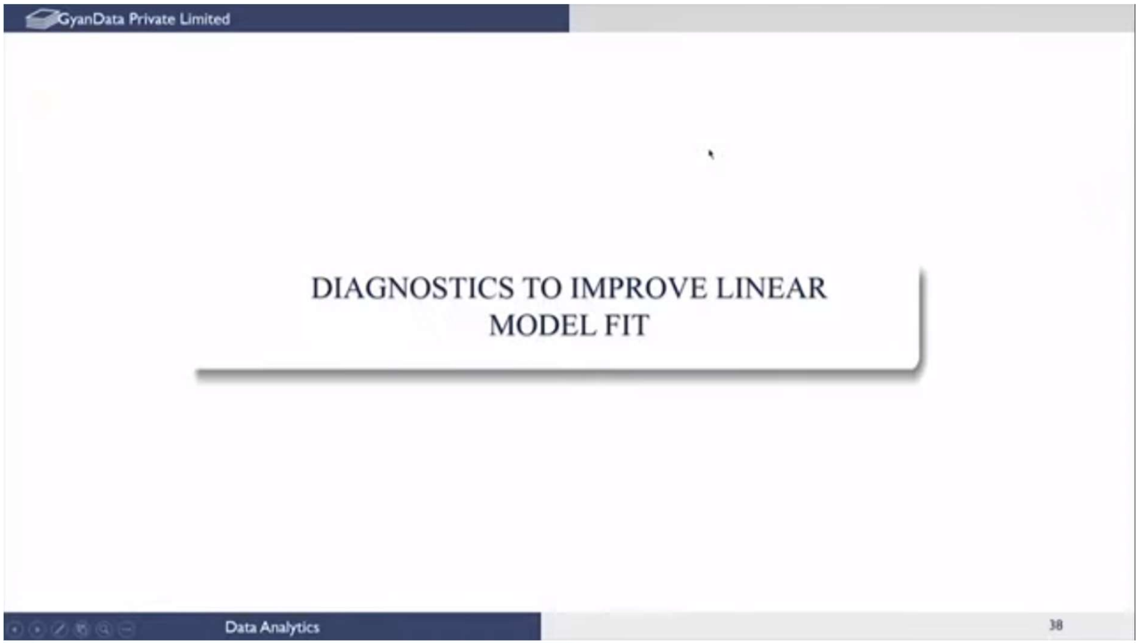
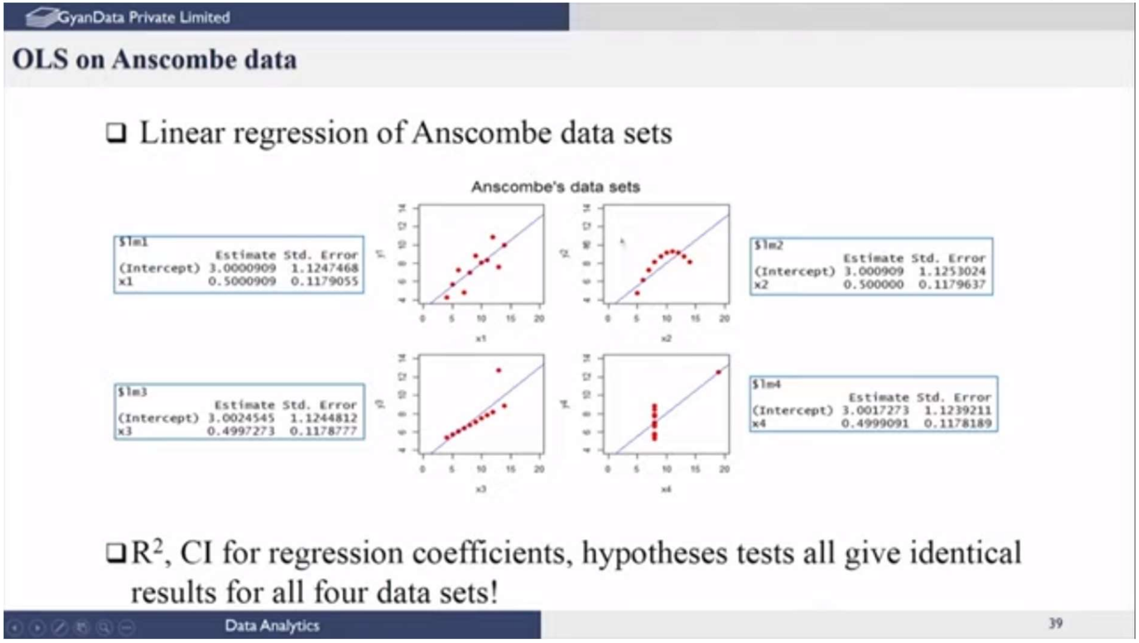


