> library(MASS)

> data(birthwt)

> str(birthwt)

'data.frame': 189 obs. of 10 variables:

$ low : int 0 0 0 0 0 0 0 0 0 0 ...

$ age : int 19 33 20 21 18 21 22 17 29 26 ...

$ lwt : int 182 155 105 108 107 124 118 103 123 113 ...

$ race : int 2 3 1 1 1 3 1 3 1 1 ...

$ smoke: int 0 0 1 1 1 0 0 0 1 1 ...

$ ptl : int 0 0 0 0 0 0 0 0 0 0 ...

$ ht : int 0 0 0 0 0 0 0 0 0 0 ...

$ ui : int 1 0 0 1 1 0 0 0 0 0 ...

$ ftv : int 0 3 1 2 0 0 1 1 1 0 ...

$ bwt : int 2523 2551 2557 2594 2600 2622 2637 2637 2663 2665 ...

To get a gentle numerical overview of the data, we can use the

summary function.

> summary(birthwt)

low age lwt race

Min. :0.0000 Min. :14.00 Min. : 80.0 Min. :1.000

1st Qu.:0.0000 1st Qu.:19.00 1st Qu.:110.0 1st Qu.:1.000

Median :0.0000 Median :23.00 Median :121.0 Median :1.000

Mean :0.3122 Mean :23.24 Mean :129.8 Mean :1.847

3rd Qu.:1.0000 3rd Qu.:26.00 3rd Qu.:140.0 3rd Qu.:3.000

Max. :1.0000 Max. :45.00 Max. :250.0 Max. :3.000 smoke ptl ht ui

Min. :0.0000 Min. :0.0000 Min. :0.00000 Min. :0.0000

1st Qu.:0.0000 1st Qu.:0.0000 1st Qu.:0.00000 1st Qu.:0.0000

Median :0.0000 Median :0.0000 Median :0.00000 Median :0.0000

Mean :0.3915 Mean :0.1958 Mean :0.06349 Mean :0.1481

3rd Qu.:1.0000 3rd Qu.:0.0000 3rd Qu.:0.00000 3rd Qu.:0.0000

Max. :1.0000 Max. :3.0000 Max. :1.00000 Max. :1.0000 ftv bwt

Min. :0.0000 Min. : 709

1st Qu.:0.0000 1st Qu.:2414

Median :0.0000 Median :2977

Mean :0.7937 Mean :2945

3rd Qu.:1.0000 3rd Qu.:3487

Max. :6.0000 Max. :4990

The very first thing to do with unknown datasets is to check for the presence of missing values (mv). This is done with is.na, but see also complete.cases. Some statistical models doesn’t accomodate well

with mv, others delete them listwise. In R, mv are stored as NA.

> apply(birthwt, 2, function(x) sum(is.na(x)))

low age lwt race smoke ptl ht ui ftv bwt

0 0 0 0 0 0 0 0 0 0

> dim(birthwt)

[1] 189 10

Recoding and checking variables

Some of the factors of interest will not be understood by R as we would like it to do because they are just treated as numerical vari- ables. So, the next step is to convert them to (unordered) fac-

tor. The within command provides a convenient to update several

columns in a data.frame within a single call.

> birthwt <- within(birthwt, {

+ low <- factor(low, labels=c("No","Yes"))

+ race <- factor(race, labels=c("White","Black","Other"))

+ smoke <- factor(smoke, labels=c("No","Yes"))

+ ui <- factor(ui, labels=c("No","Yes"))

+ ht <- factor(ht, labels=c("No","Yes"))

+ })

It is important to take care of the way R represents data, especially when we 

want it to treat some variable as a factor or discrete-valued vector.

Again, with unknown data, it is recommended to take a close look at the distribution of numerical variables, which in this particular case can be isolated using a repeated call to is.numeric.

> idx <- sapply(birthwt, is.numeric)

We then use box-and-whiskers charts to summarize each distribu-

tion. Note that we have rescaled them to avoid the problem of varying

6

y-axis.

