Data Cleaning, Data Visualization, & Functions

Fundamental Techniques in Data Science

Utrecht University

Kyle M. Lang

Department of Methodology & Statistics
Utrecht University

Outline

Data Analytic Lifecycle

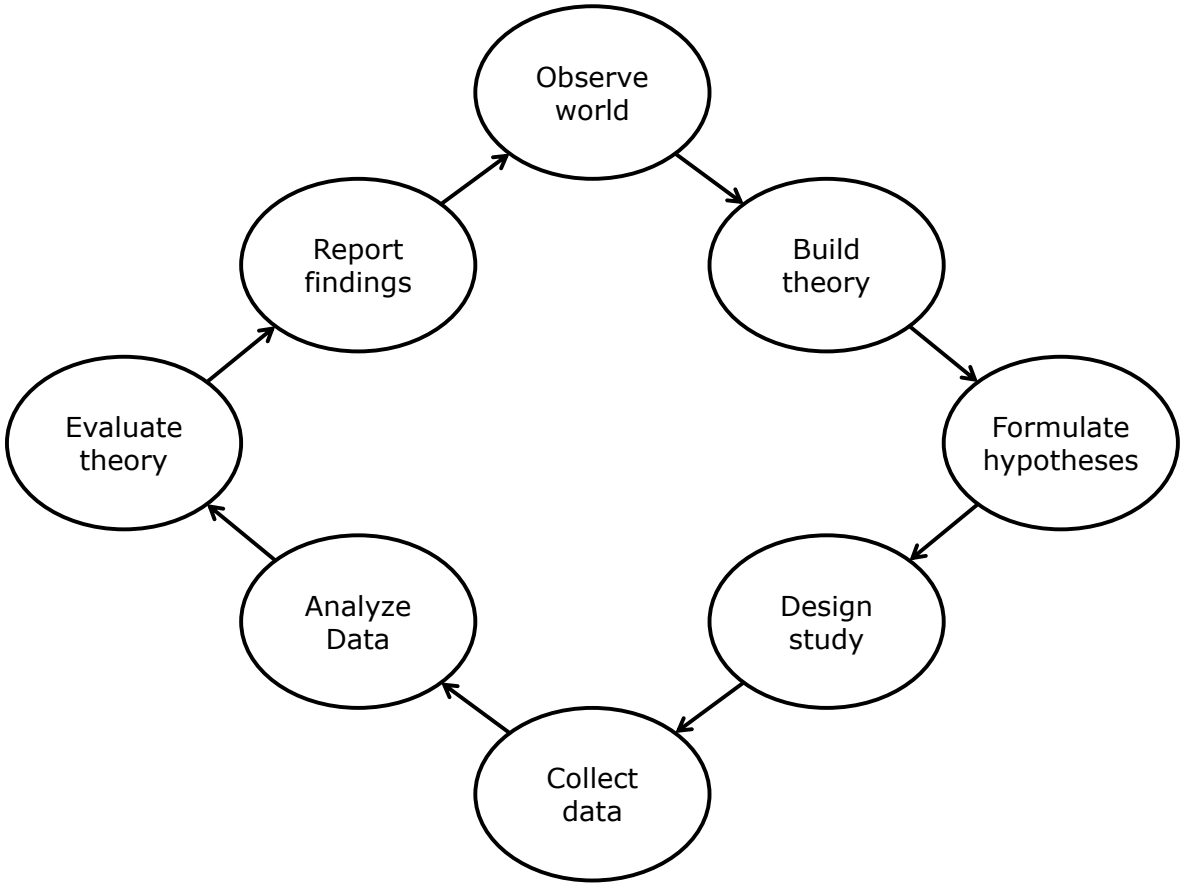
