Simple Linear Regression Model Assessment

First level model assessment- Recap

- ✓ How good is the linear model?
- ✓ Which coefficients of the linear model are significant

In this lecture

- Second level model assessment
  - Can we improve quality of linear model?
  - Are there bad measurements in the data (outliers)

Checking for outliers in data

- Outliers: Points which do not conform to the pattern in bulk of the data
- A point is considered an outlier if the corresponding standardized residuals lies outside [-2, 2] at 5 % level of significance

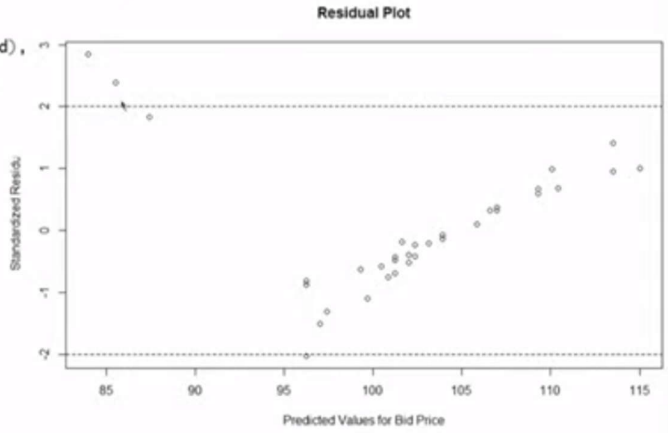
Handling outliers in data

- Even if several residuals lie outside confidence region, identify only one outlier at every iteration
- Apply regression to reduced sample set
- Iterate until no outliers are detected

Residual analysis

```
plot(bondsmod$fitted.values,rstandard(bondsmod),
     main = "Residual Plot",
     xlab = "Predicted Values for Bid Price",
     ylab = "Standardized Residuals")
abline(h=2,lty=2)
abline(h=-2,lty=2)
```

- To know the indices of the outliers we use the function `identify( )`



identify( )

- Reads the position of the graphics pointer when the mouse button is pressed.
- It then searches the coordinates given in x and y for the point closest to the pointer
- If this point is close enough to the pointer, its index will be returned as part of the value of the call

SYNTAX `identify(x,y)`

x, y	coordinates of points in a scatter plot.

Residual analysis- Identifying indices of outliers

```
plot(bondsmod$fitted.values,rstandard(bondsmod),
     main = "Residual Plot",
     xlab = "Predicted Values for Bid Price",
     ylab = "Standardized Residuals")
abline(h=2,lty=2)
abline(h=-2,lty=2)
identify(bondsmod$fitted.values,rstandard(bondsmod))
```

Residual analysis- Identifying indices of outliers

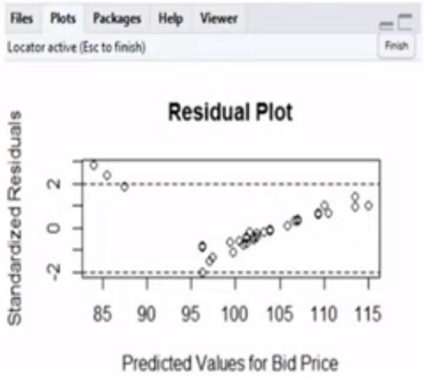
```
identify(bondsmod$fitted.values,rstandard(bondsmod))
```

- Clicking near a point adds it to the list of identified points
- Points can be identified only once
- If the point has already been identified the following message is printed immediately on the R console

```
> identify(bondsmod$fitted.values,rstandard(bondsmod))
warning: nearest point already identified
```

- If the click is not near any of the points then following message is displayed

```
> identify(bondsmod$fitted.values,rstandard(bondsmod))
warning: no point within 0.25 inches
```

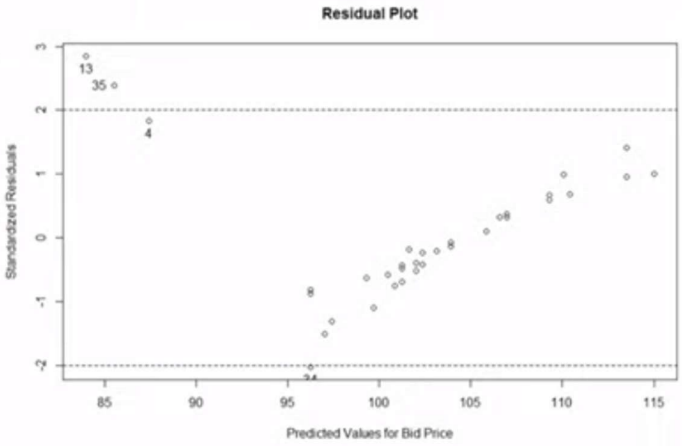


Residual analysis- Identifying indices of outliers

- The identification process is terminated by clicking 'Finish'