DIAGNOSTICS TO IMPROVE LINEAR MODEL FIT



OLS on Anscombe data

❑ Linear regression of Anscombe data sets

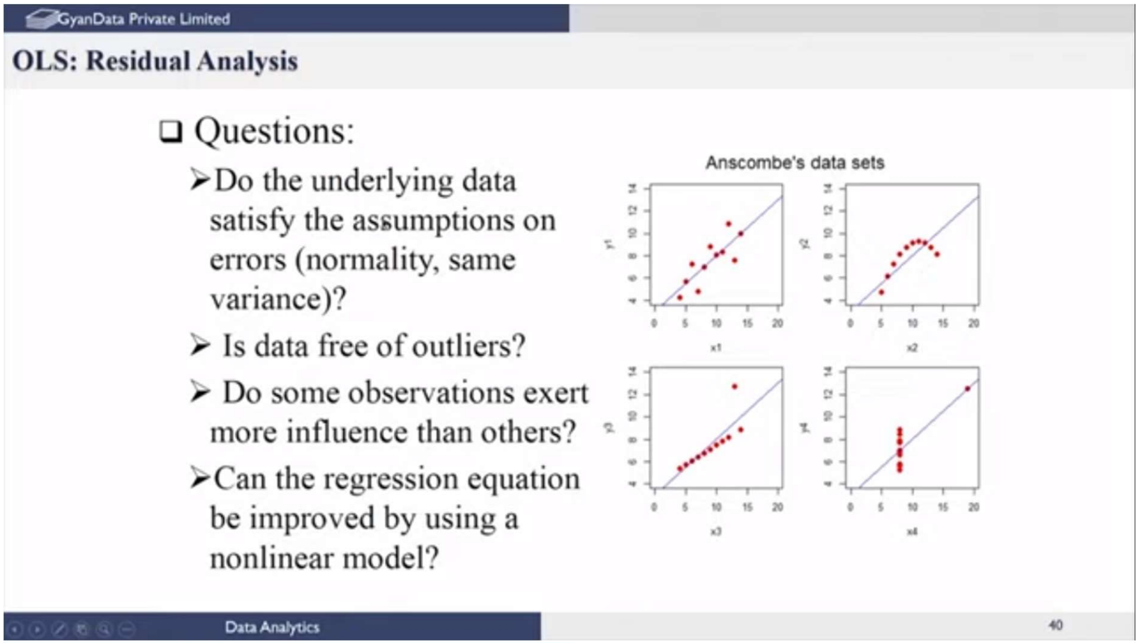
```
$lm1
            Estimate Std. Error
(Intercept) 3.0000909  1.1247468
x1          0.5000909  0.1179055
```

```
$lm2
            Estimate Std. Error
(Intercept) 3.000909  1.1253024
x2          0.500000  0.1179637
```

```
$lm3
            Estimate Std. Error
(Intercept) 3.0024545  1.1244812
x3          0.4997273  0.1178777
```

