36-617: Applied Linear Models

Multiple Regression

Brian Junker
132E Baker Hall
brian@stat.cmu.edu

Announcements

- HW01 is (finally!) available on Canvas
  - Due next Wed Sept 7, 11:59pm
    - Submit a single pdf to Gradescope within Canvas
- Reading
  - For this week: Sheather Ch 5
    (supplemental: ISLR 3.1, 3.2; G&H Ch 3)
  - For next week: ISLR 3.1, 3.2, 3.3.1, 3.3.2
    (supplemental: G&H Ch 3)
- "Monday Quiz" will be on Weds next week.
  - Available after class on Canvas
  - You have until 5pm Thu to do the quiz
  - Covers only Sheather Ch 5

Outline

- Matrix Form of Multiple Regression Model
- Regression - ML/LS Estimates
- Distributional Properties
- Standard Error vs Standard Deviation
- R's Casewise Diagnostic Plots
- Confidence Intervals and Prediction Intervals

Matrix Form of Multiple Regression Model

$$Y = X\beta + \epsilon$$

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} x_{10} & x_{11} & \cdots & x_{1p} \\ x_{20} & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n0} & x_{n1} & \cdots & x_{np} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

Usually $x_{i0} \equiv 1$, so we get

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

Regression – ML/LS Estimates

$$y = X\beta + \epsilon, \;\; \epsilon \sim N(0, \sigma^2 I)$$

- $\hat{\beta} = (X^T X)^{-1} X^T y$
- $\hat{y} = X\hat{\beta} = X(X^T X)^{-1} X^T y = Hy$
- $\hat{e} = y - \hat{y} = (I - H)y$

"Hat matrix"

- The "residual SD" is the square root of

$$s^2 = \frac{1}{n-k} \sum_{i=1}^{n} (y_i - X_i\hat{\beta})^2 = \frac{1}{n-k} (y - X\hat{\beta})^T (y - X\hat{\beta})$$

- With a little more matrix algebra,

Recall that k=p+1

$$\begin{aligned} \mathrm{Var}(\hat{\beta}) &= (X^T X)^{-1}\sigma^2 \\ \mathrm{Var}(\hat{y}) &= X(X^T X)^{-1} X^T \sigma^2 = H\sigma^2 \\ \mathrm{Var}(\hat{e}) &= (I - H)\sigma^2 \end{aligned}$$

Distributional Properties: $\hat{\beta}$

Fact: $Y \sim N(\mu, \Sigma) \Rightarrow AY \sim N(A\mu, A\Sigma A^T)$

$$
\begin{aligned}
y &\sim N(X\beta, \sigma^2 I) \\
\hat{\beta} &= (X^T X)^{-1} X^T y \\
E[\hat{\beta}] &= E[(X^T X)^{-1} X^T y] = (X^T X)^{-1} X^T E[y] \\
&= (X^T X)^{-1} X^T X \beta = \beta \\
\mathsf{Var}(\hat{\beta}) &= \mathsf{Var}((X^T X)^{-1} X^T y) \\
&= (X^T X)^{-1} X^T \mathsf{Var}(y) X (X^T X)^{-1} \\
&= (X^T X)^{-1} X^T X (X^T X)^{-1} \sigma^2 = (X^T X)^{-1} \sigma^2 \\
\Rightarrow \hat{\beta} &\sim N(\beta, (X^T X)^{-1} \sigma^2)
\end{aligned}
$$

Standard Error vs. Standard Deviation

- Because $\hat{\beta} = \left(X^T X\right)^{-1} X^T y$ is a function of the (random) data, $\hat{\beta}$ has a distribution:

$$\hat{\beta} \sim N(\beta, \left(X^T X\right)^{-1} \sigma^2)$$

- For each j, we know

$$Var(\hat{\beta}_j) = \text{diag}\left(\left(X^T X\right)^{-1} \sigma^2\right)_j \approx \text{diag}\left(\left(X^T X\right)^{-1} s^2\right)_j$$

- An estimate of the standard deviation of $\hat{\beta}_j$ is

$$\sqrt{\widehat{Var}(\hat{\beta}_j)} = \sqrt{\text{diag}\left((X^T X)^{-1} s^2\right)_j}$$

- Because $\hat{\beta}_j$ is an estimator, we call this the ***standard error***, $SE(\hat{\beta}_j)$, instead of standard deviation.

