Statistics and Data Analysis

Regression Analysis (3)

Ling-Chieh Kung

Department of Information Management
National Taiwan University

Introduction

- When doing regression:
  - We try to discover the hidden relationship among variables.
  - We assume a specific model

    $$y = \beta_0 + \beta_1 x_1 + \cdots + \epsilon$$

    and then fit our sample data to the model.
  - We validate our model based on the **degree of fitness** (R^2 and R^2_{adj}) and **significance of variables** (p-values).
- If our model is good, the random error ϵ should be really "random."
  - There should be no **systematic pattern** for ϵ.
- We need **residual analysis**.

Residual analysis

- **Residual analysis**.
- Case study: bike rentals.

Residuals

- Consider a pair of variables x and y.
- We may assume a linear relationship

$$y = \beta_0 + \beta_1 x + \epsilon$$

  for some unknown parameters β_0 and β_1. ϵ is the random error.
- Four assumptions on the random error:
  - **Zero mean**: The expected value of ϵ is zero for any value of x.
  - **Constant variance**: The variance of ϵ is the same for any value of x.
  - **Independence**: ϵ for different values of x should be independent.
  - **Normality**: ϵ is normal for any value of x.
- Once we obtain a regression model, we need to test these assumptions.
  - To predict: We need the first three.
  - To explain: We need all the four.

Testing the four assumptions

- Consider a sample data set $\{(x_i, y_i)\}_{i=1,\ldots,n}$.
- Linear regression helps us find $\hat{\beta}_0$ and $\hat{\beta}_1$ based on the sample data and obtain the regression formula

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \epsilon_i,$$

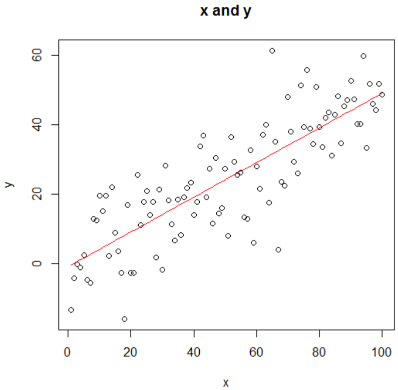
  in which the error term ϵ_i is called the **residual** between our estimate $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ and the real value y_i.
- By conducting a **residual analysis**, we check these ϵ_is to see if we have the desired properties.
- While there are rigorous statistical tests, we will only introduce some graphical approaches.

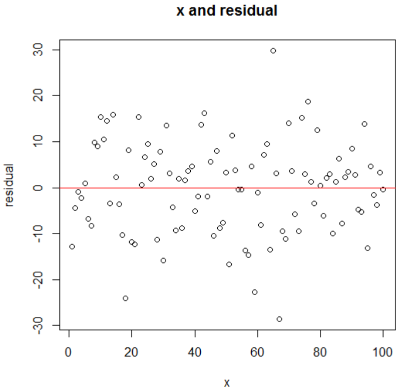
The residual plot and histogram

- We may plot the residuals ϵ_is along with x_is to form a **residual plot**.
  - This tests zero mean, constant variance, and independence.
  - There should be no systematic pattern.
- We may construct a **histogram** of residuals.
  - This tests normality.
  - The histogram should be symmetric and bell-shaped.
- In general:
  - A "good" plot does not guarantee a good model.
  - A "bad" plot **strongly suggests** that the model is bad!

The residual plot and histogram

- Consider the artificial data set as an example.

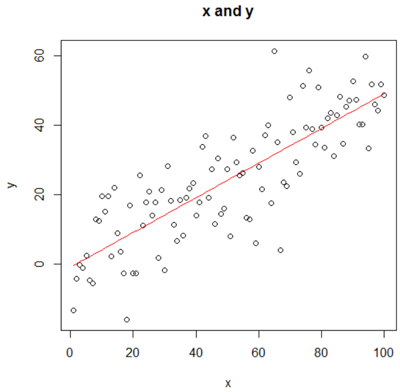


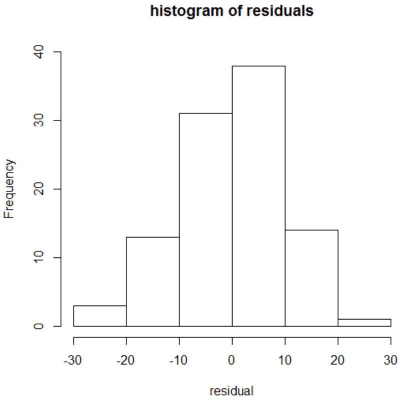


- There is no pattern in the residual plot: good!

