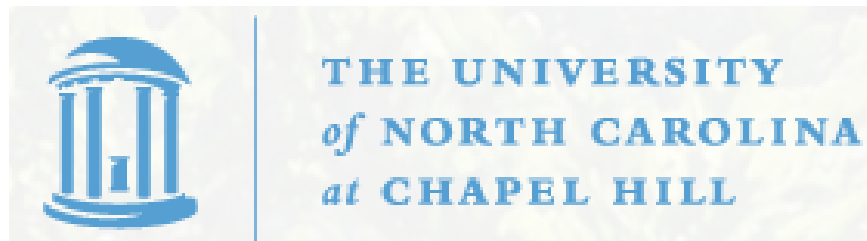


STOR 590:
ADVANCED LINEAR MODELS
Instructor: Richard L. Smith

Class Notes:
August 19, 2020



CLASS ANNOUNCEMENTS

- Homework 1 deadline (1pm Friday) — let me know if this is a problem
- Please submit on *both* the “assignments” and the “grade-scope” tabs of sakai if you can. However if for some reason gradescope doesn’t work, just use “assignments”.
- Submit as pdf if you can
- Any further disruptions caused by change in university schedule: please let me know
- Friday class will end at 2:00 pm
- Still working on assignments for office hours for next week

SAVING R MARKDOWN AS PDF

- R Markdown not required — another way to prepare your homework assignments is to create a word document and cut and paste tables and figures you want from R (however, if you do that, make sure you “save as” as pdf file)
- In R Markdown, select “Knit” from the top panel. There’s a pull-down menu, select “pdf”.
- To do this, you must have a latex package installed. If you don’t have one, please install Miktex (free package, google the name for details)
- You can also save as an html file and then convert that file to pdf. However, I think you need the Acrobat package (not just the reader) installed to do that.
- If you can’t figure out how to save your file as pdf, we will accept an html file for your homework

FINAL EXAM SCHEDULE

- Last day of classes: Tuesday November 17
- Exam period: November 18–24
- No reading day
- Our exam schedule: Monday Nov 23, 12:00–3:00 pm
- I am still planning a take-home exam with at least a 24 hour time limit
- 3 exams in 24 hours rule still applies
- Regular grades

