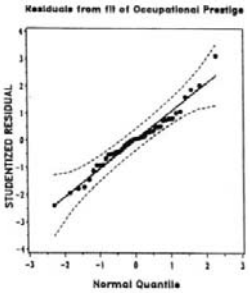
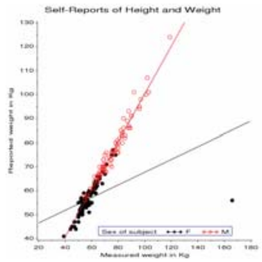
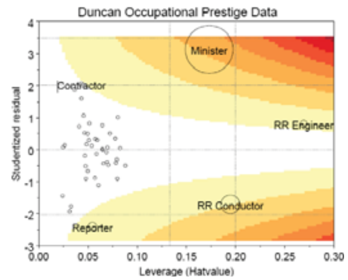






Regression diagnostics

(I thought I was done when I fit the model)

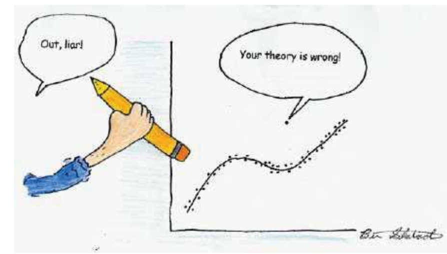
Psychology 6140

Topics

- Assumptions of the linear regression model
- Patterns in residual plots
- Assessing normality of residuals
- Diagnosing non-constant variance
- Unusual data: Leverage & Influence
- Partial plots

Sometimes a few “bad” points can ruin a theory (Duncan data)

Sometimes, they can help suggest a better one (Fuel data)



What is a regression model?

- A model is a merely a representation / description of reality.
- A regression model specifies how a quantitative variable (*Y*) is related to other variables (*X*s), with certain assumptions.

- But reality is often too complicated to be perfectly represented / described.
  - All models are wrong – or simply partial descriptions.
  - That’s OK: we don’t need perfect models – just adequate ones.
  - But, we do need to make sure that inferences / conclusions are correct!
- We should try to formulate models that closely represent reality.
  1. Fit the model
  2. Check assumptions
  3. If necessary, modify model, go back to 1.

Linear regression model

- Model:

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

- 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”: $\varepsilon_i \sim_{iid} \mathcal{N}(0, \sigma^2)$
    - **Independence**: the errors are uncorrelated
    - **Homogeneity of variance**: Var(ε_i) = σ^2 = constant
    - **Normality**: ε_i have a normal distribution

Model fitting & model criticism

- Bad news:
  - Any statistical model we fit is probably wrong (or incomplete).
  - Hope for a decent summary & valid inference.
- Good news:
  - Info about the "explained" portion → fitted values
  - Info about the "unexplained" portion → residuals
- Residual plots help to guide us:
  - Model assumptions: NQQ plots, spread vs. level plots
  - Model specifications: partial residual plots (Y, X_i | other Xs)
- Other problems:
  - Outliers, leverage → Influence plots

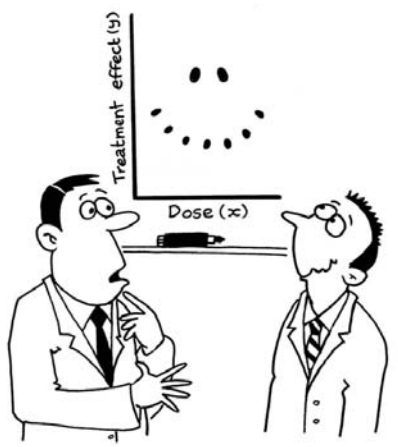
Why look at residuals?

- Our model claims that values of *Y* are the result of two components:

$$y_i = \overbrace{\beta_0 + \beta_1 x_i}^{\substack{X\text{'s contribution} \\ \text{(regression line)}}} + \overbrace{\varepsilon_i}^{\substack{\text{error or individual} \\ \text{differences}}}$$

  - The model does not say that nothing else is related to Y.
  - Only that-- anything else is random, not systematic
  - The remaining part – the residual -- is considered as random error or individual differences.
- Since we think that there is nothing systematically related to *Y* beyond *X*,
  - if there are any other variables available to us, we should explore the relationship between such variables and *e*.

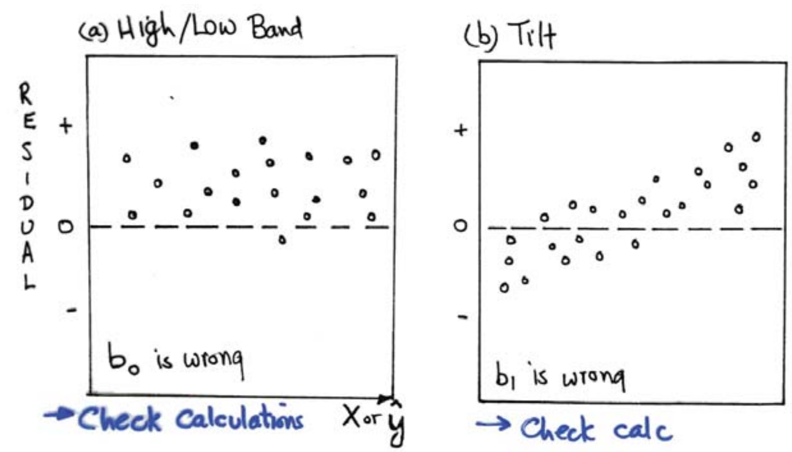
Patterns in residual plots



"It's a non-linear pattern with outliers.....but for some reason I'm very happy with the data."

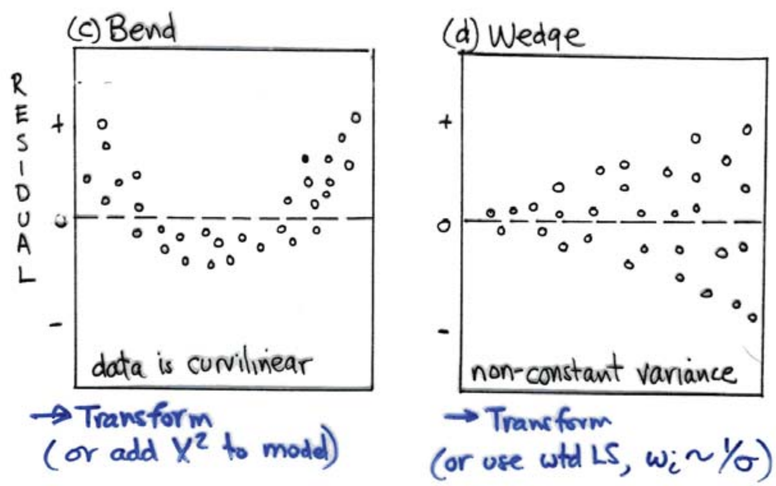
- Residual plots show what has not yet been accounted for in a model
- As such, they offer an opportunity to learn something more.
- Sometimes, we can truly be happy, learning something not shown in model summaries.
- Need to know what to look for

