mathbf{A}$.

Definition 2 Trace of a square matrix $tr(\mathbf{A}) = \sum_{i=1}^n A_{ii}$.

Properties:

1. $tr(\mathbf{A} + \mathbf{B}) = tr(\mathbf{A}) + tr(\mathbf{B})$.
2. $tr(a\mathbf{A}) = a \times tr(\mathbf{A})$.
3. $tr(\mathbf{A}') = tr(\mathbf{A})$
4. $tr(\mathbf{AB}) = tr(\mathbf{BA})$ if \mathbf{AB} and \mathbf{BA} exist.

Example 3 $tr(\mathbf{ab}') = \mathbf{a}'\mathbf{b} = \sum_{i=1}^n a_i b_i$, the scalar product (\mathbf{a} and \mathbf{b} must have the same dimension).

Example 4 $tr(\mathbf{aa}') = \|\mathbf{a}\|^2$

Definition 5 A symmetric matrix \mathbf{A} is called idempotent if $\mathbf{A}^2 = \mathbf{A}$.

$\mathbf{1}_n$ is the $n \times 1$ vector of ones.

$$\mathbf{1}_n' \mathbf{1}_n = n$$

$\mathbf{1}_n \mathbf{1}_n' = \{1, i, j = 1, \dots, n\}$ is a matrix with all entries equal to 1.

Example 6 (a) Prove that matrix

$$\mathbf{A} = \mathbf{I} - \frac{1}{n} \mathbf{1}\mathbf{1}'$$

is symmetric and idempotent. (b) Find

$$tr(\mathbf{I} - \frac{1}{n} \mathbf{1}\mathbf{1}').$$

Solution.

(a)

$$\begin{aligned}\mathbf{A}^2 &= \left(\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}'\right) \left(\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}'\right) \\ &= \mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}' - \frac{1}{n}\mathbf{1}\mathbf{1}' + \left(-\frac{1}{n}\mathbf{1}\mathbf{1}'\right) \left(-\frac{1}{n}\mathbf{1}\mathbf{1}'\right) \\ &= \mathbf{I} - \frac{2}{n}\mathbf{1}\mathbf{1}' + \frac{1}{n^2}\mathbf{1}\mathbf{1}'\mathbf{1}\mathbf{1}'\end{aligned}$$

But

$$\mathbf{1}\mathbf{1}'\mathbf{1}\mathbf{1}' = \mathbf{1}(\mathbf{1}'\mathbf{1})\mathbf{1}' = \mathbf{1}n\mathbf{1}' = n\mathbf{1}\mathbf{1}'$$

Finally,

$$\mathbf{A}^2 = \mathbf{I} - \frac{2}{n}\mathbf{1}\mathbf{1}' + \frac{1}{n}\mathbf{1}\mathbf{1}' = \mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}' = \mathbf{A}.$$

(b)

$$\begin{aligned}tr\left(\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}'\right) &= tr(\mathbf{I}) - \frac{1}{n}tr(\mathbf{1}\mathbf{1}') = n - \frac{1}{n}tr(\mathbf{1}'\mathbf{1}) \\ &= n - \frac{1}{n}n \\ &= n - 1.\end{aligned}$$

Implementation in R:

`A+B` computes $\mathbf{A} + \mathbf{B}$

`A%*%B` computes \mathbf{AB} and `A%%b` computes \mathbf{Ab} . Note that `A*B` computes an element-wise matrix product, similarly, `a*b`.

`solve(A)` finds \mathbf{A}^{-1}

`t(A)` finds \mathbf{A}'

There is no built-in trace function in R.