4

0

> boxplot(apply(birthwt[,idx], 2, scale))

2

Summarizing data

−2

Most base functions in R can be used to summarize our variables, either numerically or graphically. We have already seen the useful summary command. For numerical variables, it will output a five- number summary; for categorical data, a frequency of counts. In both cases, missing cases will be reported separately. However, there is a

full set of dedicated tools included in the [Hmisc package,](http://biostat.mc.vanderbilt.edu/wiki/Main/RS) especially the

summary.formula command.

> library(Hmisc)

> summary(low ~ ., data=birthwt, method="reverse", overall=TRUE)

> summary(bwt ~ ., data=birthwt)

> summary(low ~ smoke + ht + ui, data=birthwt, fun=table)

low N=189

+-------+---+---+---+---+

| | |N |No |Yes|

+-------+---+---+---+---+

|smoke |No |115| 86|29 |

| |Yes| 74| 44|30 |

+-------+---+---+---+---+

|ht |No |177|125|52 |

| |Yes| 12| 5| 7 |

+-------+---+---+---+---+

|ui |No |161|116|45 |

| |Yes| 28| 14|14 |

+-------+---+---+---+---+

|Overall| |189|130|59 |

+-------+---+---+---+---+

The above commands might be combined with the latex function to produce pretty-print output, which follow standards for publications in biomedical journal.

Multivariate displays can be used, essentially to study relation- ships between numerical variables or assess any systematic patterns of covariations. At this stage, it might be interesting to highlight individ- uals according to a certain characteristic (e.g., low birth weight).

> library(lattice)

> parallel(birthwt[,idx], groups=birthwt$low, horizontal.axis=FALSE)

> print(parallel(~ birthwt[,idx] | smoke, data=birthwt, groups=low,

+ lty=1:2, col=c("gray80","gray20")))

age lwt ptl ftv bwt

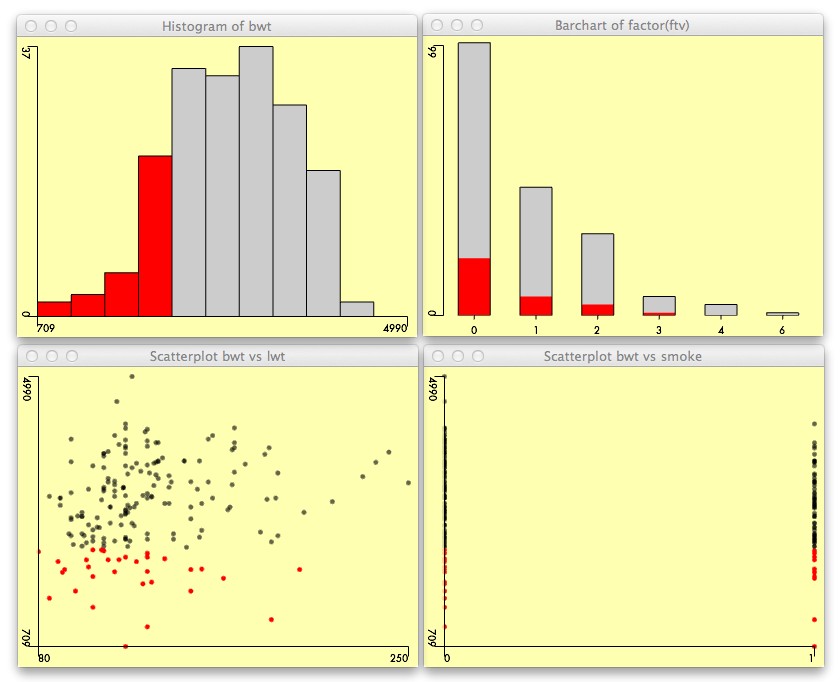
> splom(~ birthwt[,idx], data=birthwt, jitter.x=TRUE, cex=.6,

