Keele University

Statistics & Experimental Design with R

Barbara Kitchenham
Keele University

Correlation and Regression

Correlation

- The association between two variable
- Strength of association usually measured by a correlation coefficient ρ in range [-1, 1]
- Most well known
  - Pearson Product Moment Correlation coefficient
    - Arises from bi-variate normal distribution
      - If both variables are standardized then plotted
      - Elipse shape indicates an association
        - » Narrower the elipse the closer ρ~1(+ve) or -1 (-ve)
      - Circular shape indicates no associate with ρ~0

Bivariate Normal Distribution

- Bivariate Normal distribution

$$\phi(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} exp\left\{\frac{-1}{2(1-\rho^2)}[z_1^2 - 2\rho z_1 z_2 + z_2^2]\right\}$$

$$where\ z_1 = \frac{x_1 - \mu_1}{\sigma_1}\ and\ z_1 = \frac{x_2 - \mu_2}{\sigma_2}$$

- Standard Bivariate Normal z~N(0,1)

$$\phi(z_1, z_2) = \frac{1}{2\pi\sqrt{1-\rho^2}} exp\left\{\frac{-1}{2(1-\rho^2)}[z_1^2 - 2\rho z_1 z_2 + z_2^2]\right\}$$

- Generalises to n dimensions
- Pearson's ρ is a parameter of the distribution

Pearson's ρ

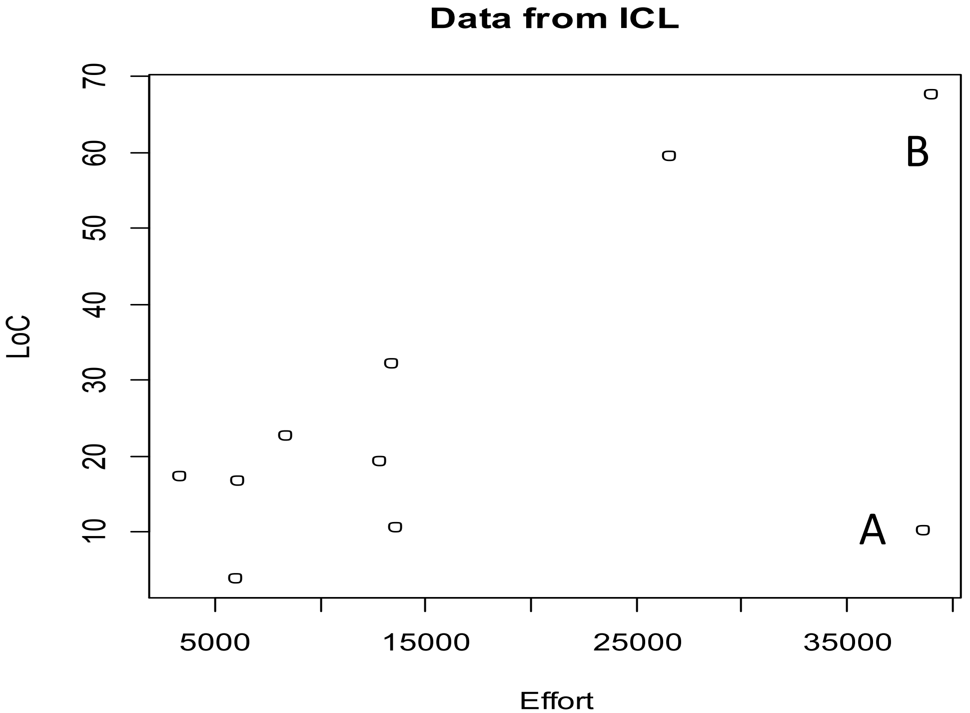
- From the bivariate normal distribution $\rho = \frac{cov(x,y)}{\sigma_x \sigma_y}$
- Estimated from data

$$\hat{\rho} = r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

- Calculating *r* does require normality
  - But statistical tests of significance do
  - Test H0 *r*=0 can be based on T having Student's t distribution n-2 df, where $T = r\sqrt{\frac{n-2}{1-r^2}}$
  - There is also a normalising transformation $z_r = 0.5\frac{log_e(1+r)}{log_e(1-r)}$
    - Which has standard error $\frac{1}{\sqrt{N-3}}$
  - Used when correlations from different sources need to be aggregated (such as during meta-analyses)

Small Data set

- Using cor.test in R ρ=0.57, T=1.9448 n.s.
- Delete A and ρ=0.57, T=5.887***
- Delete B and ρ=0.28, T=0.760 n.s.



Factors Affecting Magnitude Pearson's ρ

- The slope of the line about which points are clustered
  - If slope=0, ρ=0, the larger the slope the larger is ρ
- The magnitude of the deviations from the line
  - Closer points are to notional line the larger is ρ
- Outliers
- Restricting range of X values
  - Can increase or decrease ρ
- Curvature
  - ρ assumes a linear relationship

Robust correlation

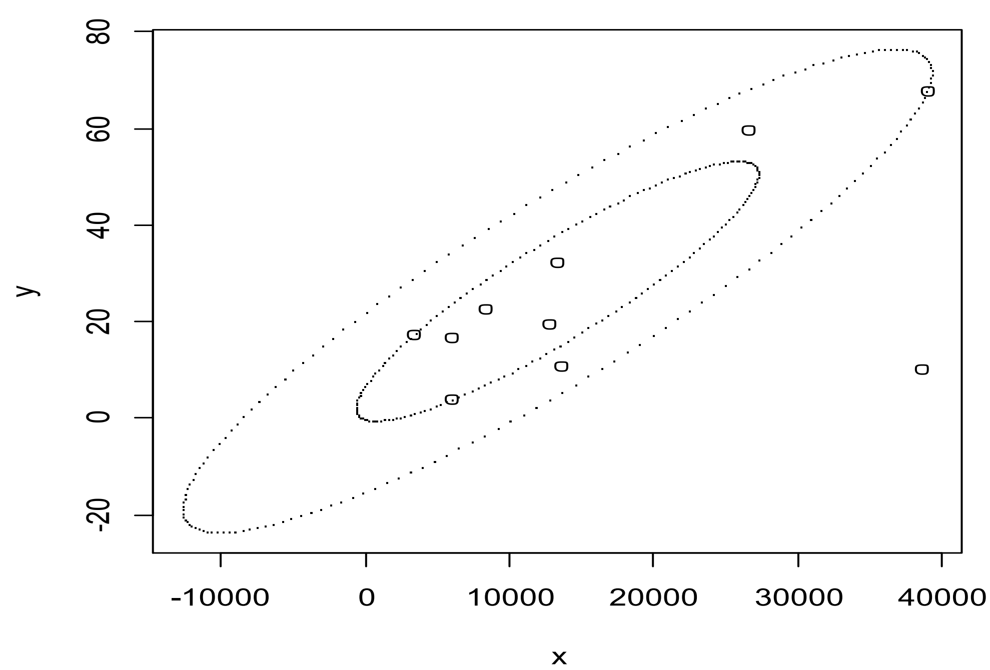
- Spearman's ρ
  - Replace data values by ranks
  - Uses same calculation as Pearson
- With previous data set
  - All data, $r=0.41$ $p=0.25$
  - With A removed, $r=0.67$, $p=0.059$
  - With B removed, $r=0.18$, $p=0.64$

Non-Parametric Correlation

- Kendall's tau (τ)
- Based on calculating slopes between all pairs of points
  - Takes median slope
- With previous data set
  - All data, r=0.33 p=0.22
  - With A removed, r=0.56, p=0.045
  - With B removed, r=0.17, p=0.61

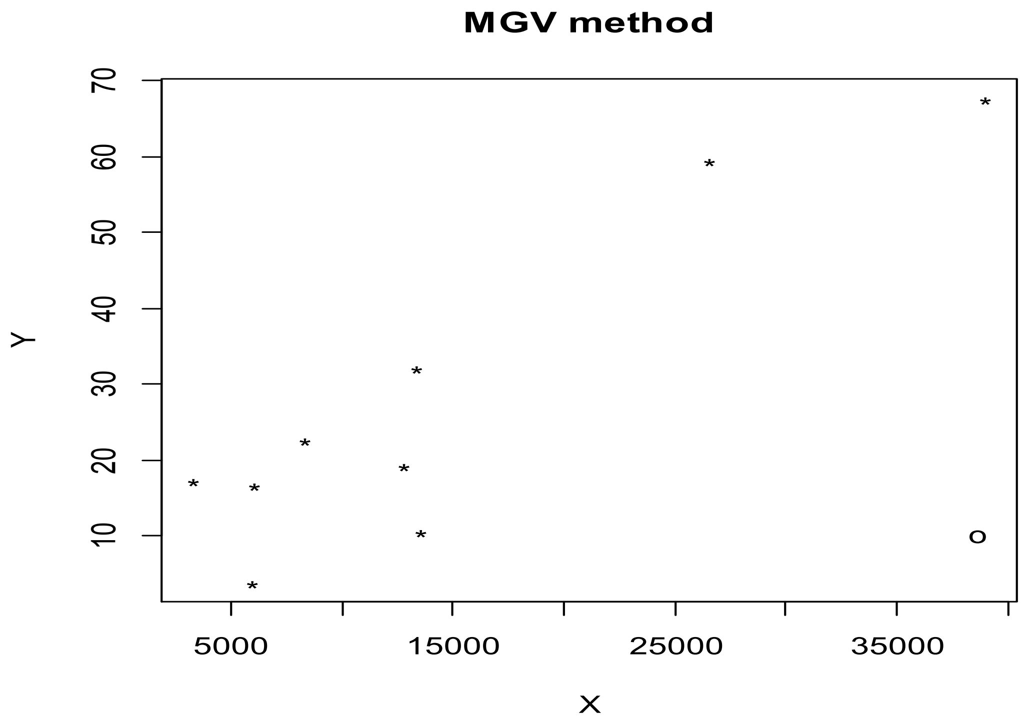
RelPlot

- relplot function is a bivariate equivalent of box plot
- Shows the central ellipsoid part of the bi-variate distribution plus outliers
- Calculates a robust estimate of r=0.90
- Does not generalise to more dimensions
- Assuming bi-variate normal means negative values are expected



