Applied Statistical Regression
HS 2010 – Week 08

Marcel Dettling

Institute for Data Analysis and Process Design

Zurich University of Applied Sciences

marcel.dettling@zhaw.ch

http://stat.ethz.ch/~dettling

ETH Zürich, November 15, 2010

Missing Data

The best thing to do is certainly to go and find the missing values. Often, this is impractical or impossible. Thus, …

→ **ask the question WHY the data are missing?**

- Just ***randomly***, non-informatively for the analysis goal. Fixing up missing data is comparatively easy.

- ***Systematically*** with respect to the goal of the analysis. Example: patients who dropped out of a drug study because they believed their treatment was not working.

Case 1 is tractable, case 2 is notoriously difficult!

Fix-Up Alternatives

If the missing are non-systematic, we can do the following:

1) Omitting incomplete cases
→ OK if only a small proportion of cases is incomplete

2) Filling-in missing data with the mean
→ quick and easy, but not always very accurate

3) Filling-in missing data by regression
→ regress a predictor on the other predictors

4) Sophisticated approaches, EM-algorithm
→ treating the missing values as nuisance parameters

Experimentation Setup

- *State dataset from last week:*

  - Life.Exp ~ Murder + Frost + HS.Grad + Pop

- *Random deletion of some five observations:*

  - Murder (2 NA introduced)
  - Frost (3 NA introduced)

- This is more interesting than to work with a dataset with true missings: *we can study the influence of different imputation methods*.

Applied Statistical Regression
HS 2010 – Week 08
Example: Plain R fit

```
> summary(lm(Life.Exp ~ Population + Murder + HS.Grad + Frost, state)

Estimate Std. Error t value Pr(>|t|)
(Intercept) 68.43923    1.91211  35.793  < 2e-16 ***
Population   0.31831    0.11248   2.830 0.007247 **
Murder      -1.43049    0.17821  -8.027 7.26e-10 ***
HS.Grad      5.75964    1.45363   3.962 0.000298 ***
Frost       -0.10537    0.03838  -2.746 0.009006 **

Residual standard error: 0.6824 on 40 degrees of freedom
  (5 observations deleted due to missingness)
Multiple R-squared: 0.7515, Adjusted R-squared: 0.7266
F-statistic: 30.24 on 4 and 40 DF,  p-value: 1.293e-11
```

Filling-in Missing Data with the Mean

The 3 missing data points in variable Frost are replaced by the overall mean value in this variable

```
> missings    <- which(is.na(state.trsf$Frost))
> mean.Frost <- mean(state.trsf$Frost, na.rm=TRUE)
> state.trsf$Frost[missings] <- mean.Frost
```

- The replacement value is 9.85, when the removed ones were 0, 10.68 and 13.19 for Hawaii, Kansas and New Hampshire.
- Apply strategy 2) only in problems where there are many predictors and in only few, data are missing – then it’s OK to profit from the information which in the other predictors.

Applied Statistical Regression
HS 2010 – Week 08

Results from Strategy 2)

```
> Coefficients: Estimate Std. Error t value Pr(>|t|)

(Intercept)      66.80292    1.98216  33.702  < 2e-16 ***

Population        0.36425    0.12058   3.021 0.004233 **

Murder           -1.34124    0.18860  -7.112 8.87e-09 ***

Frost            -0.03007    0.04800  -0.626 0.534333

HS.Grad           6.15488    1.50475   4.090 0.000185 ***

---

Residual standard error: 0.7298 on 43 degrees of freedom

  (2 observations deleted due to missingness)

Multiple R-squared: 0.7288, Adjusted R-squared: 0.7036

F-statistic: 28.89 on 4 and 43 DF,  p-value: 1.092e-11
```

Filling-in Missing Data with Regression

Predict the missing observations in Frost from a regression of the form: ***Frost ~ Population + Murder + HS.Grad***:

```
missing <- which(is.na(state.trsf$Frost))
fit.imp <- lm(Frost~Population+Murder+HS.Grad, state.trsf)
predval <- predict(fit.imp, newdata=state.trsf[missing,])
state.trsf$Frost[missing] <- pred.val
```

