Simple and Multiple Linear Regression Models
&
Regression Diagnostics

Madhuri G. Kulkarni
Department of Statistics
Savitribai Phule Pune University

March 27, 2023

What is Regression Analysis?

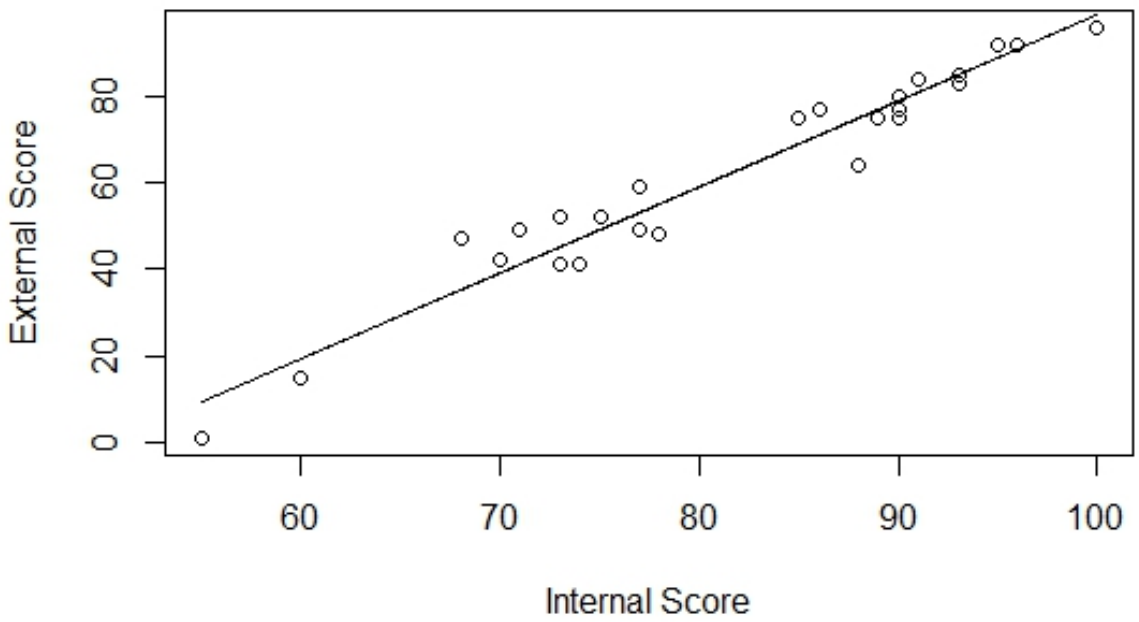
Concept

Regression analysis is a statistical technique for investigating and modeling the relationship between variables. A variable of interest is linked with other variables (influencing it) through some function.

Example Following table lists students' scores obtained during internal examination and the final examination.

INTERNAL	FINAL
75	52
93	85
90	80
100	96

Simple Linear Regression



Is there a relationship between the final score and the internal score? Is it linear?
Can we predict the final score based on the internal score?
How accurately can we estimate the effect of internal exam on the score obtained in the final exam?
How accurately can we predict future final exam scores?

- The variable of interest, student's final score, is called as response variable. It is usually denoted by Y .
- The variable which influences the final score, viz., student's internal score is the explanatory variable. It is also called as predictor or regressor variable.
- In regression analysis, a functional relationship is established between the response variable and the regressor.
- A simple linear regression model is

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

where ϵ is assumed to be a random error representing the discrepancy in the approximation. It accounts for the failure of the model to fit the data exactly.

Simple Linear Regression

Simple linear regression is a very straightforward approach for predicting a quantitative response y on the basis of a single predictor variable x.

More about Simple Linear Regression Model

The simple linear regression model is

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

β_0 and β_1 are two unknown constants that represent the intercept and slope terms in the linear model.

Together, β_0 and β_1 are known as the model coefficients or parameters.

