Gov 2000 - 10. Troubleshooting the Linear Model

Matthew Blackwell
Harvard University
mblackwell@gov.harvard.edu

Where are we? Where are we going?

- Last few weeks: estimation and inference for the linear model under Gauss-Markov assumptions (and sometimes conditional Normality)
- This week: what happens when the assumptions fail? Can we tell? Can we fix it?
- Next weeks: more of the same, then panel data

Review of the OLS assumptions

1. Linearity: $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$
2. Random/iid sample: $(y_i, \mathbf{x}_i')$ are a iid sample from the population.
3. No perfect collinearity: $\mathbf{X}$ is an $n \times (K+1)$ matrix with rank $K+1$
4. Zero conditional mean: $\mathbb{E}[\mathbf{u}|\mathbf{X}] = \mathbf{0}$
5. Homoskedasticity: $\text{var}(\mathbf{u}|\mathbf{X}) = \sigma_u^2 \mathbf{I}_n$
6. Normality: $\mathbf{u}|\mathbf{X} \sim N(\mathbf{0}, \sigma_u^2 \mathbf{I}_n)$

- 1-4 give us unbiasedness/consistency
- 1-5 are the Gauss-Markov, allow for large-sample inference
- 1-6 allow for small-sample inference

Violations of the assumptions

1. Nonlinearity
    - Result: biased/inconsistent estimates
    - Diagnose: scatterplots, added variable plots, component-plus-residual plots
    - Correct: transformations, polynomials, different model
2. iid/random sample
    - Result: no bias with appropriate alternative assumptions (structured dependence)
    - Result (ii): violations imply heteroskedasticity
    - Result (iii): outliers from different distributions can cause inefficiency/bias
    - Diagnose/Correct: next week!
3. Perfect collinearity
    - Result: can't run OLS
    - Diagnose/correct: drop one collinear term
4. Zero conditional mean error
    - Result: biased/inconsistent estimates
    - Diagnose: very difficult
    - Correct: instrumental variables (Gov 2002)
5. Heteroskedasticity
    - Result: SEs are biased (usually downward)
    - Diagnose/correct: next week!
6. Non-Normality
    - Result: critical values for t and F tests wrong
    - Diagnose: checking the (studentized) residuals, QQ-plots, etc
    - Correct: transformations, add variables to $\mathbf{X}$, different model

NONNORMALITY OF THE ERRORS

Review of the Normality assumption

$$\mathbf{u}|\mathbf{X} \sim \mathcal{N}(0, \sigma_u^2 \mathbf{I})$$

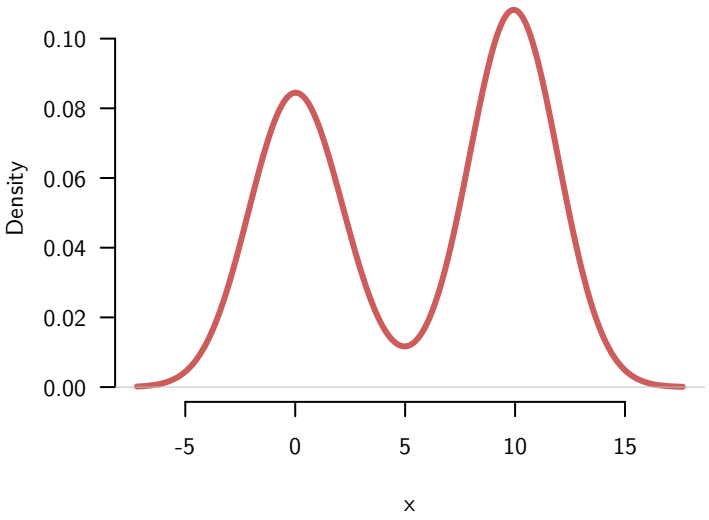
What are the consequences of non-Normal errors?

- In small samples:
  - Sampling distribution of $\widehat{\boldsymbol{\beta}}$ will not be Normal
  - Test statistics will not have t or F distributions
  - Probability of Type I error will not be α
  - $1 - \alpha$ confidence interval will not have $1 - \alpha$ coverage
- In large samples:
  - Sampling distribution of $\widehat{\boldsymbol{\beta}} \approx$ Normal by the CLT
  - Test statistics will be $\approx t$ or F by the CLT
  - Probability of Type I error $\approx \alpha$
  - $1 - \alpha$ confidence interval will have $\approx 1 - \alpha$ coverage
- The n needed for approximation to hold depends on how non-Normal the data is

Is this a violation?

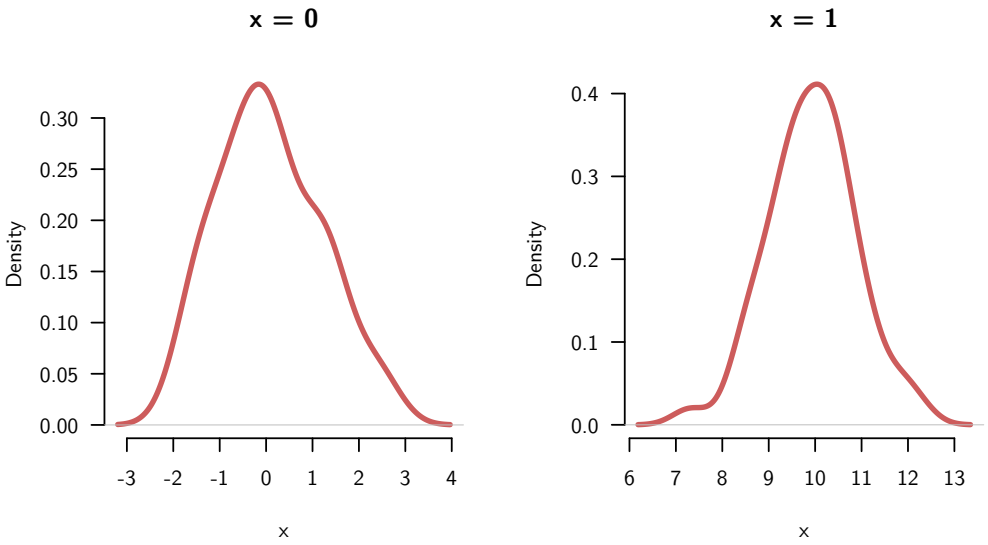
- Be careful with this assumption. Remember it's the distribution of the error, or the distribution of the outcome conditional on the independent variables.
- The **marginal distribution** of Y can be non-Normal even if the conditional distribution is Normal!
- For example, this looks bad:

```
x <- rbinom(100, 1, 0.5)
u <- rnorm(100, 0, 1)
y <- 10 * x + u
plot(density(y), lwd = 3, col = "indianred", las = 1, xlab = "x", main = "",
    bty = "n")
```



- But if we look at the conditional distributions, things look better:

```
par(mfrow = c(1, 2))
plot(density(y[x == 0]), lwd = 3, col = "indianred", las = 1, xlab = "x", main = "x = 0",
    bty = "n")
plot(density(y[x == 1]), lwd = 3, col = "indianred", las = 1, xlab = "x", main = "x = 1",
    bty = "n")
```



How to diagnose?

