Contents

Diagnostics and Remedies (Ch.10)

1. Overview

Recall the discussion on regression diagnostics in simple linear regression (see summary in handout ***Regression_Diagnostics.doc***). We want to check assumptions for the regression model:

- Model
  - Linearity
  - Choice of predictors

- The error term
  - Normality
  - Constant variance
  - Independent

- Data
  - Outliers?

2. Added-variable plots (aka, Partial Regression plot)

- Consider the partial regression plot for X_1:
  Step 1. Use the other X's to predict Y, get the residuals e(Y | other X's).
  Step 2. Use the other X's to predict X_1, get the residuals e(X_1 | other X's).
  Step 3. Plot the residuals e(Y | other X's) vs e(X_1 | other X's).

- Repeat the above process for all predictors in the model.

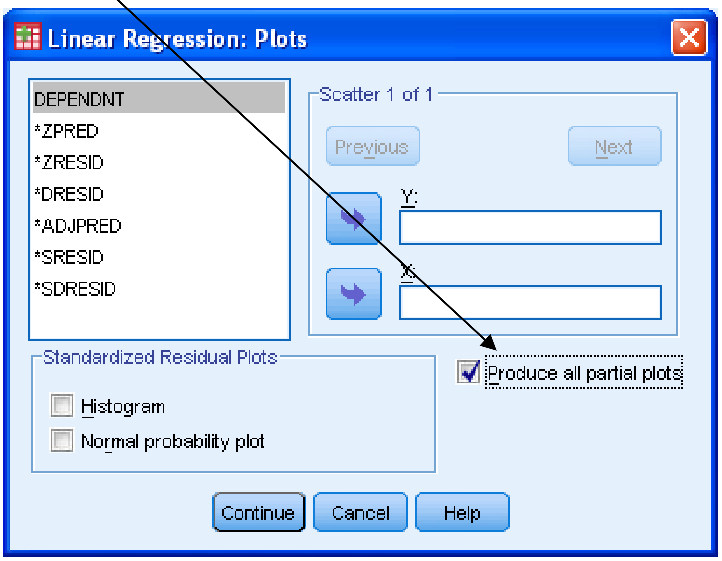
- These plots show the strength of the <u>*marginal*</u> relationship between Y and Xi in the full model.

- They can also detect:
  - Nonlinear relationships
  - Heterogeneous variances
  - Outliers

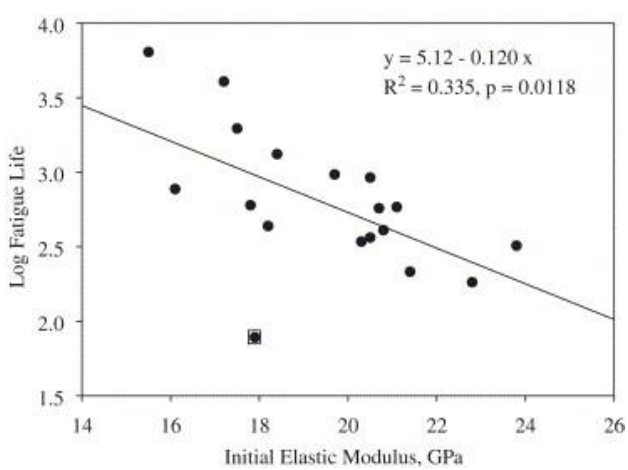
- Read the plots (similar to the residual plots)

- If there is no pattern, the predictor may be omitted from the model.
- If there is a linear pattern, the predictor shall be added to the model. Depending on the shape of the pattern, the predictor may be added using first order, polynomial, or other transformations.