Data Cleaning

- Missing Data
- Outliers

Data Analytic Lifecycle

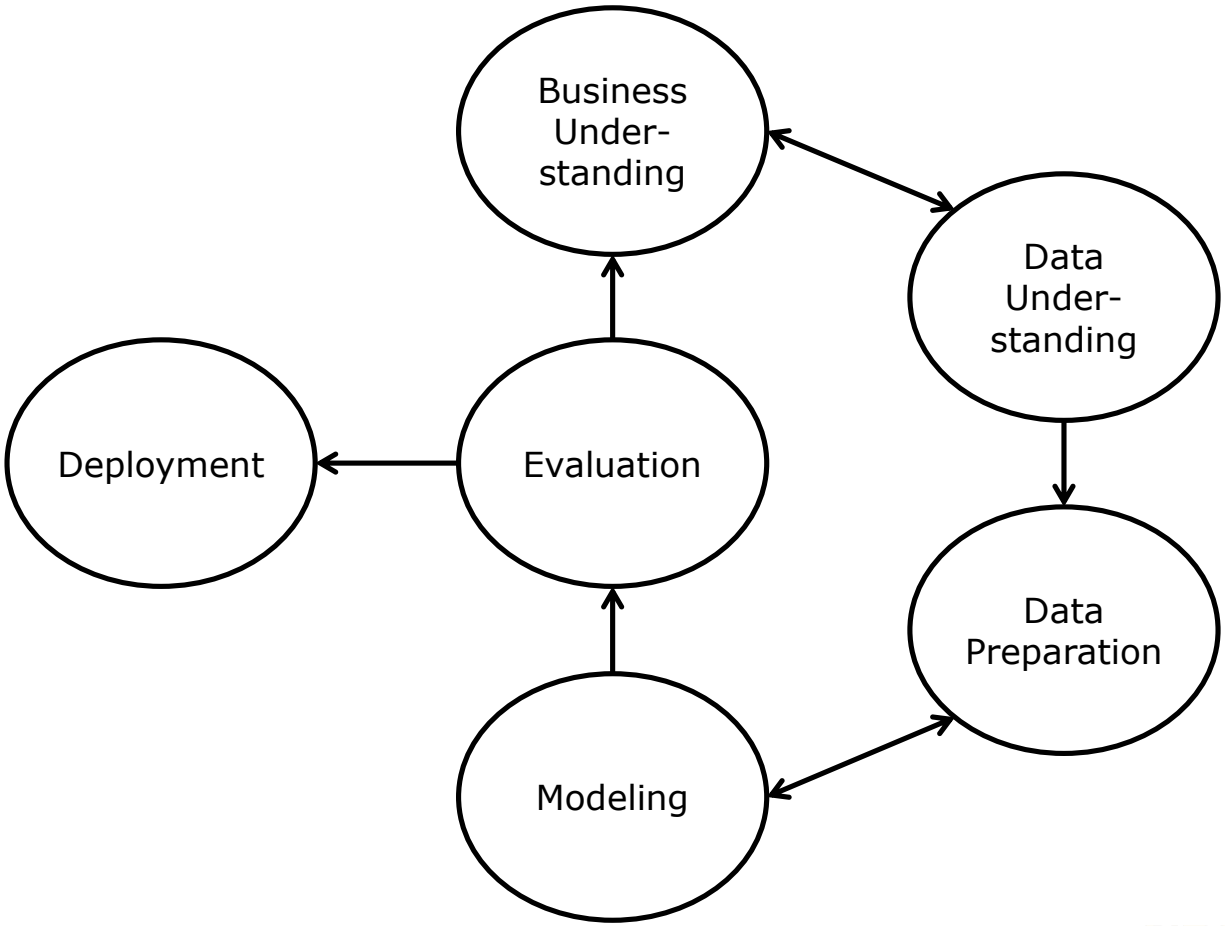
Research Cycle

The following is a representation of the *Research Cycle* used for empirical research in most of the sciences.



