Regression Graphics

Can We Diagnose Violations from Plots?

Steven W. Nydick

May 18, 2012

Why Graphical Checks

Anscombe (1973) put it succinctly:

> *"Instead of looking at a scatterplot of $\{e_i\}$ against $\{\hat{y}_i\}$, we could detect effects such as those just listed by calculating suitable test statistics, and we could assess their significance. But the plot shows a variety of features quickly and vividly, and formal tests often seem unnecessary" (p. 18).*

Residual Plots

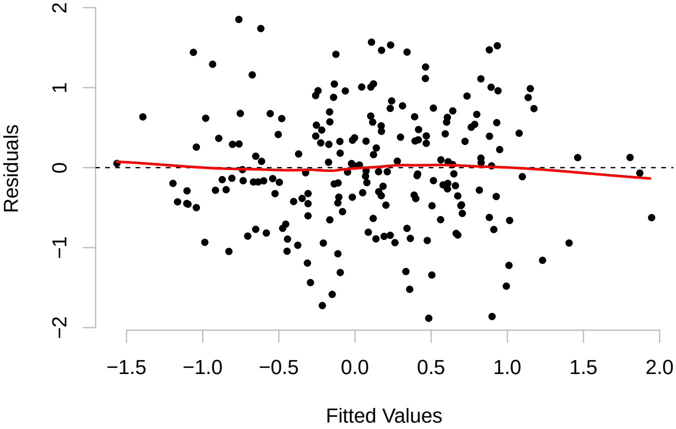
Constructing a residual plot is simple:

1. Run the regression of y on $\boldsymbol{x}$.
2. Find $\hat{y}$.
3. Extract the model-implied error: $e = y - \hat{y}$.
4. Plot e versus $\hat{y}$.

If no assumptions are violated $\implies$ no relationships in plot.

Residual Plots: No Relationship

Here is an ideal residual plot:



Residual Plots: Diagnostics

What if the residual plot is not perfect?

Statisticians are trained to look for the following things:

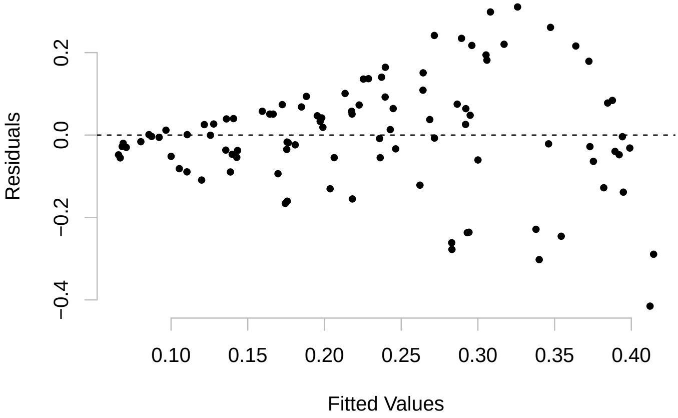
1. Outliers
2. A U-shaped pattern (curved residuals)
3. A fan shaped pattern (a change in the variance of the residuals for different values of $\hat{y}$)
4. A non-normal distribution of the residuals

Anscombe (1973) described the typical approaches:

- (2), (3), and (4) indicate possible response transformations.
- (2) indicates extra terms needed.

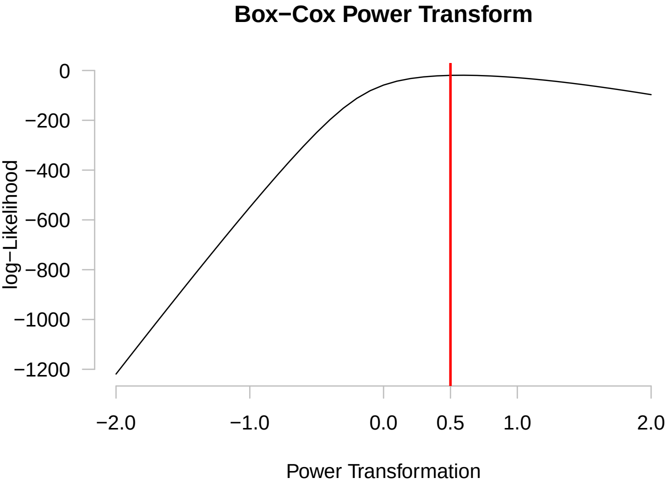
Residual Plots: Inference

What should we conclude in the following plot?



Residual Plots: Boxcox

Boxcox suggests a square-root transformation.



Residual Plots: Inference

Boxcox suggests a square-root transformation.

How the plot was constructed:

1. $\boldsymbol{x} = (x_1, x_2)$ was Pearson II distributed (Johnson, 1987).
2. $y|\boldsymbol{x} = |x_1|/[2 + (1.5 + x_2)^2] + e$
3. e was normally distributed.

The problem with inference:

> *If the form of the regression function is incorrect, then "the heteroscedastic pattern can be a manifestation of a nonlinear regression function or a non-constant variance function, or both, and we can conclude only that the homoscedastic linear regression model does not reasonably describe the data" (Cook, 1998, p. 5).*

Residual Plots: Inference

Why do wacky things happen???