- Assumption is about unobserved $\mathbf{u} = \mathbf{y} - \mathbf{X}\boldsymbol{\beta}$
- We can only observe residuals, $\widehat{\mathbf{u}} = \mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}$

Hat matrix

$$\begin{aligned}\widehat{\mathbf{u}} &= \mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}} \\ &= \mathbf{y} - \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y} \\ &\equiv \mathbf{y} - \mathbf{H}\mathbf{y} \\ &= (\mathbf{I} - \mathbf{H})\mathbf{y}\end{aligned}$$

- $\mathbf{H}$ is the **hat matrix** because it puts the "hat" on $\mathbf{y}$:

$$\widehat{\mathbf{y}} = \mathbf{H}\mathbf{y}$$

  - $\mathbf{H}$ is an $n \times n$ symmetric matrix
  - $\mathbf{H}$ is **idempotent**: $\mathbf{H}\mathbf{H} = \mathbf{H}$

Relationship between the residuals and the errors

$$\begin{aligned}\widehat{\mathbf{u}} &= (\mathbf{I} - \mathbf{H})(\mathbf{X}\boldsymbol{\beta} + \mathbf{u}) \\ &= (\mathbf{I} - \mathbf{H})\mathbf{X}\boldsymbol{\beta} + (\mathbf{I} - \mathbf{H})\mathbf{u} \\ &= \mathbf{I}\mathbf{X}\boldsymbol{\beta} - \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta} + (\mathbf{I} - \mathbf{H})\mathbf{u} \\ &= (\mathbf{I} - \mathbf{H})\mathbf{u}\end{aligned}$$

Distribution of the residuals

$$\mathbb{E}[\widehat{\mathbf{u}}] = (I - H)\mathbb{E}[\mathbf{u}] = \mathbf{0}$$
$$\text{Var}[\mathbf{u}] = \sigma_u^2(\mathbf{I} - \mathbf{H})$$

- Variance of the ith residual:

$$\text{Var}[\widehat{u}_i] = \sigma_u^2(1 - h_{ii})$$

- The residuals are not independent:

  - We know that the must sum up to o: $\sum_i \widehat{u}_i = 0$, so if I know $n - 1$ of them, I know the last one too.

- Residuals not independent, nor identically distributed, even when all the OLS assumptions hold

Standardized residuals

$$\widehat{u}_i' = \frac{\widehat{u}_i}{\widehat{\sigma}\sqrt{1 - h_{ii}}}$$

Studentized residuals

$$\widehat{u}_i^* = \frac{\widehat{u}_i}{\widehat{\sigma}_{-i}\sqrt{1 - h_{ii}}}$$

- If the errors are Normal, the studentized residuals follow a t distribution with $(n - k - 2)$ degrees of freedom.
- We can use this distribution to check to see if our residuals seem to conform

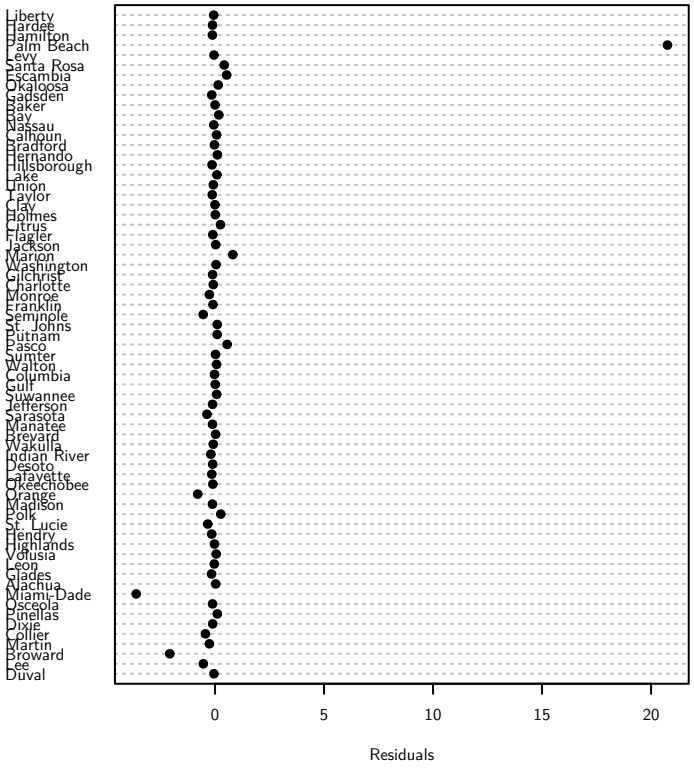
Buchanan vote example

```
flvote <- foreign::read.dta("flbuchan.dta")
mod <- lm(edaybuchanan ~ edaytotal, data = flvote)
summary(mod)
```

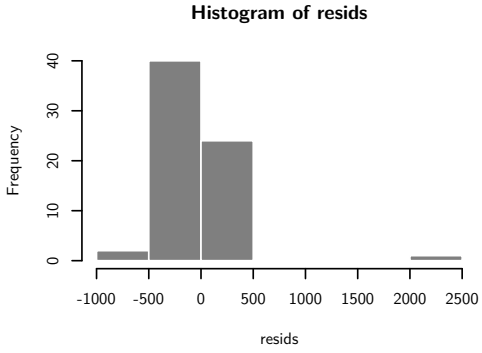
```
##
## Call:
## lm(formula = edaybuchanan ~ edaytotal, data = flvote)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -947.05  -41.74  -19.47   20.20 2350.54
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 5.423e+01  4.914e+01   1.104    0.274
## edaytotal   2.323e-03  3.104e-04   7.483 2.42e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 332.7 on 65 degrees of freedom
## Multiple R-squared:  0.4628, Adjusted R-squared:  0.4545
## F-statistic:    56 on 1 and 65 DF,  p-value: 2.417e-10
```

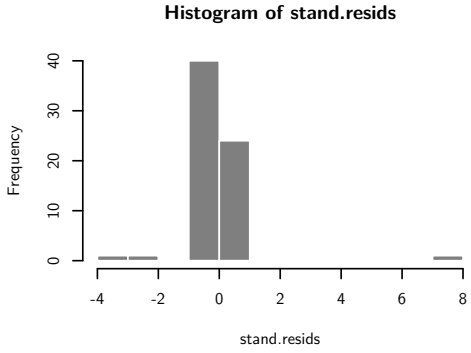
```
resids <- residuals(mod)
stand.resids <- rstandard(mod)
student.resids <- rstudent(mod)
```

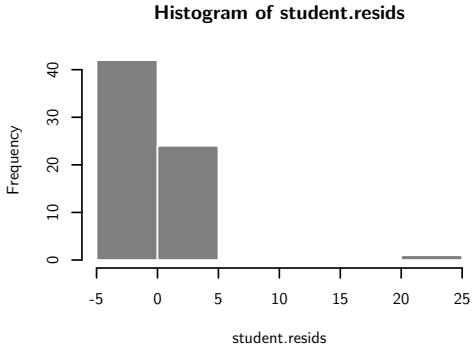
```
dotchart(student.resids, flvote$county, cex = 0.7, xlab = "Residuals", pch = 19)
```