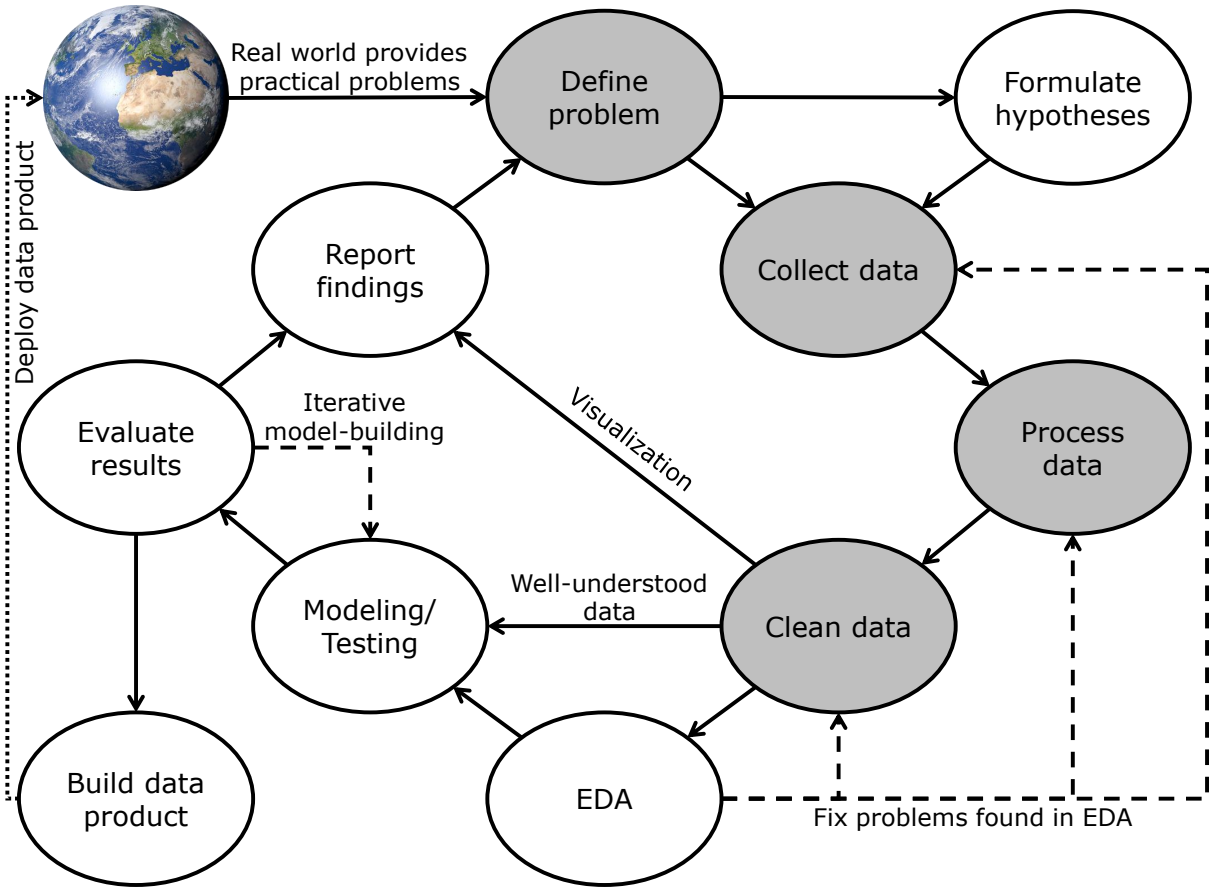
CRISP-DM

The *Cross-industry Standard Process for Data Mining* was developed to standardized the process of data mining in industry applications.



Data Science Cycle

The *Data Science Cycle* represented here was adapted from O'Neil and Schutt (2014).



Data Cleaning

Data Cleaning

When we receive new data, they are generally messy and contaminated by various anomalies and errors.

- One of the first steps in processing a new set of data is *cleaning*.
- By cleaning the data, we ensure a few properties:
  - The data are in an analyzable format.
  - All data take legal values.
  - Any outliers are located and treated.
  - Any missing data are located and treated.

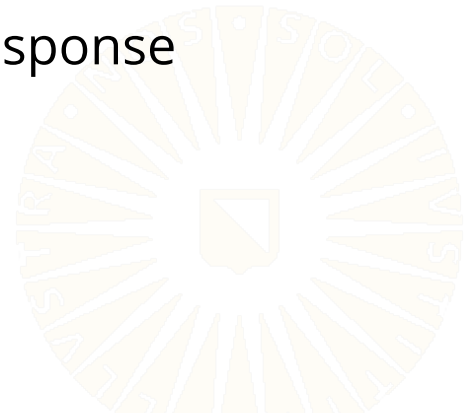
A Little Notation

Y := An $N \times P$ data matrix

Y_{mis} := The *missing* part of Y

Y_{obs} := The *observed* part of Y

R := An $N \times P$ pattern matrix encoding nonresponse



What are Missing Data?

Missing data are empty cells in a dataset where there should be observed values.

- The missing cells correspond to true population values, but we haven't observed those values.

What are Missing Data?

Missing data are empty cells in a dataset where there should be observed values.

- The missing cells correspond to true population values, but we haven't observed those values.

Not every empty cell is a missing datum.

- Quality-of-life ratings for dead patients in a mortality study
- Firm profitability after the company goes out of business
- Self-reported severity of menstrual cramping for men
- Empty blocks of data following "gateway" items

Missing Data Descriptives

Missing Data Pattern

Missing data (or response) patterns represent unique combinations of observed and missing items.

- P items $\Rightarrow 2^P$ possible patterns.

	X	Y
1	x	y
2	x	.
3	.	y
4	.	.

Patterns for $P = 2$

	X	Y	Z
1	x	y	z
2	x	y	.
3	x	.	z
4	.	y	z
5	x	.	.
6	.	.	z
7	.	y	.
8	.	.	.