+ groups=low, panel=panel.superpose, grid=TRUE,

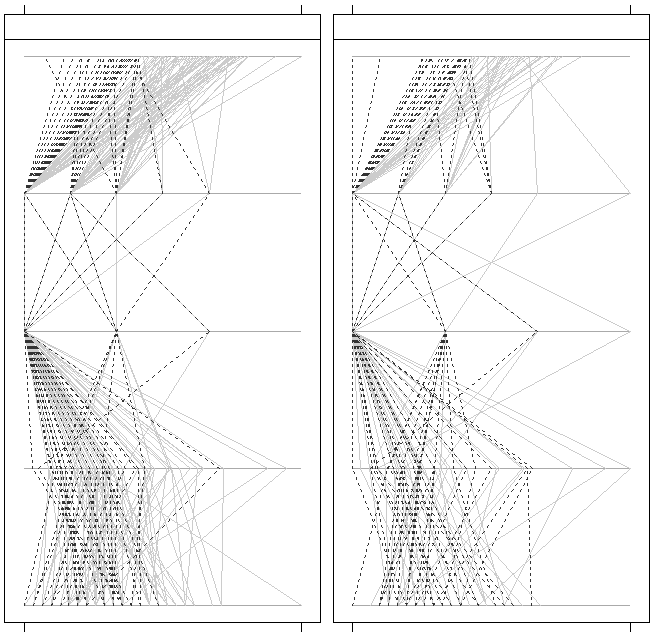
+ axis.text.cex=0.6, xlab="", pch=c(2,6))

Visual exploration of the dataset3 can also rely on brushing and 3

linking techniques available in the [ggobi software,](http://www.ggobi.org/) or directly with the Acinonyx package. Below is a screenshot after having selected low weight infants from the histogram.



No



bwt

ftv

Min Max

Yes

> library(Acinonyx)

> attach(birthwt)

> iplot(lwt, bwt)

> iplot(smoke, bwt)

> ihist(bwt)

> ibar(factor(ftv))

ptl

lwt

age

Min Max

> detach(birthwt)

Dealing with continuous outcomes

Assessing linear two-way relationships

The covariation between mother’s weight in pounds at last menstrual period and birth weight can be assessed using a correlation test. Of course, we need to display the raw data in a scatterplot, first. It is possible to add a local regression line to the cloud of points to visually assess the linearity of the relationship.4

4

> library(lattice)

> print(xyplot(bwt ~ lwt, data=birthwt, type=c("p", "smooth")))

> with(birthwt, cor.test(bwt, lwt))

Pearson's product-moment correlation data: bwt and lwt

t = 2.5848, df = 187, p-value = 0.0105

alternative hypothesis: true correlation is not equal to 0

95 percent confidence interval:

0.04417405 0.31998094 sample estimates:

cor

0.1857333

Two notes of caution about correlation. Correlation does not necessarily 

mean there exists a causal relationship between the two variables of in-

terest. The use of linear correlation coefficient, like Bravais-Pearson r, is for assessing linear relationship. See Anscombe’s famous illustration, help(anscombe).

Compare the results to the ones obtained using Spearman’s method,

which relies on ranks. I

> with(birthwt, cor.test(bwt, lwt, method="spearman"))

Spearman's rank correlation rho data: bwt and lwt

S = 845135.9, p-value = 0.0005535

alternative hypothesis: true rho is not equal to 0 sample estimates:

rho

0.2488882

Comparing two means

To test the hypothesis that there is no difference in baby weight de-

pending on the smoking status (of the mother), we can use a t-test5 5

for independent samples where the standard error is computed from

sc n−1 −1

the pooled standard deviation (weighted average of sample sds).

> t.test(bwt ~ smoke, data=birthwt, var.equal=TRUE)

Two Sample t-test data: bwt by smoke

t = 2.6529, df = 187, p-value = 0.008667

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

72.75612 494.79735 sample estimates:

mean in group No mean in group Yes

3055.696 2771.919

> t.test(bwt ~ smoke, data=birthwt)

Welch Two Sample t-test data: bwt by smoke

t = 2.7299, df = 170.1, p-value = 0.007003

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

78.57486 488.97860 sample estimates:

mean in group No mean in group Yes

3055.696 2771.919

It is always a good idea to check whether the assumptions of the test are met, by computing the variances and displaying a quantile- quantile plot. This allows to gauge the homogeneity of variances and the normality of the distributions (see discussion p. 7).