The following always holds:

- If only one predictor → no loss of information in plot.
- If more than one predictor → loss of information in 2D plot.
- The residual versus fitted plot is essentially a projection.

What does a projection mean in this context?

Residual Plots: Inference

Why do wacky things happen???

Plotting the residuals versus the fitted values only works if...

1. the linear model is correct,
2. the covariance between y and x_j is not 0,
3. and the objective function is convex.

See Cook (1994).

Residual Plot Matrices

Researchers also look at scatterplot matrices of the residuals versus each predictor to determine if the model is correct.

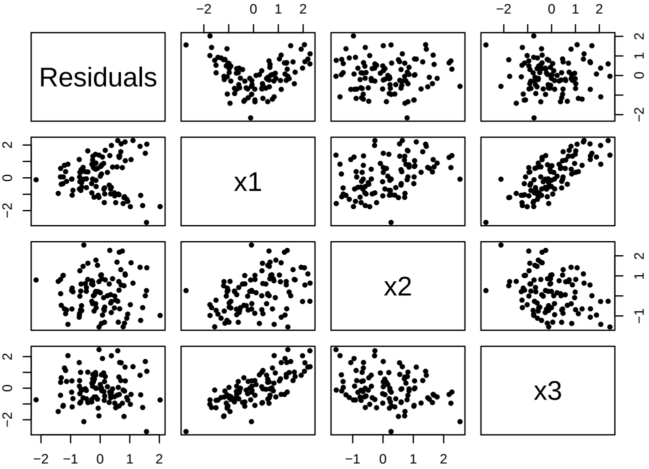
The standard interpretation:

> *"The presence of a curvilinear relationship suggests that a higher-order term, perhaps a quadratic in the explanatory variable, should be added to the model" (Atkinson, 1985, p. 3, as cited in Cook, 1998, p. 6).*

So, curvilinear $\implies$ add a term.

Residual Plot Matrices: Example

What's the problem here? What should we do?



Residual Plot Matrices: Example

Based on the plot, there is...

1. no relationship in the residual plot for x_2 and x_3,
2. a curvilinear relationship in the residual plot for x_1,
3. so we should restrict our attention to x_1 and transform.

Residual Plot Matrices: Example

How the plot was constructed:

1. $\boldsymbol{x} = (x_1, x_2, x_3)$ was normally distributed.
2. $y|\boldsymbol{x} = |x_2 + x_3| + e$
3. e was normally distributed.

Therefore, x_1 wasn't even needed in the regression. The plot was misleading because the functional form was misspecified and the predictors were correlated.

Diagnosing Problems: What to Do?

Residual plot matrices are problematic if the predictors are correlated. We should remove the correlation.

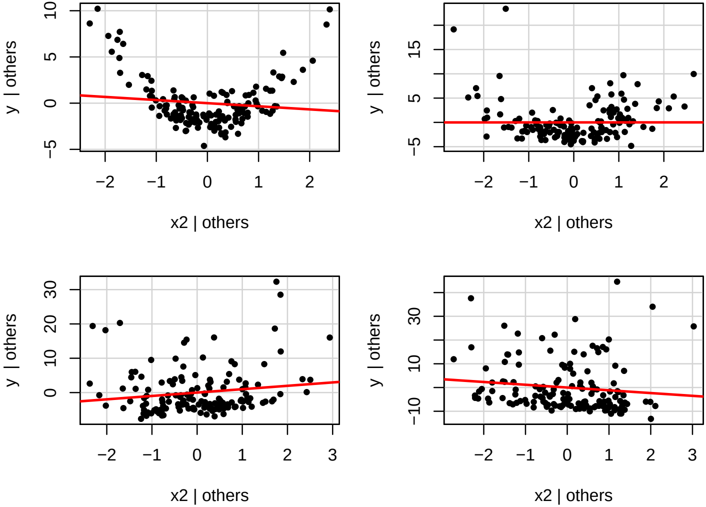
Added Variable Plots:

1. Pick the predictor of interest, x_1.
2. Regress y on $x_2, \ldots, x_p$, pull out error $e_{y|x_2,\ldots,x_p}$.
3. Regress x_1 on $x_2 \ldots, x_p$, pull out error $e_{x_1|x_2,\ldots,x_p}$.
4. Plot $e_{y|x_2,\ldots,x_p}$ versus $e_{x_1|x_2,\ldots,x_p}$.

The form of the added variable plot suggests transformations.

Added Variable Plots

Here are four added variable plots. How should we transform?



Added Variable Plots: Inference

Problem: For AVPs to work well, the regression of $e_{y|x_2,\ldots,x_p}$ on $e_{x_1|x_2,\ldots,x_p}$ needs to be of the same (or a similar) form as the predictor transformation (see Cook, 1996).

The form depends on the marginal distribution of the predictors and on the transformation needed. The AVP is biased toward suggesting linearity unless the predictors are not-highly correlated.

Added Variable Plots: Inference

In the example above (see Cook, 1996)...

