

# Visualizing Linear Models: An R Bag of Tricks Session 1: Getting Started

Michael Friendly  
SCS Short Course  
Oct. 2020

# Today's topics

- What you need for this course
- Why plot your data?
- Data plots
- Model (effect) plots
- Diagnostic plots

# What you need

- R, version  $\geq 3.6$ 
  - Download from <https://cran.r-project.org/>
- RStudio IDE, highly recommended
  - <https://www.rstudio.com/products/rstudio/>
- R packages: see course web page

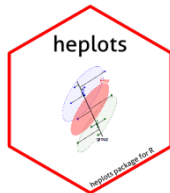
- car



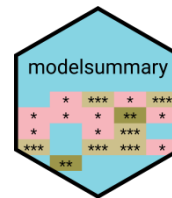
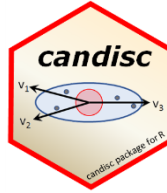
- effects



- heplots



- candisc



# Why plot your data?

*Getting information from a table is like extracting sunlight from a cucumber. --- Farquhar & Farquhar, 1891*

*Information that is imperfectly acquired, is generally as imperfectly retained; and a man who has carefully investigated a printed table, finds, when done, that he has only a very faint and partial idea of what he has read; and that like a figure imprinted on sand, is soon totally erased and defaced.*

*--- William Playfair, The Commercial and Political Atlas (p. 3), 1786*



# Cucumbers

**Table 7**  
Stevens et al. 2006, table 2: Determinants of authoritarian aggression

Variable	Coefficient (Standard Error)
Constant	.41 (.93)
Countries	
Argentina	1.31 (.33)**B,M
Chile	.93 (.32)**B,M
Colombia	1.46 (.32)**B,M
Mexico	.07 (.32) <sup>A,CH,CO,V</sup>
Venezuela	.96 (.37)**B,M
Threat	
Retrospective egocentric economic perceptions	.20 (.13)
Prospective egocentric economic perceptions	.22 (.12) <sup>#</sup>
Retrospective sociotropic economic perceptions	-.21 (.12) <sup>#</sup>
Prospective sociotropic economic perceptions	-.32 (.12)*
Ideological distance from president	-.27 (.07)**
Ideology	
Ideology	.23 (.07)**
Individual Differences	
Age	.00 (.01)
Female	-.03 (.21)
Education	.13 (.14)
Academic Sector	.15 (.29)
Business Sector	.31 (.25)
Government Sector	-.10 (.27)
$R^2$	.15
Adjusted $R^2$	.12
$N$	500

Results of a one model for authoritarian aggression

The information is overwhelmed by footnotes & significance \*\*stars\*\*

\*\*p < .01, \*p < .05, #p < .10 (twotailed)

<sup>A</sup>Coefficient is significantly different from Argentina's at p < .05;

<sup>B</sup>Coefficient is significantly different from Brazil's at p < .05;

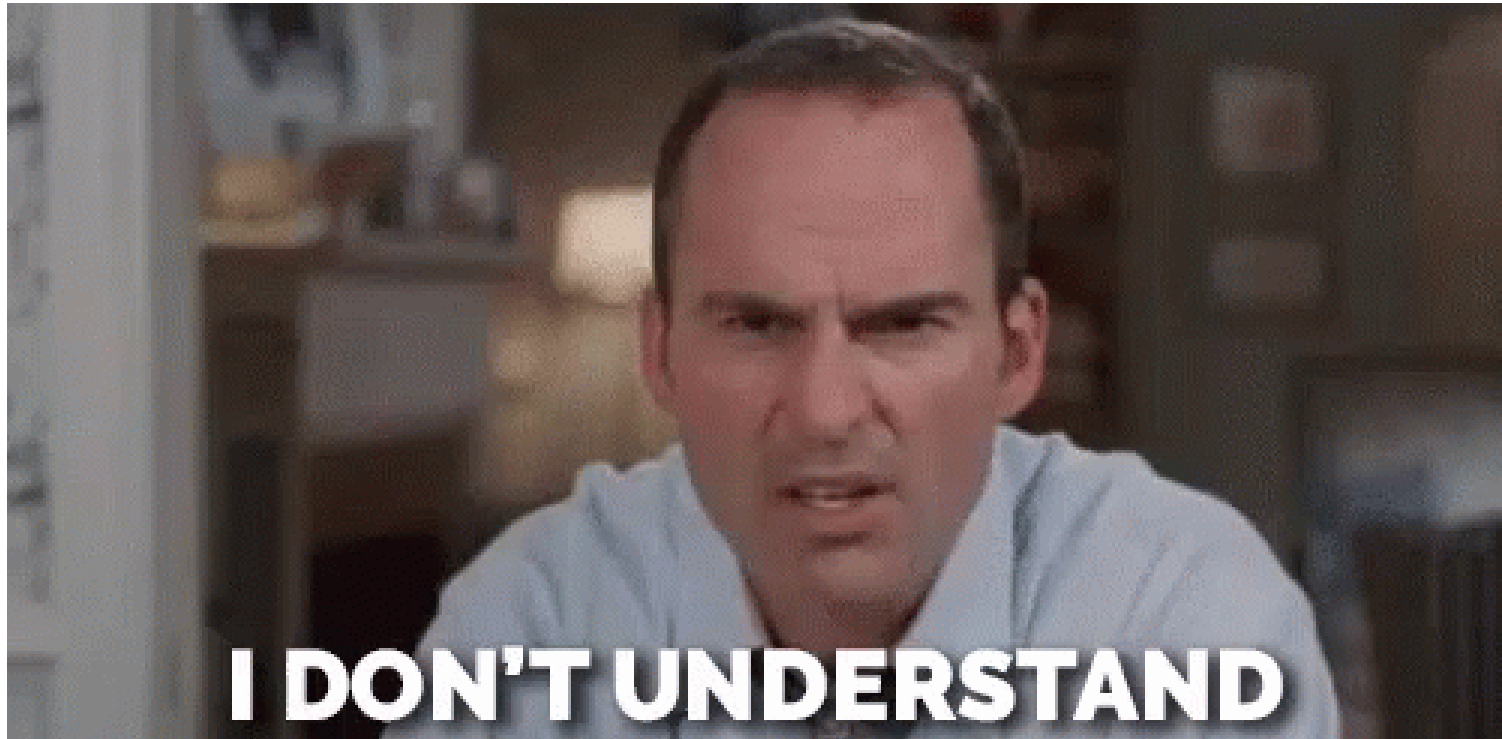
<sup>CH</sup>Coefficient is significantly different from Chile's at p < .05;

<sup>CO</sup>Coefficient is significantly different from Colombia's at p < .05;

<sup>M</sup>Coefficient is significantly different from Mexico's at p < .05;

<sup>V</sup>Coefficient is significantly different from Venezuela's at p < .05.