Once we have used our data to produce estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ for the model coefficients, we can predict future scores on the basis of a particular value of internal score by computing

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

where $\hat{y}$ indicates prediction of y on the basis of $X = x$.

Understanding Simple Linear Regression Model

The random error, ϵ determines the properties of the response, y.

Assuming that the mean and variance of ϵ are 0 and σ^2, we have

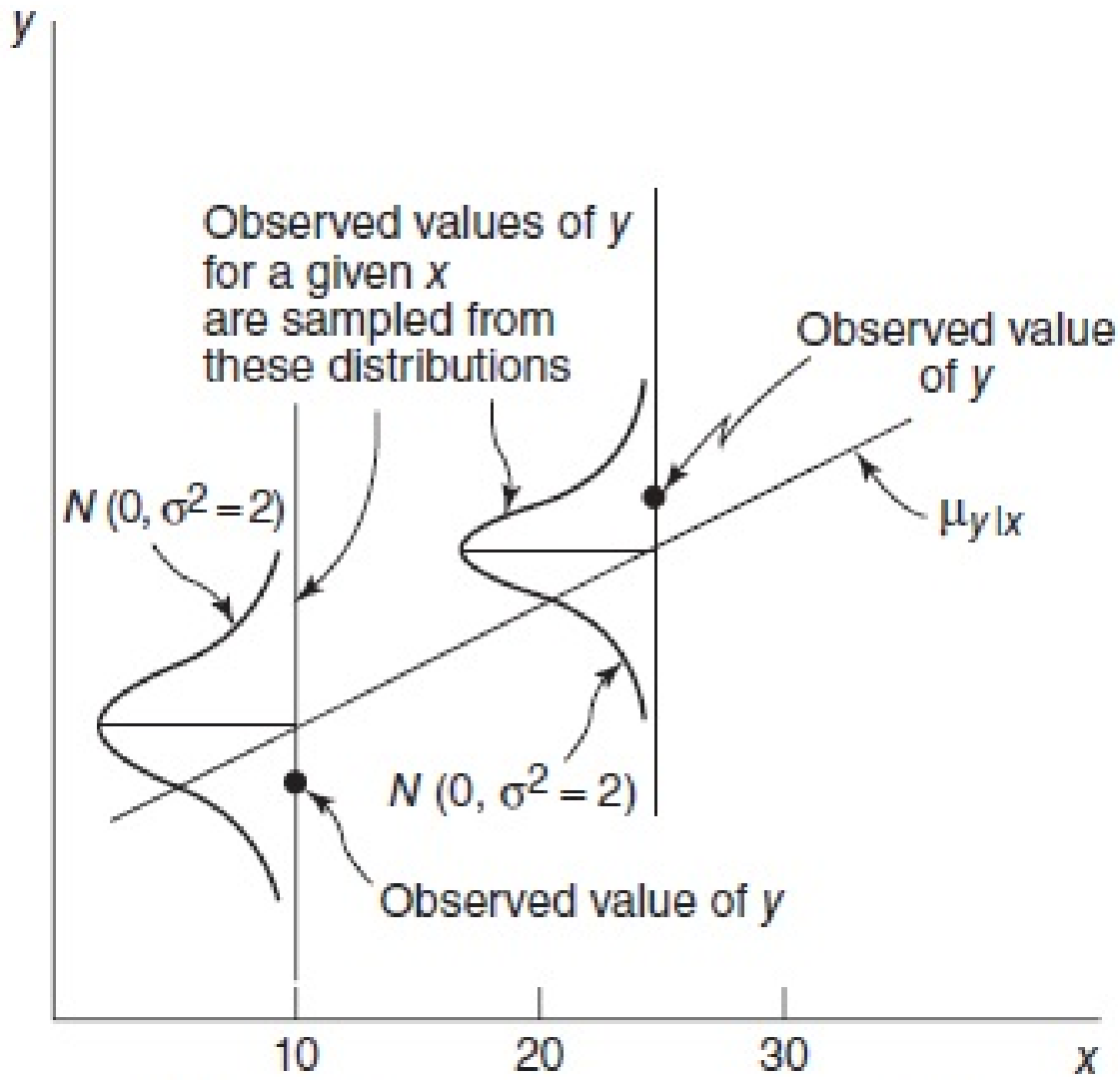
$$E(Y|x) = \mu_{y|x} = \beta_0 + \beta_1 x \quad \text{and} \quad \text{Var}(Y|x) = \sigma^2_{y|x} = \sigma^2$$