> with(birthwt, tapply(bwt, smoke, var))

No Yes

566492.0 435118.2

1000 1500 2000 2500 3000 3500 4000

> with(birthwt, qqnorm(bwt[smoke=="No"]

, main=""))

> with(birthwt, qqnorm(bwt[smoke=="Yes"], main=""))

Sample Quantiles

Sample Quantiles

In addition, we can produce a graphical summary of the two distri- butions using boxplots or histograms. To superimpose a kernel density estimate on the latter, see panel.mathdensity .

> print(bwplot(bwt ~ smoke, data=birthwt))

4000 5000

1000 2000

3000

−2 0 1 2

Theoretical Quantiles

−2 −1 0 1 2

Theoretical Quantiles

> print(histogram(~ bwt | smoke, data=birthwt, type="count"))

.

Should we want to use a non-parametric (distribution free) statistic,

we could apply a Wilcoxon test with wilcox.test.6 6 .

Using re-randomization instead of parametric testing

We could use a permutation technique instead of the t-test, and con- sider an approximate or exact distribution of the test statistic to de- cide whether there is a significant difference between central locations of the two groups.

> library(coin)

> independence\_test(bwt ~ smoke, data=birthwt, distribution="exact")

Exact General Independence Test data: bwt by smoke (No, Yes)

No

25

20

15

10

5

Count

0

1000 2000 3000 4000 5000

bwt

1000 2000 3000 4000 5000

Yes

Z = 2.6113, p-value = 0.008678 alternative hypothesis: two.sided

Comparing more than two means

In case we are interested in comparing more than two groups, we can

use an anova7 (instead of carrying out multiple unprotected t-tests). 7

1000 2000

> fm <- bwt ~ race

> aov.fit <- aov(fm, data=birthwt)

> summary(aov.fit)

Df Sum Sq Mean Sq F value Pr(>F)

race 2 5015725 2507863 4.9125 0.008336

Residuals 186 94953931 51050

> plot(aov.fit, which=1)

The plot command is a generic method for linear models objects;

It is important to understand that the “normality assumption” concerns the 

distribution of the residuals, not the raw data values. This is obvious when we reframe the anova as a linear model, yij = *µ* + *α*i + *ε*ij where the *α*i are deviations from the grand mean, for group i = 1, . . . , I, and *ε*ij ∼ N (0; *σ*2) (i.e., residuals are assumed to be i.i.d. gaussian variates).

Residuals vs Fitted

226

The summary command gives the classical anova table. Some-

times, we are interested in judging whether the between-group means

Residuals

0

exhibit large differences or not. We can use tapply, or simply the

model.table command.

−2000 −1000

> model.tables(aov.fit)

Tables of effects race

White Black Other

158.1 -224.9 -139.3 rep 96.0 26.0 67.0

> plot.design(fm, data=birthwt)

4 10

2800 2900 3000 3100

Fitted values aov(fm)

3100

White

Likewise, a two-way anova can be used to test whether birth weight depends on both smoking status and history of hypertension. If we consider a saturated model, we need to include the interaction smoke : ht in addition to the main effects, smoke + ht. Following

mean of bwt

2800

2900

3000

Wilkinson and Rogers’ notation,8 this reduces to smoke \* ht.

Other

> summary(aov.fit2 <- aov(bwt ~ smoke \* ht, data=birthwt))

Black

race

Df Sum Sq Mean Sq F value Pr(>F) smoke 1 3625946 3625946 7.1390 0.008216 ht 1 2056920 2056920 4.0498 0.045627 smoke:ht 1 324044 324044 0.6380 0.425461

Residuals 185 93962746 507907

Factors

> with(birthwt, interaction.plot(smoke, ht, bwt))

As the interaction appears to be non significant, we can remove this term and fit a model including main effects only.

