Missing data

Based on a book by Julian J. Faraway

University of Iceland

Missing data

- Missing data is the situation where some values of some cases are missing. This is not uncommon.
- Dealing with missing data is time consuming - fixing up problems caused by missing data sometimes takes longer than the analysis itself.
- What can be done? Obviously, finding the missing values is the best option but this is not always possible. Next ask why the data are missing. If the reason for a datum being missing is non-informative, then a fix is easier.

Missing data

What can be done?

- Delete the case with missing observations. This is OK if this only causes the loss of a relatively small number of cases. This is the simplest solution.
- Fill-in or *impute* the missing values. Use the rest of the data to predict the missing values.
  - Simply replacing the missing value of a predictor with the average value of that predictor is one easy method.
  - Using regression on the other predictors is another possibility. Its not clear how much the diagnostics and inference on the filled-in dataset is affected.
  - Some additional uncertainty is caused by the imputation which needs to be allowed for.

Homeprice data

```
dat<-read.table("homeprices.csv",header=T,na.strings="*")
```

```
dim(dat)
```

```
## [1] 117   8
```

```
summary(dat)
```

```
##      PRICE           SQFT           AGE            FEATS
##  Min.   : 540   Min.   : 837   Min.   : 1.00   Min.   :0.00
##  1st Qu.: 780   1st Qu.:1280   1st Qu.: 5.75   1st Qu.:3.00
##  Median : 960   Median :1549   Median :13.00   Median :4.00
##  Mean   :1063   Mean   :1654   Mean   :14.97   Mean   :3.53
##  3rd Qu.:1200   3rd Qu.:1894   3rd Qu.:19.25   3rd Qu.:4.00
##  Max.   :2150   Max.   :3750   Max.   :53.00   Max.   :8.00
##                                NA's   :49
##        NE              CUST             COR             TAX
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.000   Min.   : 223.0
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.000   1st Qu.: 600.0
##  Median :1.0000   Median :0.0000   Median :0.000   Median : 731.0
##  Mean   :0.6667   Mean   :0.2308   Mean   :0.188   Mean   : 793.5
##  3rd Qu.:1.0000   3rd Qu.:0.0000   3rd Qu.:0.000   3rd Qu.: 919.0
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.000   Max.   :1765.0
##                                                    NA's   :10
```

```r
fit<-lm(PRICE~SQFT+AGE+FEATS+NE+CUST+COR,data=dat)
summary(fit)
```

```
##
## Call:
## lm(formula = PRICE ~ SQFT + AGE + FEATS + NE + CUST + COR, data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -505.96  -78.52   13.35   97.96  603.35
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  83.14037  102.73572   0.809  0.42151
## SQFT          0.63719    0.05119  12.448  < 2e-16 ***
## AGE          -3.72095    1.80540  -2.061  0.04357 *
## FEATS         3.25714   18.93246   0.172  0.86398
## NE          -14.32888   49.23057  -0.291  0.77200
## CUST        148.47950   54.40590   2.729  0.00829 **
## COR         -83.39862   51.26812  -1.627  0.10895
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 174.4 on 61 degrees of freedom
##   (49 observations deleted due to missingness)
## Multiple R-squared:  0.8295, Adjusted R-squared:  0.8128
## F-statistic: 49.47 on 6 and 61 DF,  p-value: < 2.2e-16
```

```r
length(fit$fitted.values)
```

```
## [1] 68
```

Homeprice data

```
dat$AGE[is.na(dat$AGE)]<-mean(dat$AGE,na.rm=T)
summary(dat)
```

```
##      PRICE           SQFT           AGE            FEATS
##  Min.   : 540   Min.   : 837   Min.   : 1.00   Min.   :0.00
##  1st Qu.: 780   1st Qu.:1280   1st Qu.: 7.00   1st Qu.:3.00
##  Median : 960   Median :1549   Median :14.97   Median :4.00
##  Mean   :1063   Mean   :1654   Mean   :14.97   Mean   :3.53
##  3rd Qu.:1200   3rd Qu.:1894   3rd Qu.:15.00   3rd Qu.:4.00
##  Max.   :2150   Max.   :3750   Max.   :53.00   Max.   :8.00
##
##        NE              CUST             COR             TAX
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.000   Min.   : 223.0
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.000   1st Qu.: 600.0
##  Median :1.0000   Median :0.0000   Median :0.000   Median : 731.0
##  Mean   :0.6667   Mean   :0.2308   Mean   :0.188   Mean   : 793.5
##  3rd Qu.:1.0000   3rd Qu.:0.0000   3rd Qu.:0.000   3rd Qu.: 919.0
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.000   Max.   :1765.0
##                                                    NA's   :10
```

```r
fit<-lm(PRICE~SQFT+AGE+FEATS+NE+CUST+COR,data=dat)
summary(fit)
```

```
##
## Call:
## lm(formula = PRICE ~ SQFT + AGE + FEATS + NE + CUST + COR, data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -896.31  -90.64   -0.58   82.84  747.79
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 138.23279   75.76579   1.824  0.07079 .
## SQFT          0.53048    0.04215  12.587  < 2e-16 ***
## AGE          -4.10208    1.91307  -2.144  0.03421 *
## FEATS        18.01453   14.07588   1.280  0.20330
## NE           43.51708   39.31952   1.107  0.27081
## CUST        139.98651   49.14450   2.848  0.00524 **
## COR         -86.86030   45.77427  -1.898  0.06037 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 190.7 on 110 degrees of freedom
## Multiple R-squared:  0.7617, Adjusted R-squared:  0.7487
## F-statistic: 58.59 on 6 and 110 DF,  p-value: < 2.2e-16
```

```r
length(fit$fitted.values)
```

```
## [1] 117
```