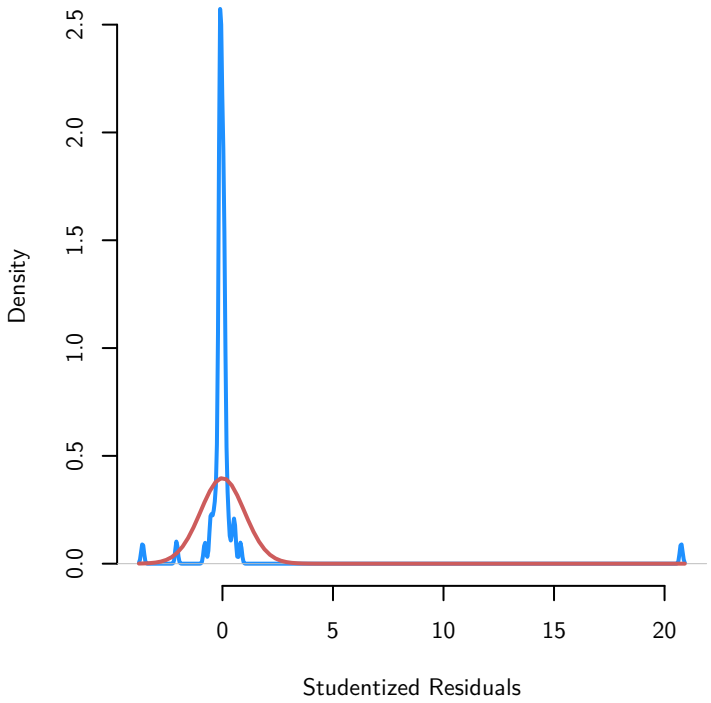

```
par(mfrow = c(2, 2))
hist(resids, col = "grey50", border = "white")
hist(stand.resids, col = "grey50", border = "white")
hist(student.resids, col = "grey50", border = "white")
```





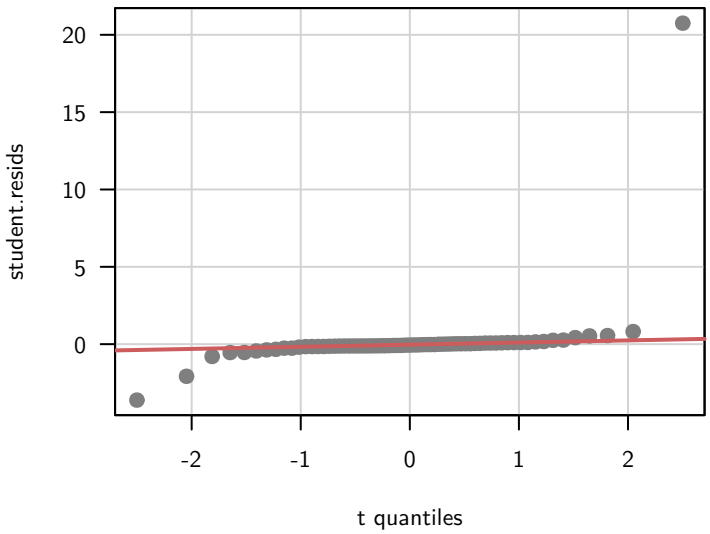



```
plot(density(student.resids), col = "dodgerblue", xlab = "Studentized Residuals",
    ylab = "Density", bty = "n", lwd = 2, main = "")
curve(dt(x, nrow(flvote) - 1 - 2), col = "indianred", add = TRUE, lwd = 2)
```



Quantile-Quantile plots

```
library(car)
qqPlot(student.resids, distribution = "t", df = nrow(flvote) - 1 - 2, envelope = TRUE,
    pch = 19, cex = 1.25, col = "grey50", col.lines = "indianred", las = 1)
```



Dealing with non-Normal errors

- Remove problematic observations (be transparent!)
- Add or drop variables in $\mathbf{X}$
- Transform $\mathbf{y}$ ($\log(\mathbf{y})$)

Buchanan revisited

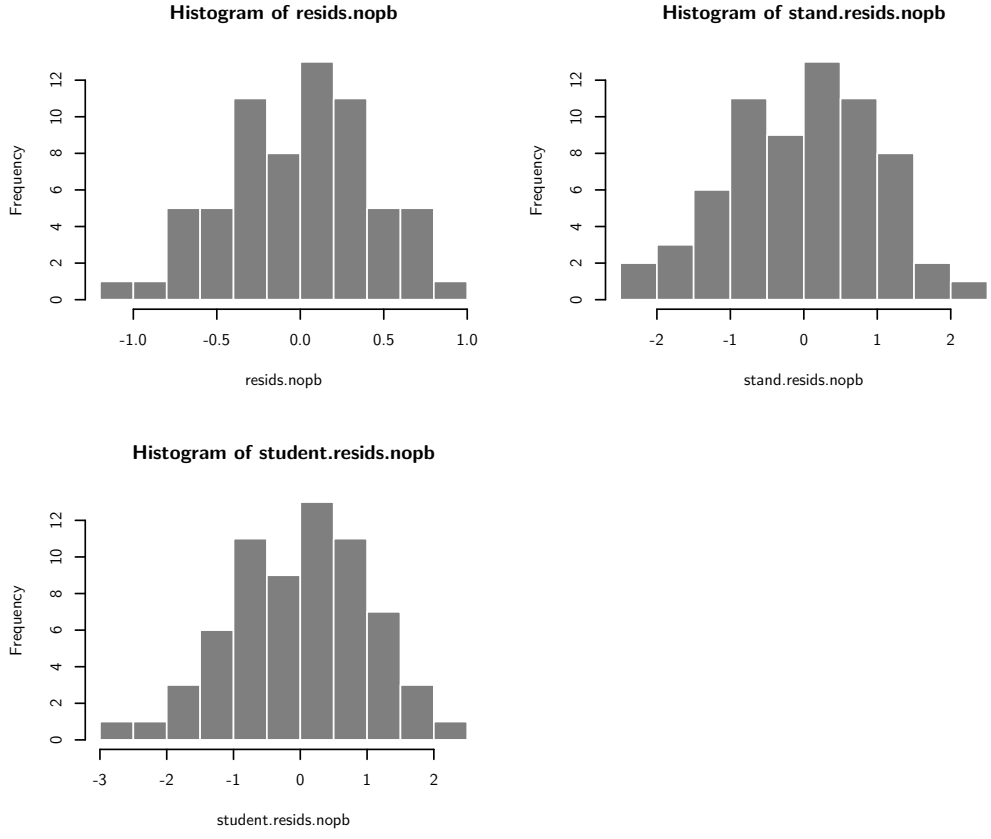
```
flvote.nopb <- flvote[flvote$county != "Palm Beach", ]
mod.nopb <- lm(log(edaybuchanan) ~ log(edaytotal), data = flvote.nopb)
summary(mod.nopb)
```

```
##
## Call:
## lm(formula = log(edaybuchanan) ~ log(edaytotal), data = flvote.nopb)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.02692 -0.23283  0.02836  0.29525  0.97317
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -2.48597    0.37889  -6.561 1.09e-08 ***
## log(edaytotal)  0.70311    0.03621  19.417  < 2e-16 ***
```