Thus, the true regression model $\mu_{y|x} = \beta_0 + \beta_1 x$ is a line of mean values, that is, the height of the regression line at any value of x is just the expected value of y for that x.

The slope, β_1 can be interpreted as the change in the mean of y for a unit change in x.

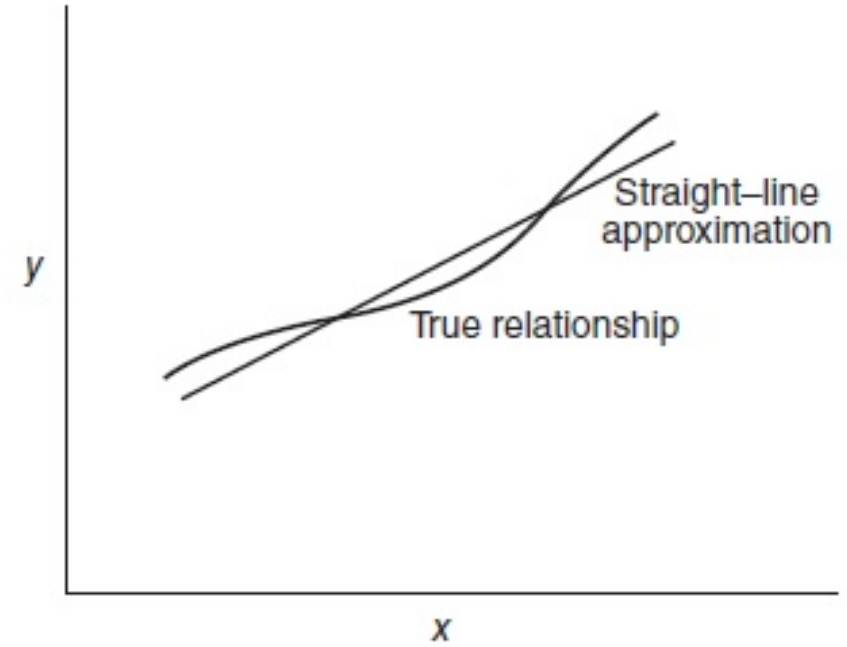
The variability of y at a particular value of x is determined by the variance of the error component of the model, σ^2. This implies that there is a distribution of y values at each x and that the variance of this distribution is the same at each x.

Understanding SLR Model



How observations are generated

When can we apply Linear Regression Models?



The functional relationships between the response and the regressors are often based on physical, chemical, or other engineering or scientific theory, that is, knowledge of the underlying mechanism.
Above figure illustrates a situation where the true relationship between y and x is relatively complex, yet it may be approximated quite well by a linear regression equation.