The residual plot and histogram

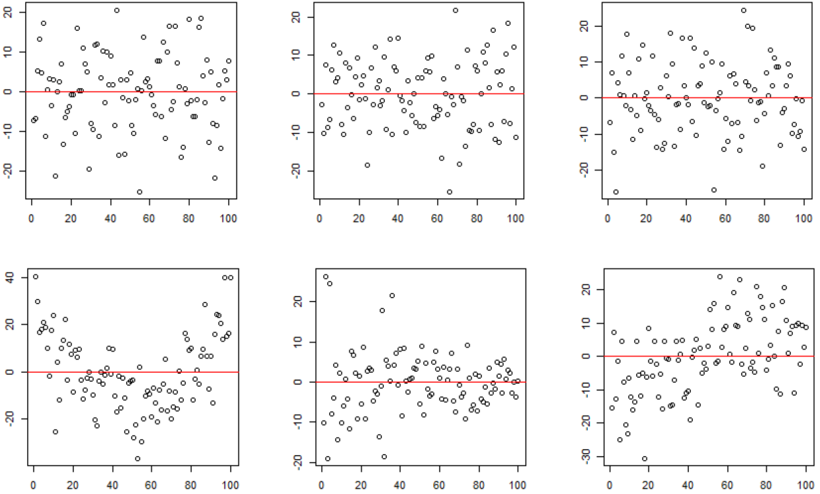
- Consider the artificial data set as an example.



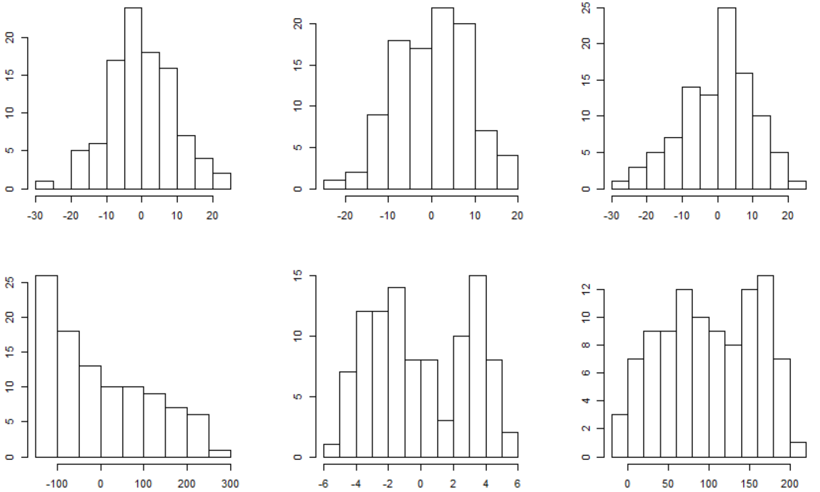


- The histogram is symmetric and bell-shaped: good!

Residual plots that pass and fail the tests



Histograms that pass and fail the tests



Residual analysis for multiple regression

- Suppose that we construct a multiple regression model

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \cdots + \hat{\beta}_p x_p + \epsilon_i.$$

- We still use residual plots and a histogram to test the assumptions.
- **Multiple** residual plots should be depicted.
  - The vertical axis is always for the residuals ϵ_is.
  - The horizontal axis is for a function of $(x_1, x_2, ..., x_p)$.
  - E.g., the kth independent variable x_k along.
  - E.g., the fitted value $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \cdots + \hat{\beta}_p x_p$.

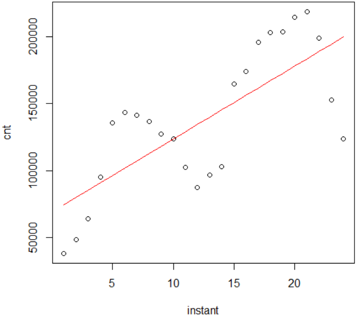
Residual analysis

- Residual analysis.
- **Case study: bike rentals**.

