
Data Analytics

Yong Zheng

Illinois Institute of Technology
Chicago, IL, 60616, USA



School of Applied Technology
ILLINOIS INSTITUTE OF TECHNOLOGY

Quick Review

- Supervised vs Unsupervised Learning
- Predictive Models, e.g., regression, classification, etc
- Regression Models
 - Simple Linear Regression Model
 - Multiple Linear Regression Model
 - Multivariate Regression Model
- Simple and Multiple Linear Regression Model



Multiple Linear Regression

Consider one Y variable and **k** independent variables X_i , e.g. X_1, X_2, X_3 .

- Data on n tuples $(y_i, x_{i1}, x_{i2}, x_{i3})$.
- Scatter plots show **linear association between Y and the X-variables**
- The observations on y can be assumed to satisfy the following model

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + e_i \text{ for } i = 1, \dots, n$$

Data

Prediction

error



Multiple Linear Regression

Important Steps in Multiple Linear Regression

- Data Splits – build a model based on train set, and evaluate it based on the test set → either hold-out or N-fold evaluations
- Determine linear relationship between y and x variables
- Build a multiple linear regression model by parameter estimates
- Goodness of fit test
- Residual analysis – the last step to tell your model is qualified
- Interpret the performance of the training process
- Evaluations and predictions – evaluate it based on test set

Multiple Linear Regression

How to determine x and y have a linear relationship?

- Draw a scatter plot Y and X-variables
 - Observe whether there is a straight line pattern
- Calculate the correlations
 - Observe the correlation values are close to 1 or -1
 - If the correlation value falls in $[-0.4, 0.4]$, we usually say it is a weak linear relationship
 - If the correlation value falls in $[-0.1, 0.1]$, we usually say there are almost no linear relationship



Simple Linear Regression

- It implies a strong correlation between X and Y.
- The **correlation coefficient** r is the measure of the **linear** association between two variables.

The correlation coefficient is defined as

$$r(X, Y) = \frac{1}{n-1} \sum \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$$

- Where X has average \bar{x} and standard deviation s_x , and Y has average \bar{y} and standard deviation s_y .

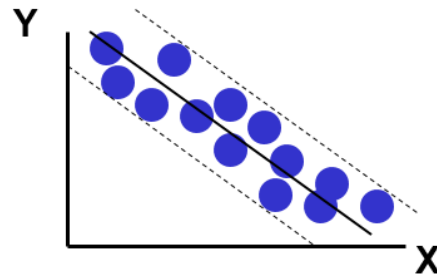
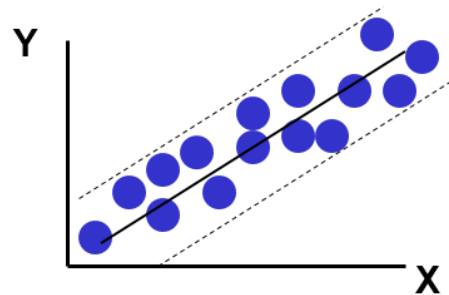
Simple Linear Regression

- It is much more obvious from the visualization

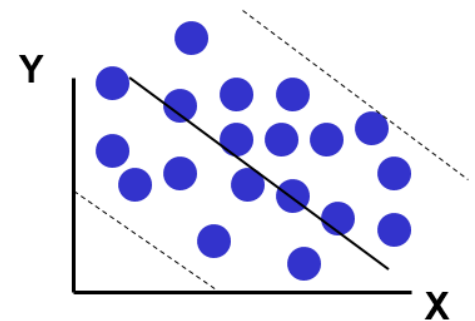
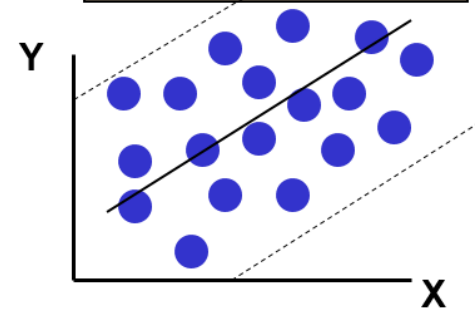
- r lies between -1 and $+1$.
 - $r > 0$ indicates a positive linear relationship.
 - $r < 0$ indicates a negative linear relationship.
 - $r = 0$ indicates no linear relationship.
 - $r = \pm 1$ indicates perfect linear relationship.

- The larger the absolute value of r , the stronger the linear relationship.

Strong relationships



Weak relationships



What if there are no linear relationship?

What if there are no linear relations or not clear linear relations between Y and X?

Solutions:

- You may need to perform transformations on either or both of the Y and X variables. Try X first, then Y.
 - **Square transformation: $X' = X * X$**
 - **Log transformation: $X' = \log X$**
 - **Inversion transformation: $X' = 1/X$**
- After transformation, evaluate the relationship between Y and X'. If they have linear relationship, use X' instead of X



Multiple Linear Regression

Important Steps in Multiple Linear Regression

- Data Splits – build a model based on train set, and evaluate it based on the test set → we use hold-out as example in the class, we will introduce N-fold cross validation later
- Draw scatter plot
- Observe whether you can fit a line to describe the pattern
- Build a multiple linear regression model
- Parameter estimates
- Goodness of fit test
- Residual analysis – the last step to tell your model is qualified
- Evaluations and predictions – evaluate it based on test set



Multiple Linear Regression

- Parameter Estimates
- Goodness of Fit Test
- Residual Analysis
- Evaluations and Predictions



Simple Linear Regression

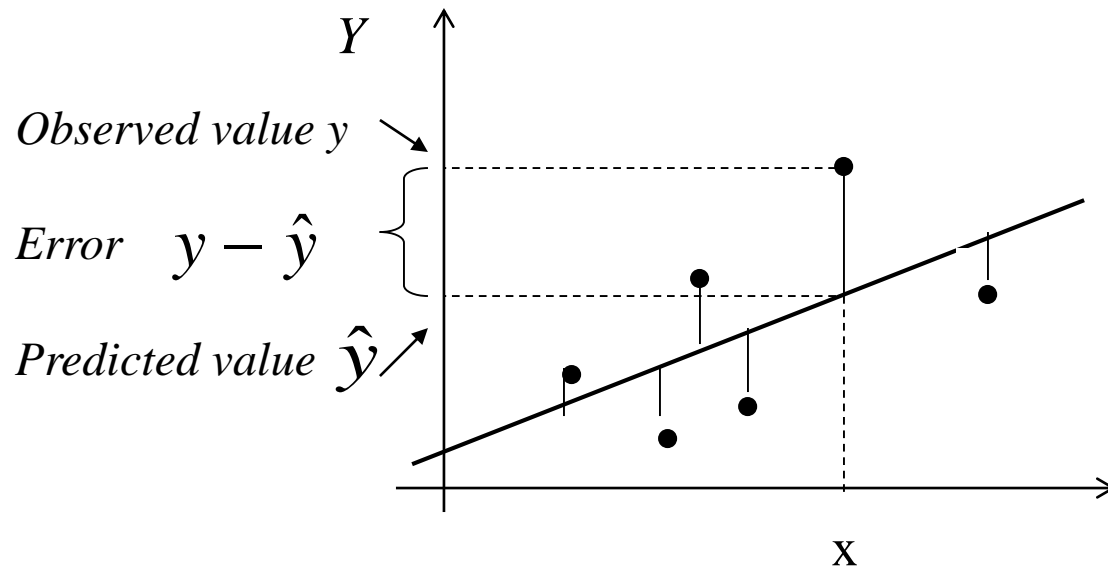
The regression line is used to predict the response \hat{y} at any given x . Regression line **minimizes the vertical distances between observed y and the point on the line**. The accuracy of the prediction depends on how much spread out the observations are around the line.

