Predictive Modeling

Chapter 3: Data Pre-processing

STA 6543

The University of Texas at San Antonio

What is data pre-processing?

- Data pre-processing generally refers to the addition, deletion, or transformation of the *training* data.
- Different predictive models usually have different sensitivities to the predictors.
- The need for data pre-processing is determined by the type of models being used.
  - Some models, such as *tree-based models*, are notably insensitive to the characteristics of the predictor data. Others, like *linear regression*, are not.
  - Care should be paid to check which, if any, pre-processing techniques can be useful.

A summary of models and some of their characteristics

Table A.1: A summary of models and some of their characteristics

Model	Allows $n < p$	Pre-processing	Interpretable	Automatic feature selection	# Tuning parameters	Robust to predictor noise	Computation time
Linear regression†	×	CS, NZV, Corr	✓	×	0	×	✓
Partial least squares	✓	CS	✓	○	1	×	✓
Ridge regression	×	CS, NZV	✓	×	1	×	✓
Elastic net/lasso	×	CS, NZV	✓	✓	1–2	×	✓
Neural networks	✓	CS, NZV, Corr	×	×	2	×	×
Support vector machines	✓	CS	×	×	1–3	×	×
MARS/FDA	✓		○	✓	1–2	○	○
K-nearest neighbors	✓	CS, NZV	×	×	1	○	✓
Single trees	✓		○	✓	1	✓	✓
Model trees/rules†	✓		○	✓	1–2	✓	✓
Bagged trees	✓		×	✓	0	✓	○
Random forest	✓		×	○	0–1	✓	×
Boosted trees	✓		×	✓	3	✓	×
Cubist†	✓		×	○	2	✓	×
Logistic regression*	×	CS, NZV, Corr	✓	×	0	×	✓
{LQRM}DA*	×	NZV	○	×	0–2	×	✓
Nearest shrunken centroids*	✓	NZV	○	✓	1	×	✓
Naïve Bayes*	✓	NZV	×	×	0–1	○	○
C5.0*	✓		○	✓	0–3	✓	×

†regression only *classification only

Symbols represent affirmative (✓), negative (×), and somewhere in between (○)

- Table A1 of the textbook.

CS: centering & scaling

NZV: near zero variance

Corr: correlation analysis.

Motivating example: cell segmentation in high-content screening

- Medical researchers often assess the cell characteristics of a living organism or plant to understand the effects of medicines or diseases on the size, shape, development status, and number of cells.
- There are two ways to do this:
  1) Experts can examine the target serum or tissue under a microscope and manually assess the desired cell characteristics. This work is tedious and requires expert knowledge of the cell type and characteristics.
  2) Another way to measure the cell characteristics from these kinds of samples is by using high-content screening.

High-content screening

- A sample is first dyed with a substance that will bind to the desired characteristic of the cells.
- The sample is then interrogated by an instrument (such as a confocal microscope),where the dye deflects light and the detectors quantify the degree of scattering for that specific wavelength.
- The light scattering measurements are then processed through imaging software to quantify the desired cell characteristics.
- Using an automated, high-throughput approach to assess samples' cell characteristics can *sometimes produce misleading results*.