→ **Needs collinear predictors, doubtful here!**

```
> pred.val
      HI       KS       NH
11.43693 11.00075 12.27640
```

Applied Statistical Regression
HS 2010 – Week 08

Results from Strategy 3)

```
Coefficients: Estimate Std. Error t value Pr(>|t|)

(Intercept) 66.57107    2.00466  33.208  < 2e-16 ***
Population   0.37502    0.12243   3.063 0.003771 **
Murder      -1.32308    0.19082  -6.934 1.60e-08 ***
Frost       -0.01595    0.04908  -0.325 0.746796
HS.Grad      6.10990    1.51291   4.039 0.000218 ***
---
Residual standard error: 0.7322 on 43 degrees of freedom
  (2 observations deleted due to missingness)
Multiple R-squared: 0.727, Adjusted R-squared: 0.7016
F-statistic: 28.62 on 4 and 43 DF, p-value: 1.256e-11
```

Applied Statistical Regression
HS 2010 – Week 08

Synopsis

- It is not so simple *regenerate* and *impute* missing information
- While the mean or regression *fill-in methods* may provide an advantage, they are often *useless* or even *make things worse*
- Their *success* depends on the collinearity of the predictors – imputed values are better with more *collinear predictors*
- Both *fill-in* techniques will introduce a *bias towards zero* in the regression coefficients while tending to *reduce the variance*.
- When a *substantial proportion of the data is missing*,1-3) tend not to work well. Use *more sophisticated approaches* then!

Applied Statistical Regression
HS 2010 – Week 08

Modeling Strategies

- In which order to apply: estimation – diagnostics – transformation – variable selection???

There is no definite answer to this: regression analysis is the search for structure in the data and there are no hard-and-fast rules about how it should be done.

Professional regression analysis can be seen as an art and definitely requires skill an expertise – one must be alert to unexpected structure in the data.

→ We here provide a rough guideline for regression analysis

Applied Statistical Regression
HS 2010 – Week 08

Guideline for Regression analysis

0) Preprocessing the data

- learning the meaning of all variables
- give short and informative names
- check for impossible values, errors
- if they exist: set them to NA
- systematic or random missings?

1) First-aid transformations

- bring all variables to a suitable scale
- use statistical and specific knowledge
- routinely apply the first-aid transformations

Applied Statistical Regression
HS 2010 – Week 08

Guideline for Regression analysis

2) Fitting a big model

First fit a big model with potentially too many predictors
- use all if $p < n/5$
- preselect manually according to previous knowledge
- preselect with forward search and a p-value of 0.2

3) Model Diagnostics

Check for normality, constant variance, uncorrelated errors:
- transformations
- robust regression
- weighted regression
- dealing with correlation

Applied Statistical Regression
HS 2010 – Week 08

Guideline for Regression analysis

6) Interactions
- try (two-way) interactions
- do only use predictors that are in the model

7) Influential data points
- attractors for the regression line
- keep them or skip them?
- compare with and without

8) Do model and coefficients make sense?
- implausible predictors, wrong signs, against theory, …
- remove if there are no drastic changes!

Guideline for Regression analysis

If there were substantial changes to the model in steps 4-8), then one should go back to 3) and repeat the diagnostics.

Hypothesis testing:

- proceed similarly
- careful: transformations, selection, collinearity
- question dictates what works and what not!

Prediction:

- guideline is still reasonable
- we are a little less picky here in selection and diagnostics
- check generalization error with test data / cross validation

Significance vs. Relevance

The larger a sample, the smaller the p-values for the very same predictor effect. Thus do not confuse a small p-values with an important predictor effect!!!

With large datasets:

- statistically significant results which are practically useless
- we have high evidence that a blood value is lowered by 0.1%

Models are approximative:

- most predictors have influence, thus $\beta_1 = 0$ never holds
- point null hypothesis is usually wrong in practice
- we just need enough data to be able to reject it