36-617: Applied Linear Models

Multivariate Regression

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

Reading, HW

- Reading
  - This week: Sheather Ch 5 (Skip Ch 4)
  - For next week: Sheather Ch 6

- HW 04 due Mon 11:59pm
  - Start early! It may be somewhat long…

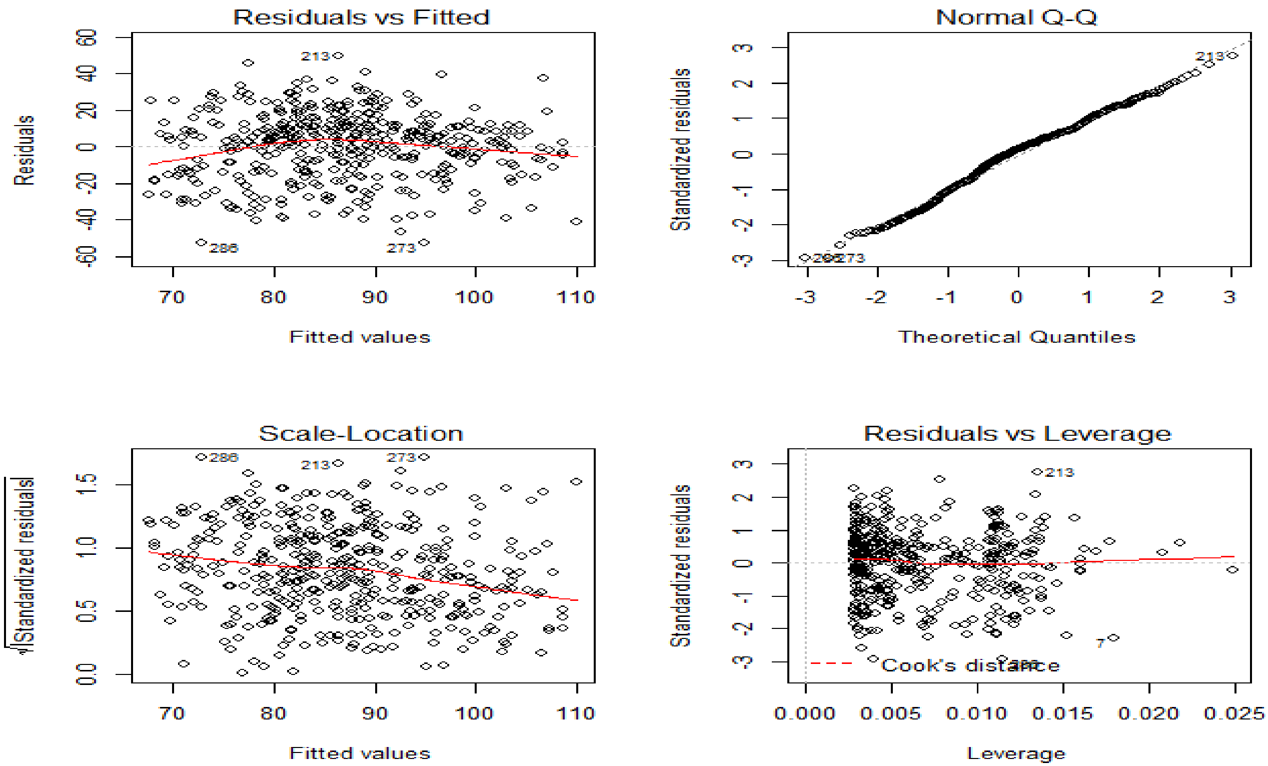
- First (midterm) project will be announced soon…

Outline

- Casewise diagnostic plots for multiple regression
- Added variable plots
- Aside: Analysis of Covariance (ANCOVA)
- Example(s)

