An Analysis

Jane Doe

Department of Biostatistics

Vanderbilt University School of Medicine
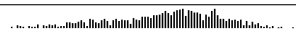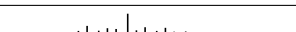
November 2, 2015

Contents

1   Descriptive Statistics

```
getHdata(support)    # Use Hmisc/getHdata to get dataset from VU DataSets wiki
d ← subset(support, select=c(age,sex,race,edu,income,hospdead,slos,dzgroup,
                             meanbp,hrt))
latex(describe(d), file='')
```


d

10 Variables　　**1000**　**Observations**

age : Age

	n	missing	unique	Info	Mean	.05	.10	.25	.50	.75	.90	.95
	1000	0	970	1	62.47	33.76	38.91	51.81	64.90	74.50	81.87	86.00

```
lowest :  18.04  18.41  19.76  20.30  20.31
highest:  95.51  96.02  96.71 100.13 101.85
```

sex

	n	missing	unique
	1000	0	2

```
female (438, 44%), male (562, 56%)
```

race

	n	missing	unique
	995	5	5

	white	black	asian	other	hispanic
Frequency	781	157	9	12	36
%		78	16	1	1	4

edu : Years of Education

	n	missing	unique	Info	Mean	.05	.10	.25	.50	.75	.90	.95
	798	202	25	0.97	11.78	6	8	10	12	14	16	18

```
lowest :  0  1  2  3  4, highest: 20 21 22 24 30
```

income

	n	missing	unique
	651	349	4

```
under $11k (309, 47%), $11-$25k (161, 25%), $25-$50k (106, 16%)
>$50k (75, 12%)
```

hospdead : Death in Hospital

n	missing	unique	Info	Sum	Mean
1000	0	2	0.57	253	0.253

slos : Days from Study Entry to Discharge

n	missing	unique	Info	Mean	.05	.10	.25	.50	.75	.90	.95
1000	0	88	1	17.86	4	4	6	11	20	37	53

lowest :   3   4   5   6   7, highest: 145 164 202 236 241

dzgroup

n	missing	unique
1000	0	8

	ARF/MOSF w/Sepsis	COPD	CHF	Cirrhosis	Coma	Colon Cancer	Lung Cancer
Frequency	391	116	143	55	60	49	100
%	39	12	14	6	6	5	10

	MOSF w/Malig
Frequency	86
%	9

meanbp : Mean Arterial Blood Pressure Day 3

n	missing	unique	Info	Mean	.05	.10	.25	.50	.75	.90	.95
1000	0	122	1	84.98	47.00	55.00	64.75	78.00	107.00	120.00	128.05

lowest :   0 20 27 30 32, highest: 155 158 161 162 180

hrt : Heart Rate Day 3

n	missing	unique	Info	Mean	.05	.10	.25	.50	.75	.90	.95
1000	0	124	1	97.87	54.0	60.0	72.0	100.0	120.0	135.0	146.1

lowest :   0 11 30 35 36, highest: 189 193 199 232 300

Race is reduced to three levels (white, black, OTHER) because of low frequencies in other levels (minimum relative frequency set to 0.05).

```
d ← upData(d,
           race = combine.levels(race, minlev = 0.05))
```

```
Input object size:        107336 bytes;    10 variables     1000 observations
Modified variable         race
New object size:          107216 bytes;    10 variables     1000 observations
```

2   Redundancy Analysis and Variable Interrelationships

```
v ← varclus(∼., data=d)
plot(v)
redun(∼age+sex+race+edu+income+dzgroup+meanbp+hrt, data=d)
```

```
Redundancy Analysis

redun(formula = ∼age + sex + race + edu + income + dzgroup +
    meanbp + hrt, data = d)

n: 617  p: 8    nk: 3

Number of NAs:   383
Frequencies of Missing Values Due to Each Variable
    age      sex     race      edu   income  dzgroup   meanbp      hrt
      0        0        5      202      349        0        0        0



Transformation of target variables forced to be linear
```

R^2 cutoff: 0.9 Type: ordinary

R^2 with which each variable can be predicted from all other variables:
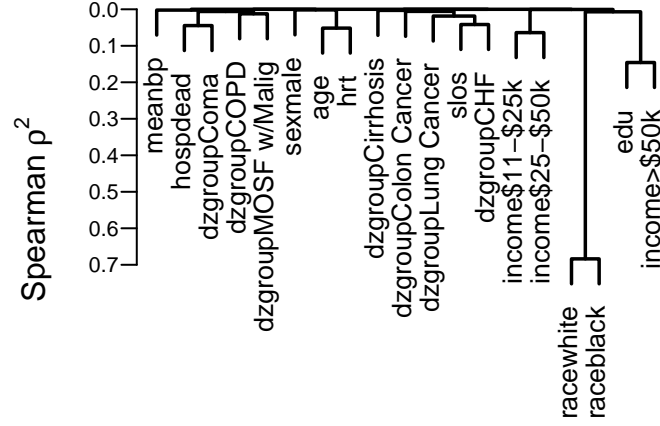
```
    age      sex     race      edu   income  dzgroup   meanbp      hrt
  0.196    0.088    0.120    0.284    0.339    0.253    0.067    0.163

No redundant variables
```

```
# Alternative: redun(~., data=subset(d, select=-c(hospdead,slos)))
```


Note that the clustering of black with white is not interesting; this just means that these are mutually exclusive higher frequency categories, causing them to be negatively correlated.

