STA302 METHODS OF DATA ANALYSIS 1

MODULE 8: THE PROBLEM WITH PROBLEMATIC OBSERVATIONS

PROF. KATHERINE DAIGNAULT

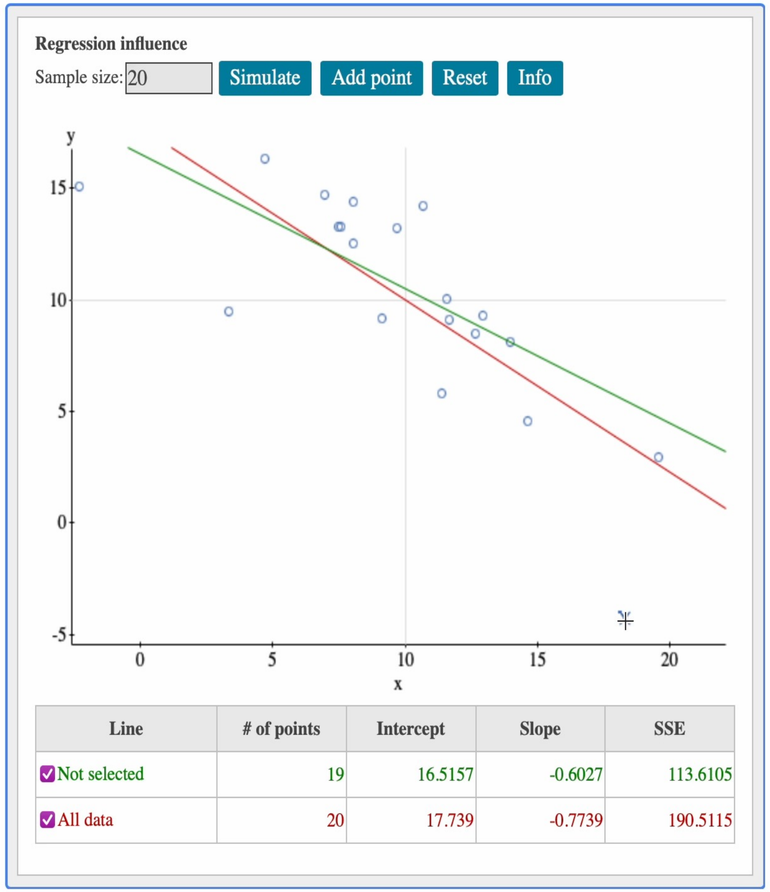
MODULE 8 OUTLINE

1. Leverage Observations
2. Outlying Points in Regression
3. Influential Observations
4. Assessing and Addressing Problematic Observations

WHY SHOULD WE LOOK FOR PROBLEMATIC POINTS?

- Likely familiar with notion of a statistical outlier
  - e.g., outlier in a box plot, defined based on distance from median
- Statistical outlier only one example of a problematic observation
  - single observation impacting ability to make accurate inference about parameter
  - e.g., affects accurate estimation of mean and variance
- In regression, many parameter estimates potentially impacted by individual observations
  - individual coefficients, individual fitted values, entire regression surface
- Need to identify points with potential to disproportionately impact estimation to understand reliability of estimates

