

Graphical Diagnosis

Paul E. Johnson¹ ²

¹Dept of Political Science

²Dept. of Psychology, University of Kansas

.

Outline

- 1 Introduction
- 2 Start Simple: Scatterplot
- 3 Standard Diagnostic Plots
- 4 Decipher Each Type Of Plot
 - Top Left: Residual Plot
 - Top Right: Q-Q plot
 - Bottom Left: Scale-Location Plot
 - Bottom-Right: Leverage, Cook's Distance
- 5 Repeat with more Predictors
 - Multiple Regression Fitted
 - Termplot
 - Diagnostic Plots
- 6 Stress Test These Diagnostics
 - Bad Nonlinearity

Outline ...

- Add Some “Outliers”
- Heteroskedasticity

7 Practice Problems

Outline

- 1 Introduction
- 2 Start Simple: Scatterplot
- 3 Standard Diagnostic Plots
- 4 Decipher Each Type Of Plot
 - Top Left: Residual Plot
 - Top Right: Q-Q plot
 - Bottom Left: Scale-Location Plot
 - Bottom-Right: Leverage, Cook's Distance
- 5 Repeat with more Predictors
 - Multiple Regression Fitted
 - Termplot
 - Diagnostic Plots
- 6 Stress Test These Diagnostics
 - Bad Nonlinearity
 - Add Some "Outliers"
 - Heteroskedasticity

You assumed

- 1 Linearity: $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + e_i$
- 2 Assumptions about Error term e_i . Either:

e_i is <i>Normal</i> (0, σ_e^2)	or this:	<ol style="list-style-type: none">1 Unbiased errors: $E[e_i] = 0$2 Homoskedasticity: $E[e_i^2] = \sigma_e^2$
---	----------	---
- 3 No “autocorrelation” between error terms, $E[e_i \cdot e_j] = 0$ for all i and j
- 4 No correlation between x 's and the error term, $E[x_{ji} \cdot e_i]$ for variables j and cases i

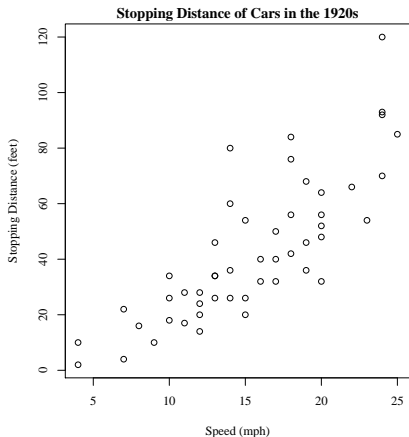
The Rest of This Course Is Diagnostics and Remedies

- Decide how inappropriate the results from the linear model and OLS might be
- If inappropriate, 3 major options
 - 1 Choose a different formula for y_i or e_i (or both)
 - nonlinear model for y_i
 - Weighted Least Squares (WLS) or Generalized Least Squares (GLS)
 - 2 Keep the same formula and estimate in a different way
 - robust regression
 - 3 Keep the same estimates but apply a post hoc correction (e.g., robust “heteroskedasticity consistent” standard errors for parameter estimates)

Outline

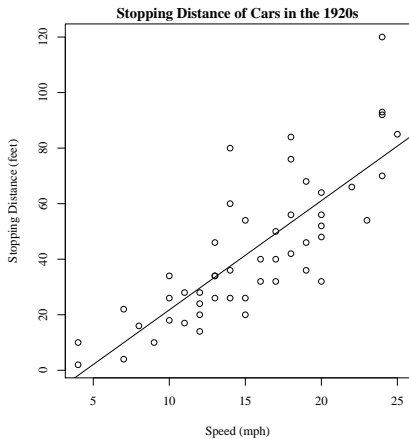
- 1 Introduction
- 2 Start Simple: Scatterplot**
- 3 Standard Diagnostic Plots
- 4 Decipher Each Type Of Plot
 - Top Left: Residual Plot
 - Top Right: Q-Q plot
 - Bottom Left: Scale-Location Plot
 - Bottom-Right: Leverage, Cook's Distance
- 5 Repeat with more Predictors
 - Multiple Regression Fitted
 - Termplot
 - Diagnostic Plots
- 6 Stress Test These Diagnostics
 - Bad Nonlinearity
 - Add Some "Outliers"
 - Heteroskedasticity

Scatterplot: Two Numeric Variables



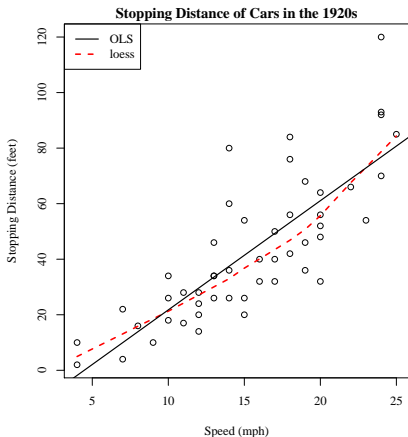
- If there is only one predictor, the best diagnostic might be the simple scatterplot.
- Look for
 - linearity
 - homoskedasticity
- This is R's "cars" data set, a set commonly used for illustrations

Superimpose a Regression Line



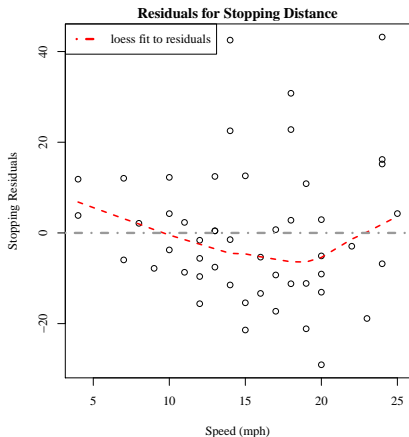
- Straight line may be OK

Superimpose a Loess Line



- Loess: locally weighted error sum of squares regression
- Fits a regression model individually for each point in data!
- Puts less weight on “far away” observations
- Predicted values are a “connect the dots” line, smoothed graphically to look pleasant.