➢ In SPSS, in the linear regression window click the "Plots …" button, then check the box to "Produce all partial plots"



3. Identifying outlying Y observations: Studentized Deleted Residuals



➢ Other forms of residuals

a. Residual: $e_i = Y_i - \hat{Y}_i$

b. Standardized (semi-studentized) residual: $e_i^* = \dfrac{e_i}{\sqrt{MSE}}$ (Recall that $\hat{\sigma}^2 = MSE$.)

c. Studentized residual: $r_i = \dfrac{e_i}{\sqrt{MSE(1-h_{ii})}}$, where h_{ii} is the "leverage" of the i-th case

(more on the leverage later.) Note that $se(e_i) = \sqrt{MSE(1-h_{ii})}$, and $r_i \sim t(n-p)$.

d. Deleted residual: $d_i = Y_i - \hat{Y}_{i(i)} = \dfrac{e_i}{1-h_{ii}}$ (Recall PRESS)

e. **Studentized deleted residual**:

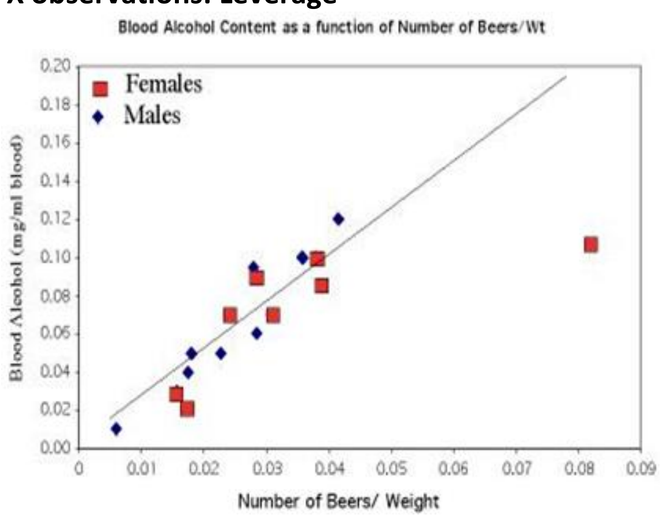
$$t_i = \frac{d_i}{se(d_i)} = \frac{d_i}{\sqrt{MSE_{(i)}/(1-h_{ii})}} = \frac{e_i}{\sqrt{MSE_{(i)}(1-h_{ii})}}$$
$$= e_i\sqrt{\frac{n-p-1}{SSE(1-h_{ii})-e_i^2}}$$
$$\sim t(n-1-p)$$

- When we examine the residuals, regardless of version, we are evaluating:
  - Outliers and influential observation.
  - Non-normal error distributions.
  - Non-constant error variance.
  - Non-linearity.

- A large $|t_i|$ ($|t_i| > t_{(1-\alpha/2,\, n-p-1)}$) suggests outlying Y observations. Since we are testing for all cases, use Bonferroni to adjust for multiple tests, or use a small α (e.g., 0.01, 0.005).

- In SPSS, save the Studentized deleted residuals (click "save" in linear regression window, then check the corresponding box.). Then use Analyze → Forecasting → Sequence Charts … to prepare an index plot. (See example on p.7.)

4. Identifying outlying X observations: Leverage



- The “leverage” of the i-th case (denoted as h_{ii}) is the i-th diagonal component of the Hat Matrix $H = X\left(X^T X\right)^{-1} X^T$.

- It is a measure of the distance between the X values for the i-th case and the means of the X values. This is why it can be used to identify outlying X observations.

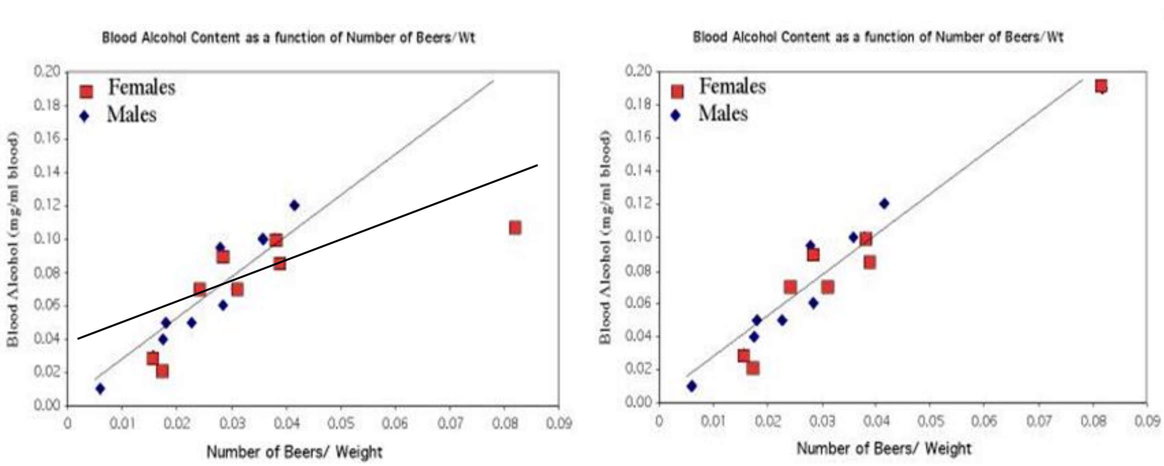
- $0 \le h_{ii} \le 1$, $\Sigma(h_{ii}) = p$, and the average value of h_{ii} is p/n.