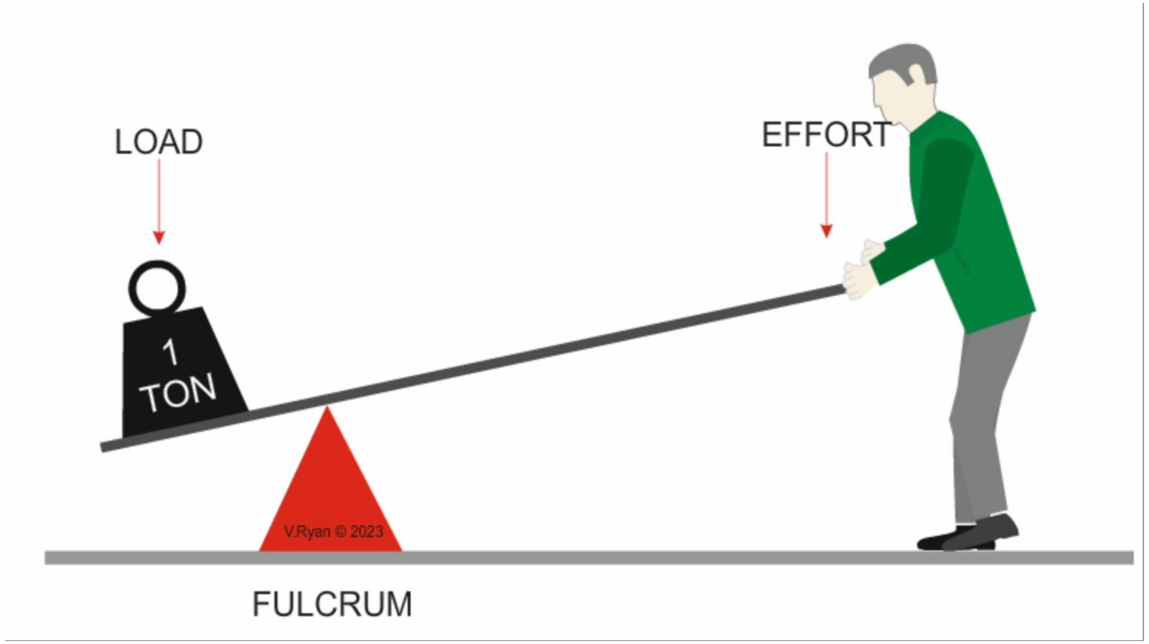
Regression influence



CONCEPT OF LEVERAGE

- A leverage observation is an observation very distant from the center of the $\boldsymbol{X}$-space that may change $\widehat{Y}$
  - i.e., far away in the horizontal $\boldsymbol{X}$ distance in SLR
- Works like a lever – move a weight with little effort because balanced on a fulcrum
  - fulcrum here is the middle of all the predictors
  - board/lever is our regression surface
  - leverage points are far from fulcrum/middle, so they require very little effort to shift the estimated regression line.
- Leverage points have potential to shift the regression line, but will not always do it.

https://www.technologystudent.com/forcmom/lever1.htm#google_vignette



HAT MATRIX IS KEY

- Need to be able to measure distances between various predictor values observed and the center of all predictors to measure leverage
  - but also incorporate how the estimated trend could change as a result
- The hat matrix $\boldsymbol{H}$ is the key: $\widehat{\boldsymbol{Y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y} = \boldsymbol{H}\boldsymbol{Y}$
  - $\boldsymbol{H}$ is a projection matrix, projecting $\boldsymbol{Y}$ onto the model space through $\boldsymbol{X}$
- The hat matrix turns $\boldsymbol{Y}$ into $\widehat{\boldsymbol{Y}}$ through matrix multiplication:

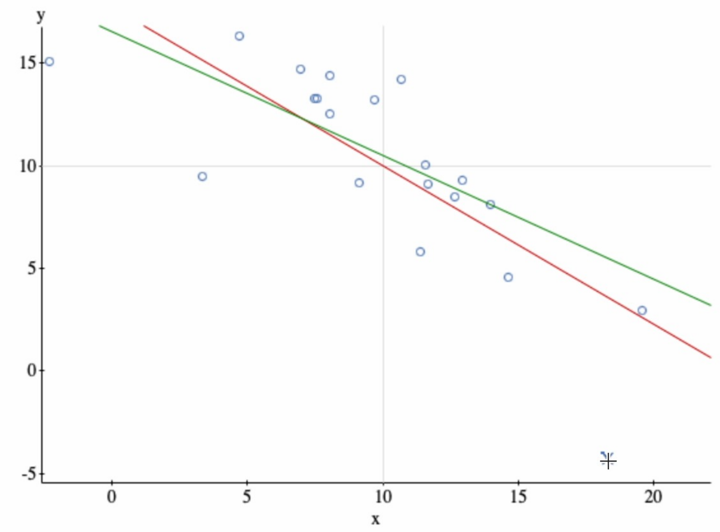
$$\widehat{\boldsymbol{Y}} = \boldsymbol{H}\boldsymbol{Y} = \begin{pmatrix} h_{11} & h_{12} & \cdots & h_{1n} \\ h_{21} & h_{22} & \cdots & h_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ h_{n1} & h_{n2} & \cdots & h_{nn} \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_n \end{pmatrix}$$

- Each fitted value $\hat{y}_i$ is a linear combination of all y_i's with coefficients h_{ij}

Regression influence

Regression influence

Sample size: 20 Simulate Add point Reset Info



Line	# of points	Intercept	Slope	SSE
Not selected	19	16.5157	-0.6027	113.6105
All data	20	17.739	-0.7739	190.5115

HAT MATRIX IS KEY (CONTINUED)

$$\hat{\boldsymbol{Y}} = \boldsymbol{HY} = \begin{pmatrix} h_{11} & h_{12} & \cdots & h_{1n} \\ h_{21} & h_{22} & \cdots & h_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ h_{n1} & h_{n2} & \cdots & h_{nn} \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_n \end{pmatrix}$$

