SDS358: Applied Regression Analysis

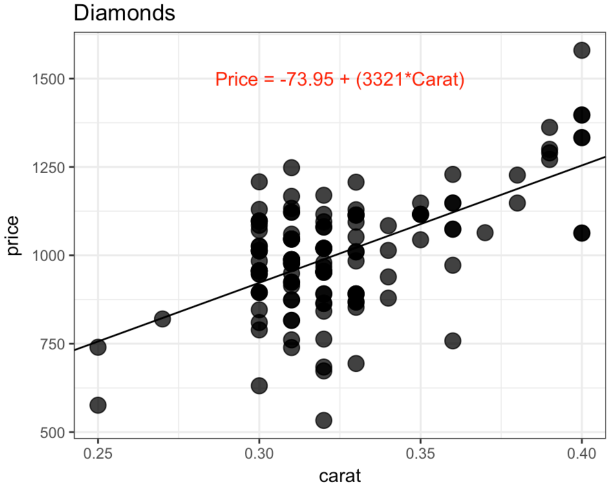
Day 6: SLR Continued

Dr. Michael J. Mahometa

Agenda for Today:

- Simple Linear Regression continued
  - Assumptions
  - Basis in residuals (e_i)
  - Outlier checks

Recall:



Interpretation:
- Slope
- Intercept

Assumptions of Simple Linear Regression

- Quantitative Data for both variables.
- The relationship is linear in nature.
- Residuals
  - Independence: Error associated with each data point is independent of every other value.
  - For a given value of x, e has a normal distribution, meaning:
    - The population mean of e is 0.
    - For a given value of x, the population variance of e is σ_e^2
  - Homoscedasticity
  - No Outliers

Residuals: Independence

- Every observation in the dataset must be unique.
- A subject (or record), cannot be recorded twice, and take two lines.
    - In other words, the same person cannot contribute twice (even with different data) to the prediction of y.

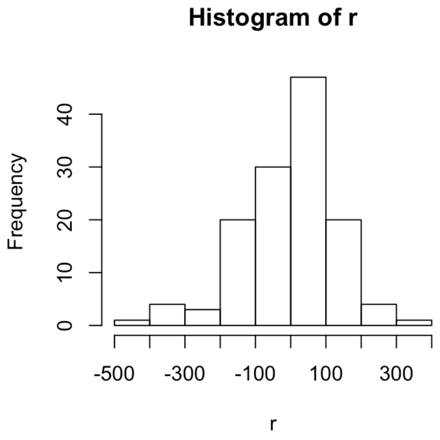
Residuals: Population mean of zero

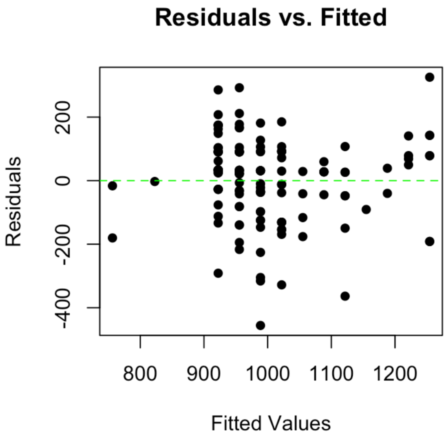
- This is easy: just record each observation's residual after fitting the model.
- Then, grab the mean…or better yet, a histogram

```r
#Diamonds
d_mod <- lm(price ~ carat, data=diamonds)
r <- d_mod$residuals
histogram(r)
residFitted(d_mod)
```