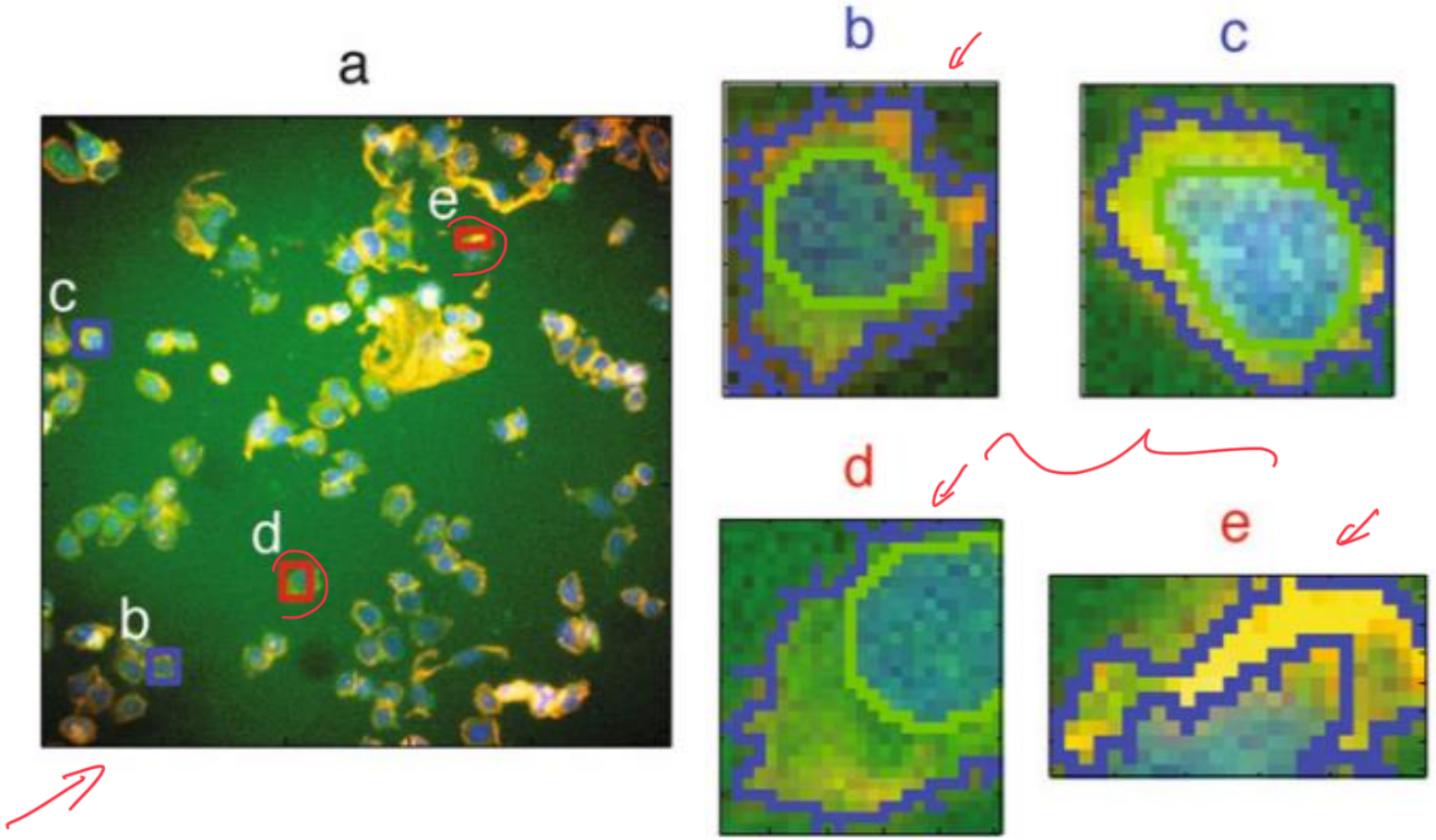


Fig. 3.1: An image showing cell segmentation from Hill et al. (2007). The *red boxes* [panels (**d**) and (**e**)] show poorly segmented cells while the cells in the *blue boxes* are examples of proper segmentation

(b) and (c) samples are good

How do we deal with these destroyed samples?

Data pre-processing techniques

- *Data transformations* (e.g., centering and scaling, resolving skewness, resolving outliers, data reduction, etc)
- *Dealing with missing values* (e.g., removal of missing sample, data imputations)
- *Removing predictors* (e.g., near-zero variance predictor, multicollinearity)
- *Creating dummy variables* (e.g., categorical variables)
- *Binning predictors* (e.g., binning a numerical predictor to two or more groups)

Age :

- ……

Centering and scaling

- The most straightforward and common data transformation is to center and scale the data.
- To center the data, the average variable value is subtracted from all the values to make the predictors have a zero mean.
- To scale the data, each value of the variable is divided by its standard deviation to make the predictors have a common standard deviation of one.

Centering and scaling

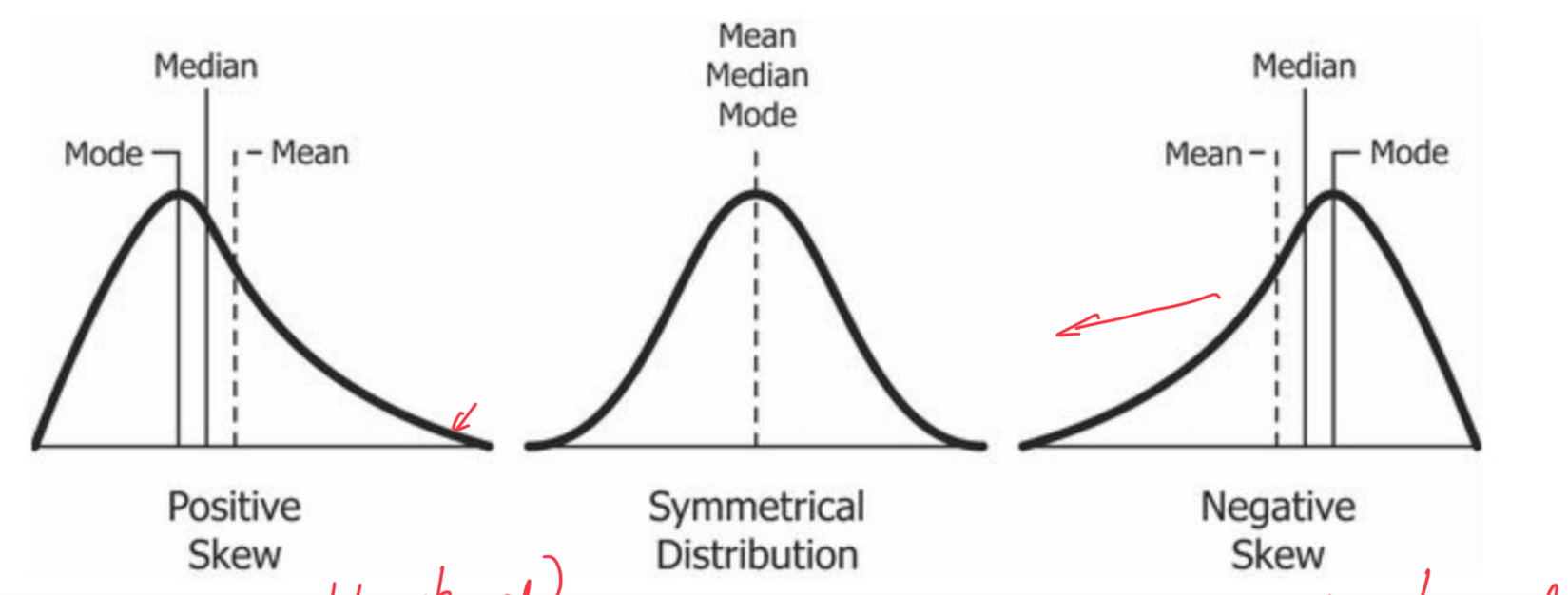
- Pros:
  - Improve the numerical stability of some calculations
  - Some methods, such as PCA or PLS benefit from the data being on a common scale
- Cons:
  - Loss of interpretability of the individual value.