Patterns for $P = 3$

Nonresponse Rates

Percent/Proportion Missing

- The proportion of cells containing missing data
- Good early screening measure
- Should be computed for each variable, not for the entire dataset

Attrition Rate

- The proportion of participants that drop-out of a study at each measurement occasion

Percent/Proportion of Complete Cases

- The proportion of observations with no missing data
- Often reported but nearly useless quantity

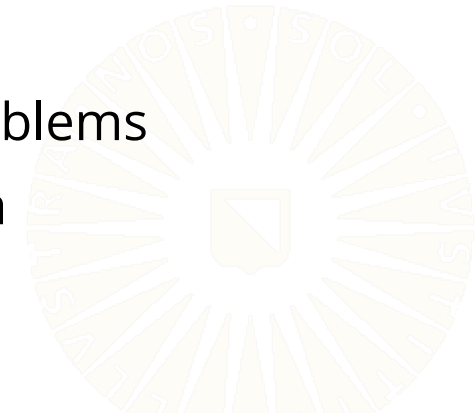
Nonresponse Rates

Covariance Coverage

- The proportion of cases available to estimate a given pairwise relationship (e.g., a covariance between two variables)
- Very important to have adequate coverage for the parameters you want to estimate

Fraction of Missing Information

- Associated with an estimated parameter, not with an incomplete variable
- Like an R^2 for the missing data
- Most important diagnostic value for missing data problems
- Can only be computed after treating the missing data



Covariance Coverage Examples

- What is the coverage for $cov(X, Y)$?
- What is the coverage for $cov(W, Y)$?
- What about $cov(X, Z)$?

	W	X	Y	Z
1	w	x	y	.
2	w	x	y	.
3	w	x	y	.
4	w	x	y	.
5	w	x	y	.
6	w	.	y	z
7	w	.	y	z
8	w	.	y	z
9	w	.	y	z
10	w	.	y	z

Nonresponse Rate Examples

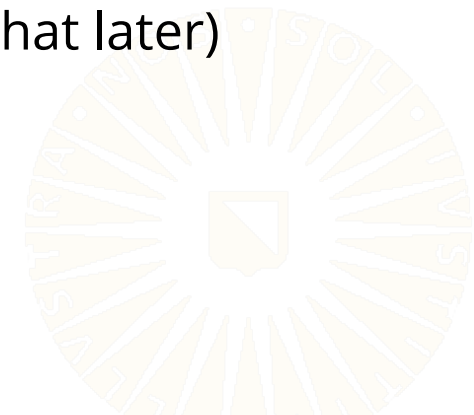
- What is the percent missing at Time 2?
- What is the attrition rate at Time 3?

	T1	T2	T3	T4
1	x1	x2	x3	x4
2	x1	x2	x3	x4
3	x1	x2	x3	x4
4	x1	x2	x3	.
5	x1	x2	x3	.
6	x1	x2	.	.
7	x1	x2	.	.
8	x1	.	.	.
9	x1	.	.	.
10	x1	.	.	.

What is an outlier?

For the time being, we're considering *univariate outliers*.

- Extreme values with respect to the distribution of a variable's other observations
  - A human height measurement of 3 meters
  - A high temperature in Tilburg of $50°$
  - Annual income of €250,000 for a student
- Not accounting for any particular model (we'll get to that later)



What is an outlier?

A univariate outlier may, or may not, be an illegal value.

- Data entry errors are probably the most common cause.
- Outliers can also be legal, but extreme, values.

Key Point: We choose to view an outlier as arising from a different population than the one to which we want to generalize our findings.

Finding Univariate Outliers

We have many methods available to diagnose potential outliers.

- Four of the simplest and most popular are:
  1. Internally studentized residuals (AKA Z-score method)
  2. Externally studentized residuals
  3. Median absolute deviation method
  4. Tukey's boxplot method

Internally Studentized Residuals

For each observation, X_n, we compute the following quantity:

$$T_n = \frac{X_n - \overline{X}}{SD_X}$$

- T_n follows a Student's t distribution with $df = N - 1$.
  - We can do a formal test for "outlier" status.
- Assuming a large sample, if $T_n > C$ (where C is usually 2 or 3), we label X_n as an outlier.

Internally Studentized Residuals

For each observation, X_n, we compute the following quantity:

$$T_n = \frac{X_n - \overline{X}}{SD_X}$$

- T_n follows a Student's t distribution with $df = N - 1$.
  - We can do a formal test for "outlier" status.
- Assuming a large sample, if $T_n > C$ (where C is usually 2 or 3), we label X_n as an outlier.

Although simple, this method has some substantial limitations.

- The cutpoint, C, can only be meaningfully chosen when X is normally distributed.
- Both $\overline{X}$ and SD_X are highly sensitive to outliers.

Externally Studentized Residual

The externally studentized residual method is essentially the same as the internally studentized residual method, but we adjust $\bar{X}$ and SD_X to remove the influence of the observation we're evaluating.