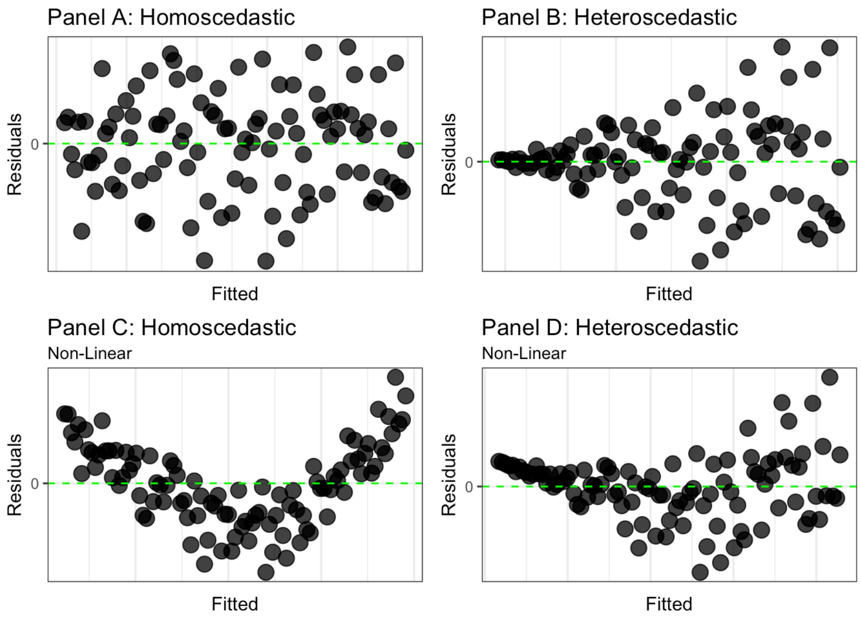
Residuals: Population mean of zero

- This is easy: just record each observation's residual after fitting the model.
- Then, grab the mean…or better yet, a histogram

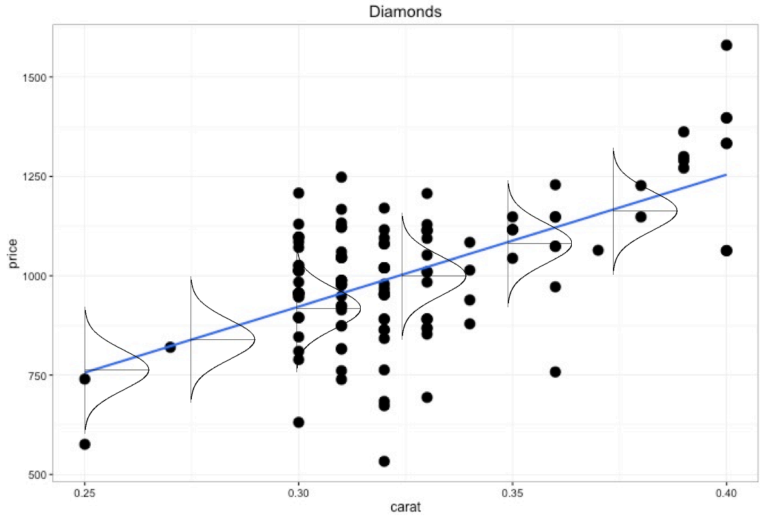




Residuals: What to watch out for….



Redsiduals: normal around each x



Becuase Residuals are NORMAL:

- Measure of "overall error" in the regression model.

$$\text{Model Variance} = \sigma_e^2$$

- We can't actually measure this (it's a population value).
- But, we can estimate it with S_e^2 .

$$S_e^2 = \text{Estimate of } \sigma_e^2$$

What is "sigma"?

- S_e^2 is the variance of the overall error
- S_e is the *average* of the overall error:

$$S_e = \sqrt{S_e^2}$$

- HINT: Think *standard deviation* but for regression model residuals.

What is "sigma"?

- AGAIN: Think *standard deviation* but for regression model residuals.

$$S_e = \sqrt{S_e^2} = \sqrt{\frac{\Sigma(y - \hat{y})^2}{(n - p)}}$$

$$\text{p} = \text{Number of "parameters"}$$

What is "sigma"?

- AGAIN: Think *standard deviation* but for regression model residuals.

$$S_e = \sqrt{S_e^2} = \sqrt{\frac{\Sigma(y - \hat{y})^2}{(n - k - 1)}}$$

$$\text{k} = \text{Number of "predictors"}$$

What else are the residuals good for?

- What's the link between r, r^2, and the linear model?
- Let's head out to RStudio and investigate….

Proportion of Variance

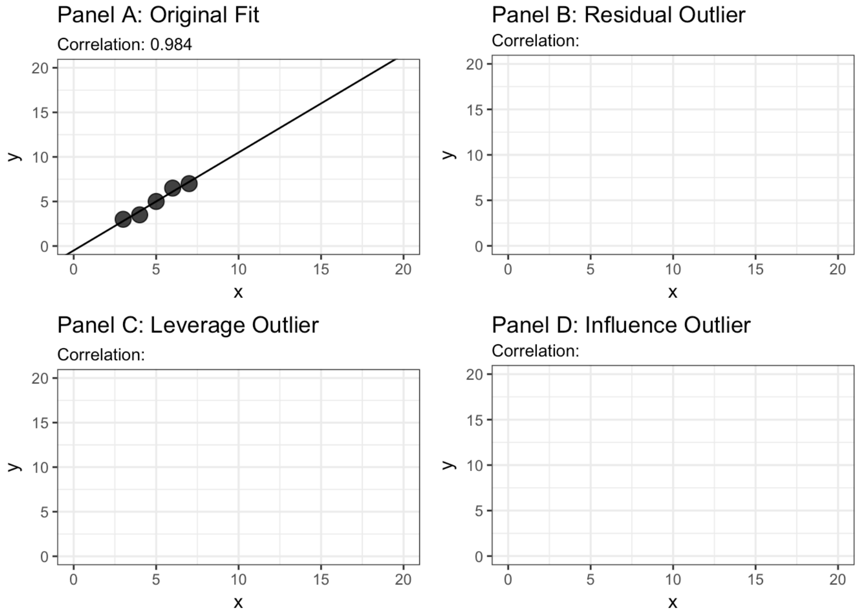
- r^2 is defined as the "Proportion of variance accounted for."
- Literally:

"How much of the variance in y can be accunted for by x (or all x's), now that we know the realtioship between y and x (or y and all x's)."

SLR Diagnostics

- Residuals help show us if something is "off" in our model/data.
- *But* they, are just a part of a larger tool set to diagnose the validity of a model and/or look for problematic data
- We'll use *three* primary tools to look for bad data:
  - Studentized deleted residuals
  - Hat values (also know as leverage points)
  - Cook's Distance

Some invented data….


Panel A: Original Fit

Correlation: 0.984

Panel B: Residual Outlier

Correlation:

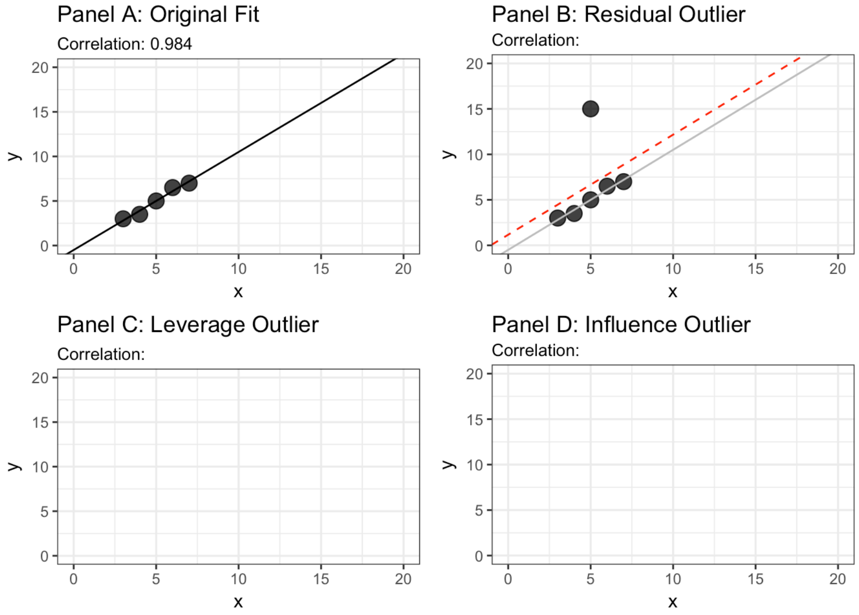
Panel C: Leverage Outlier

Correlation:

Panel D: Influence Outlier

Correlation:

Some invented data….You Guess


Panel A: Original Fit

Correlation: 0.984

Panel B: Residual Outlier

Correlation:

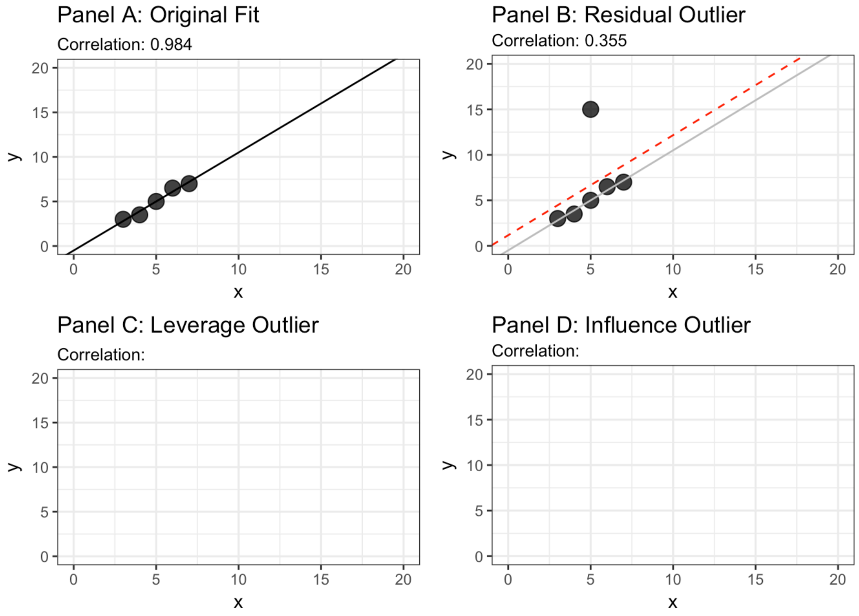
Panel C: Leverage Outlier

Correlation:

Panel D: Influence Outlier

Correlation:

Some invented data….