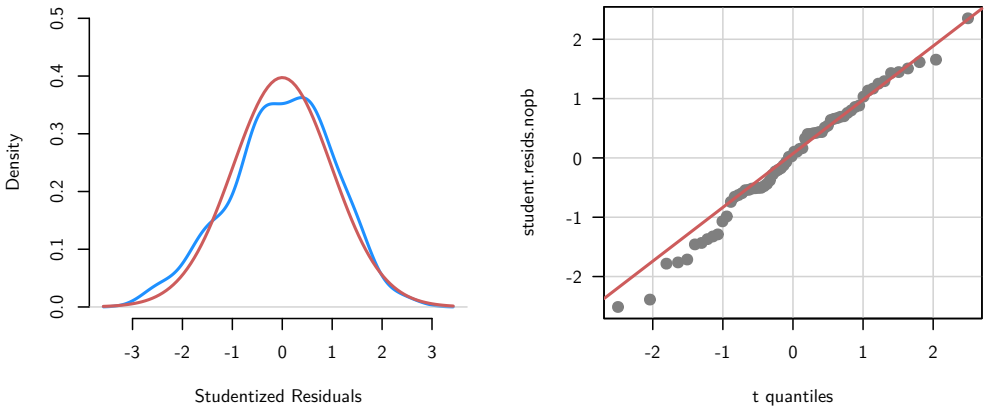
```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4362 on 64 degrees of freedom
## Multiple R-squared:  0.8549, Adjusted R-squared:  0.8526
## F-statistic:   377 on 1 and 64 DF,  p-value: < 2.2e-16
```

```r
resids.nopb <- residuals(mod.nopb)
stand.resids.nopb <- rstandard(mod.nopb)
student.resids.nopb <- rstudent(mod.nopb)
```

```r
par(mfrow = c(2, 2))
hist(resids.nopb, col = "grey50", border = "white")
hist(stand.resids.nopb, col = "grey50", border = "white")
hist(student.resids.nopb, col = "grey50", border = "white")
```



```r
par(mfrow = c(1, 2))
plot(density(student.resids.nopb), col = "dodgerblue", xlab = "Studentized Residuals",
    ylab = "Density", bty = "n", lwd = 2, main = "", ylim = c(0, 0.5))
curve(dt(x, nrow(flvote) - 1 - 2), col = "indianred", add = TRUE, lwd = 2)
library(car)
qqPlot(student.resids.nopb, distribution = "t", df = nrow(flvote) - 1 - 2, envelope = TRUE,
    pch = 19, cex = 1.25, col = "grey50", col.lines = "indianred", las = 1)
```



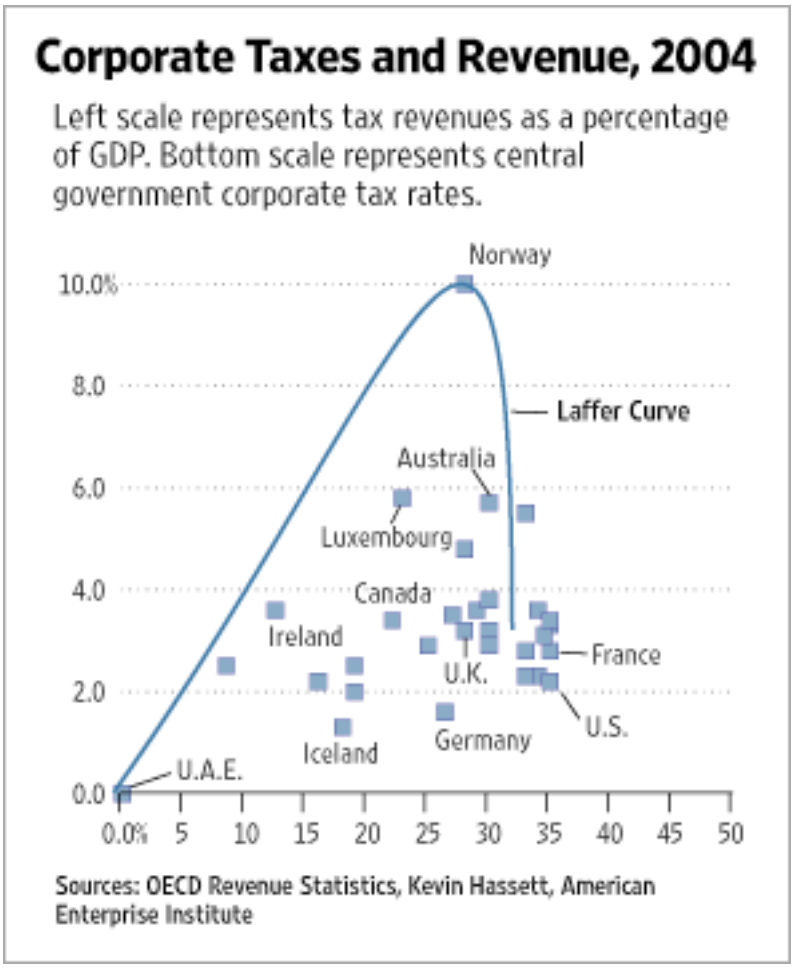
OUTLIERS, LEVERAGE POINTS, AND INFLUENTIAL OBSERVATIONS

The trouble with Norway

- Lange and Garrett (1985): organizational and political power of labor interact to improve economic growth
- Jackman (1987): relationship just due to North Sea Oil?
- Table guide:
  - x_1 = organizational power of labor
  - x_2 = political power of labor
  - Parentheses contain t-statistics

	Constant	x_1	x_2	$x_1 \cdot x_2$
Norway Obs Included	.814	-.192	-.278	.137
	(4.7)	(2.0)	(2.4)	(2.9)
Norway Obs Excluded	.641	-.068	-.138	.054
	(4.8)	(0.9)	(1.5)	(1.3)

Creative curve fitting with Norway



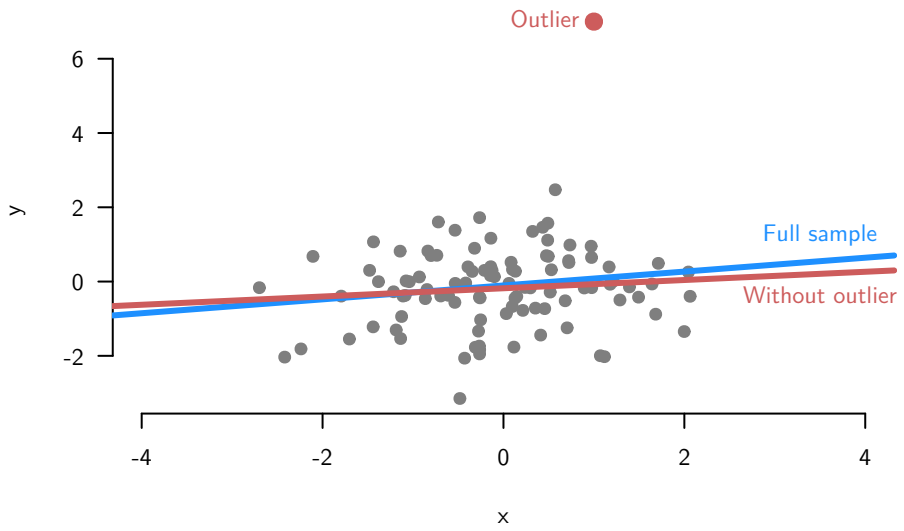
Three types of extreme values

1. Outlier: extreme in the y direction
2. Leverage point: extreme in one x direction
3. Influence point: extreme in both directions

- Not all of these are problematic
- If the data are truly "contaminated" (come from a different distribution), can cause inefficiency and possibly bias
- Can be a violation of iid (not identically distributed)

Outlier definition

```
set.seed(2138)
x <- rnorm(100, 0, 1)
y <- 0.2 * x + rnorm(100, 0, 1)
x <- c(x, 1)
y <- c(y, 7)
plot(x, y, pch = 19, col = ifelse(1:101 != 101, "grey50", "indianred"), cex = ifelse(1:101 !=
    101, 1, 1.5), bty = "n", las = 1, xlim = c(-4, 4))
abline(lm(y ~ x), col = "dodgerblue", lwd = 3)
abline(lm(y[1:100] ~ x[1:100]), col = "indianred", lwd = 3)
text(3.5, -0.35, "Without outlier", col = "indianred")
text(3.5, 1.25, "Full sample", col = "dodgerblue")
text(1, 7, "Outlier", pos = 2, col = "indianred")
```