- All responses used to compute each fitted value:

$$\hat{y}_i = h_{i1}y_1 + h_{i2}y_2 + \cdots + h_{in}y_n$$

- Separate the effect observation i has on its own mean response estimate by $\hat{y}_i = h_{ii}y_i + \sum_{j\neq i} h_{ij}y_j$
- Diagonal elements h_{ii} are the leverage of observation i
  - says how much impact the value y_i has on $\hat{y}_i$ versus the other $n-1$ responses

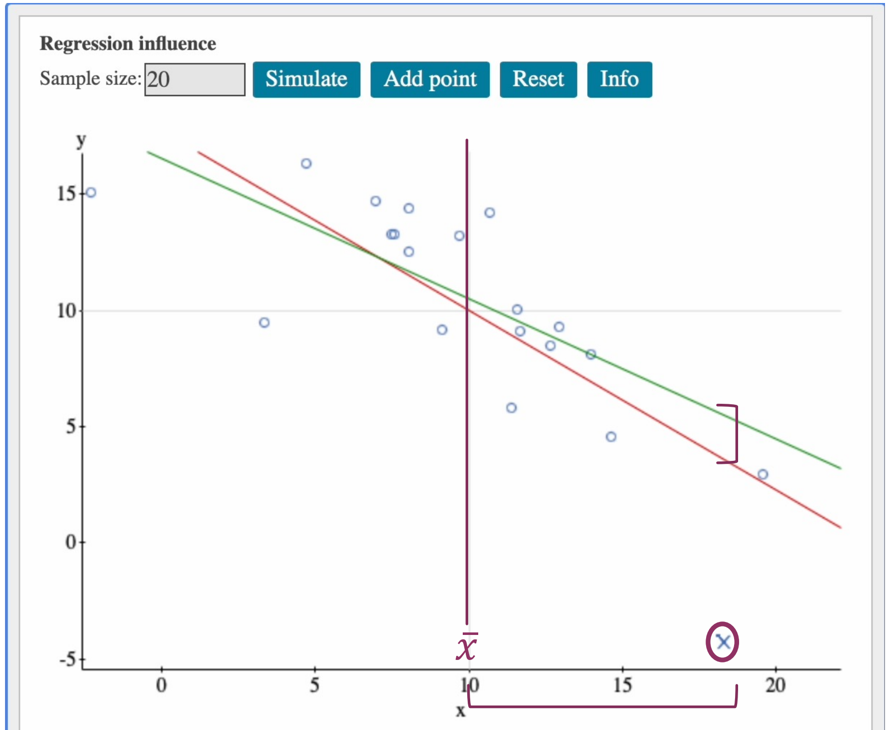
- In SLR, the h_{ii} have a nice expression:

$$h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{j=1}^{n}(x_j - \bar{x})^2}$$

  - ratio of distance of own x-value from center to total variation in predictor
- Some other properties of the hat matrix elements:
  - $\sum_{i=1}^{n} h_{ii} = p+1$, so the average leverage is ${}^{(p+1)}/_{n}$
  - $\sum_{j=1}^{n} {h_{ij}}^2 = h_{ii}$ because H is idempotent
  - as a result, $0 \le h_{ii} \le 1$, so it tells us the fraction of $\hat{y}_i$ due to y_i versus the other responses

LEVERAGE OF OBSERVATION i

- We saw h_{ii} tells us how much $\hat{y}_i$ is driven by the value of y_i
  - namely how much the regression line is potentially attracted to this one observation because of its distance from the center
- E.g., if $h_{ii} \approx 1$, then other $h_{ij} \approx 0$
  - says observation i is really far away from rest of data
  - means $\hat{y}_i = h_{ii}y_i + \sum_{j\neq i} h_{ij}y_j \approx y_i$ so the estimated line was pulled to directly intersect this point
    - therefore, a different line may have been estimated if this observation wasn't used!
- High leverage (h_{ii} close to 1) does not always mean the line has shifted – it just has the potential to do so



Line	# of points	Intercept	Slope	SSE
Not selected	19	16.5157	-0.6027	113.6105
All data	20	17.739	-0.7739	190.5115

