Regression Models
Diagnostics

Alexander Trofimov
PhD, professor, NRNU MEPhI

lab@neuroinfo.ru
http://datalearning.ru

Course "Machine Learning"

September 2021

Regression Diagnostics

Regression diagnostics is a part of regression analysis whose objective is to investigate if the trained model and the assumptions we made about the data and the model, are consistent with the observed data

Ways to regression diagnostics:

- Checking the adequacy of the assumptions of regression analysis
- Detecting extreme points (outliers) that may be dominating the regression and possibly distorting the results
- Detecting if strong relationships among the independent variables (collinearity) are affecting the results
- Assessing model structure

Approaches to regression diagnostics:

- Graphical analysis
- Quantitative analysis

Importance of Assumptions in Regression Analysis

Linear regression model:

$$Y|x = x\beta + \varepsilon(x)$$

where $x = (1, x_1, ..., x_k)$, $\beta = (\beta_0, ..., \beta_k)^T$

Assumptions in linear regression analysis:

- Linearity
- Exogeneity of regressors
- Homoscedasticity
- Independence of errors
- Normality
- Variability of regressors

If these assumptions are violated, then the statistical inference may be invalid

Checking Linearity

Assumption: the regression function is linear $\varphi(x) = x\beta$

Graphical diagnostics:

- Residual plots
- Partial regression plot

If there is a clear non-linear pattern in any of these plot, there is a problem of non-linearity

Quantitative diagnostics:

- Tests of the functional form of the model (Ramsey's regression error specification test, etc.)

Solutions:

- Transformation of variables
- Non-linear regression

Non-linearity in Multiple Regression

The plots of residuals vs each of the predictor variables, outcome or fitted values are called as residual plots

For simple regression:

The non-linearity is evident. The residual plots are usually sufficient to identify non-linearity

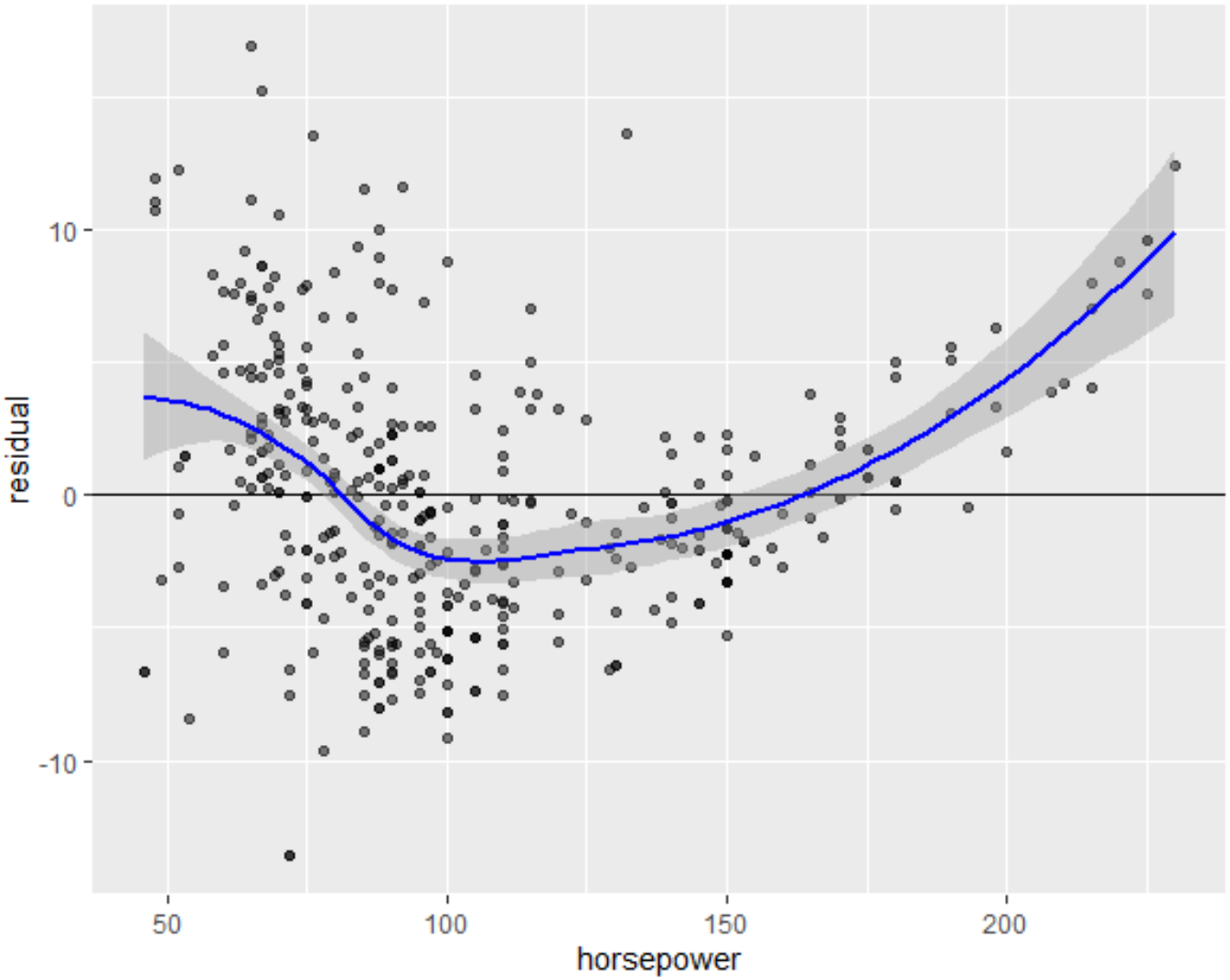
For multiple regression:

The non-linearity can be masked. The individual bivariate plots do not take into account the effect of the other explanatory variables in the model

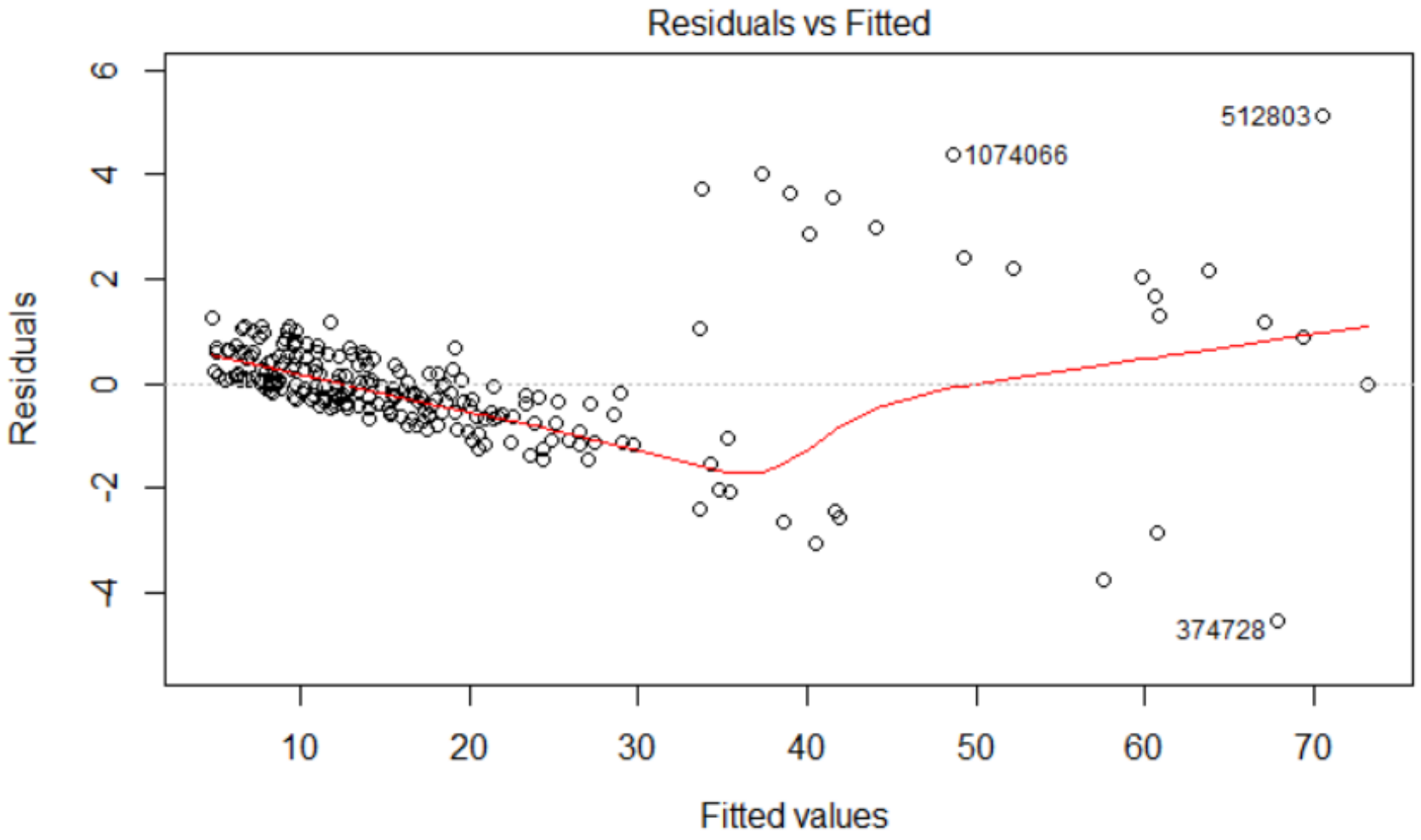
We need to look at the relationship between the outcome and explanatory variables conditional on the other explanatory variables

How to exclude the effect of some variables on another variable?

Residual Plot. Illustration 1



Residual Plot. Illustration 2



Omitted Variable Bias

The influences on the dependent variable Y which are not captured by the model are collected in the error term, which we assumed to be uncorrelated with the regressors

If there is a variable omitted in the model but influencing the outcome Y and it is related with existing regressors, then the assumption of exogeneity is violated

The bias of OLS estimates due to this model misspecification is called as omitted variable bias (OVB)

Negative effects of OVB:

- OLS estimates become biased
  Wrong interpretation of regression coefficients
- OLS estimates become inconsistent
  The OVB cannot be solved by increasing the number of observations

OVB and Model Misspecification

The true model:

$$Y|x = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon(x)$$

where $\varepsilon(x) \sim N(0, \sigma^2)$

Good model:

$$Y|x = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon_1(x)$$

The OLS-estimates $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$ will be unbiased and consistent

Under-specified model:

$$Y|x = \beta_0 + \beta_1 x_1 + \varepsilon_2(x)$$

The error term $\varepsilon_2(x)$ will take the influence of x_2

x_1 and x_2 are independent $\Rightarrow$ $\varepsilon_2(x)$ and x_1 are independent

x_1 and x_2 are correlated $\Rightarrow$ $\varepsilon_2(x)$ and x_1 are correlated, the exogeneity is violated $\Rightarrow$ $\hat{\beta}_0, \hat{\beta}_1$ are biased and inconsistent

Omitted Variable Bias. Examples

The true model:

$$salary = \beta_0 + \beta_1 education + \beta_2 skills + \varepsilon(x)$$

Under-specified model:

$$salary = \beta_0 + \beta_1 education + \varepsilon_1(x)$$

The OLS-estimate $\hat{\beta}_1$ will be overestimated (positive bias), since $education$, $skills$ and $salary$ are positively correlated

The true model:

$$score = \beta_0 + \beta_1 gametime + \beta_2 studytime + \varepsilon(x)$$

Under-specified model:

$$score = \beta_0 + \beta_1 gametime + \varepsilon_1(x)$$

The OLS-estimate $\hat{\beta}_1$ will be underestimated (negative bias), since $\rho(gametime, studytime) < 0$ and $\rho(studytime, score) > 0$

Partial Regression Plot

Regression of Y on explanation variables without x_j:

$$\hat{y} = X_{\sim j}\hat{\beta} + \varepsilon_Y$$

Regression of X_j on other explanation variables:

$$\hat{x}_j = X_{\sim j}\tilde{\beta} + \varepsilon_X$$

where $X_{\sim j}$ is a design matrix with excluded regressor x_j, $j = \overline{0, k}$

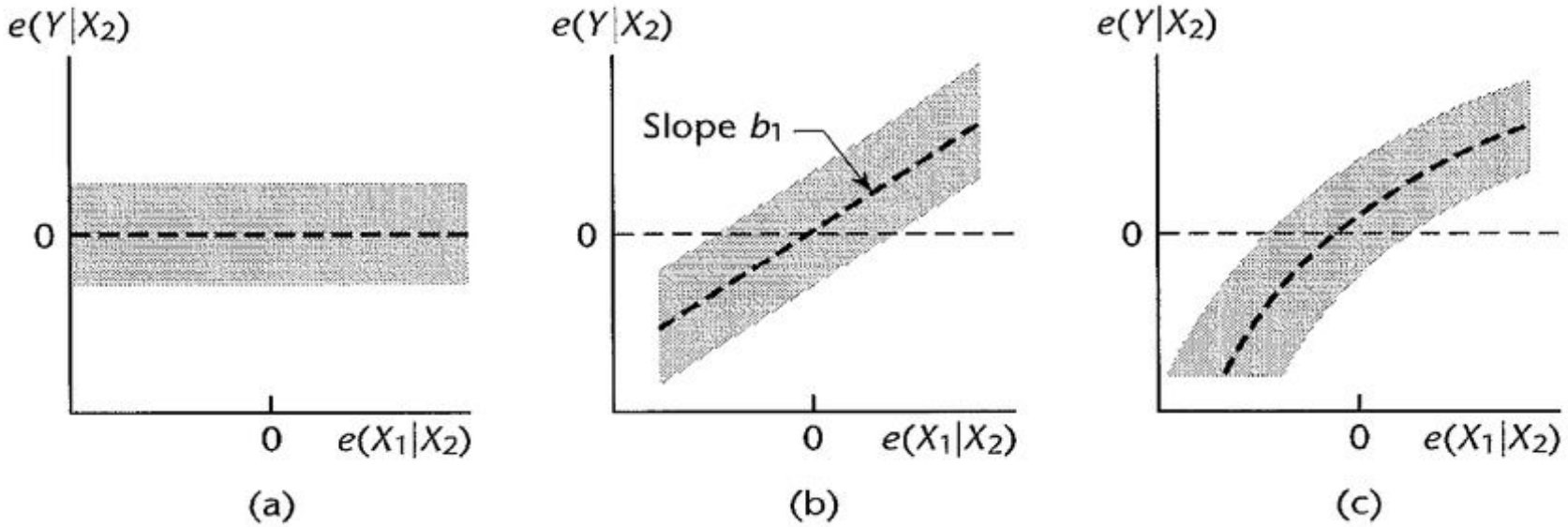
The scatter plot of ε_Y vs ε_X is called as partial regression plot, or added variable plot

ε_Y represents the part of the response values unexplained by the predictors (except x_j), and ε_X represents the part of the x_j values unexplained by the other predictors

The fitted line in ε_Y vs ε_X plane represents how the new information introduced by adding x_j can explain the unexplained part of the response values