Patterns in residual plots



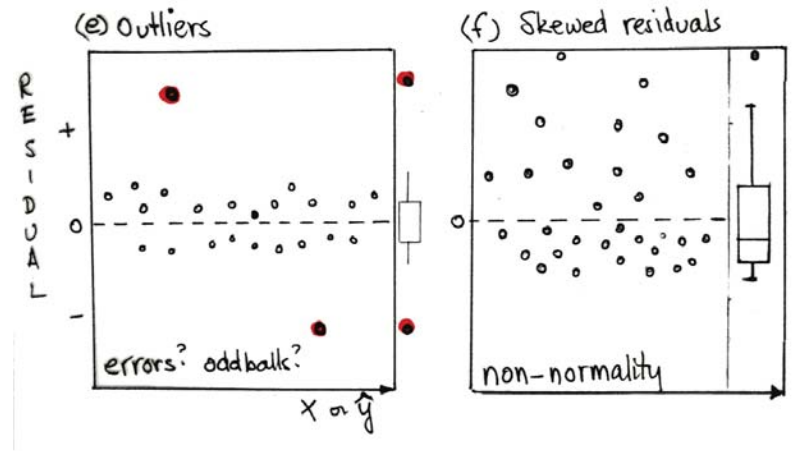
Patterns like these rarely occur, except in hand calculations or excel

Patterns in residual plots



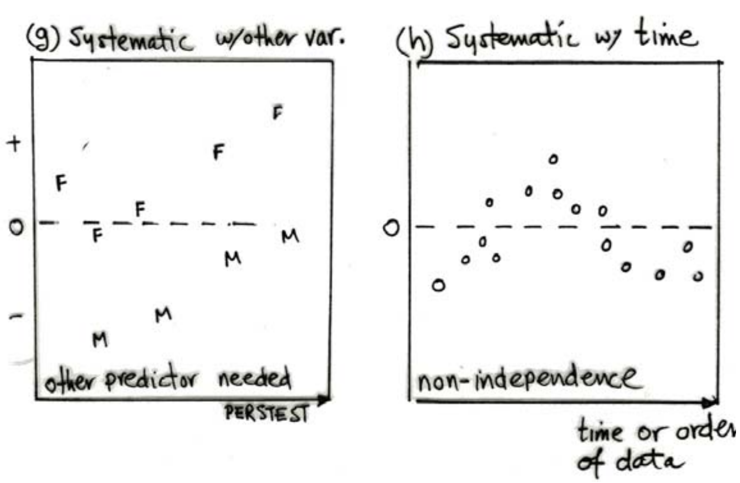
More common, but usually less pronounced than these cartoons

Patterns in residual plots



More subtle patterns, often better revealed by other plots

Patterns in residual plots



More subtle still: need to think about what not yet included

Running example: Duncan data

- Duncan (1961) studied how well one could predict occupational prestige (hard to measure) from available census measures
  - Income: proportion of males in an occupation with income > \$3500 in 1950 census
  - Educ: proportion of males with >= high school
  - Prestige: % of people rating an occupation as “good” or “excellent” (survey of 3000 people)
- Issue: relative effects of Income & Educ --- are they equally important as determinants of occ. Prestige?

Statistical results:

- Model fits well (R^2 = 0.83)
- Income & educ both significant (& approx. equal!)
- What's not to like?

Analysis of Variance

Source	*DF*	*Sum of Squares*	*Mean Square*	*F Value*	*Pr > F*
Model	2	36181	18090	101.22	**<.0001**
Error	42	7506.70	178.73		
Corrected Total	44	43688			

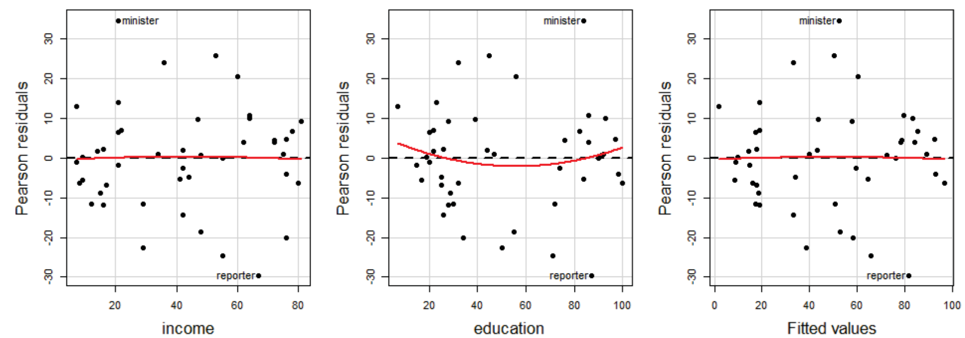
Root MSE	13.369	*R-Square*	**0.8282**
Dependent Mean	47.689	*Adj R-Sq*	0.8200

Parameter Estimates

Variable	*Label*	*DF*	*Parameter Estimate*	*Standard Error*	*t Value*	*Pr > |t|*
Intercept	Intercept	1	-6.065	4.27194	-1.42	0.1631
income	Income	1	**0.599**	0.11967	5.00	**<.0001**
educ	Education	1	**0.546**	0.09825	5.56	**<.0001**


```
library(car)
data("Duncan", package="car")
duncan.mod <- lm(prestige ~ income + education, data=Duncan)
# basic residual plots
residualPlots(duncan.mod, layout=c(1,3), id.n=2)
```

All residual plots look OK (~flat). But two points have large residuals



Smoothing is often essential to see the overall trend, particularly with small N
Automatic labeling of unusual points is helpful too.

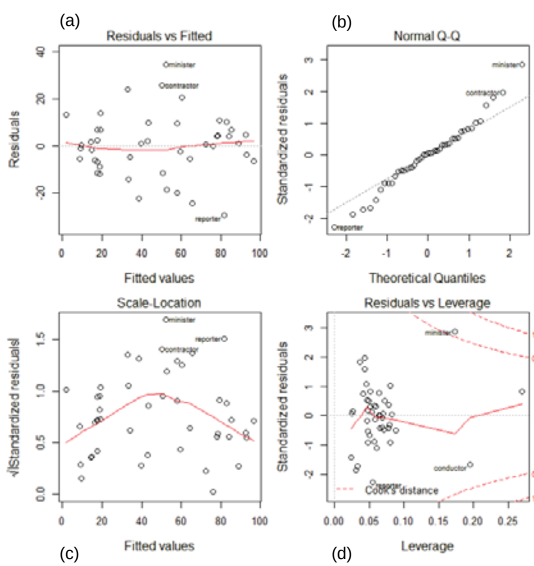
In R, always good to "plot the model" first → the regression quartet

```
plot(duncan.mod)
```

These help to diagnose:
(a) Systematic residuals?
(b) Normality?
(c) Heterogeneous variance?
(d) Influential observations?