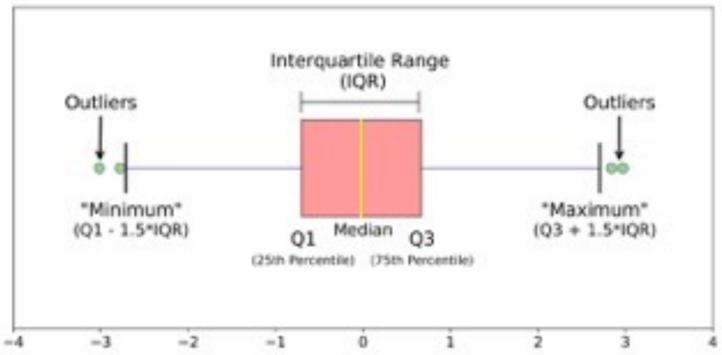
MODULE 8 OUTLINE

1. Leverage Observations
2. Outlying Points in Regression
3. Influential Observations
4. Assessing and Addressing Problematic Observations

OUTLIERS IN REGRESSION

- In addition to leverage points, we want to also identify outlying points (i.e., outliers)
- Statistical outliers are those far from the outer quartiles
  - Regression outliers are those far from the trend/conditional means
- Distance between observed response and trend/fitted value is just the residual
  - trend/fitted value: $\widehat{\boldsymbol{Y}} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y} = \boldsymbol{HY}$
  - residual: $\hat{\boldsymbol{e}} = \boldsymbol{Y} - \widehat{\boldsymbol{Y}} = \boldsymbol{Y} - \boldsymbol{HY} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}$
- Residuals are just observed responses weighted by the leverage
  - can be used to measure potential of each observation to attract the regression line

https://medium.com/@agarwal.vishal819/outlier-detection-with-boxplots-1b6757fafa21



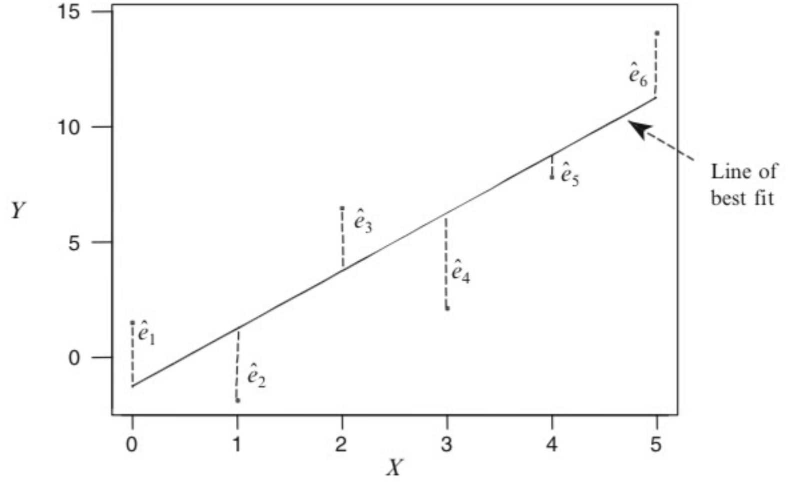


Figure 2.2 A scatter plot of data with a line of best fit and the residuals identified

From Sheather's A Modern Approach to Regression with R

LEVERAGE AND RESIDUALS

- We saw earlier that each fitted value was a weighted sum of responses: $\hat{y}_i = h_{ii}y_i + \sum_{j\neq i} h_{ij}y_j$
- The residuals involve subtracting the identity matrix from the hat matrix:

$$(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y} = \begin{pmatrix} 1-h_{11} & -h_{12} & \cdots & -h_{1n} \\ -h_{21} & 1-h_{22} & \cdots & -h_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ -h_{n1} & -h_{n2} & \cdots & 1-h_{nn} \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

- Each residual also becomes a weighted sum of responses:

$$\hat{e}_i = -h_{i1}y_1 - h_{i2}y_2 - \cdots + (1-h_{ii})y_i - \cdots - h_{in}y_n$$

- Simplifying, we find $\hat{e}_i = (1-h_{ii})y_i - \sum_{j\neq i} h_{ij}y_j$
- Since $0 \leq h_{ii} \leq 1$, we find the impact of response y_i on $\hat{e}_i$ depends on its distance in the X-space h_{ii}
  - the higher the leverage (i.e., farther from center), the lower the weight this observation has on residual
  - means observations with large h_{ii} tend to have small residuals
- As a result, our residuals do not have constant variance:

$$Cov(\hat{\boldsymbol{e}}|\boldsymbol{X}) = Cov((\boldsymbol{I}-\boldsymbol{H})\boldsymbol{Y}|\boldsymbol{X}) = \sigma^2(\boldsymbol{I}-\boldsymbol{H})$$