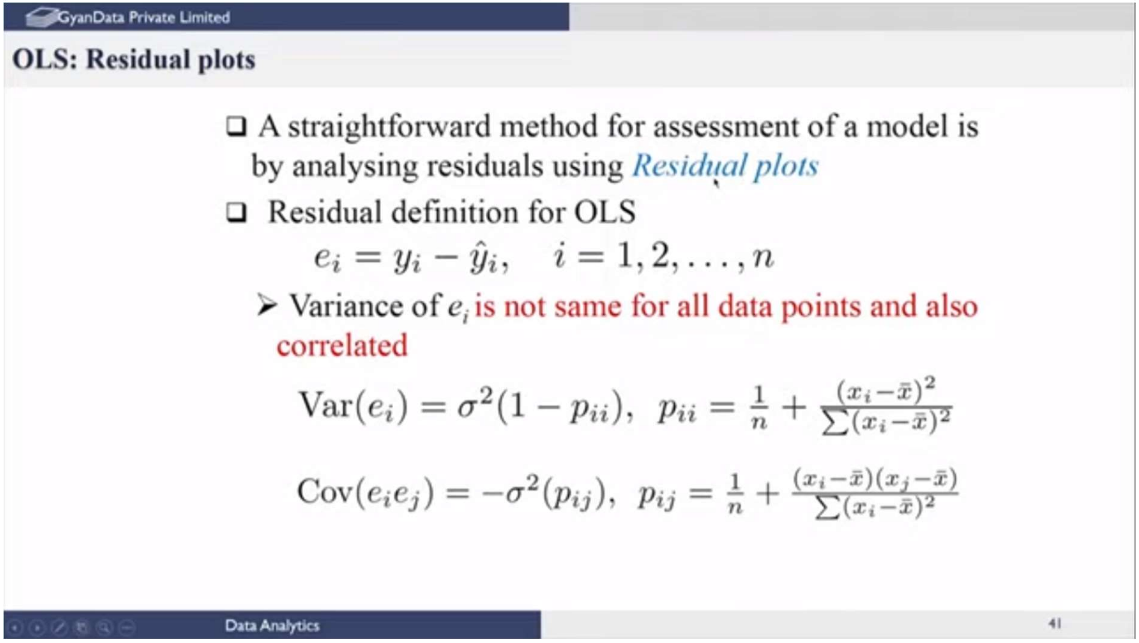
```
$lm4
            Estimate Std. Error
(Intercept) 3.0017273  1.1239211
x4          0.4999091  0.1178189
```

❑ R^2, CI for regression coefficients, hypotheses tests all give identical results for all four data sets!



OLS: Residual Analysis

- ❑ Questions:
  - ➢Do the underlying data satisfy the assumptions on errors (normality, same variance)?
  - ➢ Is data free of outliers?
  - ➢ Do some observations exert more influence than others?
  - ➢Can the regression equation be improved by using a nonlinear model?



OLS: Residual plots

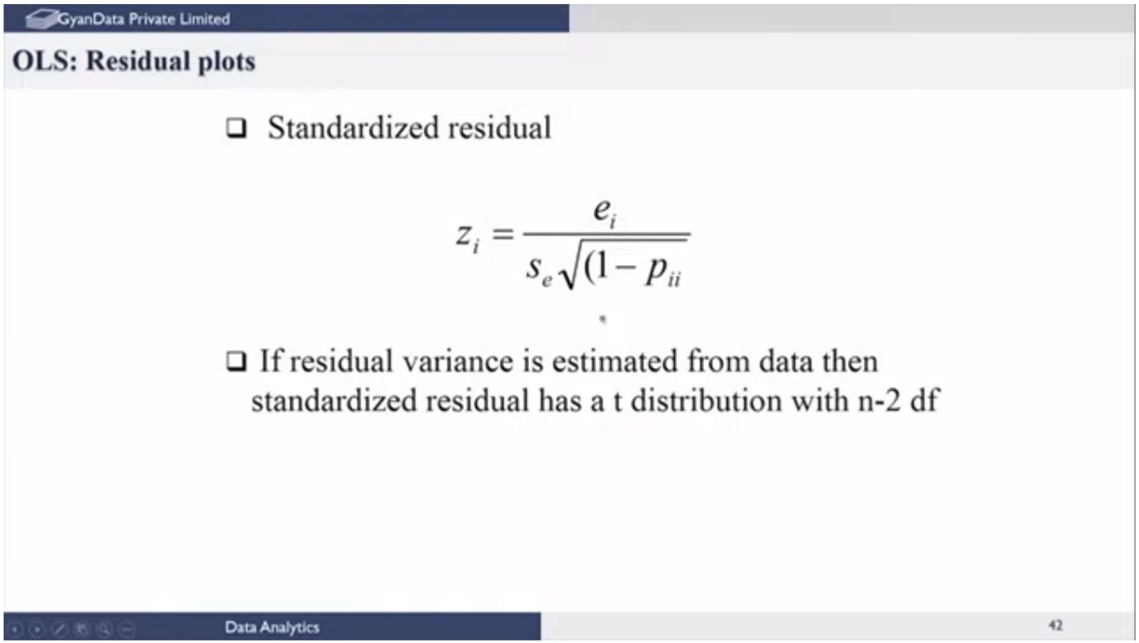
- ❑ A straightforward method for assessment of a model is by analysing residuals using *Residual plots*
- ❑ Residual definition for OLS