Plot residuals against X



- Loess: locally weighted error sum of squares regression
- Evaluate subjectively (!) or (?)
- Hints about how model might be redesigned to fit the data more accurately

Outline

- 1 Introduction
- 2 Start Simple: Scatterplot
- 3 Standard Diagnostic Plots**
- 4 Decipher Each Type Of Plot
 - Top Left: Residual Plot
 - Top Right: Q-Q plot
 - Bottom Left: Scale-Location Plot
 - Bottom-Right: Leverage, Cook's Distance
- 5 Repeat with more Predictors
 - Multiple Regression Fitted
 - Termplot
 - Diagnostic Plots
- 6 Stress Test These Diagnostics
 - Bad Nonlinearity
 - Add Some "Outliers"
 - Heteroskedasticity

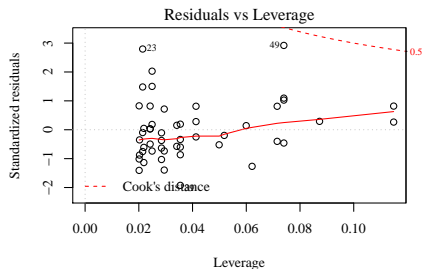
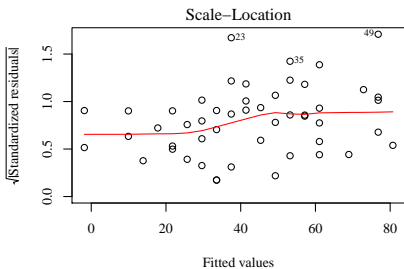
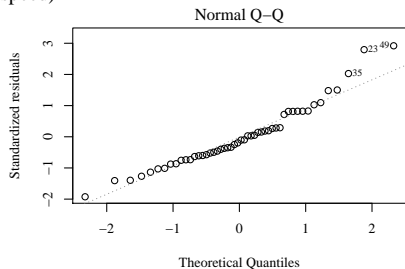
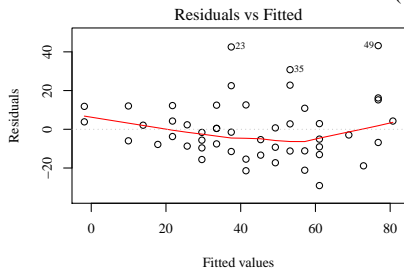
The R Diagnostic Plot Series

- In R, one can fit a model and then ask for the standard diagnostic plot:

```
mod1 <- lm(output ~ x1 +x2+x3+x4+x5, data=dat)
plot(mod1)
```

- plot is a “generic function”, does lots of different things, depending on what you give it.

lm(dist ~ speed)

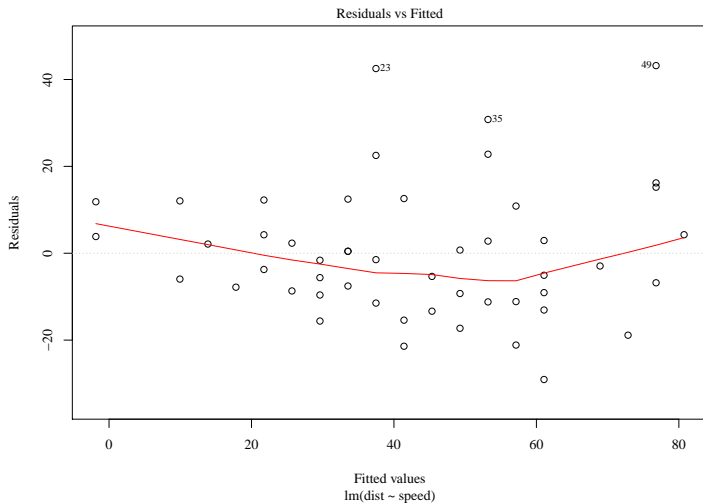


The 2x2 diagnostic plot matrix for the cars regression

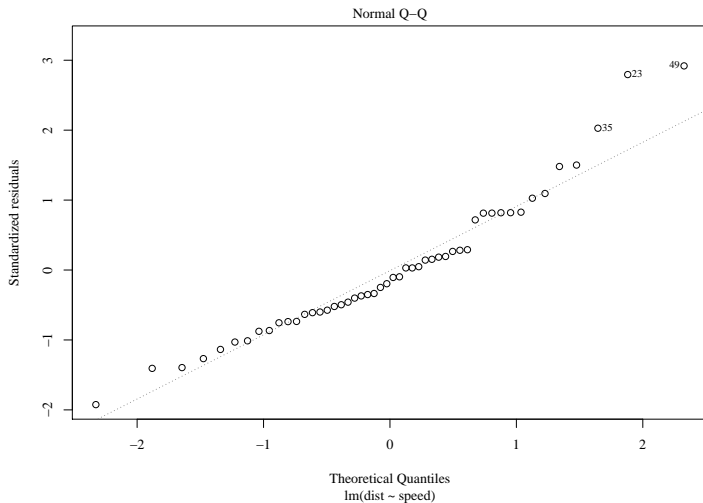
Outline

- 1 Introduction
- 2 Start Simple: Scatterplot
- 3 Standard Diagnostic Plots
- 4 Decipher Each Type Of Plot**
 - Top Left: Residual Plot
 - Top Right: Q-Q plot
 - Bottom Left: Scale-Location Plot
 - Bottom-Right: Leverage, Cook's Distance
- 5 Repeat with more Predictors
 - Multiple Regression Fitted
 - Termplot
 - Diagnostic Plots
- 6 Stress Test These Diagnostics
 - Bad Nonlinearity
 - Add Some "Outliers"
 - Heteroskedasticity

Residuals with Loess Line

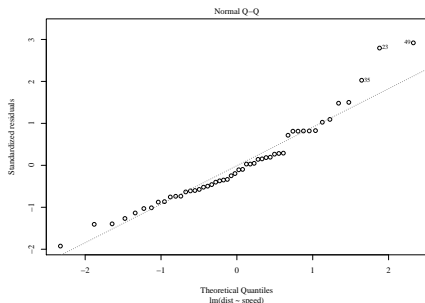


QQ plot



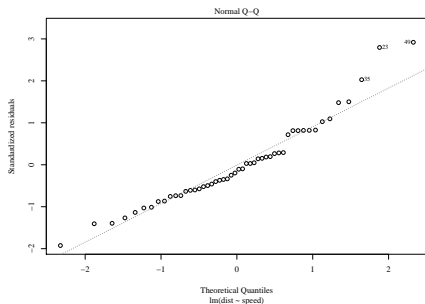
QQ plot

- Standardized (or “Studentized”) residuals
- Standardization intended to put residuals onto scale of their true variance (at a particular value of x_i)
- Proceed with assumption that these residuals are drawn from $N(0, 1)$