In R Usage: scale(x, center = TRUE, scale = TRUE)

data FALSE

Transformation to resolve skewness

- Skewness is a measure of the asymmetry of the probability distribution of a real-valued random variable about its mean.



Transformation to resolve skewness

- If the distribution is roughly *symmetric*, the skewness values will be close to *zero*.
- If the distribution becomes more *right skewed*, the skewness statistic becomes *larger*.
- If the distribution becomes more *left skewed*, the value becomes *negative*.

max(x) / min(x) > 20

- The rule of thumb: skewed data whose ratio of the highest value to the lowest value is greater than 20 have significant skewness.

Transformation to resolve skewness

- Box-cox transformation

$$x^* = \begin{cases} \frac{x^\lambda - 1}{\lambda} & \text{if } \lambda \neq 0 \\ \log(x) & \text{if } \lambda = 0 \end{cases}$$

x: original data

λ: unknown

x^2

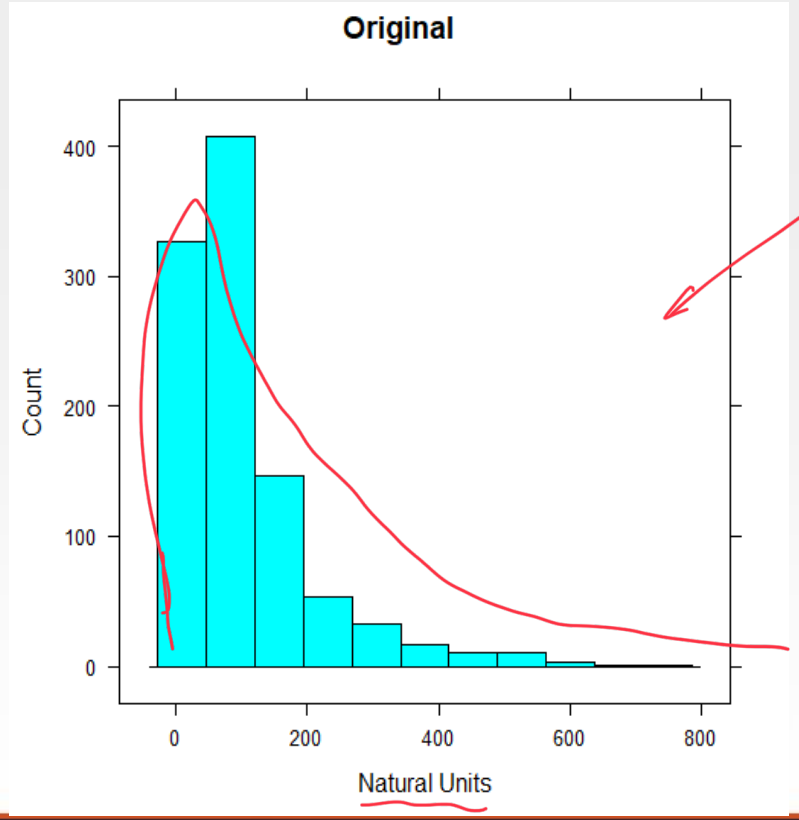
$\sqrt{x}$

- It includes square transformation ($\lambda = 2$), square root ($\lambda = 0.5$), and inverse ($\lambda = -1$). Often times, some approximations are implemented.