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

- ➢ Variance of e_i is not same for all data points and also correlated

$$\mathrm{Var}(e_i) = \sigma^2(1 - p_{ii}), \;\; p_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum(x_i - \bar{x})^2}$$

$$\mathrm{Cov}(e_i e_j) = -\sigma^2(p_{ij}), \;\; p_{ij} = \frac{1}{n} + \frac{(x_i - \bar{x})(x_j - \bar{x})}{\sum(x_i - \bar{x})^2}$$

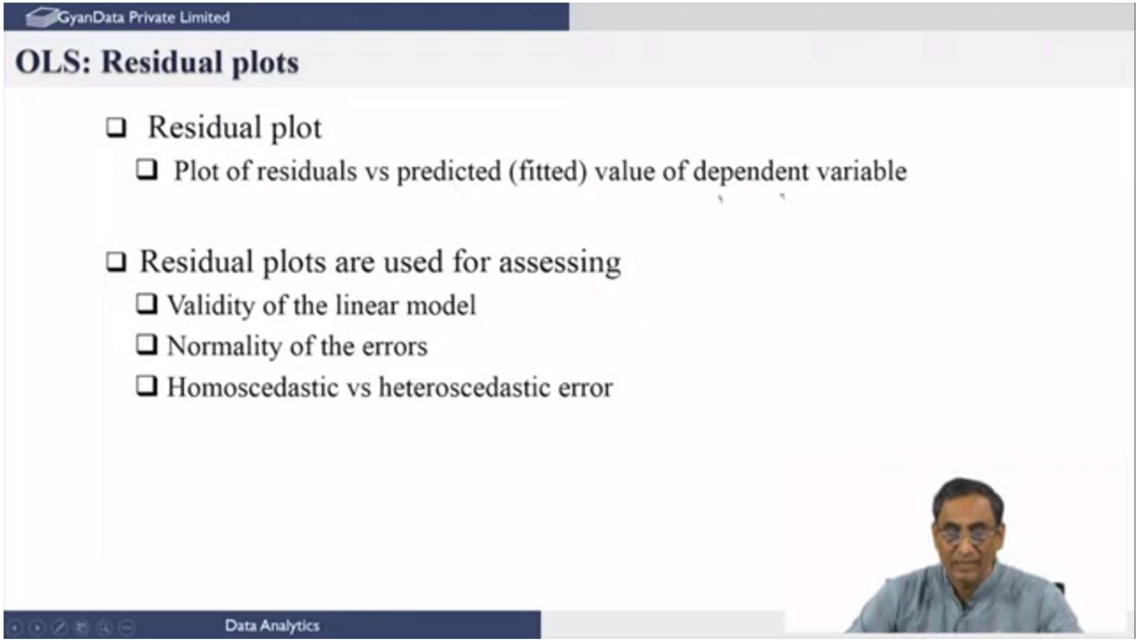


OLS: Residual plots

- ❑ Standardized residual