MGV method for outliers

- Minimum Generalised Variance method can be used with many variables

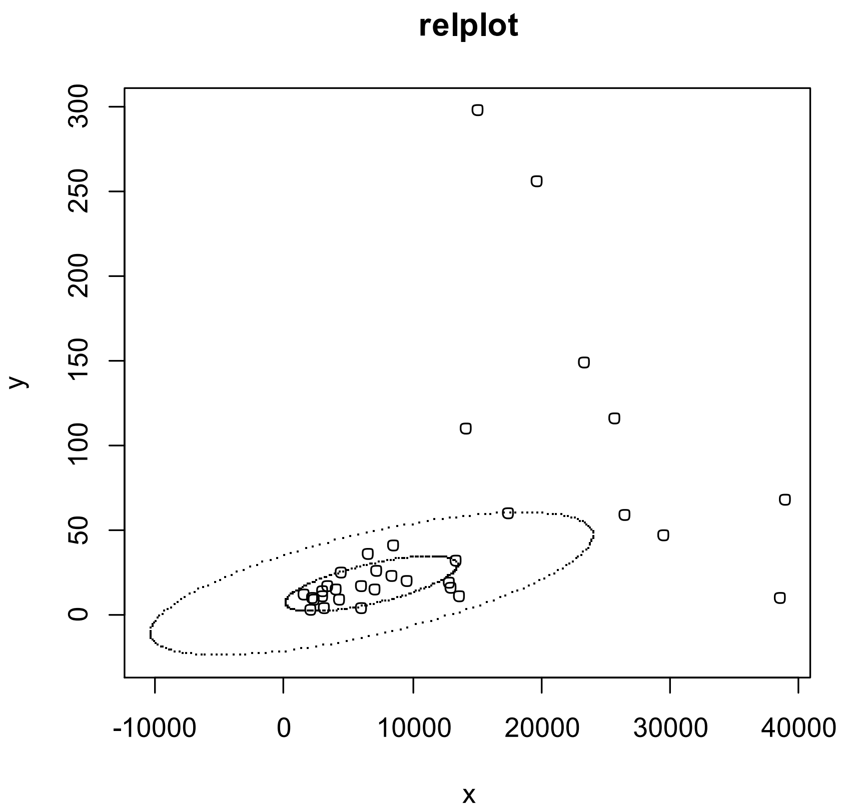


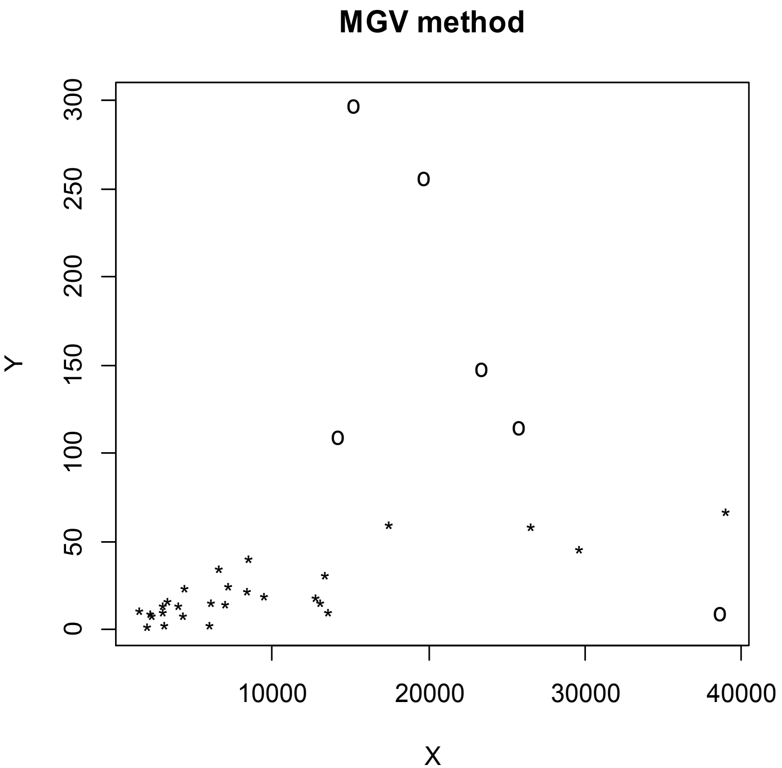
Robust Correlations

- Winsorized correlation (wincor(x,y))
  - Replace X and y values at extremes with 25 (low) 75 (high) percentile values
  - 0.407 sig.level=.276
- Percentage Bend Correlation
  - Not estimate of Pearson's r
  - New correlation robust to changes in distribution
  - Based on trimming univariate outliers
  - corb(x,y,corfun=pbcor,nboot=599)
  - r_{pb}=.441 Boostrap CI=(-0.44, 0.97)
- Skipped correlations (i.e. remove outliers)
  - Removed based on MGV then use Pearson (r=0.91)
  - Need to adjust Test value & critical value

$$T_o = r_o \sqrt{\frac{n-2}{1-r_o^2}} = 6.29, cv = \frac{6.947}{n} + 2.3197 = 3.0144$$

Comparison on full data set





Linear Regression

- Finding the parameters of a model of the form
$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$$
  – Y is the response/outcome/dependent variable
  – X_i is the *i*th of *p* stimulus/input/independent variables
  – β_i is the ith parameter of the model
- A linear model is linear w.r.t the parameters
  – Polynomial models are linear models of the *n*th order where *n* is highest power
  – I.e. a second-order regression model has form
$$Y = \beta_0 + \beta_1 X + \beta_2 X^2$$
  – A non-linear model might have form $Y = \beta_0 X^{\beta_1}$

Least Squares Principles

- Basic model for one input variable is

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

- Sum of squares of deviations from true line is

$$S = \sum_{i=1}^{n} \epsilon_1^2 = \sum_{i=1}^{n} (Y_i - \beta_0 + \beta_1 X_i)^2$$

- To estimate by least squares
  - Differentiate w.r.t each parameter in turn
  - To find the turning point (i.e. minimum) set each differential to 0
    - Solve for each parameter in turn

Parameter Estimation

- Differentials are

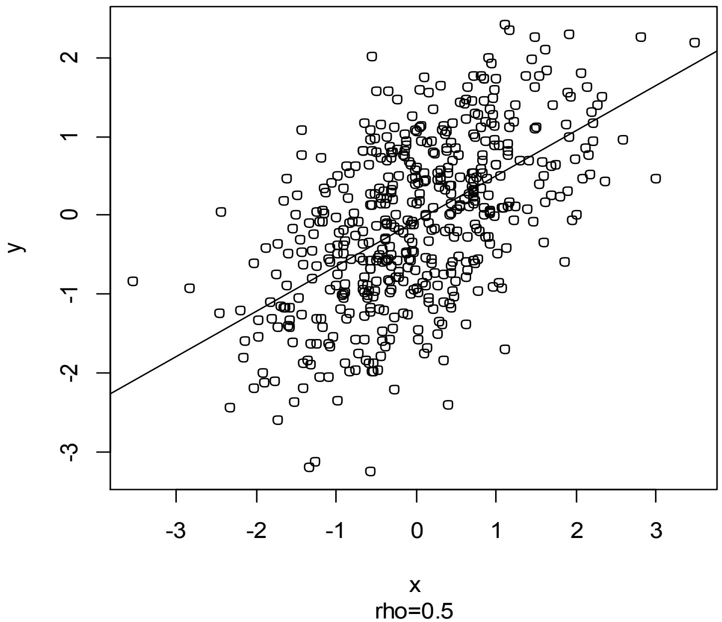
$$\frac{\delta S}{\delta \beta_0} = -2\sum(Y_i - \beta_0 + \beta_1 X_i) \qquad \frac{\delta S}{\delta \beta_1} = -2\sum X_i(Y_i - \beta_0 + \beta_1 X_i)$$