INTRODUCING LEVERAGE, INFLUENCE AND COOK'S D STATISTIC

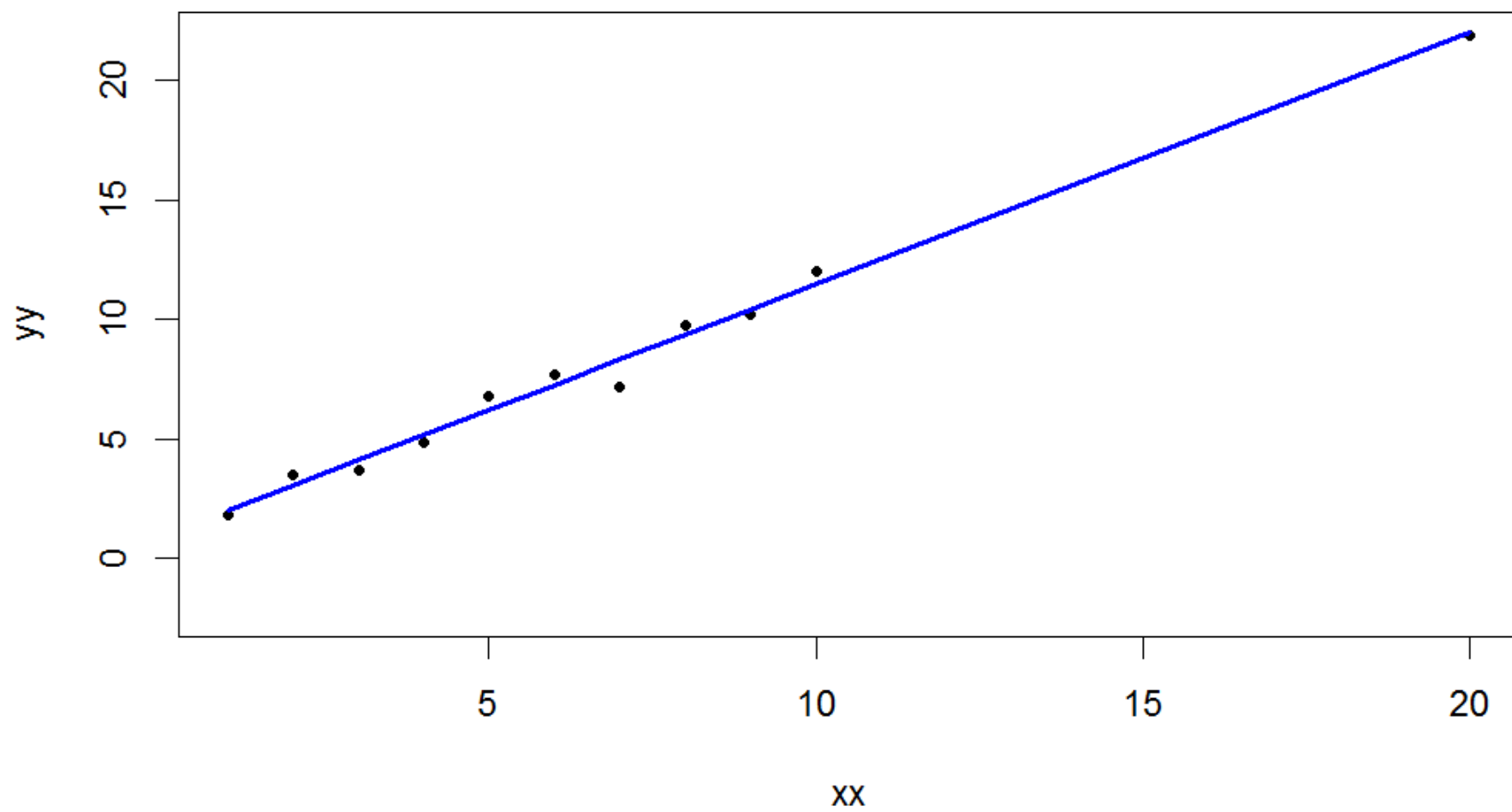
Confidence and Prediction Intervals

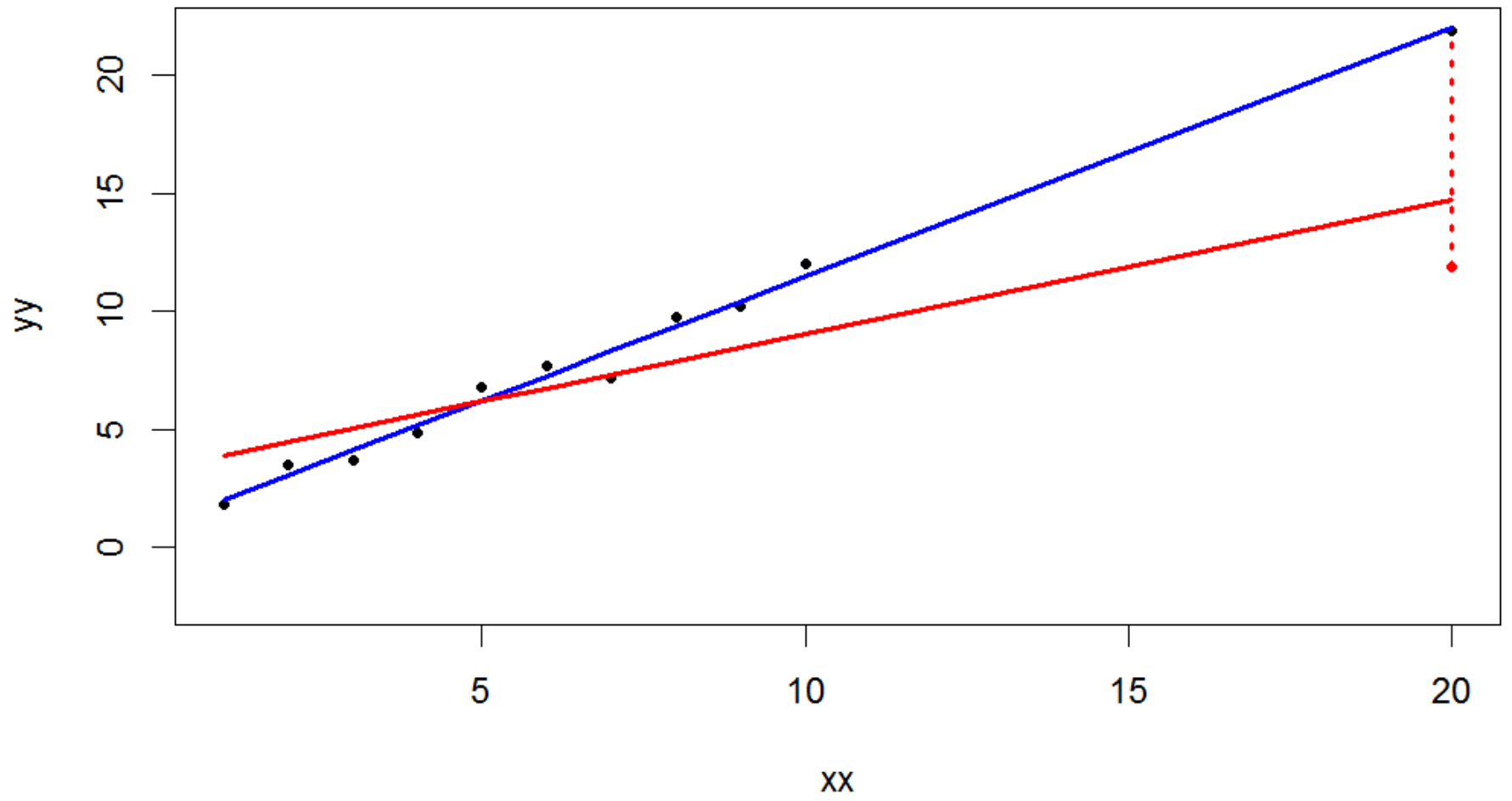
Consider a simple x-y plot with one “outlier” in the x direction.

