Basic Regression Diagnostics[1]

STA 441 Spring 2024

[1]This slide show is an open-source document. See last slide for copyright information.

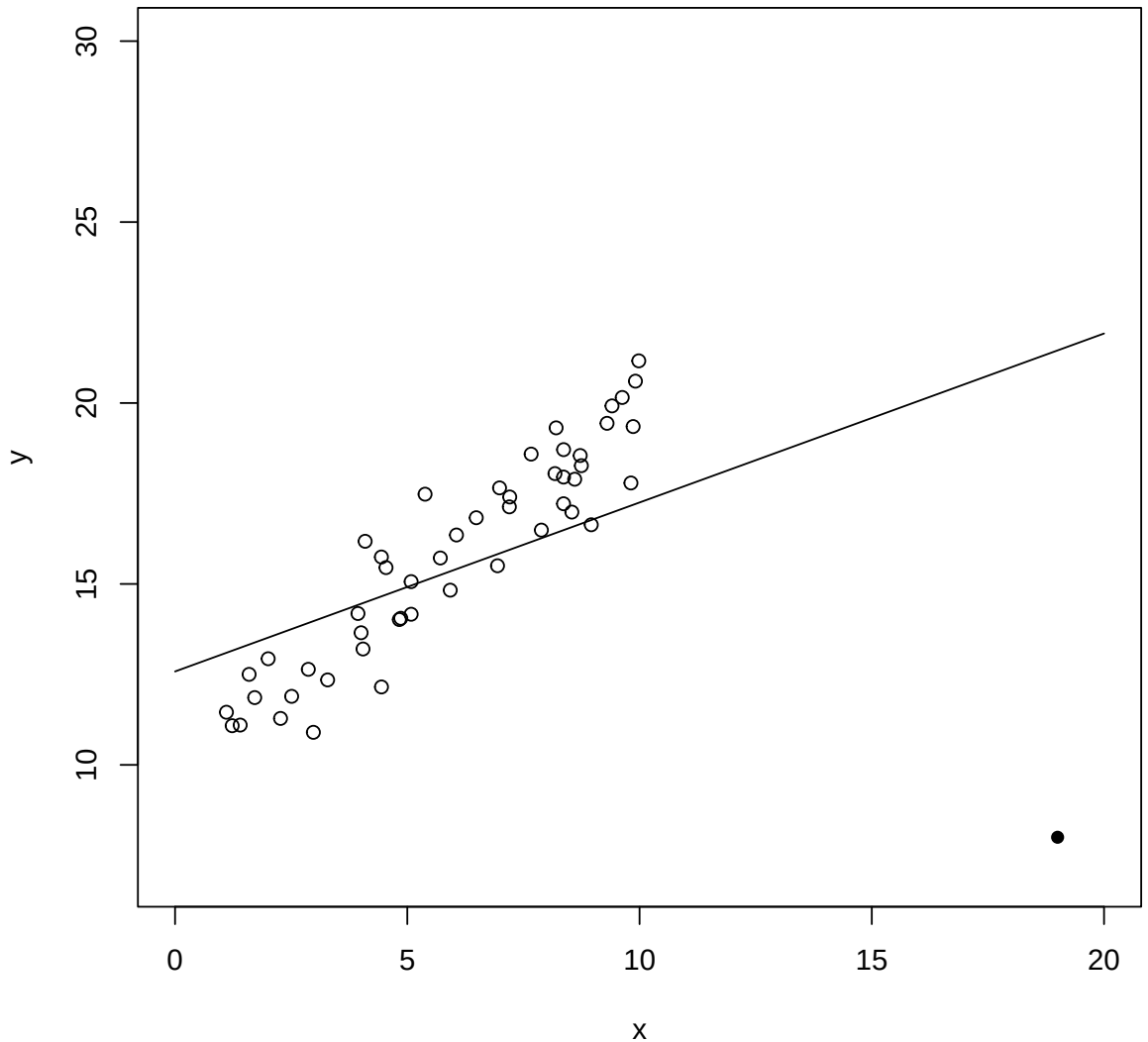
Overview

This description of regression diagnostics is very basic

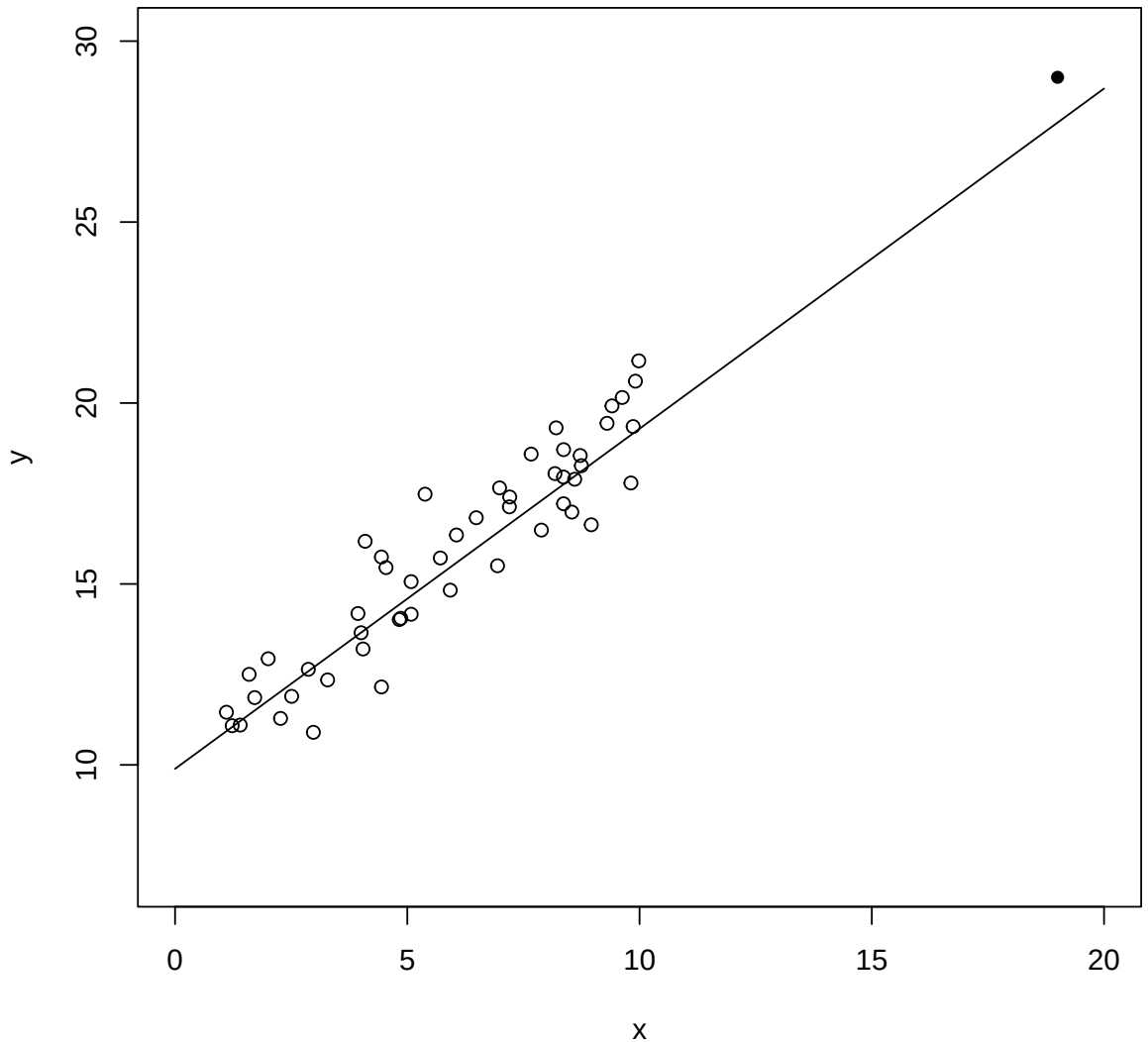
For a more complete and technical discussion, see

http://www.utstat.toronto.edu/brunner/oldclass/302f20/lectures/302f20Diagnostics.pdf

Outliers in x can be influential



Or sometimes not



Hat Values: Also called "Leverage" values

- For the record, the "hat matrix" is $\mathbf{H} = [h_{ij}] = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$
- Hat values, denoted h_{ii}, are the diagonal elements of this $n \times n$ matrix.
- Big hat values indicate (multivariate) outliers in the x variables.
- So they could be influential.

Small hat values are good

Denote a residual by $e_i = y_i - \widehat{y}_i$

- $0 \leq h_{ii} \leq 1$.
- $\sum_{i=1}^{n} h_{ii} = p$ (the number of betas), so most hat values go to zero as $n \to \infty$.
- $Var(e_i - \epsilon_i) = \sigma^2 h_{ii}$. For large samples, this variance is very small for most of the cases, and e_i is probably close to ϵ_i.
- This is a justification for residual analysis. in which we check whether e_i have the assumed properties of ϵ_i.