Partial Regression Plot. Interpretation

Assume we already have a regression model of Y on x_2 and consider if we should add x_1 into the model



a) x_1 has no additional information useful for the prediction of Y beyond that contained in and provided for by x_2

b) x_1 contains useful addition information for the prediction of Y

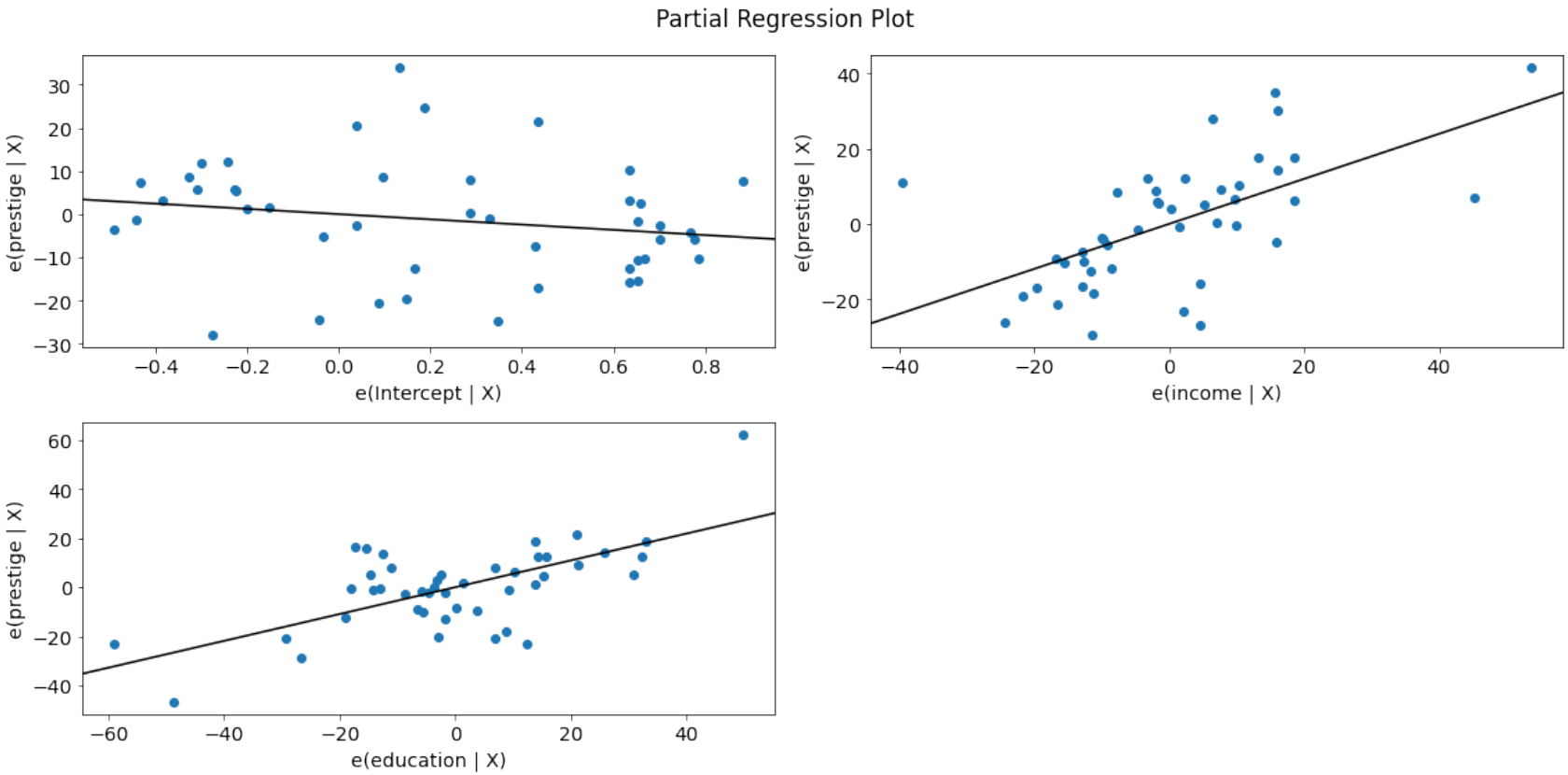
c) inclusion of x_1 is justified but some non-linear transformation are needed

Partial Regression Plot. Notes

- The partial regression plots are useful to identify heteroscedasticity, influential data points, the need to include the regressor into existing model and the need for non-linear data transformation
- If the slope of the fitted line in partial regression plot is close to zero and the confidence bounds include a horizontal line, then the new information from x_j does not explain the unexplained part of the response value. That is, x_j is not significant in the model fit
- The partial regression plots in multiple linear regression play the same role as the scatter diagrams in simple linear regression
- Some other plots (partial residual plot, CCPR plot) are related with partial regression plot

Partial Regression Plot. Illustration

Regression model: $prestige \sim 1 + income + education$



Checking Heteroscedasticity

Assumption: the variance of the errors $\mathrm{D}[\varepsilon(x)] = \sigma^2, \ \forall x \in \mathscr{X}$

Graphical diagnostics:

- Residual plots

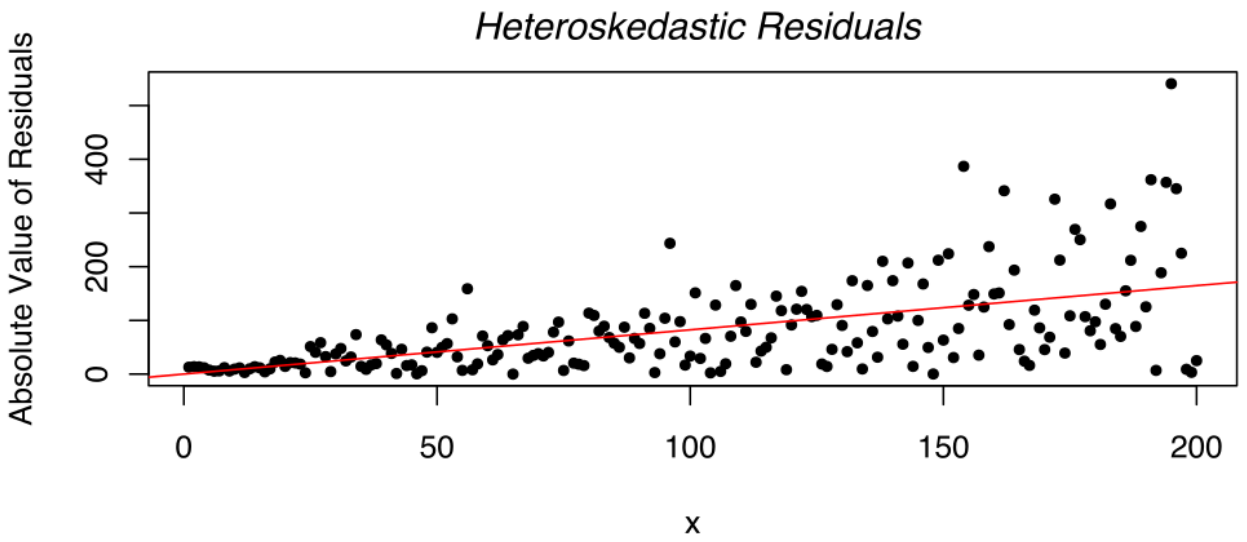
Statistical tests for heteroscedasticity:

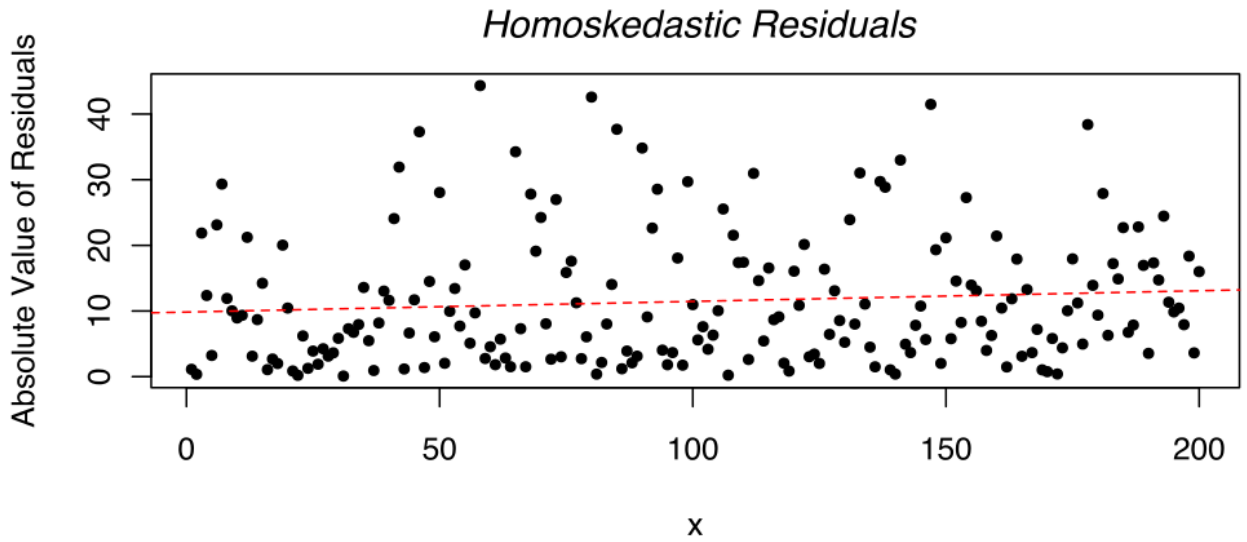
- White test
- Breush-Pagan test
- Park test
- Glejser test
- Goldfeld-Quandt test
- ...

Solutions:

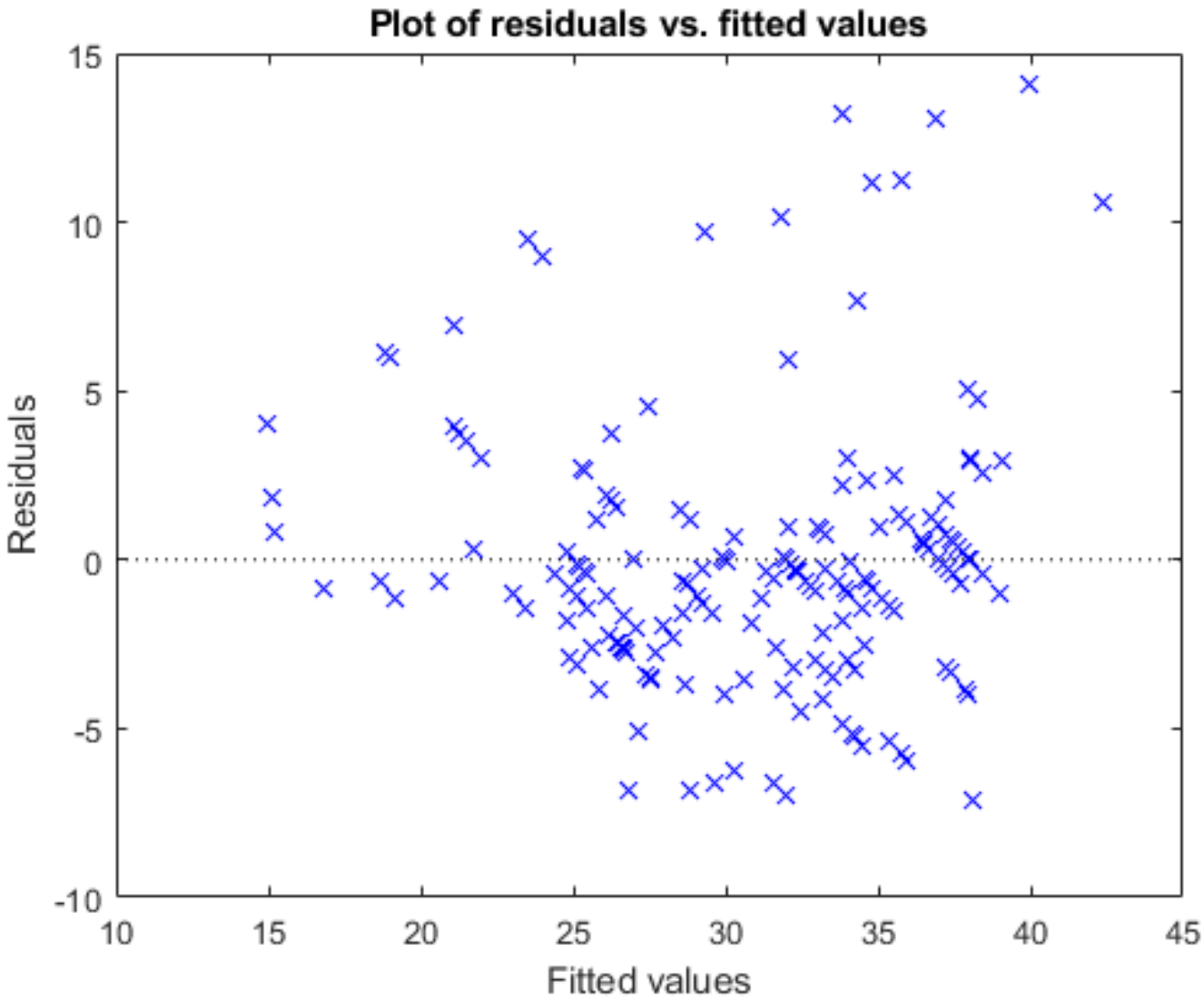
- Transformation of variables
- Weighted least squares

Heteroscedasticity of Residuals. Illustration 1





Heteroscedasticity of Residuals. Illustration 2



Exogeneity of Regressors

Linear model:

$$Y|x = x\beta + \varepsilon(x)$$

Assumption: the explanatory variable X and the error $\varepsilon(x)$ are independent $\forall x \in \mathscr{X}$. It means that the sampled values of explanatory variable are independent on model errors

The assumption of exogeneity cannot be tested without additional information about observed variables

Example: the sample obtained by using the model

$$Y|x_i = x_i\beta + \varepsilon(x_i)$$

where $x_i = y_{i-1}$, $\varepsilon(x_i)$ is a random error, is not exogenous since x_i are dependent on error $\varepsilon(x_{i-1})$

Independence of Errors

Assumption: any pair of errors $\varepsilon(x_i)$ and $\varepsilon(x_j)$ (or $Y|x_i$ and $Y|x_j$) are independent. Weaker assumption is uncorrelatedness:

$$cov[\varepsilon(x_i), \varepsilon(x_j)] = cov[Y|x_i, Y|x_j] = 0 \quad \forall x_i, x_j \in \mathscr{X},\ i \neq j$$

Graphical diagnostics:

- Residual lag plots

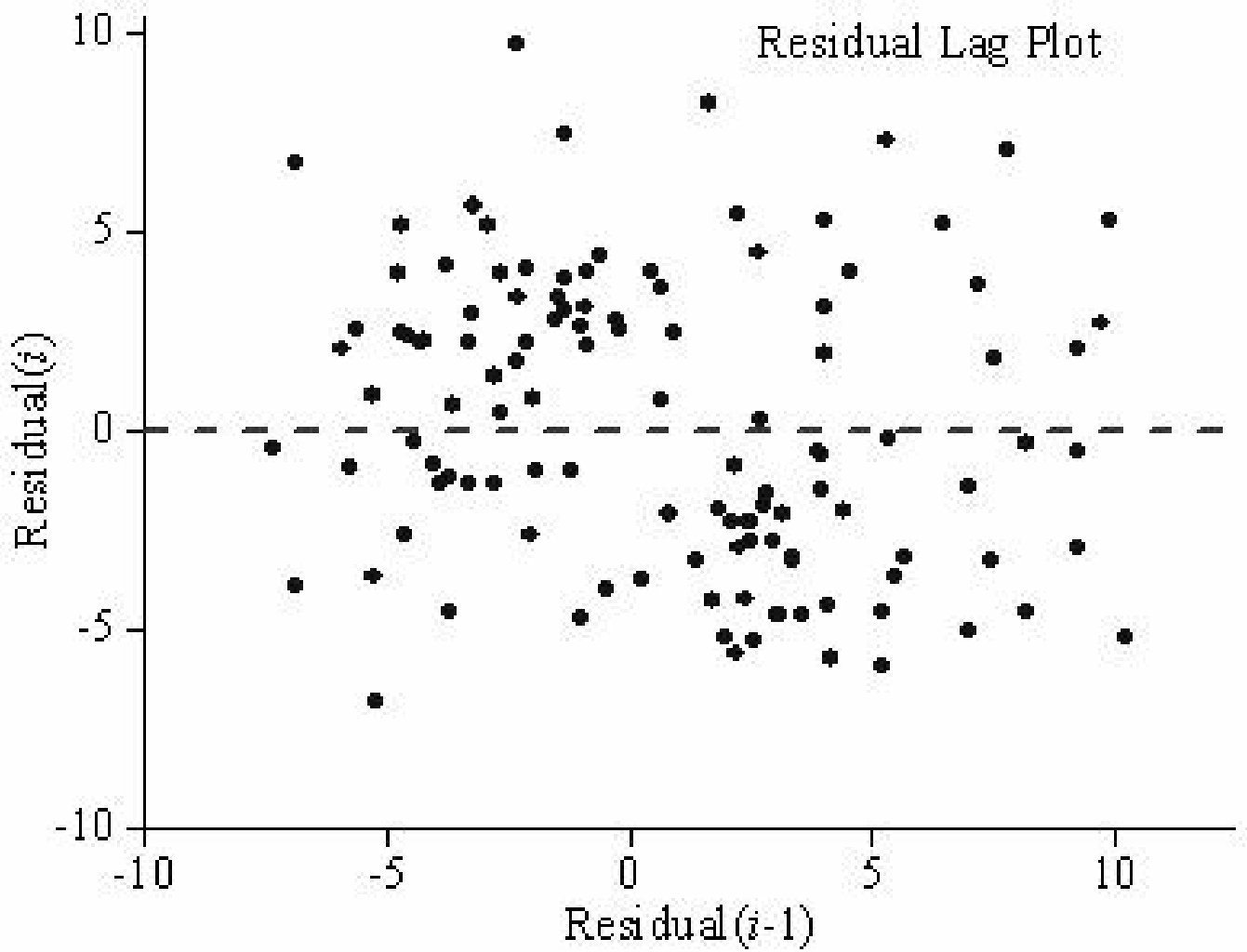
Statistical tests for autocorrelation:

- Durbin-Watson test
- Ljung-Box test
- Breusch-Godfrey test
- ...

Solutions:

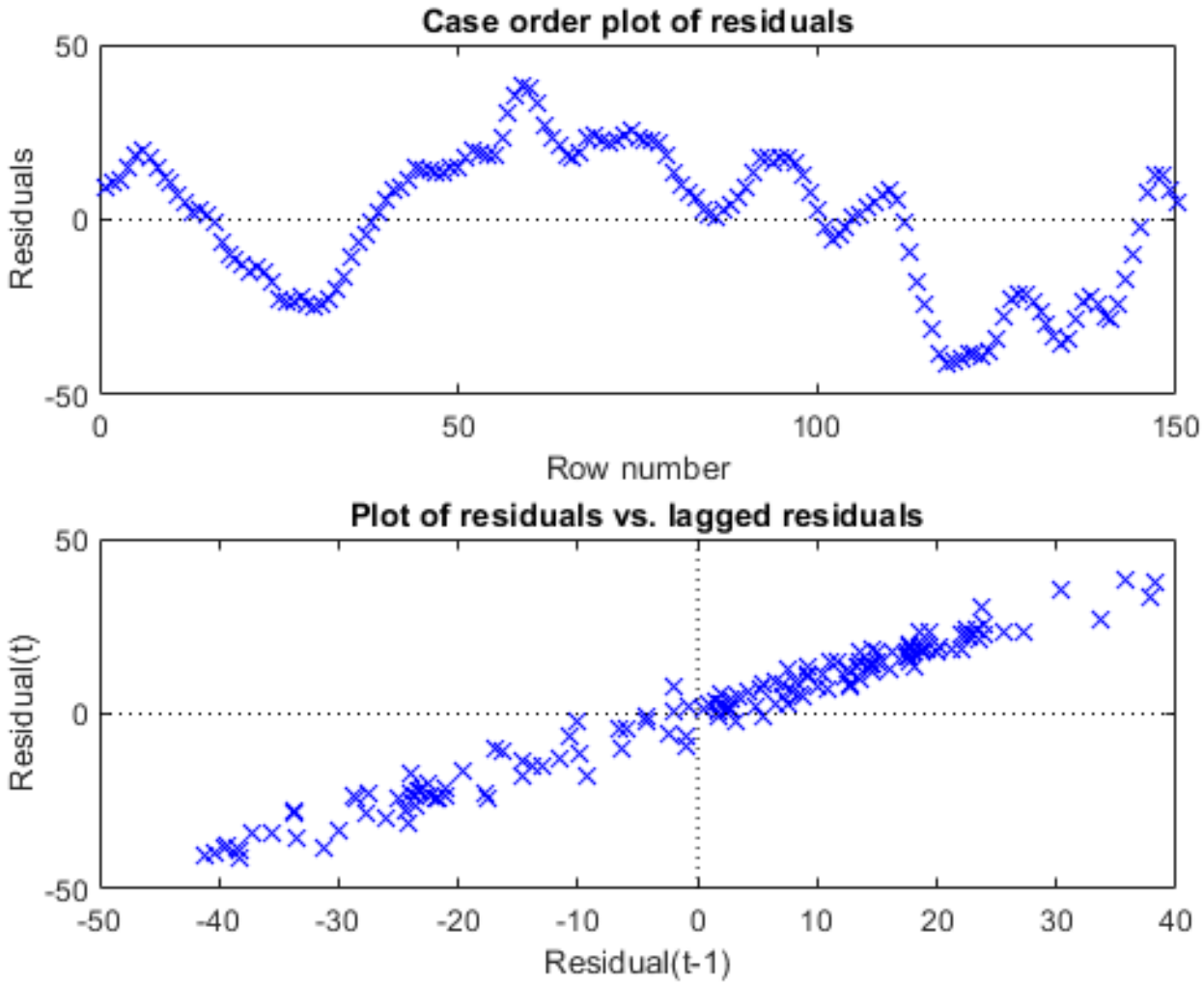
- Whitening transformation of variables
- Generalized least squares

Residual Lag Plot. Illustration 1



Residual lag plot showing that the error term is independent

Residual Lag Plot. Illustration 2



Residual lag plot showing that the error term is strongly correlated

Normality of the Residuals

Assumption: the errors $\varepsilon(x) \sim N(0, \sigma^2),\ \forall x \in \mathscr{X}$

Graphical diagnostics:

- Residual histogram
- Q-Q plot

Statistical tests for normality:

- Goodness-of-fit tests (chi-square test, Kolmogorov-Smirnov test, etc.)
- Jarque-Bera test
- Shapiro-Wilk test
- ...

Solutions:

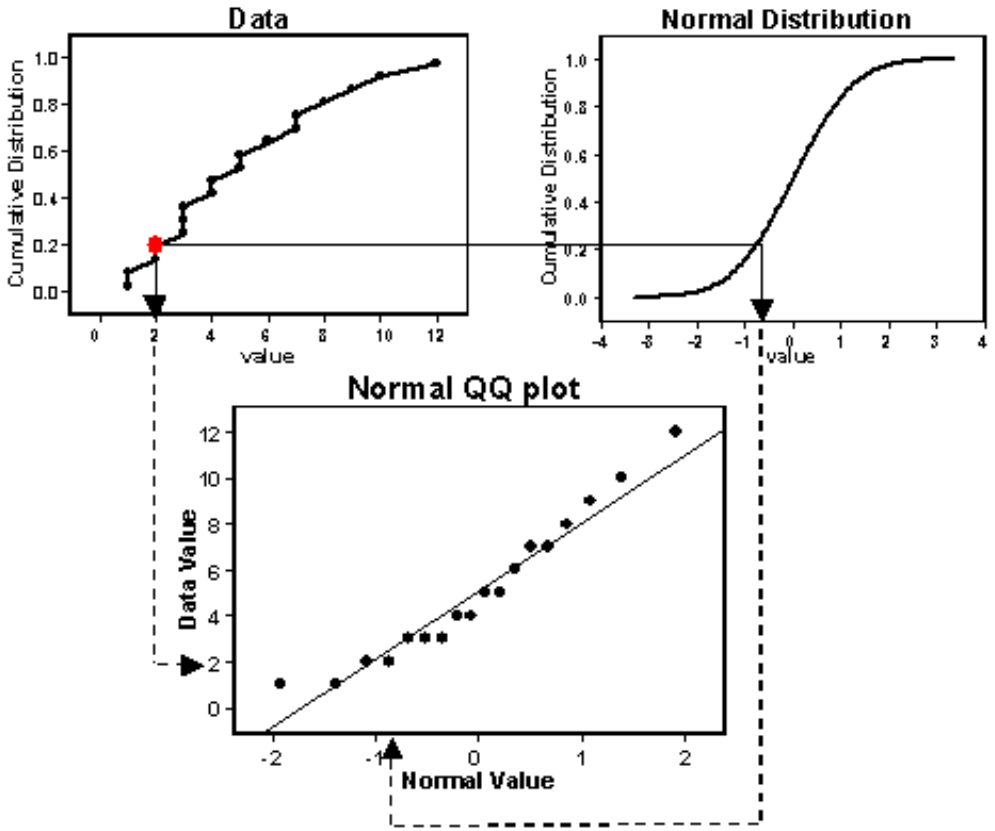
- Transformation of variables
- Generalized linear models

Q-Q Plot

Let $X \sim F(x)$ and $Y \sim G(y)$. The q-th quantiles:

$$x_q = F^{-1}(q), \quad y_q = G^{-1}(q), \quad q \in (0, 1)$$

The Q-Q (quantile-quantile) curve is a parametric curve $\{(x_q, y_q)\}$

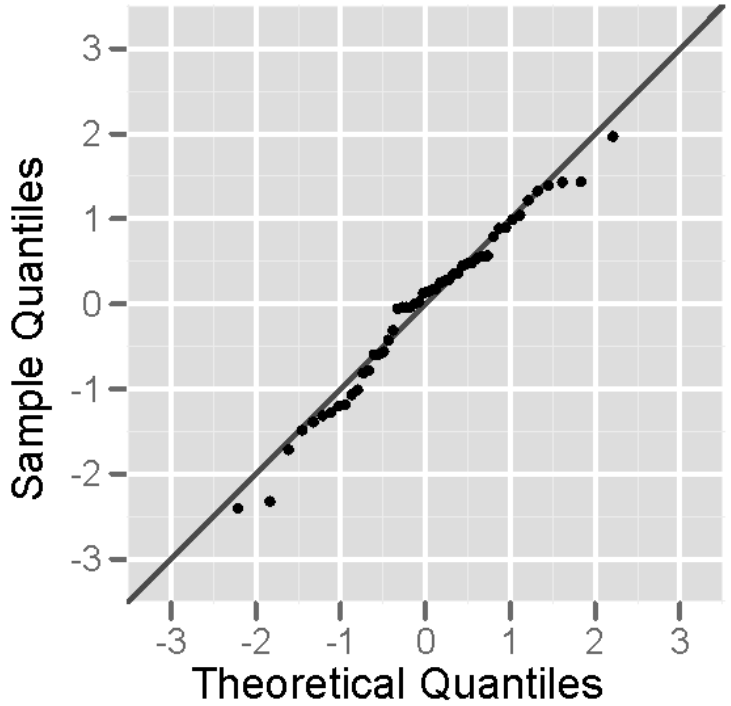


Normal Probability Plot

Q-Q plot is used to compare two probability distributions graphically

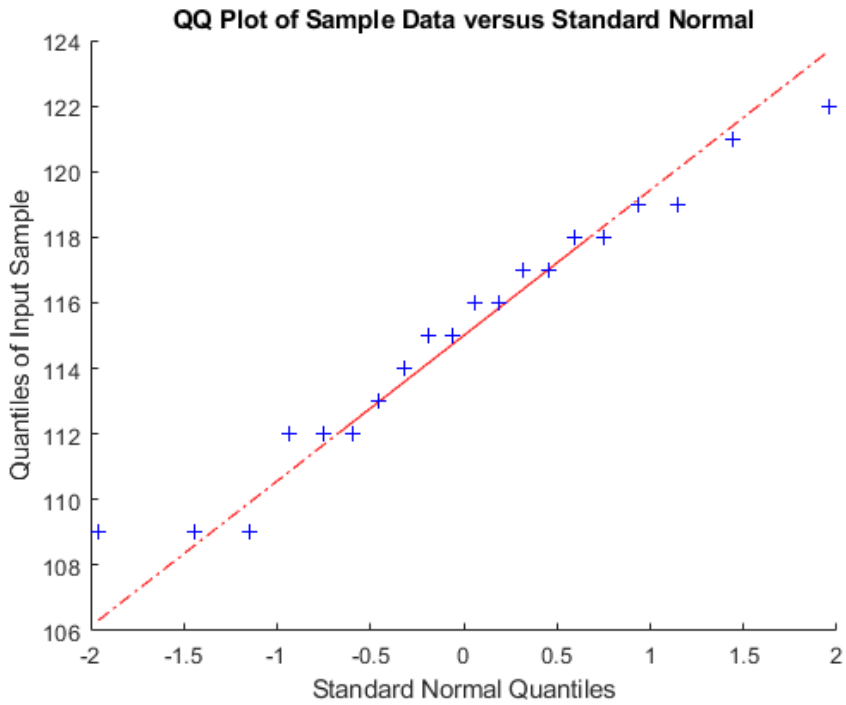
If $F(x)$ is a normal distribution and $G(y)$ is a sample distribution, then the Q-Q plot is called as normal probability plot

The expectation and variance of $F(x)$ are usually unknown, the sample mean and sample variances are used, $X \sim N(\bar{x}, S^2)$