$\frac{1}{x}$

Motivating example: cell segmentation data

- The cell segmentation data contain a predictor that measures the standard deviation of the intensity of the pixels in the actin filaments.



right-skewed (heavily)

↑

Box-cox

Motivating example: cell segmentation data

```
library(AppliedPredictiveModeling)
data(segmentationOriginal)

## Retain the original training set
segTrain <- subset(segmentationOriginal, Case == "Train")

## Remove the first three columns (identifier columns)
segTrainX <- segTrain[, -(1:3)]
segTrainClass <- segTrain$Class #two levels: PS and WS
```

access

or training data

remove the first three columns.

Motivating example: cell segmentation data

```
#Rule of thumbs > 20
max(segTrainX$VarIntenCh3)/min(segTrainX$VarIntenCh3)

#calculate the skewness of a predictor
library(e1071)
skewness(segTrainX$VarIntenCh3)

#Box-cox transformation
library(caret)
BoxCoxTrans(segTrainX$VarIntenCh3)
```

Box Cox transformation for one predictor

```
> library(caret)
> BoxCoxTrans(segTrainX$VarIntenCh3)
Box-Cox Transformation

1009 data points used to estimate Lambda

Input data summary:
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
  0.8693  37.0615  68.1316 101.6718 124.9899 757.0210

Largest/Smallest: 871
Sample Skewness: 2.39

Estimated Lambda: 0.1
With fudge factor, Lambda = 0 will be used for transformations
```

λ = 0.1

Conclusion: we may let λ = 0 indicating a log-transformation.

Box Cox transformation for one predictor

```
## Apply the transformations for predictor VarIntenCh3
VarIntenCh3BoxCox <- BoxCoxTrans(segTrainX$VarIntenCh3)
VarIntenCh3Trans <- predict(VarIntenCh3BoxCox, segTrainX$VarIntenCh3)

#Histgram comparisons before and after transformation
histogram(segTrainX$VarIntenCh3, xlab = "Natural Units",type = "count",main="Original")
histogram(VarIntenCh3Trans, xlab = "Log Units",type = "count", main="Log-transformation")
```