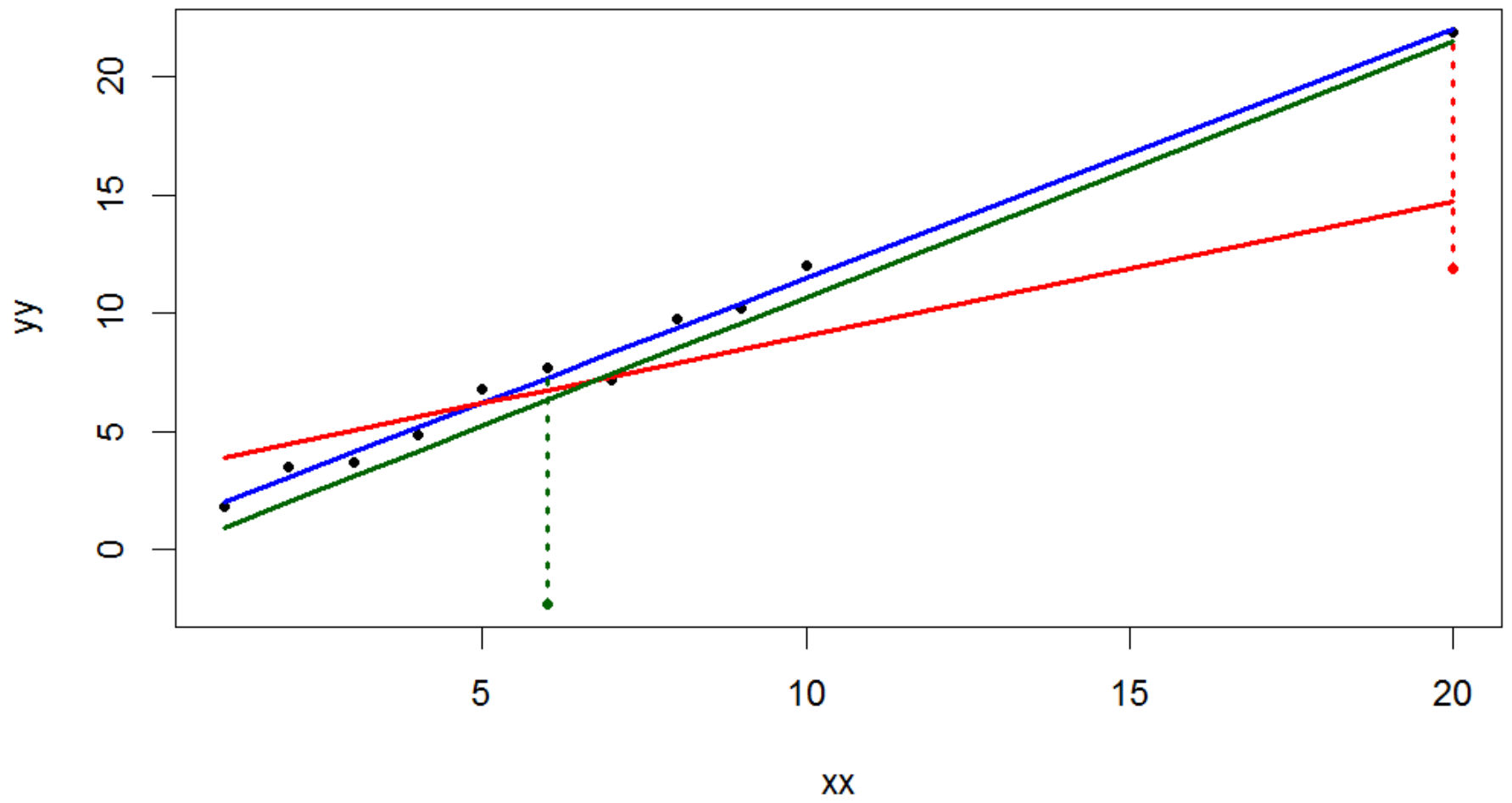
Consider the consequence of moving the corresponding y value up or down.