*******where properties of transposes and projection matrices were used*

STANDARDIZED RESIDUALS

- The covariance of the residuals depends on the leverage: $Cov(\hat{\boldsymbol{e}}|\boldsymbol{X}) = \sigma^2(\boldsymbol{I} - \boldsymbol{H})$
  - since leverage incorporates predictor information only, the covariance depends on predictors
    - then $Var(\hat{e}_i|\boldsymbol{x}_i) = \sigma^2(1 - h_{ii})$
- Want the residuals to have the same properties we assume about the errors
- To achieve this, we can standardize each $\hat{e}_i$ by its variance:

$$r_i = \frac{\hat{e}_i}{s\sqrt{1 - h_{ii}}}$$

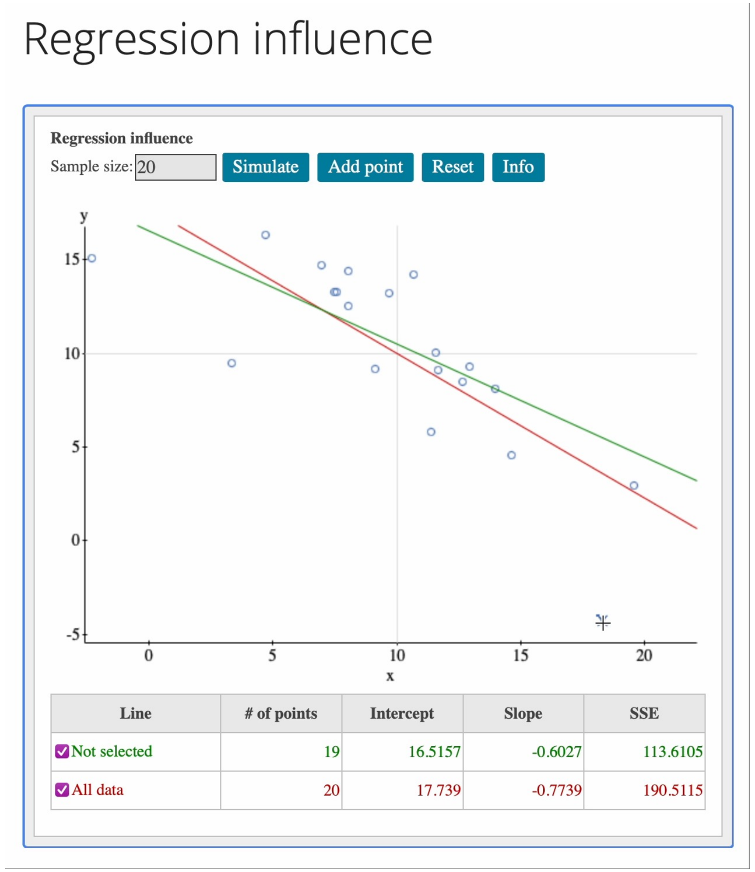
- These standardized residuals r_i will have constant variance
  - combines leverage information, residual value, and estimated error variance
- Standardized residuals more fairly measure "outlying-ness" by adjusting for leverage
  - tells us how many standard deviations from estimated trend
- Also useful for checking assumptions
  - violations tend to be more prominent at remote points but harder to detect with normal residuals
  - presence of high leverage points impact detection of model violations, so use standardized residuals

MODULE 8 OUTLINE

1. Leverage Observations
2. Outlying Points in Regression
3. Influential Observations
4. Assessing and Addressing Problematic Observations

INTRODUCTION TO INFLUENCE

- Both leverage points and outliers have the potential to change the estimated regression relationship
- Problematic observation can influence how the model is estimated in three ways:
  - affect how all fitted values are estimated
  - affect how its own fitted value is estimated
  - affect how at least one coefficient is estimated
- To identify if an observation is an influential observation, we use three different measures of influence
  - each is a delete-one measure – works by fitting new models after deleting a single observation



Line	# of points	Intercept	Slope	SSE
Not selected	19	16.5157	-0.6027	113.6105
All data	20	17.739	-0.7739	190.5115

INFLUENTIAL ON ALL FITTED VALUES

- To identify if an observation influences the estimation of all fitted values, we must quantify its influence.
- We use a measure called Cook's Distance
- To measure influence of one observation, we fit a model using all n observations
  - then refit the model using $n-1$ observations
  - the difference in the estimated trend between these two models tells us influence of deleted observation
- We can compare either the vector of estimated coefficients ($\widehat{\boldsymbol{\beta}}$) or the fitted values ($\widehat{\boldsymbol{Y}}$)

$$D_i = \frac{(\widehat{\boldsymbol{Y}}_{(i)} - \widehat{\boldsymbol{Y}})^T(\widehat{\boldsymbol{Y}}_{(i)} - \widehat{\boldsymbol{Y}})}{(p+1)s^2} = \frac{(\widehat{\boldsymbol{\beta}}_{(i)} - \widehat{\boldsymbol{\beta}})^T\boldsymbol{X}^T\boldsymbol{X}(\widehat{\boldsymbol{\beta}}_{(i)} - \widehat{\boldsymbol{\beta}})}{(p+1)s^2}$$