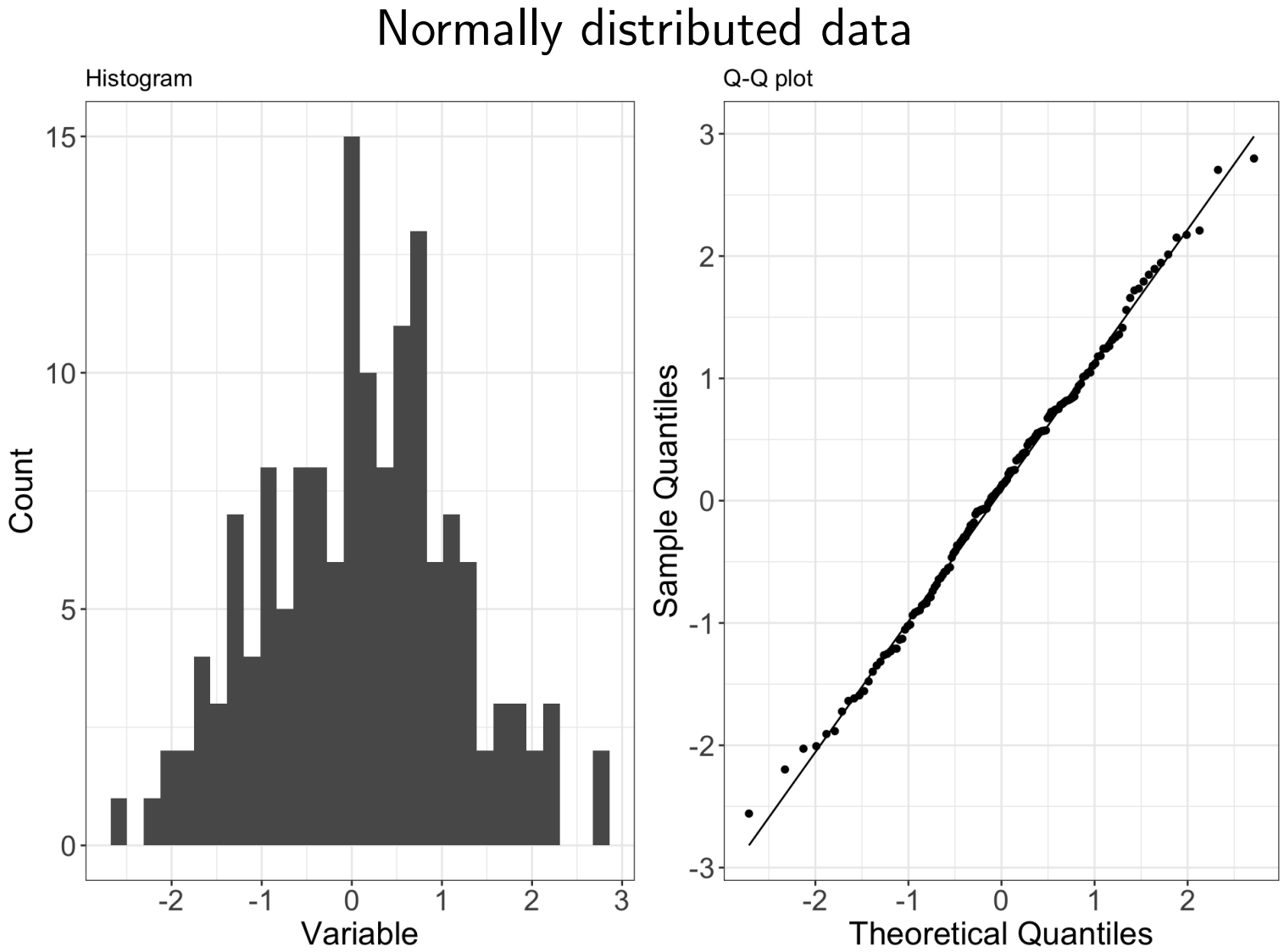
Reference Line on Q-Q Plot

The Q-Q plot usually is drawn with a reference line. The reference line connects the points corresponding to the lower and higher quartiles ($q = 0.25$ and $q = 0.75$) of the distributions