I'll take up the details of each next.

There are better versions of these plots in other packages (car), but this should be a first step



Assessing normality of residuals

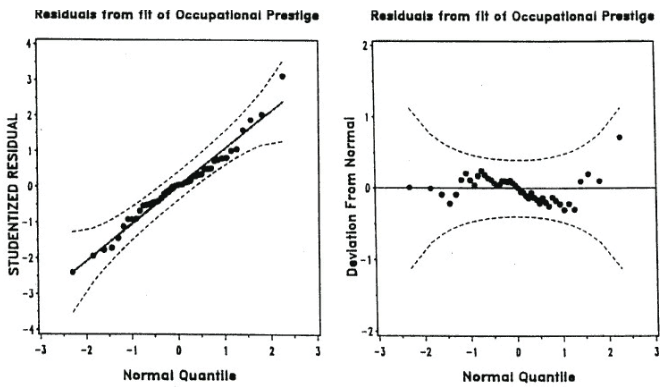
- The linear model does not require **y** to be normally distributed– only the errors, **ε**
- Neither are quantitative **X**s required to be normal
  - but highly skewed Xs may cause other problems– non-linearity
- Practical impact of violation of normality of **ε** :
  - Univariate tests of normality (e.g., K-S test) are highly sensitive to small departures; don't need exact normality
  - small effect on *p*-values, unless highly non-normal
  - High kurtosis – long tails (outliers) more a threat than skewness
- → Graphical method (NQQ plot) sufficient in practice

Assessing normality: NQQ plots

- **standard NQQ plot**: plot sorted residuals, $e_{[i]}$ vs. z_i = quantiles in a N(0,1) distribution– should follow a 45º line
- **Better**: show confidence envelope for assessing departures
- **Better yet**: detrended version plots ($e_{[i]} - z_i$) vs. z_i – should follow a flat line

SAS: nqplot macro

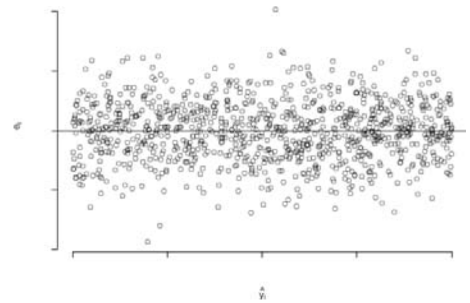
R: car::qqPlot

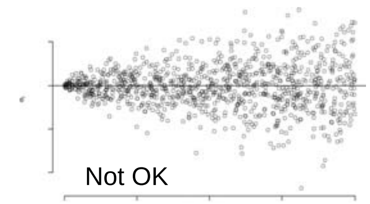


Diagnosing non-constant variance

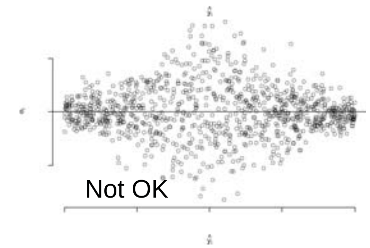
- Usual method: plot residuals vs. fitted values: look for differences in residual variance

OK





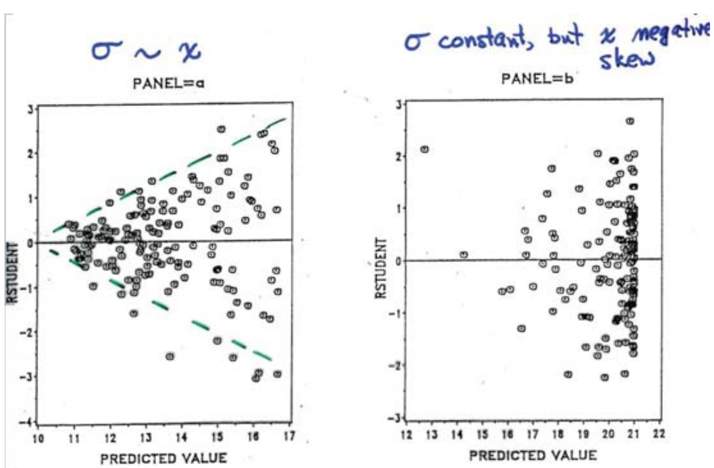
Not OK



Not OK

Diagnosing non-constant variance

- This doesn't always work, e.g., if the distribution of predicted values is highly skewed, the plot can be misleading due to number of observations.



Diagnosing non-constant variance

- Better: plot absolute value, $|e_i|$ vs. predicted, w/ smoothed curve to show variation

The smooth should be flat

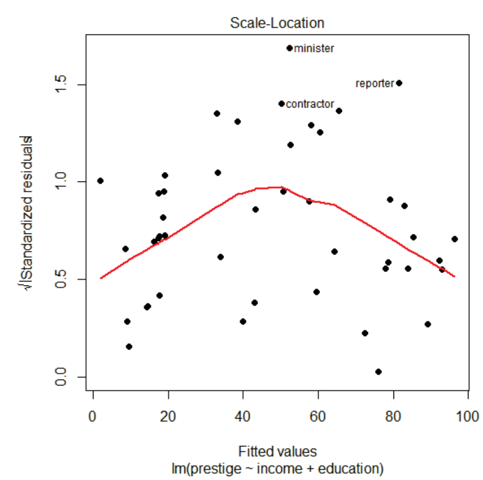
Here, occupational prestige is a proportion, and

$$\text{var}(p) = \sqrt{p(1-p)/n}$$

Is maximal at p=0.5

This suggests a transformation:

$$p \rightarrow \left\langle \begin{matrix} \log p / 1-p & \text{(logit)} \\ \sin^{-1}\sqrt{p} & \text{(arcsin)} \end{matrix} \right.$$



Correcting non-constant variance

- As always, two options:
  - Transform **y** to make σ^2 ~ constant
  - Fit a more general model that allows σ^2 to vary with E(**y**|**x**)– a *generalized* linear model (E.g.,: logistic regression, poisson regression)
- For now, the transformation route is easier– stays within the classical linear model
- A spread-level plot gives an easy way to find a power transformation, if spread varies with level