- Let $\mathbb{N}_{(n)} = \{1, \ldots, (n-1), (n+1), \ldots, N\}$.
- Define the deletion mean, $\bar{X}_{(n)}$, and deletion SD, $SD_{X(n)}$, as:

$$\bar{X}_{(n)} = \frac{1}{N-1} \sum_{i \in \mathbb{N}_{(n)}} X_i$$

$$SD_{X(n)} = \sqrt{\frac{1}{N-2} \sum_{i \in \mathbb{N}_{(n)}} \left(X_i - \bar{X}_{(n)} \right)^2}$$

Externally Studentized Residual

The externally studentized residual is defined in the same way as the internally studentized version:

$$T_{(n)} = \frac{X_n - \bar{X}_{(n)}}{SD_{X(n)}}$$

- $T_{(n)}$ follows a Student's t distribution with $df = N - 2$.
  - We can do a formal test for "outlier" status.
- Assuming a large sample, if $T_{(n)} > C$ (where C is usually 2 or 3), we label X_n as an outlier.

Externally Studentized Residual

The externally studentized residual is defined in the same way as the internally studentized version:

$$T_{(n)} = \frac{X_n - \bar{X}_{(n)}}{SD_{X(n)}}$$

- $T_{(n)}$ follows a Student's t distribution with $df = N - 2$.
  - We can do a formal test for "outlier" status.
- Assuming a large sample, if $T_{(n)} > C$ (where C is usually 2 or 3), we label X_n as an outlier.

$T_{(n)}$ is immune to the influence of the nth observation.

- Still requires X to be normally distributed
- Still sensitive to outliers other than the nth observation

Median Absolute Deviation Method

The biggest limitation of studentized residuals is that their measures of central tendency and dispersion are sensitive to outliers.

- If we can replace the (deletion) mean and the (deletion) SD with more robust statistics, we can avoid this issue.
  - Replace the mean, $\bar{X}$, with the *median*, $\text{Med}(X)$
  - Replace the SD with the *median absolute deviation*:

$$MAD_X = b \times \text{Med}\left(\left|X_n - \text{Med}(X)\right|\right)$$

  - We choose the coefficient as $b = 1/Q_{0.75}$
  - For the normal distribution, $b \approx 1/0.6745 \approx 1.4826$

Median Absolute Deviation Method

We compute our test statistic by replacing the mean with the median and the SD with the MAD in the standard Wald test formula:

$$T_{MAD} = \frac{X_n - \text{Med}(X)}{MAD_X}$$

- T_{MAD} doesn't allow for formal statistical tests.
- We can use the same general cutoffs we would use for the studentized residual methods.
  - Assuming a large sample, if $T_{(n)} > C$ (where C is usually 2 or 3), we label X_n as an outlier.

Median Absolute Deviation Method

We compute our test statistic by replacing the mean with the median and the SD with the MAD in the standard Wald test formula:

$$T_{MAD} = \frac{X_n - \text{Med}(X)}{MAD_X}$$

- T_{MAD} doesn't allow for formal statistical tests.
- We can use the same general cutoffs we would use for the studentized residual methods.
  - Assuming a large sample, if $T_{(n)} > C$ (where C is usually 2 or 3), we label X_n as an outlier.

T_{MAD} is immune to the influence of, up to, 50% outlying observations.

- Requires us to assume a parametric distribution for X
  - This assumption is necessary to compute b.

Breakdown Point

To compare robust statistics, we consider their *breakdown points*.

- The breakdown point is the minimum proportion of cases that must be replaced by ∞ to cause the value of the statistic to go to ∞.

The mean has a breakdown point of $1/N$.

- Replacing a single value with ∞ will produce an infinite mean.

The deletion mean has a breakdown point of $2/N$.

- We can replace, at most, 1 value with ∞ without producing an infinite mean.