- (i) subscript represents the values from the model with observation i deleted (i.e., on $n-1$ observations)
- Instead of fitting n different delete-one models, use

$$D_i = \frac{r_i^2}{(p+1)}\frac{h_{ii}}{(1-h_{ii})}$$

- Incorporates effect due potentially to
  - being distant in the X-space (i.e., leverage)
  - being far from estimated trend (i.e., outlying-ness)

INFLUENTIAL ON OWN FITTED VALUE

- To quantify the influence of a single observation on its own fitted value, we use a different measure.
  - called the DFFITS ("difference in fitted values")
- Less conservative than Cook's distance
- Again, we compare two models to measure effect of single observation on $\hat{y}_i$:
  - fit model using all n observations
  - delete observation i and refit same model using $n-1$ observations
- Look at the change in $\hat{y}_i$ values, accounting for variation expected

$$DFFITS_i = \frac{\hat{y}_i - \hat{y}_{i(i)}}{\sqrt{s_{(i)}^2 h_{ii}}}$$

- $\hat{y}_{i(i)}$ is fitted value for observation i using model without i
  - similarly, $s_{(i)}^2$ is estimated error variance from model omitting i
- Rather than fitting n different models, use

$$DFFITS_i = \left(\frac{h_{ii}}{1-h_{ii}}\right)^{0.5} \frac{\hat{e}_i}{s_{(i)}\sqrt{1-h_{ii}}}$$

  - Combines outlying-ness and leverage

INFLUENTIAL ON AT LEAST ONE ESTIMATED COEFFICIENT

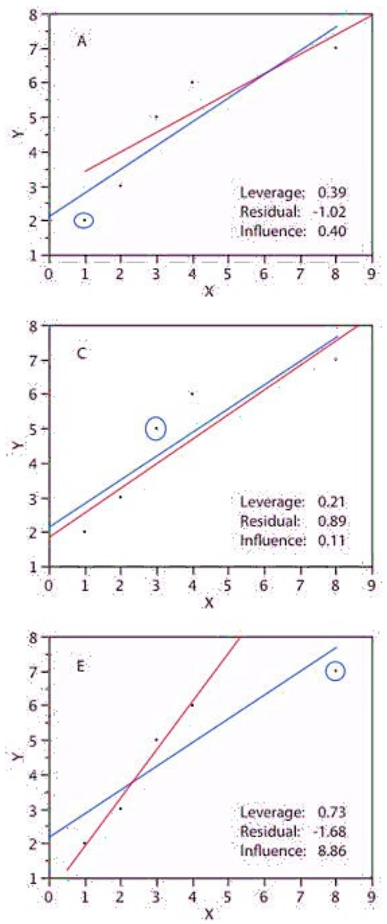
- Last way an observation can be influential is on the estimated value of at least one coefficient.
- The measure we use to quantify the amount of influence is called DFBETAS ("difference in betas")
- It is again a delete-one measure, so we fit two models: one with n observations, one with only $n-1$
- We look at how each individual coefficient changes with and without each observation.
  - means we delete each of the n observations individually
  - compare each of the $p+1$ coefficients individually

$$DFBETAS_{j(i)} = \frac{\hat{\beta}_j - \hat{\beta}_{j(i)}}{\sqrt{s_{(i)}^2 (\boldsymbol{X}^T\boldsymbol{X})^{-1}_{j+1,j+1}}}$$

- $\hat{\beta}_{j(i)}$ is coefficient j from model without point i
  - need to use the diagonal value corresponding to coefficient j
- We are comparing the change in estimated value to the expected variation in values due to sampling distribution
- Unfortunately, no simpler formula to use

MODULE 8 OUTLINE

1. Leverage Observations
2. Outlying Points in Regression
3. Influential Observations
4. Assessing and Addressing Problematic Observations



https://onlinestatbook.com/2/regression/influential.html

PROBLEMATIC OBSERVATIONS

- We've seen many different types of problematic points:
  - those with the potential to influence the estimated trend (leverage points and outliers)
  - those that do have an influence on some aspect of the estimated trend (influential points)
- Should always check for leverage, outliers, and all 3 kinds of influence
  - need to understand why estimated regression surface is estimated the way it is
  - should check every observation in your model
  - one type of problematic point may not necessarily be another type of problematic point.