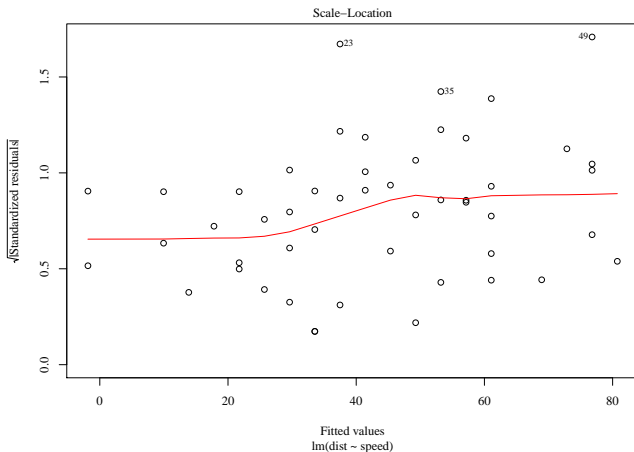


QQ Compare Residuals against Normal(0,1) CDF

- Recall Normal CDF tells us how likely each value is “theoretically” supposed to be.
- QQ plot matches theoretical distribution with observed distribution
- If all points are exactly on the line, then the observed distribution matches $N(0,1)$
- If points deviate above or below the line, we suspect error is not normal

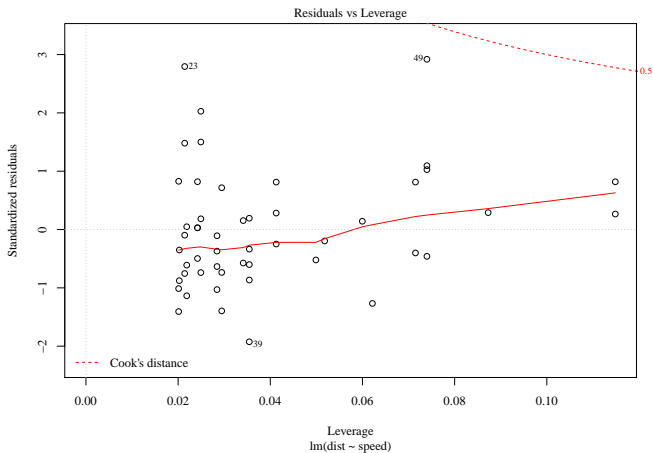


Scale-Location



Should be a homogeneous cloud, that is not taller on one side than the other

Review One At a Time



Leverage: Outlier Diagnostics

- Leverage: Case-by-case estimate of a case's potential for influence on predicted values (not just its own predicted value, but predictions for other cases).
- Recall predicted values are $\hat{y} = \{\hat{y}_1, \hat{y}_2, \dots, \hat{y}_N\}$
- It can be shown that the predicted value can be calculated as a linear combination of the observations y_i like so:

$$\hat{y}_j = h_{j1}y_1 + h_{j2}y_2 + h_{j3}y_3 + h_{j4}y_4 + \dots + h_{j(N-1)}y_{j(N-1)} + h_{jN}y_{jN}$$

(The prediction for the j' th case is a weighted sum of all observed y_i).

- The coefficients h_{ji} are from a thing called the “hat matrix”
- Intuition: In a perfect world, no observation exerts a “huge influence” on the predictions, the h_{ji} weights are all roughly the same (and will average out to $1/N$).

QQ Compare Residuals against Normal(0,1) CDF

Cook's Distance

- Cook's Distance. I interpret this as a weighted summary one case's impact on slope coefficient estimates.
 - Fit the model with all of the cases, get the regression slopes in a vector $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k)$.
 - Exclude a "row" (case) j , estimate the regression, and calculate the "leave j out estimates", $\hat{\beta}_j = (\hat{\beta}_{0j}, \hat{\beta}_{1j}, \dots, \hat{\beta}_{kj})$.
 - Square the differences between $\hat{\beta}$ and $\hat{\beta}_j$ and add them up, inserting a weighting formula that Cook proposed.
 - The end result is interpreted as a "change in the predicted value for all cases caused by case j "
- Will study more later when we do "regression diagnostics"

Outline

- 1 Introduction
- 2 Start Simple: Scatterplot
- 3 Standard Diagnostic Plots
- 4 Decipher Each Type Of Plot
 - Top Left: Residual Plot
 - Top Right: Q-Q plot
 - Bottom Left: Scale-Location Plot
 - Bottom-Right: Leverage, Cook's Distance
- 5 Repeat with more Predictors**
 - Multiple Regression Fitted
 - Termplot
 - Diagnostic Plots
- 6 Stress Test These Diagnostics
 - Bad Nonlinearity
 - Add Some "Outliers"
 - Heteroskedasticity

Occupational Prestige Data from John Fox's "car" package

```
library(car)
Prestige$income <- Prestige$income/10
presmod1 <- lm(prestige ~ income + education +
               women + type, data=Prestige)
```

My Professionally Acceptable Regression Table

	M1	
	Estimate	(S.E.)
(Intercept)	-0.814	(5.331)
income	0.104***	(0.026)
education	3.662***	(0.646)
women	0.006	(0.030)
typeprof	5.905	(3.938)
typewc	-2.917	(2.665)
N	98	
RMSE	7.132	
R^2	0.835	
adj R^2	0.826	

* $p \leq 0.05$ ** $p \leq 0.01$ *** $p \leq 0.001$

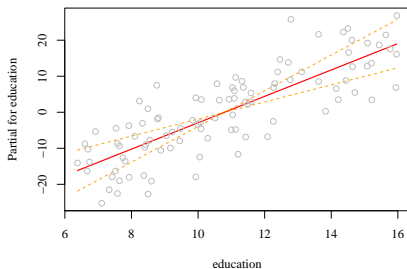
I'd usually run Termplot

Termplot is Multiple Regression Equivalent of Scatterplot

```
termplot(presmod1, se=T, partial=T)
```

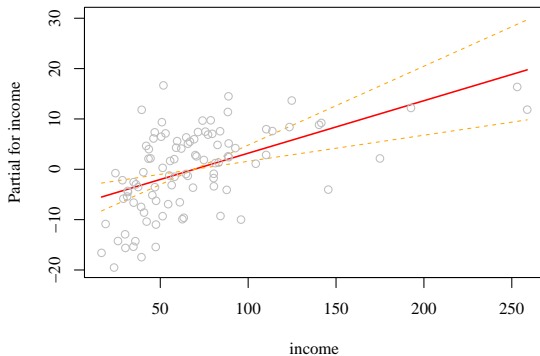
```
termplot(presmod1, se=T, partial.resid=T)
```

Education Termplot:



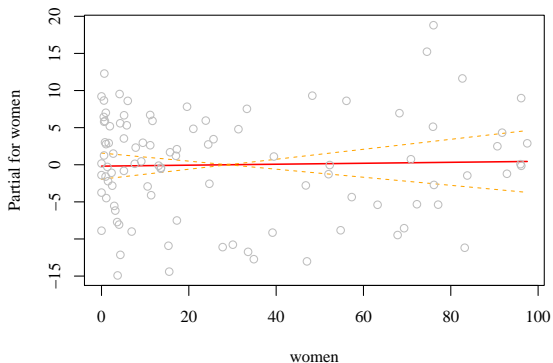
```
termplot(presmod1, se=T, partial.resid=T)
```

Income Termplot:



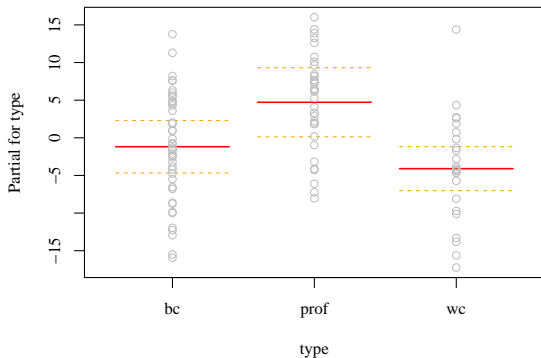
```
termplot(presmod1, se=T, partial.resid=T)
```

Women (percentage of members in field) Termplot:



```
termplot(presmod1, se=T, partial.resid=T)
```

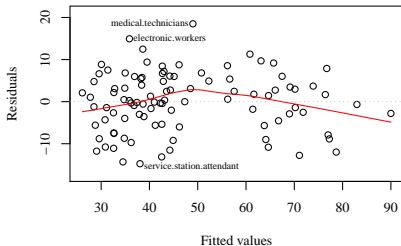
type Termplot:



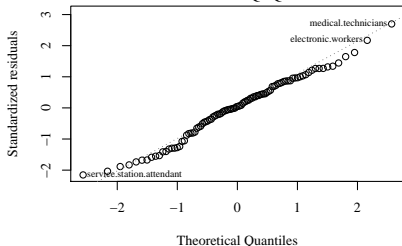
Plot Diagnostics

lm(prestige ~ income + education + women + type)