- An **outlier** is a data point with very large regression errors, u_i
- Very distant from the rest of the data in the y-dimension
- Increases standard errors (by increasing $\widehat{\sigma}^2$)
- No bias if typical in the x's

Detecting outliers

- One possible approach: Look at standardized residuals, $\widehat{u}_i'$, but this is problematic because it depends on $\widehat{\sigma}^2$, which could be biased upwards by the large residual from the outlier

- Makes detecting residuals harder
- Possible solution: use studentized residuals

$$\widehat{u}_i^* = \frac{\widehat{u}_i}{\widehat{\sigma}_{-i}\sqrt{1-h_i}}$$

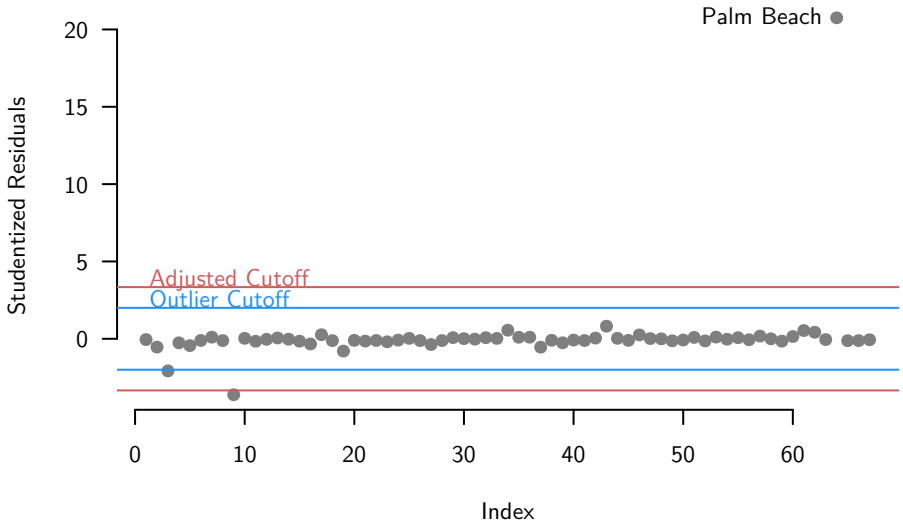
- $\widehat{\sigma} > \widehat{\sigma}_{-i}$ because we drop the large residual from the outlier, and so $\widehat{u}_i' < \widehat{u}_i^*$

Cutoff rules for outliers

- The studentized residuals follow a t distribution, $u_i^* \sim t_{n-k-2}$, when $u_i \sim N(0, \sigma^2)$
- Rule of thumb: $|\widehat{u}_i^*| > 2$ will be relatively rare, but if you have $n > 100$ this is likely to happen in the data.
- Extreme outliers, $|\widehat{u}_i^*| > 4 - 5$ are much less likely.
- Some people use more formal tests, but it is unclear what to do with those results
- Determination is ultimately subjective

Buchanan outliers

```
plot(student.resids, pch = 19, col = "grey50", bty = "n", las = 1, ylab = "Studentized Residuals")
abline(h = c(-2, 2), col = "dodgerblue")
abline(h = c(-1, 1) * qt(1 - 7e-04, df = nrow(flvote) - 1 - 2), col = "indianred")
text(x = which(flvote$county == "Palm Beach"), student.resids[which(flvote$county ==
    "Palm Beach")], "Palm Beach", pos = 2)
text(x = 0, y = 2.4, "Outlier Cutoff", col = "dodgerblue", pos = 4)
text(x = 0, y = 3.8, "Adjusted Cutoff", col = "indianred", pos = 4)
```



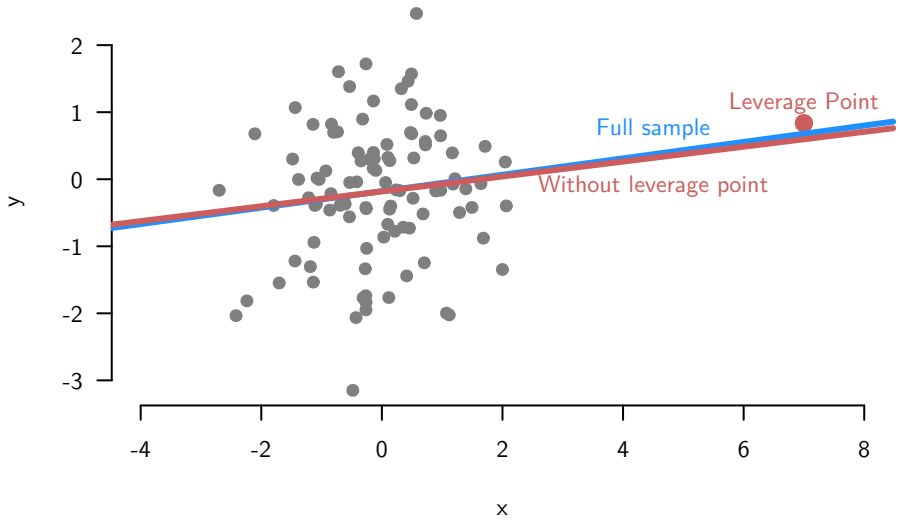
What to do about outliers

- Is the data corrupted?
  - Fix the observation (obvious data entry errors)
  - Remove the observation
  - Be transparent either way
- Is the outlier part of the data generating process?
  - Transform the dependent variable ($\log(y)$)
  - Use a method that is robust to outliers (robust regression, least absolute deviations)

Leverage point definition

```
set.seed(2138)
x <- rnorm(100, 0, 1)
x <- c(x, 7)
y <- 0.2 * x + rnorm(101, 0, 1)
plot(x, y, pch = 19, col = ifelse(1:101 != 101, "grey50", "indianred"), cex = ifelse(1:101 !=
    101, 1, 1.5), bty = "n", las = 1, xlim = c(-4, 8))
abline(lm(y ~ x), col = "dodgerblue", lwd = 3)
abline(lm(y[1:100] ~ x[1:100]), col = "indianred", lwd = 3)
text(4.5, -0.1, "Without leverage point", col = "indianred")
```

```
text(4.5, 0.75, "Full sample", col = "dodgerblue")
text(7, y[101], "Leverage Point", pos = 3, col = "indianred")
```



Hat values

$$\widehat{\mathbf{y}} = \mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \mathbf{H}\mathbf{y}$$

- For a particular observation i, we can show this means:

$$\widehat{y}_i = \sum_{j=1}^{n} h_{ij} y_j$$

- Thus, h_{ij} indicates how important observation j is for the fitted value $\widehat{y}_i$
- $h_i = h_{ii}$ are the diagonal entries of the hat matrix and have the following property

$$h_{ii} = \sum_{j=1}^{n} h_{ij}^2$$

- Summarizes how important i is for all fitted values
- With a simple linear regression, we have

$$h_i = \frac{1}{n} + \frac{(X_i - \overline{X})^2}{\sum_{j=1}^{n}(X_j - \overline{X})^2}$$