- Large value of h_{ii} suggests that i-th case is distant from the center of all X’s. Values far from the average point (e.g., $h_{ii} > 2p/n$ for large data sets) are outlying X observation and the case should be examined carefully.

- Computing leverage for a new set of X’s can help check for extrapolation.

- In SPSS, save the leverage, then plot. Note that SPSS uses a “centered leverage,” which is (h_{ii} – 1/n).

5. Identifying influential cases



- ***DFFITS$_i$*** is a measure of the influence of the i-th case on $\hat{Y}_i$ (a single case)

$$DFFITS_i = \frac{\hat{Y}_i - \hat{Y}_{i(i)}}{\sqrt{MSE_{(i)} h_{ii}}} = t_i \sqrt{\frac{h_{ii}}{1 - h_{ii}}},$$

where t_i is studentized deleted residual.

  - It is a standardized version of the difference between $\hat{Y}_i$ computed with and without the i-th case.

- The i-th case is an influential case if $|DFFITS_i|$ (i.e., the absolute value of $DFFITS_i$) is greater than 1 for small data sets, or greater than $2\sqrt{p/n}$ for large data sets.

➢ ***Cook's Distance*** is a measure of the influence of the i-th case on all of the $\hat{Y}_i$'s (all the cases).

$$D_i = \frac{\sum_{j=1}^{n}\left(\hat{Y}_j - \hat{Y}_{j(i)}\right)^2}{p \cdot MSE} = \frac{h_{ii}}{p \cdot MSE}\left(\frac{e_i}{1-h_{ii}}\right)^2$$

- It is a standardized version of the sum of squares of the differences between the predicted values computed with and without case I.

- The i-th case is an influential case if $D_i > F(0.5, p, n-p)$.