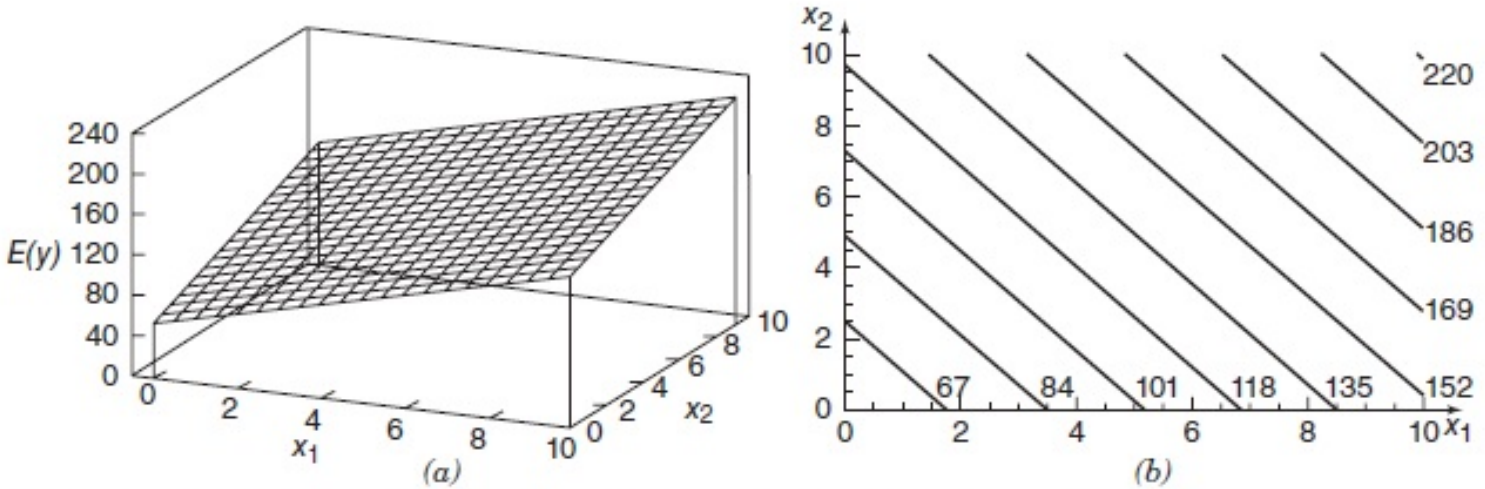
Multiple Linear Regression Model

- Simple linear regression is a useful approach for predicting a response on the basis of a single regressor variable. However, in practice we often have more than one regressor.
- The model

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + \epsilon$$

  is known as **Multiple Linear Regression Model** because more than one regressor is involved.
- The adjective 'linear' is employed to indicate that the model is linear in the parameters $\beta_0, \beta_1, \ldots, \beta_k$, not because y is a linear function of the xs.
- When $k = 2$, the regression model describes a plane in the three-dimensional space of y, x_1 and x_2.
- The parameters $\beta_j,\ j = 1, 2, \ldots, k$ are called the regression coefficients.
- The model describes a hyperplane in the k-dimensional space of the regressor variables x_j.
- The parameter β_j represents the expected change in the response y per unit change in x_j when all of the remaining regressor variables $x_i (i \neq j)$ are held

How does an MLR model look?



(*a*) The regression plane for the model $E(y) = 50 + 10x_1 + 7x_2$.