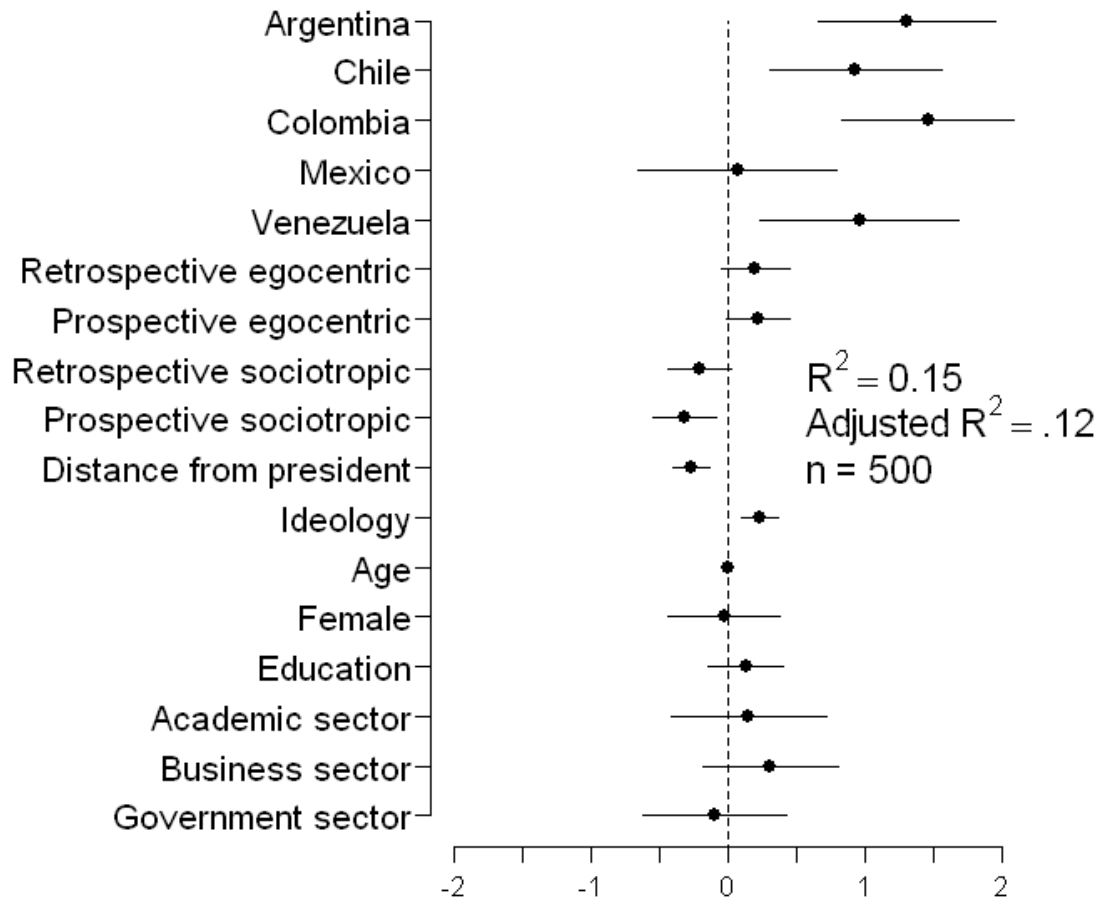
# What's wrong with this picture?





# Sunlight

`coefplot(model)`



Why didn't they say this in the first place?

NB: This is a presentation graph equivalent of the table

Shows coefficient with 95% CI

# Run, don't walk toward the sunlight





# Graphs can give enlightenment



*The greatest value of a picture is when it forces us to notice what we never expected to see.*

-- John W. Tukey

# Dangers of numbers-only output

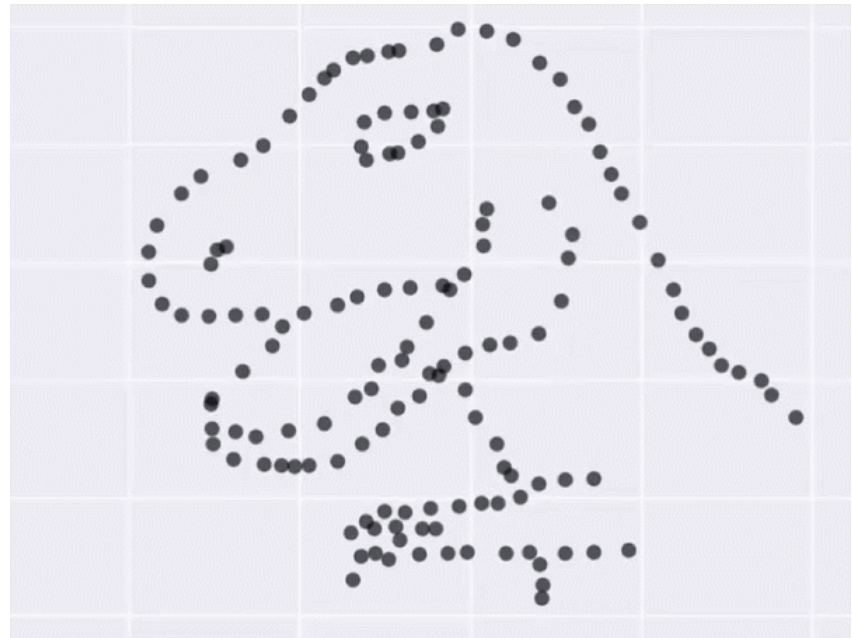
*Student:* You said to run descriptives and compute the correlation. What next?

```
X Mean: 54.26  
Y Mean: 47.83  
X SD   : 16.76  
Y SD   : 26.93  
Corr.  : -0.06
```

*Consultant:* Did you plot your data?

With **exactly** the same stats, the data could be *any* of these plots

See how this is done in R: <https://cran.r-project.org/web/packages/datasauRus/>



# Sometimes, don't need numbers at all

COVID transmission risk  $\sim$  Occupancy \* Ventilation \* Activity \* Mask? \* Contact.time

A complex 5-way table,  
whose message is clearly  
shown w/o numbers

There are 1+ unusual cells  
here. Can you see them?

Type and level of group activity	Low occupancy			High occupancy		
	Outdoors and well ventilated	Indoors and well ventilated	Poorly ventilated	Outdoors and well ventilated	Indoors and well ventilated	Poorly ventilated
<b>Wearing face coverings, contact for short time</b>						
Silent	Low	Low	Low	Low	Low	Medium
Speaking	Low	Low	Low	Low	Low	Medium
Shouting, singing	Low	Low	Medium	Medium	Medium	High
<b>Wearing face coverings, contact for prolonged time</b>						
Silent	Low	Low	Medium	Low	Medium	High
Speaking	Low	* Low	Medium	* Medium	Medium	High
Shouting, singing	Low	Medium	High	Medium	High	High
<b>No face coverings, contact for short time</b>						
Silent	Low	Low	Medium	Medium	Medium	High
Speaking	Low	Medium	Medium	Medium	High	High
Shouting, singing	Medium	Medium	High	High	High	High
<b>No face coverings, contact for prolonged time</b>						
Silent	Low	Medium	High	Medium	High	High
Speaking	Medium	Medium	High	High	High	High
Shouting, singing	Medium	High	High	High	High	High

**Risk of transmission**  
 Low ■ Medium ■ High ■







\* Borderline case that is highly dependent on quantitative definitions of distancing, number of individuals, and time of exposure

From: N.R. Jones et-al (2020). Two metres or one: what is the evidence for physical distancing in covid-19? *BMJ* 2020;370:m3223, doi: <https://doi.org/10.1136/bmj.m3223>

# If you do need tables– make them pretty

Several R packages make it easier to construct tables

Flipper lengths (mm) of the famous penguins of Palmer Station, Antarctica.