$$z_i = \frac{e_i}{s_e\sqrt{(1-p_{ii}}}$$

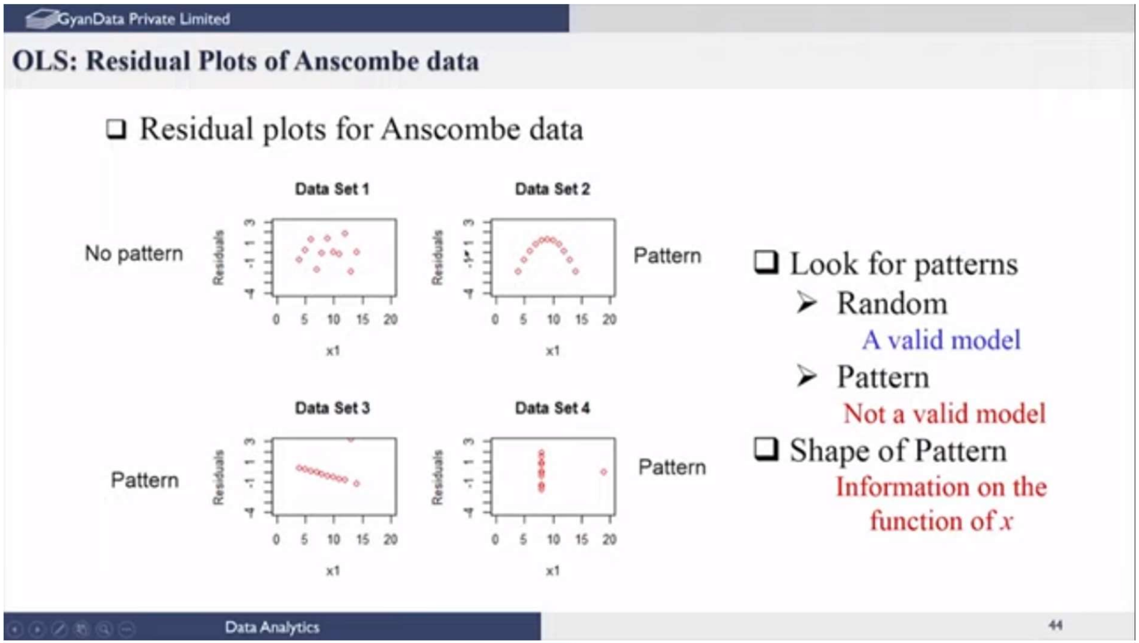
- ❑ If residual variance is estimated from data then standardized residual has a t distribution with n-2 df



OLS: Residual plots

- ❑ Residual plot
  - ❑ Plot of residuals vs predicted (fitted) value of dependent variable

- ❑ Residual plots are used for assessing
  - ❑ Validity of the linear model
  - ❑ Normality of the errors
  - ❑ Homoscedastic vs heteroscedastic error