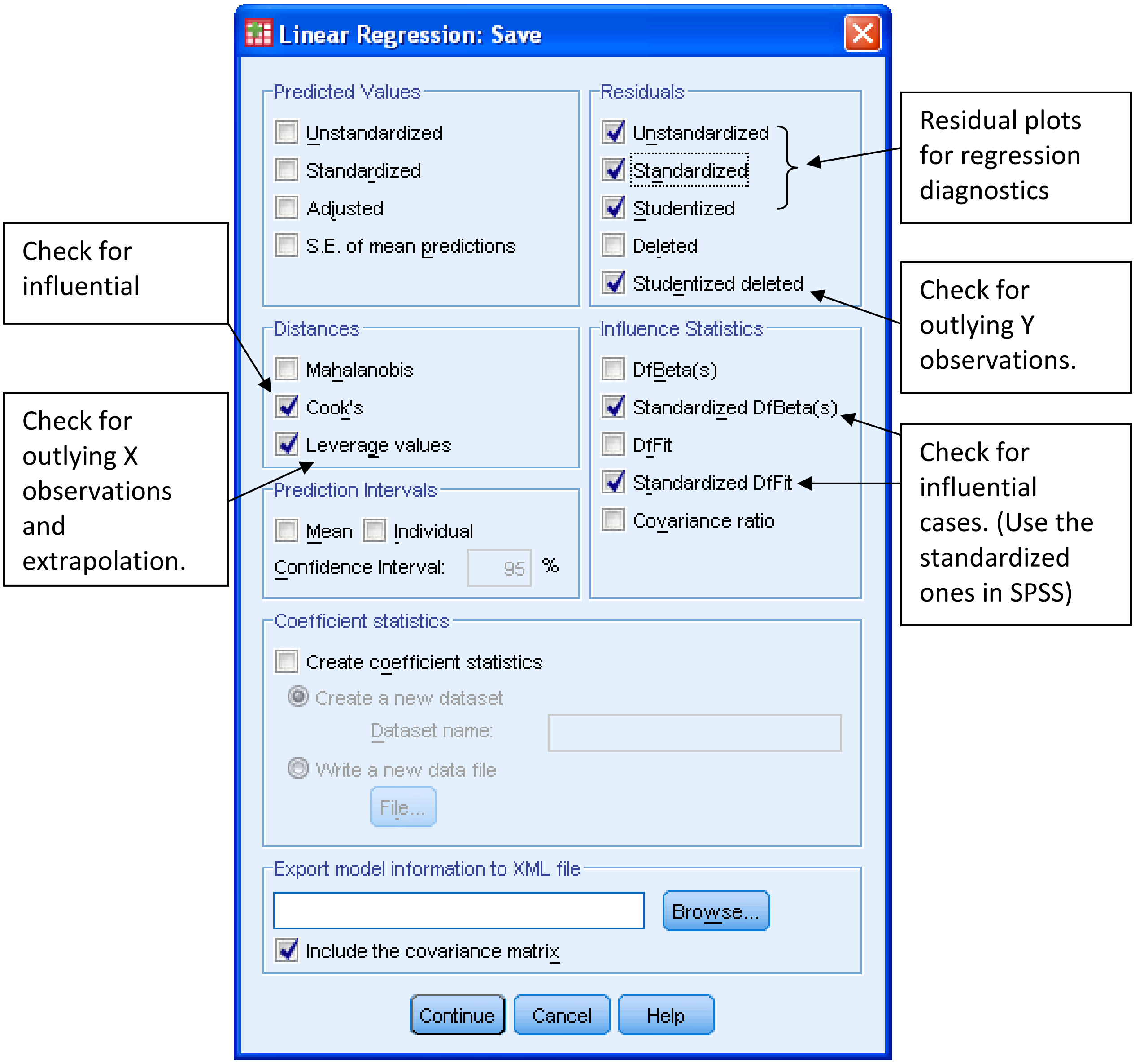
- Others also suggest use $D_i > 1$, $D_i > 4/(n-p)$, or $D_i > 4/n$.

➢ ***DFBETAS$_{k,i}$*** is a measure of the influence of the i-th case on the k-th regression coefficients.

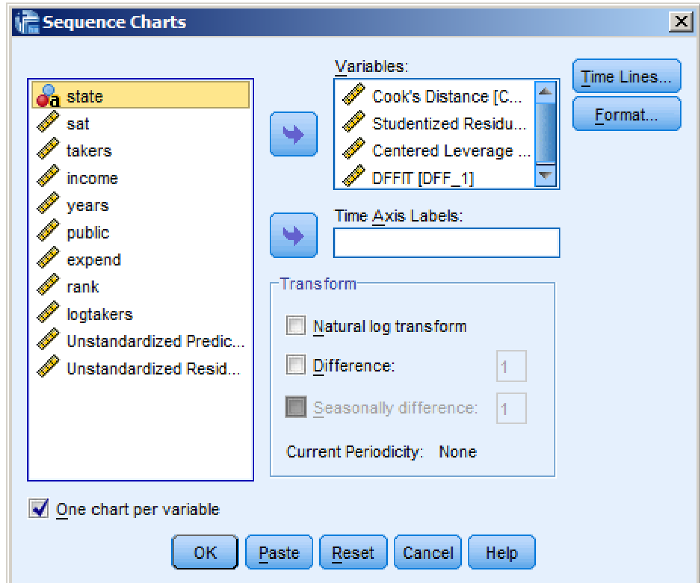
- It is a standardized version of the difference between $b_k = \hat{\beta}_k$ computed with and without the i-th case.