Panel A: Original Fit

Correlation: 0.984

Panel B: Residual Outlier

Correlation: 0.355

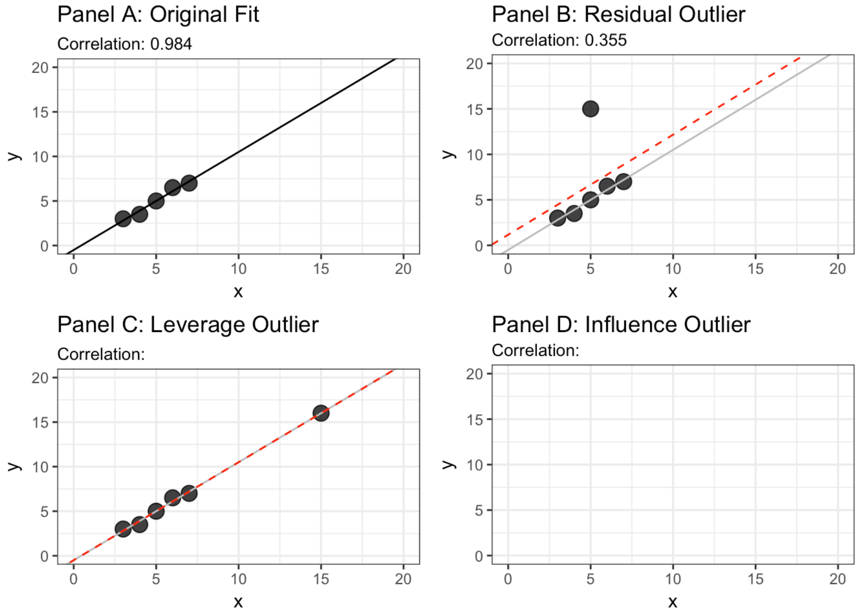
Panel C: Leverage Outlier

Correlation:

Panel D: Influence Outlier

Correlation:

Some invented data….You Guess


Panel A: Original Fit

Correlation: 0.984

Panel B: Residual Outlier

Correlation: 0.355

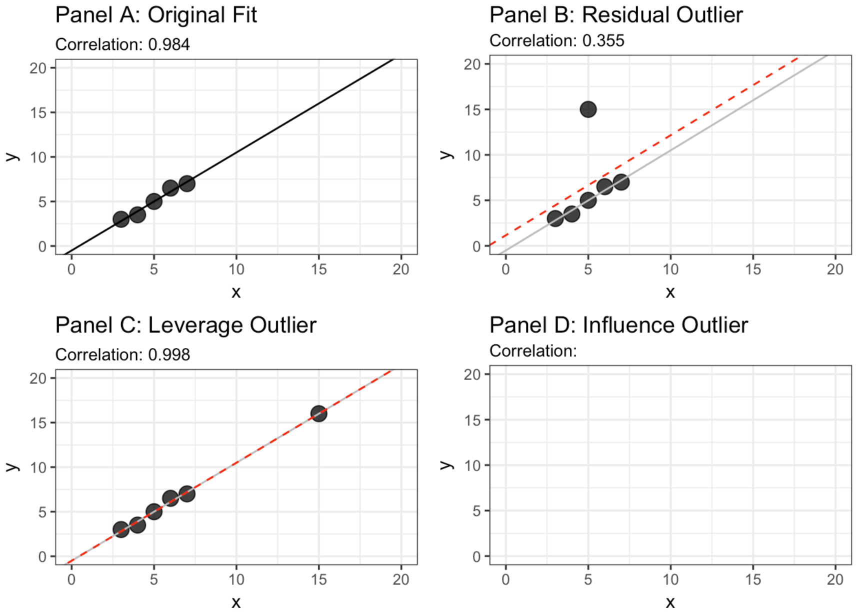
Panel C: Leverage Outlier

Correlation:

Panel D: Influence Outlier

Correlation:

Some invented data…



Panel A: Original Fit

Correlation: 0.984

Panel B: Residual Outlier

Correlation: 0.355

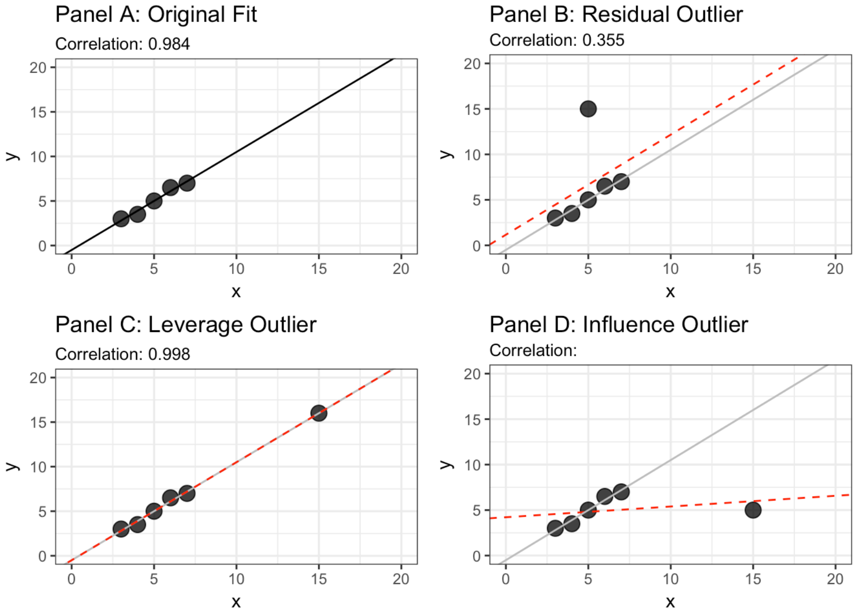
Panel C: Leverage Outlier

Correlation: 0.998

Panel D: Influence Outlier

Correlation:

Some invented data….You Guess



Panel A: Original Fit

Correlation: 0.984

Panel B: Residual Outlier

Correlation: 0.355

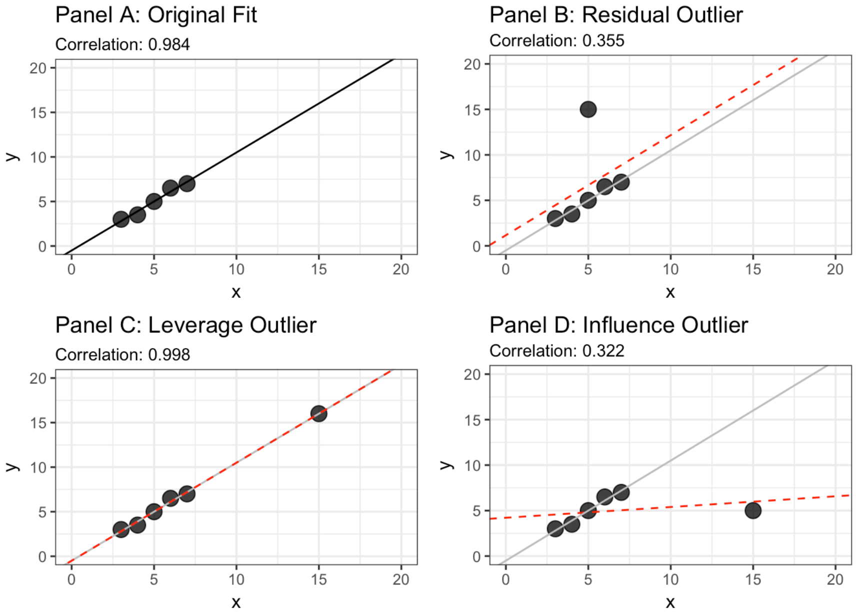
Panel C: Leverage Outlier

Correlation: 0.998

Panel D: Influence Outlier

Correlation:

Some invented data…



SLR Diagnostics

- So which diagnostic do you choose?
  - Which is *most* important?