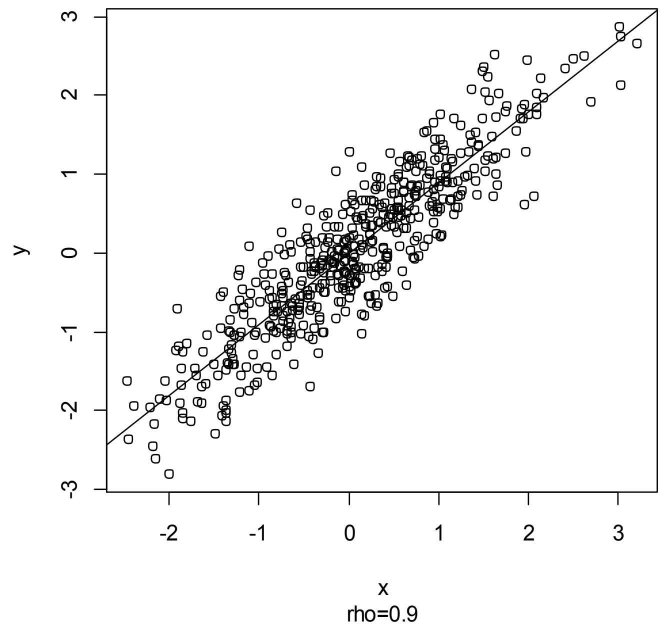
- Solutions after setting each to 0 are

$$b_1 = \frac{\Sigma(Y_i - \overline{Y})(X_i - \overline{X})}{\Sigma(X_i - \overline{X})^2} = \frac{s_{XY}}{s_X^2} = r\frac{s_Y}{s_X} \qquad b_0 = \overline{Y} - b_1\overline{X}$$

- For standardized normal variables $b_1 = r$
  - Slope must less than 1, even if Y=X
  - The larger the error term, the larger *r* and the lower the value of b_1

Bivariate Normal Distributions





b_1=0.57441
b_0 =-0.07613

b_1=0.9018
b_0=-0.0097

Multivariate Regression

- Formulate in matrix algebra terms, assuming X and Y have means removed i.e. $Y=y-\mu_y$

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

- $\boldsymbol{Y}$ is an (n×1) vector
- $\boldsymbol{X}$ is an (n×p) matrix of known form
- $\boldsymbol{\beta}$ is a (p×1) vector of parameters
- $\boldsymbol{\epsilon}$ is a (n×1) vector of error terms
- Where $E(\boldsymbol{\epsilon})=0$, $V(\boldsymbol{\epsilon}) = \boldsymbol{I}\sigma^2$
- Solution is $\boldsymbol{b} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}$

Least Squares Properties

- Fitted values are obtained from
$$\hat{Y} = Xb = X(X'X)^{-1}X'Y$$
- Vector of residuals $\epsilon = Y - \hat{Y}$
- Variance of parameters $V(b) = (X'X)^{-1}$
- Multiple Correlation Coefficient $R^2 = \dfrac{\sum_{i=1}^{n}\left(\hat{Y}_i - \bar{Y}\right)^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}$
- Adjusted $R_a^2 = 1 - (1 - R^2)\left(\dfrac{n-1}{n-p}\right)$
- Both R^2 Vulnerable to outliers
- Many diagnostic tools available based on residuals and Hat Matrix

The Hat Matrix

- Hat Matrix is defined as $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$
- Called the Hat matrix because $\hat{\boldsymbol{y}} = \boldsymbol{H}\boldsymbol{y}$
- Its important because if h_{ii} is *i*-the diagonal element of of ***H***
  - Difference between
    - Parameter with and without observation x_j is

$$\hat{\beta} - \hat{\beta}(i) = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_i' \frac{\epsilon_i}{(1 - h_{ii})}$$

    - Fitted value with and without observation x_j is

$$\hat{y}_i - \hat{y}_i(i) = \frac{\epsilon_i h_{ii}}{(1 - h_{ii})}$$

Three Types of Residual

- Residuals $\epsilon_i = r_i = y_i - \hat{y}_i$
- Standardized Residuals

$$r_i = \frac{y_i - \hat{y}_i}{s} \qquad s^2 = \frac{\sum \epsilon_i^2}{n-p}$$

- Studentized Residuals (based on omitting each data point in turn from variance)

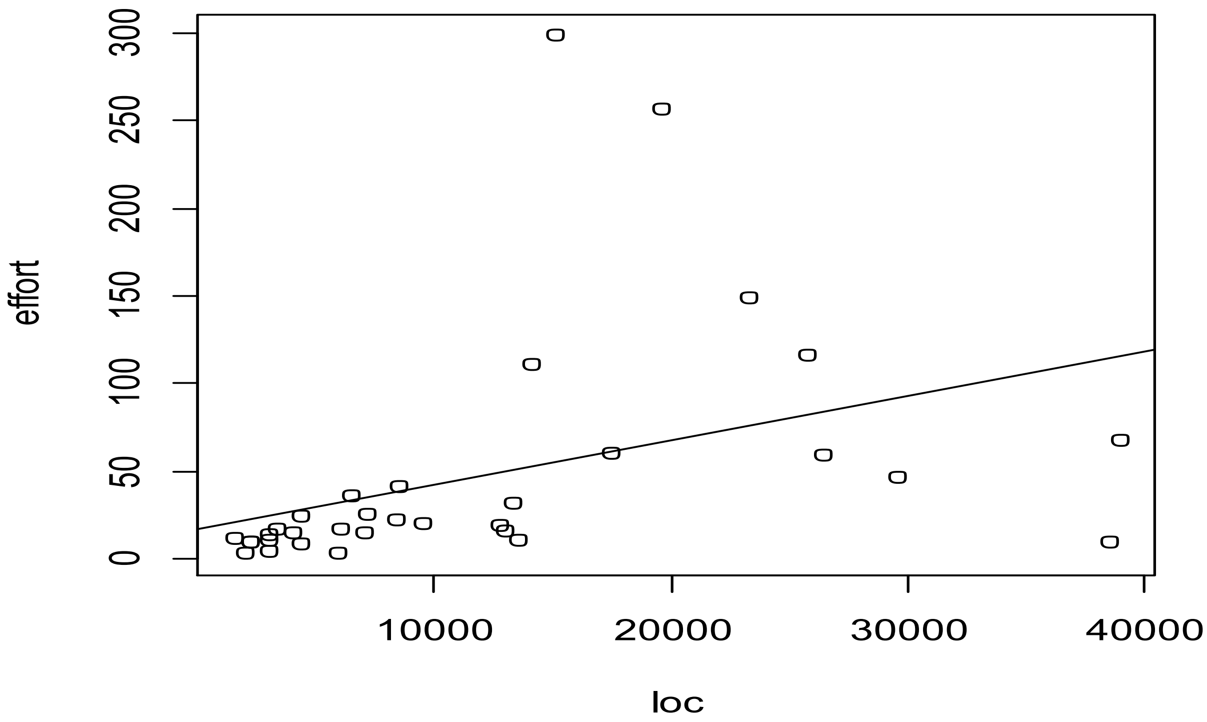
$$r_s(i) = \frac{y_i - \hat{y}_i}{s(i)} \qquad s(i)^2 = \frac{\left(\sum \epsilon_i^2\right) - \frac{\epsilon_i^2}{(1-h_{ii})}}{n-p-1}$$

- Sadly doesn't automatically provide fitted values based on i-1 points
  - However, lm provides access to the hat matrix values
    - Via the fitted model i.e. hatvalues(fit)
    - So can be calculated by writing your own R program

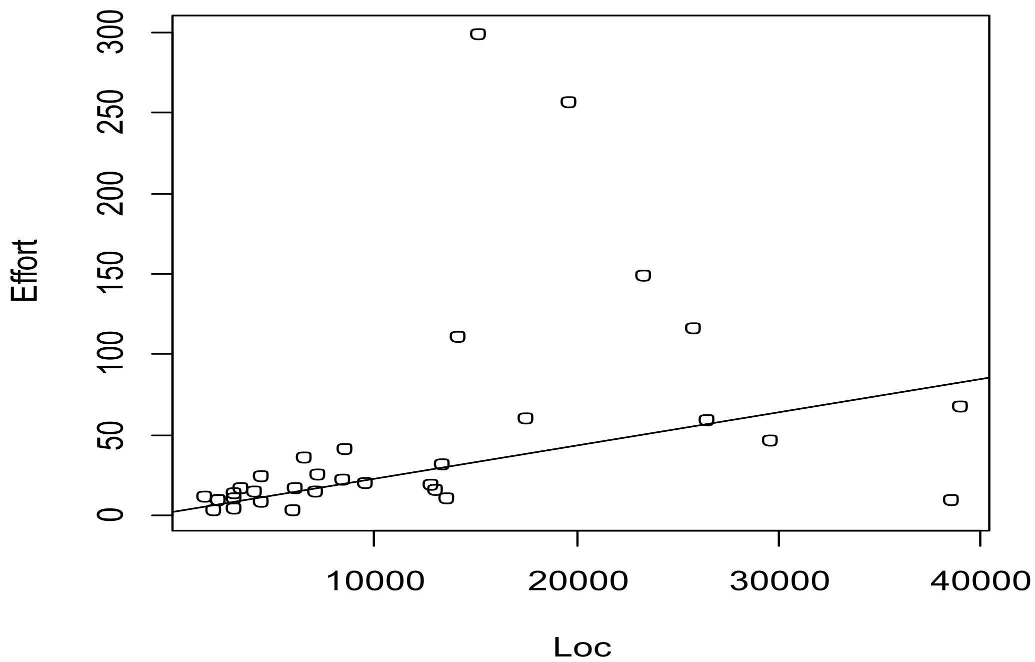
Fitting Regression Models in R

- The R command is
  - lm(y~x1+x2+..+Xn,data=mydata)
- You should save the output of the linear model e.g.
  - fit<-lm(effort~loc,data=iclbt)
  - Effort=17.22+.00253322×loc
- From the object “fit” you can access
  - Residuals
  - Hat values
  - Fitted values