Regression – Example

- Demographic factors and income…

```
> library("foreign") # use read.dta for STATA files...
> heights <- read.dta("heights.dta")
> str(heights)
'data.frame':   2029 obs. of  9 variables:
 $ earn   : num  NA NA 50000 60000 30000 NA 50000 NA 51000 ...
 $ height1: int  5 5 6 5 5 5 5 5 5 5 ...
 $ height2: num  6 4 2 6 4 5 3 8 3 4 ...
 $ sex    : int  2 1 1 2 2 2 2 2 2 2 ...
 $ race   : int  1 2 1 1 1 1 3 2 1 1 ...
 $ hisp   : int  2 2 2 2 2 2 2 2 2 2 ...
 $ ed     : num  12 12 16 16 16 17 16 18 17 15 ...
 $ yearbn : num  53 50 45 32 61 33 99 36 51 64 ...
 $ height : num  66 64 74 66 64 65 63 68 63 64 ...
```

Regression – Example

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

$$SE(\hat{\beta}) = \sqrt{\text{diag}(\widehat{\text{Var}}(\hat{\beta}))} = \sqrt{\text{diag}((X^T X)^{-1} s^2)}$$

```
> summary(lm(earn ~ height + ed, data=heights))

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -106263.5     8564.7  -12.41   <2e-16 ***
height         1376.7      126.7   10.87   <2e-16 ***
ed             2590.8      197.7   13.11   <2e-16 ***

Residual standard error: 17780 on 1376 degrees of freedom
  (650 observations deleted due to missingness)
Multiple R-squared:  0.1915,    Adjusted R-squared:  0.1904
F-statistic:   163 on 2 and 1376 DF,  p-value: < 2.2e-16
```

$$t_j = \hat{\beta}_j / SE(\hat{\beta}_j)$$

$$\begin{aligned} \text{earn} &\approx \hat{\beta}_0 + \hat{\beta}_1(\text{height}) + \hat{\beta}_2(\text{ed}) \\ &= -106263.5 + 1376.7(\text{height}) + 2590.8(\text{ed}) \end{aligned}$$

$$\frac{1}{n-k} RSS = s^2 = (17780)^2 = 316128400$$

$$R^2 = \frac{\widehat{\text{Var}}(\hat{y})}{\widehat{\text{Var}}(y)} = \underbrace{\frac{SS_{reg}}{SSY} = 1 - \frac{RSS}{SSY}}_{\text{We'll talk about this decomposition next week}} = 0.1915$$

$$k = p + 1 = 3\,; \quad n - k = 1376$$

$$n = 1379$$

Hat Matrix H, & Distribution of $\hat{y}$ and $\hat{e}$

$$\hat{y} = X\hat{\beta} = X\left[(X^TX)^{-1}X^Ty\right] = \underbrace{\left[X(X^TX)^{-1}X^T\right]}_{\text{Hat matrix, H}} y = Hy$$