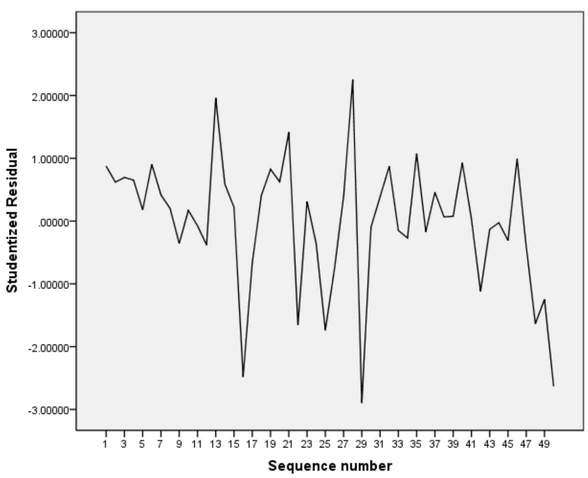
- The i-th case is an influential case if $|DFBETAS_{k,i}|$ (i.e., the absolute value of $DFBETAS_{k,i}$) is greater than 1 for small data sets, or greater than $2/\sqrt{n}$ for large data sets.

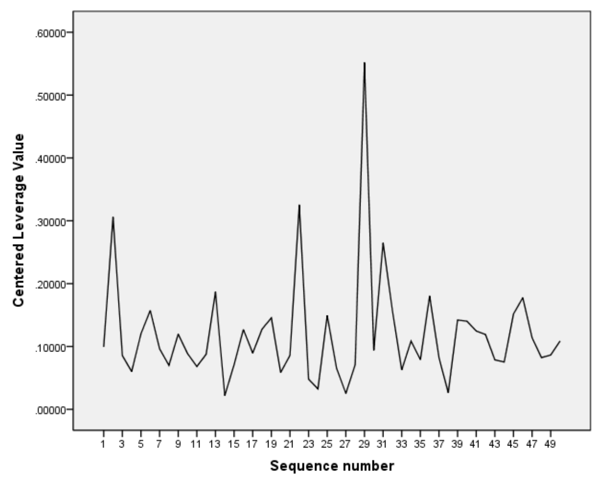
➢ In SPSS, save the above values, then plot. Note that the above formulas are the "standardized" DFFits and "standardized" DFBetas in SPSS.

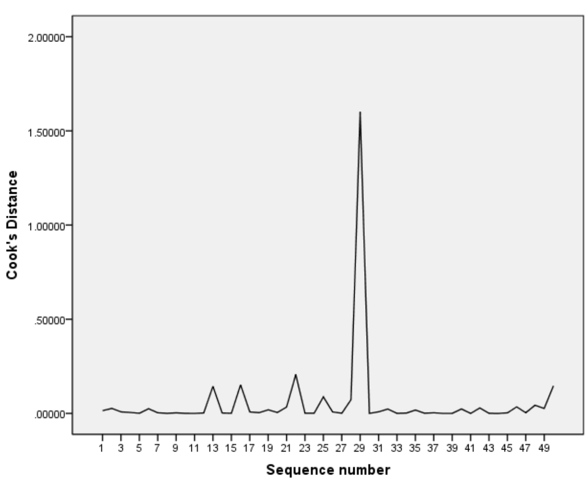

Linear Regression: Save
Predicted Values
Unstandardized
Standardized
Adjusted
S.E. of mean predictions
Residuals
Unstandardized
Standardized
Studentized
Deleted
Studentized deleted
Distances
Mahalanobis
Cook's
Leverage values
Influence Statistics
DfBeta(s)
Standardized DfBeta(s)
DfFit
Standardized DfFit
Covariance ratio
Prediction Intervals
Mean
Individual
Confidence Interval: 95 %
Coefficient statistics
Create coefficient statistics
Create a new dataset
Dataset name:
Write a new data file
File...
Export model information to XML file
Browse...
Include the covariance matrix
Continue
Cancel
Help
Residual plots for regression diagnostics
Check for influential
Check for outlying Y observations.
Check for outlying X observations and extrapolation.
Check for influential cases. (Use the standardized ones in SPSS)