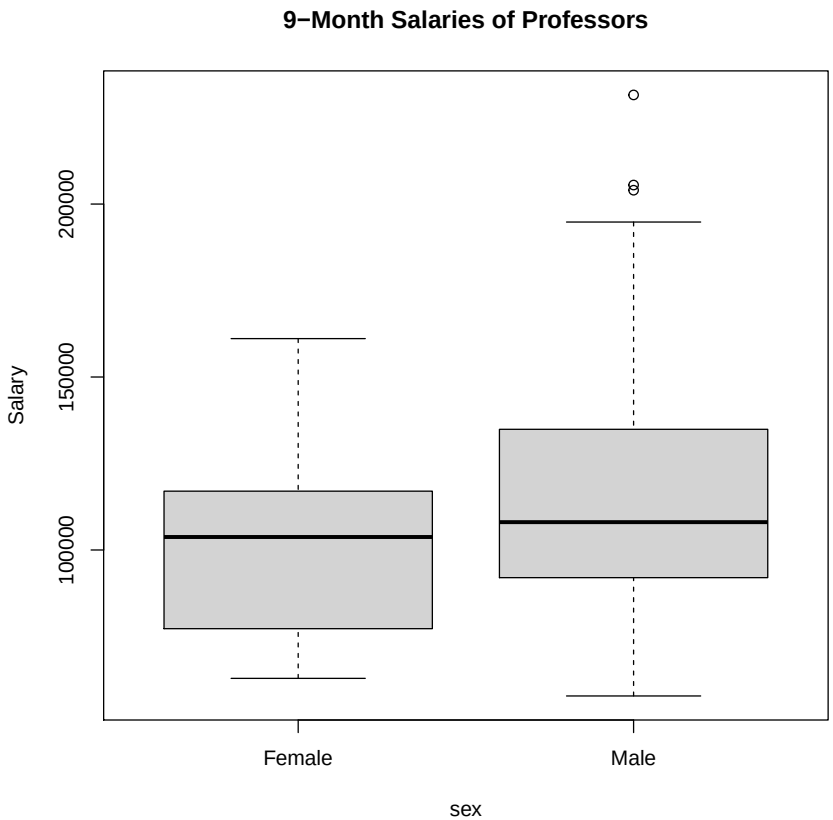
The median has breakdown point of 50%.

- We can replace $n < N/2$ of the observations with ∞ without producing an infinite median.

Boxplot Method

Tukey (1977) described a procedure for flagging potential outliers based on the familiar box-and-whiskers plot.

- Does not require normally distributed X
- Not sensitive to outliers
- Doesn't allow for formal statistical tests



Boxplot Method

A *fence* is an interval defined as the following function of the *first quartile*, the *third quartile*, and the *inner quartile range* ($IQR = Q_3 - Q_1$):

$$F = \{Q_1 - C \times IQR, Q_3 + C \times IQR\}$$

- Taking $C = 1.5$ produces the *inner fence*.
- Taking $C = 3.0$ produces the *outer fence*.

We can use these fences to identify potential outliers:

- Any value that falls outside of the inner fence is a *possible outlier*.
- Any value that falls outside of the outer fence is a *probable outlier*.

Multivariate Outliers

Sometimes, the combinations of values in an observation are very unlikely, even when no individual value is an outlier.

- These observations are *multivariate outliers*.
  - A person in the 95*th* percentile for height and the 5*th* percentile for weight
  - A person who simultaneously scores highly on scales of depression and positive affect

To detect multivariate outliers, we use *distance metrics*.

- Distance metrics quantify the similarity of two vectors.
  - Similarity between two observations
  - Similarity between an observation and the mean vector

Mahalanobis Distance

One of the most common distance metrics is the *Mahalanobis Distance*.

- The Mahalanobis distance, Δ, is a multivariate generalization of the internally studentized residual:

$$\Delta_n = \sqrt{\left(\mathbf{x}_n - \hat{\mu}_{\mathbf{X}}\right) \hat{\Sigma}_{\mathbf{X}}^{-1} \left(\mathbf{x}_n - \hat{\mu}_{\mathbf{X}}\right)^T}$$

As with studentized residuals, if $\Delta_n > C$, we label $\mathbf{x}_n$ as an outlier.

- When $\mathbf{X}$ is K-variate normally distributed, Δ_n^2 follows a χ^2 distribution with $df = K$.
- We take C to be the square-root of a suitably conservative quantile (e.g., $q \in \{99\%, 99.9\%\}$) of the χ_K^2 distribution: $C = \sqrt{\chi_{K,q}^2}$.

Problems with Mahalanobis Distance

Like the internally studentized residual, Mahalanobis distance is highly sensitive to outliers.

- The underlying estimates of central tendency, $\hat{\mu}_{\mathbf{X}}$, and dispersion, $\hat{\Sigma}_{\mathbf{X}}$, are computed using all observations.

Problems with Mahalanobis Distance

Like the internally studentized residual, Mahalanobis distance is highly sensitive to outliers.

- The underlying estimates of central tendency, $\hat{\mu}_{\mathbf{X}}$, and dispersion, $\hat{\Sigma}_{\mathbf{X}}$, are computed using all observations.

We want robust analogues of $\hat{\mu}_{\mathbf{X}}$ and $\hat{\Sigma}_{\mathbf{X}}$.