Do it yourself:

```
tr=function(M) sum(diag(M))
```

Example

```
A=matrix(rnorm(10^2),ncol=10,nrow=10)
```

```
tr(A)
```

Definition 7 *Euclidean norm and distance. Squared norm*

$$\|\mathbf{a}\|^2 = \mathbf{a}'\mathbf{a}$$

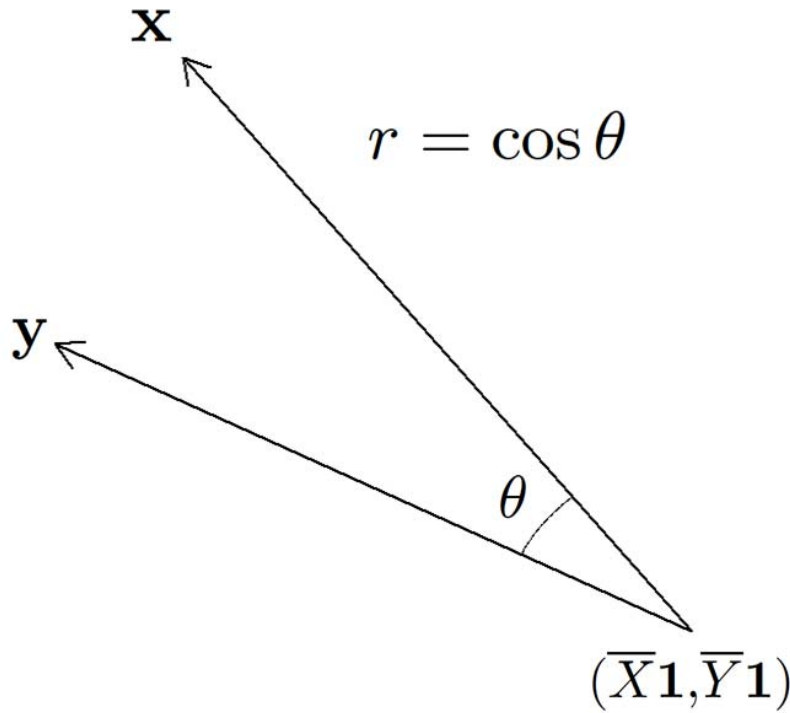
The norm of vector

$$\|\mathbf{a}\| = \sqrt{\sum_{i=1}^n a_i^2}.$$

Definition 8 *Scalar product of two vectors of the same length, $\mathbf{a}'\mathbf{b}$. The cosine angle between two vectors*

$$\cos \theta = \frac{\mathbf{a}'\mathbf{b}}{\|\mathbf{a}\| \|\mathbf{b}\|}.$$

We write $\mathbf{a} \perp \mathbf{b}$ if $\mathbf{a}'\mathbf{b} = 0$.



$$r = \text{correlation coefficient} = \cos \theta$$

Theorem 9 *The Cauchy inequality*

$$-\|\mathbf{a}\| \|\mathbf{b}\| \leq \mathbf{a}'\mathbf{b} \leq \|\mathbf{a}\| \|\mathbf{b}\|.$$

Definition 10 *Rank of matrix $\mathbf{A}^{n \times m}$ is the maximum number of linearly independent columns (rows). We say that this matrix has full rank if $\text{rank}(\mathbf{A}) = m \leq n$.*

Properties:

1. $\text{rank}(\mathbf{A}) = \text{rank}(\mathbf{A}')$.
2. $\text{rank}(\mathbf{A}) = \text{rank}(\mathbf{M}\mathbf{A})$ where \mathbf{M} is a nonsingular matrix.
3. $\text{rank}(\mathbf{A}\mathbf{B}) = \min(\text{rank}(\mathbf{A}), \text{rank}(\mathbf{B}))$