SLR Diagnostics

- So which diagnostic do you choose?
  - Which is *most* important?
- Cook's distance
  - All three markers would take precedence
  - It's NOT black and white…

SLR Diagnostics: Leverage

- The leverage of a single point is measured by the *hat* value.
- In simple regression:

$$\text{Leverage}_i = h_i = \frac{1}{n} + \frac{(x - \bar{x})^2}{\sum(x - \bar{x})^2}$$

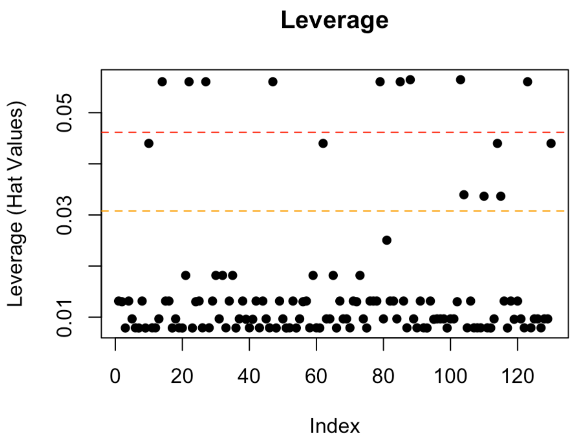
- Think: "The farther away from the mean of x, the higher the hat value."
- Look for: 2 or 3 time the "average h_i "

$$h_i > 2 * ((k + 1)/n)$$

$$h_i > 2 * \left(\frac{\text{Number of } \beta \text{ predictors in the model incl. } \beta_0}{n} \right)$$

SLR Diagnostics: Leverage

```r
levPlot(d_mod)
```



SLR Diagnostics: Leverage

```r
levPlot(d_mod, key.variable = "sampleDiamond",
        print.obs=TRUE, sort.obs = TRUE)
```

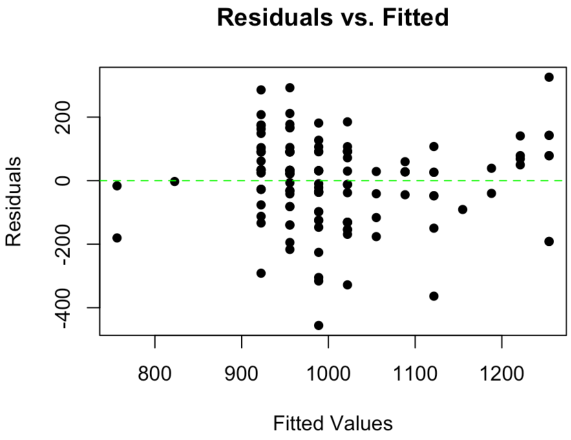
```
##   sampleDiamond price carat Predicted_Y Hat_Values
## 1            88   740  0.25    756.2989 0.05644717
## 2           103   576  0.25    756.2989 0.05644717
## 3            14  1397  0.40   1254.4485 0.05604876
## 4            22  1333  0.40   1254.4485 0.05604876
## 5            27  1397  0.40   1254.4485 0.05604876
## 6            47  1580  0.40   1254.4485 0.05604876
## 7            79  1063  0.40   1254.4485 0.05604876
## 8            85  1333  0.40   1254.4485 0.05604876
## 9           123  1063  0.40   1254.4485 0.05604876
```

SLR Diagnostics: Residuals

- Looking at residuals is good too.
- Even looking at fitted values and examining values them in relation to some error of e (*Homoscedacticity*)

SLR Diagnostics: Residuals

```
residFitted(d_mod)
```



SLR Diagnostics: Standardized Residuals

- But, there's a **problem** with residuals: *how to evaluate them.*
- Comparing them to a value like a *z-score* might work.
- Then we can use the *sigma* of the model (the RMSE = 134.68):

$$S_e = \sqrt{\frac{(y - \hat{y})^2}{n - k - 1}}$$

- If we use this though, we can run into an issue: what happens to sigma with an observation with a large residual?

SLR Diagnostics: Studentized Residuals

- On *TOP* of that, we know that Leverage plays a key role in influencing the SLR Model.
- So, we should recognize that the RMSE (s_e) is, *by itself* not great as a denominator.
- We can *incorprate* Leverage into the denominator and get:

$$\text{Studentized Residual}_i = E_i' = \frac{e_i}{S_e\sqrt{1 - h_i}}$$

with:

$$e_i = (y - \hat{y})^2$$

SLR Diagnostics: Studentized DELETED Residuals

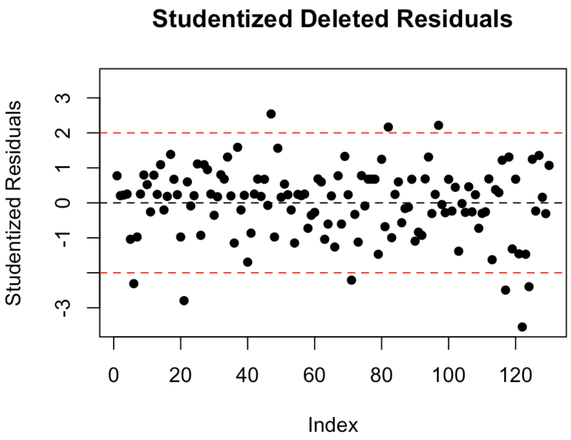
- Studentized Residuals are *better* than Standardized, but there's *still the issue* of a high residual outlier.
- To help "correct" for the effect of a single problematic residual, while still taking into account for Leverage, we can use a better calculation—through *DELETION*.