new X : $X^* = \log(X)$ Since $\lambda = 0$

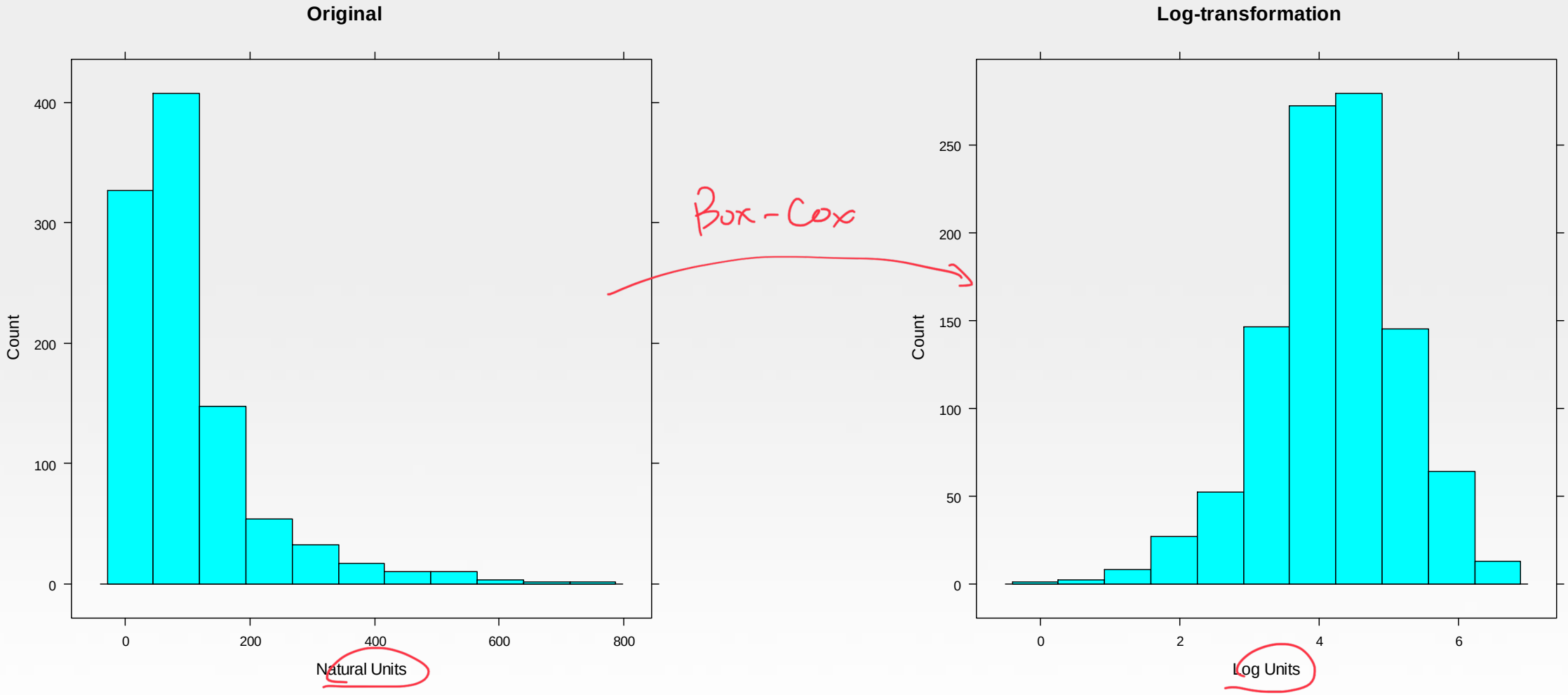
original x

original x

transformed X : X^*

↓ title

Histograms of before and after trans.



Box Cox transformation for another predictor PerimCh1

```
## Apply the transformations for predictor VarIntenCh3
PerimCh1BoxCox <- BoxCoxTrans(segTrainX$PerimCh1)
PerimCh1Trans <- predict(PerimCh1BoxCox, segTrainX$PerimCh1)

#Histgram comparisons before and after transformation
histogram(segTrainX$PerimCh1, xlab = "Natural Units",type = "count",main="Original")
histogram(PerimCh1Trans, xlab = “Log Units",type = "count", main="Log-transformation")
```

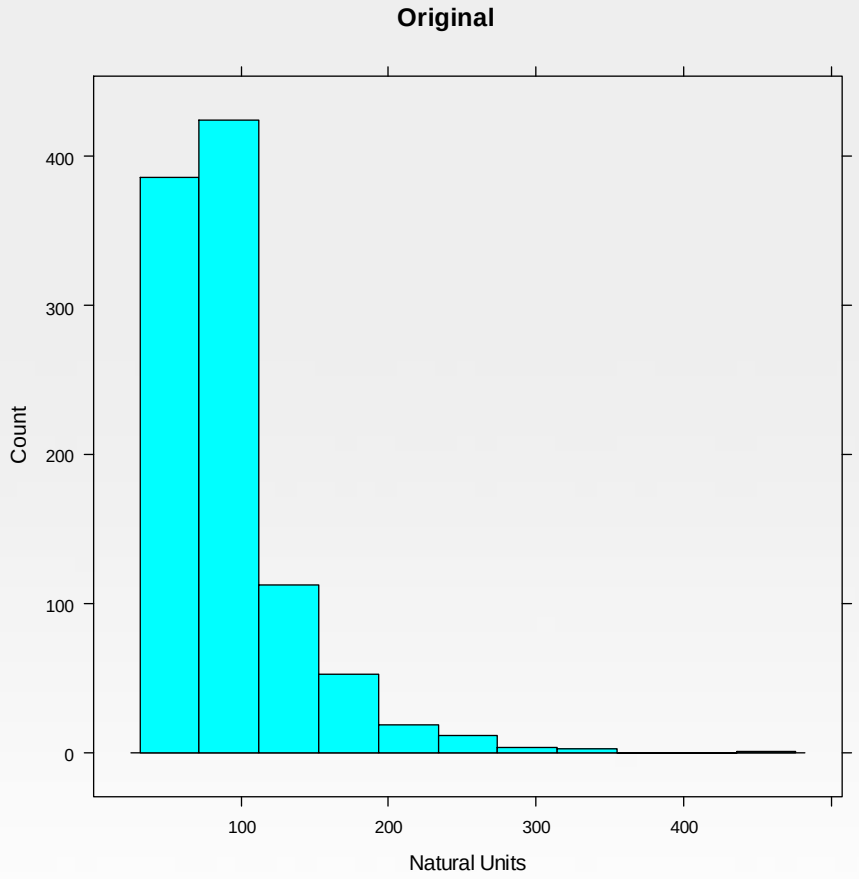
Box Cox transformation for another predictor PerimCh1