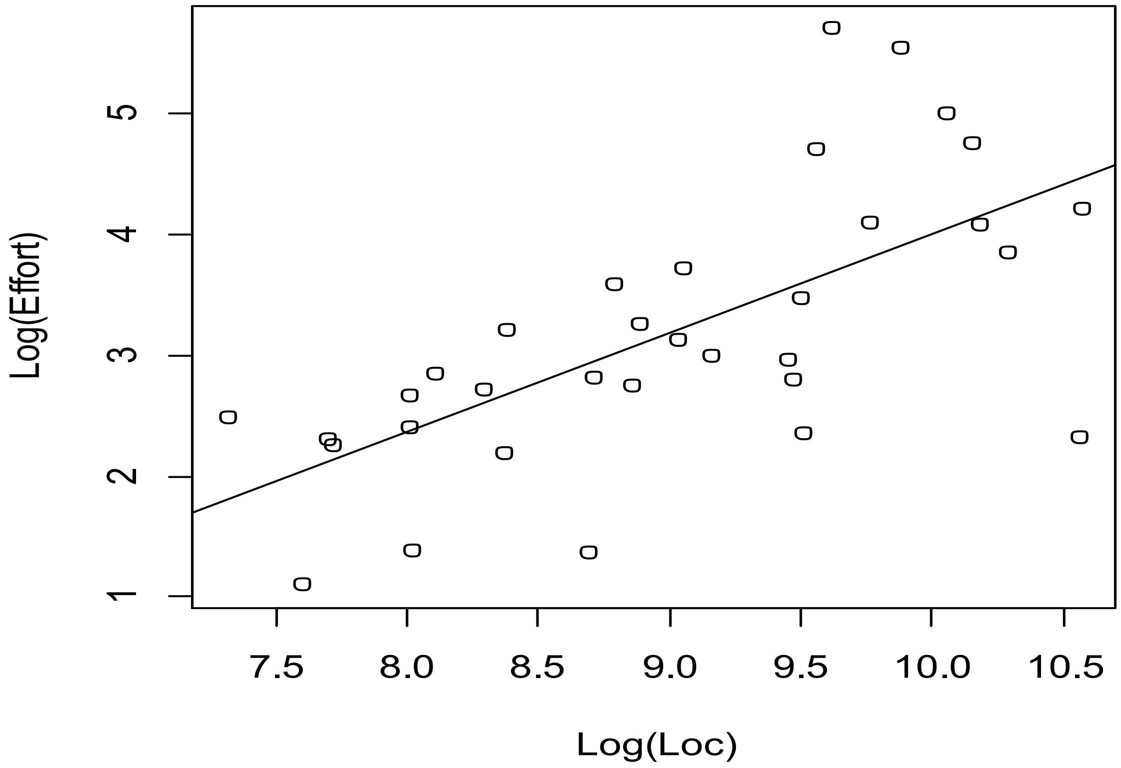
Plotting Effort and Loc showing Regression Line



Theil-Sen Regression



Using Log Transformation



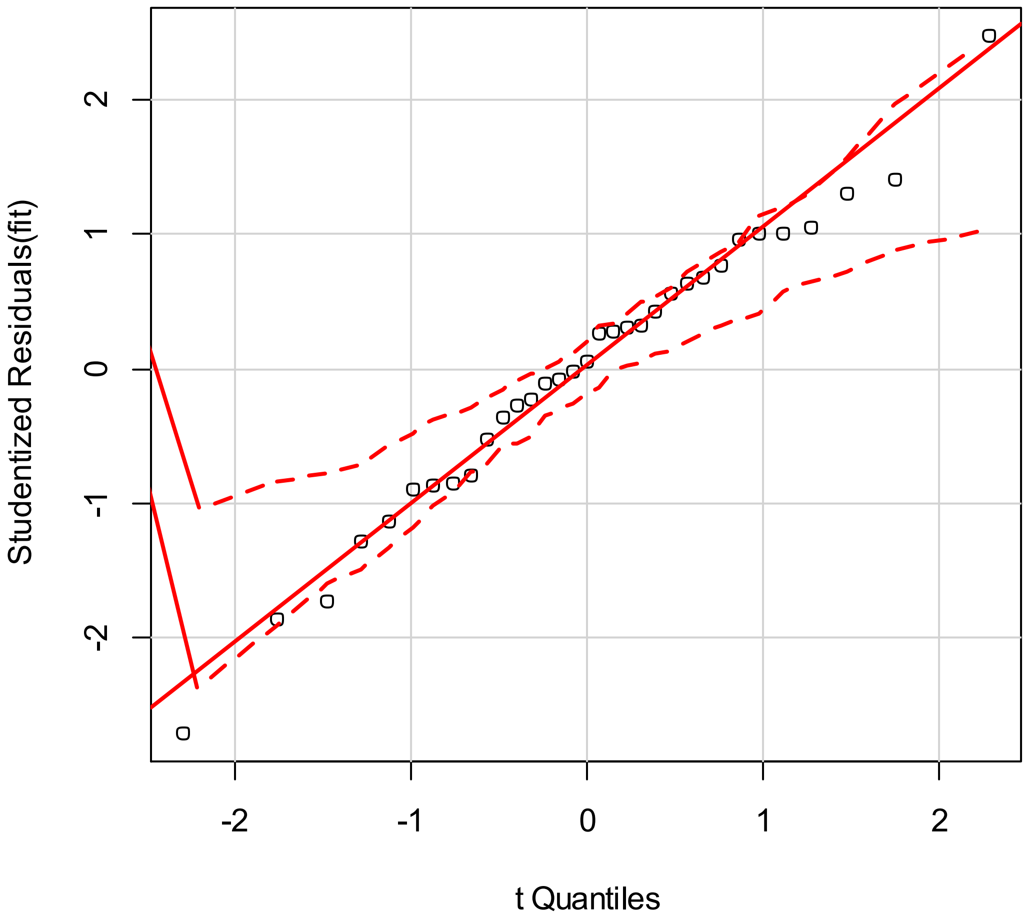
Diagnostics

- Many diagnostic facilities assume fitting via the linear model function
- To evaluate diagnostics can use
  - Log(effort)=Log(loc)+log(dur)+co
- “co” is a factor that defines the source of the data
- Needs to be defined as a factor to
  - iclbt$co<-factor(c("1","2","3"))