$$HX = X(X^TX)^{-1}X^TX = X$$

$$\Rightarrow \forall \beta^*, HX\beta^* = X\beta^*$$

$$H^T = H \text{ and } (I-H)^T = (I-H) \quad \text{(symmetric)}$$

$$HH = H \text{ and } (I-H)(I-H) = (I-H) \quad \text{(idempotent)}$$

$$E[\hat{y}] = E[Hy] = HE[y] = HX\beta = X\beta$$

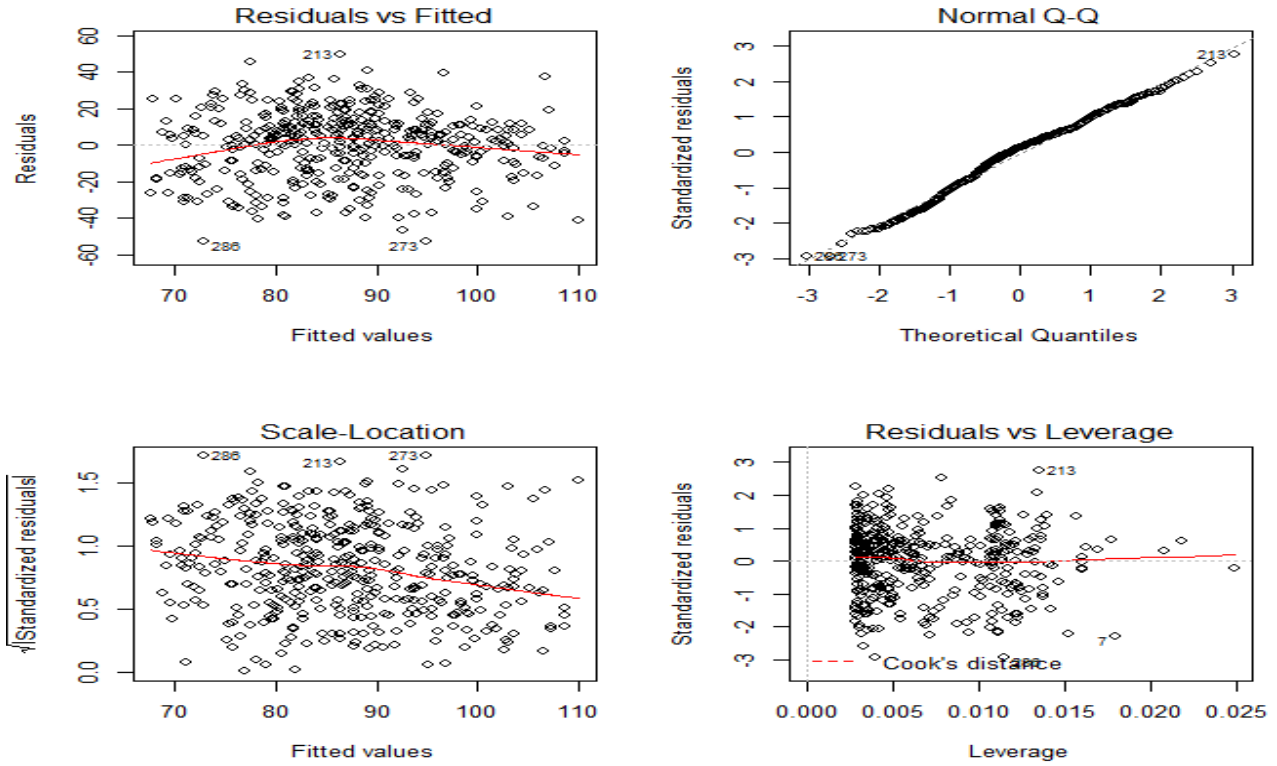
$$\text{Var}(\hat{y}) = \text{Var}(Hy) = H\text{Var}(y)H^T = HH\sigma^2 = H\sigma^2$$

$$\Rightarrow \hat{y} \equiv Hy \sim N(X\beta, H\sigma^2)$$

Similarly, $\hat{e} = y - \hat{y} \equiv (I-H)y \sim N(0, (I-H)\sigma^2)$

Casewise diagnostic plots

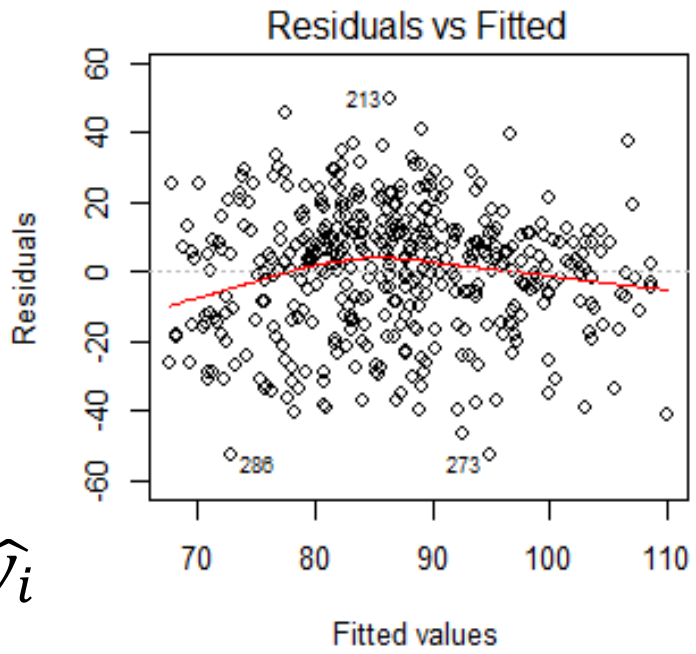
```
> kidiq <- read.csv("kidiq.csv",header=T)
> lm.3 <- lm(kid.score ~ mom.iq + mom.hs, data=kidiq)
> par(mfrow=c(2,2)); plot(lm.3)
```



Raw residual plot: $\hat{e}$ vs $\hat{y}$

- If y = Xβ + ε holds, then

$$\begin{aligned} \hat{e} &= (I - H)y \\ \hat{e} &\sim N(0, (I - H)\sigma^2) \\ \mathrm{Cov}(\hat{e}, \hat{y}) &= 0 \end{aligned}$$



- So expect to see
  - Mean 0
  - No functional dependence on $\hat{y}_i$
  - No outliers
  - Modest correlation between residuals (usually not a problem for interpreting plot)
- Violations suggest nonlinearity / non-normality…

Leverage h_{ii} ...