3   Logistic Regression Model

Here we fit a tentative binary logistic regression model. The coefficients are not very useful so they are not printed.

```
dd ← datadist(d); options(datadist='dd')
f ← lrm(hospdead ~ rcs(age,4) + sex + race + dzgroup + rcs(meanbp,5),
        data=d) # see Section \ref{descStats} for descriptive statistics
print(f, latex=TRUE, coefs=FALSE)
```

Logistic Regression Model

```
lrm(formula = hospdead ~ rcs(age, 4) + sex + race + dzgroup +
    rcs(meanbp, 5), data = d)
```

Frequencies of Missing Values Due to Each Variable

```
hospdead       age      sex     race  dzgroup   meanbp
       0         0        0        5        0        0
```

		Model Likelihood Ratio Test		Discrimination Indexes		Rank Discrim. Indexes	
Obs	995	LR χ^2	245.83	R^2	0.323	C	0.800
0	744	d.f.	17	g	1.605	D_{xy}	0.601
1	251	$\Pr(> \chi^2)$	< 0.0001	g_r	4.980	γ	0.602
$\max \lvert \frac{\partial \log L}{\partial \beta} \rvert$	1×10^{-9}			g_p	0.228	τ_a	0.227
				Brier	0.144		

```
latex(anova(f), where='h', file='')    # can also try where='htbp'
```

Table 1: Wald Statistics for `hospdead`

	χ^2	d.f.	P
age	7.12	3	0.0683
Nonlinear	2.91	2	0.2338
sex	2.16	1	0.1413
race	1.38	2	0.5005
dzgroup	78.77	7	< 0.0001
meanbp	65.62	4	< 0.0001
Nonlinear	48.11	3	< 0.0001
TOTAL NONLINEAR	50.15	5	< 0.0001
TOTAL	151.71	17	< 0.0001

4   Test Calculations

```
x ← 3; y ← 2
if(x ≤ y) 'this' else 'that'
```
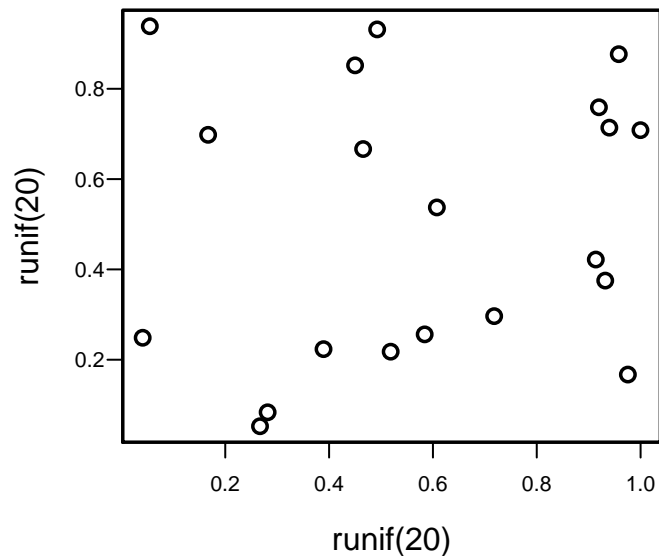
```
[1] "that"
```

```
if(y ≥ x) 'that' else 'this'
```

```
[1] "this"
```

```
x^y
```

```
[1] 9
```

```
plot(runif(20),runif(20))
```

5   Computing Environment

These analyses were done using the following versions of R[3], the operating system, and add-on packages Hmisc[1], rms[2], and others:

- R version 3.2.2 (2015-08-14), `x86_64-pc-linux-gnu`

- Base packages: base, datasets, graphics, grDevices, grid, methods, stats, utils

- Other packages: Formula 1.2-1, ggplot2 1.0.1, Hmisc 3.17-0, knitr 1.11, lattice 0.20-33, rms 4.4-1, SparseM 1.7, survival 2.38-3