Error of prediction, $err = y - \hat{y}$

Squared error = err^2

Sum of error (SE) = $\sum err$

Sum of squared errors (SSE) = $\sum err^2$



Multiple Linear Regression

The parameter estimates are those values for β 's that minimize the sum of the square errors: [Least Square Optimization]

$$\sum_i (y_i - \hat{y}_i)^2 = \sum_i [y_i - (\beta_0 + \beta_1 x_1 + \dots + \beta_k x_k)]^2$$

- Thus the parameter estimates $\hat{\beta}$ are those values for β 's that will make the model residuals as small as possible!

The fitted model to compute predictions for Y is

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \dots + \hat{\beta}_k x_k$$



EXAMPLE: CPU usage

A study was conducted to examine what factors affect the CPU usage. A set of 38 processes written in a programming language was considered. For each program, data were collected on the

Y = CPU usage (time) in seconds of time,

X1= the number of lines (linet) in thousands generated by the process execution.

X2 = number of programs (step) forming the process

X3 = number of mounted computer devices (device).

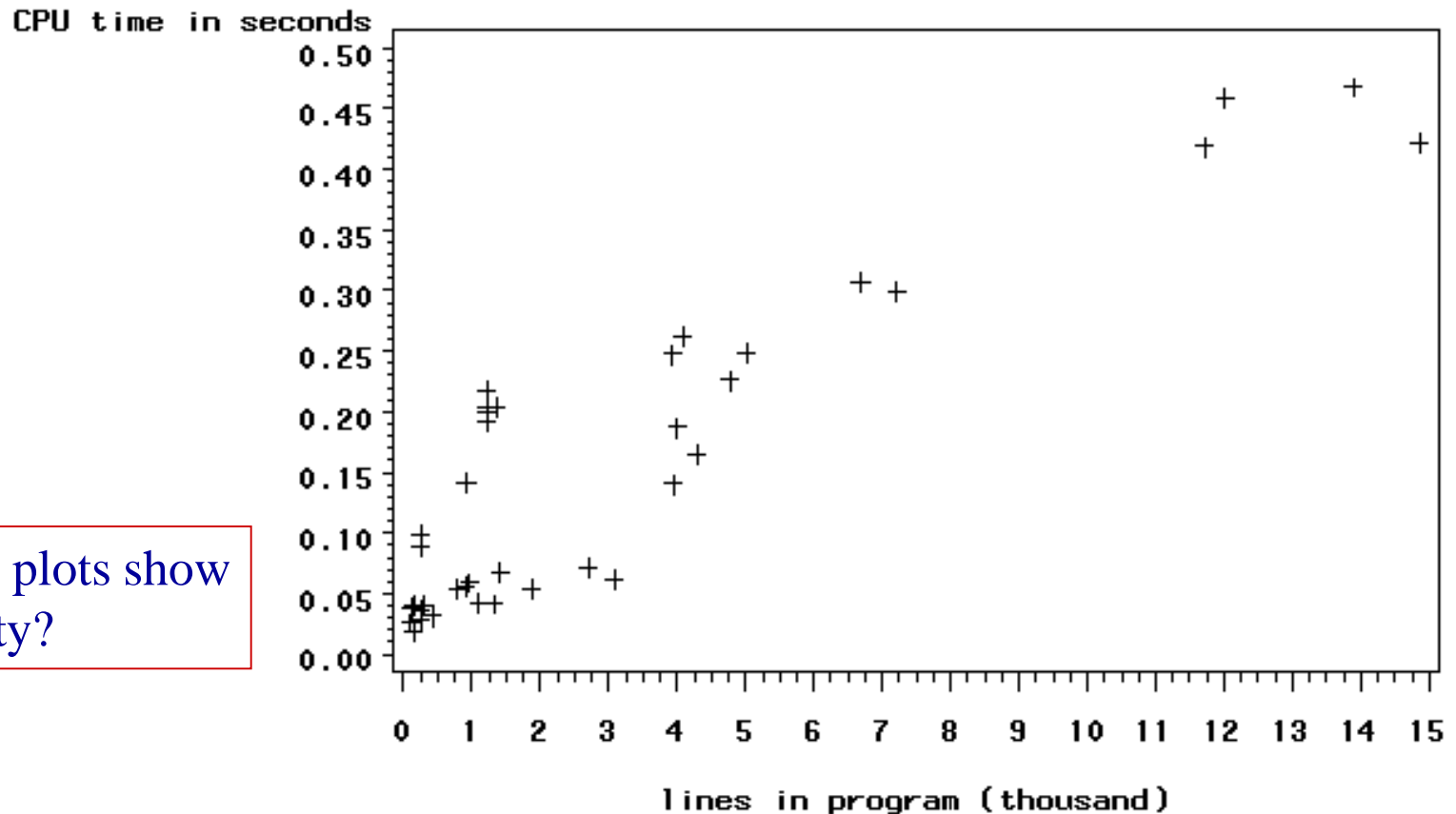
Problem: Estimate the regression model of Y on X1,X2 and X3

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 LINET + \hat{\beta}_2 STEP + \hat{\beta}_3 DEVICE$$



I) Exploratory data step: Are the associations between Y and the x-variables linear?