- Write $H = [h_{ij}]_{i,j=1}^{n}$; then h_{ii} is the *leverage* of the i^{th} data point.
- We just saw that $HX = X$. If X has a column of 1's (intercept!), then for every i, $\sum_j h_{ij} = 1$.
- $\sum_i h_{ii} = tr(H) = tr(X(X^TX)^{-1}X^T) = tr(X^TX(X^TX)^{-1})$ *
  $= tr(I_{(p+1)\times(p+1)}) = p + 1.$
- $Var(\hat{y}) = H\sigma^2 \Rightarrow Var(\hat{y}_i) = h_{ii}\sigma^2 \Rightarrow h_{ii} \geq 0.$
- $Var(\hat{e}) = (I - H)\sigma^2 \Rightarrow Var(\hat{e}_i) = 1 - h_{ii}\sigma^2 \Rightarrow h_{ii} \leq 1.$
- $h_{ii} = M_i/(n-1) + 1/n \geq 1/n.$ **
  - M_i is the Mahalanobis distance from row X_i to the mean vector of the rows of X.
  - Since $\hat{y} = Hy$, then $\hat{y}_i = \sum_j h_{ij}y_j = h_{ii}y_i + \sum_{j\neq i} h_{ij}y_j$.

* Using the fact that $tr(AB) = tr(BA)$ when the multiplications both work

** See https://en.wikipedia.org/wiki/Leverage_(statistics)

Normal Q-Q Plot of Stdized Resids r_i

Since $\hat{e}_i \sim N(0, (1-h_{ii})\sigma^2)$, the standardized residuals

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

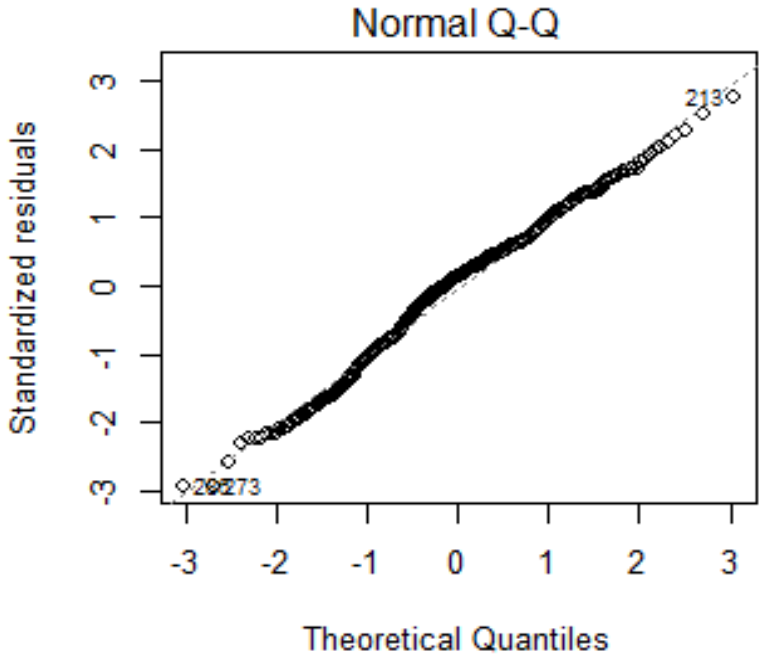
should be approx $N(0,1)$ under $y = X\beta + \epsilon$, where again