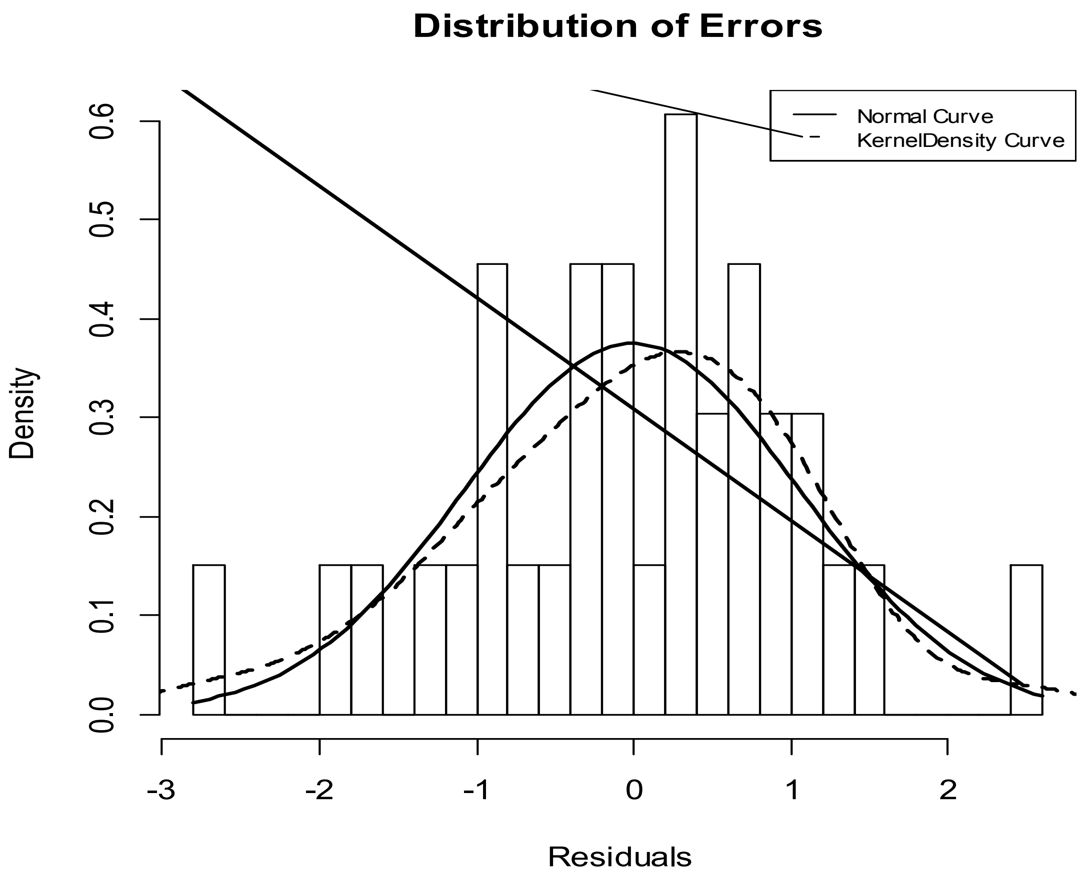
Diagnostic Aids - 1

- Q-Q Plots
  - Plots Studentized residuals against a t distribution with n-p-1 degrees of freedom
- Histogram of residuals (all types)

QQPlot for ICLBT data



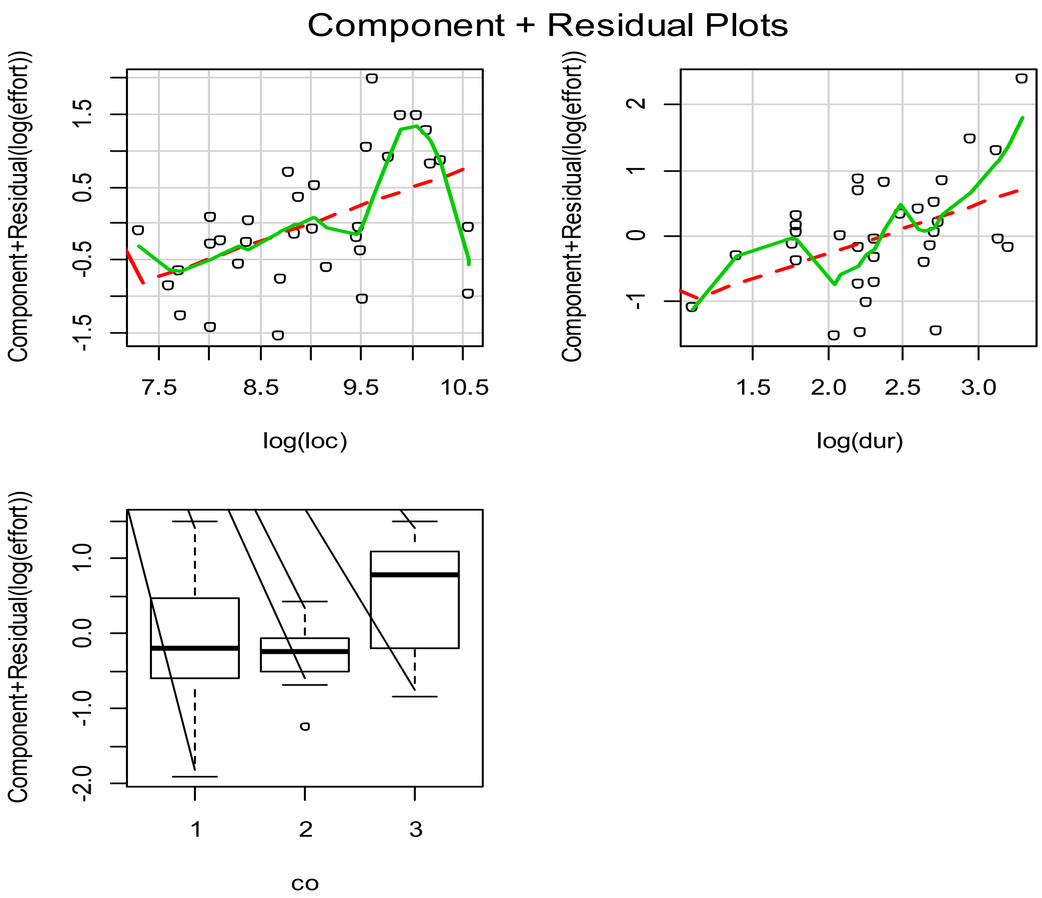
Residual Plot



Diagnostic Aids - 2

- Component + Residual plots
  - Partial residual plots
  - For each j-variable plots $\epsilon_i + (\beta_j X_{ij})$ against X_{ij}
    - where ϵ_i are based on full model
  - The straight line on graph is the least squares fit
  - The other line is the "lowess" line
    - A nonparametric weighted fit line based on locally weighted polynomial regression

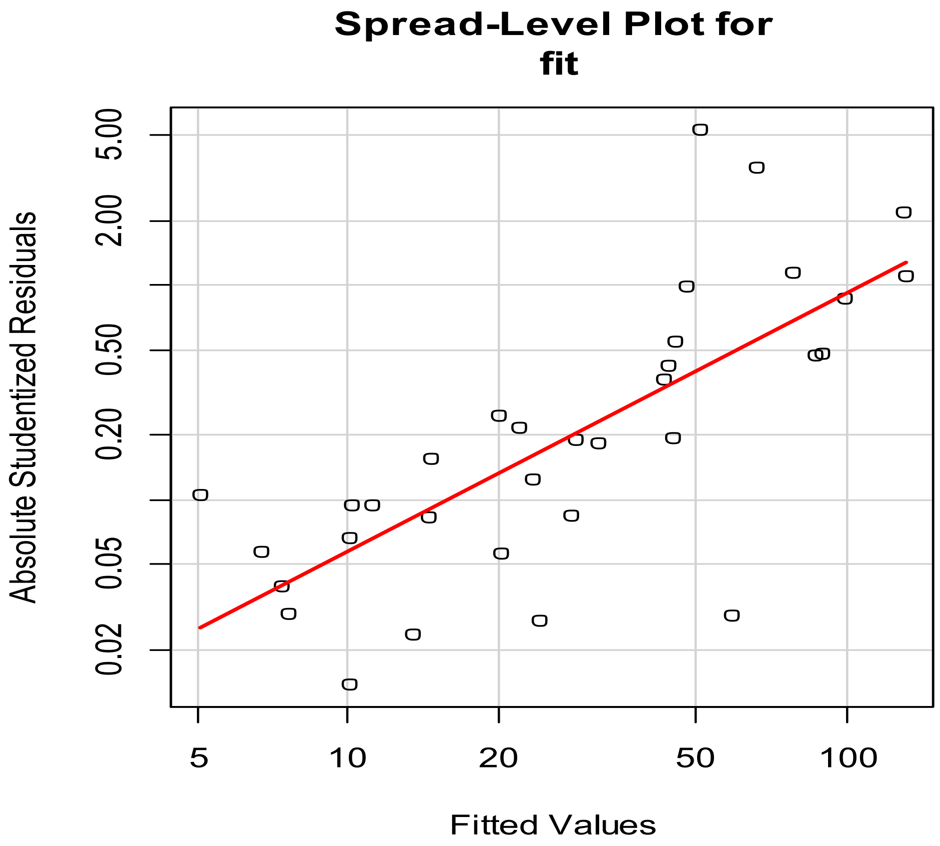
CrPlots for ICLBT data



Diagnostic aids - 3

- Test for non-constant error variance
  - ncvTest() function
    - For ICLBT data, ChiSquare = 1 4.055072 p=0.044*
- Plot of absolute standardized residuals versus fitted values with best fitting line (Spread-Level Plot)
  - Can indicate possible non-linearity in Y variable
    - Suggests power transform
    - 0 suggestion identifies log transform
    - Suggests –0.33
- Multcollinearity vif() function
  - Only when multiple X variables
  - Measure extent to which parameter standard deviation for a parameter is expanded
    - Relative to model with independent variables
  - If square root of vif >2 there may be a problem
    - No problem for this model

Spread Level Plot



Major Diagnostic Concepts

- Outliers
  - Observations that are not predicted well by model
  - Have large residuals
- High leverage points
  - Are outliers with respect to other predictors
  - Found using the Hat Matrix
- Influential points
  - Observations that have an major impact on parameter values
  - High leverage points that are also outliers
    - Added Value plots
    - Cook's Distance

Cook's Distance

- Aim to summarize the information in
  - Leverage
  - Residual-squared plot
- Into single number index

$$D_i = \frac{\epsilon_i}{k} \frac{s_{(i)}^2}{s^2} \sqrt{\frac{(n-1)}{(1-h_{ii})}}$$