Monthly rentals

- Recall our monthly bike rental example. Our sample data gives us

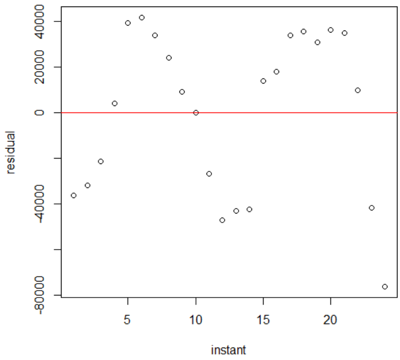
$$cnt_i = 69033 + 5453 instant_i + \epsilon_i.$$

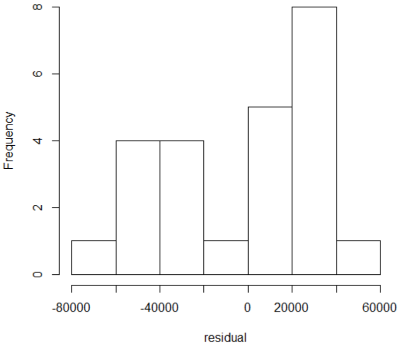


instant	*cnt*	$\hat{y}_i$	ϵ_i
1	38189	74486	−36297
2	48215	79939	−31724
3	64045	85392	−21347
4	94870	90845	4025
5	135821	96298	39523
6	143512	101751	41761
7	141341	107204	34137
8	136691	112657	24034
9	127418	118110	9308
10	123511	123563	−52
11	102167	129016	−26849
12	87323	134469	−47146
13	96744	139922	−43178
14	103137	145375	−42238
	⋮		
23	152664	194452	−41788
24	123713	199905	−76192

Residual analysis reveals poor quality

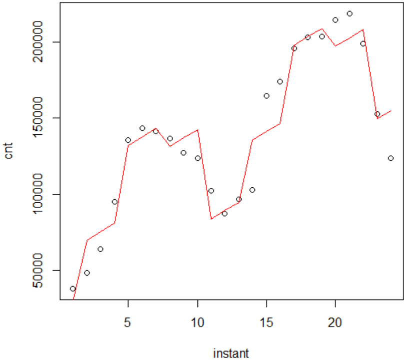
- This simple linear modal $cnt = 69033 + 5453 instant$ is very bad!





Using *instant* plus *month*

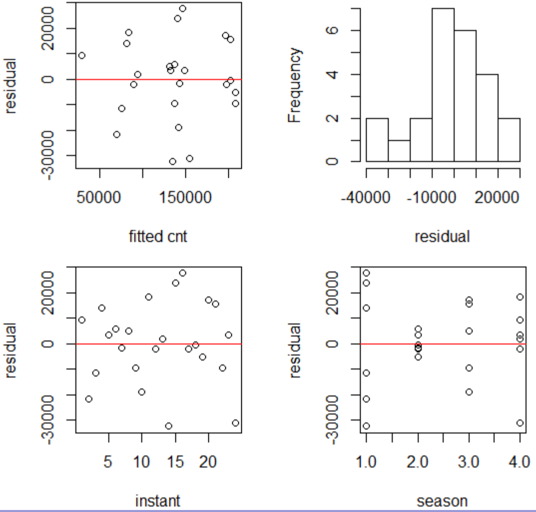
- Let's add *month* into our model.



- This model is better. How about the residuals?

Using *instant* plus *month*

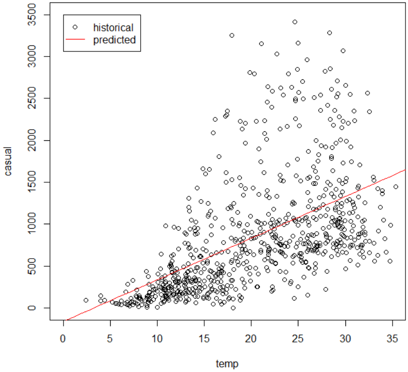
- We may now look at three residual plots.
  - Not perfect, but now much better.
  - There may still be missing factors.
- The histogram is also not perfect.
- This may be due to the **lack of data**.