$$S^2 = \frac{1}{n-p-1}RSS$$

Note that

$$\hat{e}_i = (I-H)_i y = (1-h_{ii})y_i - \sum_{j \neq i} h_{ij} y_j$$

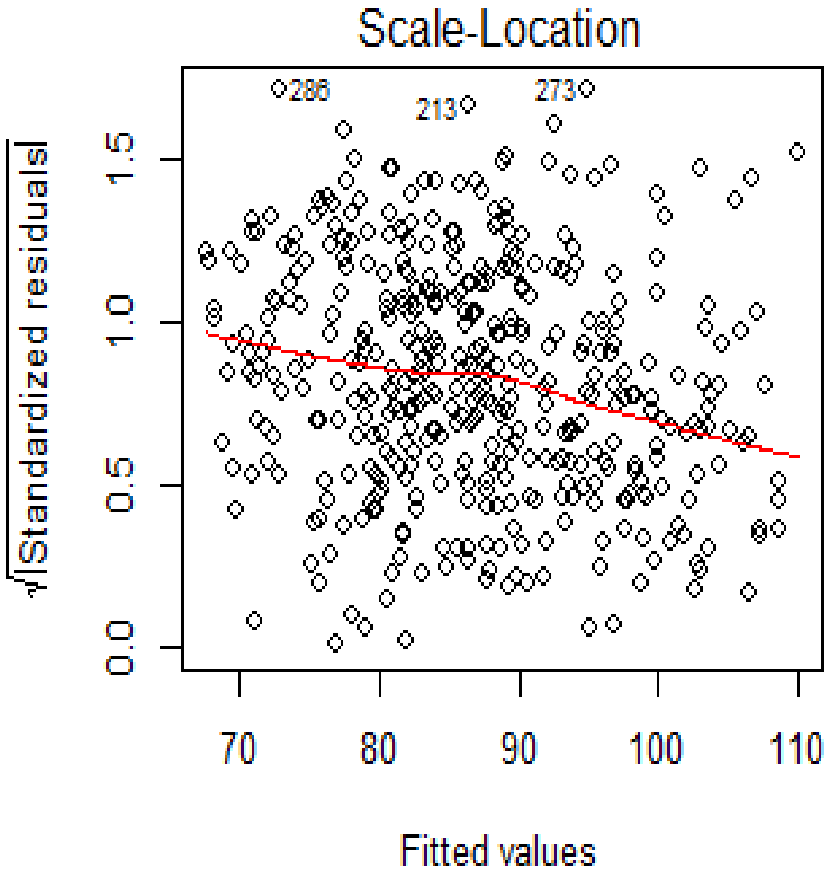
so $\hat{e}_i$ can look normal because y is, **or** because the sum obeys the CLT.



- Look for
  - Normal?
  - Outliers?

Scale-Location Plot: $\sqrt{r_i}$ vs $\hat{y}_i$

- If $\text{Var}(\epsilon_i) \equiv \sigma^2$ then r_i should have constant variance ≈ 1
  - Dependence on $\hat{y}_i$?
  - Loess line helps eye
  - Careful of edge effects
- Designed to catch patterns that depend on x_i (or y_i?)
- Patterns can be caused by
  - Nonconstant variance in ϵ_i
  - Nonlinear relationship between x_i and y_i



Leverage h_{ii} & Cooks' Distance D_i

- Leverage $h_{ii} = M_i/(n-1) + 1/n$ measures how far x_i is from $\bar{x}$.
- To have an effect, $\hat{e}_i = y_i - \hat{y}_i$ must be large also.
- How much effect can be measured by Cook's D_i:

$$D_i = \frac{\sum_{j=1}^{n}(\hat{y}_{j(i)} - \hat{y}_j)^2}{2S^2}$$

$$= \frac{r_i^2}{p+1} \cdot \frac{h_{ii}}{1-h_{ii}} \quad \text{(not obvious)!}$$

where $\hat{y}_{j(i)}$ is the fitted value for y_j, omitting the pair (x_i, y_i) from the data set.