Robustness to normality

Robust means strong

- If $\lim_{n\to\infty} \max_i h_{ii} = 0$, then the distribution of $\widehat{\boldsymbol{\beta}}$ approaches a multivariate normal $N_p(\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1})$, even if the distribution of the ϵ_i is not normal.

- In this case, tests and confidence intervals based on the normal distribution are roughly okay for large samples.

What is a "small" hat value?

- Rule of thumb: Worry about $h_{ii} > \frac{3p}{n}$ (or maybe $\frac{2p}{n}$).
- Another rule of thumb (for multivariate normality of $\widehat{\boldsymbol{\beta}}$) is worry about $h_{ii} > 0.2$.
- Or just look at a histogram of hat values.

What to do about large hat values?

- Investigate. Maybe it's a data error.
- If the largest hat value is greater than 0.2, don't be so casual about normality.
- Do the cases with large hat values also have large residuals?
- Try temporarily setting the cases aside. Do the conclusions change?

Residuals: $e_i = y_i - \widehat{y}_i$

$$\begin{aligned} e_i &= y_i - \widehat{y}_i \\ y_i &= \widehat{y}_i + e_i \\ &= b_0 + b_1 x_{i1} + b_2 x_{i2} + \cdots + b_{p-1} x_{i,p-1} + e_i \end{aligned}$$