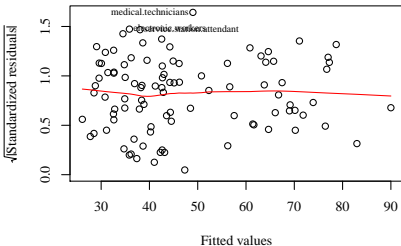
Residuals vs Fitted



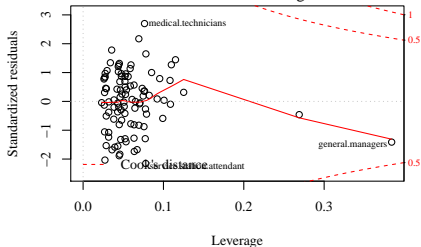
Normal Q-Q



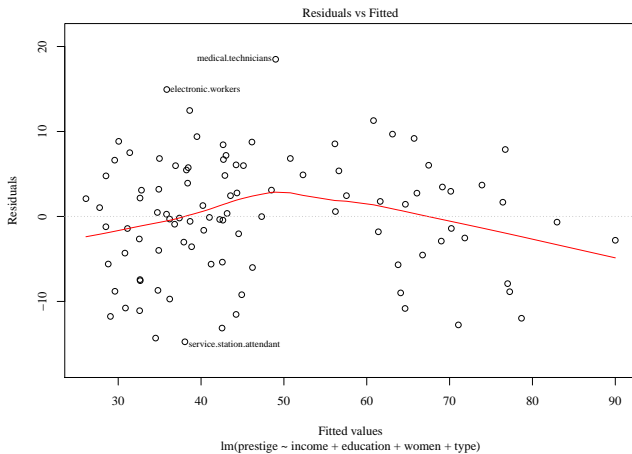
Scale-Location



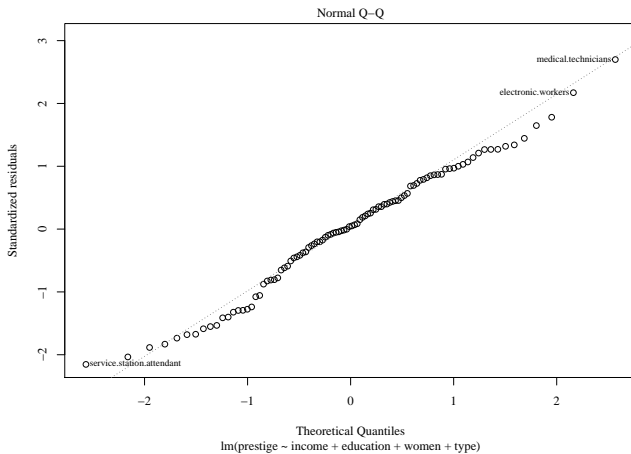
Residuals vs Leverage



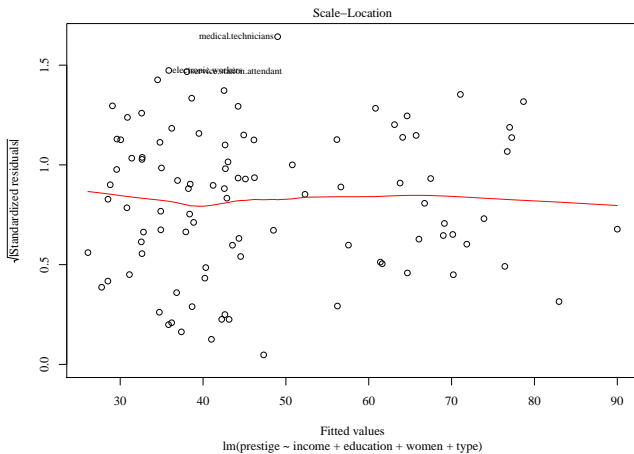
Diagnostic Plot 1



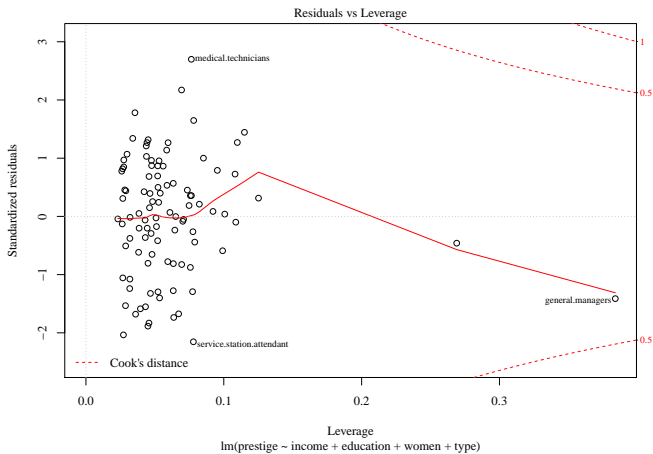
Diagnostic Plot 2



Diagnostic Plot 3



Diagnostic Plot 4



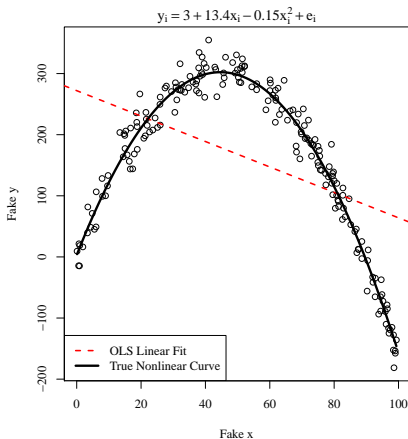
Outline

- 1 Introduction
- 2 Start Simple: Scatterplot
- 3 Standard Diagnostic Plots
- 4 Decipher Each Type Of Plot
 - Top Left: Residual Plot
 - Top Right: Q-Q plot
 - Bottom Left: Scale-Location Plot
 - Bottom-Right: Leverage, Cook's Distance
- 5 Repeat with more Predictors
 - Multiple Regression Fitted
 - Termplot
 - Diagnostic Plots
- 6 Stress Test These Diagnostics
 - Bad Nonlinearity
 - Add Some "Outliers"
 - Heteroskedasticity

Experience Required To Interpret Plots

- Usually (for me), a plot will
 - reveal a “really bad, obvious” problem
 - look “not grossly wrong.”
- Only way I can think of to “get some practice” is to make up data with flaws and then study the diagnostic plots.
- So I worked out a couple of experiments to illustrate visual effect of
 - nonlinearity
 - outliers

Demo with Manufactured Quadratic Relationship



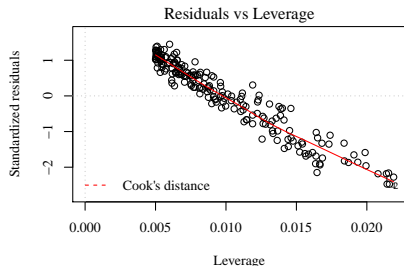
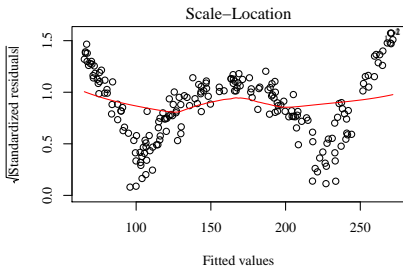
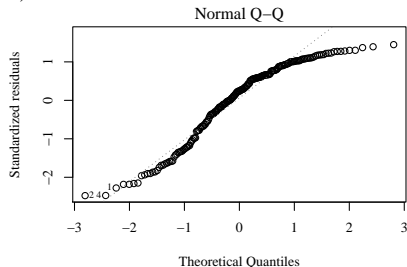
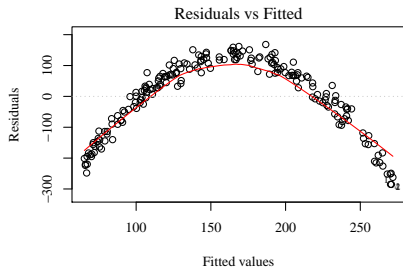
- The “true model” is a parabola (quadratic equation)

$$y_i = 3 + 13.4x_1 - 0.15x_i^2 + e_i$$

- And the error term is drawn from $N(\mu_e = 0, \sigma_e^2 = 22^2)$

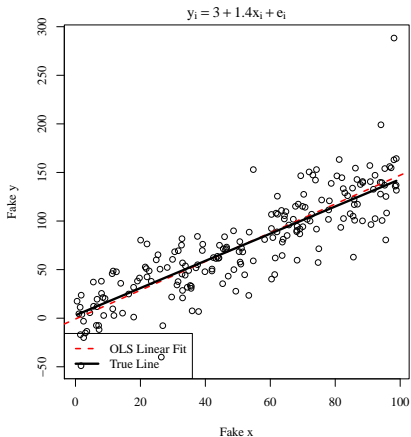
The Manufactured Quadratic Data

$\text{lm}(y \sim x)$



The 2x2 diagnostic plot matrix for the manufactured quadratic data

Demo with Manufactured Outliers



- The “true model” is a straight line

$$y_i = 3 + 1.4x_1 + e_i$$

- And the error term $e_i \sim N(\mu_e = 0, \sigma_e^2 = 22^2)$
- 10 Randomly Drawn Cases have magnified error $e_i = 4 \times e_i$
- The 10 “bad cases” are:

[1]	5	47	93	115	119	161
	162	166	185	196		

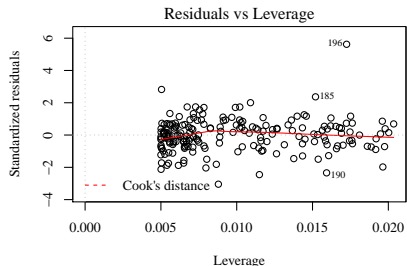
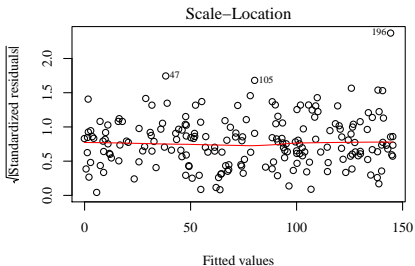
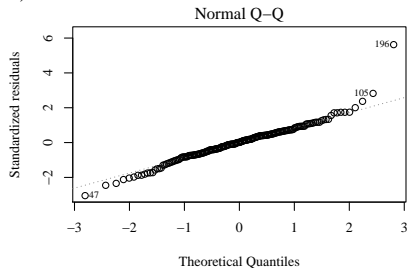
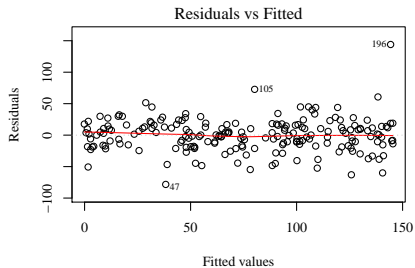
The Regression Table (just for the record)