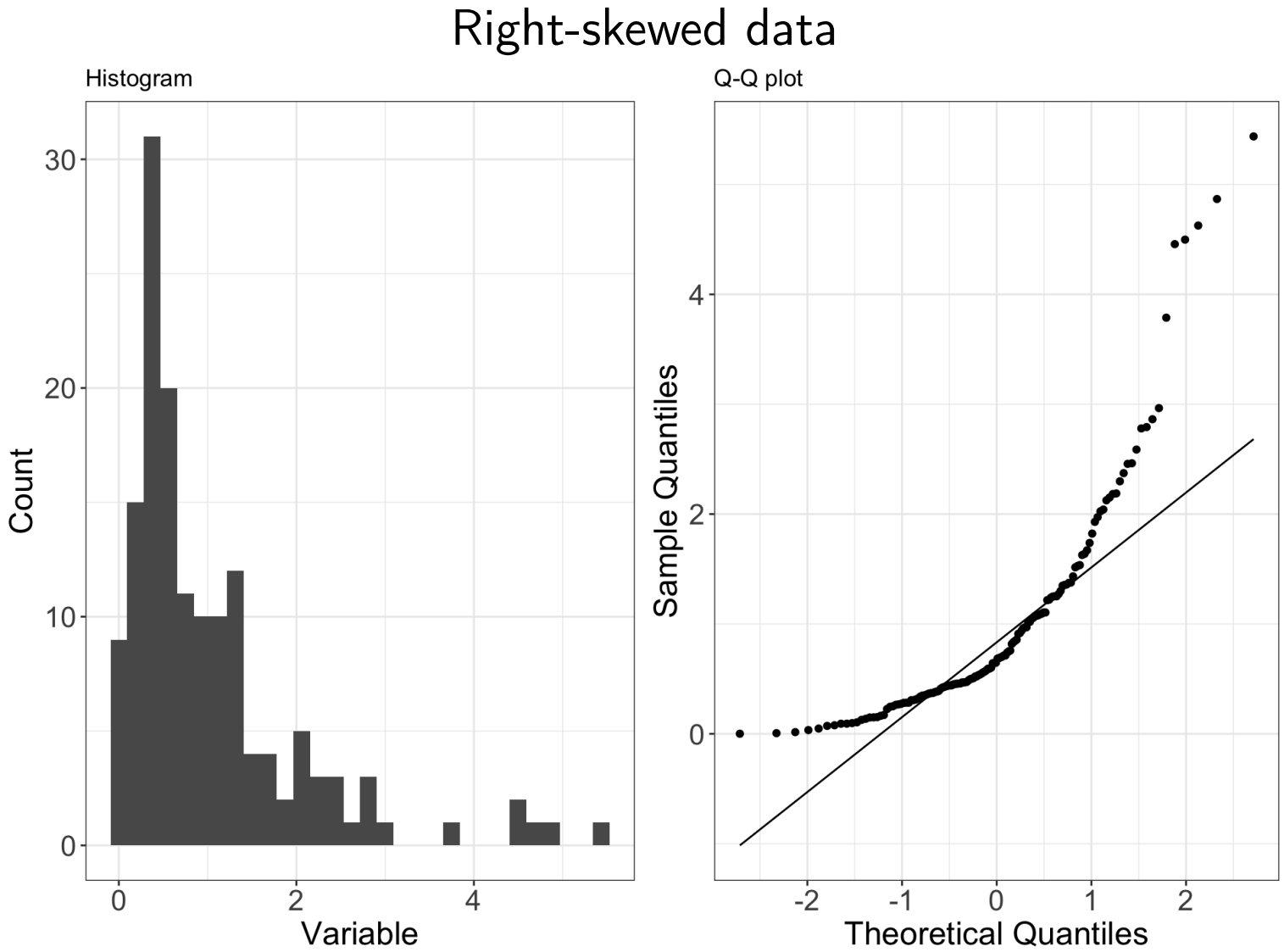


The reference line can also be a linear regression of y_q on x_q or line $y = x$

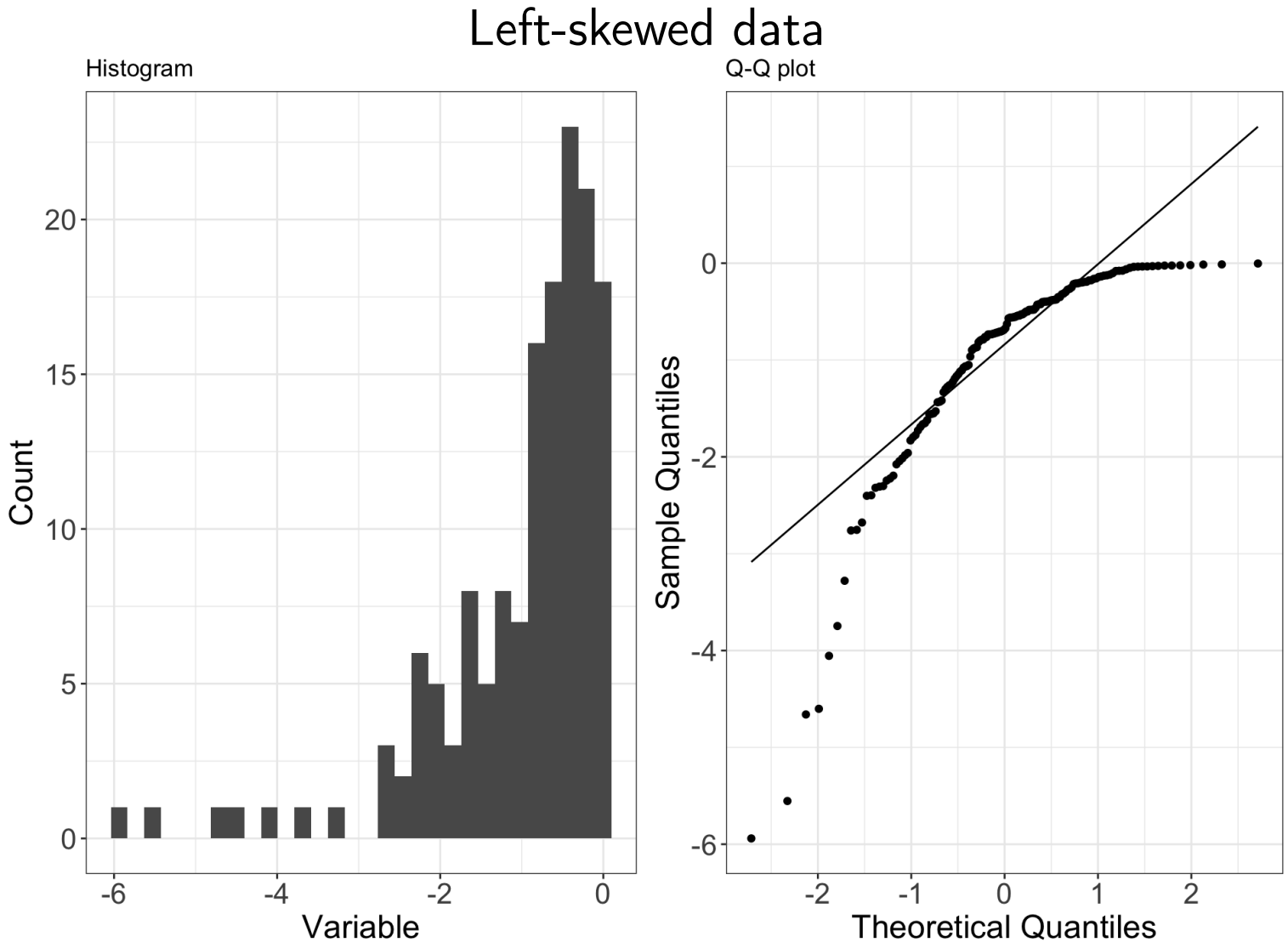
Normal Probability Plot. Illustration 1



Normal Probability Plot. Illustration 2

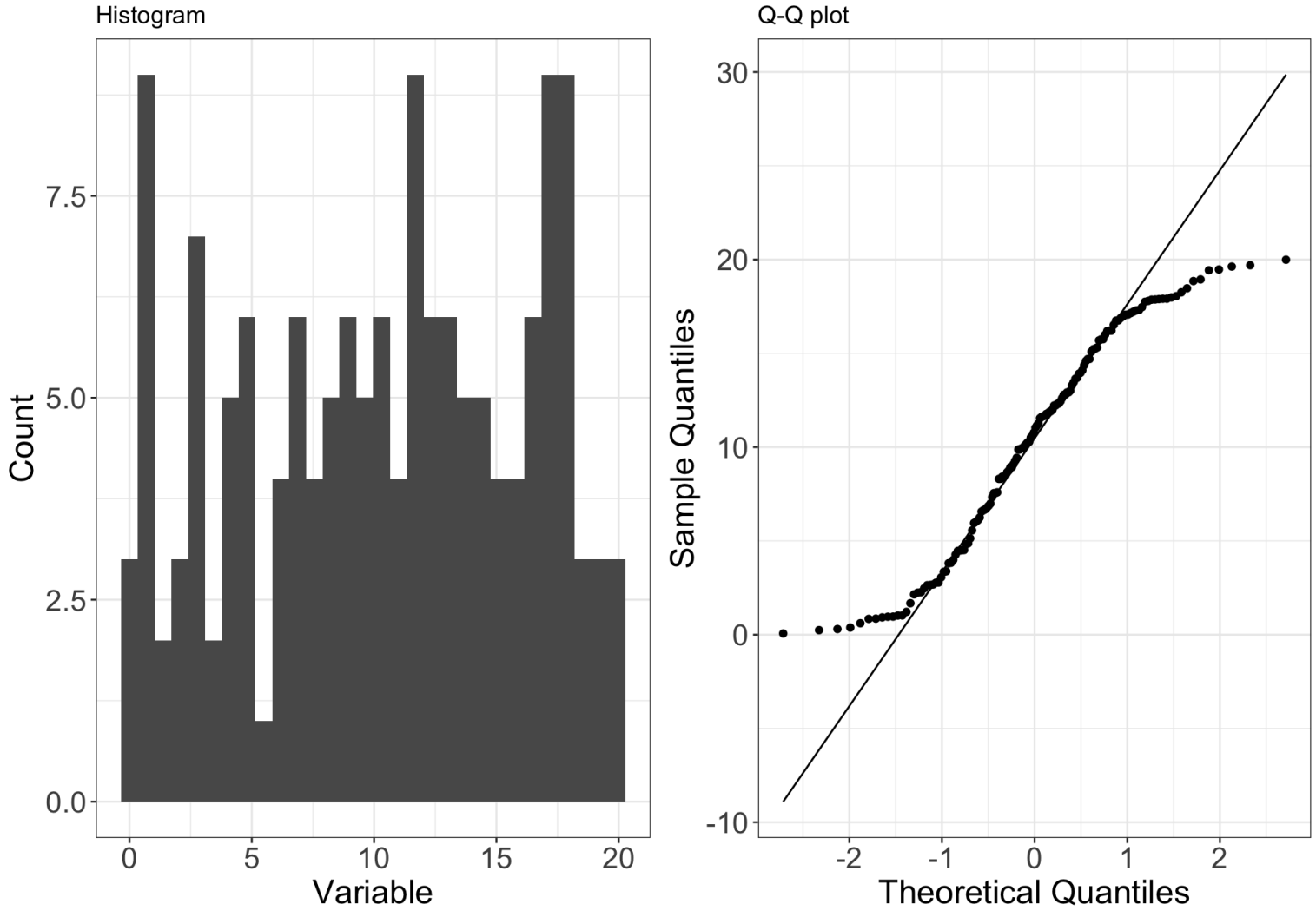


Normal Probability Plot. Illustration 3



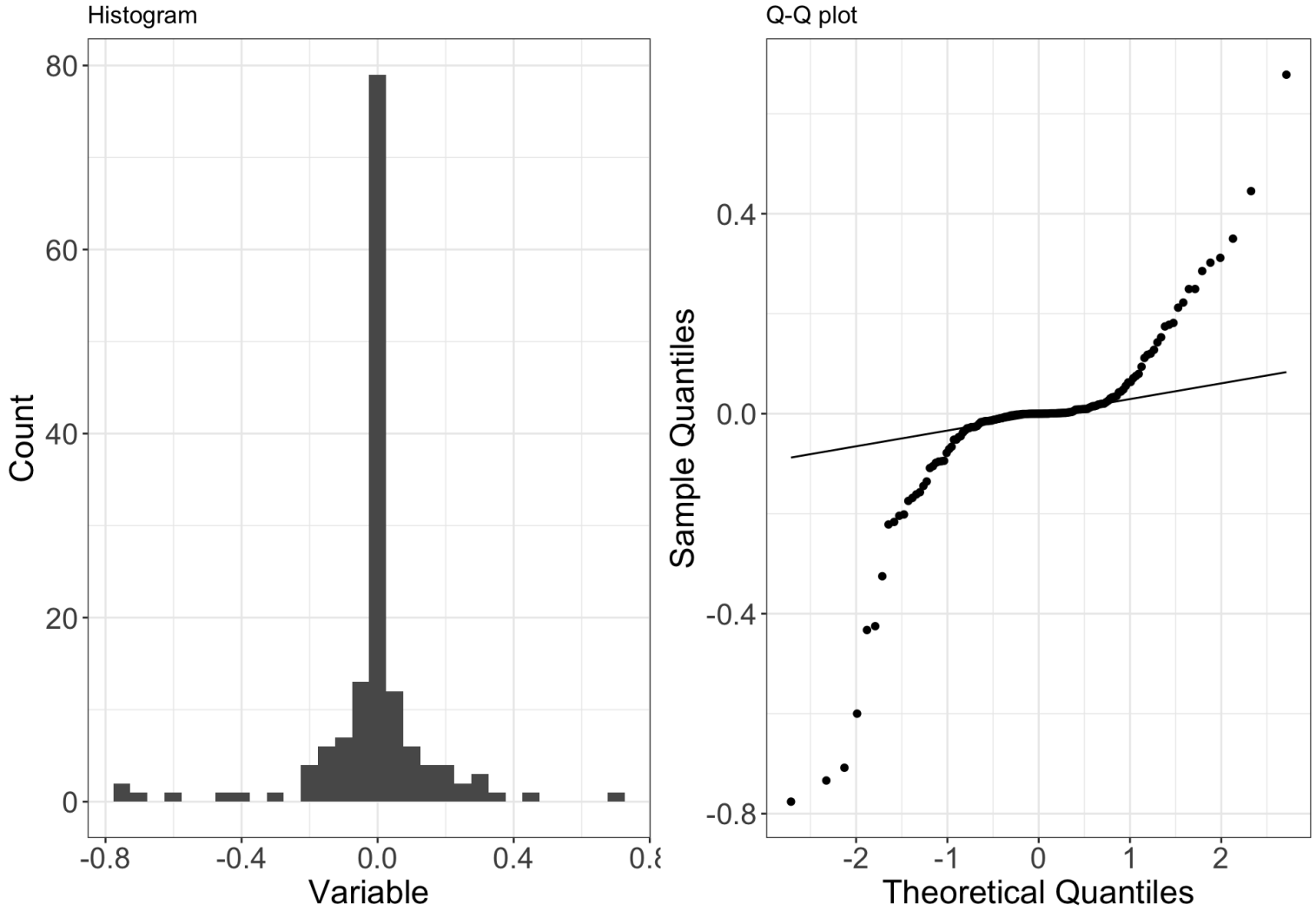
Normal Probability Plot. Illustration 4

Under-dispersed data (thin tails)



Normal Probability Plot. Illustration 5

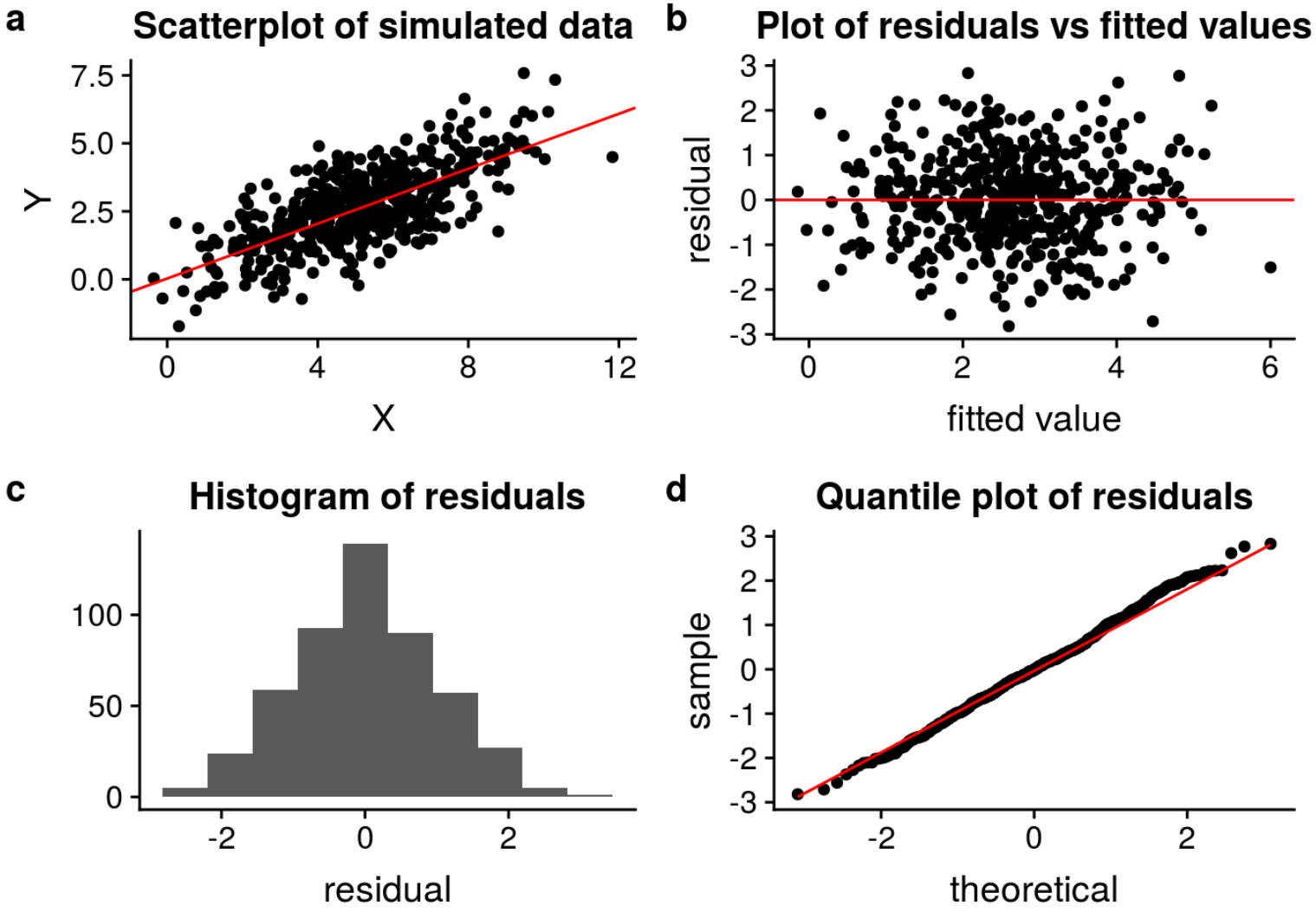
Over-dispersed data (fat tails)



Normal Probability Plot. Interpretation

- Each point on the Q-Q plot corresponds to a certain quantile coming from both distributions
- If the distributions being compared are similar, the points in the Q-Q plot will approximately lie on the line $y = x$
- If the distributions are linearly related, the points in the Q-Q plot will approximately lie on a line, but not necessarily on the line $y = x$
- The points below the reference line to the left (or above the reference line to the right) suggest a heavier tail (more outliers) than a normal distribution
- Flatter reference line indicates that the sample distribution has fat tails and positive kurtosis. Steeper reference line indicates that the sample distribution has thin tails and negative kurtosis
- Outliers in Q-Q plot correspond to the outliers in the sample

Residual Graphs. Illustration



Regression Outliers

Regression outlier is an observation with large residual, i.e. it has an unusual value of the outcome Y, conditioned on the values of the exploratory variables $x_1, ..., x_k$

Graphical diagnostics:

- Standartized residual plots
- Box plots

Measures of outlierness:

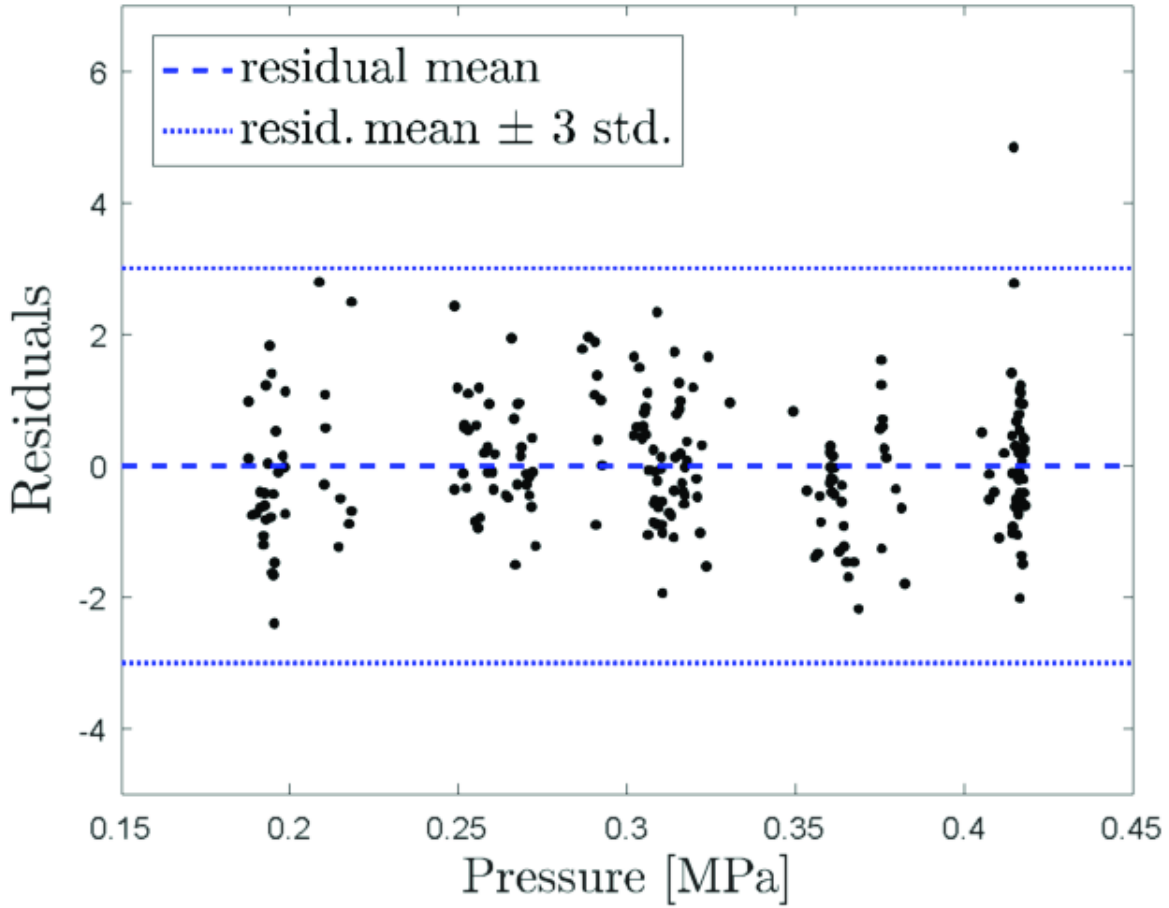
- Cook's distance
- DFBETAS, DFFITS
- ...

Solutions:

- Drop/treat outliers
- Robust regression techniques

Standartized Residual Plot

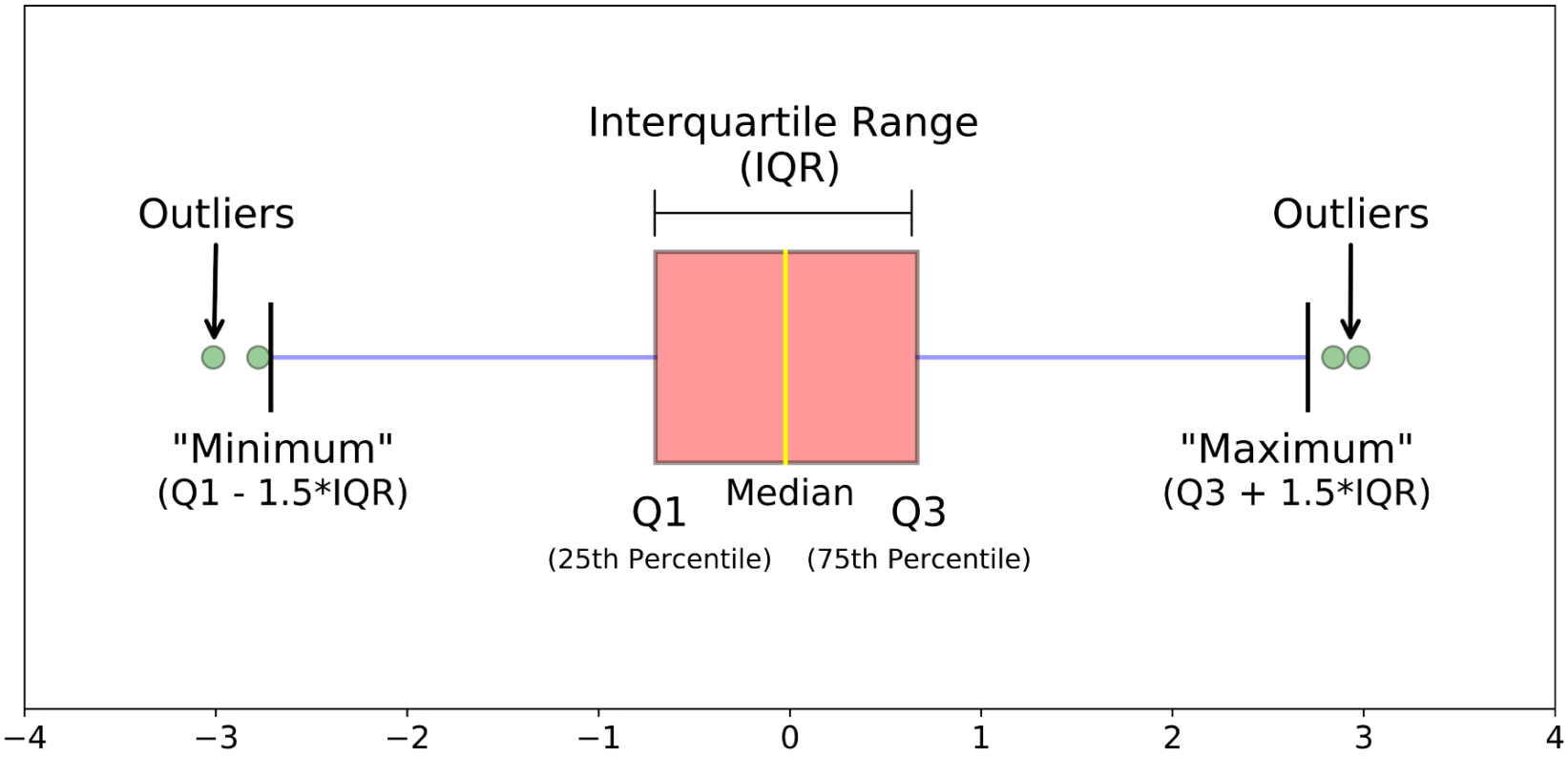
The standartized residuals have the distribution $T(n-k-1)$



Outliers can have a large residual value, but not necessarily affect the estimated slope or intercept