Casewise diagnostic plots

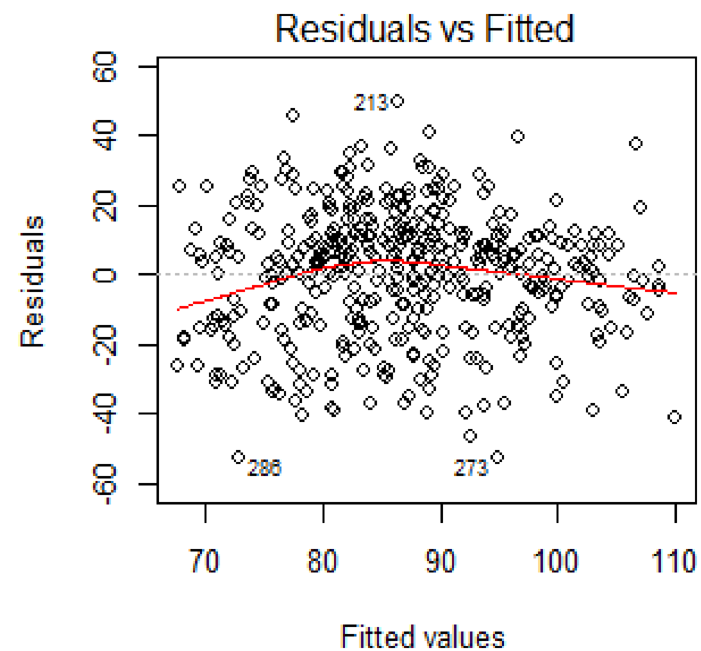
```
> kidiq <- read.csv("kidiq.csv",header=T)
> lm.3 <- lm(kid.score ~ mom.iq + mom.hs, data=kidiq)
> par(mfrow=c(2,2)); plot(lm.3)
```



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

- If y = Xβ + ε holds, then

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



- So expect to see
  - Mean 0
  - No vertical patterns
  - No outliers
  - Modest correlation between residuals (usually not a problem for interpreting plot)
- Violations suggest nonlinearity / non-normality…

Normal Q-Q Plot of Stdized Resids r_i

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

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

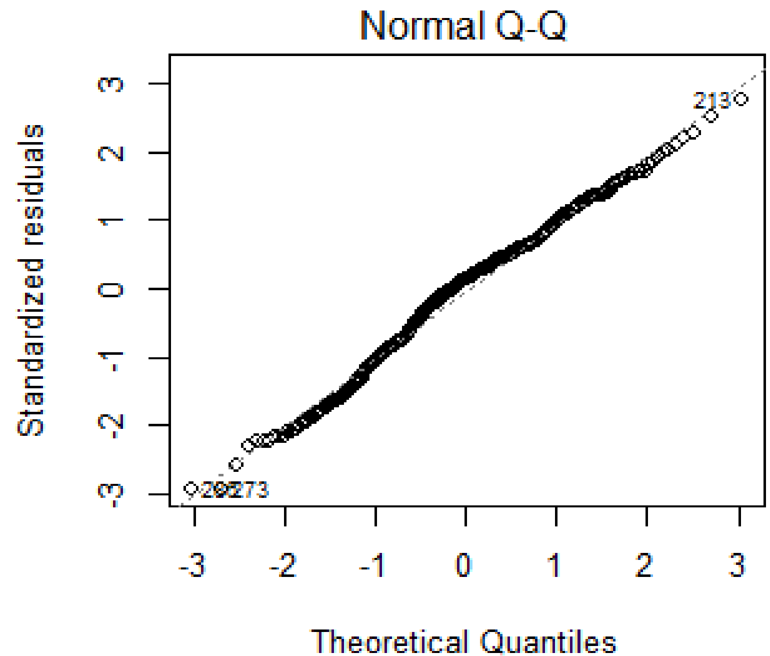
should be approx $N(0,1)$ under $y = X\beta + \epsilon$, where again