	M1	
	Estimate	(S.E.)
(Intercept)	-0.845	(3.712)
x	1.480***	(0.062)
N	200	
RMSE	25.831	
R^2	0.742	

* $p \leq 0.05$ ** $p \leq 0.01$ *** $p \leq 0.001$

R Plot for lm With The 4× Manufactured Outlier Data

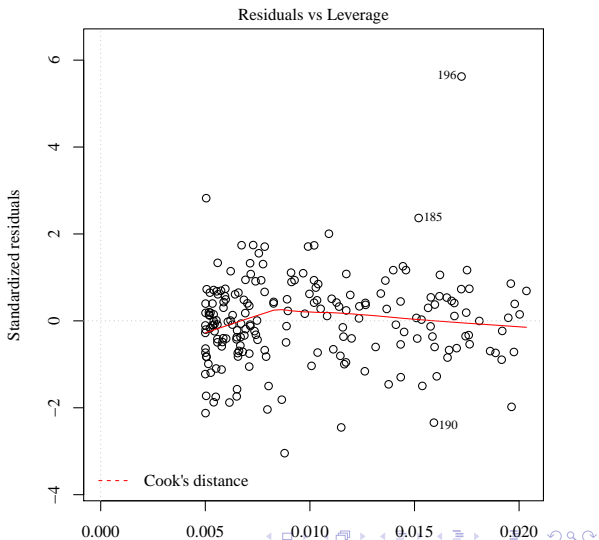
lm(y ~ x)



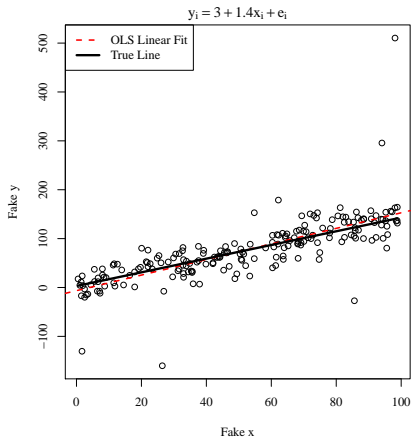
10 errors magnified by factor of 4

Concern: Leverage Plot Doesn't Notice Effect of Outliers

- I expected points would be in the extreme "bad Cook's distance" area
- I'm going to torture this until it "works" (should I say "breaks"?)



First Retry: Magnify the Outliers $\times 10$



- The “true model” is a straight line

$$y_i = 3 + 1.4x_1 + e_i$$

- And the error term $e_i \sim N(\mu_e = 0, \sigma_e^2 = 22^2)$
- 10 Randomly Drawn Cases have magnified error $e_i = 10 \times e_i$
- The 10 “bad cases” are:

[1]	5	47	93	115	119	161
	162	166	185	196		

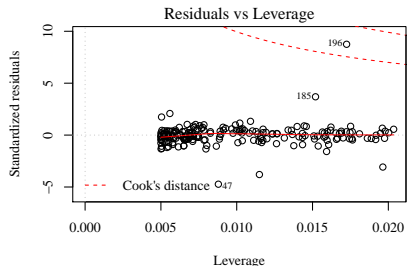
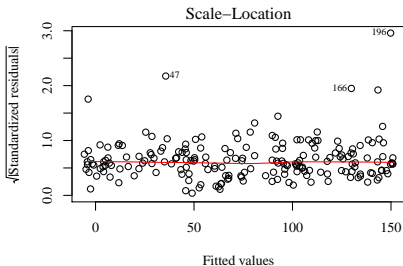
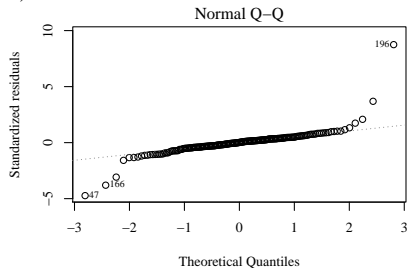
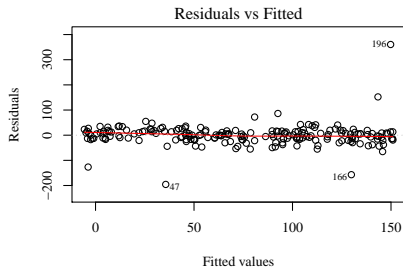
The Regression Table (just for the record)

	M1	
	Estimate	(S.E.)
(Intercept)	-6.474	(5.973)
x	1.593***	(0.100)
N	200	
RMSE	41.572	
R^2	0.562	

* $p \leq 0.05$ ** $p \leq 0.01$ *** $p \leq 0.001$

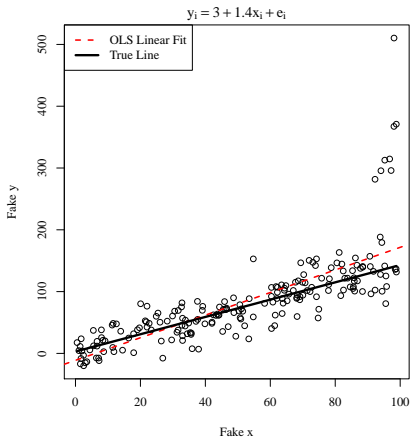
Manufactured Outliers Magnified x 10

lm(y ~ x)



10 magnified errors: Still nothing in the Leverage Plot

Cluster and Magnify the Outliers x 10



- The “true model” is a straight line

$$y_i = 3 + 1.4x_1 + e_i$$

- And the error term $e_i \sim N(\mu_e = 0, \sigma_e^2 = 22^2)$
- 10 “rightmost” positive errors: $e_i = 10 \times e_i$
- The 10 “bad cases” are:

[1]	181	183	185	186	188	189
	193	194	195	196	199	

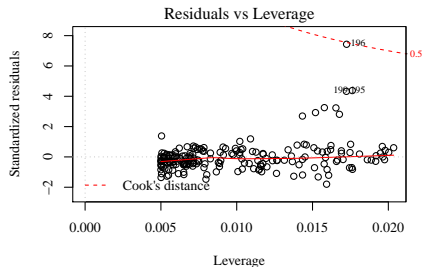
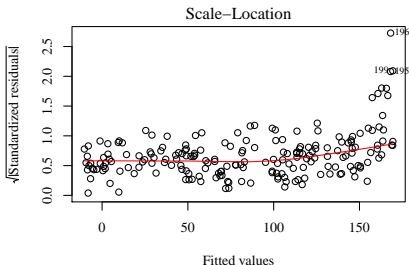
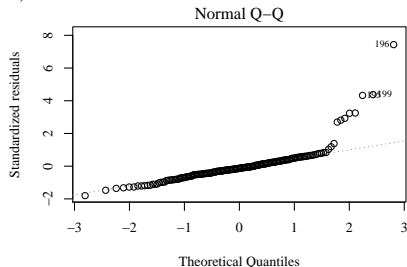
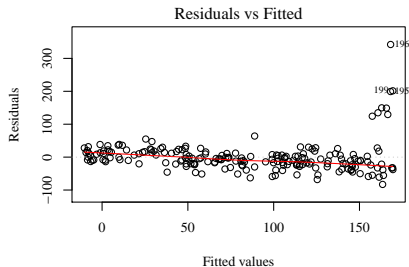
The Regression Table (just for the record)

	M1	
	Estimate	(S.E.)
(Intercept)	-11.238	(6.675)
x	1.828***	(0.112)
N	200	
RMSE	46.453	
R^2	0.575	

* $p \leq 0.05$ ** $p \leq 0.01$ *** $p \leq 0.001$

R Plot: 10 "Rightmost Positive" Magnified Errors

lm(y ~ x)

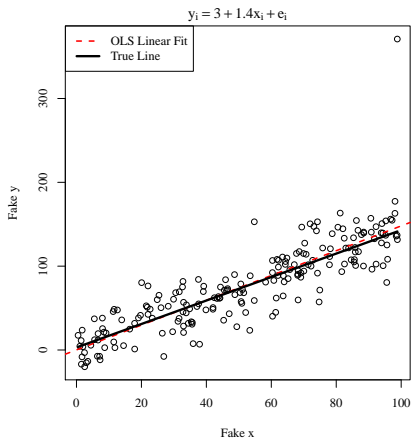


10 magnified positive errors clustered on right. Still nothing in Leverage!

Am Becoming Angry

- Had believed leverage plot would isolate the outliers after they were magnified (it did not)
- Had believed leverage plot would isolate the outliers after they were clustered (it did not)
- Drop back to simplest test: create just one outlier

Insert One Outlier On High Right Side



- The “true model” is a straight line

$$y_i = 3 + 1.4x_1 + e_i$$

- And the error term $e_i \sim N(\mu_e = 0, \sigma_e^2 = 22^2)$
- 1 “rightmost” positive errors: $e_i = 10 \times e_i$
- The “bad case” is:

[1] 199

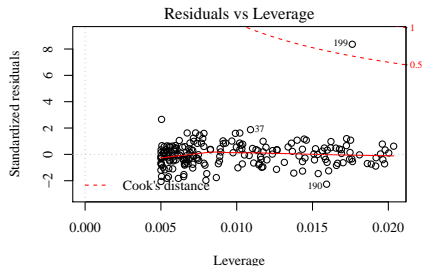
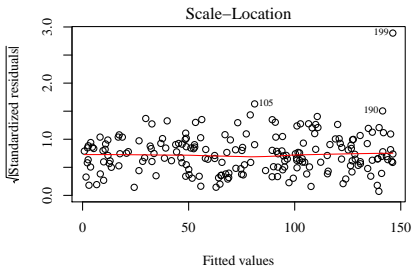
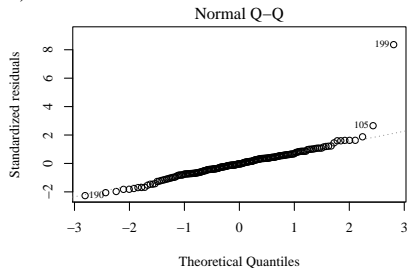
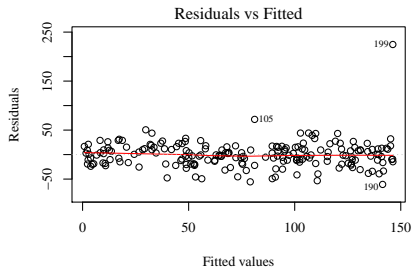
The Regression Table (just for the record)

	M1	
	Estimate	(S.E.)
(Intercept)	0.098	(3.896)
x	1.479***	(0.065)
N	200	
RMSE	27.116	
R^2	0.722	

* $p \leq 0.05$ ** $p \leq 0.01$ *** $p \leq 0.001$

Just One Bad Outlier

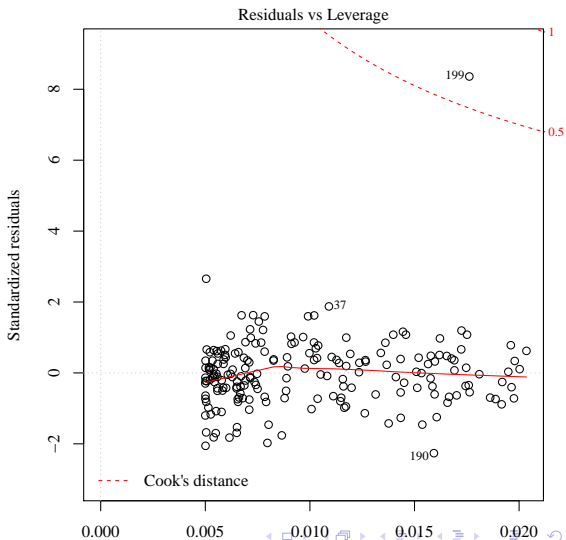
lm(y ~ x)



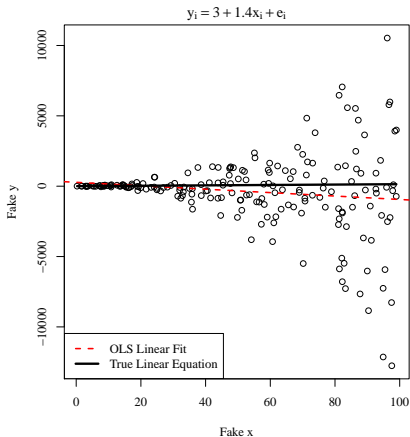
1 super magnified positive error on right

Leverage Plot Finally Spots an Outlier

- If there's just one extreme case, procedure spots it
- Conclusion: mechanical application of "spot one outlier" method not "powerful" with multiple outliers
- Appears outliers can "hide in plain sight" if there are enough of them