- After terminating, the indices are displayed on the console and on the plot

```
> identify(bondsmod$fitted.values,
+          rstandard(bondsmod))
[1]  4 13 34 35
```



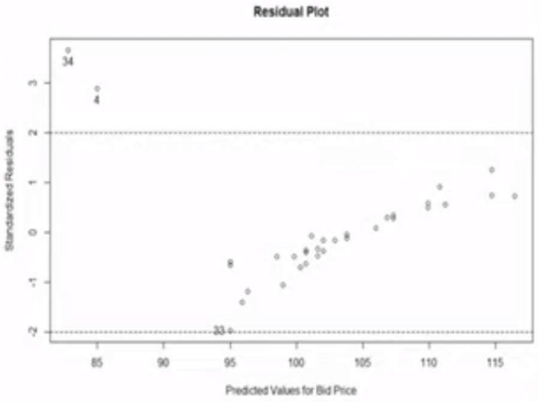
Removing outiers

- Lets start by removing the farthest outlier i.e. sample 13 and building a new model

```
bonds_new<-bonds[-13,]
bondsmod1<-lm(bonds_new$BidPrice~
                bonds_new$CouponRate)
```

- Identify the indices of the outliers on the residual plot

```
> identify(bondsmod1$fitted.values,
+          rstandard(bondsmod1))
[1]  4 33 34
```



Comparison between old and new model

With outliers

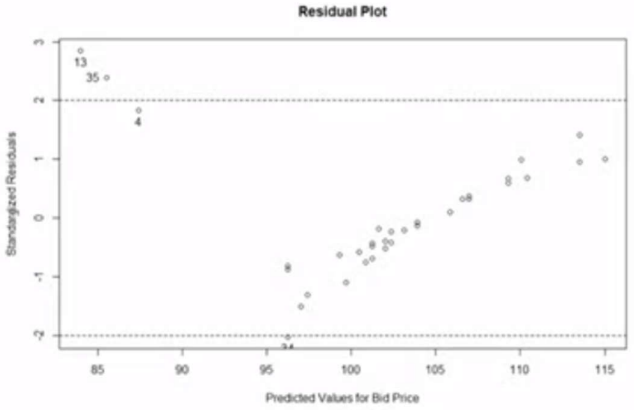
```
> summary(bondsmod)

Call:
lm(formula = BidPrice ~ CouponRate, data = bonds)

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
```

Without sample 13

```
> summary(bondsmod1)

Call:
lm(formula = bonds_new$BidPrice ~ bonds_new$CouponRate)

Residuals:
    Min      1Q  Median      3Q     Max
-7.0393 -1.7780 -0.5931  1.6511 11.7264

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)          70.5679     2.8147   25.07  < 2e-16
bonds_new$CouponRate  3.4959     0.3016   11.59 5.42e-13

(Intercept)          ***
bonds_new$CouponRate ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.683 on 32 degrees of freedom
Multiple R-squared:  0.8077,    Adjusted R-squared:  0.8017
F-statistic: 134.4 on 1 and 32 DF,  p-value: 5.417e-13
```

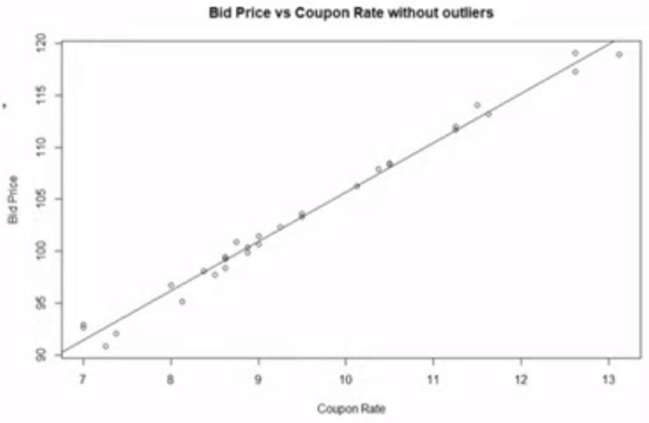
Confirm if further development is needed

- Remove the remaining sample points one by one after removing 13
- After removing,
  - 35th, R^2=0.8846
  - 4th, R^2=0.9852
  - 34th, R^2=0.9891



Plot

```
plot(bonds$CouponRate[-c(4,13,34,35)],
     bonds$BidPrice[-c(4,13,34,35)],
     main = "Bid Price vs Coupon Rate without outliers",
     xlab = "Coupon Rate",
     ylab = "Bid Price")
abline(bondsmod1)
```



Summary

- Steps in building simple linear regression models
- Model summary
- Residual analysis
- Checking need for refinement
- Refined model building