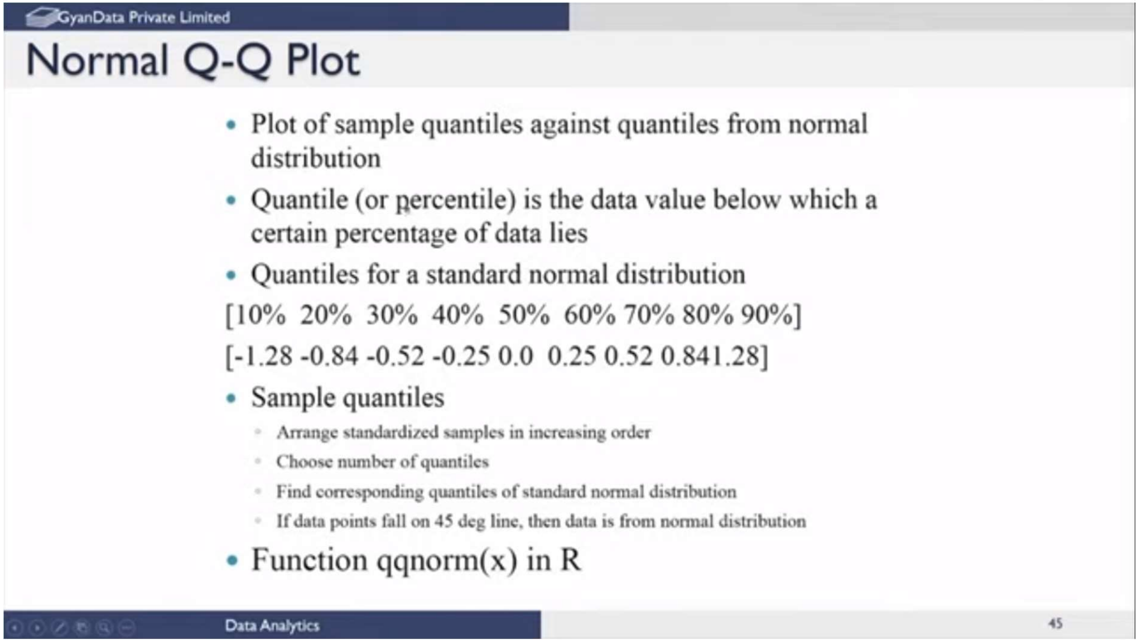
OLS: Residual Plots of Anscombe data

❑ Residual plots for Anscombe data

❑ Look for patterns
- ➢ Random
  - A valid model
- ➢ Pattern
  - Not a valid model

❑ Shape of Pattern

Information on the function of *x*



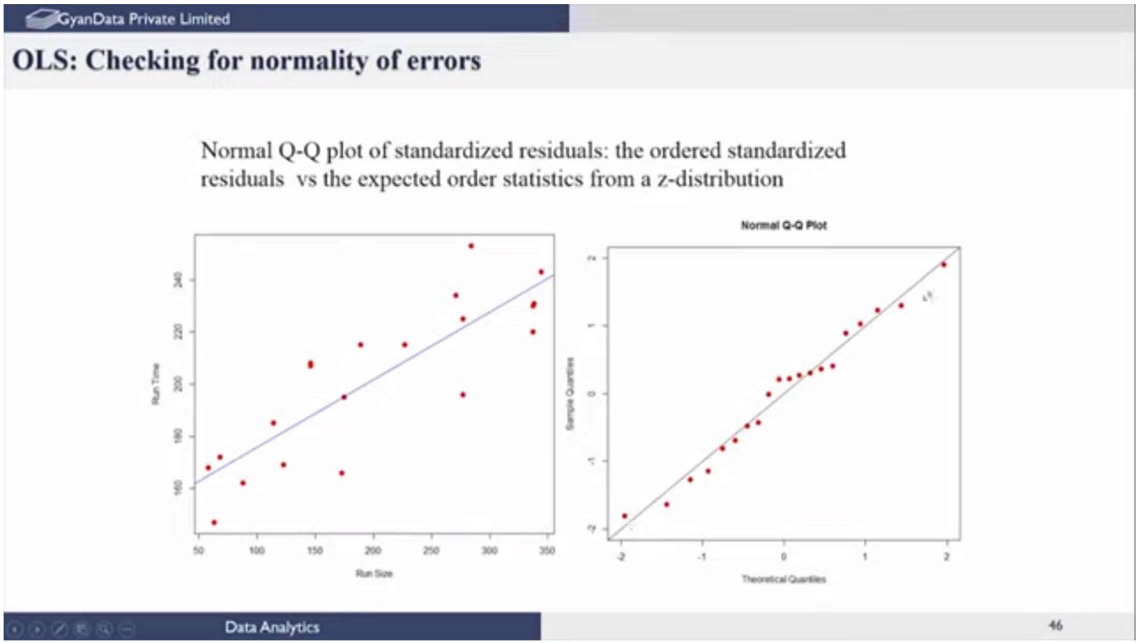
Normal Q-Q Plot

- Plot of sample quantiles against quantiles from normal distribution
- Quantile (or percentile) is the data value below which a certain percentage of data lies
- Quantiles for a standard normal distribution