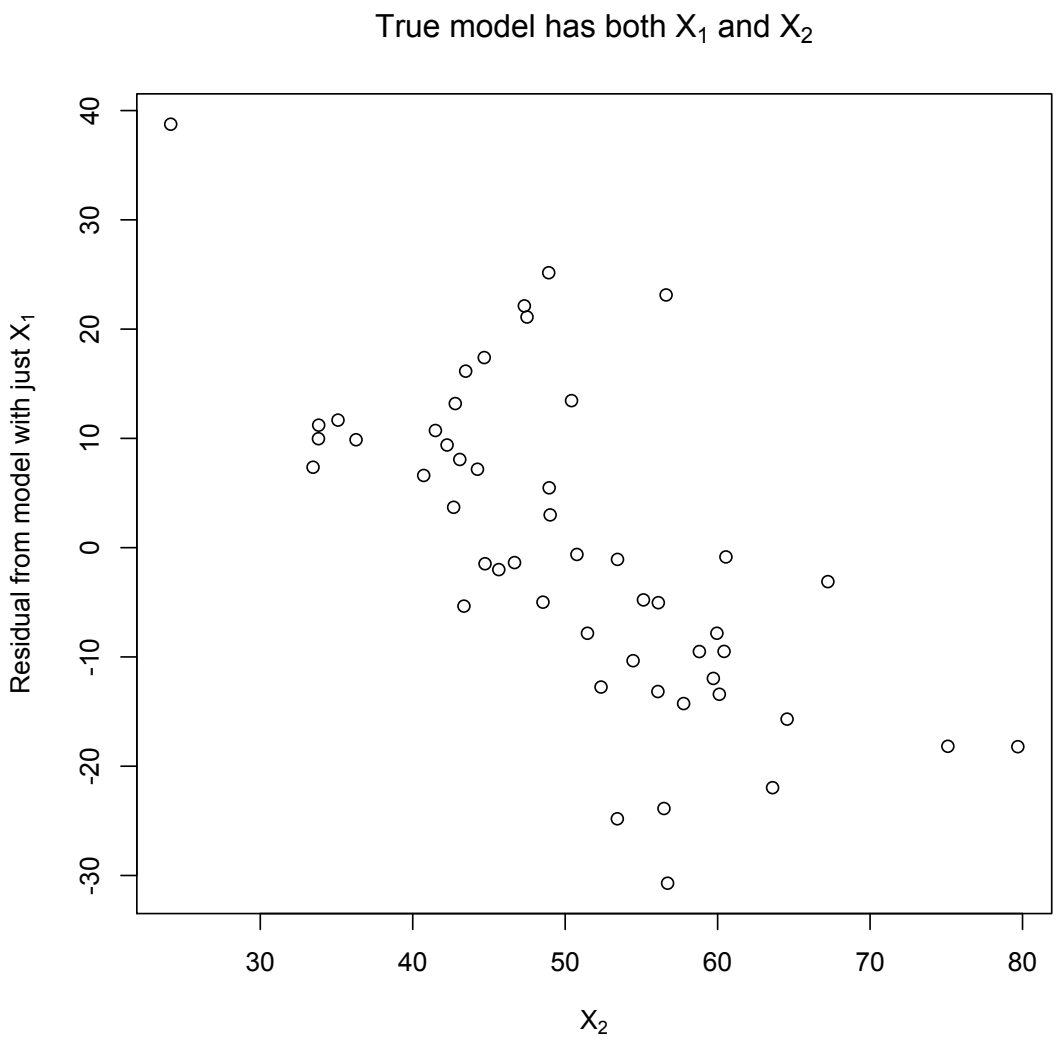
So we have

$$\begin{aligned} y_i &= b_0 + b_1 x_{i1} + b_2 x_{i2} + \cdots + b_{p-1} x_{i,p-1} + e_i \\ y_i &= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_{p-1} x_{i,p-1} + \epsilon_i \end{aligned}$$

Residuals: $e_i = y_i - \widehat{y}_i$

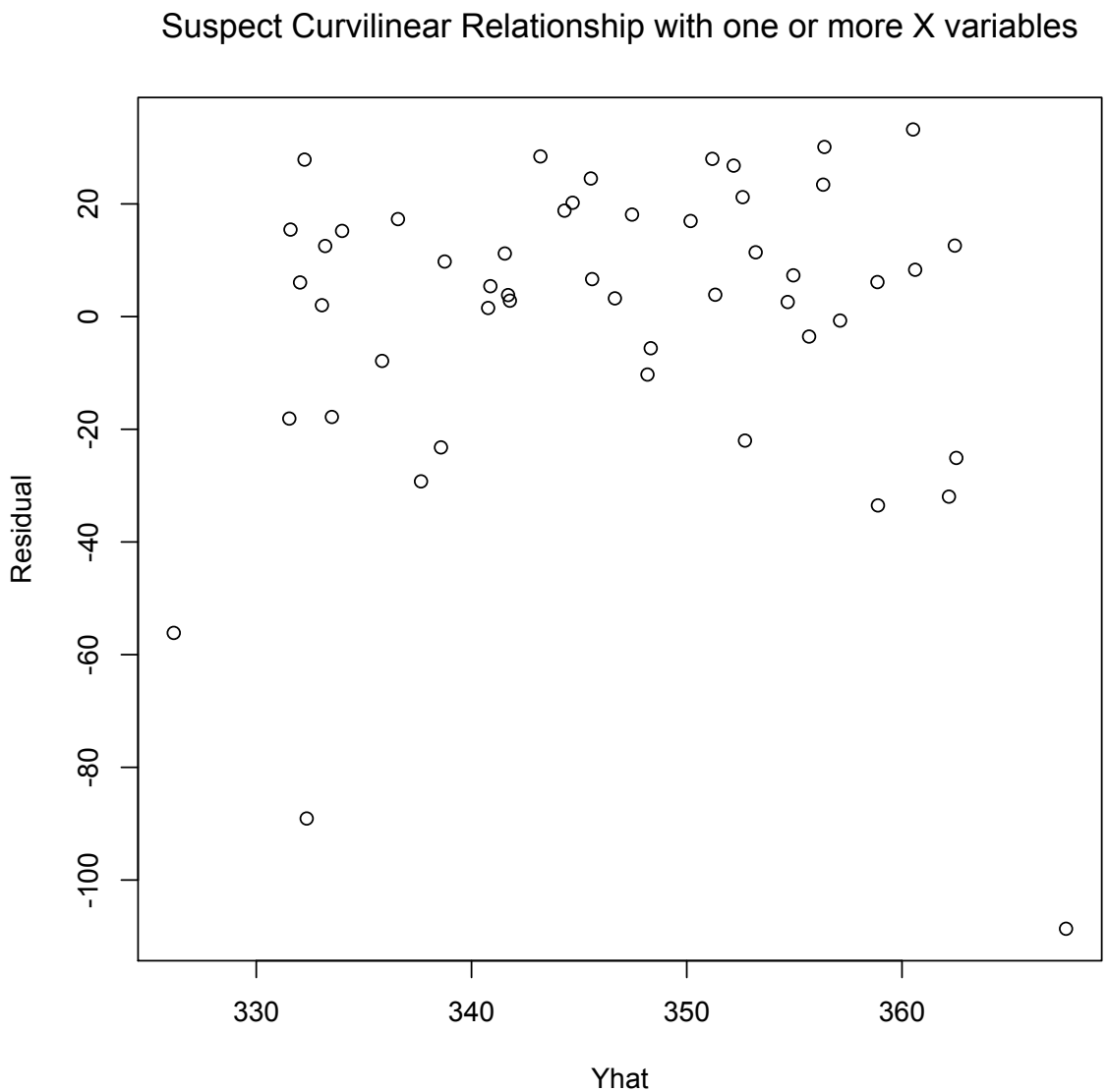
- Investigate points that are unusually far from the regression plane.
- Residual plots can reveal a lot.
  - Against predicted y.
  - Against explanatory variables not in the equation.
  - Against explanatory variables in the equation.
  - Against time.
  - Look for serious departures from normality, outliers.