Species	Distribution	Female		Male	
		Avg.	Std. Dev.	Avg.	Std. Dev.
		188	5.6	192	6.6
		192	5.8	200	6.0
		213	3.9	222	5.7


Artwork by @allison\_horst

# Linear models

- Model:

$$y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip} + \varepsilon_i$$

- Xs: quantitative predictors, factors, interactions, ...
- Assumptions:
  - **Linearity**: Predictors (possibly transformed) are linearly related to the outcome,  $y$ . [This just means linear in the **parameters**.]
  - **Specification**: No important predictors have been omitted; only important ones included. [This is often key & overlooked.]
  - The “holy trinity”:
    - **Independence**: the errors are uncorrelated
    - **Homogeneity of variance**:  $\text{Var}(\varepsilon_i) = \sigma^2 = \text{constant}$
    - **Normality**:  $\varepsilon_i$  have a normal distribution


$$\varepsilon_i \sim_{iid} N(0, \sigma^2)$$

# Plots for linear models

- Data plots:
  - plot response ( $y$ ) vs. predictors, with smooth summaries
  - scatterplot matrix --- all pairs
- Model (effect) plots
  - plot predicted response ( $\hat{y}$ ) vs. predictors, controlling for variables not shown.
- Diagnostic plots

# Occupational Prestige data

- Data on prestige of 102 occupations and
  - average education (years)
  - average income (\$)
  - % women
  - type (Blue Collar, Professional, White Collar)

```
> head(Prestige)
```

	education	income	women	prestige	census	type
gov.administrators	13.11	12351	11.16	68.8	1113	prof
general.managers	12.26	25879	4.02	69.1	1130	prof
accountants	12.77	9271	15.70	63.4	1171	prof
purchasing.officers	11.42	8865	9.11	56.8	1175	prof
chemists	14.62	8403	11.68	73.5	2111	prof
physicists	15.64	11030	5.13	77.6	2113	prof

# Informative scatterplots

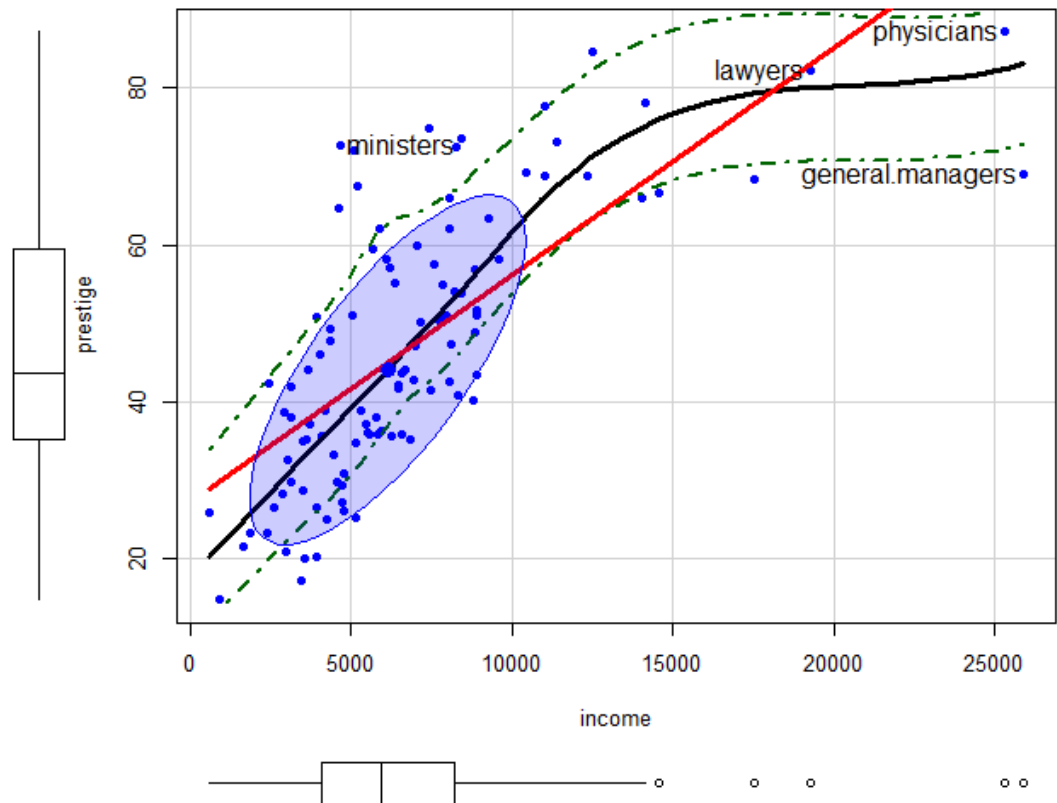
Scatterplots are most useful when enhanced with annotations & statistical summaries

Boxplots show marginal distributions

Data ellipse and regression line show the linear model,  $\text{prestige} \sim \text{income}$

Point labels show possible outliers

Smoothed (loess) curve and CI show the trend





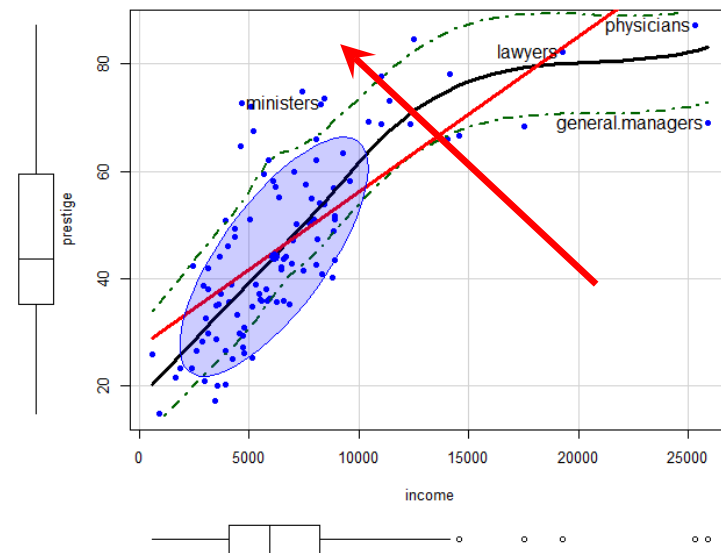
# Informative scatterplots

`car::scatterplot()` provides all of these enhancements

```
scatterplot(prestige ~ income, data=Prestige,  
            pch = 16,  
            regLine = list(col = "red", lwd=3),  
            smooth = list(smoother=loessLine,  
                           lty.smooth = 1, col.smooth = "black",  
                           lwd.smooth=3, col.var = "darkgreen"),  
            ellipse = list(levels = 0.68),  
            id = list(n=4, col="black", cex=1.2))
```

Skewed distribution of income & non-linear relation suggest need for a transformation