$$\text{Studentized Deleted Residual}_i = E_i^* = \frac{e_i}{S_{e(-i)}\sqrt{1-h_i}}$$

- Look for: 2 times a "studentized" residual (absolute value).

SLR Diagnostics: Studentized DELETED Residuals

```
studResidPlot(d_mod)
```



SLR Diagnostics: Studentized DELETED Residuals

```r
studResidPlot(d_mod, key.variable = "sampleDiamond",
              print.obs = TRUE, sort.obs = TRUE)
```

```
##   sampleDiamond price carat Predicted_Y Student_Resid
## 1            47  1580  0.40   1254.4485      2.540426
## 2            97  1248  0.31    955.5587      2.215025
## 3            82  1208  0.30    922.3488      2.165628
## 4            71   631  0.30    922.3488     -2.210468
## 5             6   684  0.32    988.7687     -2.310115
## 6           124   673  0.32    988.7687     -2.397198
## 7           117   694  0.33   1021.9787     -2.494309
## 8            21   758  0.36   1121.6086     -2.796336
## 9           122   533  0.32    988.7687     -3.548077
```

SLR Diagnostics: Cook' Distance

- Measure of "Influence": Leverage x "Outlyingness"

$$\text{Cook's Distance}_i = D_i = \frac{h_i}{1 - h_i} \times \frac{E_i'^2}{k+1}$$

- Actually measures the change in estimates when i^{th} observation is removed.
- Rule of thumb:

$$D_i > 4/(n - k - 1)$$

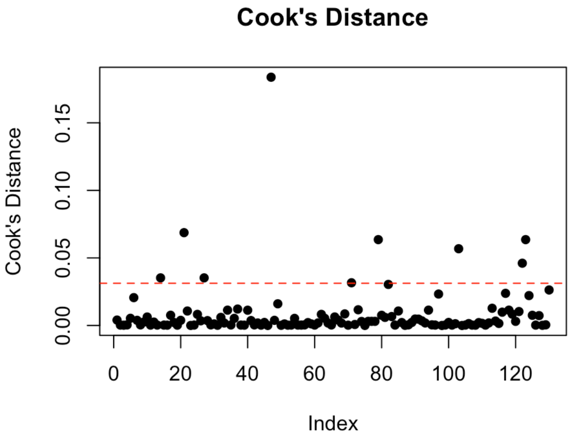
$$D_i > 4/(n - p)$$

SLR Diagnostics: Cook' Distance

- Another good option for evaluating Cook's Distance: the F-distribution.
- Use (p) and $(n - p)$ degrees of freedom (that's $(k + 1)$ and $(n - k - 1)$).
- If D_i falls *NEAR* or *ABOVE* the 50th percentile, then it is *most likely* influential.
- If D_i falls *BELOW* the 50th percentile, but *ABOVE* the 20th percentile, then it *may be* influential.
- If D_i falls *BELOW* the 20th percentile, then it's *not* influential.
- But, really, a good visual examination takes care of it.

SLR Diagnostics: Cook' Distance

```
cooksPlot(d_mod)
```



SLR Diagnostics: Cook' Distance

```
cooksPlot(d_mod, key.variable = "sampleDiamond",
          print.obs = TRUE, sort.obs = TRUE)
```

```
##   sampleDiamond price carat Predicted_Y Cooks_Distance      F_per
## 1            47  1580  0.40   1254.4485     0.18377167 0.16765506
## 2            21   758  0.36   1121.6086     0.06870973 0.06636795
## 3            79  1063  0.40   1254.4485     0.06355401 0.06154698
## 4           123  1063  0.40   1254.4485     0.06355401 0.06154698
## 5           103   576  0.25    756.2989     0.05681566 0.05520798
## 6           122   533  0.32    988.7687     0.04608736 0.04502562
## 7            14  1397  0.40   1254.4485     0.03523569 0.03461278
## 8            27  1397  0.40   1254.4485     0.03523569 0.03461278
## 9            71   631  0.30    922.3488     0.03160504 0.03110326
```