Matrix algebra and geometry

- Line in R^n through the origin is defined as the set of points $\{\mathbf{y} \in R^n : \mathbf{y} = \lambda \mathbf{x}\}$ where $\lambda \in (-\infty, \infty)$ and $\mathbf{x} \in R^n$ is called the direction vector (without loss of generality we may let $\|\mathbf{x}\| = 1$).
- Line in R^n passing through vector $\mathbf{a} \in R^n$ is defined as the set of points $\{\mathbf{y} \in R^n : \mathbf{y} = \mathbf{a} + \lambda \mathbf{x}\}$ where $\lambda \in (-\infty, \infty)$ and $\mathbf{x} \in R^n$ is called the direction vector.
- The plane of dimension=2 in R^n through the origin is defined as the set of points $\{\mathbf{y} \in R^n : \mathbf{y} = \lambda_1 \mathbf{x}_1 + \lambda_2 \mathbf{x}_2\}$ where $\lambda_1, \lambda_2 \in (-\infty, \infty)$ and $\mathbf{x}_1, \mathbf{x}_2 \in R^n$ is called span vectors. In an equivalent matrix formulation, the plane is defined as $\{\mathbf{y} \in R^n : \mathbf{y} = \mathbf{X}\boldsymbol{\lambda}\}$ where $\boldsymbol{\lambda}^{2 \times 1} = (\lambda_1, \lambda_2) \in R^2$ and $\mathbf{X}^{n \times 2} = [\mathbf{x}_1, \mathbf{x}_2]$ with $\text{rank}(\mathbf{X}) = 2$
- The plane of dimension = 2 in R^n passing through vector $\mathbf{a} \in R^n$ is defined as $\{\mathbf{y} \in R^n : \mathbf{y} = \mathbf{a} + \mathbf{X}\boldsymbol{\lambda}\}$ where $\boldsymbol{\lambda}^{2 \times 1} = (\lambda_1, \lambda_2) \in R^2$ and $\mathbf{X}^{n \times 2} = [\mathbf{x}_1, \mathbf{x}_2]$.
- The plane of dimension = m in R^n passing through vector $\mathbf{a} \in R^n$ is defined as $\{\mathbf{y} \in R^n : \mathbf{y} = \mathbf{a} + \mathbf{X}\boldsymbol{\lambda}\}$ where $\boldsymbol{\lambda}^{m \times 1} \in R^m$ and $\mathbf{X}^{n \times m}$ with $\text{rank}(\mathbf{X}) = m$
- The plane of dimension = $n-1$ in R^n passing through vector $\mathbf{a} \in R^n$ is defined as $\{\mathbf{y} \in R^n : (\mathbf{y} - \mathbf{a})' \mathbf{n}\}$ where $\mathbf{n}^{n \times 1} \in R^n$ is called *normal* vector. Typically, we chose $\|\mathbf{n}\| = 1$.

Proposition 11 *A vector \mathbf{y} orthogonal to the plane is orthogonal to any vector in the plane.*

Proof. To simplify, consider the plane passing the origin. Since \mathbf{a} belongs to the plane there exists \mathbf{b} such that $\mathbf{a} = \mathbf{X}\mathbf{b}$. The fact that \mathbf{y} is orthogonal to the plane means it is orthogonal to all vector-columns of \mathbf{X} , as the spanning vectors of the plane, that is,

$$\mathbf{X}'\mathbf{y} = \mathbf{0}.$$

Now we use matrix algebra

$$\mathbf{a}'\mathbf{y} = (\mathbf{X}\mathbf{b})'\mathbf{y} = \mathbf{b}'\mathbf{X}'\mathbf{y} = \mathbf{0}.$$