[10% 20% 30% 40% 50% 60% 70% 80% 90%]

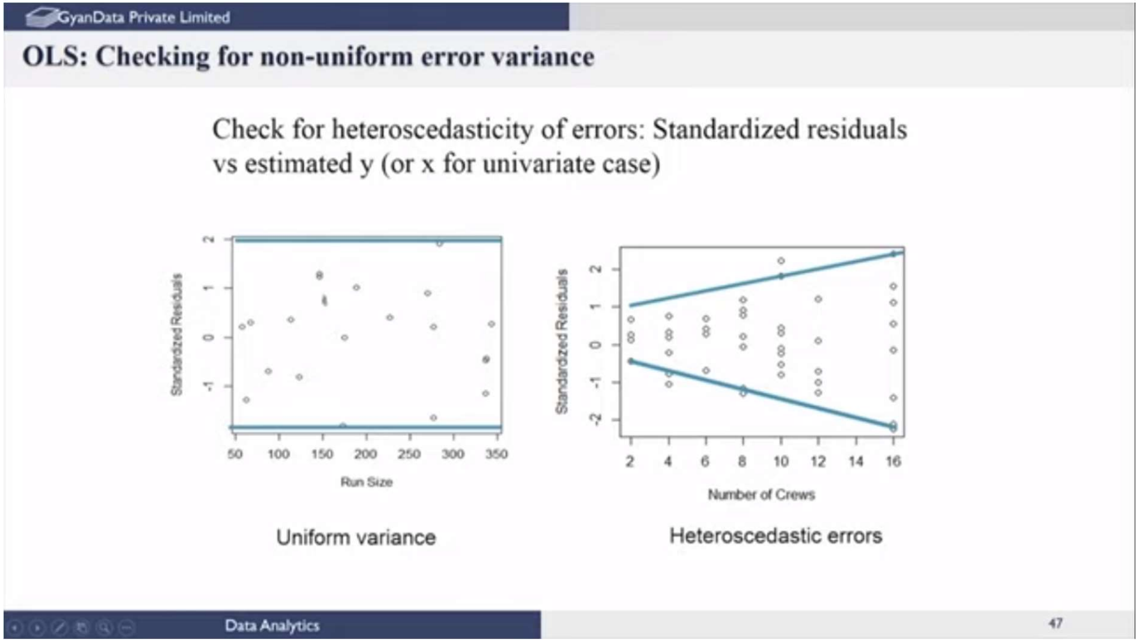
[-1.28 -0.84 -0.52 -0.25 0.0 0.25 0.52 0.841.28]

- Sample quantiles
  - Arrange standardized samples in increasing order
  - Choose number of quantiles
  - Find corresponding quantiles of standard normal distribution
  - If data points fall on 45 deg line, then data is from normal distribution
- Function qqnorm(x) in R



OLS: Checking for normality of errors

Normal Q-Q plot of standardized residuals: the ordered standardized residuals vs the expected order statistics from a z-distribution



OLS: Checking for non-uniform error variance

Check for heteroscedasticity of errors: Standardized residuals vs estimated y (or x for univariate case)

Uniform variance

Heteroscedastic errors

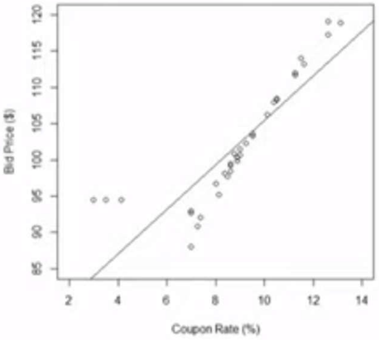
OLS: Checking for outliers in data

- ❑ Outliers: Points which do not conform to the pattern in bulk of the data
- ❑ Outliers can be identified using hypotheses test of residual of each sample
  - For a 5% level of significance a sample is considered an outlier if the corresponding standardized residuals of lie outside [-2, 2]
  - Even if several residuals lie outside confidence region, identify only one outlier at every iteration (corresponding to the sample with largest standard residual magnitude) – An outlier in a sample can 'smear' to other samples due to the regression
  - Apply regression to reduced sample set and iterate until no outliers are detected