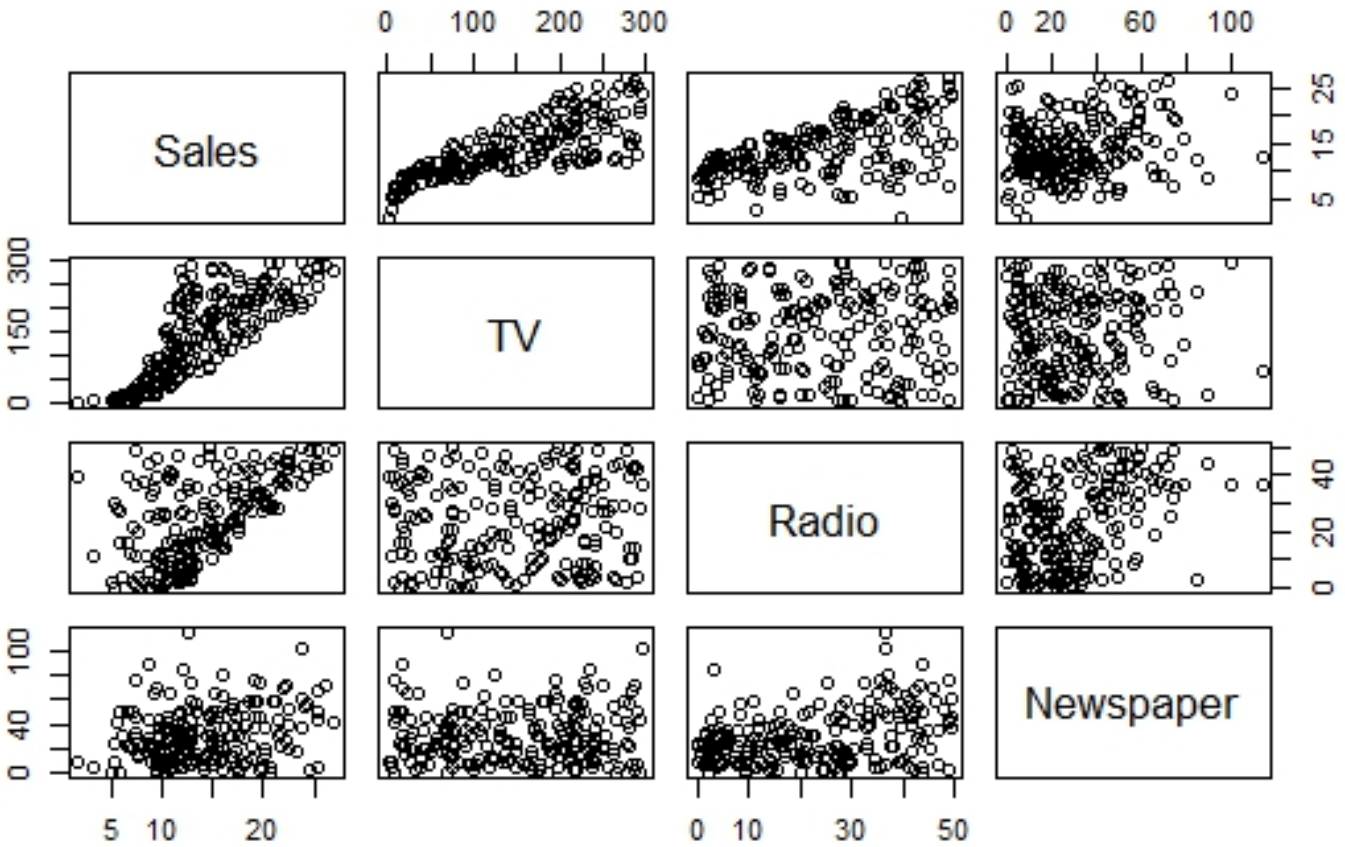
(*b*) The contour plot.

Illustration

Problem at hand

Suppose that we are statistical consultants hired by a client to provide advice on how to improve sales of a particular product. The 'Advertising' data set consists of the sales of that product in 200 different markets, along with advertising budgets for the product in each of those markets for three different media: TV, radio, and newspaper.

It is not possible for our client to directly increase sales of the product. On the other hand, they can control the advertising expenditure in each of the three media. Therefore, if we determine that there is an association between advertising and sales, then we can instruct our client to adjust advertising budgets, thereby indirectly increasing sales. In other words, our goal is **to develop an accurate model that can be used to predict sales on the basis of the three media budgets.** So, the response variable is the sales and the regressors are advertising budgets for TV, radio and newspaper.