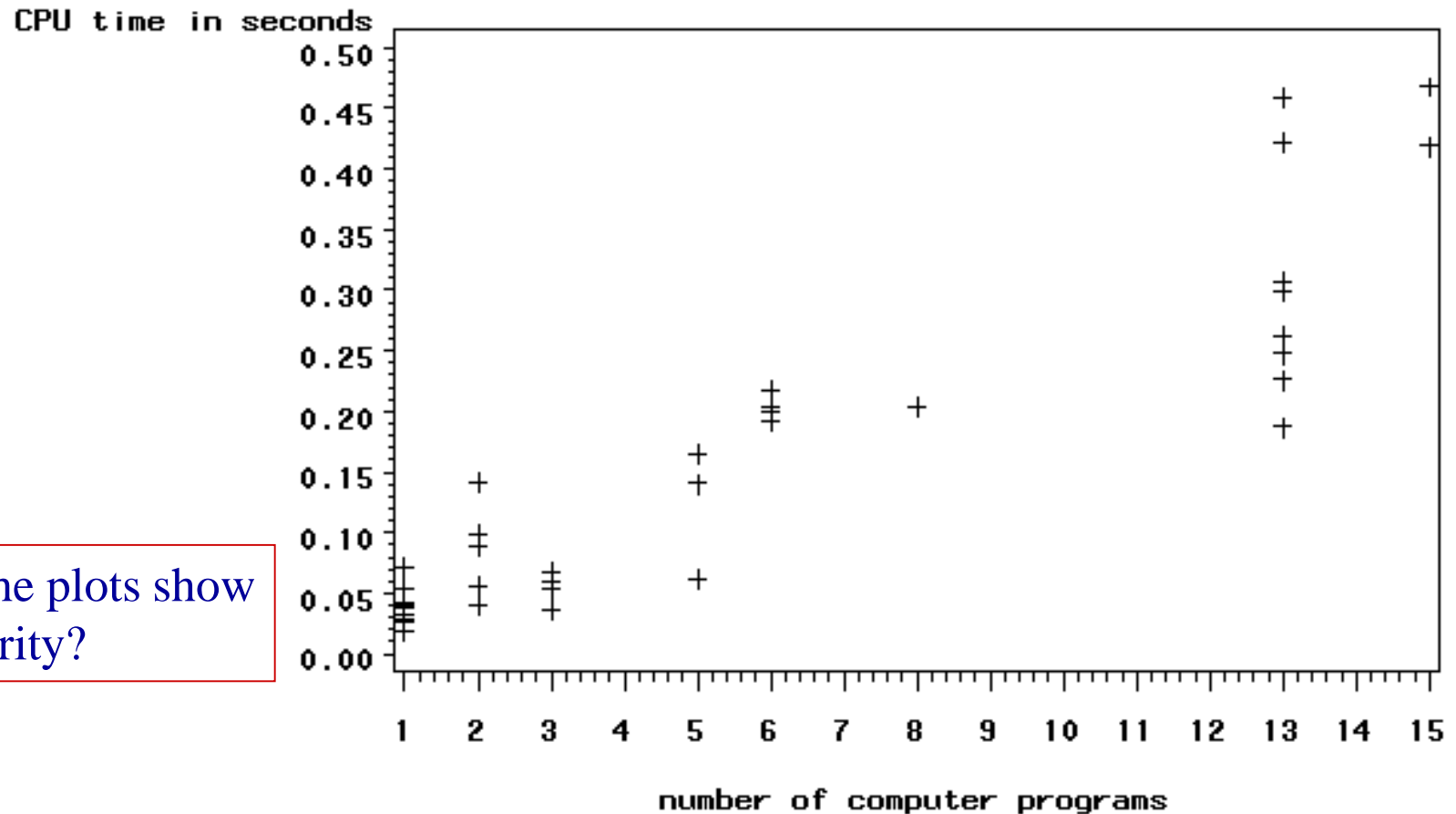
Draw the scatter plot for each pair (Y, X_i)



Do the plots show linearity?

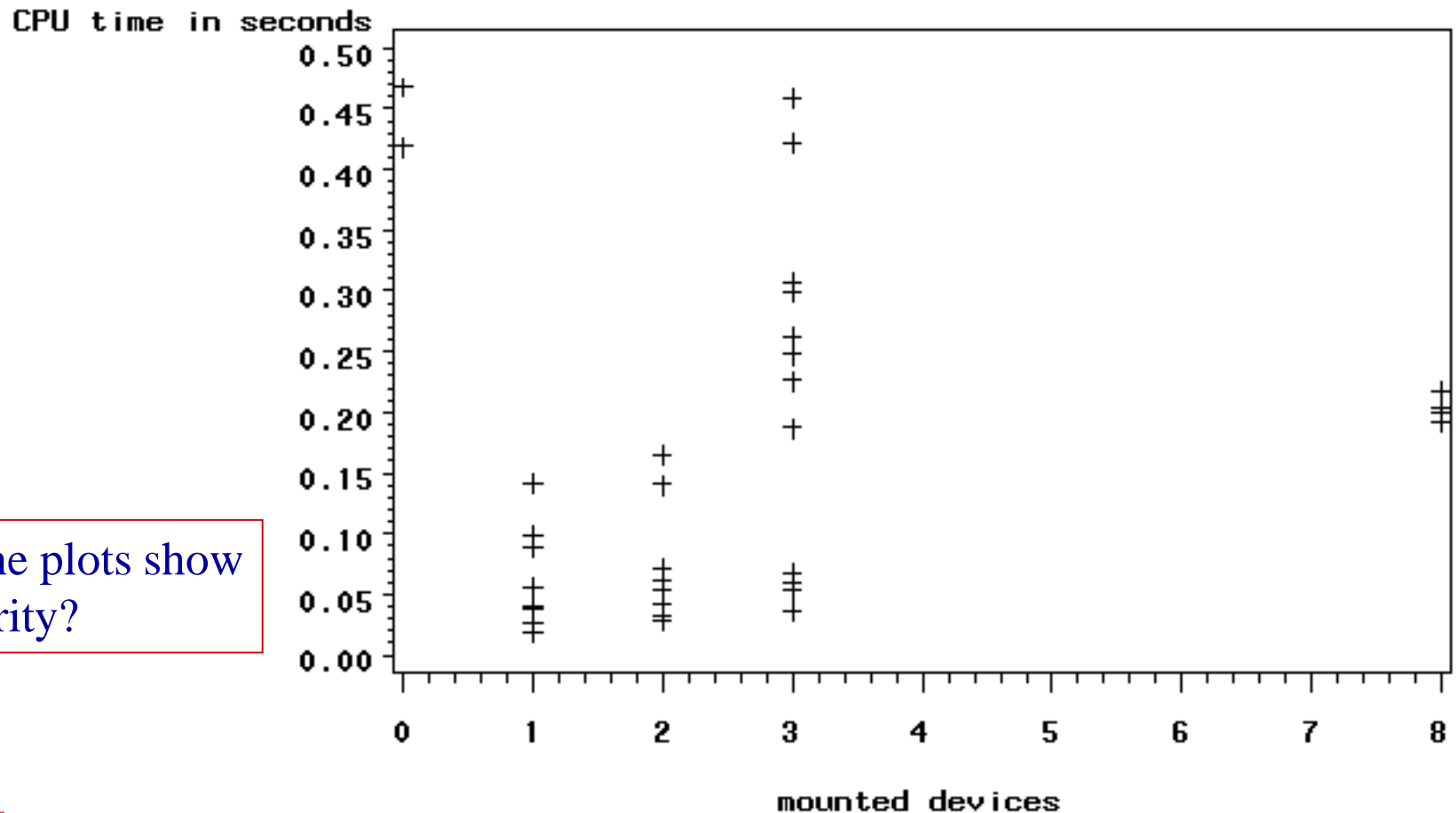
I) Exploratory data step: Are the associations between Y and the x-variables linear?

Draw the scatter plot for each pair (Y, X_i)



I) Exploratory data step: Are the associations between Y and the x-variables linear?