Plot Residuals Against Explanatory Variables Not in the Equation



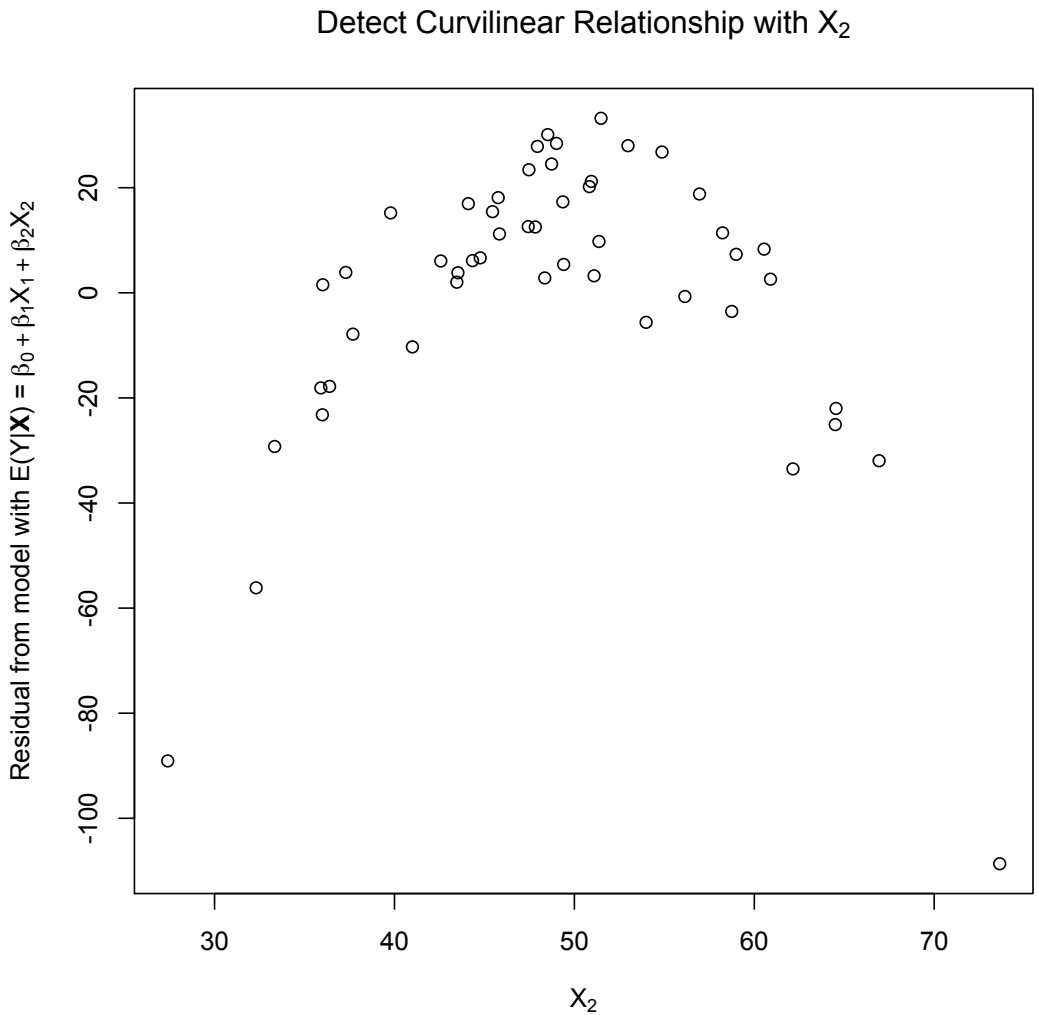
Plot Residuals Against $\widehat{y}$

e_i and $\widehat{y}$ should be independent



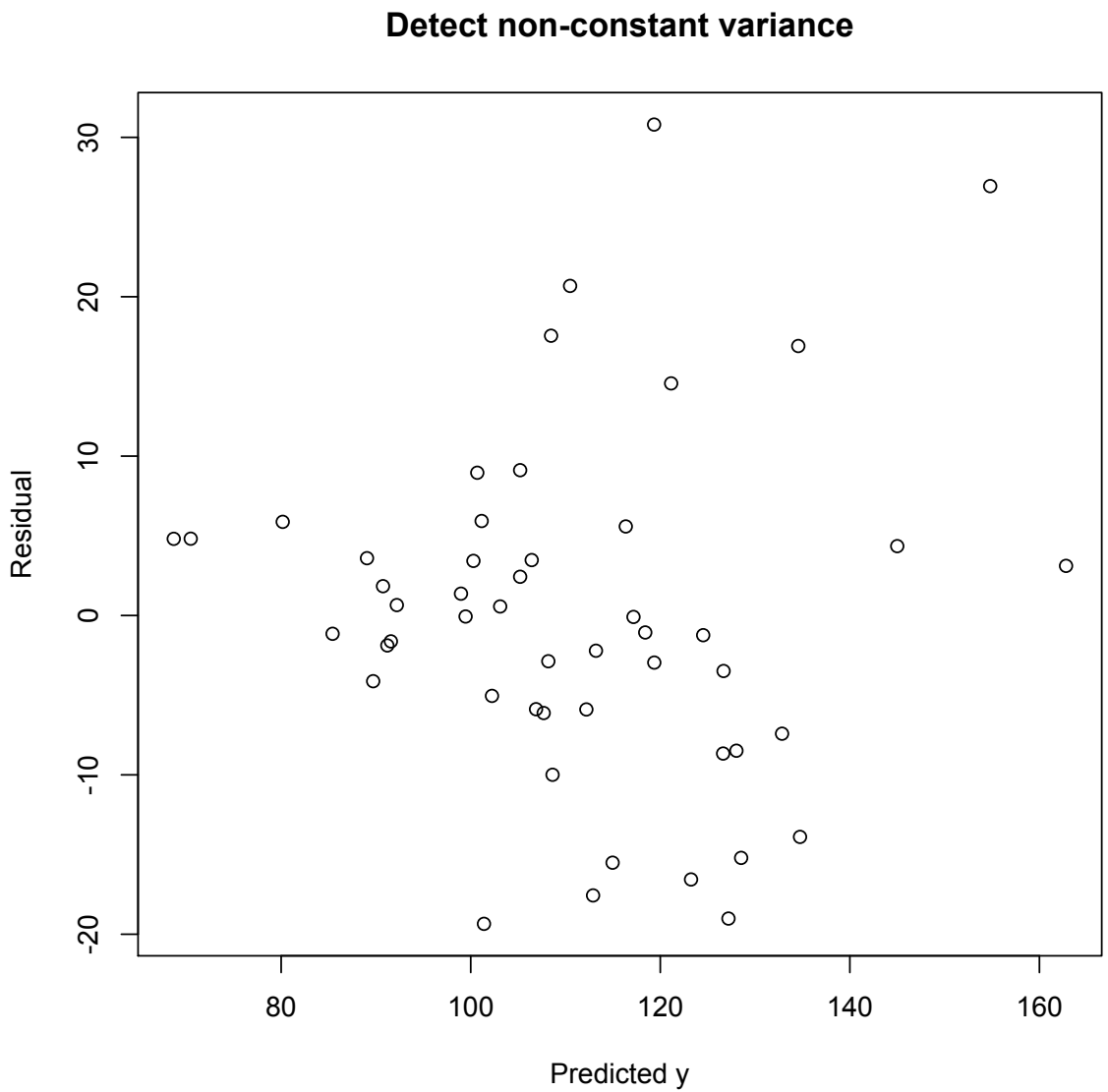
Plot Residuals Against Explanatory Variables in the Equation