Arrow rule: move on the scale of powers in direction of the bulge

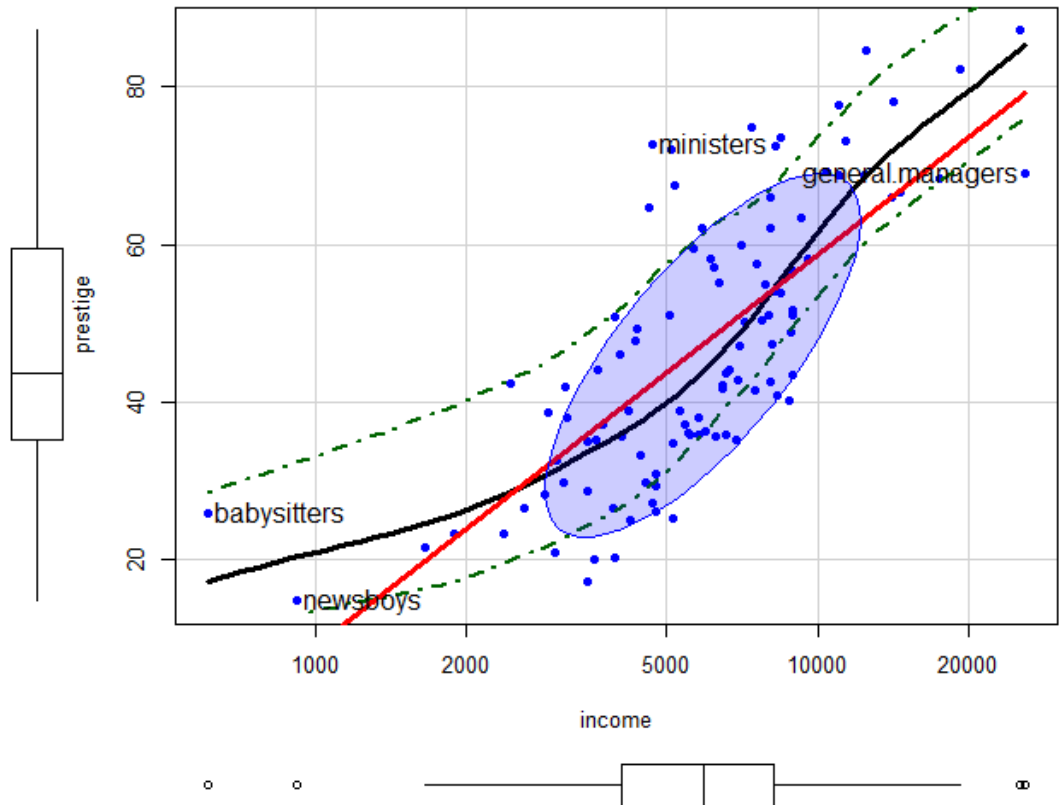


# Try log(income)

```
scatterplot
```

Income now ~ symmetric

Relation closer to linear



# Stratify by type?

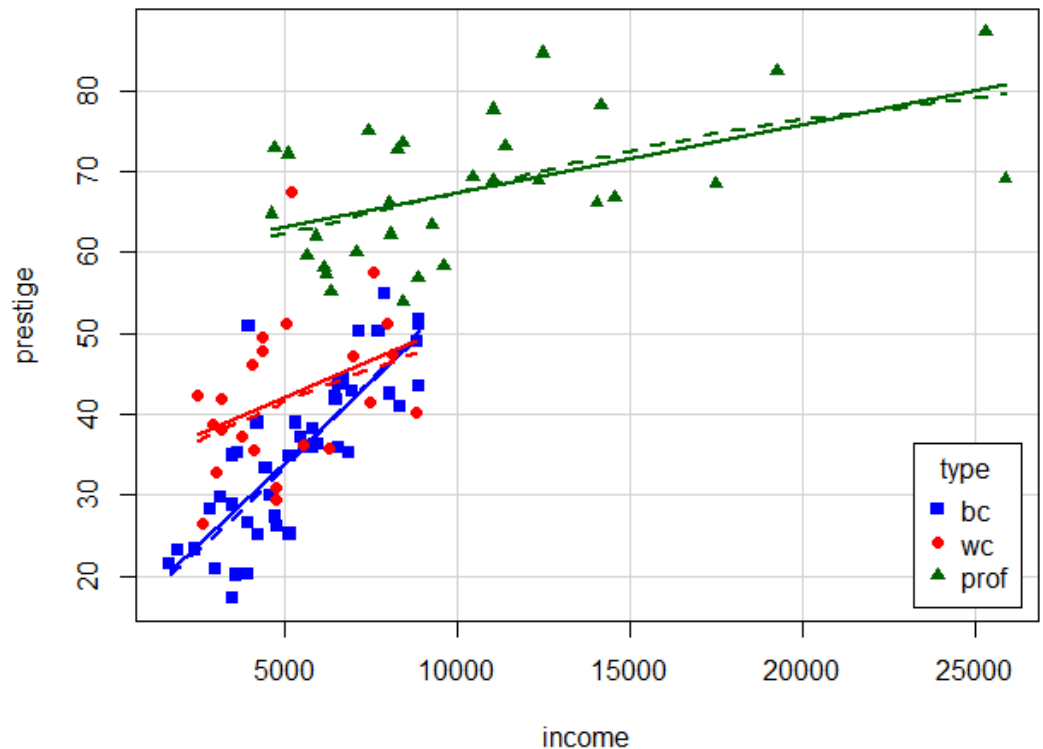
```
scatterplot(prestige ~ income | type, data=Prestige,  
  col = c("blue", "red", "darkgreen"),  
  pch = 15:17,  
  legend = list(coords="bottomright"),  
  smooth=list(smoother=loessLine, var=FALSE, span=1, lwd=4))
```

Formula: | **type** → “given type”

Different slopes: interaction of  
income \* type

Provides another explanation  
of the non-linear relation

This is a new finding!



# Scatterplot matrix

```
scatterplotMatrix(~ prestige + education + income + women ,  
  data=Prestige,  
  regLine = list(method=lm, lty=1, lwd=2, col="black"),  
  smooth=list(smoother=loessLine, spread=FALSE,  
    lty.smooth=1, lwd.smooth=3, col.smooth="red"),  
  ellipse=list(levels=0.68, fill.alpha=0.1))
```