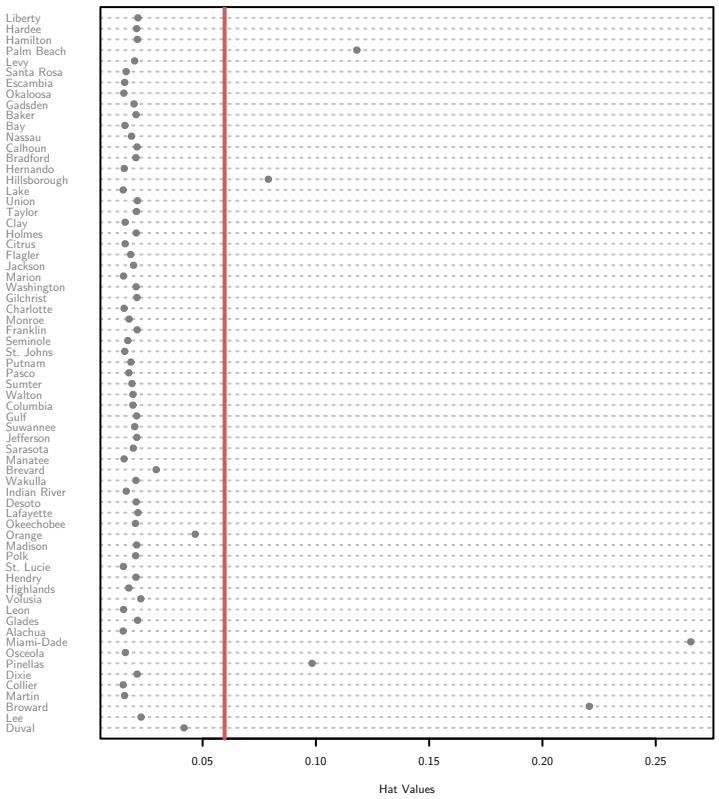
- Basically, it measures how far i is from the center of the data, relative to the overall variation

Facts about hat values

- $\sum_{i=1}^{n} h_i = k + 1$
- $1/n \geq h_i \geq 1$ for all i
- $\text{Var}[\widehat{u}_i] = (1 - h_i)\sigma^2$
- Rule of thumb: examine hat values greater than $2(k + 1)/n$

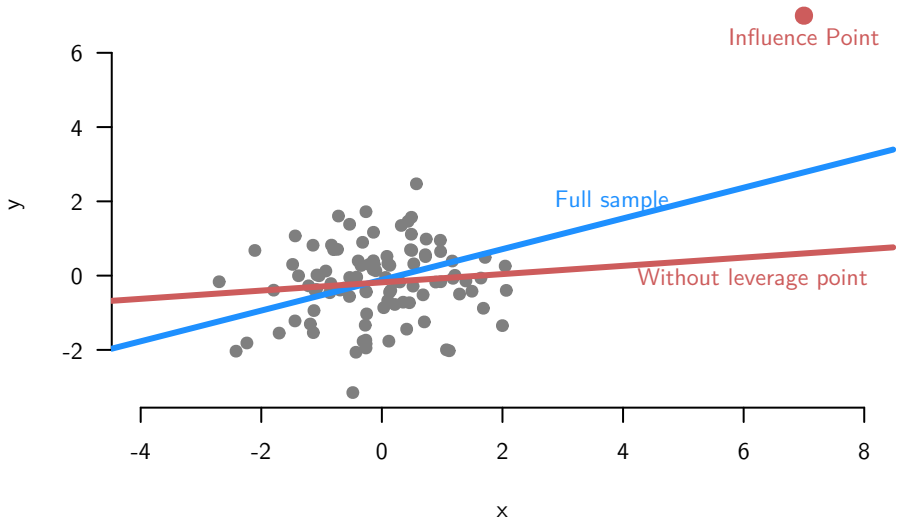
Buchanan hats

```
dotchart(hatvalues(mod), flvote$county, cex = 0.5, xlab = "Hat Values", pch = 19,
    col = "grey50")
abline(v = 2 * (2)/nrow(flvote), col = "indianred", lwd = 2)
```



Influence points

```
set.seed(2138)
x <- rnorm(100, 0, 1)
y <- 0.2 * x + rnorm(100, 0, 1)
x <- c(x, 7)
y <- c(y, 7)
plot(x, y, pch = 19, col = ifelse(1:101 != 101, "grey50", "indianred"), cex = ifelse(1:101 !=
    101, 1, 1.5), bty = "n", las = 1, xlim = c(-4, 8))
abline(lm(y ~ x), col = "dodgerblue", lwd = 3)
abline(lm(y[1:100] ~ x[1:100]), col = "indianred", lwd = 3)
text(4, -0.1, "Without leverage point", pos = 4, col = "indianred")
text(5, 2, "Full sample", pos = 2, col = "dodgerblue")
text(7, 7, "Influence Point", pos = 1, col = "indianred")
```



- An **influence point** is one that is both an outlier and a leverage point.
- Extreme in both the x and y dimensions
-

Influence

$$D_{ij} = \widehat{\beta}_j - \widehat{\beta}_{j(-i)}$$

- $\widehat{\beta}_{j(-i)}$ is the estimated coefficient dropping observation i

- D_{ij} sometimes called DFbeta

Standardized influence

$$D_{ij}^* = \frac{\widehat{\beta}_j - \widehat{\beta}_{j(-i)}}{\widehat{SE}_{-i}[\widehat{\beta}_j]}$$

- $\widehat{SE}_{-i}[\widehat{\beta}_j]$ is the SE from "leave-one-out" regression
- $D_{ij}^* > 0$ implies that deleting i will decrease the coefficient (i has a positive influence)
- $D_{ij}^* < 0$ implies that deleting i will increase the coefficient (i has a negative influence)
- $|D_{ij}^*| > 2/\sqrt{n}$ are an indication of high influence

Buchanan influence

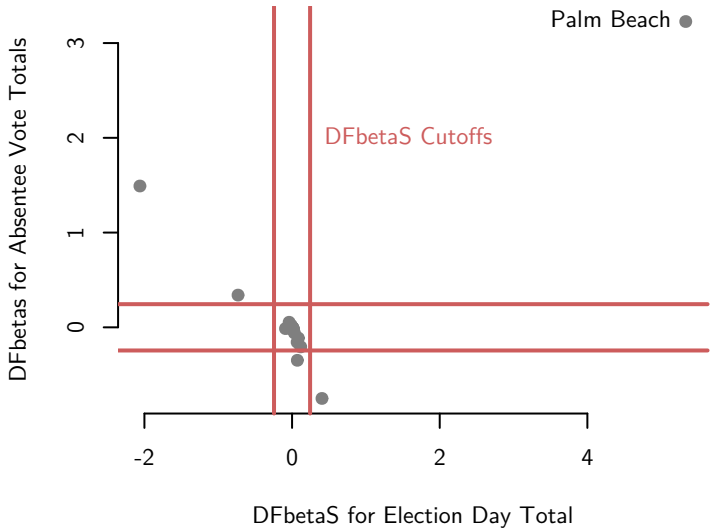
```
mod3 <- lm(edaybuchanan ~ edaytotal + absnbuchanan, data = flvote)
summary(mod3)
```

```
##
## Call:
## lm(formula = edaybuchanan ~ edaytotal + absnbuchanan, data = flvote)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -444.12  -44.30    7.42   35.84 2267.38
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -2.935e+01  5.520e+01  -0.532  0.59686
## edaytotal     1.100e-03  4.797e-04   2.293  0.02529 *
## absnbuchanan  6.895e+00  2.129e+00   3.238  0.00195 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 317.2 on 61 degrees of freedom
##   (3 observations deleted due to missingness)
## Multiple R-squared:  0.5361, Adjusted R-squared:  0.5209
## F-statistic: 35.24 on 2 and 61 DF,  p-value: 6.711e-11
```

```
head(dfbetas(mod3))
```

```
##     (Intercept)     edaytotal  absnbuchanan
## 1  0.3454475146  0.4050504921 -0.7505222758
## 2 -0.0234266617 -0.0241000045 -0.0131672181
## 3  0.0650795039 -0.7319311820  0.3401669862
## 4 -0.0333980968  0.0133802934 -0.0087505576
## 5 -0.0397626659 -0.0073746223  0.0096551713
## 6 -0.0009277798  0.0001505476  0.0002210247
```

```
plot(dfbetas(mod3)[, "edaytotal"], dfbetas(mod3)[, "absnbuchanan"], pch = 19,
    col = "grey50", bty = "n", xlab = "DFbetaS for Election Day Total", ylab = "DFbetas for Absentee Vote Totals"
pb.index <- which(flvote$county == "Palm Beach")
text(dfbetas(mod3)[pb.index, "edaytotal"], dfbetas(mod3)[pb.index, "absnbuchanan"],
    "Palm Beach", pos = 2)
abline(v = c(-2, 2)/sqrt(nrow(flvote)), col = "indianred", lwd = 2)
abline(h = c(-2, 2)/sqrt(nrow(flvote)), col = "indianred", lwd = 2)
text(x = 2/sqrt(nrow(flvote)), y = 2, "DFbetaS Cutoffs", pos = 4, col = "indianred")
```