Error Term Variance Proportional to x_i^2



- The “true model” is

$$y_i = 3 + 1.4x_i + e_i$$

- And the error term is drawn from

$$N(\mu_e = 0, \sigma_e^2 = 0.5 \times x_i^2)$$

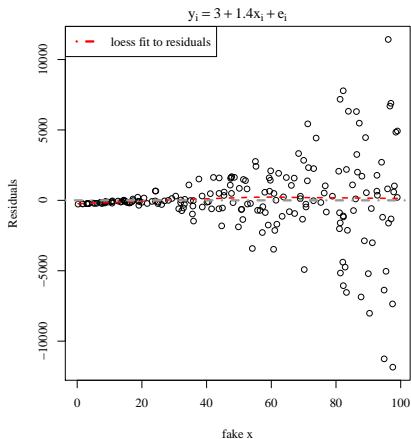
The Regression Table (just for the record)

	M1	
	Estimate	(S.E.)
(Intercept)	275.226	(393.779)
x	-12.191	(6.546)
N	200	
RMSE	2750.429	
R^2	0.017	

* $p \leq 0.05$ ** $p \leq 0.01$ *** $p \leq 0.001$

- Will work on heteroskedasticity later
- Causes high variance in estimates of $\hat{\beta}_1$ and $std.err.(\hat{\beta}_1)$ is lower than it should be

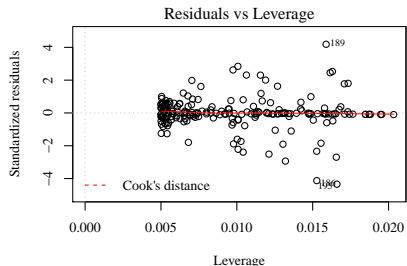
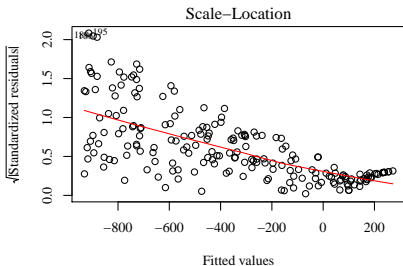
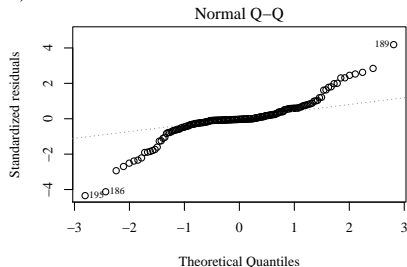
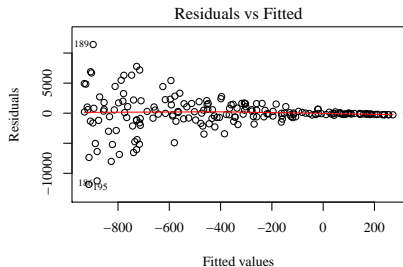
Loess / Residual Plot for Manufactured Data



- dispersion is greater on the right when plotting against x
 - will see “flip flop” when plotting against predicted value

The Manufactured Heteroskedastic Data

$\text{lm}(y \sim x)$



The 2x2 diagnostic plot matrix for the manufactured heteroskedasticity



Outline

- 1 Introduction
- 2 Start Simple: Scatterplot
- 3 Standard Diagnostic Plots
- 4 Decipher Each Type Of Plot
 - Top Left: Residual Plot
 - Top Right: Q-Q plot
 - Bottom Left: Scale-Location Plot
 - Bottom-Right: Leverage, Cook's Distance
- 5 Repeat with more Predictors
 - Multiple Regression Fitted
 - Termplot
 - Diagnostic Plots
- 6 Stress Test These Diagnostics
 - Bad Nonlinearity
 - Add Some "Outliers"
 - Heteroskedasticity

Problems

- 1 Fit a regression with several predictors. This is more interesting if some predictors are “factors” (categorical variables according to R) and some are continuous numeric variables. Then run the command `termplot(mod1)` on your model, which I assume is called “mod1”. Note that R will wait for you to click “next” before it shows the graph.
- 2 On the same fitted model as you used in the previous example, run the command `plot(mod1)`.
- 3 For the R Summer Course, I made several presentations about the plotting features of R. If you didn't know about them yet, this might be a good time to take a quick survey. In my guides folder, they are under Rcourse.
 - 1 [plot-1](#)
 - 2 [plot-2](#)
 - 3 [plot-3d](#)
 - 4 [regression-plots-1](#)

Problems ...

For regression plots, I suppose you will be mainly interested in plot-1 (scatterplots and histograms) and regression-plots-1. (I made a pretty vigorous effort to do 3-D plotting in plot-3d, but in many ways, it is just too hard for R beginners.)

- 4 Here's a trick question for you. Consider this display of a q-q plot. [imagine qq-plot that shows several points that are way far off the straight line.] Does this mean the regression results are wrong? Please explain. (This is a final exam sort of question. Its one that should make you connect theory and practice. One trick here is that I've used the word "wrong" and leave you to decide what wrong means. Did I mean "biased?" "Inconsistent"? Another trick here is that you can do almost all of the usual regression exercises without assuming that e_i is normal. So think about how a regression can still be unbiased and consistent if the error is not normal.)

Problems ...

- 5 Here is one way to find out which cases are “outliers.” R has a function called “identify” and I’ve never gotten very good at it. But maybe you are better. The idea is that you can create a scatterplot and then click on certain points to identify them. Run `?identify` to read more. Here’s a working example.

```
x <- c(1,2,3,4,5)
y <- c(5,4,3,4,5)
nam <- c("Bill", "Charles", "Jane", "Jill", "Jaime")
dat <- data.frame(x,y,nam)
rm(x,y,nam)
plot(y~x, dat=dat)
with(dat, identify(x,y, nam) )
```

Problems ...

As soon as you hit return on the last line, the R session will seem to “freeze”. It is waiting for you to left-click on the points in the graph. You left-click on a point to insert “nam” next to it, and when you are finished, you can right-click to “release control” from the identify function. This is one way to spot outliers.

This one frustrates me so much I made a `WorkingExample` for it. That’s in my collection “`plot-identify_points-1.R`” (Recall, you can get there either though pj.freefaculty.org/R or via the Rcourse notes (end up at same place). If you don’t soup up your computers, I bet you will have more fun with it than I do. My video driver is constantly out of whack, so a click does not do what I expect.