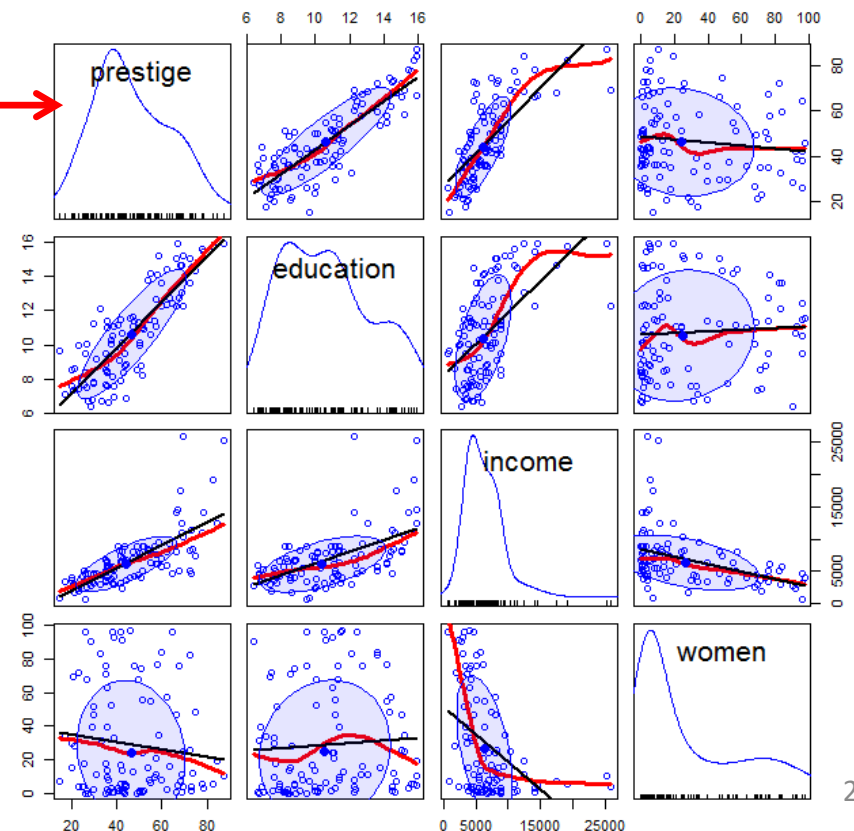
prestige vs. all predictors



diagonal: univariate distributions

- income: + skewed
- %women: bimodal

off-diagonal: relations among predictors



# Fit a model

```
> mod1 <- lm(prestige ~ education + poly(women, 2) +  
+           log(income)*type, data=Prestige)  
> summary(mod1)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )	
(Intercept)	-137.500	23.522	-5.85	8.2e-08	***
education	2.959	0.582	5.09	2.0e-06	***
poly(women, 2)1	28.339	10.190	2.78	0.0066	**
poly(women, 2)2	12.566	7.095	1.77	0.0800	.
log(income)	17.514	2.916	6.01	4.1e-08	***
typeprof	74.276	30.736	2.42	0.0177	*
typewc	0.969	39.495	0.02	0.9805	
log(income):typeprof	-7.698	3.451	-2.23	0.0282	*
log(income):typewc	-0.466	4.620	-0.10	0.9199	

---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Multiple R-squared: 0.879, Adjusted R-squared: 0.868  
F-statistic: 81.1 on 8 and 89 DF, p-value: <2e-16

- allow women<sup>2</sup> term
- interaction of log(income) and type

Fits very well!

# Model (effect) plots

- We'd like to see plots of the predicted value ( $\hat{y}$ ) of the response against predictors
  - But must control for other predictors not shown in a given plot
  - Variables not shown in a given plot are averaged over.
  - Slopes of lines reflect the partial coefficient in the model
  - Partial residuals can be shown also

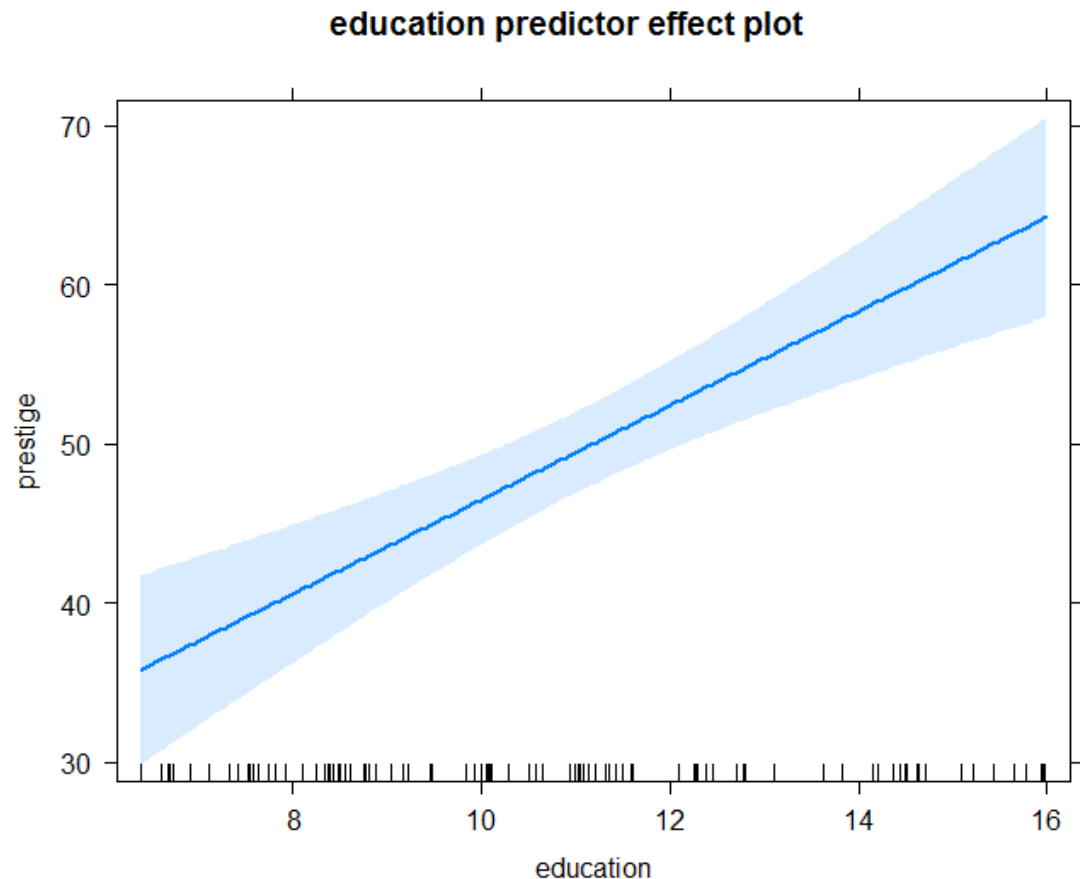
For details, see `vignette("predictor-effects-gallery", package="effects")`

# Model (effect) plots: education

```
library("effects")  
mod1.e1 <- predictorEffect("education", mod1)  
plot(mod1.e1)
```

This graph shows the partial slope for education.

For each  $\uparrow$  year in education, fitted prestige  $\uparrow$  2.96 points, (other predictors held fixed)

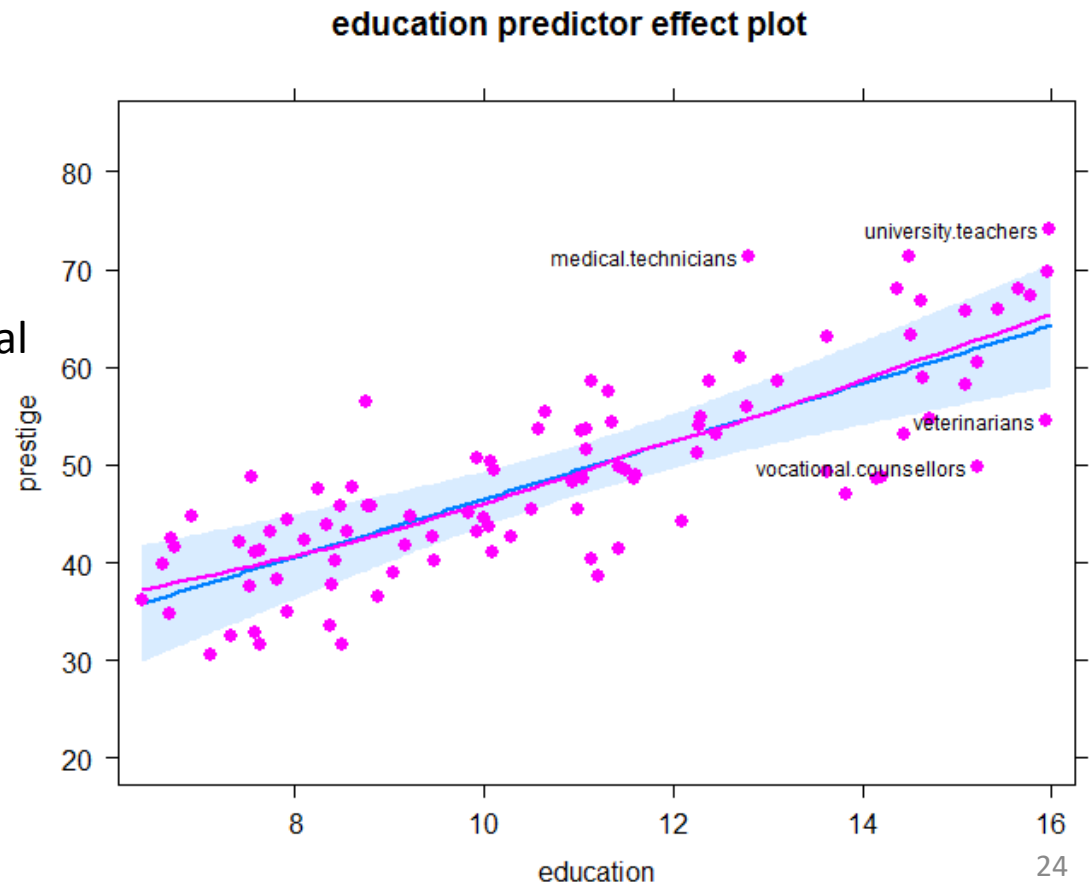


# Model (effect) plots