OLS: Example for outlier detection

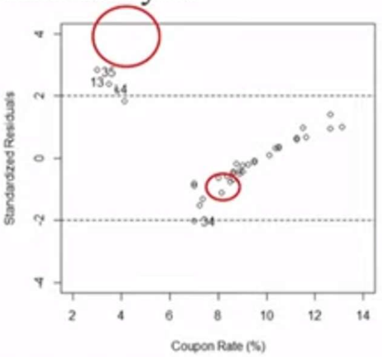
- ❑ US Bonds example (35 samples of coupon rate % of $100 face value bond vs market bid price)

```
Call:
lm(formula = BidPrice ~ CouponRate)

Residuals:
   Min     1Q Median     3Q    Max 
-8.249 -2.470 -0.838  2.550 10.515 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  74.7866     2.8267  26.458  < 2e-16 ***
CouponRate    3.0661     0.3068   9.994 1.64e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.175 on 33 degrees of freedom
Multiple R-squared:  0.7516, Adjusted R-squared:  0.7441 
F-statistic: 99.87 on 1 and 33 DF,  p-value: 1.645e-11
```

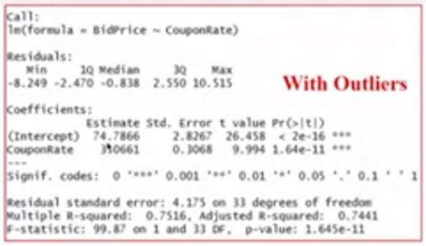


Residual plot for US Bonds data

- ❑ Samples number 4, 13, 34 and 35 are outside CI and can be considered as outliers. We can remove sample 35 and repeat regression analysis

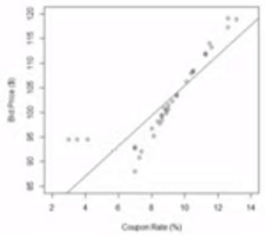


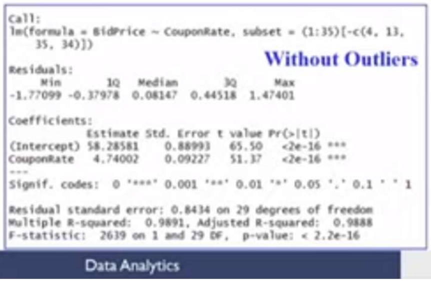
OLS on US bonds example after removing outliers



```
Call:
lm(formula = BidPrice ~ CouponRate)

Residuals:
   Min     1Q Median     3Q    Max 
-8.249 -2.470 -0.838  2.550 10.515 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  74.7866     2.8267  26.458  < 2e-16 ***
CouponRate    3.0661     0.3068   9.994 1.64e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.175 on 33 degrees of freedom
Multiple R-squared:  0.7516, Adjusted R-squared:  0.7441 
F-statistic: 99.87 on 1 and 33 DF,  p-value: 1.645e-11
```

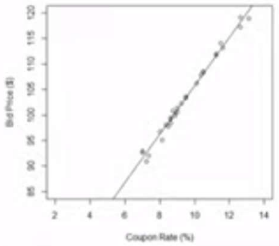
With Outliers





```
Call:
lm(formula = BidPrice ~ CouponRate, subset = (1:35)[-c(4, 13, 
    35, 34)])

Residuals:
     Min       1Q   Median       3Q      Max 
-1.77099 -0.37978  0.08147  0.44518  1.47401 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 58.28581    0.88993   65.50   <2e-16 ***
CouponRate   4.74002    0.09227   51.37   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8434 on 29 degrees of freedom
Multiple R-squared:  0.9891, Adjusted R-squared:  0.9888 
F-statistic:  2639 on 1 and 29 DF,  p-value: < 2.2e-16
```

Without Outliers


Bid Price ($)
Coupon Rate (%)