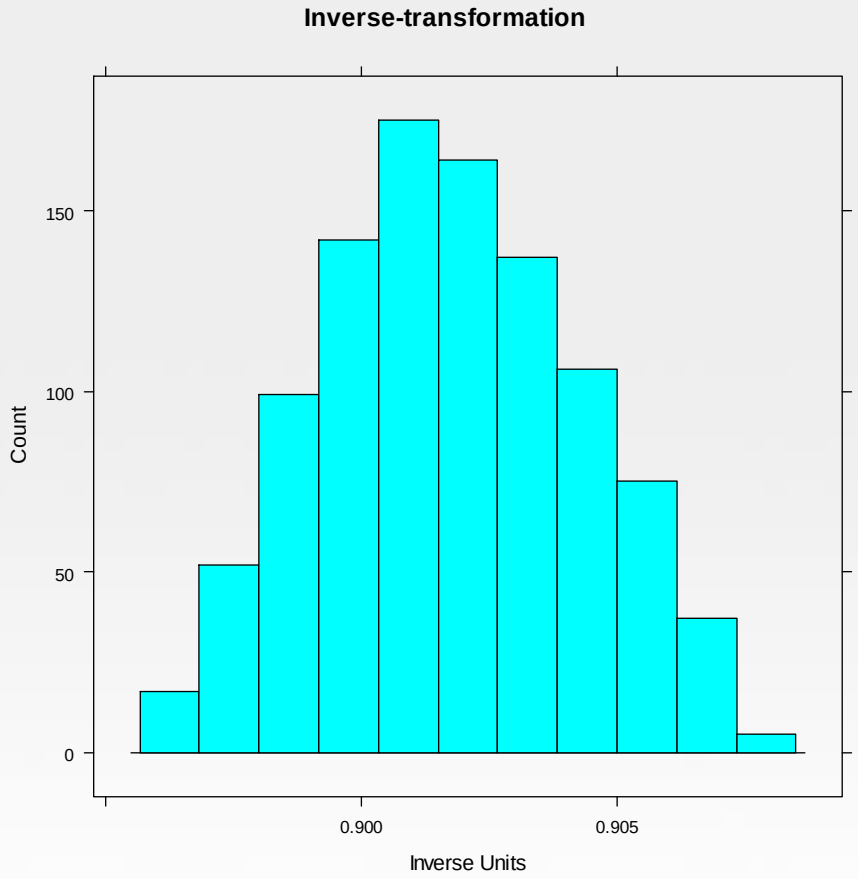
```
> BoxCoxTrans(segTrainX$PerimCh1)
Box-Cox Transformation

1009 data points used to estimate Lambda

Input data summary:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  47.74   64.37   79.02   91.61  103.24  459.77

Largest/Smallest: 9.63
Sample Skewness: 2.59

Estimated Lambda: -1.1
```

$\lambda \approx -1$ → inverse - transformation!

Conclusion: we may let $\lambda = -1.1$ indicating an inverse transformation.

Histograms of before and after trans.