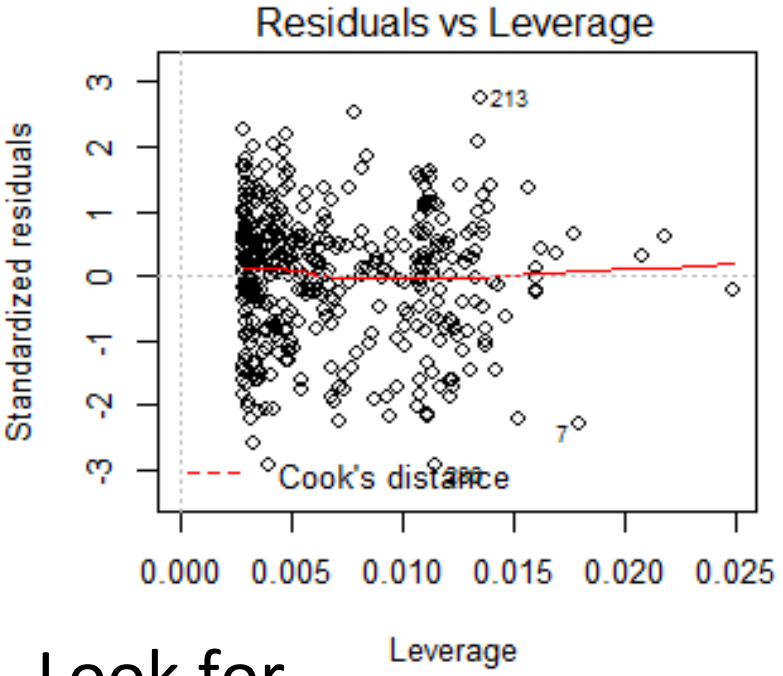
Leverage plot: h_{ii} vs D_i

$$\hat{y} \sim N(X\beta, H\sigma^2)$$
$$\hat{e} \sim N(0, (I-H)\sigma^2)$$

$$\begin{aligned}
\text{tr}\,(H_{n\times n}) &= \text{tr}\,(X(X^TX)^{-1}X^T) \\
&= \text{tr}\,(X^TX(X^TX)^{-1}) \\
&\quad (\text{since } \text{tr}\,(AB) = \text{tr}\,(BA)) \\
&= \text{tr}\,(I_{(p+1)\times(p+1)}) \\
&= p+1,
\end{aligned}$$

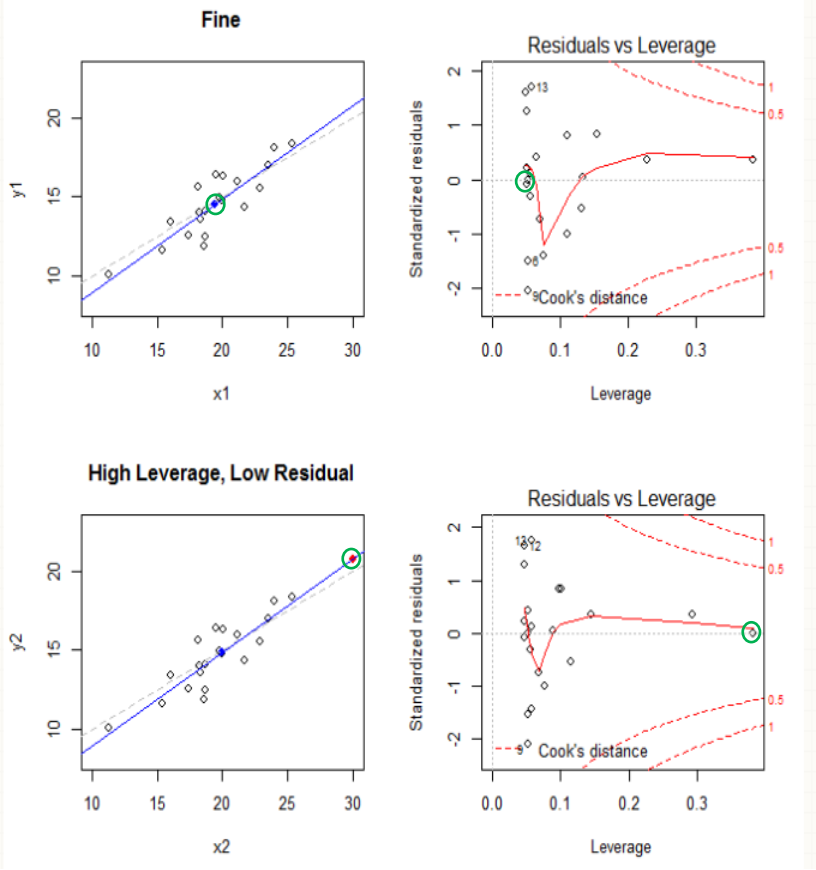
$$\text{so } 0 \le h_{ii} \le 1 \quad \& \quad \bar{h} = (p+1)/n$$

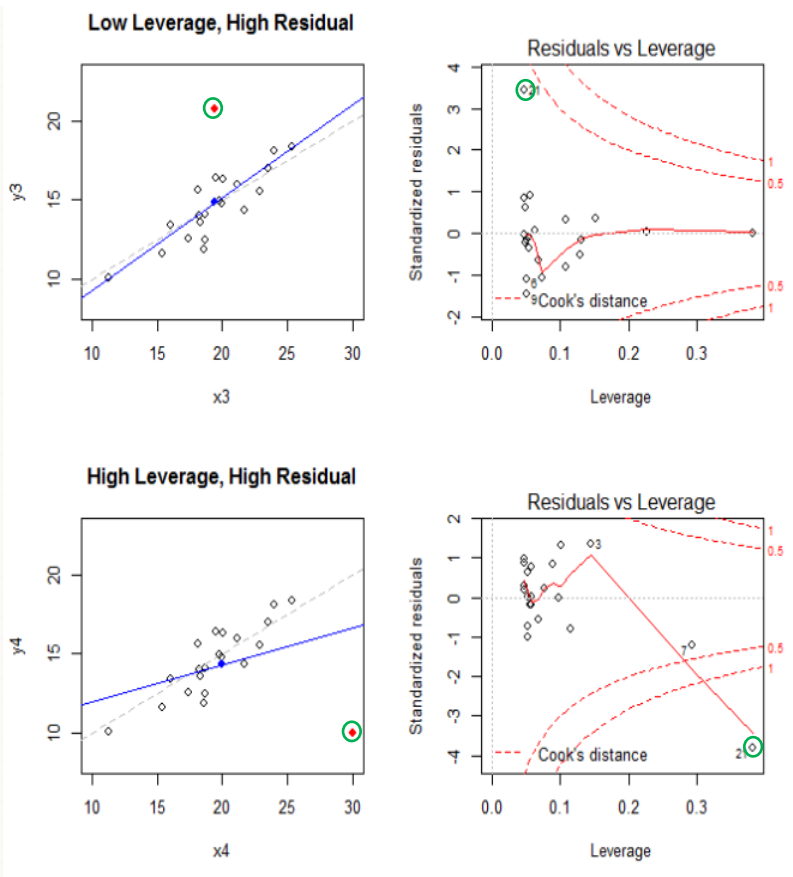
$$\text{Cook's } D_i \equiv \frac{r_i}{p+1} \cdot \frac{h_{ii}}{1-h_{ii}}$$