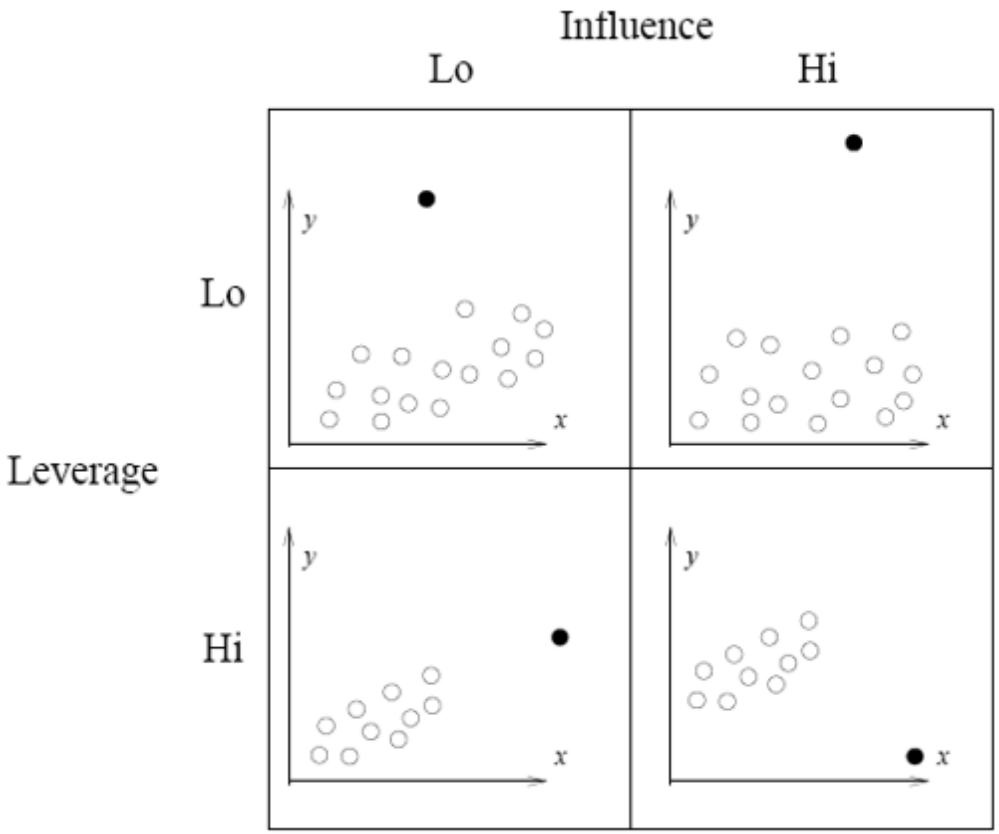
Box Plot

Box plot (or box-and-whisker diagram) is a standardized way of displaying the distribution of data based on five statistics ("minimum", first quartile (Q1), median, third quartile (Q3), and "maximum")



Leverage and Influence

Observations with extreme values of predictors have high leverage



High leverage points not necessarily affect the estimated slope or intercept

Cook's Distance

An observation is said to be influential if removing the observation substantially changes the estimate of coefficients

The Cook's distance of i-th observation:

$$D_i = \frac{\sum_{l=1}^{n} \left(\hat{y}_l - \hat{y}_{l,\sim i}\right)^2}{(k+1)S_e^2}, \quad i = \overline{1,n}$$

where $\hat{y}_l$ is the l-th fitted response value, $\hat{y}_{l,\sim i}$ is the l-th fitted response value where the fit does not include i-th observation

It can be shown that the Cook's distance

$$D_i = \frac{\hat{\varepsilon}_i^2}{(k+1)S_e^2} \frac{h_i}{(1-h_i)^2}, \quad i = \overline{1,n}$$

where $\hat{\varepsilon}_i$ is a residual, h_i is a leverage if i-th observation

Cook's distance D_i shows the influence of i-th observation on the fitted response values

Cook's Distance. Illustration

An observation with Cook's distance larger than a threshold D_0 might be an outlier

The recommended threshold $D_0 = 3\bar{D}$, $\bar{D} = \sum_{i=1}^{n} D_i$

Multicollinearity

Definition

The explanatory variables $x_1, ...x_k$ are called multicollinear if they are linearly related:

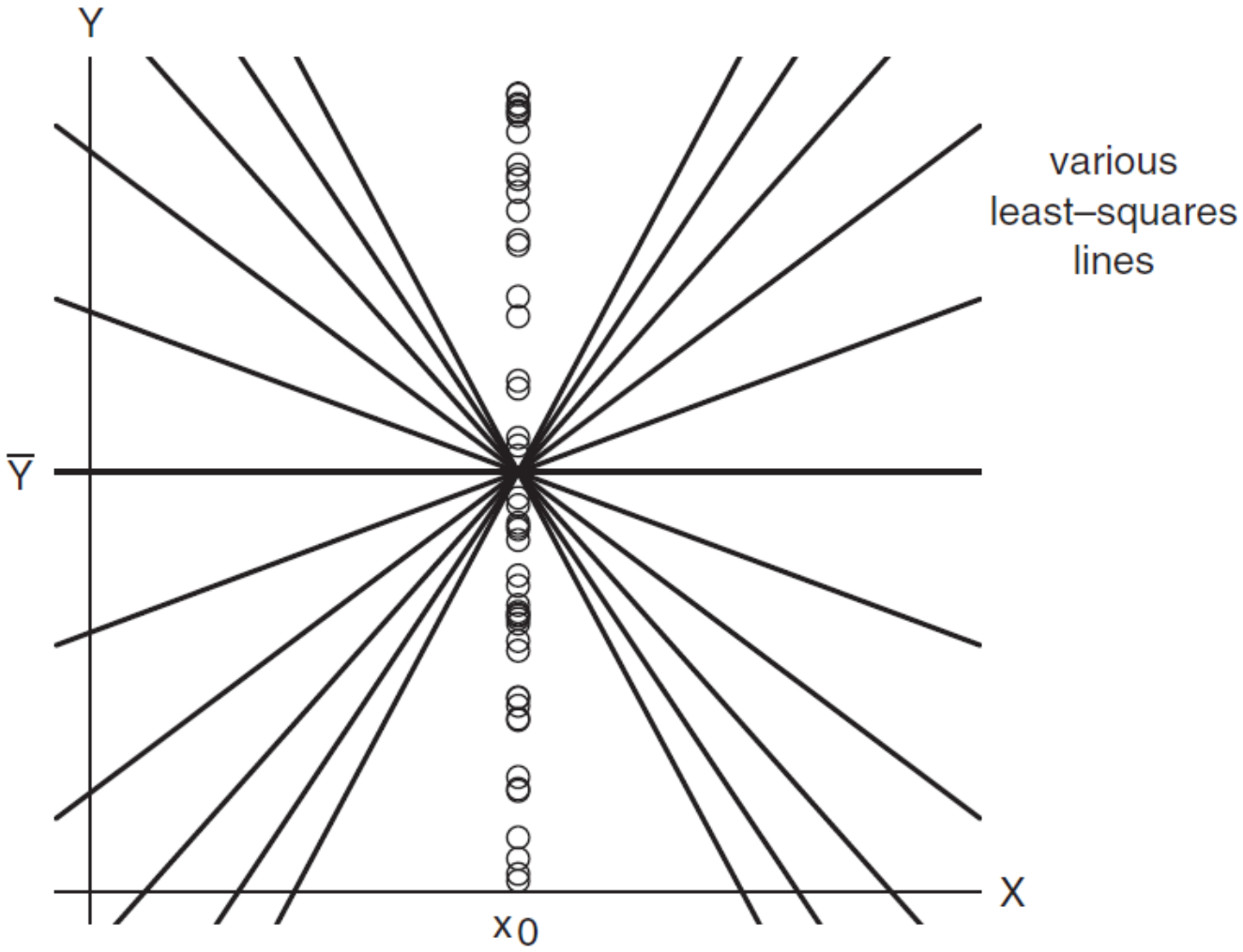
$$c_0 + c_1 x_{1i} + ... + c_k x_{ki} = 0 \quad \forall i = \overline{1, n}$$

where $c_0, c_1, ..., c_k$ are constants and $c_1^2 + ... + c_k^2 > 0$

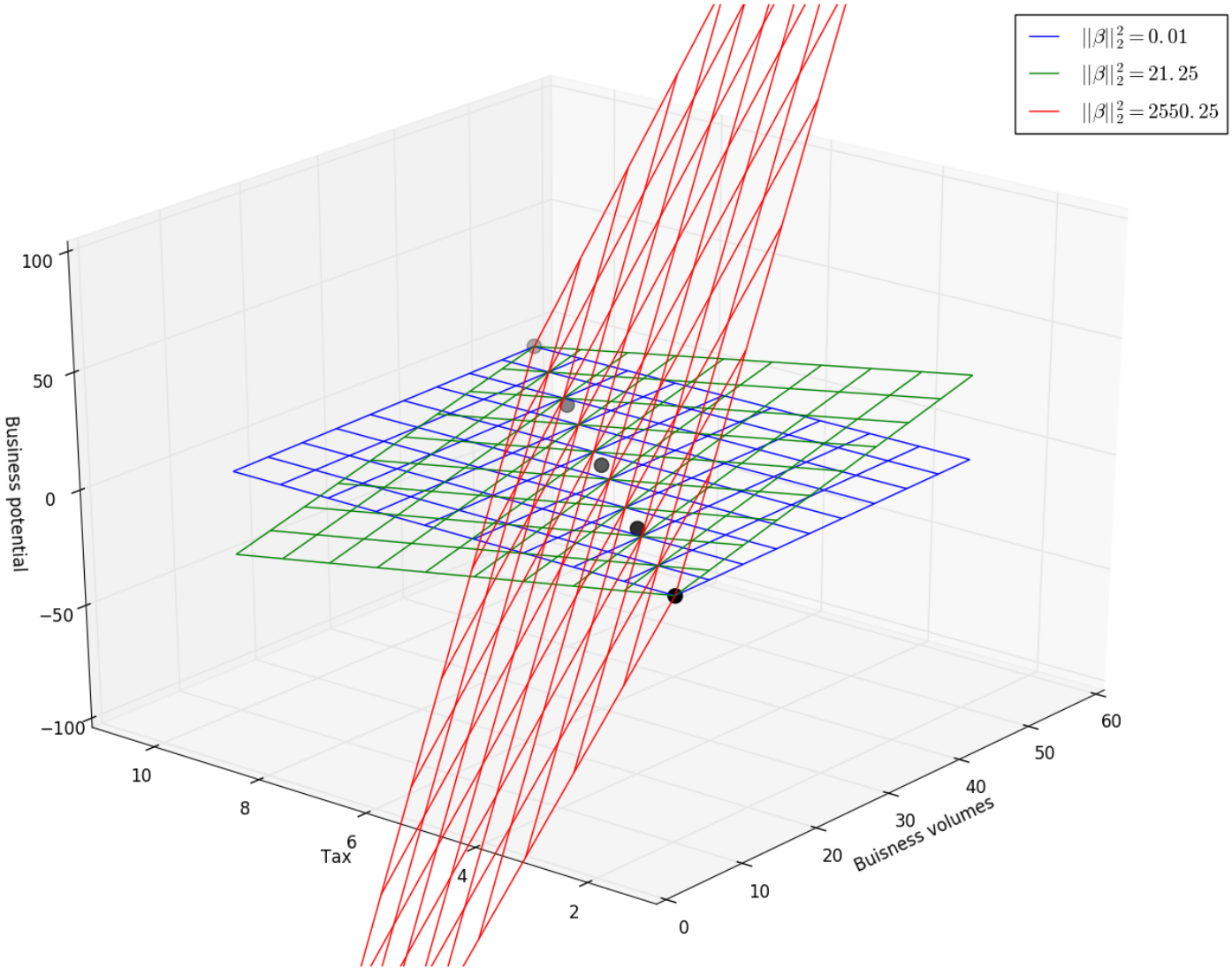
The perfect multicollinearity is rare in practice, but high multicollinearity results in loss of statistical resolution:

- Large standardized residuals
- Broad confidence intervals
- Low t-statistics values, high p-values in hypotheses tests
- Enormous sensitivity to small changes in data and model specification

Variability of Regressor. Illustration



Multicollinearity. Illustration



Multicollinearity Problem

Multiple linear regression model:

$$Y|x = x\beta + \varepsilon(x)$$

OLS-estimates:

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

where X is design matrix:

$$X = \begin{pmatrix} 1 & x_{11} & ... & x_{k1} \\ ... & ... & ... & ... \\ 1 & x_{1n} & ... & x_{kn} \end{pmatrix}$$

If $x_1, ...x_k$ are perfectly multicollinear, then at least one of the columns of X is a linear combination of the others,

$$\text{rank}(X^T X) = \text{rank}(X) < k + 1$$

and the matrix $X^T X$ is not invertible

Nearly Multicollinearity

The variables $x_1, ...x_k$ are nearly multicollinear if

$$c_0 + c_1 x_{1i} + ... + c_k x_{ki} + \nu_i = 0 \quad \forall i = \overline{1, n}$$

where $c_0, c_1, ..., c_k$ are constants, $c_1^2 + ... + c_k^2 > 0$, and ν_i is a random noise

For nearly multicollinear regressors the matrix $X^T X$ has an inverse, but it is ill-conditioned, so that numerical inversion algorithms may be unstable, i.e. a small change in matrix elements results in a large change in approximated inverse matrix

The condition number of matrix A:

$$\kappa(A) = \frac{|\lambda_{\max}|}{|\lambda_{\min}|}$$

where $\lambda_{\min}$, $\lambda_{\max}$ are minimal and maximal eigenvalues of A

$\kappa(A) \gg 0 \Leftrightarrow A$ is ill-conditioned

Multicollinearity Diagnostics

Graphical diagnostics:

- Scatter plots

Quantitative diagnostics:

- Pair-wise correlations between explanatory variables, R^2
- Condition number of $X^T X$
- Measures of multicollinearity (VIF, tolerance, etc.)
- Farrar-Glauber test (F-G test)
- ...