Plot versus X_1 showed nothing



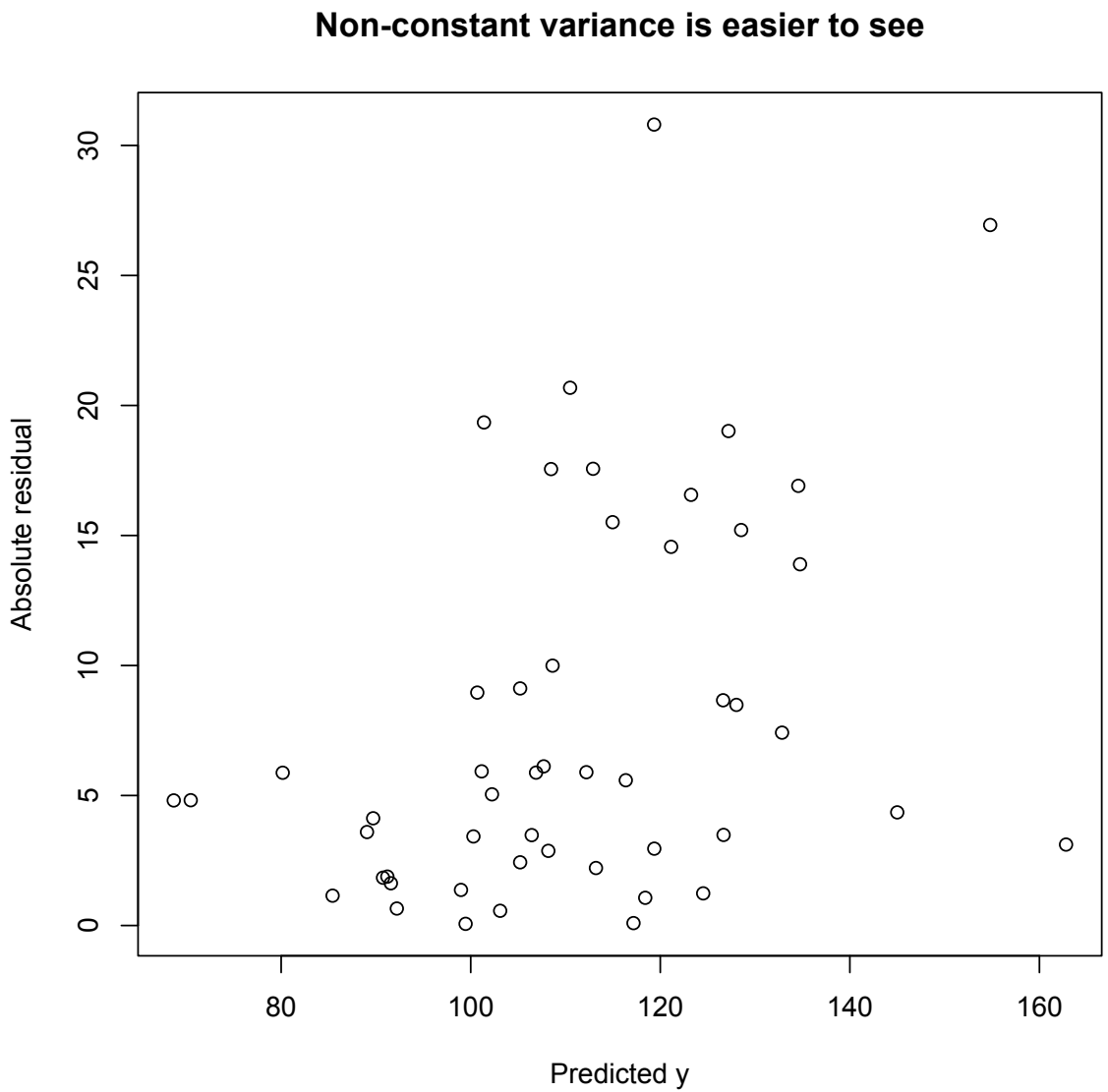
Plot Residuals Against Predicted y

Can show non-constant variance



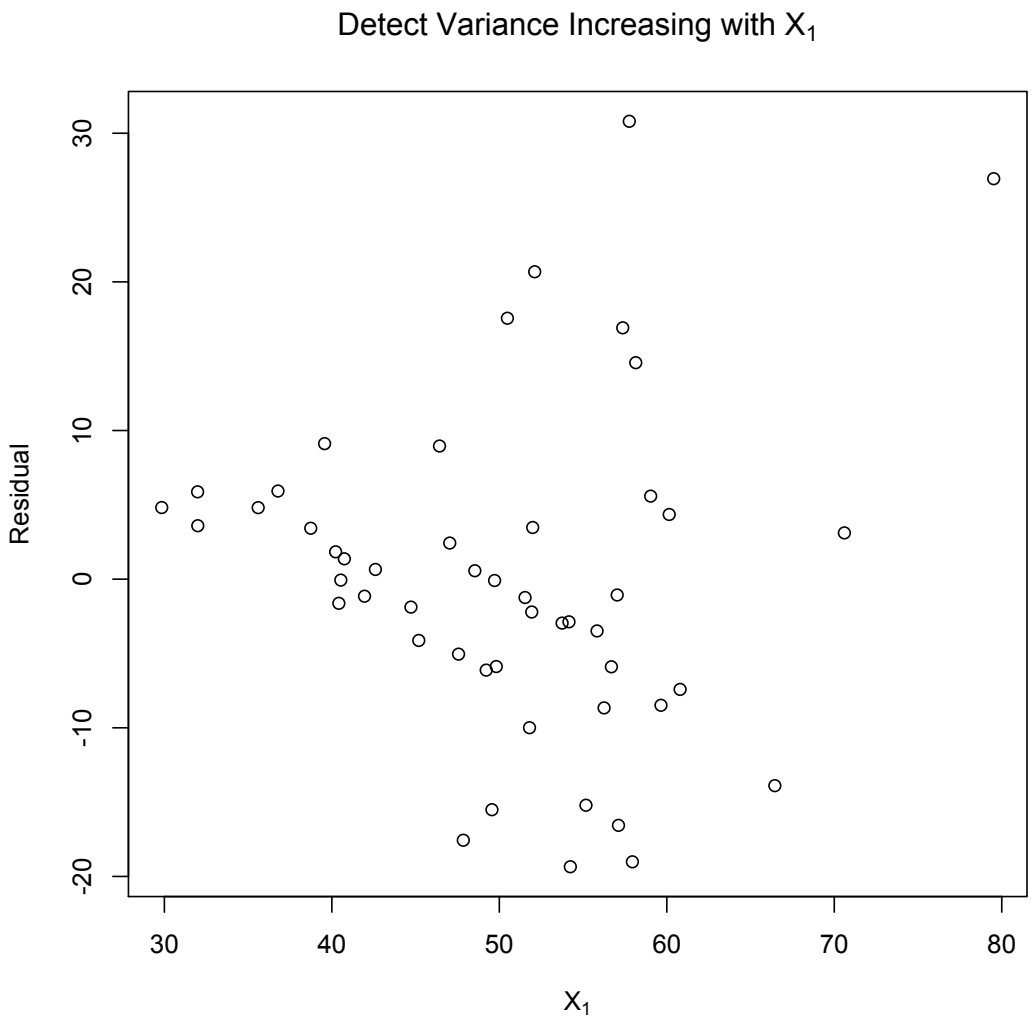
Plot **Absolute** Residuals Against Predicted y

Looking for non-constant variance



Plot Residuals Against Explanatory Variables in the Equation

Can show non-constant variance



Plot Residuals Against Time, if the data are time ordered

- You really need to watch out for time ordered data.
- Standard regression methods may not be appropriate.
- The problem is that ϵ represents all other variables that are left out of the regression equation.
- Some of them could be time dependent.
- This would make the ϵ_i non-independent, possibly yielding misleading results.
- Check the Durbin-Watson statistic.

Outlier detection

- Big residuals may be outliers. What's "big?"
- Standardized residuals
- Deleted residuals
- Various combinations
- Studentized deleted residuals.

Studentized deleted residuals

The idea

- Calculate $\widehat{y}$ for each observation based on all the other observations, but not that one. Leave one out.
- Predict each observed y based on all the others, and assess error of prediction (divided by standard error).
- Big values suggest that the expected value of y_i is not what it should be.