Spread-Level plots: theory

Commonly used variance stabilizing techniques

Relationship of σ^2 to $E[y]$	Transformation	comment
$\sigma^2 \propto$ constant	$y' = y$	no transformation
$\sigma^2 \propto E[y]$	$y' = \sqrt{y}$	Poisson data
$\sigma^2 \propto E[y](1 - E[y])$	$y' = sin^{-1}(\sqrt{y})$	binomial proportions,
$\sigma^2 \propto (E[y])^2$	$y' = \log(y)$	$y > 0$
$\sigma^2 \propto (E[y])^3$	$y' = y^{-1/2}$	$y > 0$
$\sigma^2 \propto (E[y])^4$	$y' = y^{-1}$	

$$\sigma \propto E[y]^b \Rightarrow \log(\sigma) \propto b \cdot \log(E[y])$$

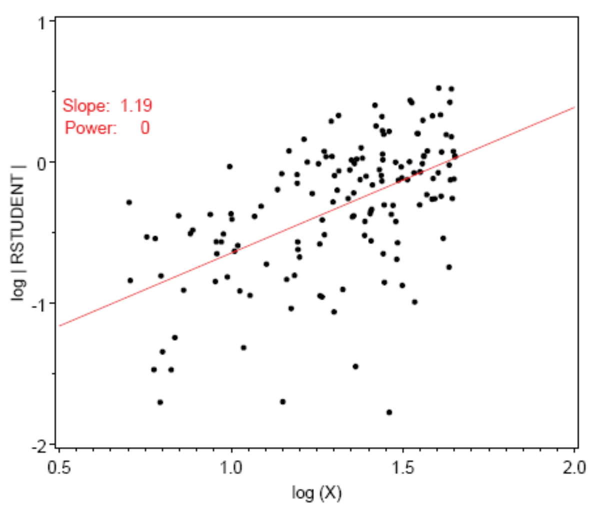
- These suggest: transform y with the power p = 1-b
- Thus, plot log(spread) vs. log(level) & use 1-slope as the power
- (Works *if* the plot is reasonably linear)
- (Proportions require something different– folded power transformations)

Spread-Level plots

- Spread vs. level plots: Plot $\log(|e_i|/\hat{\sigma})$ vs. $\log(x)$
- If linear, with slope b, transform $y \rightarrow y^p$, with $p = 1 - b$.

SAS: spredplot macro

R:
car::spreadLevelPlot()



- Artificial data, generated so that $\sigma \sim x$: Power = 0 $\rightarrow$ analyze $\log(y)$

This method doesn't work for the Duncan data, because log (spread) is not linearly related to log (fitted value)

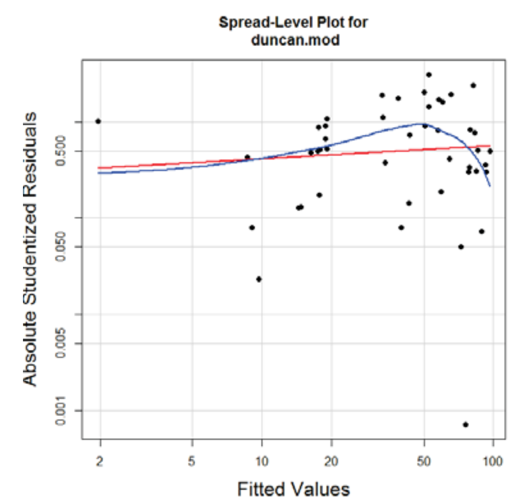
```
spreadLevelPlot(duncan.mod)
```

```
Suggested power transformation:  0.865
```

i.e., no power transform helps

The loess smoothed curve shows that residual variance is not constant, but a power transformation can't cure this.

As suggested earlier, a better analysis would have used a logit or arcsine transform of prestige to stabilize variance



NB: both are plotted on log scales

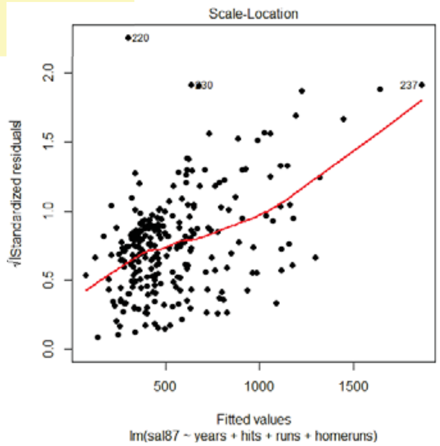
Baseball data: scale-location plot

```
data("Baseball", package="vcd")
bb.mod <- lm(sal87 ~ years + hits + runs + homeruns,
          data=Baseball)
# standard plot method
plot(bb.mod, which=3, pch=16, lwd=2)
```

This plot shows that residual variance increases with fitted value

It doesn't diagnose a corrective power transformation, but:

Increasing → $p < 1$ (down scale of powers)



car::spreadLevelPLot

Spread-level plot for a model object:

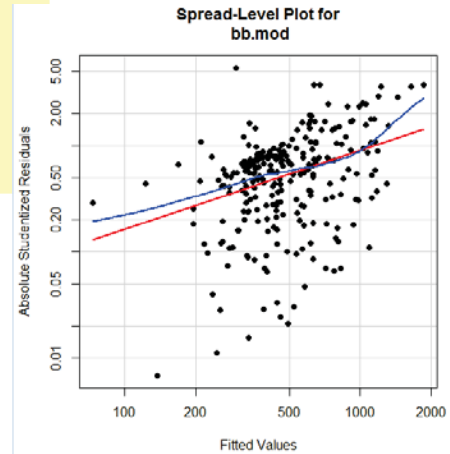
```
data("Baseball", package="vcd")
bb.mod <- lm(sal87 ~ years + hits + runs + homeruns,
          data=Baseball)
library(car)
spreadLevelPlot(bb.mod, pch=16)

# show smooth fit
fit <- fitted(bb.mod)
res <- abs(rstudent(bb.mod))
lines(loess.smooth(fit, res), col="blue", lwd=2)
```