- We have several options for robust estimation of $\hat{\mu}_{\mathbf{X}}$ and $\hat{\Sigma}_{\mathbf{X}}$. E.g.:
  - Minimum covariance determinant method (MCD; Rousseeuw, 1985)
  - Minimum volume ellipsoid method (MVE; Rousseeuw, 1985)
  - M-estimation (Maronna, 1976)
- Conceptually, robust methods operate by either:
  - Using only a "good" subset of data to estimate $\hat{\mu}_{\mathbf{X}}$ and $\hat{\Sigma}_{\mathbf{X}}$.
  - Downweighting outlying observations when estimating $\hat{\mu}_{\mathbf{X}}$ and $\hat{\Sigma}_{\mathbf{X}}$.

Robust Mahalanobis Distance

Equipped with robust estimates of central tendency, $\hat{\mu}_{R,\mathbf{X}}$, and dispersion, $\hat{\Sigma}_{R,\mathbf{X}}$, we define the robust Mahalanobis distance in the natural way:

$$\Delta_{R,n} = \sqrt{\left(\mathbf{x}_n - \hat{\mu}_{R,\mathbf{X}}\right) \hat{\Sigma}_{R,\mathbf{X}}^{-1} \left(\mathbf{x}_n - \hat{\mu}_{R,\mathbf{X}}\right)^T}$$

We use $\Delta_{R,n}$ in the same way as Δ_n.

- If $\Delta_{R,n} > C$, we label $\mathbf{x}_n$ as an outlier.
- Again, we take C to be the square-root of some quantile of the χ^2_K distribution: $C = \sqrt{\chi^2_{K,q}}$.

Practicalities: Univariate vs. Multivariate

Univariate outlier checks are safe for most variables.

Practicalities: Univariate vs. Multivariate

Univariate outlier checks are safe for most variables.

Don't include too many variables in multivariate outlier checks.

- More variables increases the chances of false positives.
- E.g., don't run a multivariate outlier test on your entire dataset.

Practicalities: Univariate vs. Multivariate

Univariate outlier checks are safe for most variables.

Don't include too many variables in multivariate outlier checks.

- More variables increases the chances of false positives.
- E.g., don't run a multivariate outlier test on your entire dataset.

Do use multivariate outlier checks for scales.

- E.g., if you have a psychometric scale measuring depression, you should check the items of that scale for multivariate outliers.

Practicalities: Univariate vs. Multivariate

Univariate outlier checks are safe for most variables.

Don't include too many variables in multivariate outlier checks.

- More variables increases the chances of false positives.
- E.g., don't run a multivariate outlier test on your entire dataset.

Do use multivariate outlier checks for scales.

- E.g., if you have a psychometric scale measuring depression, you should check the items of that scale for multivariate outliers.

Maybe check the variables in a single model for multivariate outliers.

- E.g., if you have a small set of items that you will include in a regression model, it could make sense to check these variables for multivariate outliers.

Practicalities: Outliers for Categorical Data

Nominal, ordinal, and binary items *can* have outliers.

- Outliers on categorical variables are often more indicative of bad variables than outlying cases.

Practicalities: Outliers for Categorical Data

Nominal, ordinal, and binary items *can* have outliers.

- Outliers on categorical variables are often more indicative of bad variables than outlying cases.

Ordinal

- Most participant endorse one of the lowest categories on an ordinal item, but a few participants endorse the highest category.
- The participants who endorse the highest category may be outliers.

Practicalities: Outliers for Categorical Data

Nominal, ordinal, and binary items *can* have outliers.

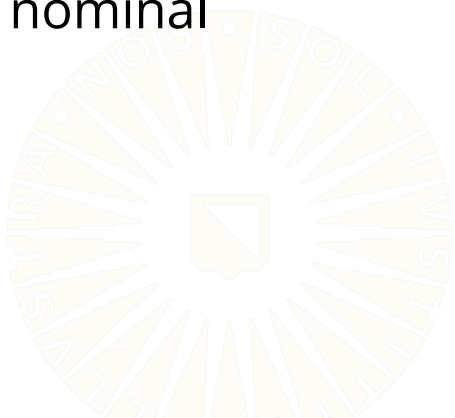
- Outliers on categorical variables are often more indicative of bad variables than outlying cases.

Ordinal

- Most participant endorse one of the lowest categories on an ordinal item, but a few participants endorse the highest category.
- The participants who endorse the highest category may be outliers.

Nominal

- Groups with very low membership may be outliers on nominal grouping variables.



Practicalities: Outliers for Categorical Data

Nominal, ordinal, and binary items *can* have outliers.

- Outliers on categorical variables are often more indicative of bad variables than outlying cases.