Example. Slope model in matrix formulation and simulations

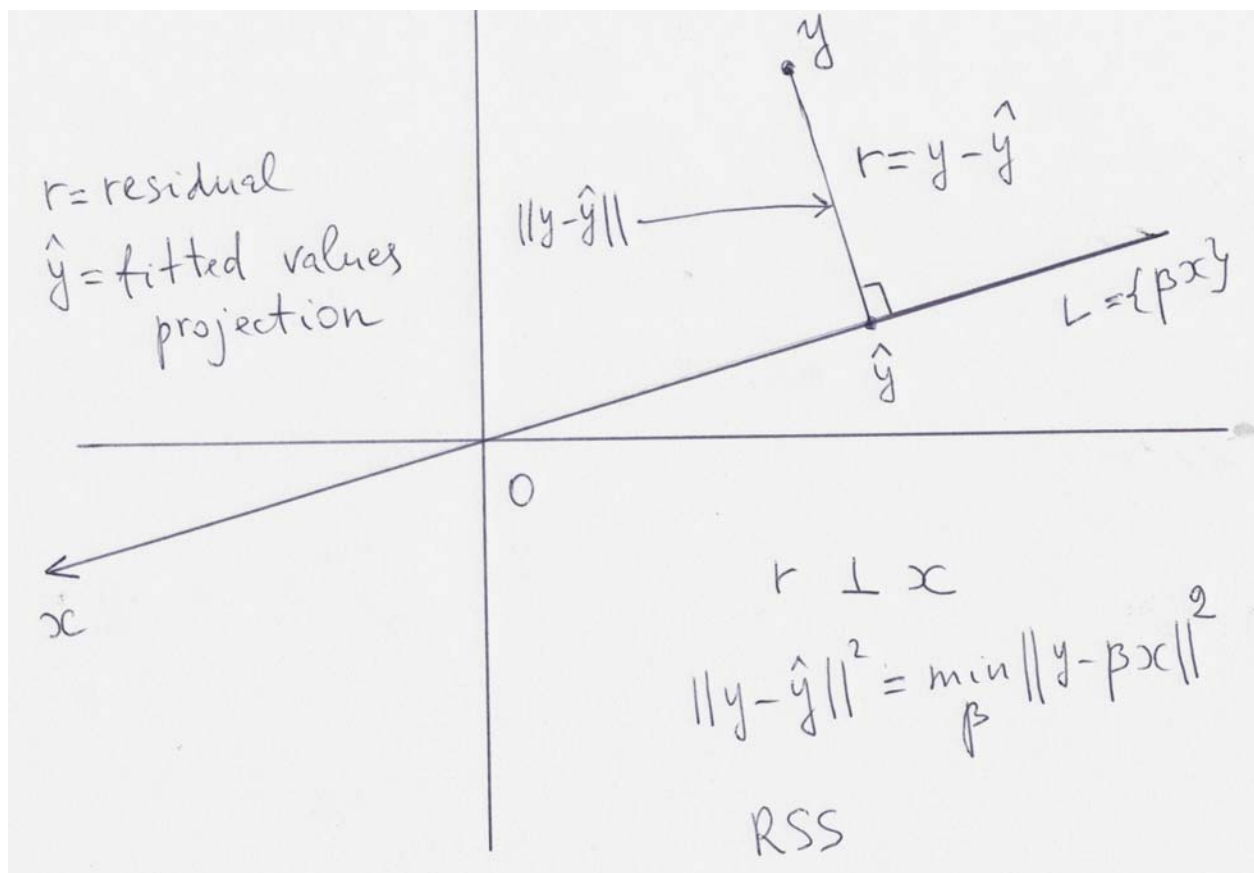
Section 6.5.2

A. Given a data vector $\mathbf{y}^{n \times 1} \in R^n$ find the minimum distance to a straight line defined by the direction vector $\mathbf{x}^{n \times 1} \in R^n$, that is,

$$L = \{\mathbf{u} \in R^n : \mathbf{u} = \beta \mathbf{x}, \quad -\infty < \beta < \infty\}.$$

We need to find

$$\min_{\beta} \|\mathbf{y} - \beta \mathbf{x}\|.$$



Easier to work with the squared distance.