Draw the scatter plot for each pair (Y, X_i)



Regression analysis in R

```
> fit <- lm(time ~ linet+step+device, data=mydata)
```

```
> summary(fit) # show results
```

Call:

```
lm(formula = time ~ linet + step + device, data = mydata)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-0.074914	-0.020733	0.001676	0.016939	0.090459

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	0.001466	0.010713	0.137	0.891959
linet	0.021091	0.002709	7.786	4.64e-09 ***
step	0.009241	0.002096	4.408	9.90e-05 ***
device	0.012183	0.002883	4.225	0.000169 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.03459 on 34 degrees of freedom

Multiple R-squared: 0.9362, Adjusted R-squared: 0.9306

F-statistic: 166.4 on 3 and 34 DF, p-value: < 2.2e-16



Next Steps

- Once you build the model, you should validate this model is qualified or not
- If the model is NOT qualified, you need to figure out the problem, fix it and re-build the model



Goodness of fit

- **Goodness of fit** refers to the examination of how well your fitted model can be used to explain the observations.
- Usually, there are three measures
 - Goodness of fit test (F-test) → whether x variables are useful
 - Individual Parameter Test → which variables are useful
 - Coefficient of determination R^2 → training performance
- **How to determine a model is qualified?**
 - Pass the F-test
 - Meet the requirement of residual analysis


Goodness of fit

- Usually, there are three measures
 - Goodness of fit test (F-test)
 - Individual Parameter Test
 - Coefficient of determination R^2

Goodness of fit: F-test

The test is on the hypothesis that the model is completely wrong!

Null hypothesis: None of the x-variables included in the model have any association with Y:

***H₀*:** $\beta_1 = \beta_2 = \beta_3 = \dots = \beta_p = 0$  No linear relationship!!

Alternative hypothesis: At least one X-variable has a significant effect on changes in Y:

***H_a*:** At least one coefficient $\beta_j \neq 0$  At least one X variable can affect Y

It is used to measure whether at least one independent variable is useful to predict the dependent variable

Goodness of fit: F-test

- Test statistic:

$$F_{STAT} = \frac{MSR}{MSE} = \frac{\frac{SSR}{k}}{\frac{SSE}{n - k - 1}}$$



Larger F value,
Better Model

where F_{STAT} has numerator d.f. = k and
denominator d.f. = $(n - k - 1)$

k = the total number of X variables

Goodness of fit: F-test

```
> fit <- lm(time ~ linet+step+device, data=mydata)
```

```
> summary(fit) # show results
```

Call:

```
lm(formula = time ~ linet + step + device, data = mydata)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-0.074914	-0.020733	0.001676	0.016939	0.090459

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	0.001466	0.010713	0.137	0.891959
linet	0.021091	0.002709	7.786	4.64e-09 ***
step	0.009241	0.002096	4.408	9.90e-05 ***
device	0.012183	0.002883	4.225	0.000169 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.03459 on 34 degrees of freedom

Multiple R-squared: 0.9362, Adjusted R-squared: 0.9306

F-statistic: 166.4 on 3 and 34 DF, **p-value: < 2.2e-16**

P-value < 0.05

Conclusion:

At 95% confidence level, we say that at least one X variable has significant linear relationship with Y, and it can affect the value of the Y



Goodness of fit

- Usually, there are three measures
 - Goodness of fit test (F-test)
 - Individual Parameter Test
 - Coefficient of determination R^2

F-Test vs Individual Parameter Test

- Based on F-Test, we know whether at least one x variable is useful to predict y
- But, which x variables are useful, while which ones are NOT useful? → the individual parameter test (it is a t -test usually) can tell you!

Goodness of fit: Individual Parameter Test

Consider the simple straight line case.

We often test the null hypothesis that *the slope is equal to zero which is equivalent to “X has no effect on Y”!*

Or in statistical terms :

Ho: $\beta_1 = 0$ vs Ha: $\beta_1 \neq 0$ X has a significant effect

The test is computed using the t-statistic

$$t = \frac{\hat{\beta}_1 - 0}{s.e.(\hat{\beta}_1)} = \frac{\hat{\beta}_1}{s_e \sqrt{1 / \sum (x_i - \bar{x})^2}}$$

With t-distribution with n-2 degrees of freedom!



Goodness of fit: Individual Parameter Test

```
> fit <- lm(time ~ linet+step+device, data=mydata)
```

```
> summary(fit) # show results
```

Call:

```
lm(formula = time ~ linet + step + device, data = mydata)
```

Residuals:

Min	1Q	Median	3Q	Max
-0.074914	-0.020733	0.001676	0.016939	0.090459

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	0.001466	0.010713	0.137	0.891959
linet	0.021091	0.002709	7.786	4.64e-09 ***
step	0.009241	0.002096	4.408	9.90e-05 ***
device	0.012183	0.002883	4.225	0.000169 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.03459 on 34 degrees of freedom

Multiple R-squared: 0.9362, Adjusted R-squared: 0.9306

F-statistic: 166.4 on 3 and 34 DF, p-value: < 2.2e-16



Goodness of fit

- Usually, there are three measures
 - Goodness of fit test (F-test)
 - Individual Parameter Test
 - Coefficient of determination R^2



Goodness of fit

- Coefficient of determination R^2

It is used to measure how many variation in Y can be explained by the X variables. It tells how good your model is based on the training set → it is NOT the metric to evaluate how good the model is

Goodness of Fit: r^2

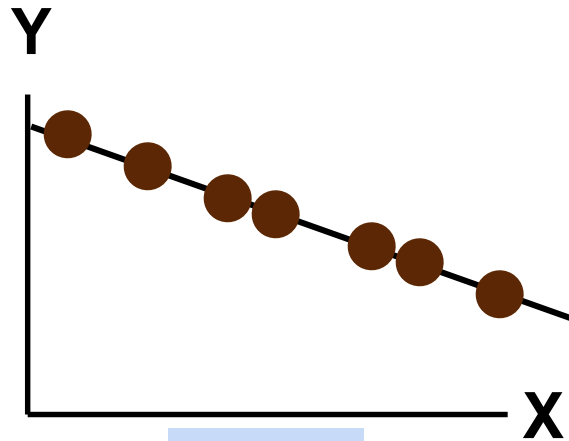
- The **coefficient of determination** is the portion of the total variation in the dependent variable that is explained by variation in the independent variable
- The coefficient of determination is also called **r-square** and is denoted as r^2
- It is considered as the metric for goodness of fit

$$r^2 = \frac{SSR}{SST} = \frac{\text{regression sum of squares}}{\text{total sum of squares}}$$

note:

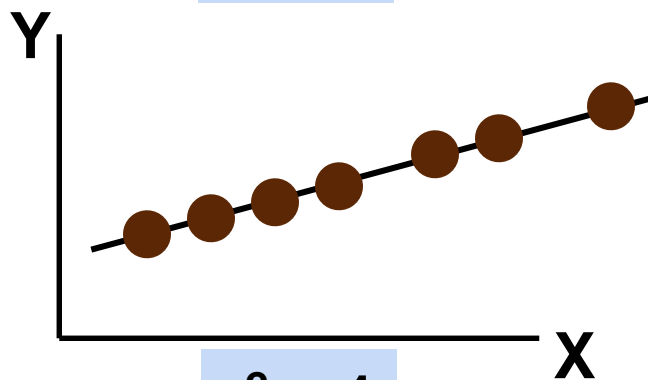
$$0 \leq r^2 \leq 1$$

Goodness of Fit: r^2



$$r^2 = 1$$

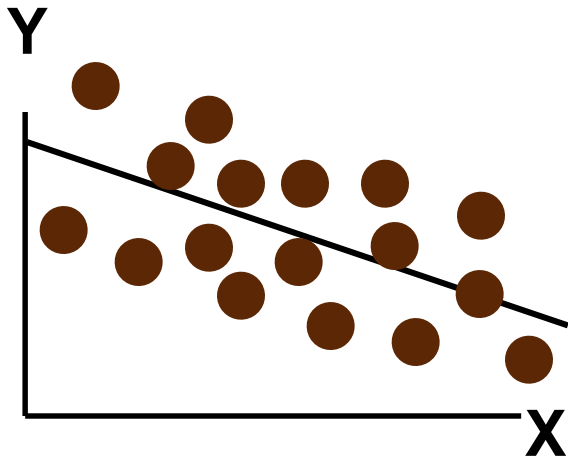
**Perfect linear relationship
between X and Y:**