Sales
TV
Radio
Newspaper
0 100 200 300
0 20 60 100
5 15 25
0 150 300
0 20 40
0 40 100
5 10 20
0 10 30 50

Model Fitting in R

Model with Summary

reg.lm = lm(Sales ~ TV+Radio+Newpaper, data = Ad)

```
Call:
lm(formula = Sales ~ TV + Radio + Newspaper, data = Ad)

Residuals:
    Min      1Q  Median      3Q     Max
-8.8277 -0.8908  0.2418  1.1893  2.8292

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.938889   0.311908   9.422   <2e-16 ***
TV           0.045765   0.001395  32.809   <2e-16 ***
Radio        0.188530   0.008611  21.893   <2e-16 ***
Newspaper   -0.001037   0.005871  -0.177     0.86
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.686 on 196 degrees of freedom
Multiple R-squared:  0.8972,    Adjusted R-squared:  0.8956
F-statistic: 570.3 on 3 and 196 DF,  p-value: < 2.2e-16
```

```
##
## Call:
## lm(formula = FINAL ~ EXAM1 + EXAM2 + EXAM3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.7452 -1.6328 -0.2984  0.8046  7.3111
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -4.3361     3.7642  -1.152  0.26230
## EXAM1         0.3559     0.1214   2.932  0.00796 **
## EXAM2         0.5425     0.1008   5.379 2.46e-05 ***
## EXAM3         1.1674     0.1030  11.333 2.08e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.614 on 21 degrees of freedom
## Multiple R-squared:  0.9897, Adjusted R-squared:  0.9882
## F-statistic: 670.1 on 3 and 21 DF,  p-value: < 2.2e-16
```

Model Assumptions

1. The relationship between the response y and the regressors is linear, at least approximately.
2. The error term ϵ has zero mean.
3. The error term ϵ has constant variance σ^2.
4. The errors are uncorrelated.
5. The errors are normally distributed.

Above assumptions can be verified through residuals, that is,

$e_i = y_i - \hat{y}_i, \;\; i = 1, 2, \ldots, n.$

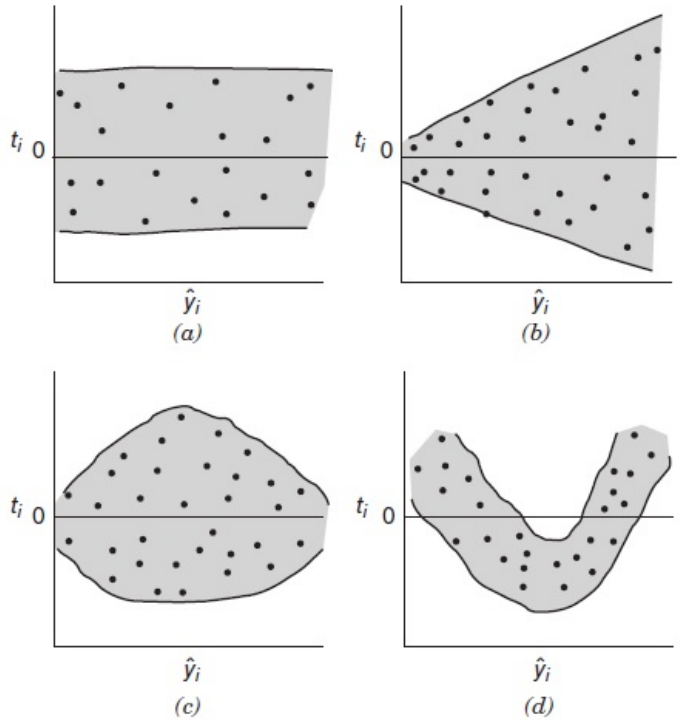
Residual Analysis

- Since a residual may be viewed as the deviation between the data and the fit, it is also a measure of the variability in the response variable not explained by the regression model. It is also convenient to think of the residuals as the realized or observed values of the model errors.
- The normality of errors is verified through QQ plot wherein the sample quantiles are plotted against the population quantiles obtained assuming normality. If most of the points fall around a straight line, the normality of errors is established. Alternative way is to apply tests to examine the normality such as Shapiro test.