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

As in the simple regression case,

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

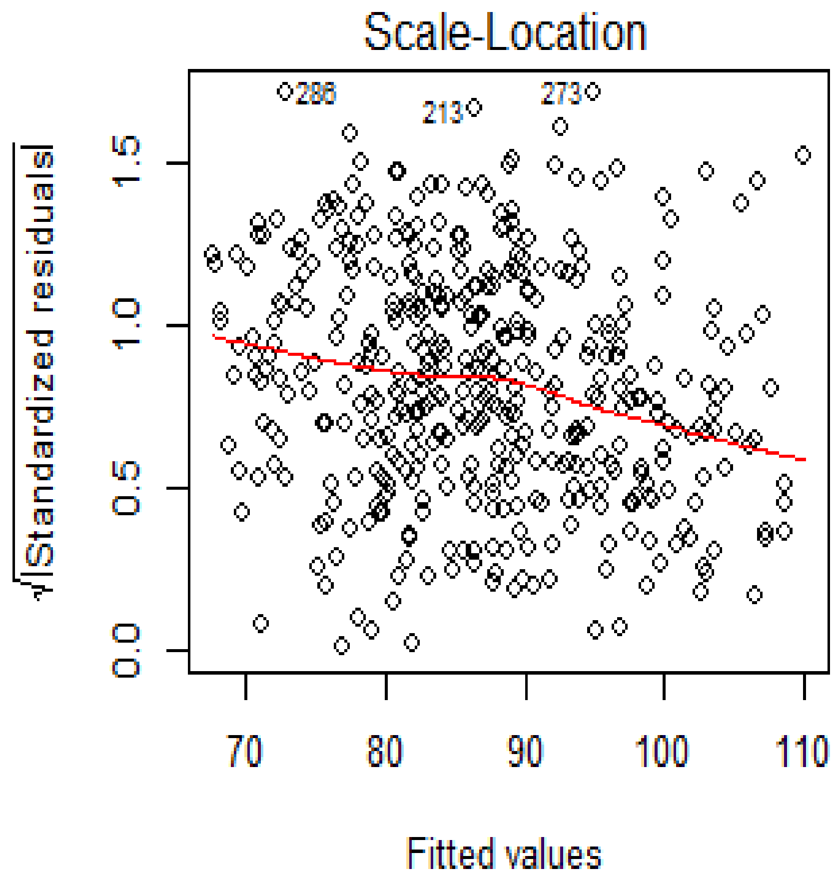
so for modest samples the residuals may look more normal than they "really are"...



- Look for
  - Normal?
  - Outliers?
  - Large enough sample?

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
  - should see no vertical patterns
  - 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



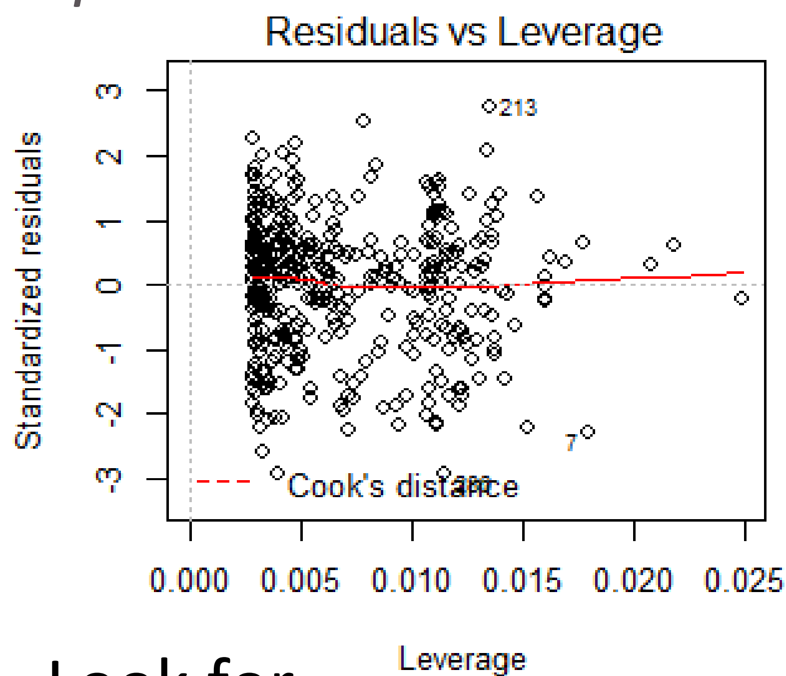
Leverage plot: h_{ii} vs D_i

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

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

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

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



- 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?*

Added-Variable Plots (add Z? or f(Z)?)