We have

$$\begin{aligned}\|\mathbf{y} - \beta \mathbf{x}\|^2 &= (\mathbf{y} - \beta \mathbf{x})'(\mathbf{y} - \beta \mathbf{x}) = \mathbf{y}'\mathbf{y} - \beta \mathbf{y}'\mathbf{x} - \beta \mathbf{x}'\mathbf{y} + \beta^2 \|\mathbf{x}\|^2 \\ &= \|\mathbf{y}\|^2 - 2\beta \mathbf{y}'\mathbf{x} + \beta^2 \|\mathbf{x}\|^2\end{aligned}$$

Find the stationary point of the quadratic function

$$\frac{d}{d\beta} \|\mathbf{y} - \beta \mathbf{x}\|^2 = -2\mathbf{y}'\mathbf{x} + 2\beta \|\mathbf{x}\|^2 = 0,$$

called **normal equation**. Finally,

$$\hat{\beta} = \frac{\mathbf{y}'\mathbf{x}}{\|\mathbf{x}\|^2},$$

the least squares solution. An important property of this solution

$$\mathbf{y} - \hat{\beta} \mathbf{x} \perp \mathbf{x}.$$

It is a straightforward implication of the normal equation

$$0 = -\mathbf{y}'\mathbf{x} + \beta \|\mathbf{x}\|^2 = -(\mathbf{y} - \beta \mathbf{x})'\mathbf{x}.$$

The minimum distance from \mathbf{y} to L is

$$\|\mathbf{y} - \hat{\beta} \mathbf{x}\|.$$

B. Prove that the minimum distance is the length of the perpendicular dropped onto the line.

Define the projection of \mathbf{y} onto the line as

$$\hat{\mathbf{y}} = \hat{\beta}\mathbf{x},$$

It's called the vector of **fitted values**. Define the vector which connects \mathbf{y} and $\hat{\mathbf{y}}$ as the vector of **residuals**:

$$\mathbf{r} = \mathbf{y} - \hat{\mathbf{y}}.$$

$\mathbf{r} \perp \mathbf{x}$ because (\mathbf{r} is the **perpendicular**)

$$\begin{aligned} (\mathbf{y} - \hat{\beta}\mathbf{x})' \mathbf{x} &= 0, \\ \mathbf{r}' \mathbf{x} &= 0. \end{aligned}$$

Residuals are orthogonal to predictors/independent variables.

Express the perpendicular through the projection (hat) matrix

$$\mathbf{P} = \frac{1}{\|\mathbf{x}\|^2} \mathbf{x}\mathbf{x}'.$$

\mathbf{P} is the projection matrix (idempotent). Then

$$\hat{\mathbf{y}} = \left(\frac{1}{\|\mathbf{x}\|^2} \mathbf{x}\mathbf{x}' \right) \mathbf{y} = \mathbf{P}\mathbf{y}.$$

Define perpendicular/residual vector

$$\mathbf{r} = \mathbf{y} - \hat{\mathbf{y}} = \left(\mathbf{I} - \frac{1}{\|\mathbf{x}\|^2} \mathbf{x}\mathbf{x}' \right) \mathbf{y}.$$

Matrix

$$\mathbf{H} = \mathbf{I} - \frac{1}{\|\mathbf{x}\|^2} \mathbf{x}\mathbf{x}'$$

is called the **annihilator** matrix (also idempotent):

$$\mathbf{r} = \mathbf{H}\mathbf{y}.$$

C. Simulations: show that $\hat{\beta}$ is unbiased for the slope regression model

$$\mathbf{y} = \beta\mathbf{x} + \boldsymbol{\varepsilon}$$

where

$$\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n)'$$

are normally iid with zero mean and variance σ^2 .

R code:

```
simslope=function(b.true=0.5,sigma2=1.6,n=5,nSim=100000)
{
  dump("simslope","c:\\M7021\\simslope.r")
  x=1:n
  sqx=sum(x^2)
  b.hat=rep(NA,nSim)
```



```

for(i in 1:nSim)
{
  eps=rnorm(n,sd=sqrt(sigma2))
  y=b.true*x+eps
  scxy=sum(x*y)
  b.hat[i]=scxy/sqx
}
sim.mean=mean(b.hat)
print(c(sim.mean,b.true))
}

```

Remember: to read your text file with function `simslope` you issue `source("c:\\M7021\\simslope.r")` in the R console.