Solutions:

- Transformation of explanatory variables (PCA, ICA, etc.)
- Regularized regression techniques
- Partial least squares (PLS) regression

Variance Inflation Factor

Idea: if the multiple linear regression of X_j, $j = \overline{1,k}$, on all other explanatory variables has high coefficient of determination R_j^2, then there is multicollinearity between $x_1, ..., x_k$

Variance inflation factor (VIF) of X_j:

$$VIF_j = \frac{1}{1 - R_j^2}, \quad j = \overline{1,k}$$

where R_j^2 is the coefficient of determination of multiple regression model:

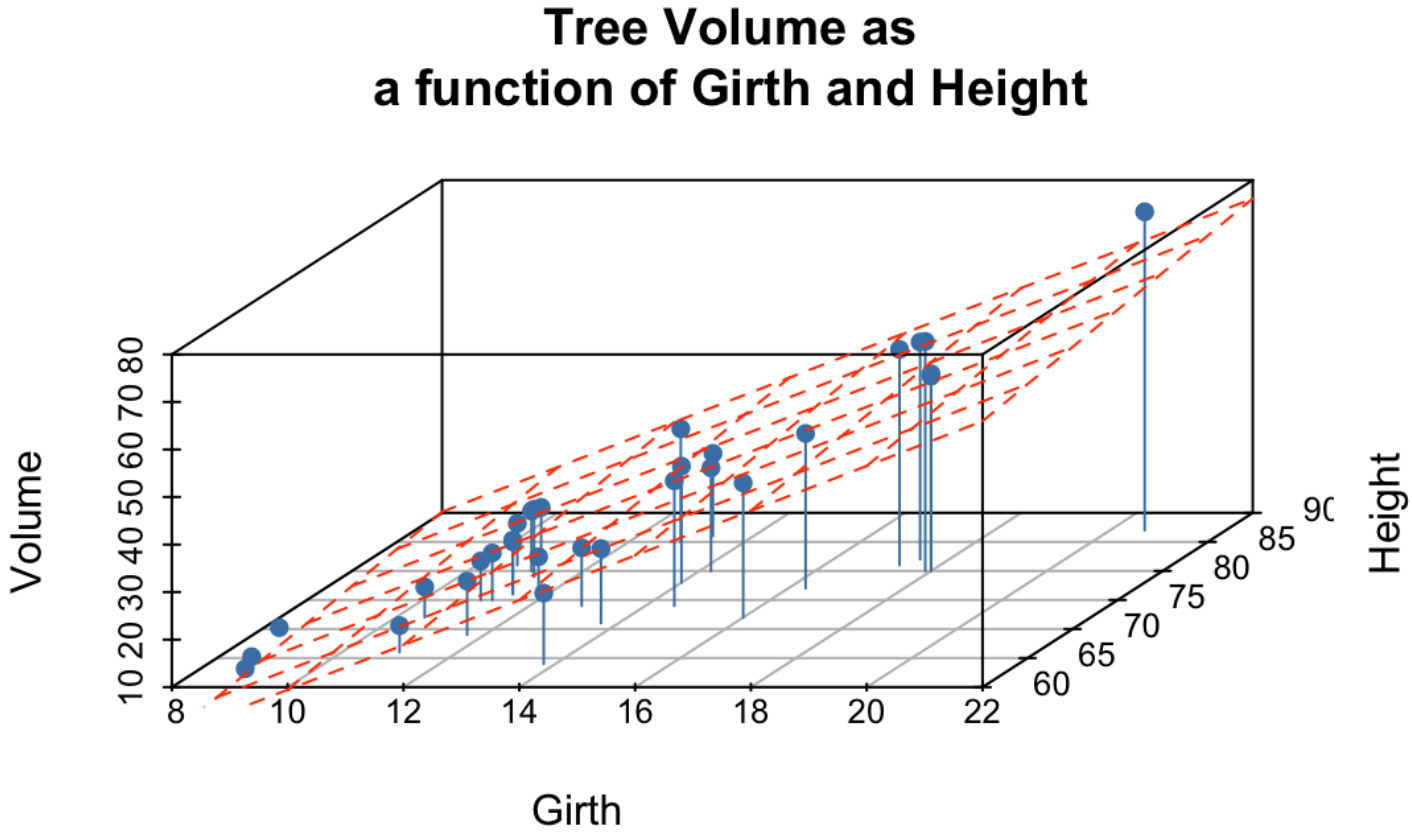
$$X_j | x_{\sim j} = x_{\sim j}\beta_{\sim j} + \varepsilon(x_{\sim j})$$

where $x_{\sim j}$ is the vector of regressors with excluded x_j

High values of any VIF_j, $j = \overline{1,k}$, indicate the multicollinearity ($VIF_j \gtrsim 5$ is considered as severe multicollinearity)

Pair-Wise Correlations vs Multicollinearity

Pair-wise correlations between the explanatory variables may be considered as the sufficient, but not the necessary condition for the multicollinearity



Regularized Regression

If there is multicollinearity between explanatory variables, the training of multiple regression model is ill-posed problem, the OLS method is unstable

The possible solution is to regularize the objective function to make the optimal solution unique:

$$E'(\beta) = E(\beta) + r(\beta) = \frac{1}{n}\sum_{i=1}^{n}(y_i - x_i\beta)^2 + r(\beta)$$

where $r(\beta)$ is a regularizer (doesn't depend on data)

Types of regularizers:

- L_2 regularizer
- L_1 regularizer

L_2 Regularization

Linear regression model:

$$Y|x = x\beta + \varepsilon(x)$$

L_2-regularized criterion:

$$E(\beta) = \sum_{i=1}^{n}(y_i - \varphi(x_i, \beta))^2 + \mu \sum_{j=0}^{k} \beta_j^2 = \sum_{i=1}^{n}(y_i - x_i\beta)^2 + \mu\beta^T\beta$$

where $\mu > 0$ is a regularization parameter

The system of normal equations:

$$(X^TX + \mu I)\beta = X^Ty$$

OLS-estimates:

$$\hat{\beta} = (X^TX + \mu I)^{-1}X^Ty$$

Ridge Regression. Notes

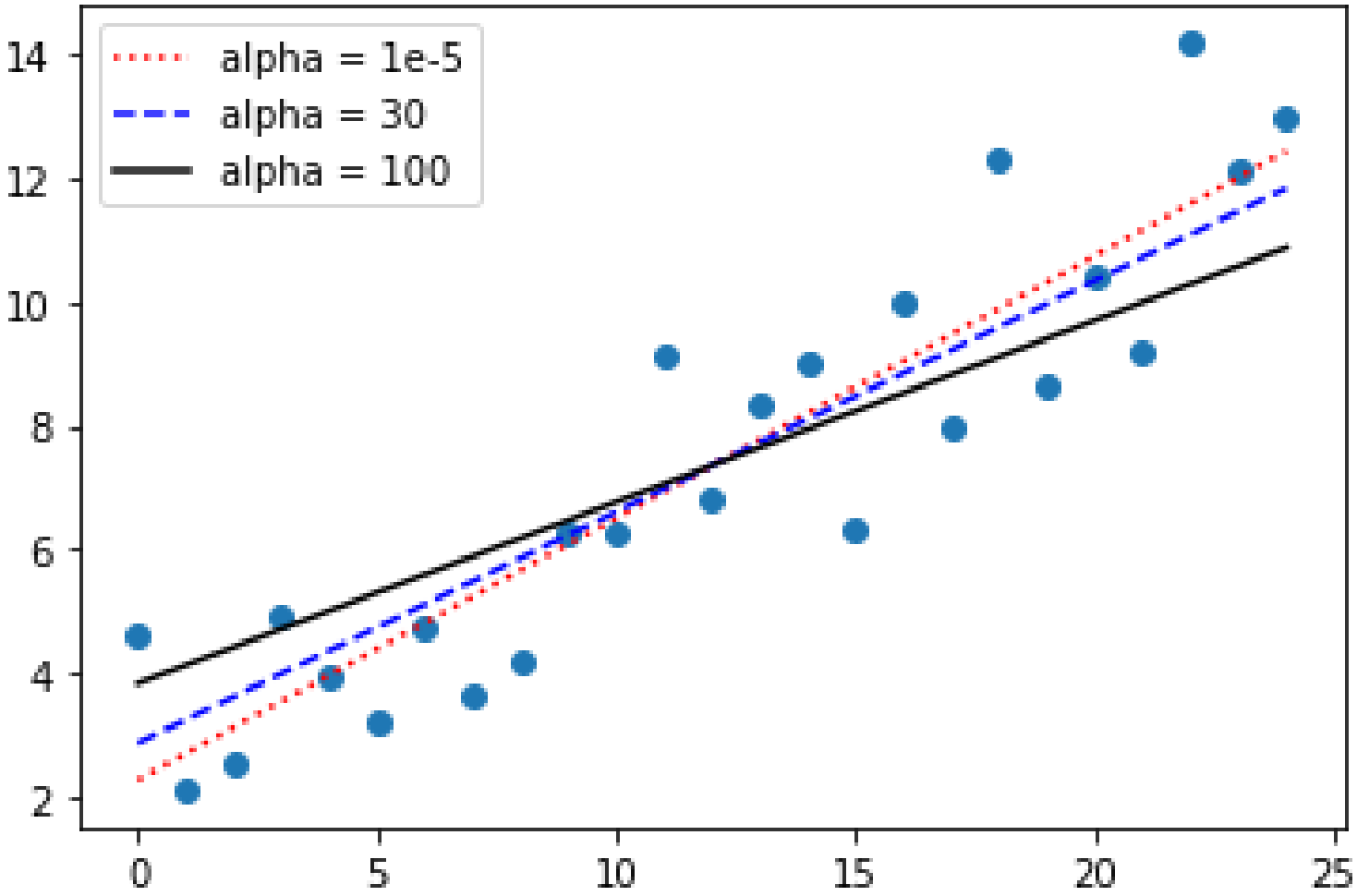
The regression trained with according to L_2-regularized criterion is called as ridge regression

- The estimates $\hat{\beta}$ are biased, the statistical inference is usually not considered, and the assumption of normality of the residuals is no longer necessary
- The ridge parameter μ manages bias-variance trade-off of regression model:
  μ increases $\Rightarrow$ bias increases, variance decreases
  μ decreases $\Rightarrow$ bias decreases, variance increases
- For high μ the regression function becomes approximately constant:

$$\mu \gg 0 \;\Rightarrow\; \varphi(x) \approx \bar{y}$$

- The regularization is usually not applied on the intercept β_0
- Cross-validation can be used in choosing μ: select μ that yields the smallest cross-validation prediction error

Ridge Regression. Illustration



Coefficient Scaling in Ridge Regression

The intercept β_0 can be excluded from the regression model by centering and scaling:

$$Y'|x' = x'\beta' + \varepsilon'(x')$$

where $x' = (x'_1, ..., x'_k)$, $\beta' = (\beta'_1, ..., \beta'_k)^T$ and

$$x'_{ji} = \frac{x_{ji} - \bar{x}_j}{s_j}, \quad i = \overline{1,n}, \quad j = \overline{1,k}$$

$$y'_i = y_i - \bar{y}, \quad i = \overline{1,n}$$

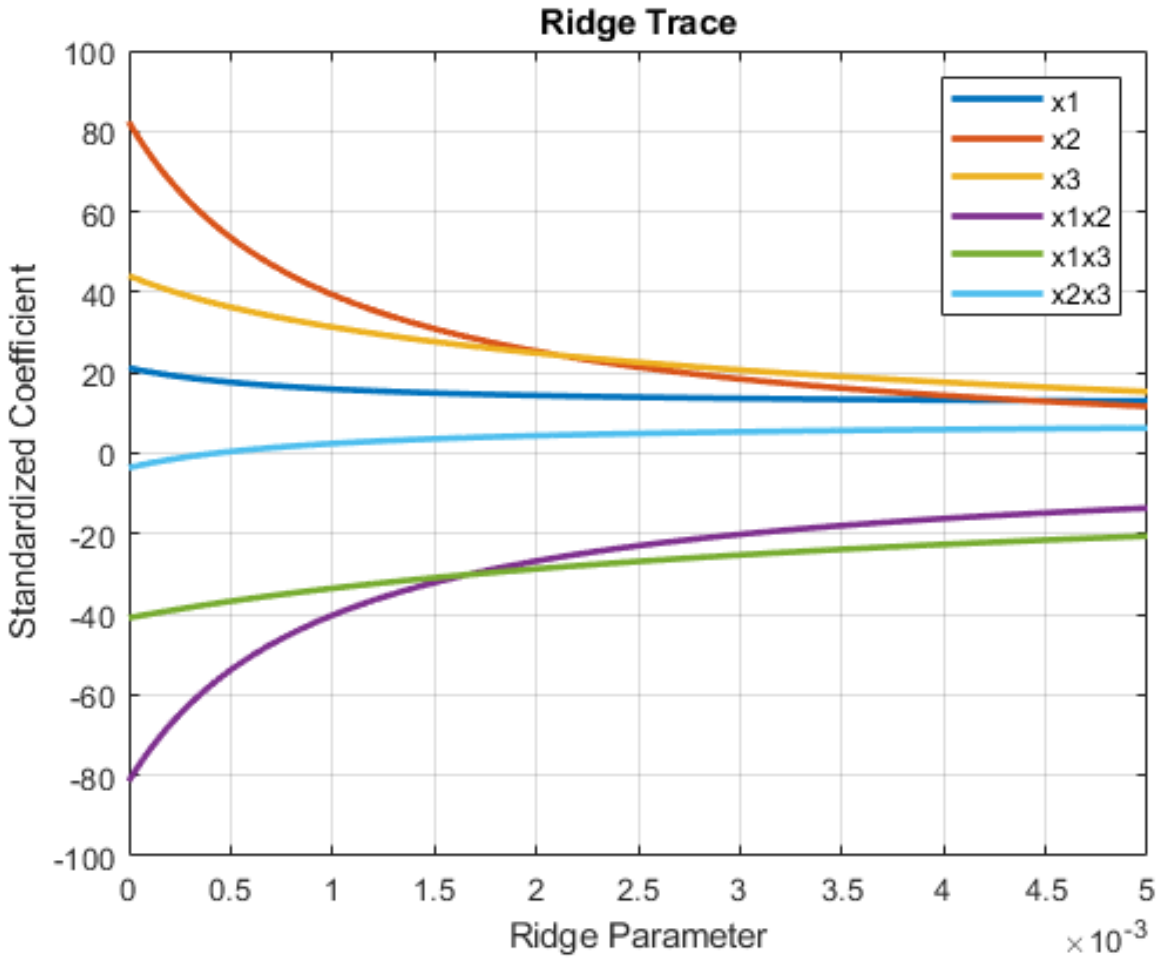
The original coefficients:

$$\beta_0 = \bar{y} - \sum_{j=1}^{k} \frac{\beta'_j \bar{x}_j}{s_j}, \quad \beta_j = \frac{\beta'_j}{s_j}, \quad j = \overline{1,k}$$

The coefficients $\beta'_1, ..., \beta'_k$ are comparable, they are useful for visual analysis of regressors' "importance" on ridge trace plot

Ridge Trace Plot

Ridge trace plot is a graph of estimates $\hat{\beta}'_1(\mu), ..., \hat{\beta}'_k(\mu)$ as functions of the ridge parameter μ



L_1 Regularization

Linear regression model:

$$Y|x = x\beta + \varepsilon(x)$$

L_1-regularized criterion:

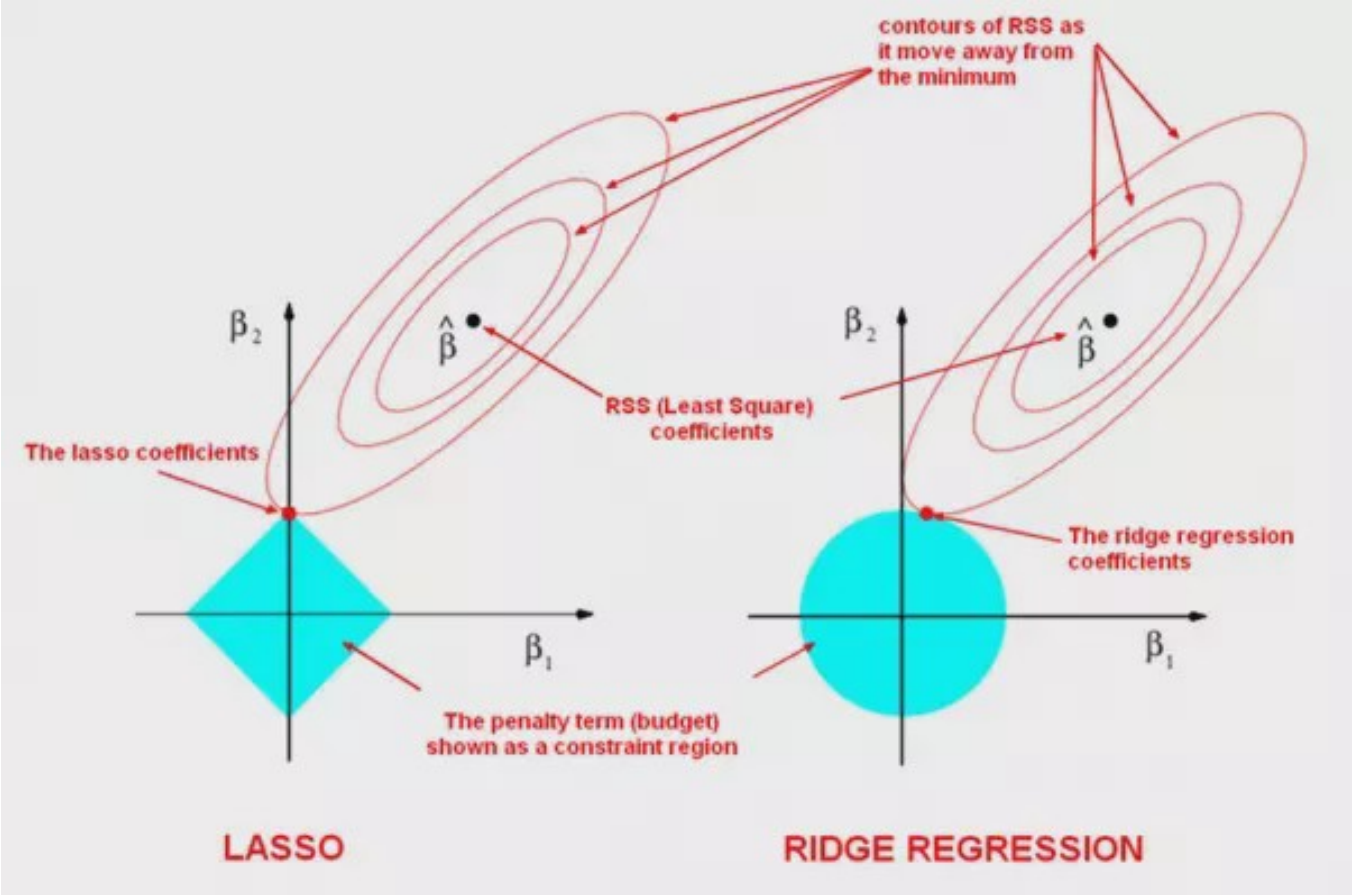
$$E(\beta) = \frac{1}{2}\sum_{i=1}^{n}(y_i - \varphi(x_i, \beta))^2 + \mu\sum_{j=0}^{k}|\beta_j|$$

where $\mu > 0$ is a regularization parameter

The regression trained with according to L_1-regularized criterion is called as LASSO (Least Absolute Shrinkage and Selection Operator) regression