- Loaded via a namespace (and not attached): acepack 1.3-3.3, cluster 2.0.3, codetools 0.2-14, colorspace 1.2-6, digest 0.6.8, evaluate 0.8, foreign 0.8-66, formatR 1.2.1, gridExtra 2.0.0, gtable 0.1.2, latticeExtra 0.6-26, magrittr 1.5, MASS 7.3-44, Matrix 1.2-2, MatrixModels 0.4-1, multcomp 1.4-1, munsell 0.4.2, mvtnorm 1.0-3, nlme 3.1-122, nnet 7.3-11, plyr 1.8.3, polspline 1.1.12, proto 0.3-10, quantreg 5.19, RColorBrewer 1.1-2, Rcpp 0.12.1, reshape2 1.4.1, rpart 4.1-10, sandwich 2.3-4, scales 0.3.0, splines 3.2.2, stringi 1.0-1, stringr 1.0.0, TH.data 1.0-6, tools 3.2.2, zoo 1.7-12
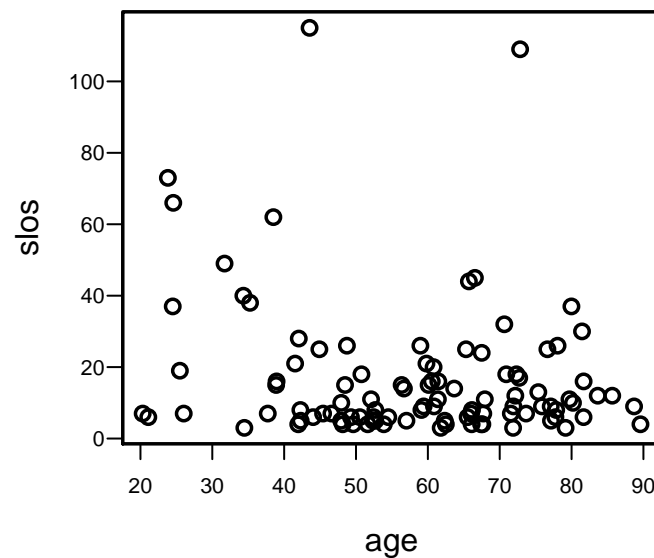
6   Descriptive Statistics Again

```
getHdata(support)    # Use Hmisc/getHdata to get dataset from VU DataSets wiki
d ← subset(support, select=c(age,sex,race,edu,income,hospdead,slos,dzgroup,
                 meanbp,hrt))
summary(d)
```

```
      age                sex              race              edu
 Min.    : 18.04    female:438    white   :781    Min.    : 0.00
 1st Qu.: 51.81    male  :562    black   :157    1st Qu.:10.00
 Median : 64.90                  asian   :  9    Median :12.00
 Mean    : 62.47                 other   : 12    Mean    :11.78
 3rd Qu.: 74.50                  hispanic: 36    3rd Qu.:14.00
 Max.    :101.85                 NA's    :  5    Max.    :30.00
                                                 NA's    :202

      income         hospdead           slos                      dzgroup
 under $11k:309    Min.    :0.000    Min.    :  3.00    ARF/MOSF w/Sepsis:391
 $11-$25k  :161    1st Qu.:0.000    1st Qu.:  6.00    CHF              :143
 $25-$50k  :106    Median :0.000    Median :  11.00    COPD             :116
 >$50k     : 75    Mean    :0.253    Mean    :  17.86    Lung Cancer      :100
 NA's      :349    3rd Qu.:1.000    3rd Qu.:  20.00    MOSF w/Malig     : 86
                   Max.    :1.000    Max.    :241.00    Coma             : 60
                                                        (Other)          :104

      meanbp              hrt
 Min.    :  0.00    Min.    :  0.00
 1st Qu.: 64.75    1st Qu.: 72.00
 Median : 78.00    Median :100.00
 Mean    : 84.98    Mean    : 97.87
 3rd Qu.:107.00    3rd Qu.:120.00
 Max.    :180.00    Max.    :300.00
```

```
with(d, hist(age, nclass=25, main=''))
```
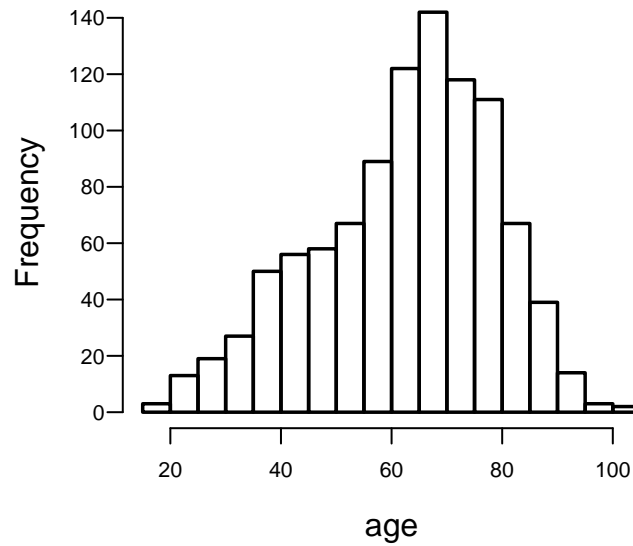
```
with(d[1:100,], plot(age, slos))
```


That was a very concise set of descriptive statistics.

References

[1] Frank E. Harrell. Hmisc: *A package of miscellaneous* R *functions.* 2015. URL: http://biostat.mc.vanderbilt.edu/Hmisc.

[2] Frank E. Harrell. *rms: R functions for biostatistical/epidemiologic modeling, testing, estimation, validation, graphics, prediction, and typesetting by storing enhanced model design attributes in the fit.* Implements methods in *Regression Modeling Strategies, 2nd edition.* 2015. URL: http://biostat.mc.vanderbilt.edu/rms.

[3] R Development Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing. Vienna, Austria, 2015. URL: http://www.R-project.org.