> aov.fit3 <- aov(bwt ~ smoke + ht, data=birthwt)

mean of bwt

> anova(aov.fit2, aov.fit3)

Analysis of Variance Table Model 1: bwt ~ smoke \* ht Model 2: bwt ~ smoke + ht

Res.Df RSS Df Sum of Sq F Pr(>F)

1 185 93962746

2500 2600 2700 2800 2900 3000 3100

2 186 94286790 -1 -324044 0.638 0.4255

No Yes smoke

ht

No

Yes

Adding a third term, race, with its two second order interactions, indicates that ethnicity also impacts baby weight, with no dependence

.

on smoking status or history of hypertension.

> summary(aov.fit4 <- update(aov.fit3, . ~ . \* race))

Df Sum Sq Mean Sq F value Pr(>F) smoke 1 3625946 3625946 7.8780 0.005555 ht 1 2056920 2056920 4.4690 0.035892 race 2 8291518 4145759 9.0074 0.000187 smoke:race 2 1754625 877313 1.9061 0.151646 ht:race 2 1393149 696575 1.5134 0.222943

Residuals 180 82847497 460264

We can use multiple comparisons, controlling for fwer,9 on the

final model using Tukey’s hsd contrasts. In fact, only the race effect

is of interest here, because the factor has more than two levels. Results can be compared with what would be obtained using t-tests

> with(birthwt, pairwise.t.test(bwt, race))

Pairwise comparisons using t tests with pooled SD

data: bwt and race

White Black Black 0.033 - Other 0.029 0.605

P value adjustment method: holm

and a step-down method to adjust p-values, see help(p.adjust) for more information on correction methods available in R.

> aov.fit5 <- update(aov.fit3, . ~ . + race)

> model.tables(aov.fit5)

Tables of effects smoke

No Yes

111.1 -172.7 rep 115.0 74.0

ht

No Yes

27.16 -400.6 rep 177.00 12.0

race

White Black Other

195.4 -204.6 -200.6 rep 96.0 26.0 67.0

> se.contrast(aov.fit5, list(birthwt$race=="White",

+ birthwt$race=="Black"))

[1] 151.1423

> TukeyHSD(aov.fit5, which="race")

Tukey multiple comparisons of means

95% family-wise confidence level

Fit: aov(formula = bwt ~ smoke + ht + race, data = birthwt)

$race

diff lwr upr p adj

Black-White -400.059794 -757.1842 -42.93537 0.0238683

Other-White -396.023345 -653.1711 -138.87560 0.0010320

Other-Black 4.036449 -369.1959 377.26884 0.9996401

> plot(TukeyHSD(aov.fit5, which="race"))

**95% family−wise confidence level**

Other−Black

Other−White

Black−White

Linear regression

It should be clear that the above models can be tested using classical linear regression.10 In this case, we use the lm command. Note that baseline categories will be the first lexicographic entry for each factor levels. Model diagnostics can be assessed using the plot method, as discussed page 7.

> lm.fit5 <- lm(bwt ~ smoke + ht + race, data=birthwt)

> summary(lm.fit5)

Call:

lm(formula = bwt ~ smoke + ht + race, data = birthwt) Residuals:

Min 1Q Median 3Q Max

-2331.70 -462.03 -6.03 474.30 1637.30

Coefficients:

Estimate Std. Error t value Pr(>|t|) (Intercept) 3352.70 91.65 36.580 < 2e-16 smokeYes -424.68 108.33 -3.920 0.000125 htYes -383.06 204.73 -1.871 0.062932 raceBlack -425.47 152.68 -2.787 0.005882 raceOther -448.49 115.72 -3.876 0.000148

Residual standard error: 683.6 on 184 degrees of freedom

Multiple R-squared: 0.1398, Adjusted R-squared: 0.1211

F-statistic: 7.475 on 4 and 184 DF, p-value: 1.335e-05

> anova(lm.fit5)

−800 −600 −400 −200 0 200 400

Differences in mean levels of race

Analysis of Variance Table

Response: bwt

Df Sum Sq Mean Sq F value Pr(>F) smoke 1 3625946 3625946 7.7583 0.005906 ht 1 2056920 2056920 4.4011 0.037282 race 2 8291518 4145759 8.8705 0.000210

Residuals 184 85995271 467366

> plot(lm.fit5, which=4)

The above command just shows influential observations as assessed by Cook’s distances. Other measures do exist, but all are based on

residuals (standardized or not). R provides a convenient formula-

It is interesting to take a closer look at the so-called “design matrix”

for this particular model since it is composed of a mix of numerical and categorical predictors.