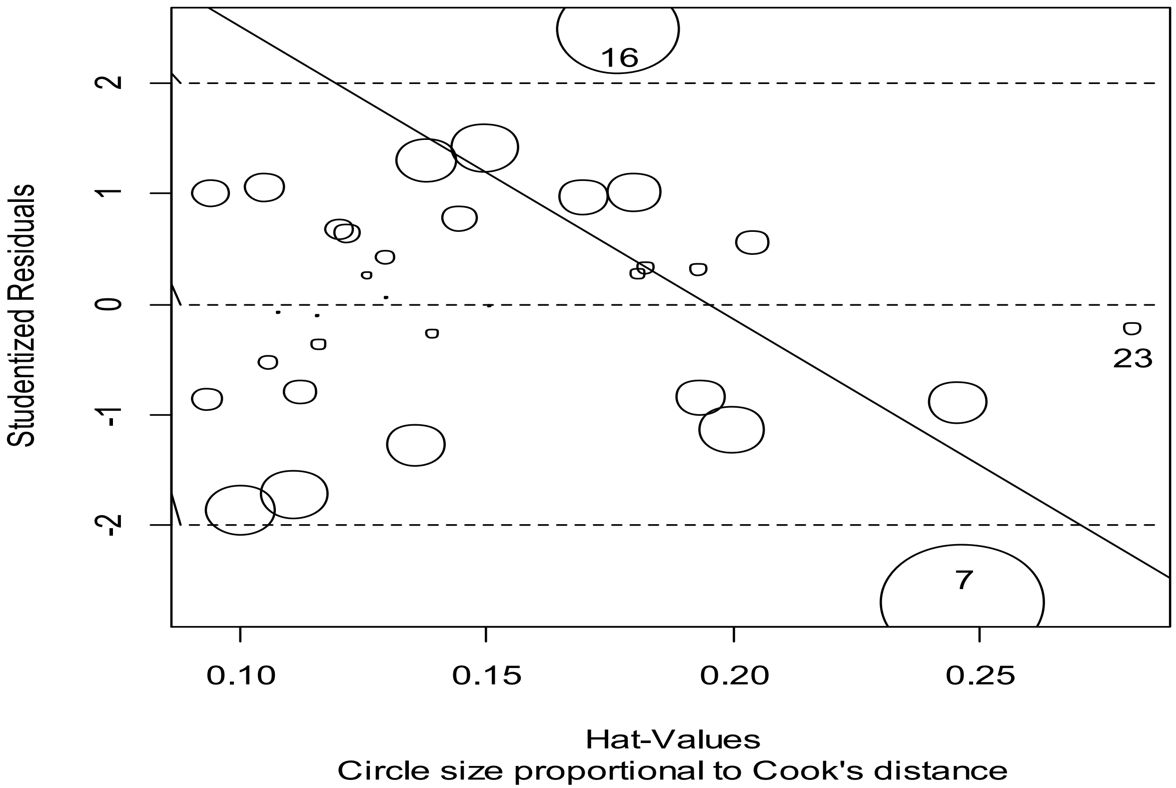
- Unusually large Cook's D greater than $2\sqrt{k/n}$
  - k is number of parameters including constant
  - N is number of observations

Aids for Outlier Detection -1

- Outlier detection based on Studentized residuals using outlierTest() function
  - Reports Bonferoni adjusted p-value for the largest absolute residuals
  - Identifies points 61 & 21 as significant outliers
- Added Value Plots
  - For each Xj
    - Show impact of regressing Y on other variables against X_j regressed on other variables
  - Can be used to assess impact of specific data points
- Influence plot
  - Studentized Residals against Hat-values with circles indicating Cook's distance

Influence plot for ICLBT Dataset

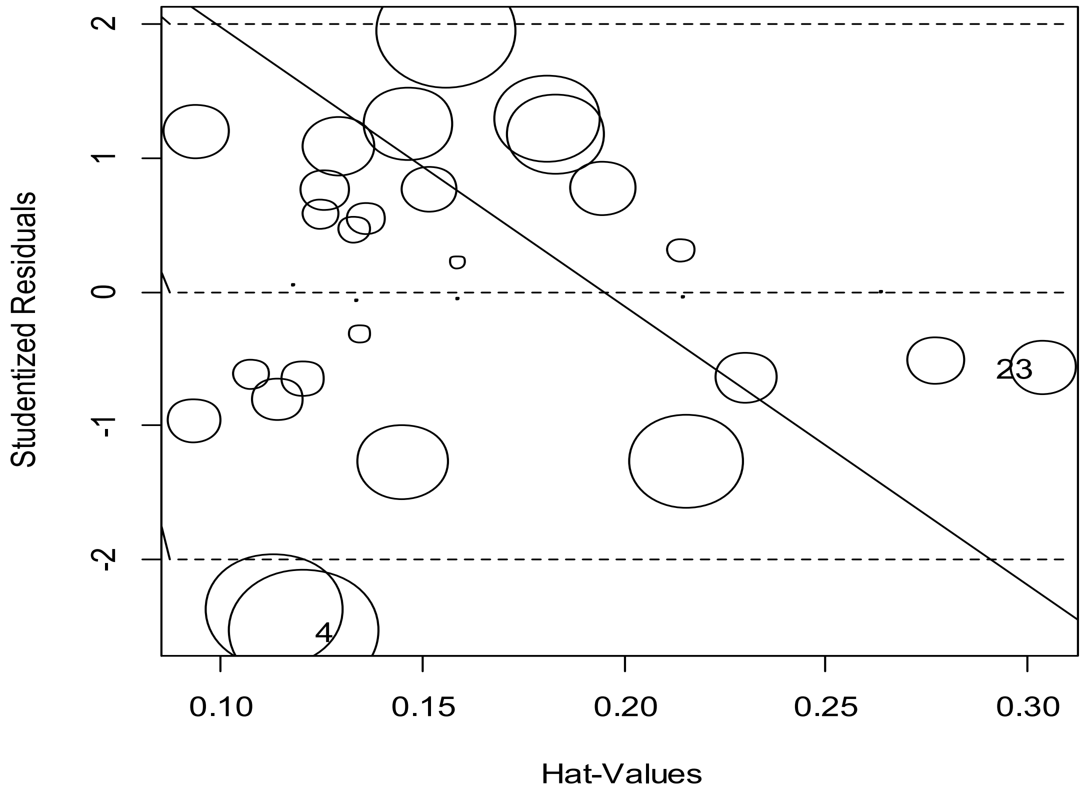
- Compare main outliers (i.e. 1, and 7) with outlier detection results Slide (12)
- Effect of removing points easy use:
  - update(fit,subset=-c(7,16))



Impact of Removing Outliers

Coefficients	Original	New
Intercept	-3.1804*	-4.1907**
Log(loc)	0.4895*	0.7089**
Log(dur)	0.7534·	0.390
Co2	-0.1049	-0.1976
Co3	0.631	0.4219
Adj R^2	0.481	0.5876

Influence Plot of reduced model



Models with Dummy Variables

- Exactly equivalent to Analysis of Covariance (ANCOVA)
- Uses variables that partition the dataset
  - E.g. Co (which stands for company) in the ICLBT database
- Co is coded as an integer and need to be specified to R as a factor
- R maps k different levels per factor into k-1 dummy variables
  - The effect of the "missing" dummy variable is included in the intercept
  - If only one dummy variable
    - The Intercept corresponds to the effect of the missing variable
    - The parameter values given to other dummy variables are
      - Effect of missing dummy variable – Effect of dummy variable

Dummy Variables - 2

- A dummy variable shifts the intercept of the regression line
  - To give a separate regression line for each data partition
- If we want to change the slope as well as the intercept we need to change the model to a model with interactions
  - lm(log(effort)~co*(log(dur)+log(loc)),data=iclbt)
- Multiple factors in a model with no variables produces a multi-way ANOVA

Interactions with Company

Coefficients	Estimate	Std.Error	t value	Pr(>|t|)
(Intercept)	-3.646	2.8032	-1.301	0.2057
co2	-2.722	3.7120	-0.733	0.4705
co3	0.7553	3.7311	0.202	0.8413
log(dur)	1.3364	0.6680	2.000	0.0569 (.)
log(loc)	0.3217	0.2839	1.133	0.2683
co2:log(dur)	-0.7094	0.8775	-0.808	0.4268
co3:log(dur)	-1.1025	0.8165	-1.350	0.1895
co2:log(loc)	0.6292	0.4405	1.428	0.1660
co3:log(loc)	0.2763	0.4095	0.675	0.5063

Removing X-variables

- May need to select most plausible model with least number of X-variables
- Stepwise regression available in R
  - Forwards stepwise starts with no variables and adds one at a time
  - Backwards starts with all variables and removes them one at a time
  - Stepwise goes forward but re-assesses all variables as each new one is added
  - Based on Akaike Information Criteria (AIC)
- Can also inspect all possible regressions
  - With limited number of variables

Akaike Information Criterion (AIC)