- Studentized deleted residuals have t distributions if the model is correct.
- Treat as test statistics.

Studentized deleted residual

$$t_i = \frac{y_i - \widehat{y}_{(i)}}{se_{(i)}} = \frac{y_i - \mathbf{x}_i'\widehat{\boldsymbol{\beta}}_{(i)}}{se_{(i)}} \sim t(n-1-p)$$

- If the model is correct, numerator has expected value zero.
- Use t_i as a test statistic. If H_0 is rejected, investigate.
- We are doing n tests.
- If all null hypotheses are true (no outliers), there is still a good chance of rejection at least one H_0.
- Type I errors are very time consuming and disturbing.
- How about a Bonferroni correction?

Bonferroni Correction for Multiple Tests

- Do the tests as usual, obtaining n test statistics.
- Could multiply the p-values by n.
- Easier is to use the critical value $t_{\frac{\alpha}{2n}}(n-1-p)$.
- Even for large n it is not overly conservative.
- If you locate an outlier, **investigate**!

Normality

- Don't worry too much if maximum hat value is less than 0.2.
- Instead of checking the residuals for normality, I like to check the Studentized deleted residuals.
- Their variances are all equal.
- And for a healthy sample size, t is almost z.
- Look at a histogram. There are also formal tests in `proc univariate normal`.

Copyright Information