Daily rentals

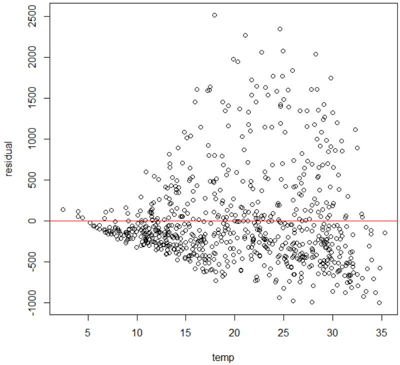
- Recall our daily bike rental example. Our sample data gives us

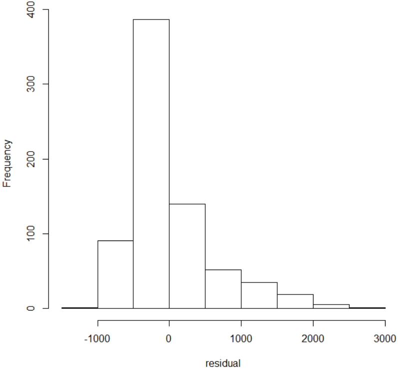
$$casual_i = -161.329 + 49.702 temp_i + \epsilon_i.$$



Residual analysis reveals poor quality

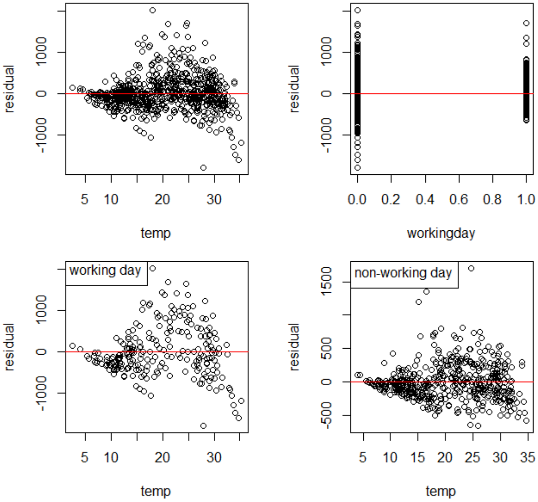
- This simple linear modal $casual = -161.329 + 49.702temp$ is very bad!





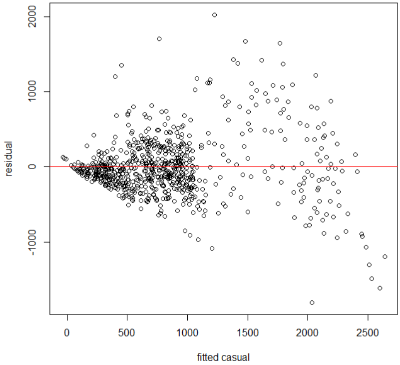
Adding *workingday* and *workingday* × *temp*

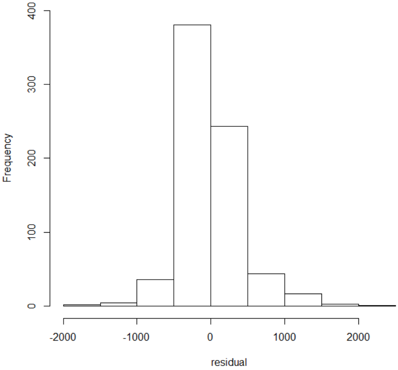
- Let's add *workingday* and *workingday* × *temp* into our model.
- It helps, but does not help too much.



Adding *workingday* and *workingday* $\times$ *temp*

- It helps, but does not help too much.





- May we do better?

Remarks

- When there is a systematic pattern in our residuals, there may be some essential factors missing.
- If we can include most essential factors into our regression model, residuals will be "more random."
  - *instant*?
  - *month*?
  - $temp^2$?
  - Interaction?
- For realistic business problems in practice, it can be hard to get "perfect" residuals.
  - Always try to improve your model.
  - But stop when it is time to make a decision.