- Palm Beach county moves each of the coefficients by more than 3 standard errors!

Overall measures of influence

- The previous measures were for a single covariate. What about across all covariates?
- **Cook's distance**:

$$D_i = \frac{\widehat{u}_i'}{k+1} \times \frac{h_i}{1-h_i}$$

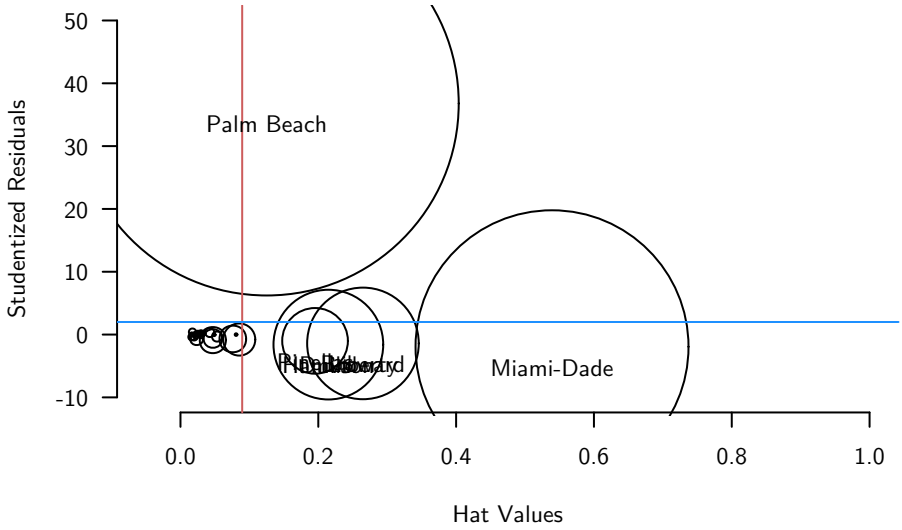
- Remember here that $\widehat{u}_i'$ is the standardized residual and h_i is the hat value.
- Basically this is "outlier $\times$ leverage"
- $D_i > 4/(n-k-1)$ considered "large"

Influence plot

- x-axis: hat values, h_i
- y-axis: studentized residuals, $\widehat{u}_i^*$
- size of points: Cook's distance

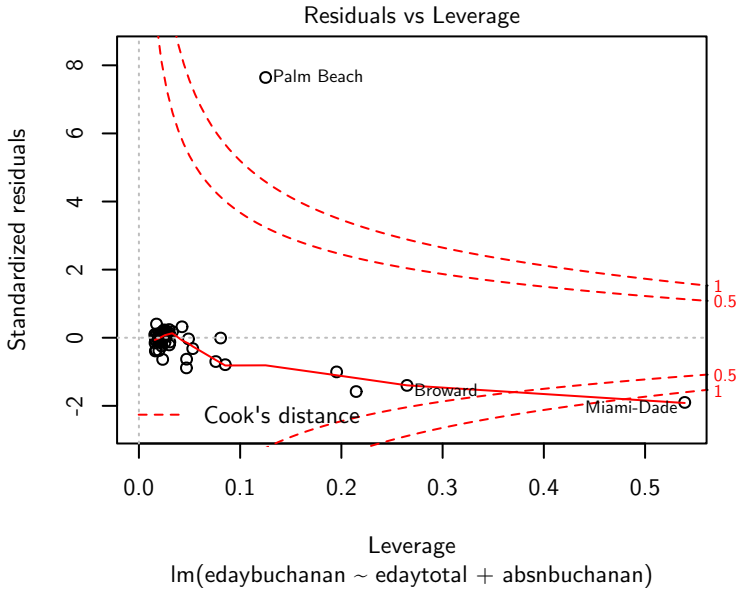
Influence Plot Buchanan

```
symbols(y = rstudent(mod3), x = hatvalues(mod3), circles = sqrt(cooks.distance(mod3)),
    ylab = "Studentized Residuals", xlab = "Hat Values", xlim = c(-0.05, 1),
    ylim = c(-10, 50), las = 1, bty = "n")
cutoffstud <- 2
cutoffhat <- 2 * (3)/nrow(flvote)
abline(v = cutoffhat, col = "indianred")
abline(h = cutoffstud, col = "dodgerblue")
filter <- rstudent(mod3) > cutoffstud | hatvalues(mod3) > cutoffhat
text(y = rstudent(mod3)[filter], x = hatvalues(mod3)[filter], flvote$county[filter],
    pos = 1)
```



Influence plot from lm output

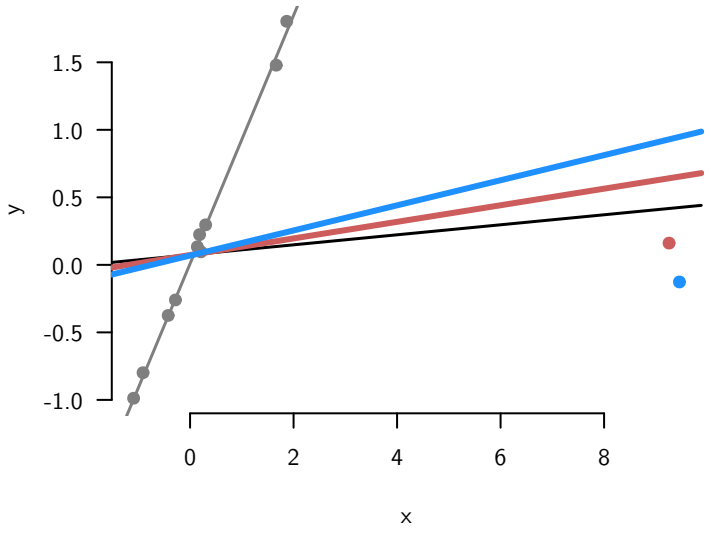
```
plot(mod3, which = 5, labels.id = flvote$county)
```