- Suppose the true model is

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

- Let us fit the models

$$\begin{aligned} Y &= X\beta + \epsilon^{(1)} \text{ with residuals } \hat{e}^{(1)} = (I - H_X)Y \\ Z &= X\beta + \epsilon^{(2)} \text{ with residuals } \hat{e}^{(2)} = (I - H_X)Z \end{aligned}$$

- If we multiply the true model by *(I-H$_X$)*, we get

$$\begin{aligned} (I - H_X)Y &= (I - H_X)X\beta + (I - H_X)Z\gamma + (I - H_X)\epsilon \\ \hat{e}^{(1)} &= 0 + \hat{e}^{(2)}\gamma + \epsilon^* \end{aligned}$$

so, plotting (or regressing) $\hat{e}^{(1)}$ on $\hat{e}^{(2)}$ will reveal γ !

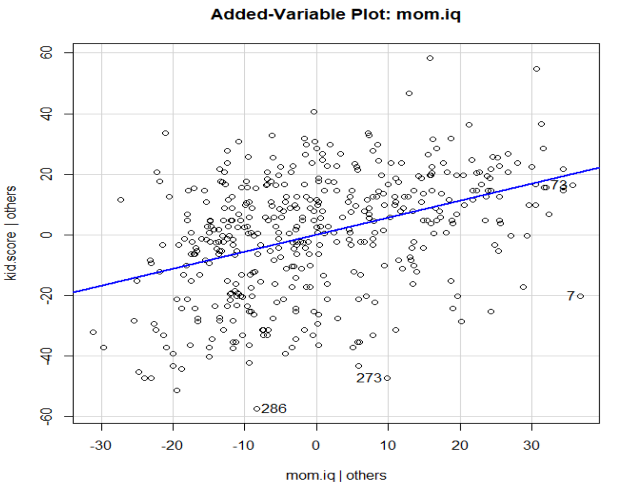
Added-Variable Plots – Example…

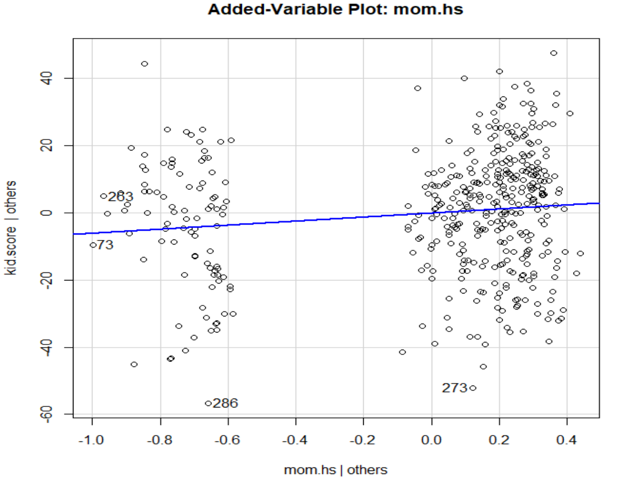
```
> library(car)
> lm.3

Call:
lm(formula = kid.score ~ mom.iq +
mom.hs, data = kidiq)

Coefficients:
(Intercept)    mom.iq    mom.hs
    25.7315    0.5639    5.9501

> avPlot(lm.3,"mom.iq")
> avPlot(lm.3,"mom.hs")
```





Added-Variable Plots – Interpretations

- Shows γ as the effect of *Z* after controlling for *X*, on *Y*, after controlling for *X*
- Allows you to visually assess the importance of γ
- Also allows you to check for nonlinearity in predicting *Y* from *Z*, after controlling for *X*
- Another plot that allows us to assess nonlinearity is the *marginal model plot* – next week!

Aside: Analysis of Covariance (ANCOVA)

- Suppose we want to predict Y from X_1 and X_2
  - X_1 is continuous (e.g. mom's iq)
  - X_2 is a 0/1 dummy variable (e.g. did mom finish HS?)
- Some possible models
  - $y_i = \beta_0 + \beta_1 x_{i1} + \epsilon_i$
  - $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i$
  - $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1} x_{i2} + \epsilon_i$
- What do they look like?
- How can we test what matters?

Summary

- Casewise diagnostic plots for multiple regression
- Added variable plots
- Aside: Analysis of Covariance (ANCOVA)
- Example(s)