- Look for
  - NE and SE "corners":
  - $|r_i| > 2$ or so?
  - $h_{ii} > 2(p+1)/n$ or so?
  - $D_i > 0.5$ or so?

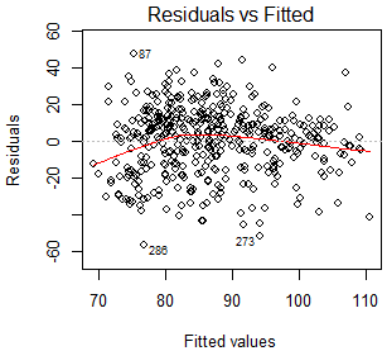
Some Leverage Examples



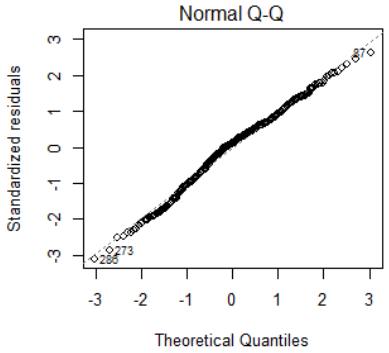


Source: https://stats.stackexchange.com/questions/58141/interpreting-plot-lm

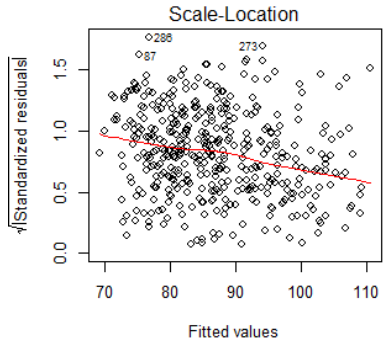
Casewise Diagnostics and Patterns



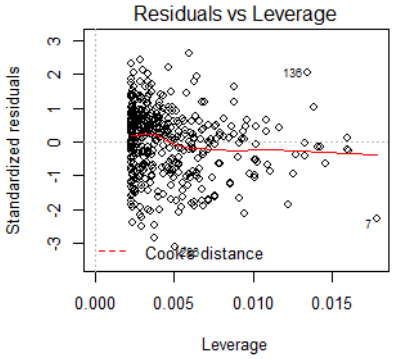
- Mean zero?
- Outliers?
- Functional dependence on $\hat{y}_i$?



- Normal?
- Outliers?



- Constant variance?
- Outliers?
- Functional dependence on $\hat{y}_i$?



- NE & SE corners:
  - High leverage h_{ii}
  - High std resid r_i
- D_i > 0.5 or so?

- Generally these are conversation points
  - Could reveal things investigator cares about!
  - Otherwise, look for data collection/recording errors
- Delete data only with a good justification!