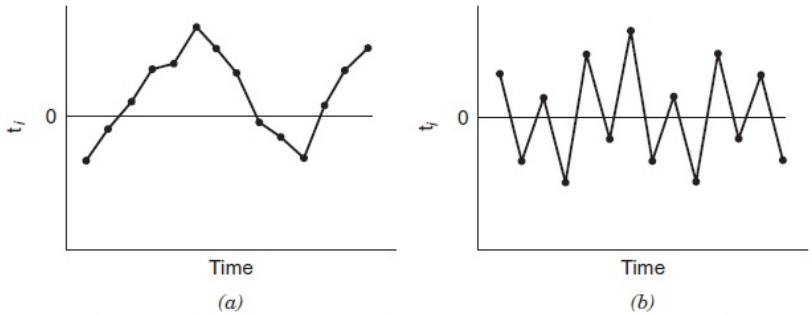
Residual Analysis

- The assumption of uncorrelatedness of the errors is verified through the sequential plot of the residuals. Ideally, a horizontal band will enclose all of the residuals, and the residuals will fluctuate in a more or less random fashion within this band. If there is no pattern in the graph, the residuals are uncorrelated.
- The assumption of constant variance is verified through the plot of residuals against fitted values. If the plot exhibits patterns like opening funnel or a bow, it indicates that the variance is not the same for all the points. This situation is referred to as 'Heteroscedasticity'. A curved plot indicates non-linearity.
- A plot of the residuals against fitted values may also reveal one or more unusually large residuals. These points are, of course, potential outliers. Large residuals that occur at the extreme fitted values could also indicate that either the variance is not constant or the true relationship between y and x is not linear.