> head(model.matrix(lm.fit5))

(Intercept) smokeYes htYes raceBlack raceOther

85 1 0 0 1 0

86 1 0 0 0 1

87 1 1 0 0 0

88 1 1 0 0 0

like notation to express the relation- ship between the outcome of interest and each predictors, as shown below.

**formula:** response ~ predictors **residuals:** distribution and link function

glm(low ~ age + lwt + race + ftv, data = birthwt, family = binomial(logit), subset = smoke == "No", na.action = na.omit)

**restriction:** subsample **missing values:** listwise deletion

0.00 0.02 0.04 0.06 0.08 0.10

Cook's distance

202

197

89 1 1 0 0 0

Cook's distance

91 1 0 0 0 1

As can be seen, categorical variables are coded as a set of dummy variables; for k categories, we only need k − 1 vectors of 0/1, the base- line category (first level of the factor in R) being omitted. Compare the preceding output with

> head(birthwt[,c("smoke","ht","race")])

smoke ht race

85 No No Black

86 No No Other

87 Yes No White

88 Yes No White

89 Yes No White

91 No No Other

13

0 50 100 150

Obs. number

lm(bwt ~ smoke + ht + race)

Figure 11: Diagnostic plot after a linear regression.

Besides basic assumptions of the linear regression, i.e. normal distribution 

of the residuals, independence of the observations, homoskedasticity, it is im- portant to remember that we further assume the linearity of the relationship between the response and each of the predictors.

Conditional means can be obtained with the aggregate command,

and plotted using a trellis display. The effects package is dedicated to the post-processing of glm fits: It allows to conveniently display mean predictions with confidence intervals, specific effects (like interac-

tions or contrasts).11 11 J Fox. Effect displays for gener-

Here, we have omitted sds, but they should be added to get an idea of the birth weight variability in each partition of the dataset.

alized linear models. Sociological Methodology, 17:347–361, 1987; and J Fox. Effect displays in r for gen- eralised linear models. Journal of Statistical Software, 8(15), 2003

> all.means <- aggregate(bwt ~ smoke + ht + race,

No

Yes

+ data=birthwt, FUN=mean)

> print(xyplot(bwt ~ smoke | race, data=all.means, groups=ht,

+ layout=c(3,1), type=c("p","l"), auto.key=TRUE))

3500

3000

White

No Yes

Black

Other

Using a regression approach also allows to include continuous pre- dictor, like age.

bwt

2500

2000

> fm <- bwt ~ (smoke + ht) \* scale(age, scale=F) + race

> lm.fit6 <- lm(fm, data=birthwt)

> summary(lm.fit6)

all:

lm(formula = fm, data = birthwt)

Residuals:

Min 1Q Median 3Q Max

-2383.76 -452.19 41.56 466.87 1356.28

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 3307.17 94.00 35.183 < 2e-16

smokeYes -406.02 108.26 -3.750 0.000238

htYes -419.89 203.76 -2.061 0.040757

scale(age, scale = F) 16.94 12.30 1.377 0.170174

raceBlack -323.07 159.12 -2.030 0.043786

raceOther -387.79 119.14 -3.255 0.001354

smokeYes:scale(age, scale = F) -27.17 20.21 -1.344 0.180598

htYes:scale(age, scale = F) -91.00 48.34 -1.883 0.061367

Residual standard error: 677.8 on 181 degrees of freedom

Multiple R-squared: 0.1681, Adjusted R-squared: 0.136

F-statistic: 5.227 on 7 and 181 DF, p-value: 1.899e-05