Limitations of the standard tools

```
set.seed(2143)
x <- c(rnorm(10, 0, 1), rnorm(2, 9, 0.5))
y <- ifelse(x > 6, 0, x) + rnorm(12, 0, 0.1)

plot(x, y, pch = 19, col = c(rep("grey50", 10), "dodgerblue", "indianred"),
    las = 1, bty = "n")
abline(lm(y ~ x), lwd = 1.5)
abline(lm(y[1:10] ~ x[1:10]), col = "grey50", lwd = 1.5)
abline(lm(y[-12] ~ x[-12]), col = "indianred", lwd = 3)
abline(lm(y[-11] ~ x[-11]), col = "dodgerblue", lwd = 3)
```



- What happens when there are two influence points?
- Red line drops the red influence point
- Blue line drops the blue influence point
- Neither of the "leave-one-out" approaches helps recover the line

NONLINEARITY OF THE REGRESSION FUNCTION

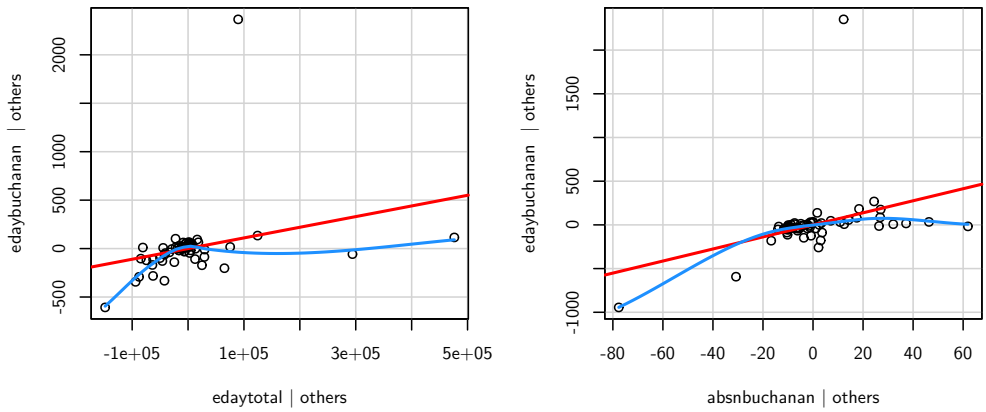
Added variable plot

- Need a way to visualize conditional relationship between Y and X_j
- How to construct an **added variable plot**:

1. Get residuals from regression of Y on all covariates except X_j
2. Get residuals from regression of X_j on all other covariates
3. Plot residuals from (1) against residuals from (2)

- In R: `avPlots(model)` from the `car` package
- OLS fit to this plot will have exactly $\widehat{\beta}_j$ and 0 intercept
- Use local smoother (loess) to detect any non-linearity

```
par(mfrow = c(1, 2))
out <- avPlots(mod3, "edaytotal")
lines(loess.smooth(x = out$edaytotal[, 1], y = out$edaytotal[, 2]), col = "dodgerblue",
    lwd = 2)
out2 <- avPlots(mod3, "absnbuchanan")
lines(loess.smooth(x = out2$absnbuchanan[, 1], y = out2$absnbuchanan[, 2]),
    col = "dodgerblue", lwd = 2)
```



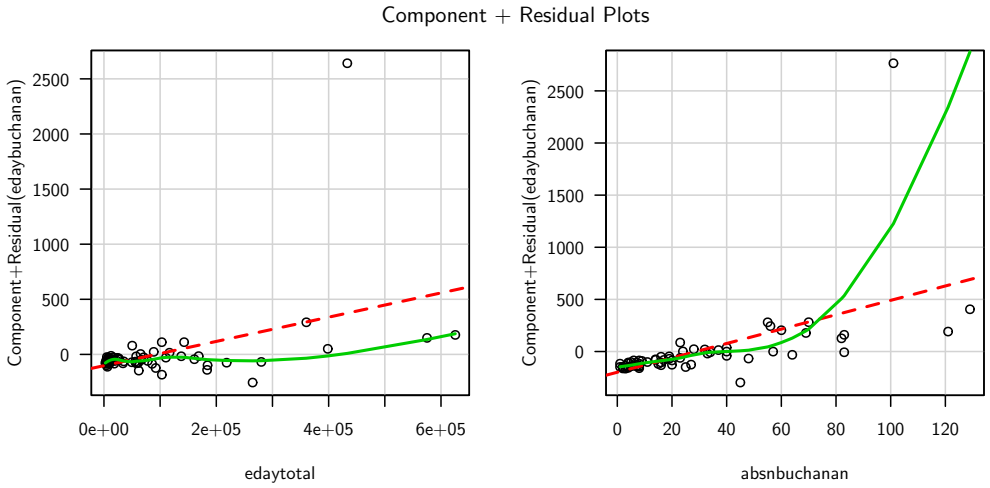
Component-Residual plots

- CR plots are a refinement of AV plots:
  1. Compute residuals from full regression:
  $$\widehat{u}_i = Y_i - \widehat{Y}_i$$
  2. Compute "linear component" of the partial relationship:
  $$C_i = \widehat{\beta}_j X_{ij}$$
  3. Add linear component to residual:
  $$\widehat{u}_i^j = \widehat{u}_i + C_i$$

4. Plot partial residual $\widehat{u}_i^j$ against X_j

- Same slope as AV plots
- X-axis is the original scale of X_j, so slightly easier for diagnostics
- Use local smoother (loess) to detect non-linearity

```
crPlots(mod3, las = 1)
```

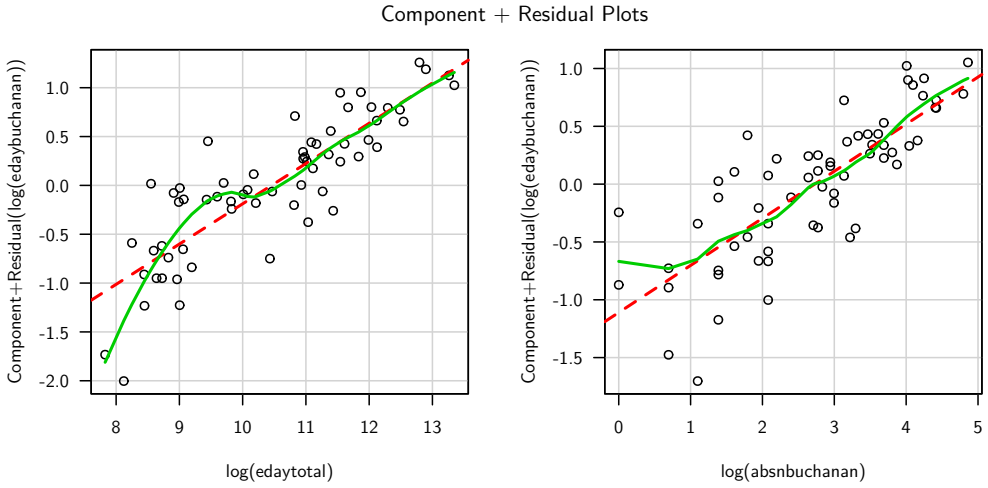


How to deal with non-linearity

- Breaking up categorical variables into dummy variables
- Including interaction terms
- Including polynomial terms
- Using transformations
- Using more flexible models (GAMs, nonlinear models, Gov 2001+)

Transformed Buchanan regression

```
mod.nopb2 <- lm(log(edaybuchanan) ~ log(edaytotal) + log(absnbuchanan), data = flvote,
    subset = county != "Palm Beach")
crPlots(mod.nopb2, las = 1)
```


Component + Residual Plots
Component+Residual(log(edaybuchanan))
log(edaytotal)
Component+Residual(log(edaybuchanan))
log(absnbuchanan)