IDENTIFYING PROBLEMATIC OBSERVATIONS

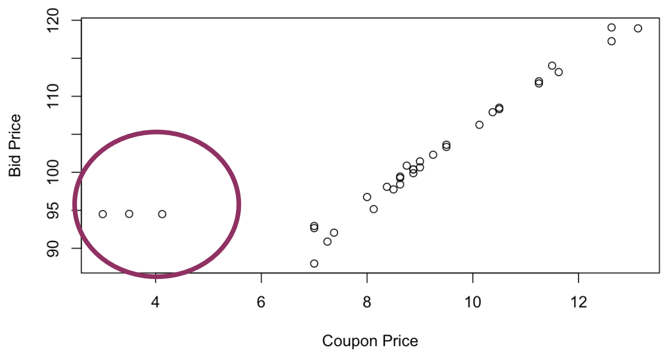
- Each measure quantifies the extent of each potential issue, so we define cutoff values to know when the amount of leverage, outlying-ness and influence is substantial.
- The cutoffs change in value depending on size of dataset (n) and number of predictors (p)

Type of Point	Measure	Cutoff
Leverage	h_{ii}	$h_{ii} > 2\left({}^{(p+1)}/_{n}\right)$
Outlier	r_i	$r_i \notin [-2,2]$ if dataset "small" (e.g. $n < 50$) $r_i \notin [-4,4]$ if dataset "large" (e.g. $n \geq 50$)
Influence	D_i	$D_i >$ median of $F(p+1, n-p-1)$
	$DFFITS_i$	$\lvert DFFITS_i \rvert > 2\sqrt{{}^{(p+1)}/_{n}}$
	$DFBETAS_{j(i)}$	$\lvert DFBETAS_{j(i)} \rvert > {}^{2}/_{\sqrt{n}}$

EXAMPLE BY HAND

Relationship between Coupon Rate (X) and Bid Price (Y) of 35 bonds. Plot of data below, and values from various models in table. Also, $s_x^2 = 5.45$, and $\bar{x} = 8.92$.

Model	Coefficients	Response	Fitted	s	$(\boldsymbol{X}^T\boldsymbol{X})^{-1}_{j+1,j+1}$
With all i	$\hat{\beta}_0 = 74.79, \hat{\beta}_1 = 3.07$	94.50	83.98	4.175	0.46, 0.01
Without i	$\hat{\beta}_0 = 70.57, \hat{\beta}_1 = 3.50$	94.50	81.06	3.683	0.58, 0.01



Investigate the observation with Coupon Rate of 3.

$$h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{j=1}^n (x_j - \bar{x})^2} = \frac{1}{35} + \frac{(3 - 8.92)^2}{34(5.45)} = 0.218$$

cutoff: $2\left(\frac{p+1}{n}\right) = 2\frac{2}{35} = 0.11 < 0.218$ leverage point

$$r_i = \frac{\hat{e}_i}{s\sqrt{1 - h_{ii}}} = \frac{\boxed{94.5 - 83.98}}{4.175\sqrt{1 - 0.218}} = 2.85$$

cutoff: 2.85 is outside $[-2,2]$ outlier point

influential on all fitted values

$$D_i = \frac{r_i^2}{(p+1)}\frac{h_{ii}}{(1 - h_{ii})} = \frac{2.85^2}{2} \times \frac{0.218}{1 - 0.218} = 1.13$$

cutoff: $1.13 >$

```
> qf(0.5, 2, 33)
[1] 0.7079124
```

influential on own fitted value

$$DFFITS_i = \left(\frac{h_{ii}}{1 - h_{ii}}\right)^{0.5} \frac{\hat{e}_i}{\boxed{s_{(i)}}\sqrt{1 - h_{ii}}} = \left(\frac{0.218}{1 - 0.218}\right)^{0.5} \frac{94.5 - 83.98}{3.683\sqrt{1 - 0.218}} = 1.33$$

$>$ cutoff: $2\sqrt{(p+1)/n} = 0.48$

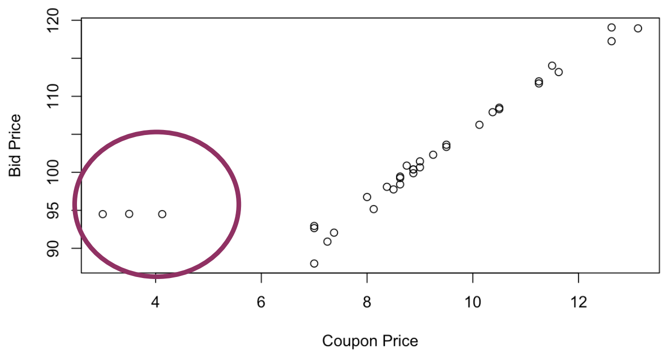
influential on estimated slope

$$DFBETAS_{1(i)} = \frac{\boxed{\hat{\beta}_1 - \hat{\beta}_{1(i)}}}{\sqrt{\boxed{s_{(i)}^2}\boxed{(\boldsymbol{X}^T\boldsymbol{X})^{-1}_{1+1,1+1}}}} = \frac{3.07 - 3.50}{3.683\sqrt{0.01}} = -1.17$$