Homework 1

(20 points). Replace the slope model example with a slope model with intercept, $\mathbf{y} = \alpha \mathbf{1} + \beta \mathbf{x} + \varepsilon$. (A) Find the minimum distance between \mathbf{y} and the plane $\{\alpha \mathbf{1} + \beta \mathbf{x}, -\infty < \alpha < \infty, -\infty < \beta < \infty\}$. (B) Demonstrate by simulations that $\hat{\beta}$ is unbiased. (C) Demonstrate graphically by simulations that with increasing number of simulated observations the bias converges stochastically to zero (5 extra points). [Hint: Get rid of the intercept and reduce to the slope model.]

Homework 1 solution

(A). The squared distance is

$$S(\alpha, \beta) = \|\mathbf{y} - \alpha \mathbf{1} - \beta \mathbf{x}\|^2.$$

To find the minimum over α and β we fix β and denote $\mathbf{z} = \mathbf{y} - \beta \mathbf{x}$. Find the minimum of

$$\|\mathbf{z} - \alpha \mathbf{1}\|^2 = \|\mathbf{z}\|^2 - 2\alpha \mathbf{1}'\mathbf{z} + \alpha^2 n$$

because $\|\mathbf{1}\|^2 = n$. Differentiating with respect to α we obtain.

$$\hat{\alpha} = \frac{1}{n} \mathbf{1}'\mathbf{z} = \frac{1}{n} \mathbf{1}'(\mathbf{y} - \beta \mathbf{x}) = \bar{y} - \beta \bar{x}.$$

Indeed, this gives us the minimum because quadratic function is convex ($n > 0$). Plugging it into S we obtain

$$S(\alpha, \beta) = \|\mathbf{y} - (\bar{y} - \beta \bar{x}) \mathbf{1} - \beta \mathbf{x}\|^2 = \|(\mathbf{y} - \bar{y} \mathbf{1}) - \beta(\mathbf{x} - \bar{x} \mathbf{1})\|^2 = \|\mathbf{y}_0 - \beta \mathbf{x}_0\|^2$$

where

$$\mathbf{y}_0 = \mathbf{y} - \bar{y} \mathbf{1}, \quad \mathbf{x}_0 = \mathbf{x} - \bar{x} \mathbf{1}.$$

But the minimum of $\|\mathbf{y}_0 - \beta \mathbf{x}_0\|^2$ is the minimum of the slope model with

$$\begin{aligned} \hat{\beta} &= \frac{\mathbf{y}_0' \mathbf{x}_0}{\|\mathbf{x}_0\|^2} \\ &= \frac{(\mathbf{y} - \bar{y} \mathbf{1})'(\mathbf{x} - \bar{x} \mathbf{1})}{\|\mathbf{x} - \bar{x} \mathbf{1}\|^2}. \end{aligned}$$

B. Modify function `simslope`

```
source("c:\\M7021\\simslope.r")
simslint=edit(simslope)
```

```
simslint=function(b.true=0.5,a.true=-1,sigma2=1.6,n=5,nSim=10000)
```

```
{
  dump("simslint","c:\\M7021\\simslint.r")
```

```
  x=1:n
```

```
  avx=mean(x)
```

```
  x.m.av=x-avx
```

```
  sqx=sum(x.m.av^2)
```

```
  b.hat=rep(NA,nSim)
```

```
  for(i in 1:nSim)
```

```
  {
```

```
    eps=rnorm(n,sd=sqrt(sigma2))
```

```
    y=a.true+b.true*x+eps
```

```
    yav=y-mean(y)
```

```
    scxy=sum(yav*x.m.av)
```