The effect is much greater than if we took some arbitrary x in the middle of the plot.

The difference is measured by *leverage*.







THEORY

$$\begin{aligned}\hat{y} &= X\hat{\beta} \\ &= X(X^T X)^{-1}X^T y \\ &= Hy\end{aligned}$$

where $H = X(X^T X)^{-1}X^T$ is known as the *hat matrix*.

H is an $n \times n$ matrix whose *trace* (sum of diagonal entries) is $p+1$, the number of unknown parameters (including the intercept). The diagonal entries h_i , $i = 1, \dots, n$ are called the *hatvalues*. “On average,” the leverages are about $\frac{p+1}{n}$. Any point substantially larger than that is called a *point of high leverage*.

If you have previously fit a linear model to create an object “lmod”, then `hatvalues(lmod)` will create the hat values.

MY EXAMPLE

$$X = \begin{pmatrix} 1 & 1 \\ 1 & 2 \\ \vdots & \vdots \\ 1 & 10 \\ 1 & 20 \end{pmatrix}, \quad X^T X = \begin{pmatrix} 11 & 75 \\ 75 & 785 \end{pmatrix},$$
$$(X^T X)^{-1} = \frac{1}{3010} \begin{pmatrix} 785 & -75 \\ -75 & 11 \end{pmatrix},$$
$$h_i = 0.21, 0.18, 0.14, 0.12, 0.10, 0.09, 0.09, 0.10, 0.11, 0.13, 0.73$$

Observation 11 has about eight times the leverage of observation 6.

R code for this example

```
X=matrix(c(rep(1,11),1:10,20),ncol=2)

# display  $X^T X$ 
t(X) %*% X

library('MASS')

# display inverse of  $X^T X$ 
ginv(t(X) %*% X)

# diagonal entries of the hat matrix, rounded to 2 decimal places
round(diag(X %*% (ginv(t(X) %*% X) %*% t(X))),2)
```

Confidence and Prediction Intervals, 1

Properties of H :

1. H is *symmetric*

$$\text{Proof: } H^T = \{X(X^T X)^{-1} X^T\}^T = (X^T)^T (X^T X)^{-1} X^T = H.$$

2. H is *idempotent* ($H^2 = H$)

$$\text{Proof: } H^2 = X(X^T X)^{-1} X^T X(X^T X)^{-1} X^T = X(X^T X)^{-1} X^T.$$

3. $HX = X$

$$\text{Proof: } \{X(X^T X)^{-1} X^T\}X = XI = X.$$

Confidence and Prediction Intervals, 2

Properties of \hat{y} :

1. $\hat{y} = Hy = H(X\beta + \epsilon) = X\beta + H\epsilon$. Mean is $X\beta$.

2. The covariance matrix of \hat{y} is

$$\begin{aligned} E \left\{ (\hat{y} - X\beta)(\hat{y} - X\beta)^T \right\} &= E \left\{ H\epsilon(H\epsilon)^T \right\} \\ &= H \cdot E \left\{ \epsilon\epsilon^T \right\} \cdot H^T \\ &= H \cdot \sigma^2 I \cdot H^T \\ &= \sigma^2 H. \end{aligned}$$

3. In particular, $Var(\hat{y}_i) = h_i\sigma^2$.

4. For the standard linear model setup, σ is estimated by the residual standard deviation s , for which $\frac{s^2}{\sigma^2} \sim \frac{\chi_{n-p}^2}{n-p}$ *independently of $\hat{\beta}$* . Here n is the number of observations and p the number of covariates (including the intercept).

Confidence and Prediction Intervals, 3

Suppose we want a $100(1 - \alpha)\%$ confidence interval for $x_i^T \beta$, x_i the i 'th column of X . We have that \hat{y}_i is an unbiased estimator with variance $h_i \sigma^2$. Therefore:

$$\frac{\hat{y}_i - x_i^T \beta}{\sqrt{h_i} \sigma} \sim N[0, 1],$$
$$\frac{\hat{y}_i - x_i^T \beta}{\sqrt{h_i} s} \sim t_{n-p}$$

where n is the number of observations and p the number of parameters (including intercept). Therefore the desired *confidence interval* is

$$\hat{y}_i \pm t_{n-p, 1-\alpha/2} \cdot \sqrt{h_i} \cdot s.$$

where $t_{n-p, 1-\alpha/2}$ is the $1 - \alpha/2$ probability point of the t_{n-p} distribution (In R: `qt(1-alpha/2, n-p)`).