```
mod1.e1a <- predictorEffect("education", mod1, residuals=TRUE)
plot(mod1.e1a,
     residuals.pch=16, id=list(n=4, col="black"))
```

Partial residuals show the residual of prestige controlling for other predictors

Unusual points here would signal undue influence

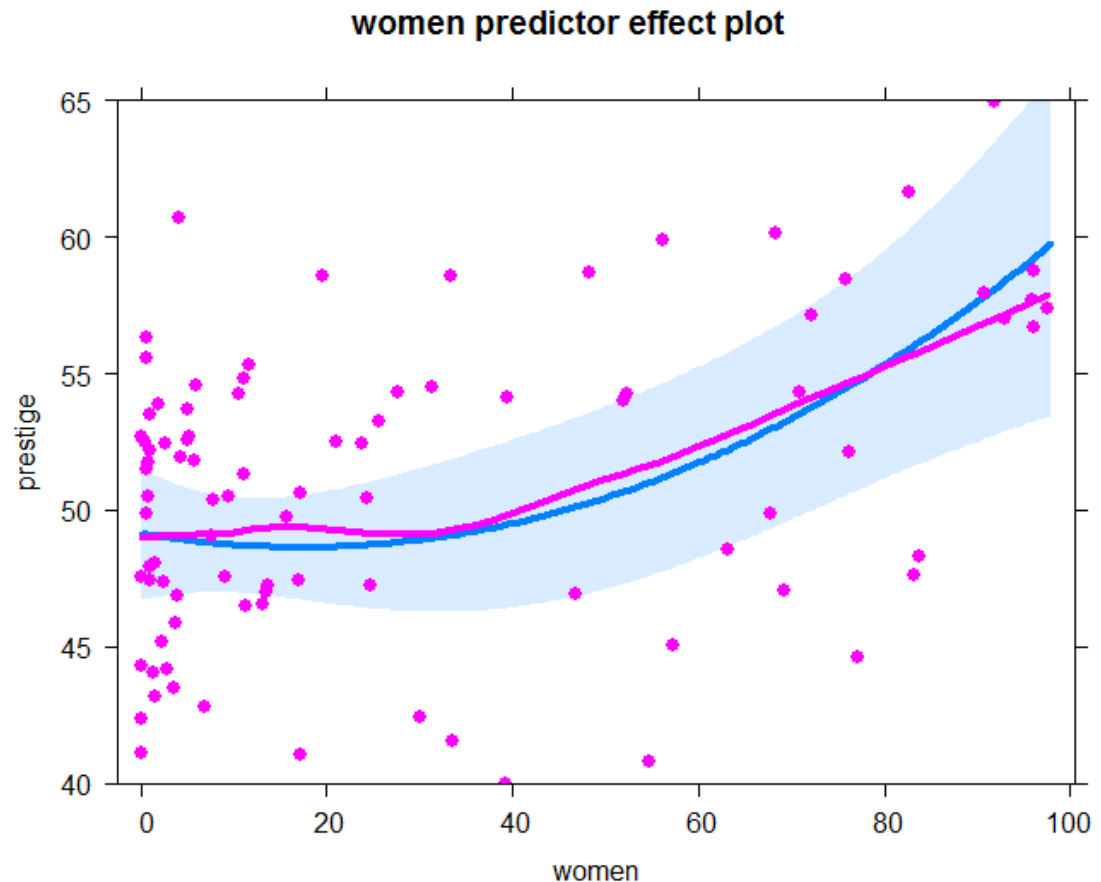




# Model (effect) plots: women

```
mod1.e2 <- predictorEffect("women", mod1, residuals=TRUE)  
plot(mod1.e2, ylim=c(40, 65), lwd=4,  
     residuals.pch=16)
```

Surprise!  
Prestige of occupations ↑  
with % women (controlling  
for other variables)

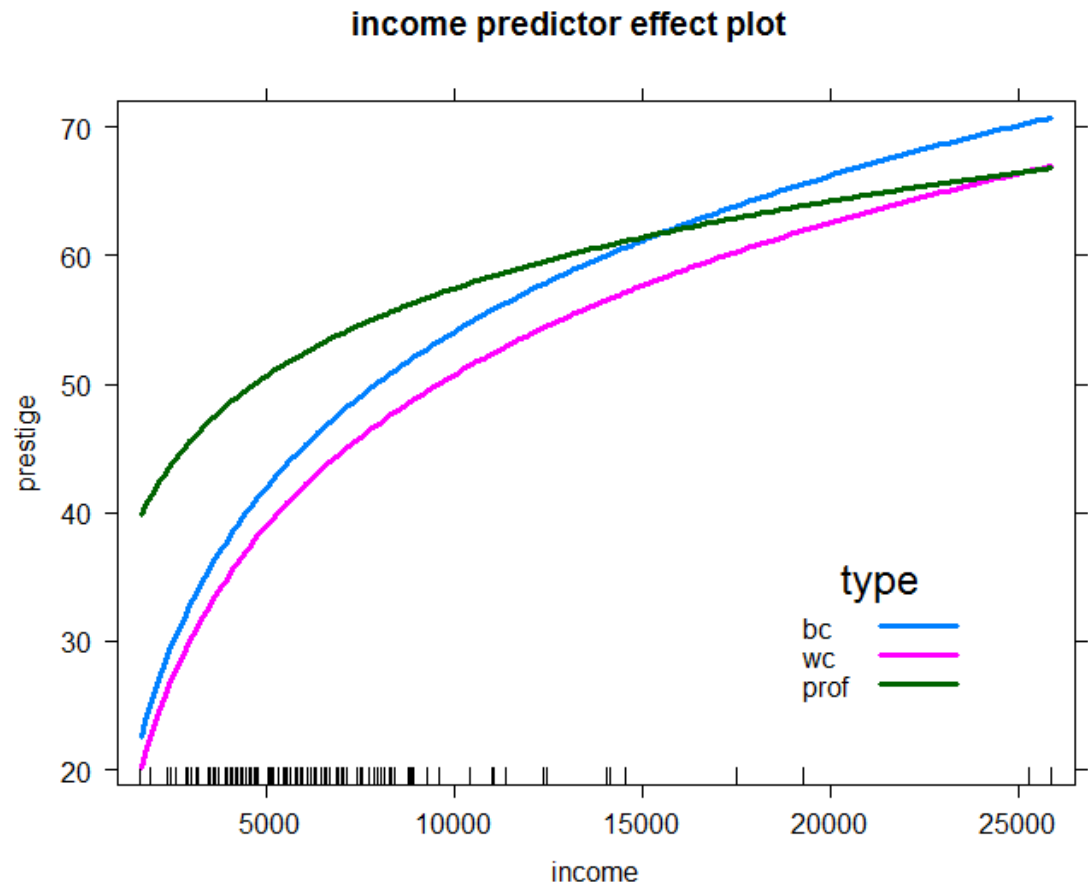


# Model (effect) plots: income