This gives:

```
Suggested power transformation:
0.261
```

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

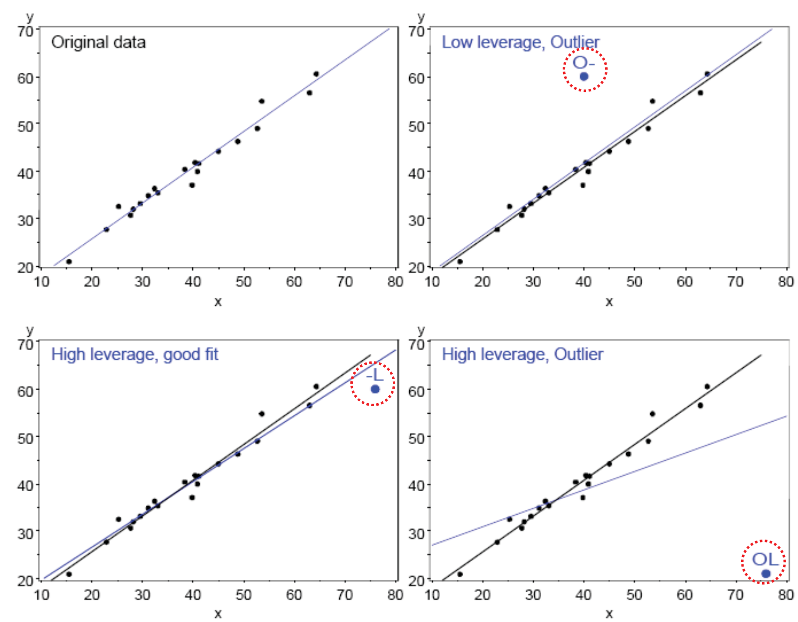


Unusual data: Leverage & Influence

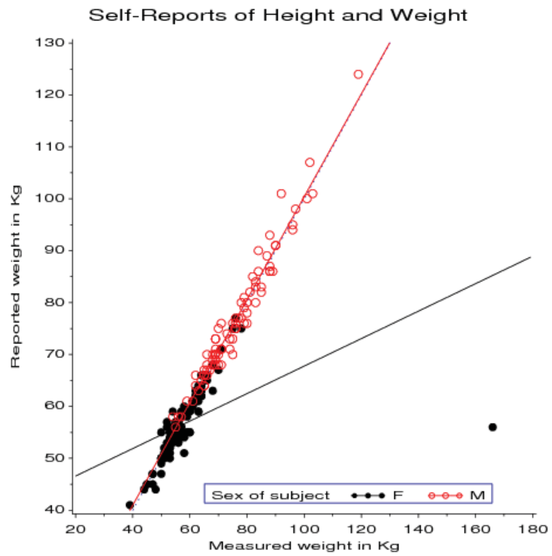
- "Unusual" observations can have dramatic effects on least-squares estimates in linear models
- Three archtypical cases:
  - Typical X (low leverage), bad fit -- Not much harm
  - Unusual X (high leverage), good fit -- Not much harm
  - Unusual X (high leverage), bad fit -- BAD, BAD, BAD
- Influential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

  Influence = X leverage x Y residual

Effect of adding one more point (new line in blue):



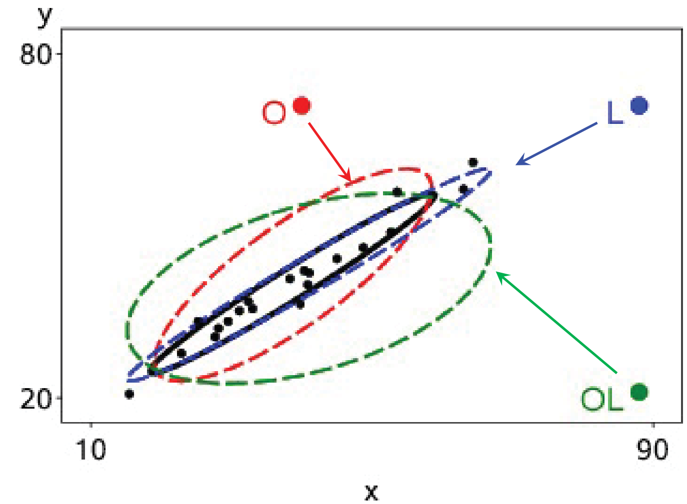
Dramatic example: Davis' data on reported and measured weight of men and women



This one bad point is both of high leverage (X) and has a huge residual (Y)

Unusual points also affect precision of estimates:

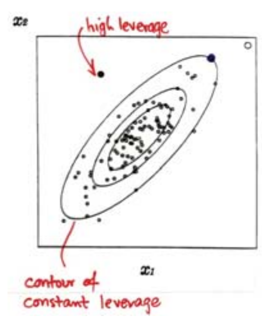
- **OL**: biases slope & increases std. error
- **O**: no bias, but increases std. error
- **L**: decreases std. error ("good leverage" point)



Measuring leverage

- **Leverage:** measured by "Hat values," h_i.
  - so-called because fitted values can be expressed as $\widehat{y} = \boldsymbol{H} y$
  - For simple linear regression, $h_i \sim (x - \bar{x})^2$
  - For p predictors, $h_i \propto$ squared distance of x_i from centroid, $\bar{x}$ (Mahalanobis squared distance)
  - All hat values range from $1/n$ to 1, and average is $\bar{h} = (p+1)/n$.
  - $\rightarrow$ observations with $h_i > 2\bar{h}$ (or $h_i > 3\bar{h}$ in small samples) are typically considered "high leverage" points

- In general, leverage is ~ Mahalanobis squared distance for the predictors from their means



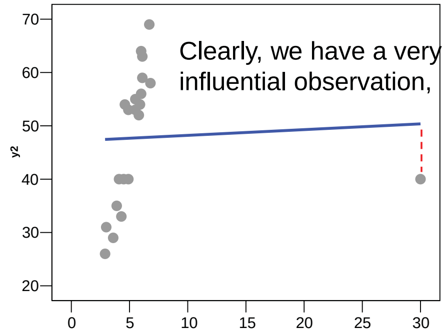
Detecting outliers: Studentized residuals

- **Ordinary residuals:** $e_i = y_i - \hat{y}_i$, not useful because:
  - Even if errors, ϵ_i have constant variance (as assumed), residuals *do not*—variance of e_i varies inversely with leverage— Var$(e_i) = \sigma^2(1 - h_i)$
  - Outliers on Y pull the regression line (surface) toward them
- **Studentized residuals:**
  - Standardized residual (RSTUDENT) calculated for y_i *deleting* observation i. Using subscript $(-i)$ to mean deleting i,

$$\text{RSTUDENT} \equiv e_i^\star = \frac{e_i}{s_{(-i)}\sqrt{1 - h_i}}$$

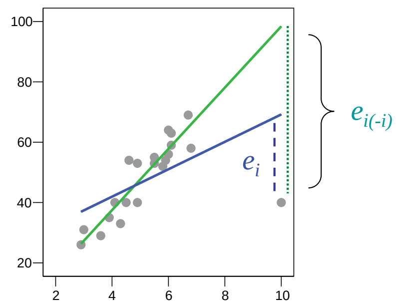
  - Gives a test for "mean-shift" outlier model, $H_0 : \mathcal{E}(y_i \,|\, X) \neq \mathcal{E}(y_{(-i)} \,|\, X)$
  - $e_i^\star \sim t(n - p - 2)$
    - $\rightarrow |e_i^\star| > t_{1-\alpha/2}(n - p - 2)$ signifcant *a priori*
    - $\rightarrow |e_i^\star| > t_{1-\alpha/2n}(n - p - 2)$ signifcant *a posteriori* (Bonferroni)

Ordinary residuals



yet residual is not big at all, because LS makes it small!