$$r^2 = 1$$

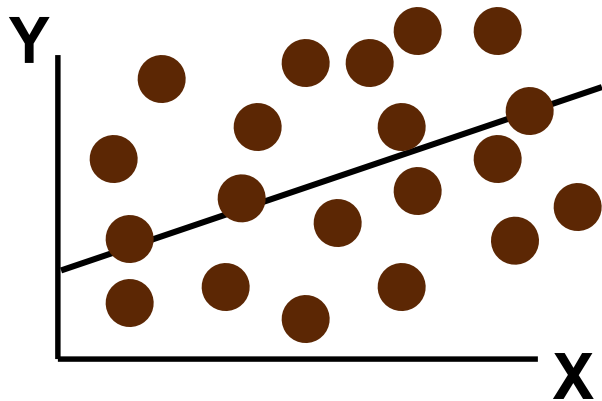
**100% of the variation in Y is
explained by variation in X**

Goodness of Fit: r^2



$$0 < r^2 < 1$$

**Weaker linear relationships
between X and Y:**



**Some but not all of the
variation in Y is explained
by variation in X**

It is similar to “accuracy”

Goodness of Fit: r^2

```
> fit <- lm(time ~ linet+step+device, data=mydata)
```

```
> summary(fit) # show results
```

Call:

```
lm(formula = time ~ linet + step + device, data = mydata)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-0.074914	-0.020733	0.001676	0.016939	0.090459

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	0.001466	0.010713	0.137	0.891959
linet	0.021091	0.002709	7.786	4.64e-09 ***
step	0.009241	0.002096	4.408	9.90e-05 ***
device	0.012183	0.002883	4.225	0.000169 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.03459 on 34 degrees of freedom

Multiple R-squared: 0.9362, Adjusted R-squared: 0.9306

F-statistic: 166.4 on 3 and 34 DF, p-value: < 2.2e-16

93.62% variation in Y can be explained by the variations in X variables based on our fitted regression model



Goodness of Fit: adjusted-r²

adj-R² is useful when comparing two models with a different set of x-variables.

$$adjR^2 = 1 - \frac{(n-1)}{n-(k+1)} (1 - R^2)$$

Unlike the R², the Adj-R² value does not increase with the addition of a x-variable that does not improve the regression model.

A higher adj-R² typically indicates a better model, in terms of the training set

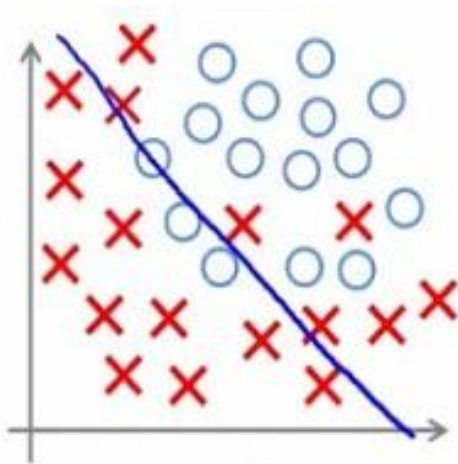


Important Note

- R^2 and Adj- R^2 can only tell you how good the model is based on the training set
- A model that performs well on the training set is NOT guaranteed to perform well on the testing set too → overfitting problem
- How to evaluate a model? We never trust the performance on training set. You have to use the test set to evaluate a model to tell how good it is.

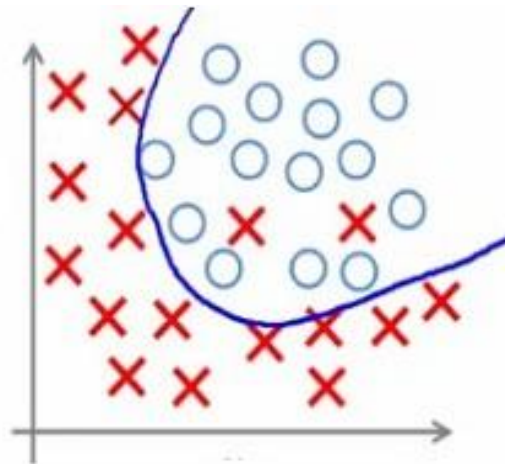
Overfitting Problem

Problem: The model is over-trained by the training set; it may show a high accuracy on training set, but bad performance on test set.

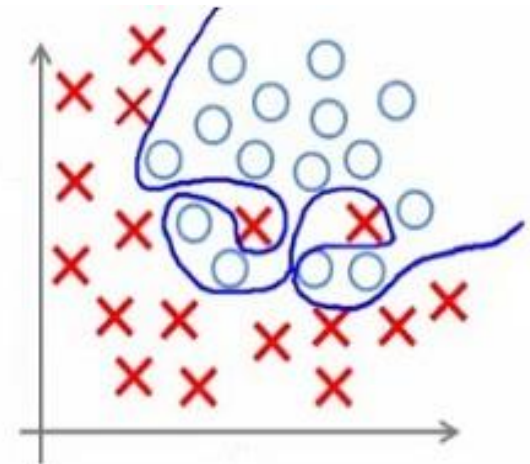


Under-fitting

(too simple to explain the variance)



Appropriate-fitting



Over-fitting

(forcefitting -- too good to be true)

Goodness of fit

- Summary

- Overall Goodness of fit test (F-test)

It is used to measure whether the linear relationship is necessary or influential enough for Y with at least one X variable

- Individual Parameter Test

Whether each X variable is influential enough

- Coefficient of determination R^2

It is used to measure how many variation in Y can be explained by the X variables. **It tells how good your model is based on the training set**

Multiple Linear Regression

- Parameter Estimates
- Goodness of Fit Test
- Residual Analysis
- Evaluations and Predictions



Assumptions on the regression model

1. The relationship between the Y-variable and the X-variables is linear
2. The error terms e_i (measured by the residuals)
 - have zero mean ($E(e_i)=0$)
 - have the constance variable or standard deviation
 - are close to normal distribution- Typically true for large samples!
 - are independent (true if sample is from simple random sampling)

Such assumptions are necessary to derive the inferential methods for testing and prediction (to be examined by residual analysis)

WARNING: if the sample size is small ($n < 30$) and residuals are not normal, you can't use regression methods!

Goals in Residual Analysis

1. Validate the constant variance
2. Validate the linearity relationship
3. Validate normal distribution of residuals
4. Identify potential outliers (optional)

~~5. Zero mean~~

~~6. Independent~~



We do not need to validate these two conditions, since we use standardized residual for analysis

Standard residuals and standardized residuals

Standard residuals measure the difference between the actual y-values and the predicted values using the regression model:

$$r_i = (y_i - \hat{y}_i)$$

We often use the **standardized residuals** to identify *outliers*, i.e. points that do not appear to be consistent with the rest of the data.

$$e_i = \frac{(y_i - \hat{y}_i)}{s.e.(y_i - \hat{y}_i)} = \frac{(y_i - \hat{y}_i)}{\sqrt{MSE}}$$

A standardized residual is computed as the i-th residual divided by its standard error. Some statistical software (such as SAS) may produce **studentized residual** which is assumed to follow t distribution.



Plots for Residual Analysis

Residual analysis shows problems in the regression analysis, e.g. it shows if there is some important variation in Y that is not explained by the regression model.

- **Plot residuals vs predicted values:** *To check constant variance for the residuals*
- **Plot residuals vs each x-variable:** *To check linearity assumptions for Y and the x-variable;*
- **Draw normal probability plot of residuals:** *To check normality assumption for the error terms; if points lie close to a line, the errors can be assumed to be approximately normal. Otherwise the assumption of normality is not satisfied.*



Residual Analysis in R

First define regression model