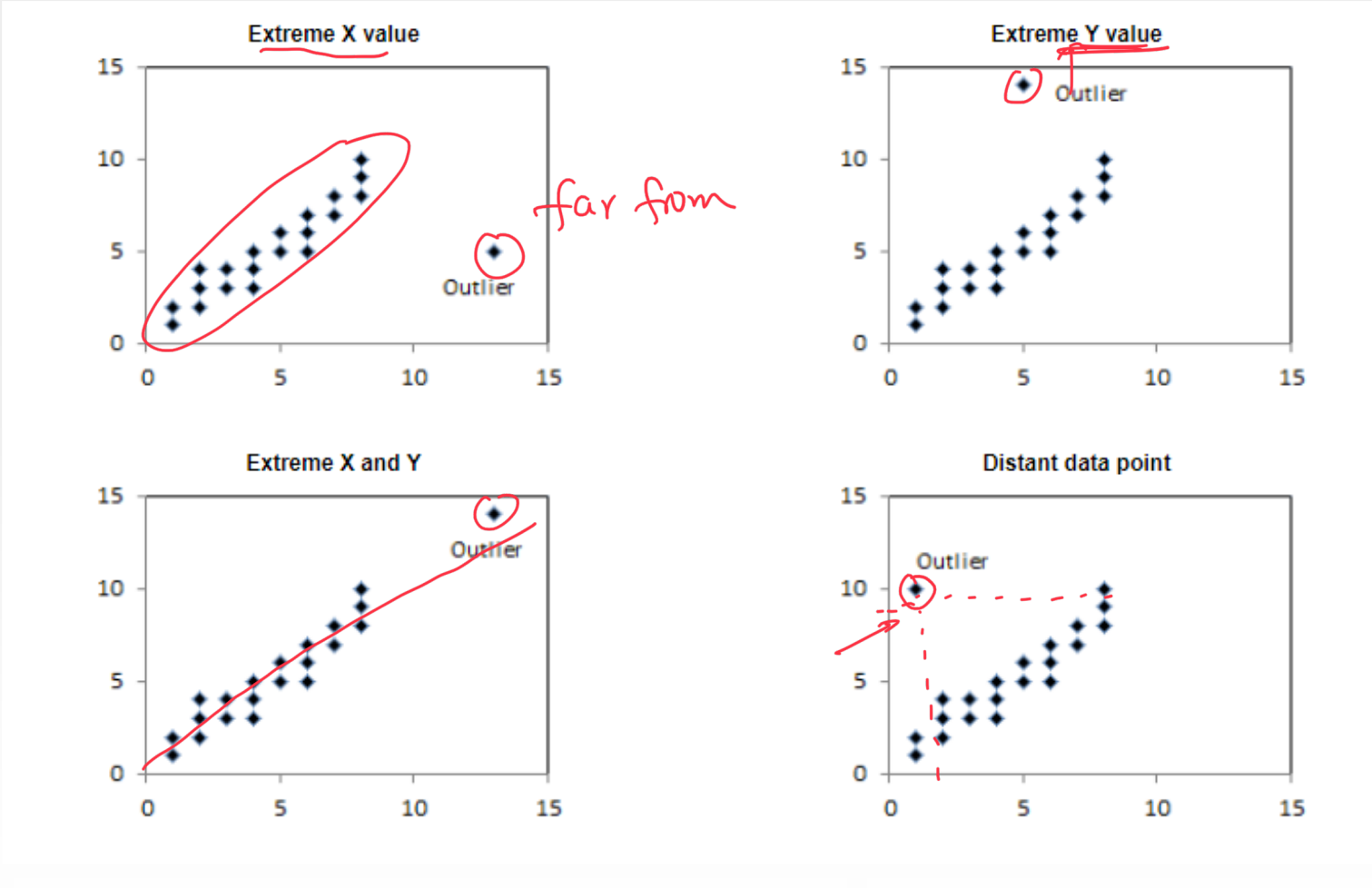




Outliers

- Outliers are defined as samples that are exceptionally far from the mainstream of the data.
- The outlying data may be an indication of a special part of the population under study that is just starting to be sampled. (Salary of UTSA faculty)
- There are four ways that a data point might be considered an outlier.
  - It could have an extreme *X* value compared to other data points.
  - It could have an extreme *Y* value compared to other data points.
  - It could have extreme *X* and *Y* values.
  - It might be distant from the rest of the data, even without extreme *X* or *Y* values.

Outliers


Extreme X value
Outlier
far from
Extreme Y value
Outlier
Extreme X and Y
Outlier
Distant data point
Outlier

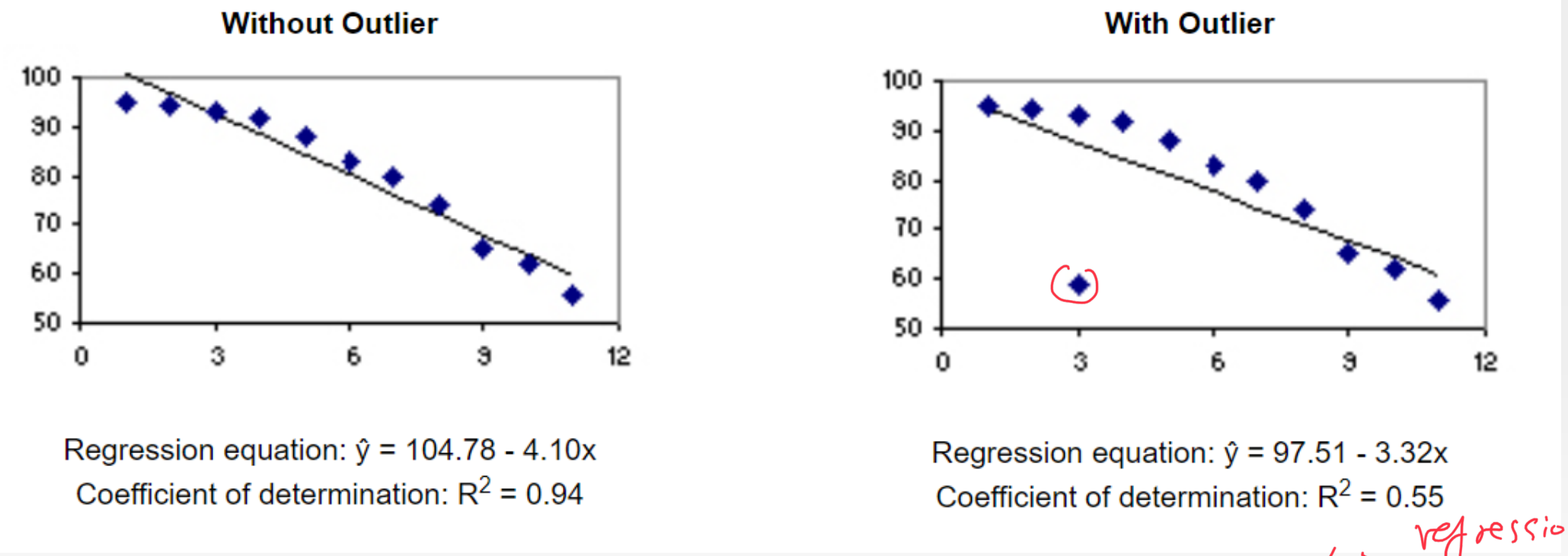
Influential points

- An influential point is an outlier that greatly affects the *slope* of the regression line.
- One way to test the influence of an outlier is to compute the regression equation with and without the outlier.

β_1

Income data in Ex 1:

Influential points



regression slope $\hat{\beta}_1$

- The scatterplots are identical, except that one plot includes an outlier. When the outlier is present, the slope is flatter (–4.10 vs. –3.32); so this outlier would be considered an influential point.