Studentized residuals



Studentized residual ignores that point

Influence = Leverage x Residual

- **Cook's D:** Scale-invariant (*squared*) measure of "distance" between β (all) and $\beta_{(-i)}$ (deleting obs. i)

$$\text{COOKD}_i \equiv D_i = \left(\frac{e_i^2}{(p+1)s^2} \right) \times \frac{h_i}{1-h_i^2}$$

  - "Large" values: $D_i > 4/n$ [or $D_i > 4/(n-p-1)$]

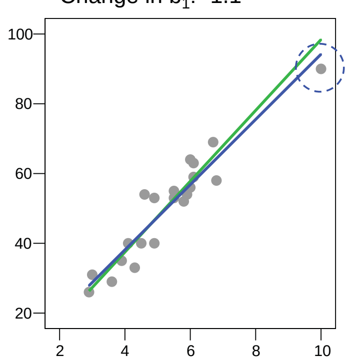
- **DFFITS:** Scaled measure of (*signed*) change in predicted value for y_i, deleting obs. i

$$\text{DFFITS}_i = \frac{\hat{y}_i - \hat{y}_{(-i)}}{s_{(-i)}\sqrt{h_i}} = \left(\frac{e_i}{s_{(-i)}} \right) \times \frac{\sqrt{h_i}}{1-h_i^2}$$

  - "Large" values: $|\text{DFFITS}_i| > 2\sqrt{(p+1)/n}$

Example: Consider the circled observation

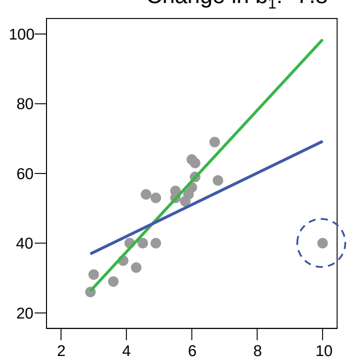
Change in b_0: 1.0
Change in b_1: -1.1



Change in $\hat{y}$: -1.2

Low influence – a "good" leverage point

Change in b_0: 6.8
Change in b_1: -7.8



Change in $\hat{y}$: -8.2

High influence – a "bad" leverage point

Influence diagnostics with SAS

- `PROC REG` (ODS GRAPHICS → regression diagnostic plots)
  - `influence` option on `model` statement gives printed values
- `inflplot` **macro**
  - Fits model using `PROC REG`, influence statistics → output dataset
  - Plots RSTUDENT vs. Hat value, bubble size ∼ Cook's D or DFFITS
  - Labels "noteworthy" observations— large RSTUDENT and/or Hat value
  - Shows nominal cutoffs for "unusual" values
- **Similar macros**
  - `inflogis` macro— logistic regression (`PROC LOGISTIC`)
  - `inflglim` macro— generalized linear models (`PROC GENMOD`)

See: `http://www.math.yorku.ca/SCS/sssg/inflplot.html`

Example: Duncan's Occupational prestige data

PROC REG step, with influence option

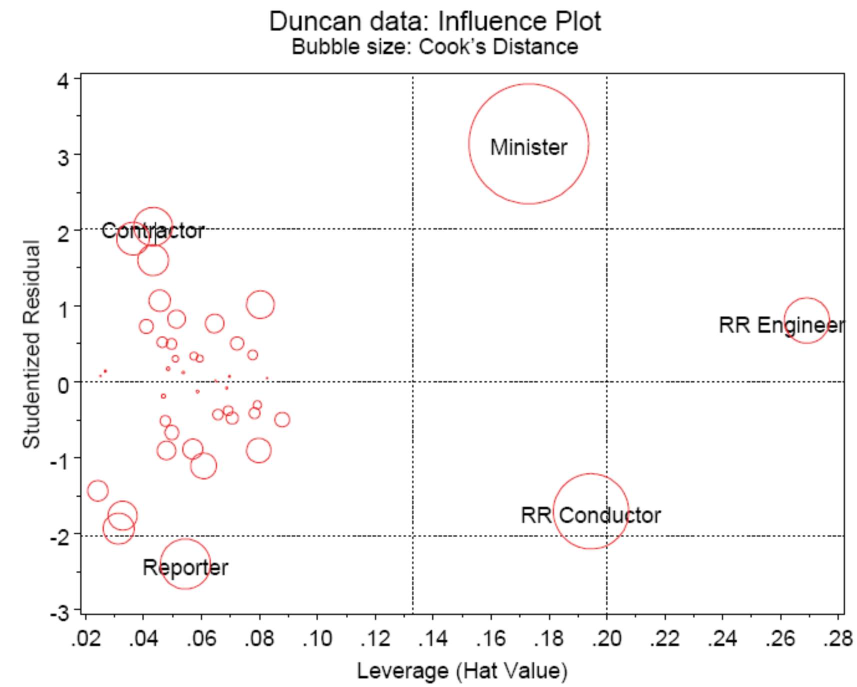
duncinfl2.sas ···

```
%include data(duncan);
proc reg data=duncan;
   model prestige = Income Educ / influence;
   id job;
   run;
```

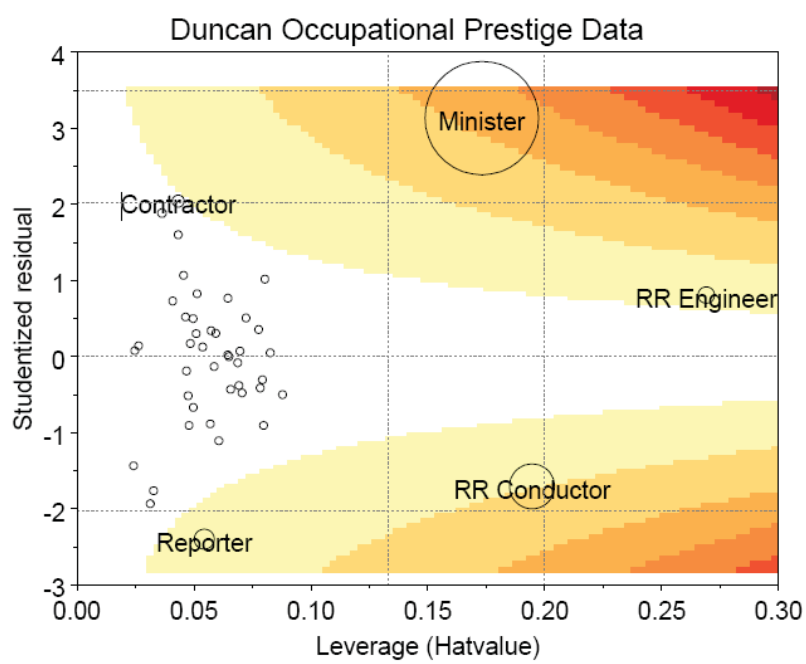
inflplot macro:

··· duncinfl2.sas

```
title 'Duncan data: Influence Plot';
title2 "Bubble size: Cook's Distance";
%inflplot(data=duncan,
   y=Prestige,       /* response        */
   x=Income Educ,    /* predictors      */
   id=job,           /* ID variable     */
   bubble=cookd      /* bubble ~ Cook's D */
);
```

Showing contours of Cook's D:



Example: Duncan's Occupational prestige data

Influence on coefficients is substantial:

- All $n = 45$ cases

Parameter Estimates

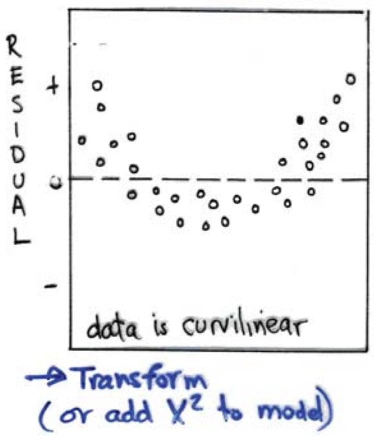
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	-6.06466	4.27194	-1.42	0.1631
income	Income	1	0.59873	0.11967	5.00	<.0001
educ	Education	1	0.54583	0.09825	5.56	<.0001

- Deleting Minister, RR Conductor, RR Engineer

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	-6.31736	3.67962	-1.72	0.0939
income	Income	1	0.93066	0.15375	6.05	<.0001
educ	Education	1	0.28464	0.12136	2.35	0.0242

Patterns in residual plots: Marginal vs. partial relations



- For a one predictor model, this plot is helpful.
- But with two+ predictors, such plots only show the marginal relationships (ignoring other Xs)
- The multiple regression model is about partial relationships – controlling for other Xs
- → We need to see the partial relation between Y and X_i, holding other Xs constant.

Partial regression plots

- **Problems**
  - Correlated predictors— Ordinary scatterplots cannot show the *unique* effects of one predictor, *controlling* for others
  - Joint influence— Single deletion diagnostics cannot show whether sets of observations are *jointly influential*, or *offset* each other
- **Solution: Partial regression (added-variable) plots**
  - For x_k, plot y | other xs vs. x_k | other xs. (others $\equiv \boldsymbol{X}[-k]$)

$$\begin{aligned} y \mid \text{others} \equiv y_k^\star &= y - \hat{y}_{\boldsymbol{X}[-k]} \\ x_k \mid \text{others} \equiv x_k^\star &= x - \hat{x}_{\boldsymbol{X}[-k]} \end{aligned}$$

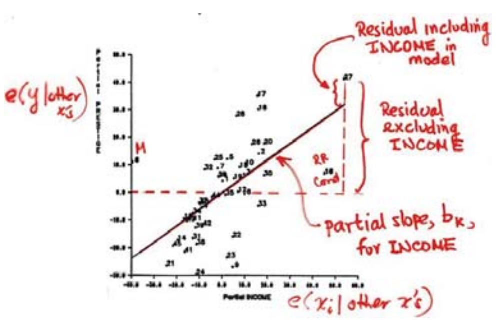
  - $y_k^\star$ = residuals from regression of y on $\boldsymbol{X}[-k]$
  - $x_k^\star$ = residuals from regression of x_k on $\boldsymbol{X}[-k]$
  - $\rightarrow$ unique relation of y to x_k, controlling/adjusting for all other xs.

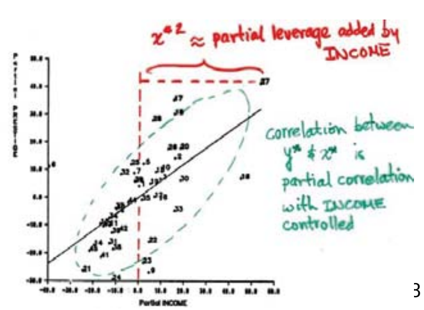
Partial regression plots: Properties

- slope of $y_k^\star$ on $x_k^\star = b_k$, the estimate of the (partial) regression coefficient, β_k, in the full model.
- residuals from the regression line in this plot $\equiv$ residuals for y in the full model, i.e.,

$$y_k^\star = b_k x_k^\star + e$$

- simple correlation between $y_k^\star$ and $x_k^\star$ = partial correlation between y and x_k with the other x variables partialled out or controlled.
- plot shows *partial* leverage ($\sim {x_{ik}^\star}^2$) and influence





Partial regression plots: Example

PROC REG step, with `partial` option $\longrightarrow$ printer plots

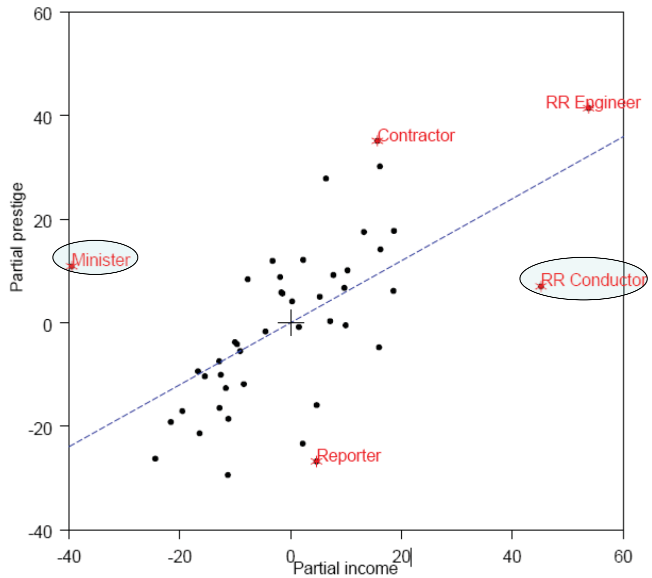
duncan4.sas ···

```
%include data(duncan);
proc reg data=duncan;
   model prestige = Income Educ / partial;
   id job;
   run;
```