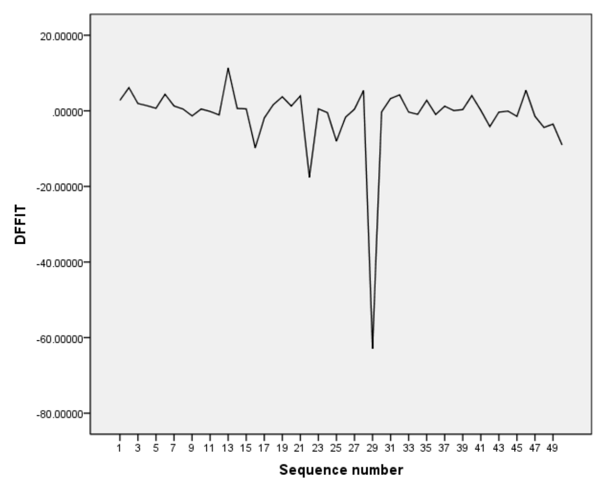
After saving these statistics, use Analyze → Forecasting → Sequence Charts





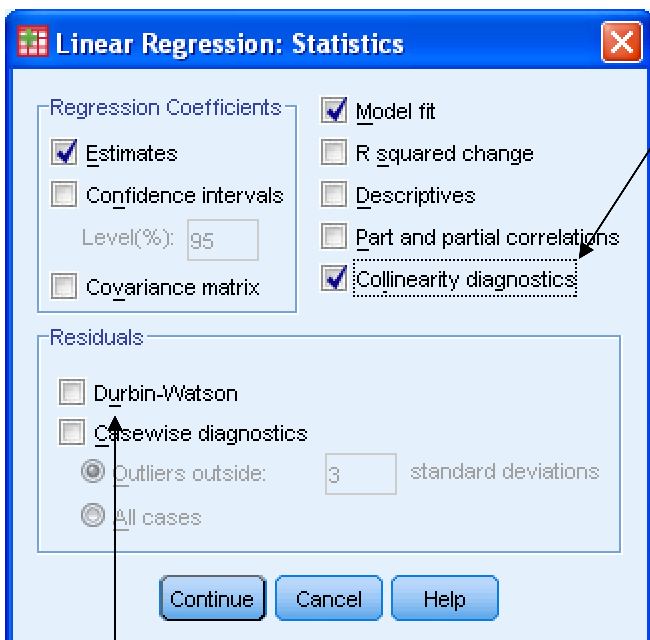






6. Multicollinearity: Variance Inflation Factor (VIF)

- Recall: Why is multicollinearity an issue in multiple linear regression?
- Use scatter plots between predictors to check for correlation.
- Compute the Variance Inflation Factor for each predictor: $VIF_k = 1/(1-R_k^2)$, where R_k^2 is the coefficient of determination from a regression between X_k (treated as the response) and all other predictors.
- One suggested rule is that a value of 10 or more for VIF indicates excessive multicollinearity.
- Some texts use $Tolerance_k = 1 - R_k^2 = 1/VIF_k$.
- In SPSS, click "Statistics…" in the linear regression window and check Collinearity diagnostics



7. Serial Correlation: Durbin-Watson Test

- Only test for serial correlation (dependence) when data have time order.
- Use Durbin-Waston to test for 1st order auto-correlation.

- Read text, p.487~490 for more details. Note that:

  - The calculation is done by software, so we just need to know when and how to use the resulting values.

  - This test conclusion can be Ho, Ha, or "inconclusive."

  - Adjust the statistic and alpha level for different Ha.

  - Table B.7 at the end of the text has the test bounds.

8. What to do next, if there are concerns with assumptions?

- In Ch.3, we discussed a few remedial measures, such as transforming the data.

- For data with non-constant error variance: Weighted Least Squares, etc.

- For data with multicollinearity: Ridge Regression, Principle Component Analysis, Factor analysis, etc.

- For outliers and influential cases: Robust Regression, Weighted Least Squares, etc.

- For non-Normal and/or non-linear data: Generalized Linear model, Loess methods, Regression and classification tree, etc.

- For dependent errors: Time series, General Linear model (repeated measure), etc.

- Use other methods (such as bootstrap) for statistical inference.