```
plot(predictorEffect("income", mod1),  
     lines=list(multiline=TRUE, lwd=3),  
     key.args = list(x=.7, y=.35))
```

Income interacts with type in the model

The plot is curved because  $\log(\text{income})$  is in the model



# Diagnostic plots

- The linear model,  $y = X\beta + \epsilon$  assumes:
  - Residuals,  $\epsilon_i$  are normally distributed,  $\epsilon_i \sim N(0, \sigma^2)$
  - (Normality not required for  $X$ s)
  - Constant variance,  $\text{Var}(\epsilon_i) = \sigma^2$
  - Observations  $y_i$  are statistically independent
- Violations  $\rightarrow$  inferences may not be valid
- A variety of plots can diagnose all these problems
- Other methods (boxCox, boxTidwell) diagnose the need for transformations of  $y$  or  $X$ s.

# The “regression quartet”

In R, plotting a `lm` model object → the “regression quartet” of plots

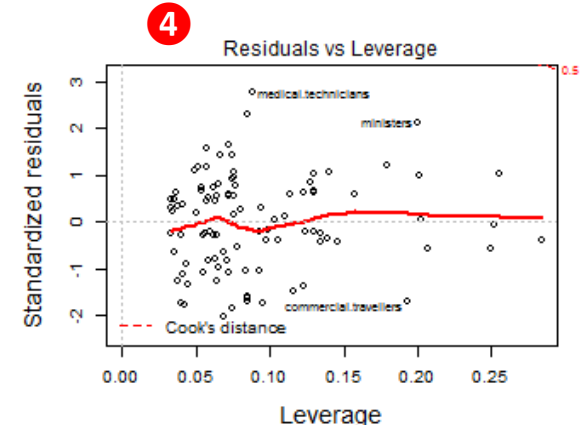
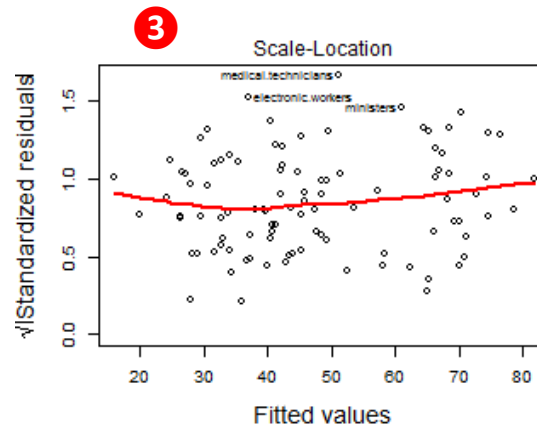
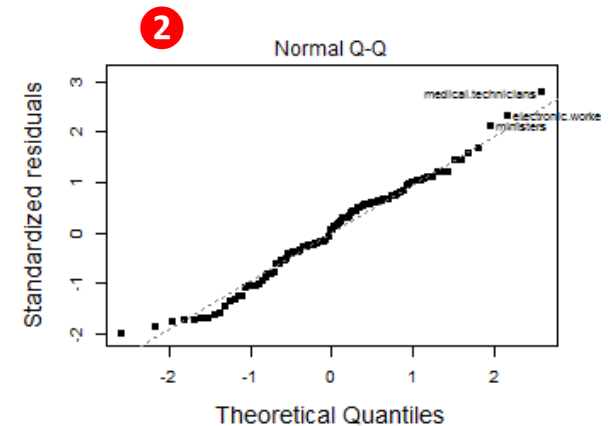
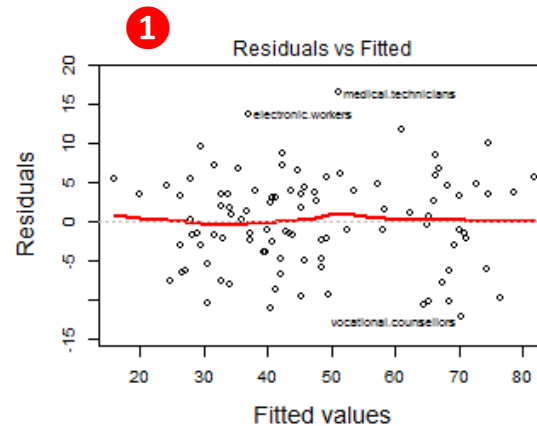
```
plot(mod1, lwd=2, cex.lab=1.4)
```

**1** Residuals: should be flat vs. fitted values

**2** Q-Q plot: should follow the 45° line

**3** Scale-location: should be flat if constant variance

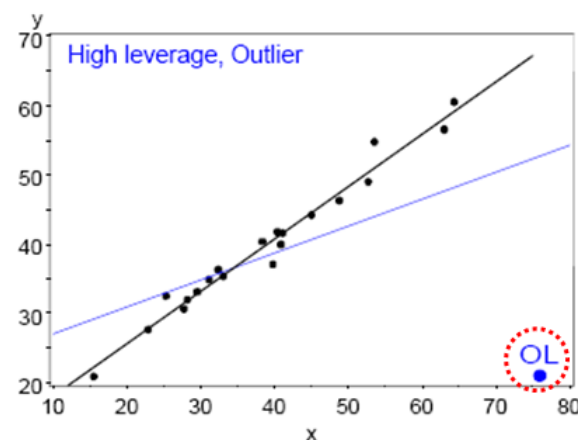
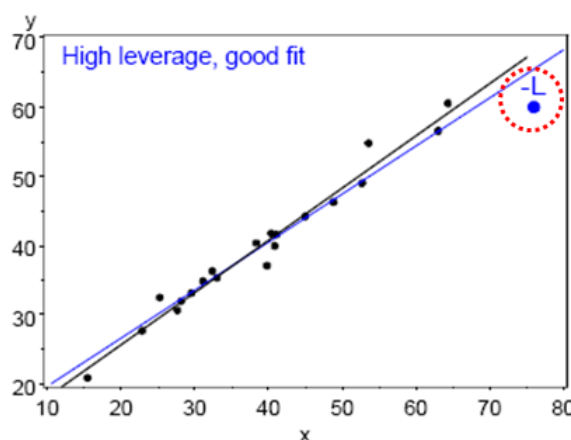
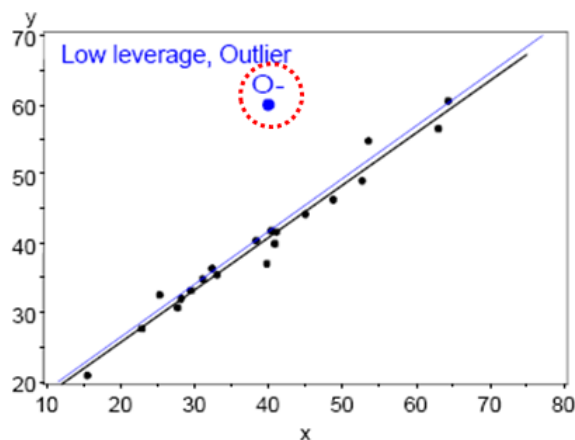
**4** Resids vs. leverage: can show influential observations



# Unusual data: Leverage & Influence

- “Unusual” observations can have dramatic effects on least-squares estimates in linear models
- Three archetypal cases:
  - Typical X (low leverage), bad fit -- Not much harm
  - Unusual X (high leverage), good fit -- Not much harm
  - Unusual X (high leverage), bad fit -- **BAD, BAD, BAD**