Spatial sign to resolve outliers

- There are several predictive models that are resistant to outliers; such as tree-based classification models, support vector machines (SVM) for classification.
- If a model is sensitive to outliers, a data transformation that can minimize the problem is the spatial sign.
- It projects the predictor values onto a multidimensional sphere. This has the effect of making all the samples the same distance from the center of the sphere. Mathematically, each sample is divided by its squared norm:

$$x_{ij}^{*} = \frac{x_{ij}}{\sum_{j=1}^{P} x_{ij}^{2}}.$$

new obs → x_{ij}^{*} ; x_{ij} ← original obs ; P: # of predictors

Remark:

- It is important to center and scale the data prior to using this transformation, since the denominator is intended to measure the squared distance to the center of the distribution.
- The *spatial sign* transforms the predictors as a group. Removing predictors after applying the spatial sign may be problematic.

In R Usage: spatial.sign(X, center = TRUE, shape = TRUE, na.action = na.fail, ...)

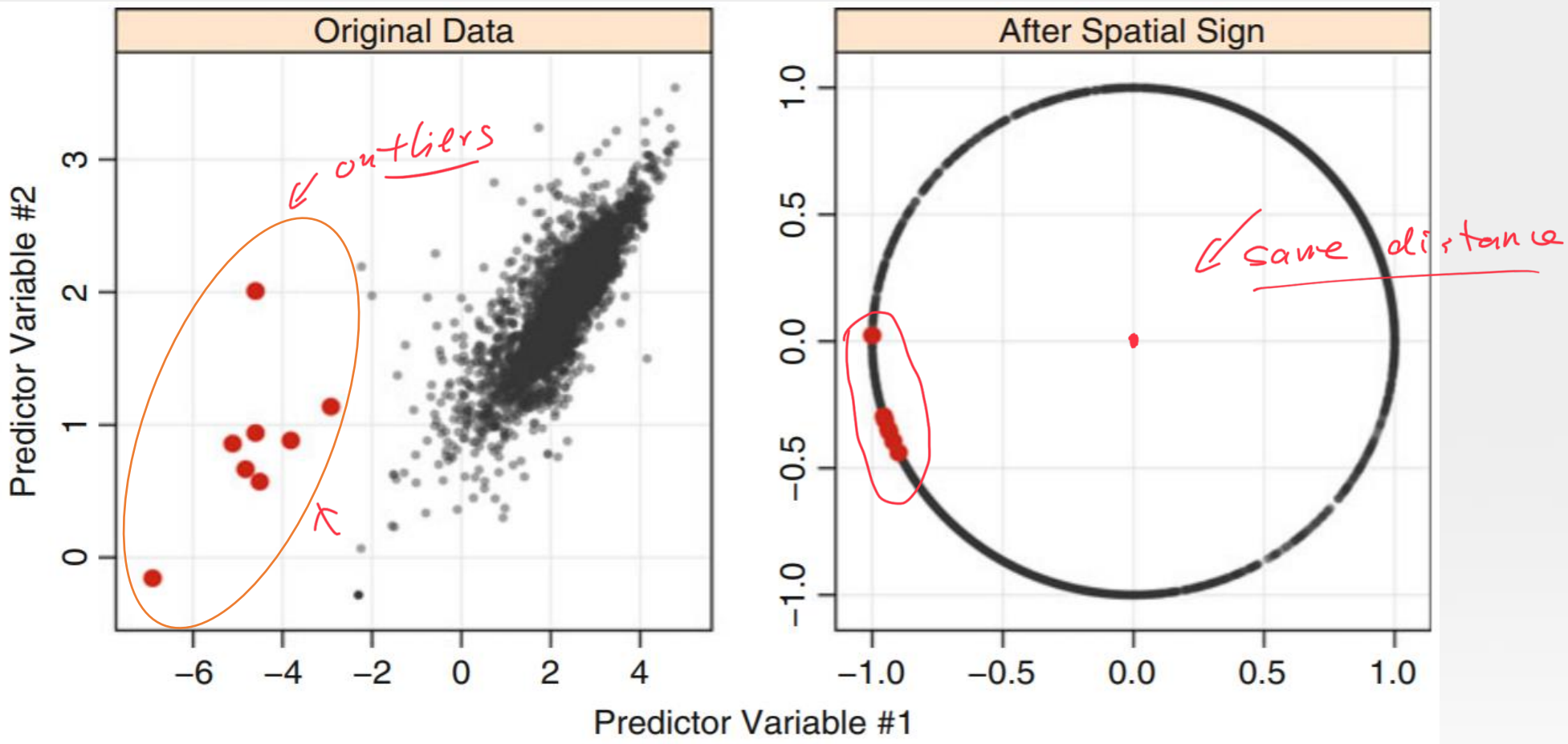


Fig. 3.4: *Left*: An illustrative example with a group of outlying data points. *Right*: When the original data are transformed, the results bring the outliers towards the majority of the data

R demonstration 3(1)

- R demonstration 3(1): Data scaling, transformation, and outliers