cutoff: $|-1.17| > 2/\sqrt{n} = 0.34$

EXAMPLE USING R

Relationship between Coupon Rate (X) and Bid Price (Y) of 35 bonds. Plot of data below, and values from various models in table. Also, $s_x^2 = 5.45$, and $\bar{x} = 8.92$.



Investigate the observation with Coupon Rate of 3.

```
> model <- lm(BidPrice ~ CouponRate, data=data)
> summary(model)

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

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
```

influential on coefficients:

```
> dfbetas <- dfbetas(model)
> dim(dfbetas)
[1] 35  2
> cutoff_dfbetas <- 2/sqrt(35)
```

```
> which(abs(dfbetas[,1])>cutoff_dfbetas)
 4 13 34 35
 4 13 34 35
```

```
> which(abs(dfbetas[,2])>cutoff_dfbetas)
 4 13 19 35
 4 13 19 35
```

leverage points:

```
> hii <- hatvalues(model)
> cutoff_hii <- 2*2/35
> which(hii > cutoff_hii)
 4  5 13 35
 4  5 13 35
```

outlier points:

```
> ri <- rstandard(model)
> which(ri > 2 | ri < -2)
13 34 35
13 34 35
```

influential on all fitted values:

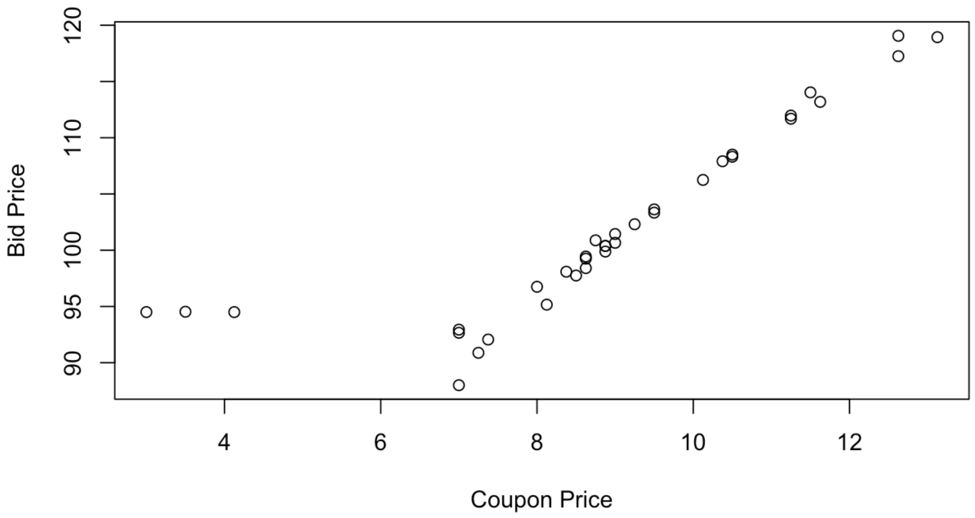
```
> di <- cooks.distance(model)
> cutoff_di <- qf(0.5, 2, 35-2)
> which(di > cutoff_di)
13
13
```

influential on own fitted values:

```
> dffits <- dffits(model)
> cutoff_dffits <- 2*sqrt(2/35)
> which(abs(dffits) > cutoff_dffits)
 4 13 19 34 35
 4 13 19 34 35
```

ADDRESSING PROBLEMATIC OBSERVATIONS

- Identification is key as you need to understand if individual observations affect your estimated relationship.
- Unless there is a contextual reason, do not remove problematic observations from data
  - e.g. in the Bonds example, these 3 datapoints are a special kind of bond so it would be justifiable to remove them
    - this changes generalizability of model (only applies to other bonds)
- Removal simply to improve model is akin to p-hacking (changing data, hypotheses, model just to make it look better)
- Note their presence and impact as limitation of model



MODULE TAKE-AWAYS

1. What is the difference between each type of problematic observations?
2. How do we quantify the various problematic observations?
3. How do we then identify the presence of each type of problematic observation?
4. What do we do if problematic observations are present?