1. $y|\boldsymbol{x} = x_1 + 2x_2^2 + e$,
2. e was normally distributed,
3. and $r_{x_1,x_2} = 0, .5, .75, .9$.

Multicollinearity is a problem.

Component Plus Residual Plots

Weisberg (1985) and Cook (1996) suggest Component Plus Residual plots work better than Added Variable plots.

Component Plus Residual Plots:

1. Pick the predictor of interest, x_1.
2. Regress y on $x_1, \ldots, x_p$, pull out error e.
3. Find C + R $= \hat{b}_1 x_1 + e$.
4. Plot C + R versus x_1.

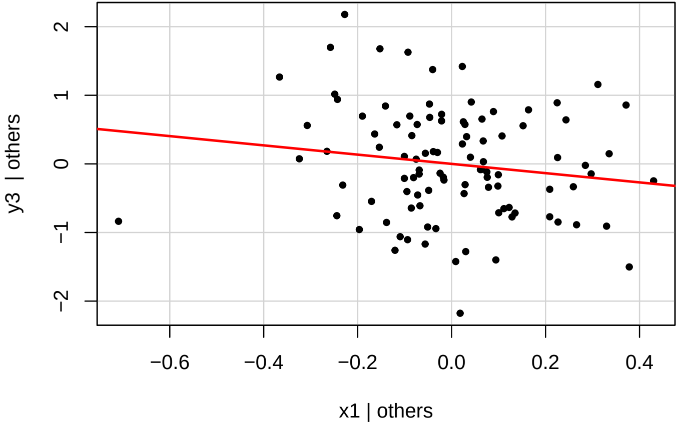
An Odd Problem

Now we run into an odd problem.

1. Two Examples Ago:
   - x_1 wasn't needed in the regression.
   - Residual plot matrices said x_1 needed a transform.
   - C + R plots said the same thing.
   - The AVP said x_1 **didn't** need a transform (next slide).
2. Last Example:
   - x_2 needed a transform.
   - The AVP said x_2 possibly **didn't** need a transform.
   - C + R plots **easily** detected the transform (two slides).

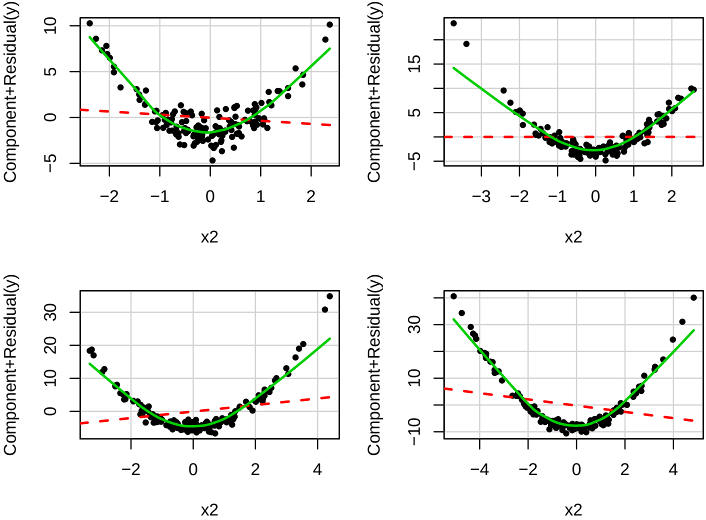
Added Variable Plots: Working!

Two examples ago, if we would have used an av-plot:



Component Plus Residual Plots: Working!

Last example, if we would have used a cr-plot:



Marginal Model Plots

Marginal Model Plots:

1. Plot y on the y-axis.
2. Ghost plot $\hat{y}$ on the y-axis.
3. Plot a linear combination of the predictors on the x-axis.
   - Usually plot $\hat{y}$ or x_j where $j = 1, \ldots, p$.
4. Smooth y versus the linear combination of the predictors.
5. Smooth $\hat{y}$ versus the linear combination of the predictors.
6. If model is correct → smooths should align!

Problem: It can only tell you that there **is** a problem (see Cook, 1998, pp. 317-327).

Suggestions

So, what should you do??

If the model is not correct, then graphical checks will conflate model correctness and violations of other assumptions.

Make sure that the linear model is appropriate before you begin to check any of the other assumptions.

References I

- Anscombe, F. J. (1973). Graphs in statistical analysis. *The American Statistician*, *27*, 17-21.
- Atkinson, A. (1985). *Plots, Transformations, and Regression.* Oxford, UK: Oxford University Press.
- Cook, D. (1994). On the interpretation of regression plots. *Journal of the American Statistical Association*, *89*, 177-189.
- Cook, D. (1996). Added-variable plots and curvature in linear regression. *Technometrics*, *38*, 275-278.
- Cook, D. (1998). *Regression Graphics: Ideas for Studying Regressions through Graphics*. New York, NY: John Wiley & Sons.
- Johnson, M. E. (1987). *Multivariate Statistical Simulation*. New York, NY: John Wiley & Sons.

References II

- Weisberg, S. (1985). *Applied Linear Regression*. New York, NY: John Wiley & Sons.