- Used to judge competing models
  - Function of the Log Likelihood function

$$AIC = 2k - 2\ln(L)$$

  - k = number of parameters in model
  - Smaller values are preferable
- Version adjusted for sample size n is preferable $AIC_C = AIC + \frac{2k(k+1)}{n-k-1}$
- Assesses impact of changing number of parameters (not functional form of model)

Other capabilities

- Kabacoff published R functions
  - For Cross-validation
    - Checking a model by splitting the data into validation and training data sets
    - Predicting the outcome value for the validation data
    - Perform k-fold cross validation
      - I.e. creates k different training & validation sets at random
      - Based on changes to the R-square statistic
  - To assess the relative importance of different variables
    - Model must not have categorical variables

Robust Regression

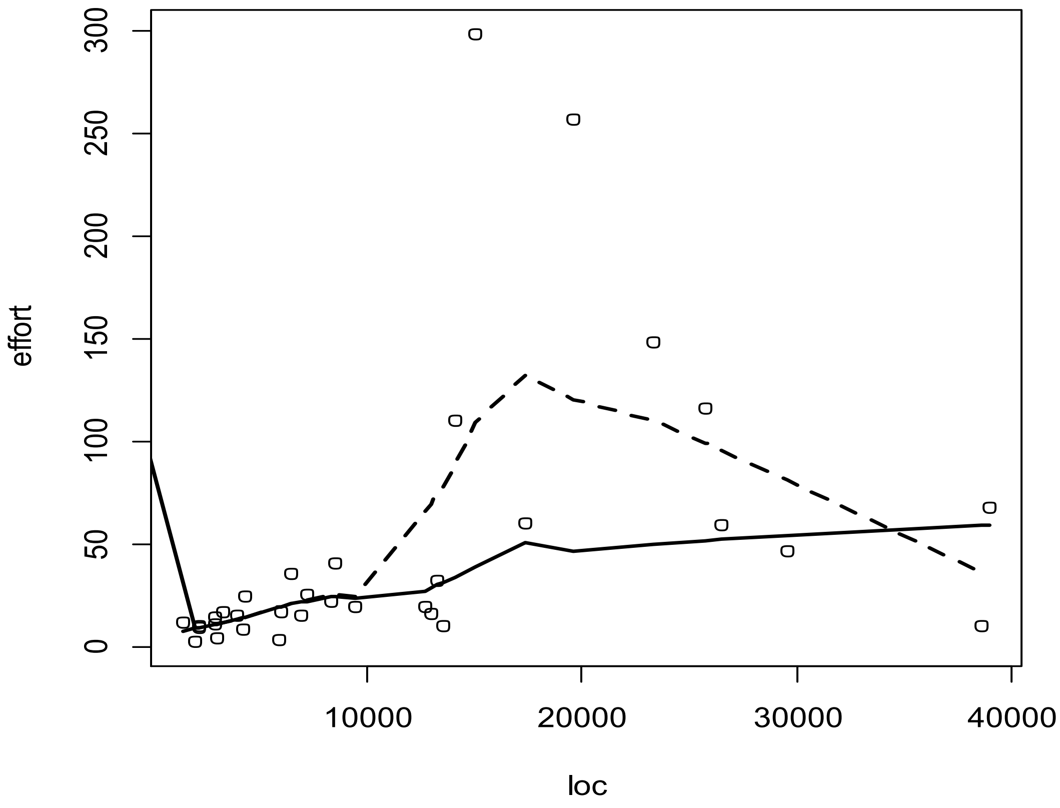
- Lowess Local Polynomial Regression $y = m(x) + \epsilon$

$$y = b_0 + b_1(x - x_0) + b_2(x - x_0)^2 + \cdots + b_p(x - x_0)^p$$

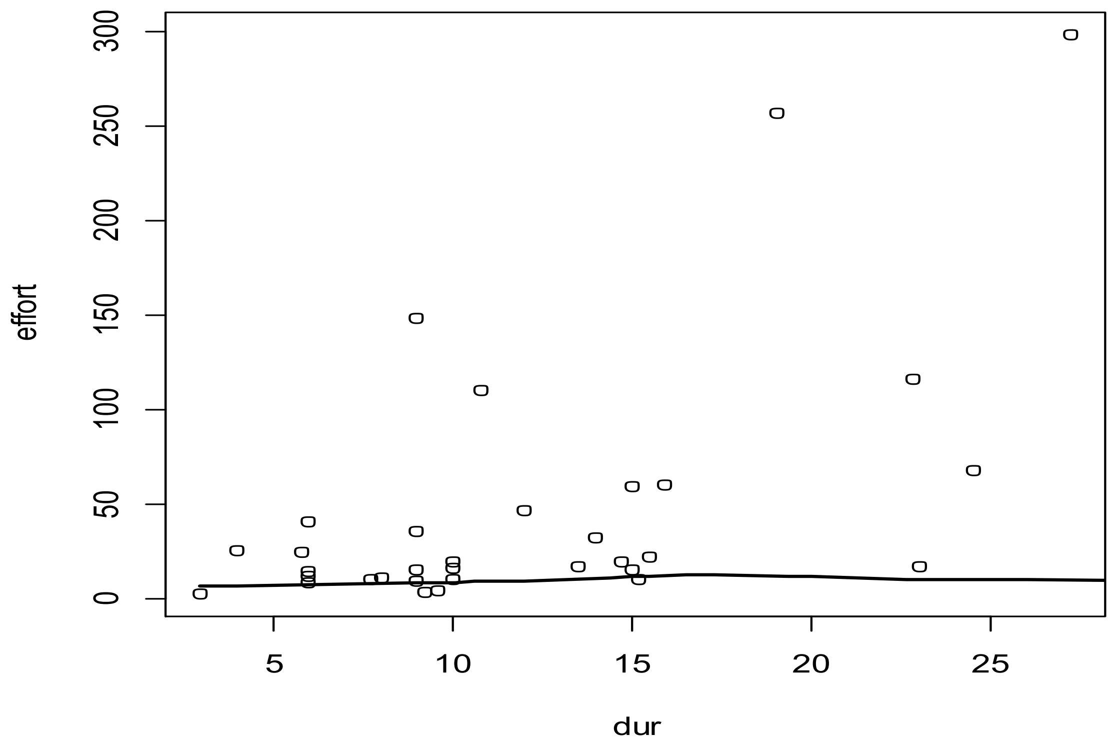
$$W(z) = \begin{cases} (1 - |z^3|)^3 & if\ |z| < 1 \\ 0 & if\ |z| \geq 1 \end{cases} \qquad z = (x - x_0)/h$$

  - *h* is half-width of a window enclosing observations for local regression
  - At x_0 estimate height of regression curve is $\widehat{y_0} = b_0$
  - Typical to adjust *h* so each local regression includes a fixed *s* proportion of data
  - *s* is span of local-regression smoother
  - Large span smoother fit but larger order of local regression
    - Require a trade-off

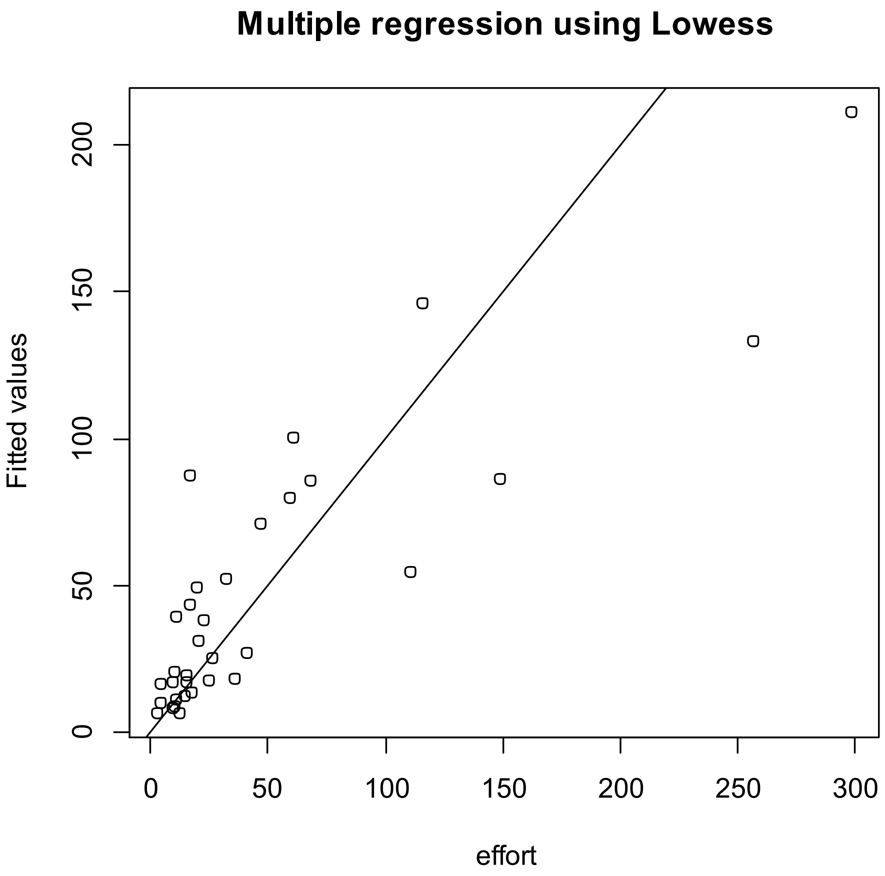
Fitted line for ICLBT data
Size v. Effort