Confidence Interval for a fitted value $\hat{y}^*$

- Let x* be a new data point (new row of the X matrix) and let y* be the new y value:

$$y^* = x^*\beta + \epsilon^*$$

- A CI for the point $E[y^*|x^*] = x^*\beta$ is

$$(\hat{y}^* - t \cdot SE(\hat{y}^*), \hat{y}^* + t \cdot SE(\hat{y}^*))$$

where

  - $\hat{y}^* = x^*\hat{\beta}$
  - $SE(\hat{y}^*) = \sqrt{\widehat{Var}(x^*\hat{\beta})} = \sqrt{x^*(X^TX)^{-1}x^{*T}s^2}$
  - t is an appropriate cutoff (around 2 for a 95% interval)

Prediction Interval for a new obs. y^*

- Again let x* be a new X row, and now consider predicting y* itself at that x*:

$$\hat{y}^*_{pred} = x^*\hat{\beta} + \epsilon^*_{pred}$$

- Using the same sorts of calculations as before,

$$\begin{aligned} E[\hat{y}^*_{pred}] &= E[\hat{y}^*] = x^*\beta \\ \text{Var}\,(\hat{y}^*_{pred}) &= (x^*(X^TX)^{-1}x^{*T} + 1)\sigma^2 \end{aligned}$$

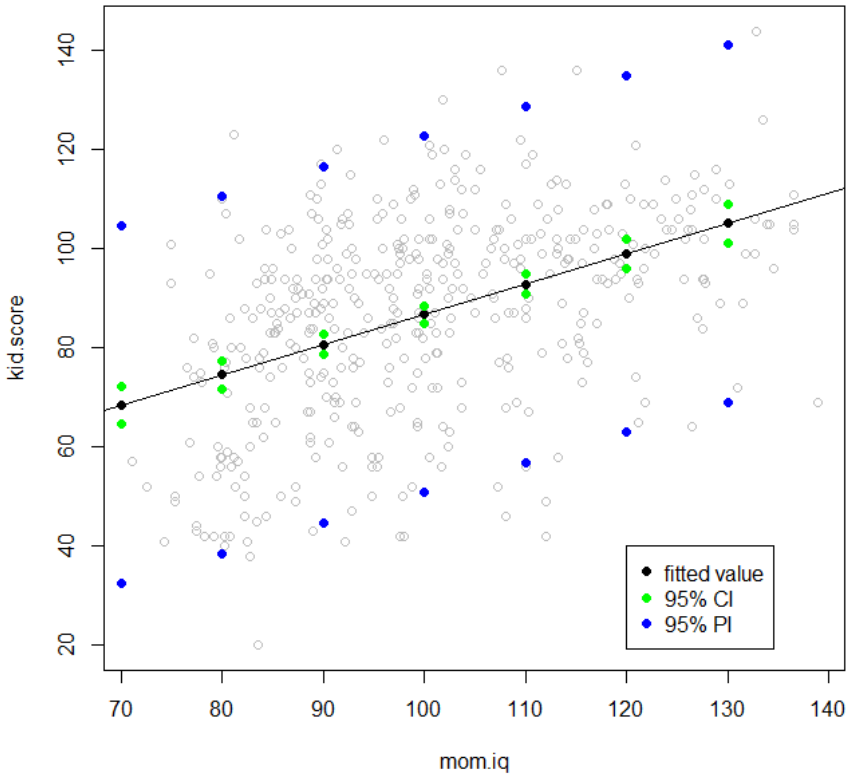
- So a prediction interval for y^*_{pred} would be

$$(\hat{y}^* - t \cdot SE(\hat{y}^*_{pred}), \hat{y}^* + t \cdot SE(\hat{y}^*_{pred})),$$

where $SE(\hat{y}^*_{pred}) = \sqrt{(x^*(X^TX)^{-1}x^{*T} + 1)s^2}$

Example…

```
kidiq <- read.csv("kidiq.csv",header=TRUE)
fit.lm.1 <- lm(kid.score ~ mom.iq, data=kidiq)

plot(kid.score ~ mom.iq, data=kidiq,col="grey")
abline(fit.lm.1)

new.Xs <- data.frame(kid.score=0, mom.iq=(7:13)*10)

new.Ys.CI <- predict(fit.lm.1,new.Xs,interval="confidence")
points((7:13)*10,new.Ys.CI[, "fit"],pch=19,col="black")
points((7:13)*10, new.Ys.CI["lwr"],pch=19,col="green")
points((7:13)*10, new.Ys.CI["upr"],pch=19,col="green")

new.Ys.PI <- predict(fit.lm.1,new.Xs,interval="prediction")
points((7:13)*10, new.Ys.PI["lwr"],pch=19,col="blue")
points((7:13)*10, new.Ys.PI["upr"],pch=19,col="blue")

legend(120,40,legend=c("fitted value", "95% CI", "95% PI"),
  pch=19,col=c("black", "green", "blue"))
```



Summary

- Matrix Form of Multiple Regression Model
- Regression - ML/LS Estimates
- Distributional Properties
- Standard Error vs Standard Deviation
- R's Casewise Diagnostic Plots
- Confidence Intervals and Prediction Intervals