SLR Diagnostics

- So which diagnostic do you choose?
- All three

SLR Diagnostics

- So which diagnostic do you choose?
- All three

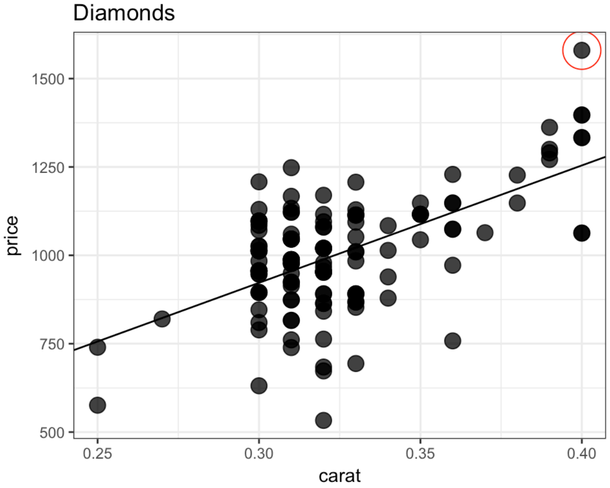
```
threeOuts(d_mod, key.variable = "sampleDiamond",)
```

```
##    sampleDiamond Student_Resid Hat_Values Cooks_Distance      F_per
## 1             47      2.540426 0.05604876     0.18377167 0.16765506
## 2             21     -2.796336         NA     0.06870973 0.06636795
## 3             79            NA 0.05604876     0.06355401 0.06154698
## 4            123            NA 0.05604876     0.06355401 0.06154698
## 5            103            NA 0.05644717     0.05681566 0.05520798
## 6            122     -3.548077         NA     0.04608736 0.04502562
## 7             14            NA 0.05604876     0.03523569 0.03461278
## 8             27            NA 0.05604876     0.03523569 0.03461278
## 9             71     -2.210468         NA     0.03160504 0.03110326
## 10             6     -2.310115         NA             NA         NA
```

"How much of an Influence?"

- When an observation has a high Cook's value (or is in general, an outlier), how much does it effect the model?
- Row 47 is our current "problem"
    - In all three measures of "outliers"
    - High Cook's value

Observation 47



Re-run without the observation:

- Remove the problem observation(s):

```
diamonds_noout <- diamonds %>%
  filter(sampleDiamond %not in% c(47))
```

Re-run without the observation:

- Re-run the model

```
d_mod_noout <- lm(price ~ carat, data=diamonds_noout)
d_mod_noout
```

```
##
## Call:
## lm(formula = price ~ carat, data = diamonds_noout)
##
## Coefficients:
## (Intercept)        carat
##      -4.152     3098.177
```

Re-run without the observation:

- With the "outlier":

```
##
## Call:
## lm(formula = price ~ carat, data = diamonds)
##
## Coefficients:
## (Intercept)        carat
##      -73.95      3321.00
```

- Without the "outlier":

```
##
## Call:
## lm(formula = price ~ carat, data = diamonds_noout)
##
## Coefficients:
## (Intercept)        carat
##      -4.152     3098.177
```

Re-run without the observation:

- What about σ_e ?
- With outlier:

```
summary(d_mod)$sigma
```

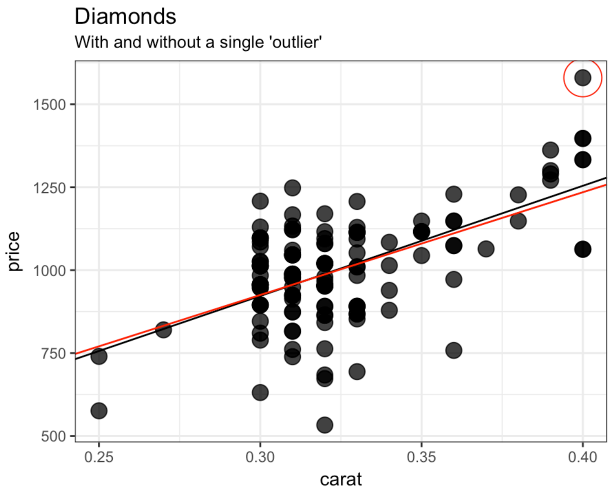
```
## [1] 134.6787
```

- Without outlier:

```
summary(d_mod_noout)$sigma
```

```
## [1] 131.8981
```

Re-run without the observation:



Summary

- Simple Linear Regression
- Interpretation and tie in to Pearson r
- Residuals and their usefulness
- Diagnostics Tools
- Types and Impact of Outliers
- Dealing Outliers with and Reporting