```
fit = lm(yvar ~ xvar1 +xvar2 +xvar3)
```

- **Plot residuals vs predicted values:**

```
plot( fitted(fit), rstandard(fit), main="Predicted vs  
residuals plot")
```

```
abline(a=0, b=0, col='red') #add zero line
```

- **Plot residuals vs each x-variable:**

```
plot(mydata$xvar, rstandard(fit), main="Margin vs  
residuals plot")
```

```
abline(a=0, b=0,col='red') #add zero line
```

- **Draw normal probability plot of residuals:**

```
qqnorm(rstandard(fit))
```

```
qqline(rstandard(fit), col = 2)
```

<pre>rstandard(fit) computes standardized residuals for model fit</pre> <pre>fitted(fit) computes predicted values for model fit</pre>



Goals in Residual Analysis

1. Validate the constant variance

Plot residuals vs predicted values

2. Validate the linearity relationship

3. Validate normal distribution of residuals

4. Identify potential outliers

How do I know I have a problem?

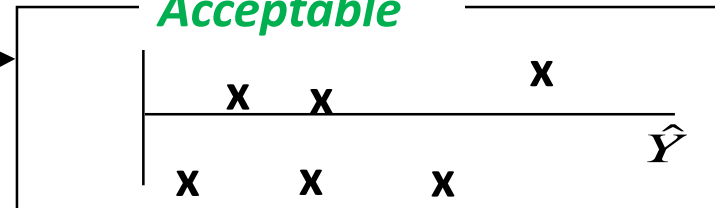
Assumption of constant variance

- 1) Plot residuals vs predicted values
- 2) Plot residuals vs each x-variable

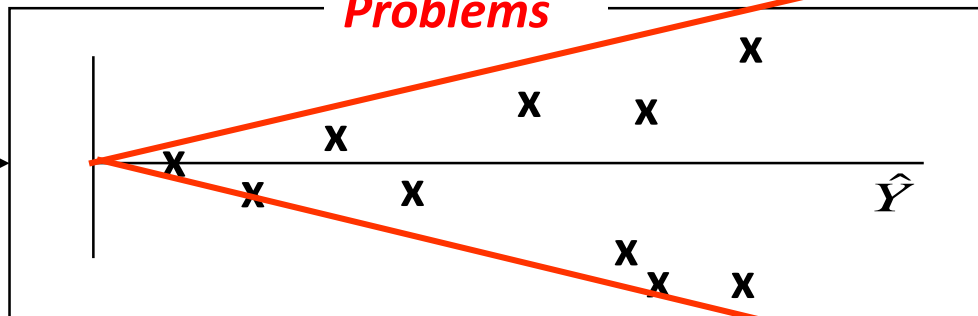
What is the pattern of the spread in the residuals as the predicted values increase?

- Spread constant.
- Spread increases.
- Spread decreases then increases.

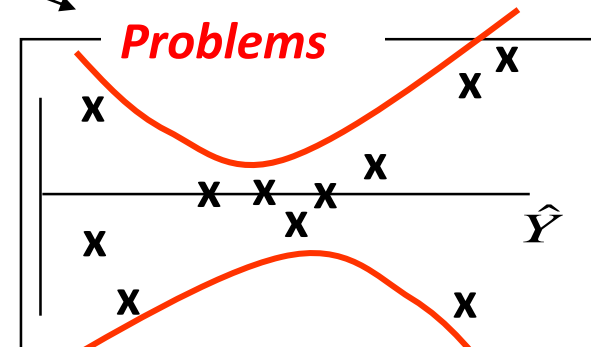
Acceptable



Problems

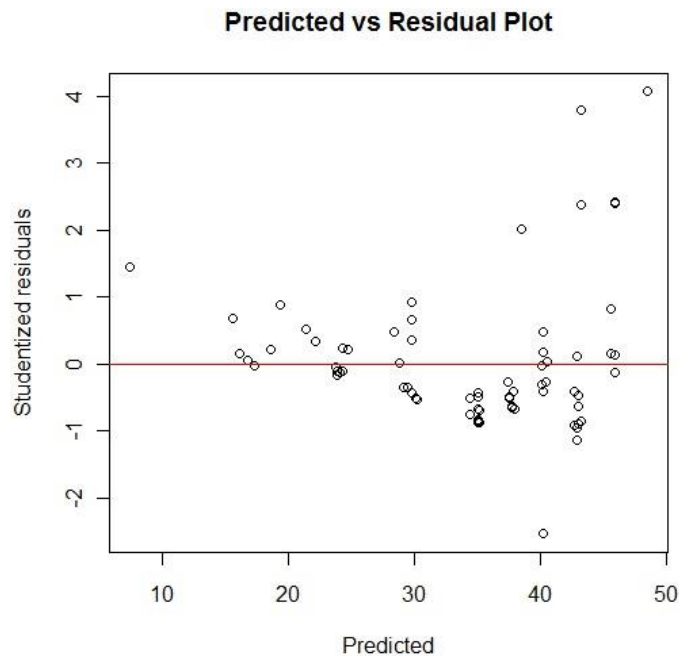


Problems

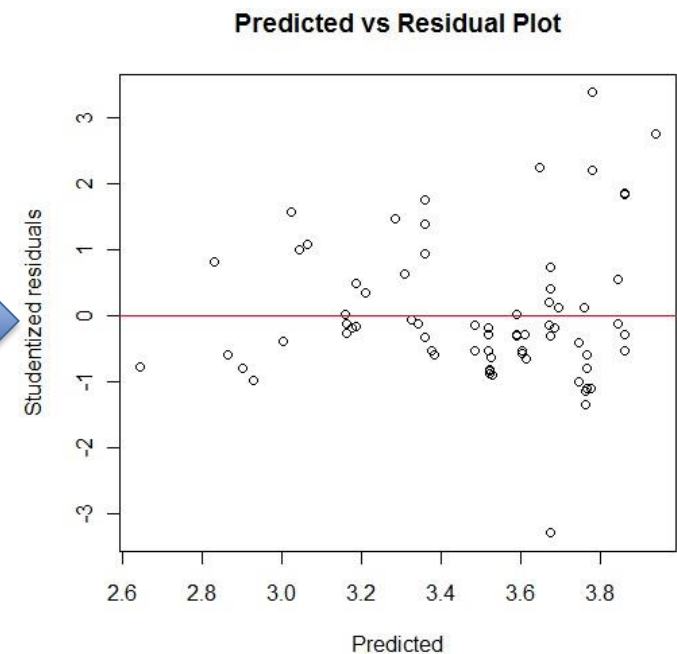


Residual Analysis: Constant Variance

- What if the variance is not constant?
You may need to apply a transformation on y , such as log transformation, and then re-fit the regression model.



Log transformation
on variable y



Goals in Residual Analysis

1. Validate the constant variance

2. Validate the linearity relationship

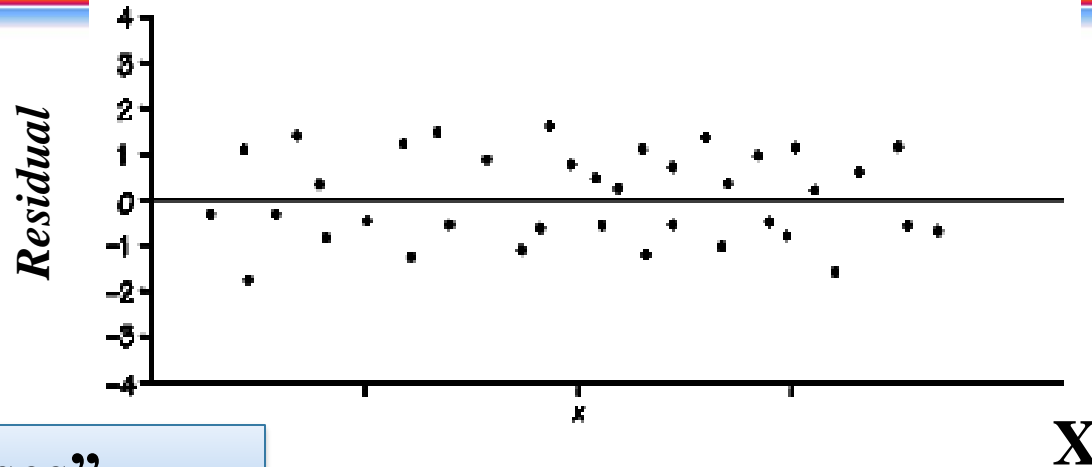
Plot residuals vs each x-variable in your model