Confidence and Prediction Intervals, 4

Suppose, however, what we are really interested in is a *new* observation at x_i , say $y^* = x_i^T \beta + \epsilon^*$ where $\epsilon^* \sim N[0, \sigma^2]$ to mimic the errors in the regression already fitted. In that case,

$$y^* - \hat{y}_i \sim N[0, \sigma^2(h_i + 1)]$$

where the $+1$ in the variance term reflects the variance of ϵ^* .

So the *prediction standard error* is $s\sqrt{1 + h_i}$ and not $s\sqrt{h_i}$. The $100(1 - \alpha)\%$ *prediction interval* for y^* is

$$y^* \pm t_{n-p, 1-\alpha/2} \cdot \sqrt{1 + h_i} \cdot s. \quad (1)$$

In R, you can do this one of two ways: either explicitly evaluate formula (1) using `summary(lmod)$sigma` for s and `hatvalues(lmod)` for the vector of h_i , or use

`predict(lmod, interval='prediction', level=1-alpha).`

The latter is usually easier to remember and use!

Relevance to the Missing Votes Problem

1. To estimate the PNR in Bladen or Robeson county, we absolutely must take into account the natural variability of PNR, as well as the regression component.
2. The `se.fit` option with the `predict` command computes the confidence interval SE, not the prediction interval SE.
3. Therefore, we must either multiply the vector of confidence interval SEs by `sqrt((1+hatvalues(lmod))/hatvalues(lmod))` or (simpler) use the `interval='prediction'` option to compute prediction intervals directly.
4. This comment applies both to the original formulation of the question (estimate the probability of the observed value in Bladen and Robeson county) and the revised formulation (estimate number of lost votes), but the latter is simpler (and more meaningful) because it works with the prediction intervals directly.

COOK'S D STATISTIC

The most influential observations are those that have both large residuals and high leverage.

Cook's D statistic combines them both into a single measure.

Define p as the number of regressors (including intercept), \hat{y} the vector of predicted values, $\hat{y}_{(i)}$ the vector of predicted values when the i 'th observation is omitted, and $\hat{\sigma}^2$ the estimated residual variance.

D_i , the influence of observation i , is defined equivalently by

$$D_i = \frac{(\hat{y} - \hat{y}_{(i)})^T (\hat{y} - \hat{y}_{(i)})}{p \hat{\sigma}^2} = \frac{1}{p} \cdot r_i^2 \cdot \frac{h_i}{1 - h_i}.$$

Usually a D_i close to 1 is considered meaningful — in other words, we should investigate whether that observation really does need to be corrected (or omitted).

Comments on the rest of Chapter 1 (page 17 onwards)

1. *Robust Regression* (fit through the R package MASS) — this method became very popular for a while in the 1980s, but is less widely used now. You should be aware of it, but no need to study in depth.
2. *Weighted Least Squares*. Another method of accounting for heteroscedasticity is to weight each observation proportional to the sample size for that observation (in our example, the AbsBal variable). Faraway recommends *against* this option because the variances don't follow a simple scaling formula in practice. I suspect the same caveat applies with our example.
3. *Transformations of y*. Also called Box-Cox transformation. Could consider replacing PNR by logarithm or square root of PNR. One test: does this improve R^2 ?

4. *Transformations of x variables.* This could be a good idea if it improves the overall R^2 (or, equivalently, reduces the RSS). Faraway gives several examples. (Another variant would be to include interactions, e.g. cross-products of existing x variables. I know some of you tried that with our voting example. The criterion is whether the new variables improve the fit to a statistically significant extent.)
5. *Variable selection methods.* Several possibilities, e.g.
- (a) Maximize adjusted R^2 (simplest but not necessarily best)
 - (b) Minimize AIC (or BIC, DIC,.....)
 - (c) Forward, backward or stepwise regression (numerous variants)
 - (d) Newer “machine learning” methods, e.g. *lasso*

None of these methods is universally “best” — choice is partly a matter of personal preference (and the size of the dataset)

Homework 2: Due Friday, August 28

Questions 2 and 5 of the problems on page 24
(“rock” and “prostate” datasets)

- Submit through sakai “Assignments” tab
- Resubmissions are allowed
- Deadline 1:00 pm, Friday, August 28
- pdf file is preferred to html, but if you can’t figure that out, hand in the html.
- Please edit your output: don’t hand in all your raw code and output, only what’s relevant to your final conclusion