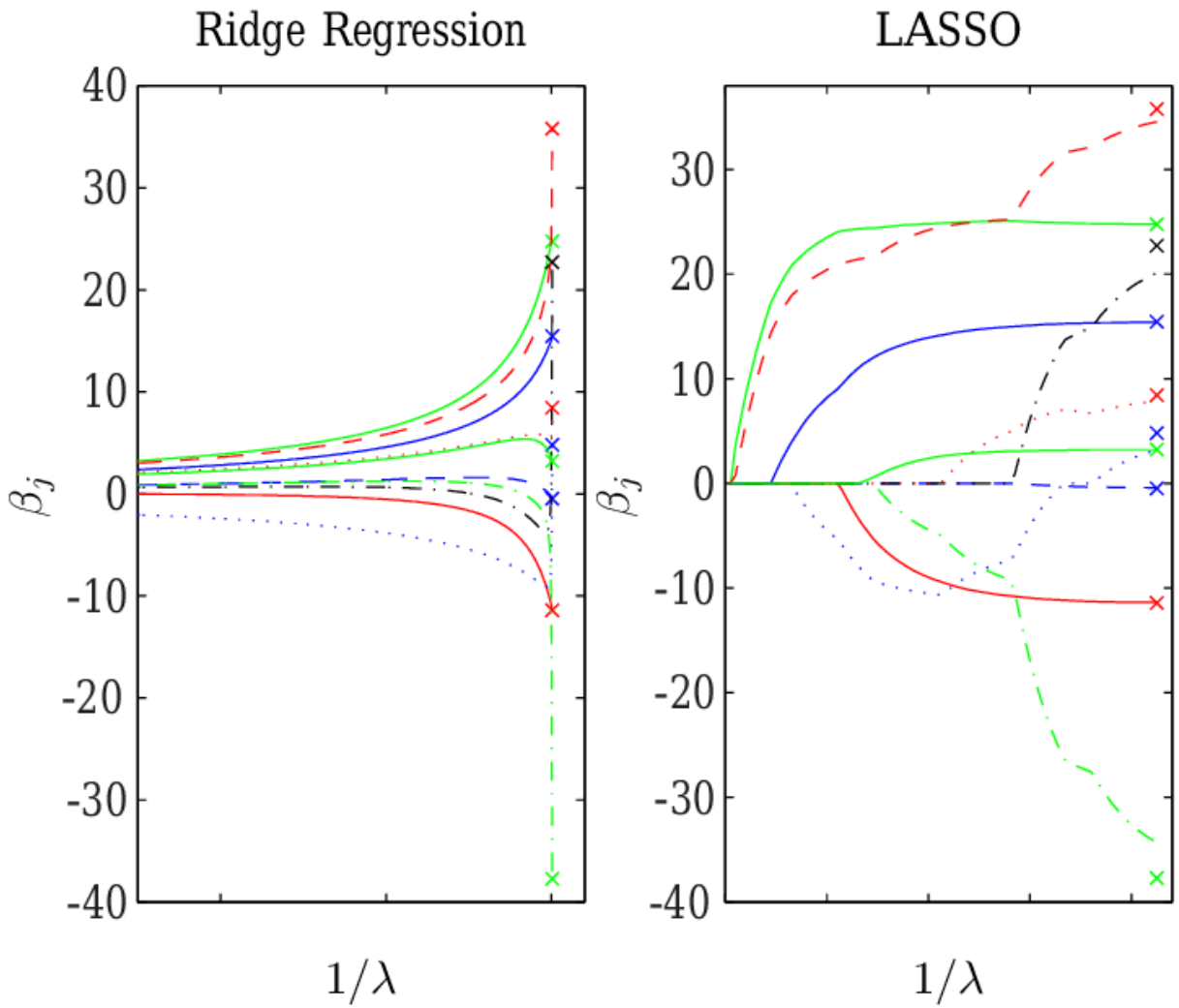
LASSO regression does not have a closed-form solution, iterative approach is required for training (based on descend methods, etc.)

L_2 vs L_1 Regularization. Illustration



L_1 regularization tends to concentrate the regression parameters in a relatively small number of high-important parameters, while others are driven toward zero

Ridge vs LASSO Trace Plots



LASSO approach is useful for feature selection

Elastic Net Regression

Linear regression model:

$$Y|x = x\beta + \varepsilon(x)$$

Elastic net criterion:

$$E(\beta) = \frac{1}{2}\sum_{i=1}^{n}(y_i - \varphi(x_i, \beta))^2 + \mu\sum_{j=0}^{k}\left(\frac{1-\alpha}{2}\beta_j^2 + \alpha|\beta_j|\right)$$

where $\mu > 0$ is a regularization parameter, α is a ridge/LASSO ratio, $0 < \alpha < 1$

Elastic net regression is a hybrid of ridge regression and LASSO regression. Like LASSO, elastic net can generate reduced models by generating zero-valued coefficients

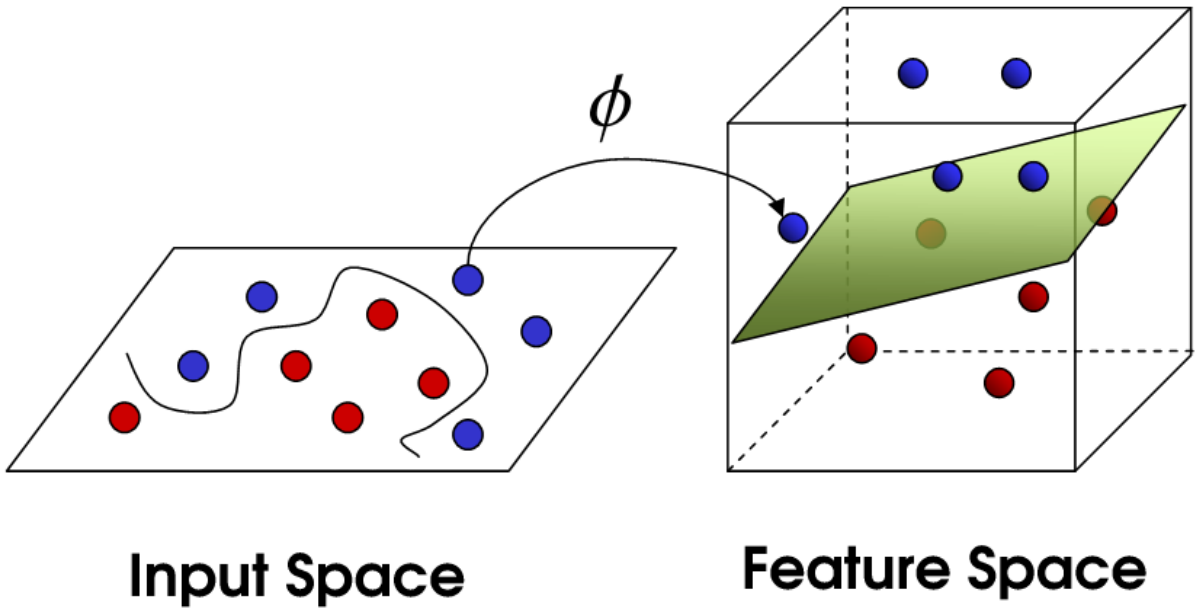
Empirical studies have suggested that the elastic net technique can outperform LASSO on data with highly correlated predictors

Feature Selection

Feature selection is the process of selecting a subset of relevant features for use in model construction

Why select features?

- To improve model prediction performance
- To provide faster training
- To provide a better interpretation of the trained model



Model Misspecification

Model misspecification refers to all of the ways that the regression model might fail to represent the true underlying model (or data generating process)

Types of model misspecification:

- Under-specification
  Omitted explanatory variables may cause omitted variable bias in OLS estimates
- Over-specification
  Redundant explanatory variables may cause multicollinearity and wrong statistical inferences
- Functional form misspecification
  Model has the appropriate explanatory variables, but the functional relationship is wrongly specified

The assumptions of regression analysis may be violated for misspecified model

Adding a New Variable into the Model

Suppose we have a regression model $Y|x = \beta_0 + \beta_1 x + \varepsilon(x)$ and observations for some new variable z

Should this variable z be included into the model?

- z is related to both x and y
  It's reasonable to include z and then solve possible multicollinearity
- z is unrelated to x but related to y
  Adding z will reduce residual variance, it should be included into the model
- z is related to x but unrelated to y
  Adding z to the regression won't reduce OVB and residual variance, moreover can cause multicollinearity
- z is unrelated to both x and y
  It doesn't matter much whether you include it or exclude it

Adding a New Variable into the Model

What happens when a new variable is added to the model?

- The estimate $\hat{\beta}_1$ changes a lot
  z is related to x and y so it should be included to avoid OVB and then we need to solve possible multicollinearity
- The estimate $\hat{\beta}_1$ doesn't change
  z is unrelated to x or unrelated to y or both
  - $\mathrm{D}[\hat{\beta}_1]$ is increased
    z causes multicollinearity, it should be excluded
  - $\mathrm{D}[\hat{\beta}_1]$ is decreased
    z and y are related. Regardless including z doesn't decrease OVB, it should be included to decrease estimation errors
  - $\mathrm{D}[\hat{\beta}_1]$ is not affected
    z is unrelated to x and unrelated to y, it's not a problem by omitting z

Feature Selection Approaches

- Filter methods
  Assign a scoring to each feature, usually univariate and consider the features independently. The selection is performed before the training algorithm
- Wrapper methods
  Manipulate with features (add or remove them) during the training process. The selection criterion directly measures the change in model performance that results from adding or removing the features
- Embedded methods
  Learn feature importance as part of the model training process. Once you train a model, you obtain the importance of the features in the trained model

Sequential Feature Selection

The goal: to find the feature subset that minimizes the prediction error

If the number k of features is small, the exhaustive feature selection can be applied, but it's infeasible for large k

Sequential feature selection is a type of greedy search algorithms

For sequential feature selection we need:

- Objective function (criterion)
  Measure of the quality of the model and, hence, of the feature subset used (usually, MSE)
- Search algorithm
  How to add or remove features from a feature subset

Filter methods: the criterion is independent of the training process

Wrapper methods: the criterion is related to machine learning model and loss function used in training

Sequential Feature Selection Approaches

- Sequential forward selection
  Features are sequentially added to an empty candidate set until the addition of further features does not decrease the criterion
- Sequential backward selection
  Features are sequentially removed from a full candidate set until the removal of further features increase the criterion
- Sequential forward floating selection
  Features are sequentially added, but can also be removed at some steps
- Sequential backward floating selection
  Features are sequentially removed, but can also be included at some steps

Stepwise Regression

Stepwise regression is a type of sequential floating selection for regression model

Forward stepwise regression: we start from the simplest model

Backward stepwise regression: we start from the complex model with all possible regressors

Stepwise regression algorithm

- **Step 1.** Fit the initial model
- **Step 2.** Add the most significant feature into the model, if any. Repeat the step until there are no significant features to add
- **Step 3.** Remove the most insignificant feature from the model, if any, and go to step 2. If there are no insignificant features, end the process

What is the measure of feature significance for the model?

F-test to Compare Two Models

Suppose we have two set of features $\mathscr{X}_0$ and $\mathscr{X}_1$ and the corresponding trained regression models

Do these models differ significantly?

Statistical hypothesis: $H_0 : R_1^2 = R_0^2$

Test statistic:

$$Z = \frac{(D^*_{res0} - D^*_{res1})/(k_1 - k_0)}{D^*_{res1}/(n - k_1 - 1)}, \quad Z|H_0 \sim F(k_1 - k_0, n - k_1 - 1)$$

where k_0, k_1 are the numbers of explanatory variables in the models, and D^*_{res0}, D^*_{res1} are residual variances of the models:

$$D^*_{res} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \varphi(x_i, \beta))^2$$

$p > \alpha \Rightarrow H_0$ accepted, the difference is insignificant
$p \leq \alpha \Rightarrow H_0$ rejected, the difference is significant

Stepwise Regression. Notes

- Stepwise feature selection approach can be used with OLS, WLS, GLS. For robust regressions the F-test is no longer valid
- It's recommended to check and remove outliers before stepwise regression
- Stepwise regression finds suboptimal subset of features. The global optimum of objective function is not guaranteed
- Stepwise selection uses many repeated hypothesis tests to make decisions on the inclusion or exclusion of individual predictors. It leads to inflation of false positive findings, the suboptimality of the subset of features may be violated
- The validation sample should be used to estimate the quality of the subset of features at each iteration
- Stepwise regression encounters a lot of criticism*

*Harrell F. Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis. New York: Springer, 2015.

Stepwise Regression. Example

```
1. Adding x4, FStat = 22.7985, pValue = 0.000576232
2. Adding x1, FStat = 108.2239, pValue = 1.105281e-06
3. Adding x2, FStat = 5.0259, pValue = 0.051687
4. Removing x4, FStat = 1.8633, pValue = 0.2054

mdl =
Linear regression model:
    y ~ 1 + x1 + x2

Estimated Coefficients:
                   Estimate       SE       tStat       pValue
                   ________    ________    ______    __________

    (Intercept)      52.577      2.2862    22.998    5.4566e-10
    x1               1.4683      0.1213    12.105    2.6922e-07
    x2              0.66225    0.045855    14.442     5.029e-08


Number of observations: 13, Error degrees of freedom: 10
Root Mean Squared Error: 2.41
R-squared: 0.979,  Adjusted R-Squared: 0.974
F-statistic vs. constant model: 230, p-value = 4.41e-09
```