3. Validate normal distribution of residuals

4. Identify potential outliers

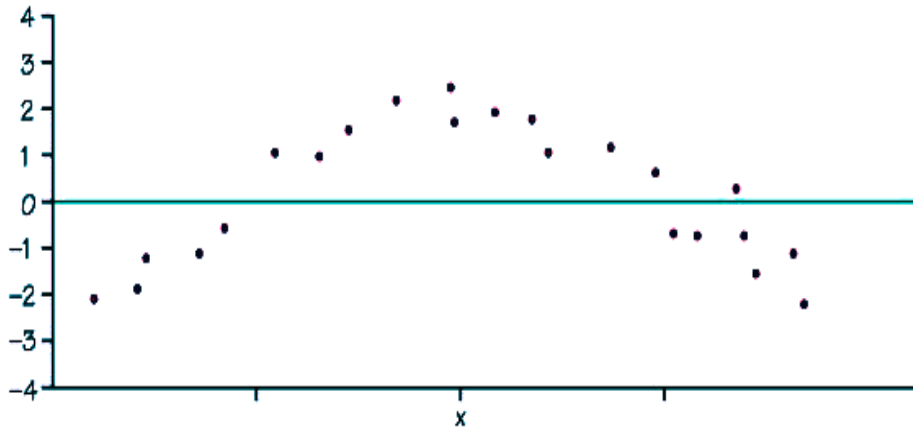
“Good case”

Points are randomly scattered around the zero line

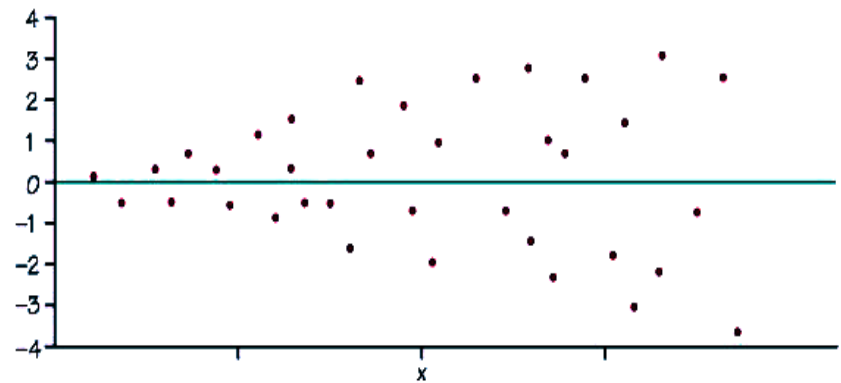


“Bad cases”

Non linear relationship



Non constant variance



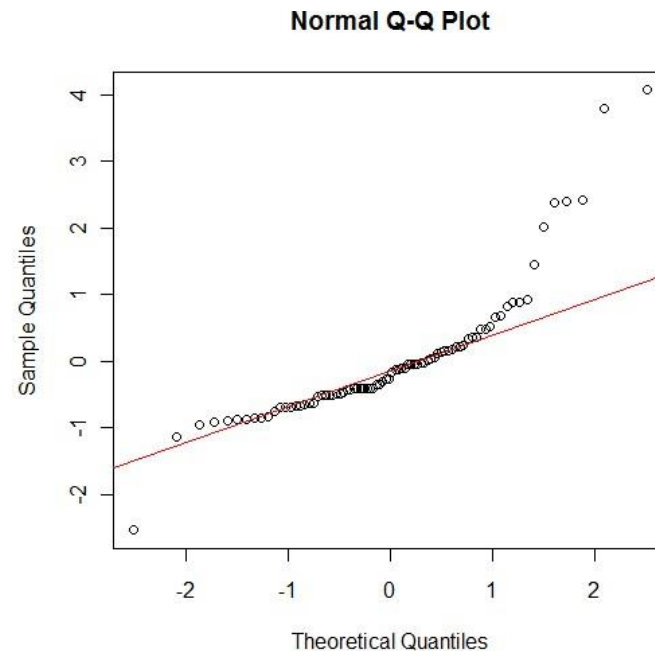
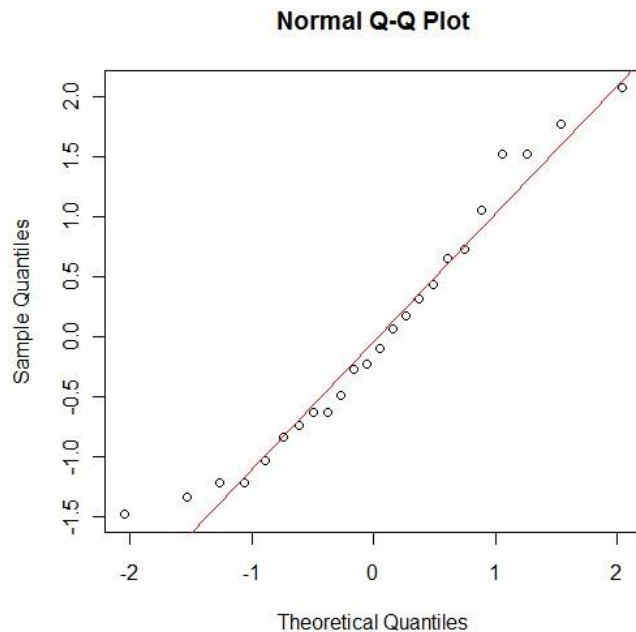
Goals in Residual Analysis

1. Validate the constant variance
 2. Validate the linearity relationship
 3. Validate normal distribution of residuals
 QQ-plot or normality test
 4. Identify potential outliers
-

Normal probability plot of residuals

There are two ways to check whether a variable (including residual) follows normal distribution or not.

Solution-1: by QQPlot ("Q" stands for quantile)



Normal probability plot of residuals

There are two ways to check whether a variable (including residual) follows normal distribution or not.

Solution-2: Normality Test

For example, Shapiro-Wilk Normality Test

In R, `shapiro.test(x)`

If $p\text{-value} > 0.05$, we say it follows normal distribution at 95% confidence level

The null hypothesis in this test is the variable is normal distributed



Goals in Residual Analysis

1. Validate the constant variance
 2. Validate the linearity relationship
 3. Validate normal distribution of residuals
 4. Identify potential outliers
-

Outlier Identification

We often use the standardized residuals to identify outliers, i.e. points that do not appear to be consistent with the rest of the data.

Use scatter plots to identify large residuals:

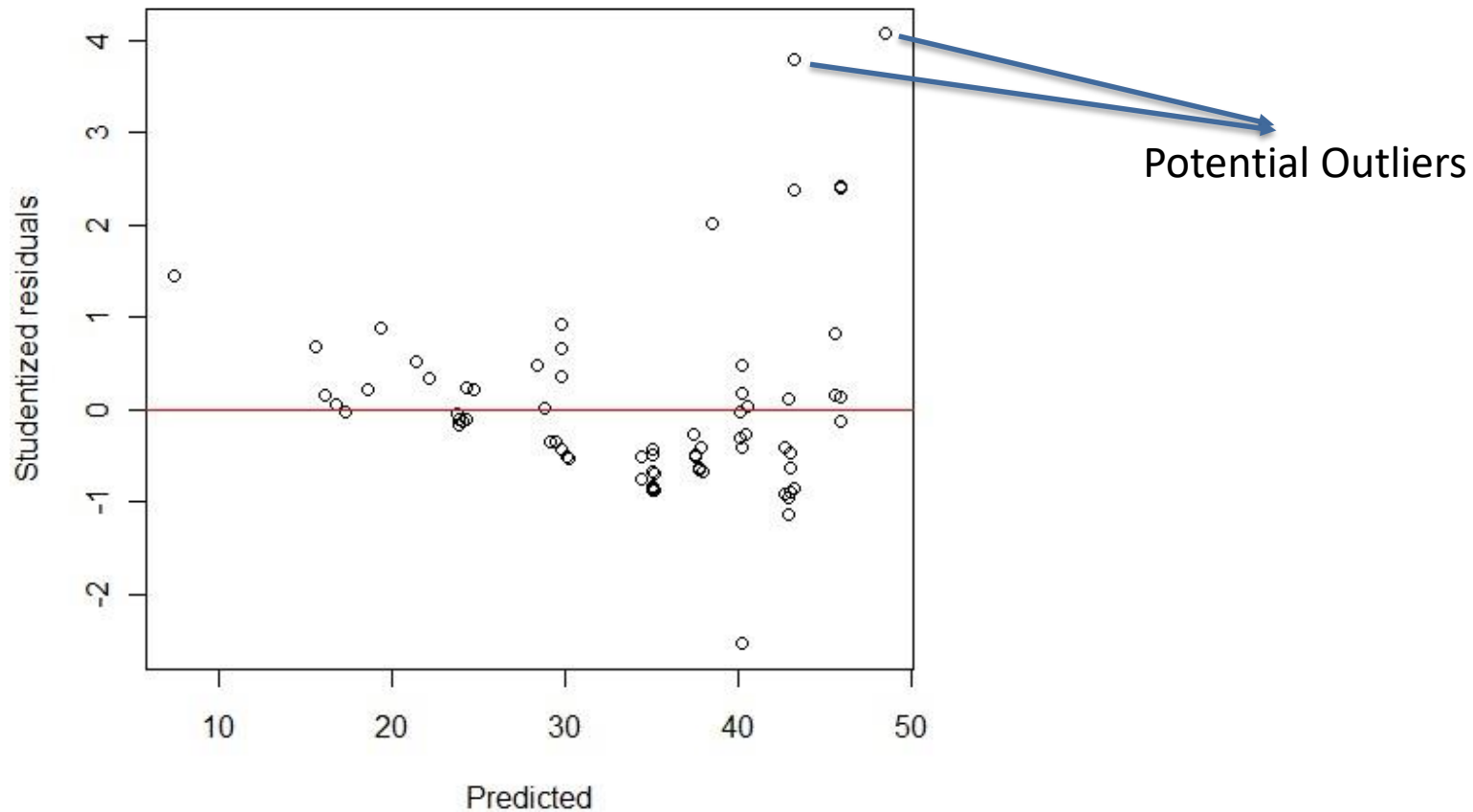
- Standardized residuals vs predicted values.
- Standardized residuals vs x-values

Possible outliers are observations with standardized/studentized residuals $|e_i| > 3$ (i.e. e_i larger than 3 or smaller than -3)