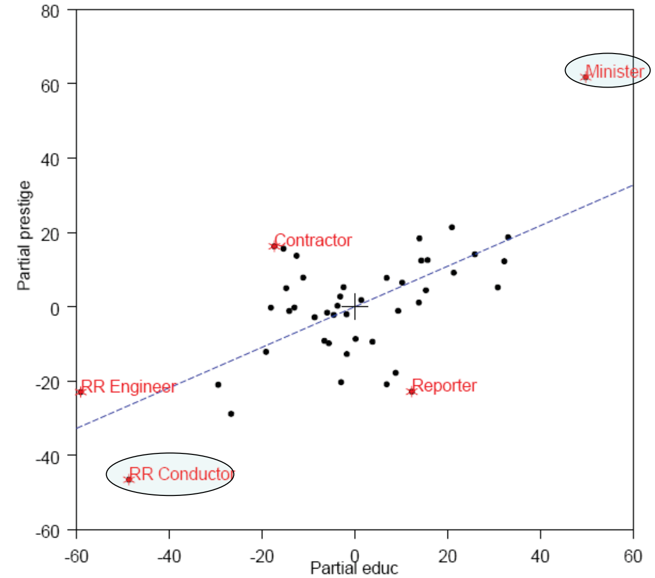
`partial` macro: high-res plots

··· duncan4.sas

```
%partial(data=duncan,
   yvar=Prestige,        /* response          */
   xvar=Income Educ,     /* predictors        */
   id=job,               /* ID variable       */
   label=INFL            /* label influential pts */
);
```



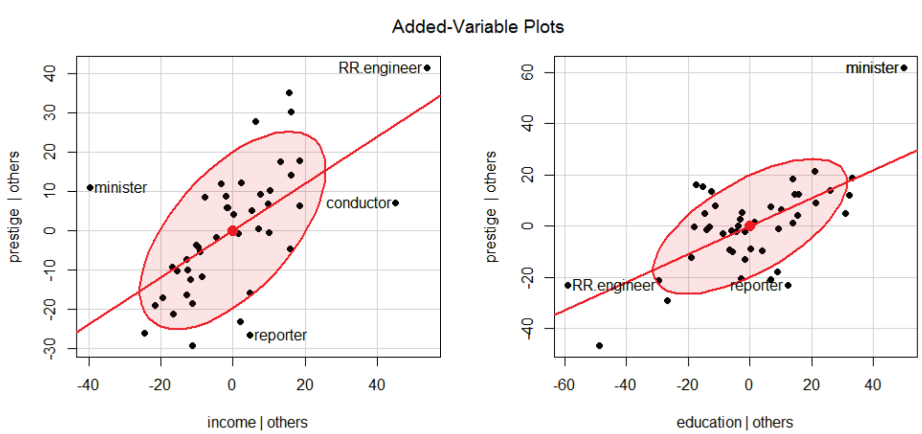
- Minister and RR Conductor are *jointly influential*— decrease slope for Income



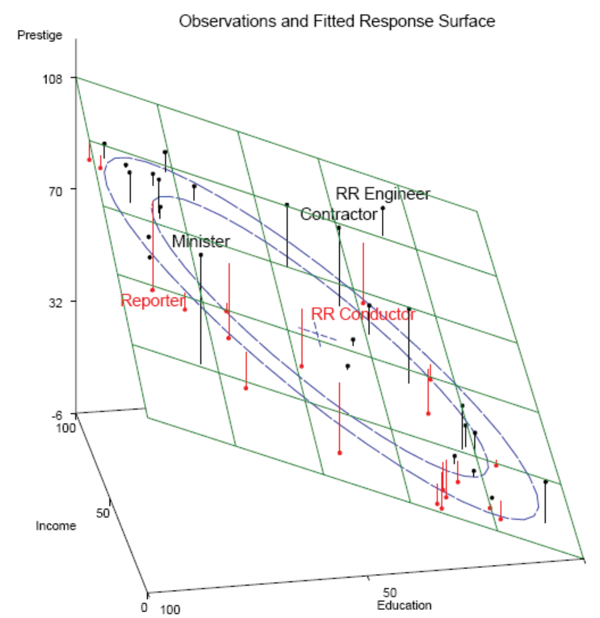
- Minister and RR Conductor are *jointly influential*— increase slope for Education

car::avPlots

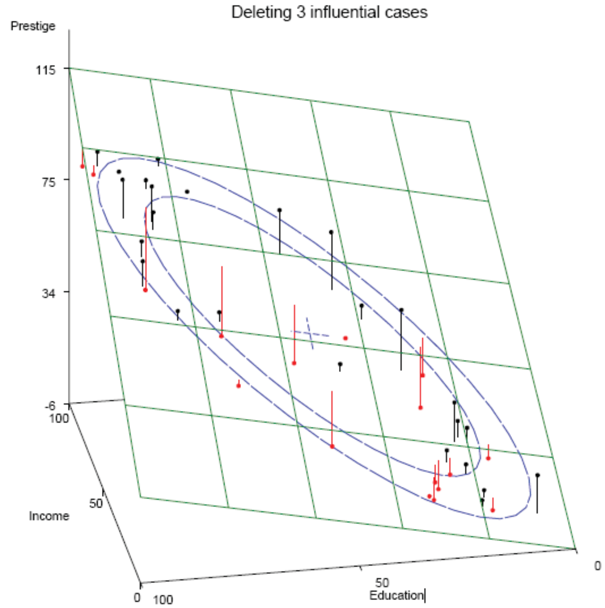
```
duncan.mod <- lm(prestige ~ income + education, data=Duncan)
avPlots(duncan.mod, id.n=2, ellipse=TRUE, …)
```



3D view:



3D view:



How to handle influential cases

- Observations in error, or from an extraneous population
  - delete or exclude them ?
  - Recall that in the fuel data, outliers suggested a better model!
- Robust methods – fit using a method that down-weights outliers
  - SAS: PROC ROBUSTREG
  - R: `MASS::rlm()`; `robust::lmRob()`
- Sensitivity analysis – effect on your conclusions?
  - Compare Q_{all} vs $Q_{(-i)}$ for any statistic, Q
  - Are there likely to be more observations like x_i in future samples?
  - Duncan data: Minister, RR Conductor clearly special – report main results excluding them, footnote Q_{all}

Summary

- Fitting a model is just the first step
  - Need to check whether assumptions are satisfied
  - If not, revise/change the model, or transform/modify the data
- Residuals: what you have not (yet) accounted for!
- Residual plots are your friend
  - Residuals vs. X or fitted Y: patterns?
  - NQQ plots to check for normality
  - Spread-level plots to check for constant variance

Summary

- Outliers & influential observations
  - Distinguish between “good leverage” points and “bad leverage” points
  - Influence = X-Leverage x Y-residual
  - Influence plots show effects of both
- Partial regression (added variable) plots
  - Show the relation of Y to X_i, controlling for all other Xs
  - Help you see exactly what the model is fitting
  - Visualize how and why observations are influential