Ordinal

- Most participant endorse one of the lowest categories on an ordinal item, but a few participants endorse the highest category.
- The participants who endorse the highest category may be outliers.

Nominal

- Groups with very low membership may be outliers on nominal grouping variables.

Binary

- If most endorse the item, the few who do not may be outliers.

Treating Outliers

If we locate any outliers, they must be treated.

- Outliers cause by errors, mistakes, or malfunctions (i.e., *error outliers*) should be directly corrected.
- Labeling non-error outliers is a subjective task.
  - A (non-error) outlier must originate from a population separate from the one we care about.
  - Don’t blindly automate the decision process.

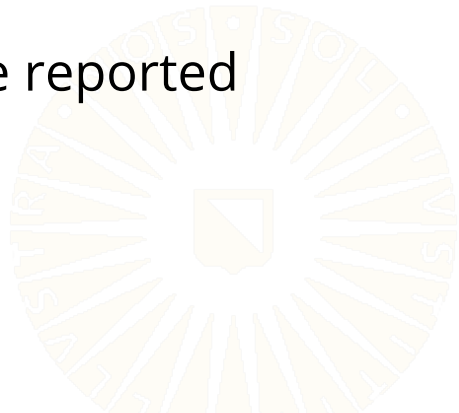
Treating Outliers

If we locate any outliers, they must be treated.

- Outliers cause by errors, mistakes, or malfunctions (i.e., *error outliers*) should be directly corrected.
- Labeling non-error outliers is a subjective task.
  - A (non-error) outlier must originate from a population separate from the one we care about.
  - Don't blindly automate the decision process.

The most direct solution is to delete any outlying observation.

- If you delete non-error outliers, the analysis should be reported twice: with outliers and without.



Treating Outliers

For univariate outliers, we can use less extreme types of deletion.

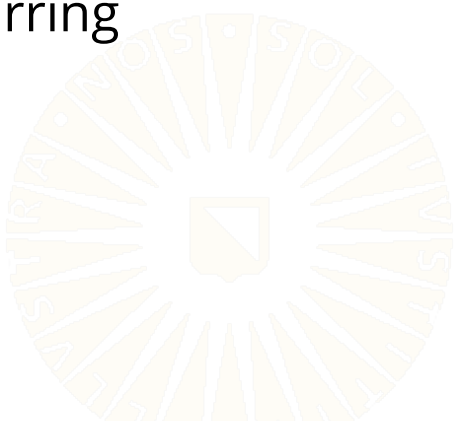
- Delete outlying values (but not the entire observation).
- These empty cells then become missing data.

Winsorization:

- Replace the missing values with the nearest non-outlying value.

Missing data analysis:

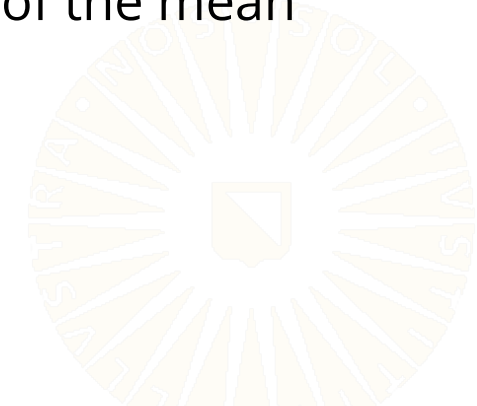
- Treat the missing values along with any naturally-occurring nonresponse.



Treating Outliers

We can also use robust regression procedures to estimate the model directly in the presence of outliers.

- Weight the objective function to reduce the impact of outliers
  - M-estimation
- Trim outlying observations during estimation
  - Least trimmed squares, MCD, MVE
- Take the median, instead of the mean, of the squared residuals
  - Least median of squares
- Model some quantile of the DV's distribution instead of the mean
  - Quantile regression
- Model the outcome with a heavy-tailed distribution
  - Laplacian, Student's T



References

Maronna, R. A. (1976). Robust m-estimators of multivariate location and scatter. *The Annals of Statistics*, *4*(1), 51–67. doi: 10.1214/aos/1176343347

O'Neil, C., & Schutt, R. (2014). *Doing data science: Straight talk from the frontline*. Sebastopol, CA: O'Reilly Media, Inc.

Rousseeuw, P. J. (1985). Multivariate estimation with high breakdown point. In W. Grossmann, G. Pflug, I. Vincze, & W. Wertz (Eds.), *Mathematical statistics and applications* (Vol. B, pp. 283–297). Netherlands: Reidel.

Tukey, J. W. (1977). *Exploratory data analysis* (Vol. 2). Reading, MA: Addison–Wesley.