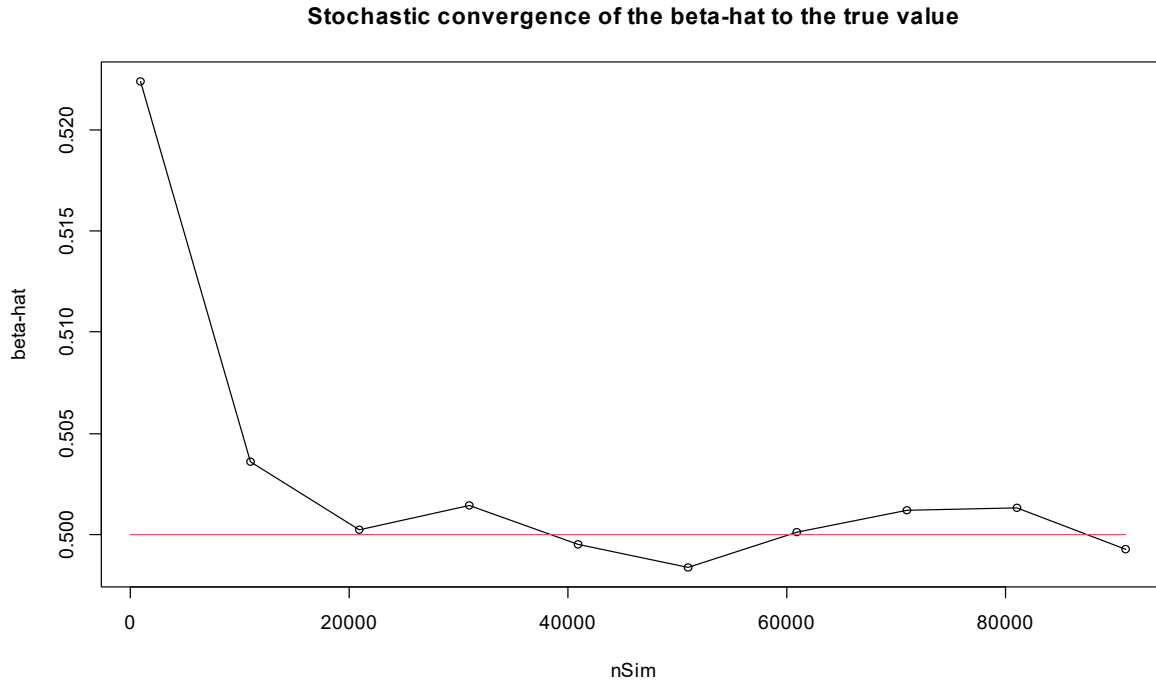
```
    b.hat[i]=scxy/sqx
```

```
  }
```

```
  sim.mean=mean(b.hat)
```

```
  print(c(sim.mean,b.true))
```

C. The program looks like the previous one with the double loop with the outer loop over the array of `nSims`.



```

simslintG=function(b.true=0.5,a.true=-1,sigma2=1.6,n=5,nSim=seq(from=1000,to=100000,by=10000))
{
  dump("simslintG","c:\\M7021\\simslintG.r")
  x=1:n
  avx=mean(x)
  x.m.av=x-avx
  sqx=sum(x.m.av^2)
  nnsim=length(nSim)
  sim.mean=rep(NA,nnsim)
  for(inn in 1:nnsim)
  {
    b.hat=rep(NA,nSim[inn])
    for(i in 1:nSim[inn])
    {
      eps=rnorm(n,sd=sqrt(sigma2))
      y=a.true+b.true*x+eps
      yav=y-mean(y)
      scxy=sum(yav*x.m.av)
      b.hat[i]=scxy/sqx
    }
    sim.mean[inn]=mean(b.hat)
  }
  plot(nSim,sim.mean,type="o",xlab="nSim",ylab="beta-hat",main="Stochastic convergence of the beta-hat
to the true value")
  segments(0,b.true,max(nSim),b.true,col=2)
}

```