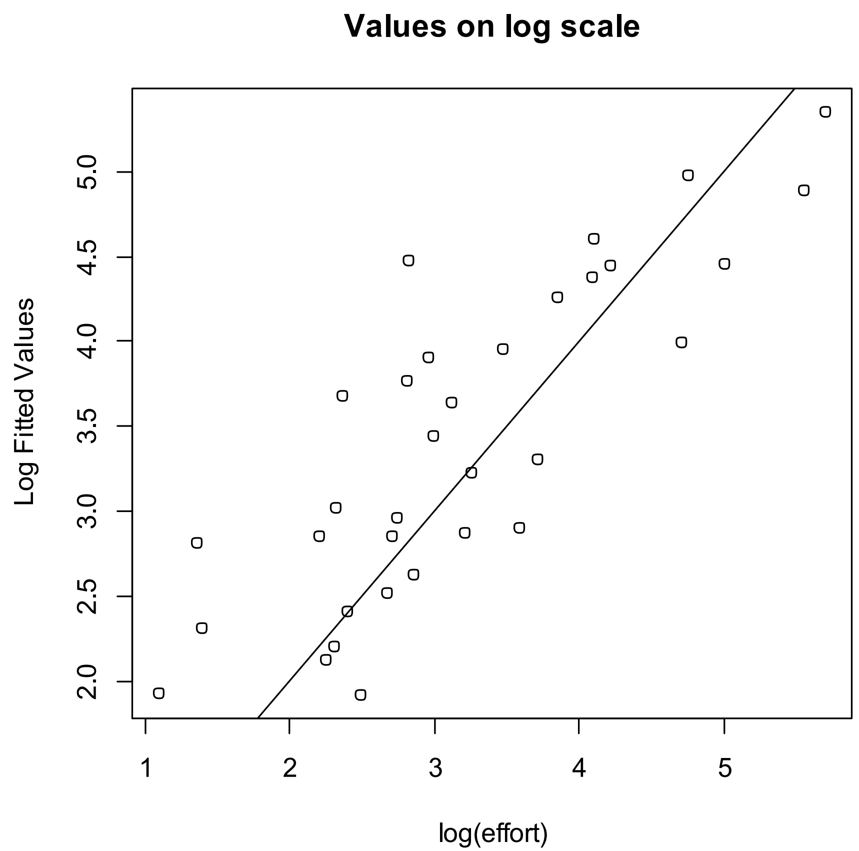


Duration v. Effort



Multiple lowess Regression





Kernel Regression

- Kernel estimators estimate some measure of location for y given x
- w_i is a measure of how close x_i is to x

$$\widehat{m}(x) = \sum w_i y_i$$

- K(u) is a contours, bounded and symmetric function $\int K(u)du = 1$

Kernel Regression - Continued

- m(x) estimated from

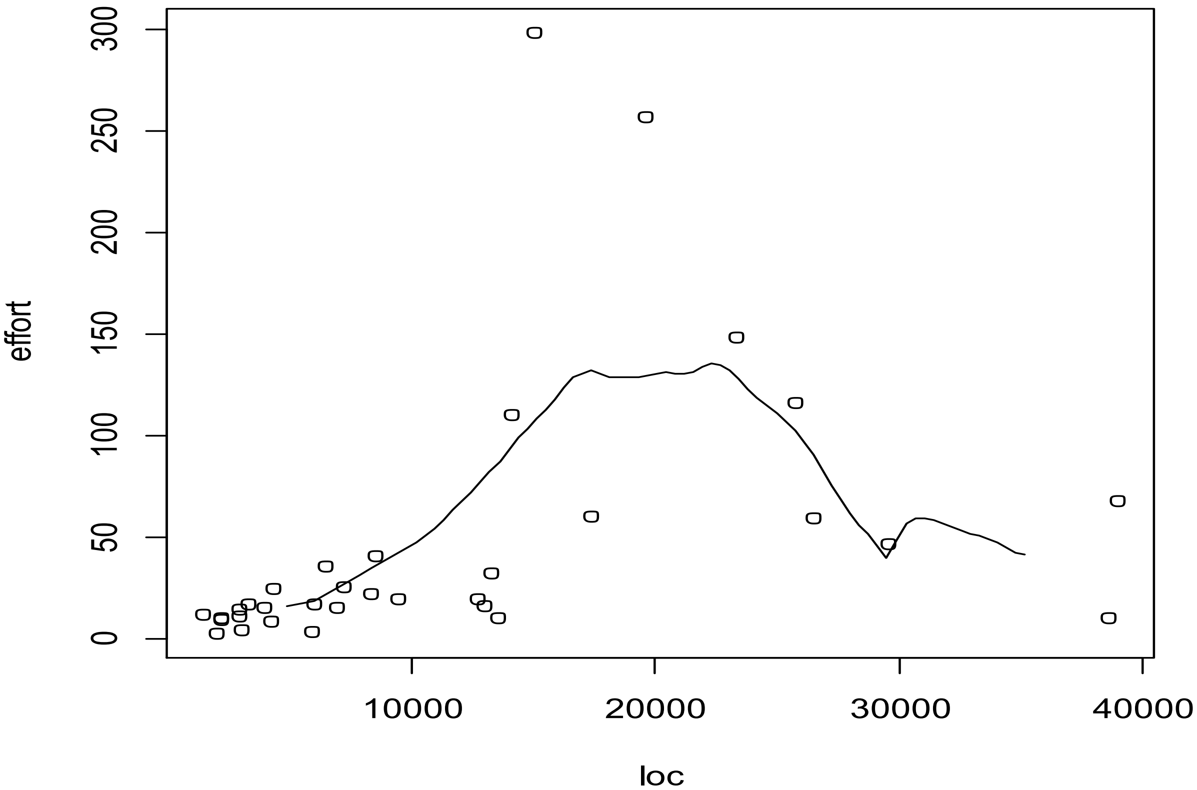
$$w_i = \frac{1}{W_s} K\left(\frac{x - x_i}{h}\right) \qquad W_s = \sum K\left(\frac{x - x_i}{h}\right)$$

- h is span $h = \min\left(s, ISQ/1.34\right)$
  - ISQ is interquartile range
- Given x ,
  - b_0 and b_1 estimated using weighted regression

$$\widehat{m}(x) = b_0 + b_1 x \qquad w_i = K\left(\frac{x_i - x}{h}\right)$$

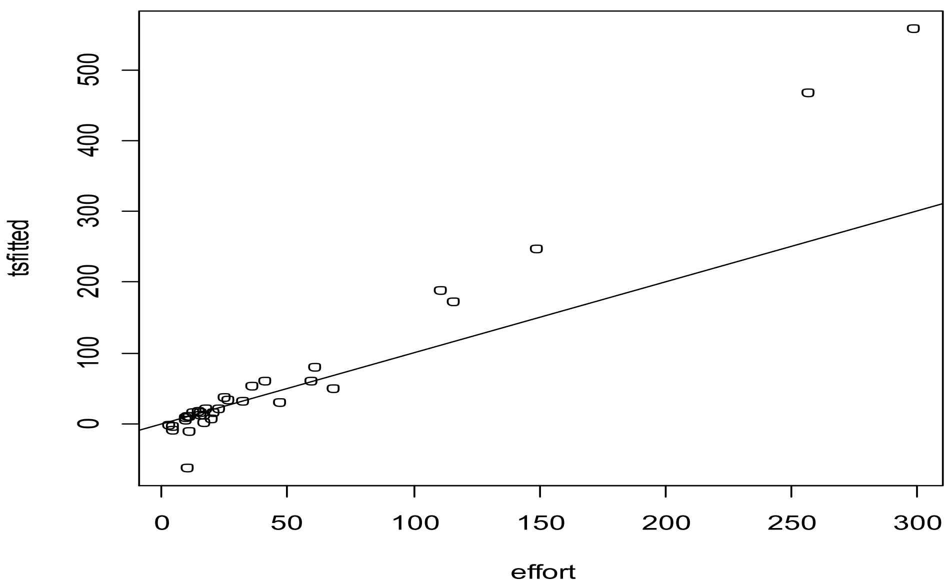
- Smooth is created by taking x to be a grid of points and plotting results

Kernel Regression



Non-Parametric Regression

- Theil-Sen can handle multiple regression
  - Not with dummy variables
  - Fitted line fitted mass of data points
    - 5 fitted values were negative



Conclusions

- Combination of transforming variables and extensive diagnostic facilities
  - Seem to reduce the need for robust regression
    - At least in the case of linear models
- Non-parametric approaches don't always work well
  - Don't permit group variables
  - Are not integrated with diagnostics library(car)
- Lowess is promising
  - Not yet well integrated with diagnostics