Outlier Identification

Predicted vs Residual Plot



Potential Outliers



Fitted model

How to examine your model is qualified or not? You need to go through the following steps:

- F-test → p-value must be $< \alpha$, so that at least one x variable is useful to make predictions → **MUST DO**
- Individual parameter test [Optional] → it is related to the process of feature selection, we will introduce later
- Read Adjust-R2 [Optional] → read it once the model is qualified
- **Residual analysis → MUST DO**
 - Constant variance
 - Linear relationship with x variables
 - Follow normal distribution
- Outlier detection [Optional] → It is used to improve your model



Fitted model

Once your model is qualified, you can write down the model, and also evaluate the model on the test set.

The fitted regression model is

$$\hat{y} = 0.0014 + 0.021LINET + 0.009STEP + 0.012DEVICE$$

- The β 's estimated values measure the changes in Y for changes in X's.
- For instance, for each increase of 1000 lines executed by the process (keeping the other variables fixed), the CPU usage time will increase of 0.021 seconds.
- *Fixing the other variables, what happens on the CPU time if I add another device?*



Interpretation of model parameters

In multiple regression

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + e_i \text{ for } i = 1, \dots, n$$

The coefficient value β_i of X_i measures the predicted change in Y for any unit increase in X_i while the other independent variables stay constant.

For instance: β_2 measures the changes in Y for a unit increase of the variable X_2 if the other x -variables X_1 and X_3 are fixed.



Assignment #2



Next Class

- Feature Selection
 - How to select the independent variables
 - How to use feature selection to build different models
- Multiple Linear Regression by Using R