Illustration

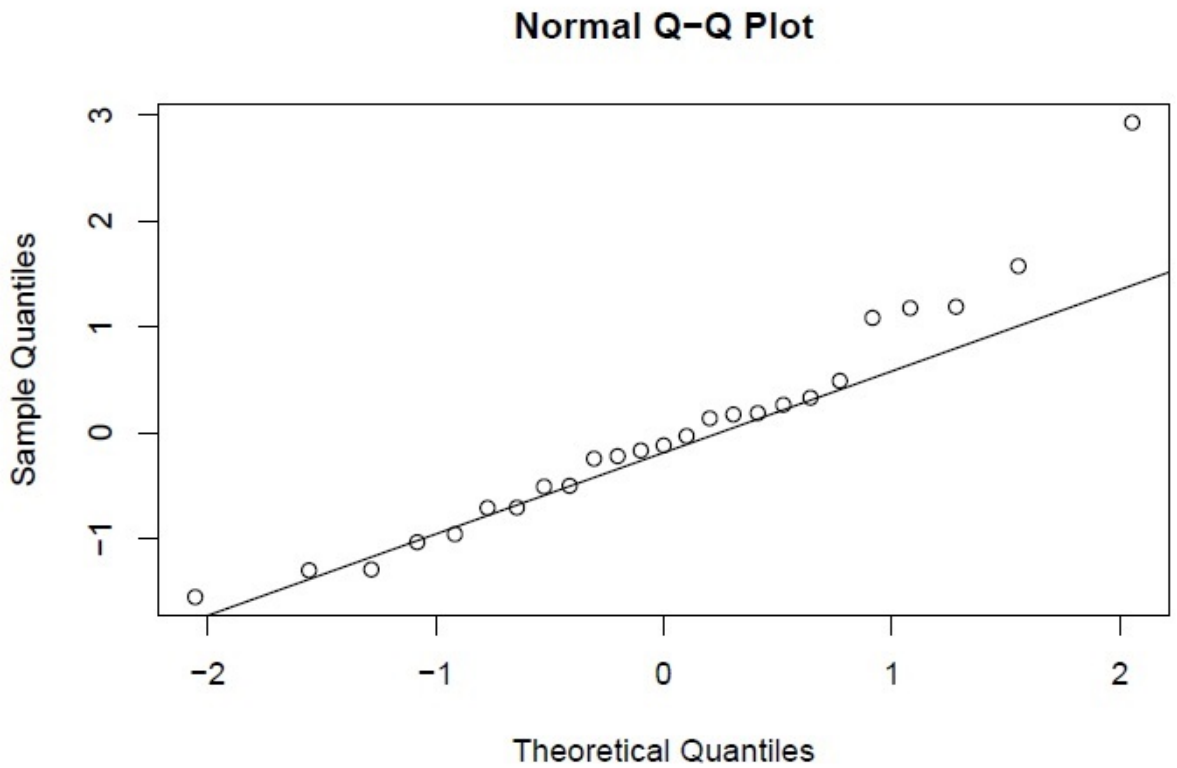


Patterns for residual plots: (*a*) satisfactory; (*b*) funnel; (*c*) double bow; (*d*) nonlinear.

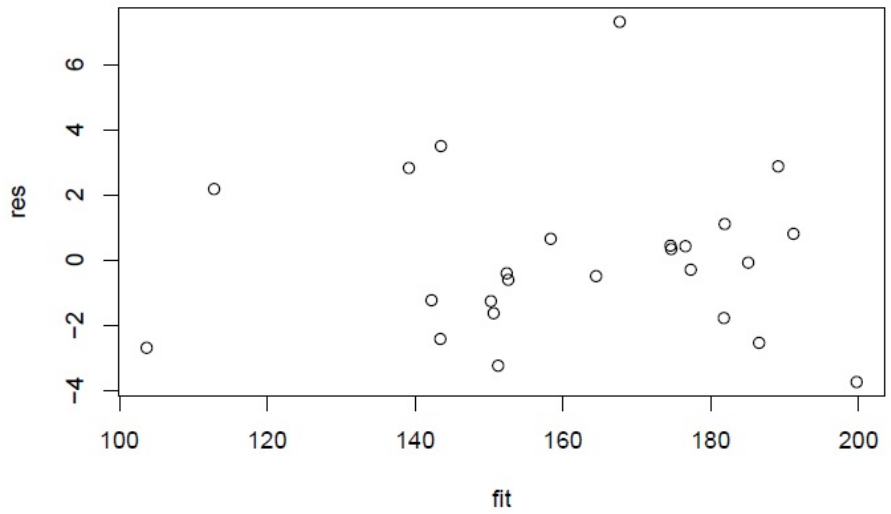


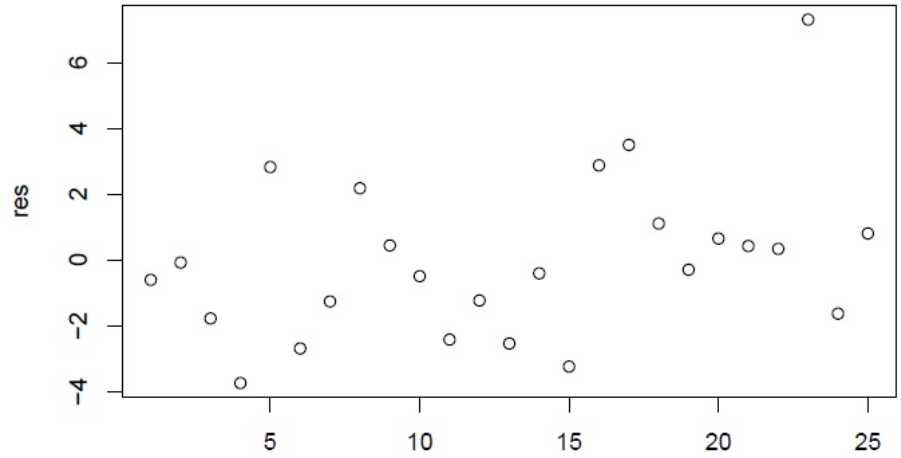
Residual Analysis of Exam Data

```
##
##  Shapiro-Wilk normality test
##
## data:  stdres
## W = 0.9432, p-value = 0.1758
```



Residual Analysis of Exam Data





What if the response is not quantitative?

What if it is binary?

Answers will be revealed in the next lecture!

THANK YOU