- Influential observations: unusual in *both* X & Y
- Heuristic formula:

$$\text{Influence} = X \text{ leverage} \times Y \text{ residual}$$



# Influence plots

Influence (Cook's D) measures impact of individual obs. on coefficients, fitted values

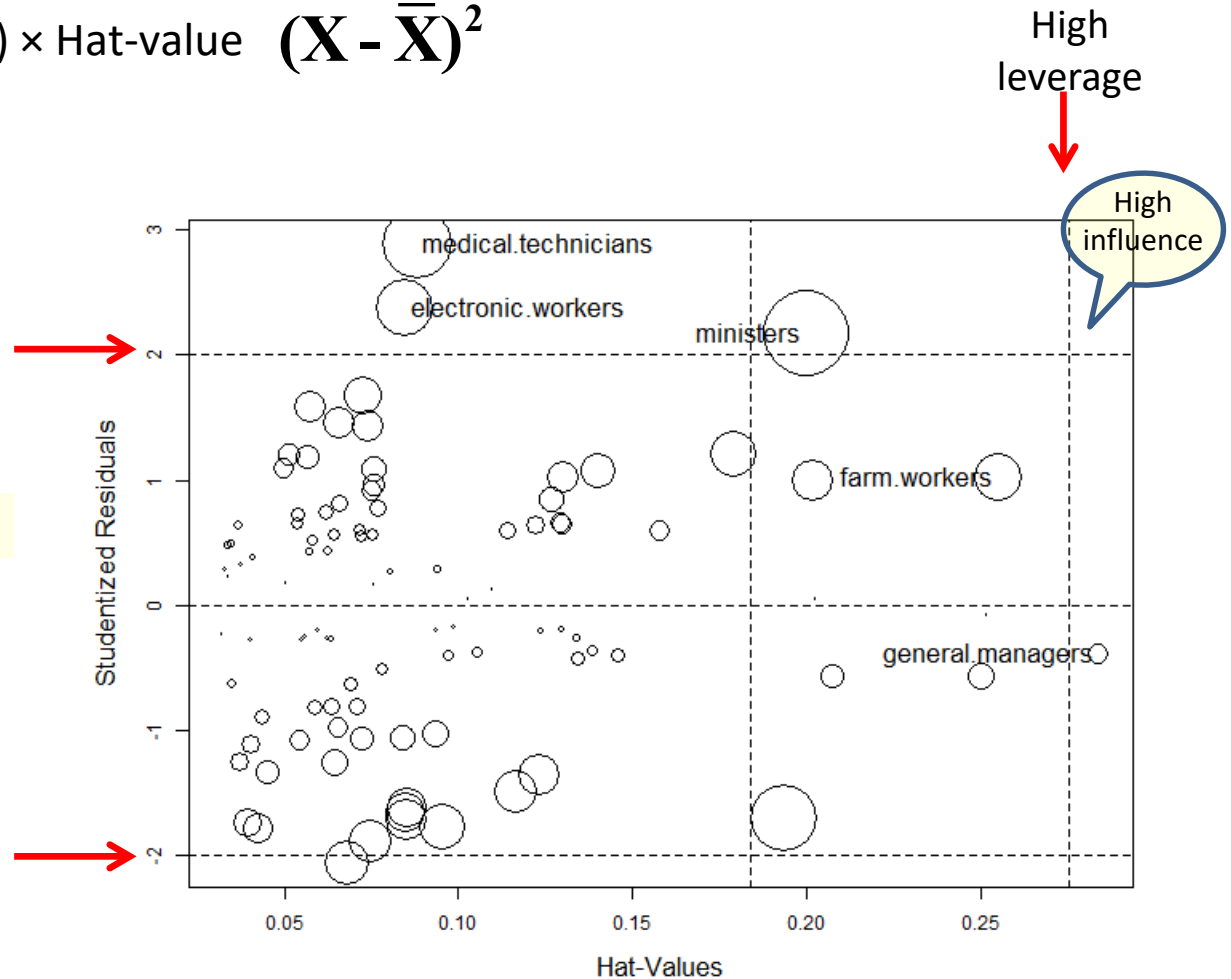
Influence  $\sim$  Residual  $(y - \hat{y}) \times \text{Hat-value}$   $(X - \bar{X})^2$

Bubble size  $\sim$  influence

`influencePlot(mod1)`

Bad fit

Bad fit

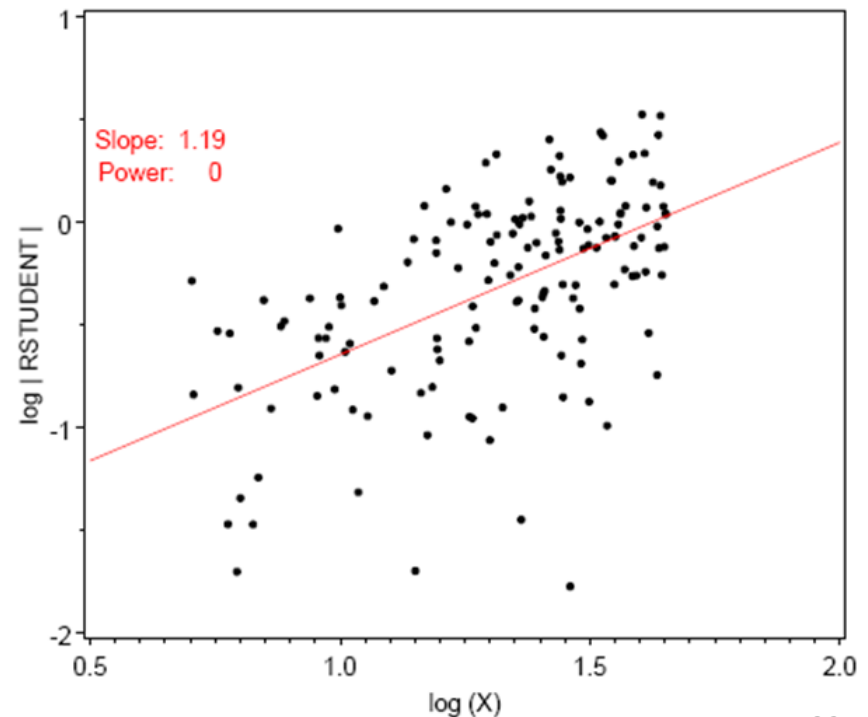


# Spread-level plots

- To diagnose non-constant variance, plot:
  - $\log |\text{Std. residual}|$  vs.  $\log(x)$
  - $\log(\text{IQR})$  vs  $\log(\text{median})$  [for grouped data]
- If  $\approx$  linear w/ slope  $b$ , transform  $y \rightarrow y^{(1-b)}$

Artificial data, generated so  $\sigma \sim x$

- $b \approx 1 \rightarrow \text{power} = 0$
- $\rightarrow$  analyze  $\log(y)$



# Spread-level plot: baseball data

Data on salary and batter performance from 1987 season

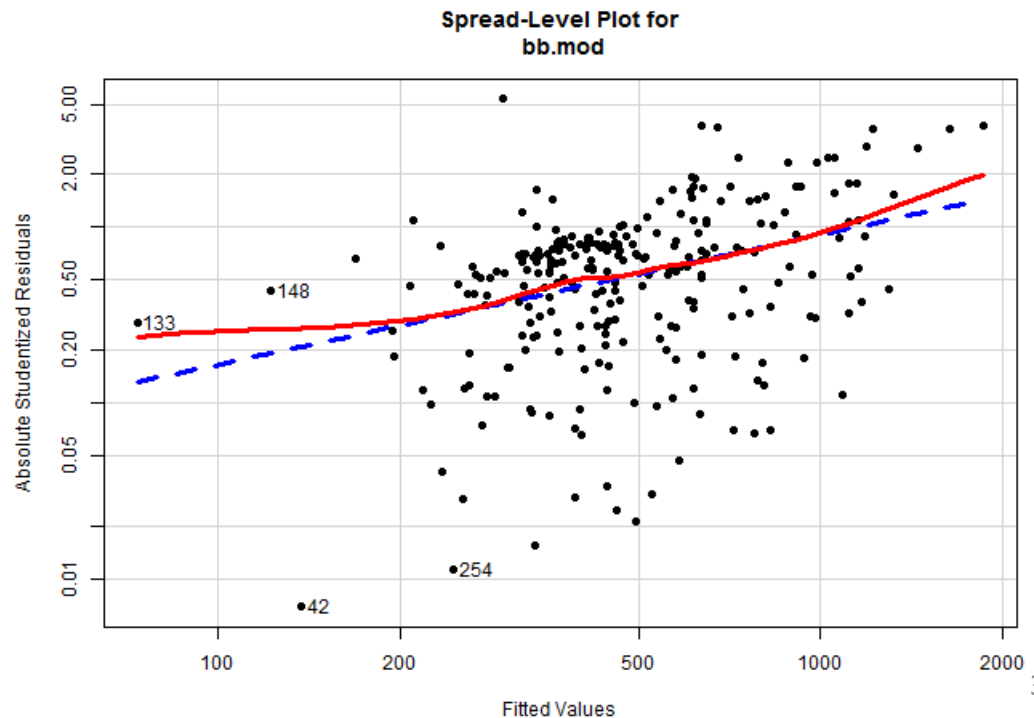
```
data("Baseball", package="vcd")
bb.mod <- lm(sal87 ~ years + hits + runs + homeruns, data=Baseball)
spreadLevelPlot(bb.mod, pch=16, lwd=3,
                id=list(n=2))
```

## Suggested power transformation: 0.2609

slope = .74  $\rightarrow$  p = .26

i.e.,  $y \rightarrow \log(y)$  or  $y^{1/4}$

NB: both axes plotted on log scale





# Summary

- Tables are for look-up; graphs can give insight
- Data plots are more effective when enhanced
  - data ellipses → strength & precision of correlation
  - regression lines and smoothed curves
  - point identification → noteworthy observations
- Effect plots show informative views of models
- Diagnostic plots can reveal influential observations and need for transformations.