Robust Regression

Jonathan Bryan

April 26, 2018

Packages https://cran.r-project.org/web/views/Robust.html

Problem

Common regression methods such as ordinary least squares (OLS) make strong assumptions about the the behavior of errors. Such methods often assume errors have constant variance (homoskedasticity), no autocorrelation, and normality. This is expected given the parametric forumlation of OLS regression $Y \overset{iid}{\sim} N(X^T\beta, \sigma_\epsilon^2)$. Commonly outliers and influential observations cand drasticaly lower the optimality of OLS regression. Divergence from these modelling assumptions is not rare and making OLS regression sensitive to response variable outliers, high leverage points, heteroskedastic errors, and autocorrelation.

Approach

Robust regression methods have been developed to overcome these challenges through parametric and non-parametric solutions. This report surveys divergences from OLS regression assumptions and which robust regression methods are best used to model these divergences for both inference and prediction.

https://www.mathworks.com/help/econ/compare-robust-regression-techniques.html

Non-normal errors

OLS regression assumes that errors are independent and identiticaly normally distributed. Normal iid errors is often a reasonable assumption given, ______ and if data is _________. However, empirical errors may show more extreme values than expected with a normal distribution. So-called fat-tailed distributions such as Student's t-distribution and the Cauchy distribution are symmetric distributions with greater probabilities assigned to extreme values.

Table 1: Percentage Capture of 95% CI

	Normal	Cauchy	t.dist
Intercept	0.94	0.95	0.97
x1	0.97	0.94	0.97
x2	0.97	0.95	0.93
x3	0.96	0.95	0.96
x4	0.96	0.93	0.95
x5	0.92	0.94	0.96

Table 2: Percentage Capture of P-values < 0.05

	Normal	Cauchy	t.dist
Intercept	0.06	0.05	0.03
x1	1.00	0.46	0.99
x2	1.00	0.45	0.99

	Normal	Cauchy	t.dist
x3	1.00	0.91	1.00
x4	0.04	0.07	0.05
x5	0.08	0.06	0.04

Response Outliers

Data may contain observations that, for various underlying reasons, have extreme values in the response variable. This may be due to data collection problems, measurement error, or represent some true data-generating process separate from the rest of the data. Outliers that arise from data collection or measurement error are rarely desriable and if at the extremes of the predictor space can cause poor parameter inference.

https://www.mathworks.com/help/econ/compare-robust-regression-techniques.html#d119e40454 https://stats.idre.ucla.edu/r/dae/robust-regression/

Table 3: Percentage Capture of 95% CI

	No.Outliers	Front.Outliers	Middle.Outliers	End.Outliers
Intercept	0.96	0.15	0.96	0.00
x1	0.95	0.59	0.96	0.25
x2	0.95	0.51	0.97	0.29
x3	0.96	0.95	0.74	0.01
x4	0.96	0.44	0.90	0.19
x5	0.96	0.56	0.98	0.27

Table 4: Percentage Capture of P-values < 0.05

	No.Outliers	Front.Outliers	Middle.Outliers	End.Outliers
Intercept	0.04	0.85	0.04	1.00
x1	1.00	0.13	0.11	0.98
x2	1.00	1.00	0.43	0.31
x3	1.00	1.00	1.00	1.00
x4	0.04	0.56	0.10	0.81
x5	0.04	0.44	0.02	0.73

Influential Observations

Influential observations are data points that would substantially change the model given their absence. An observations influence is function of the extremity of the data points covariate values and the residual for that data point. Influential observations can significantly change the estimated paramters of a regression model. Cook's distance and DFBETA are commonly used tests for influence. (Give brief mathematical description). It's clear in figure XX below that even data generated from a linear model with normally distributed errors can contain natural points of high influence. We will again look at the confidence intervals and significance tests for the estimates of the coefficients before and after adjusting for observations 73 and 66 to be of even greater influence.

Fig. Influence Plot for Normal Error OLS Regression

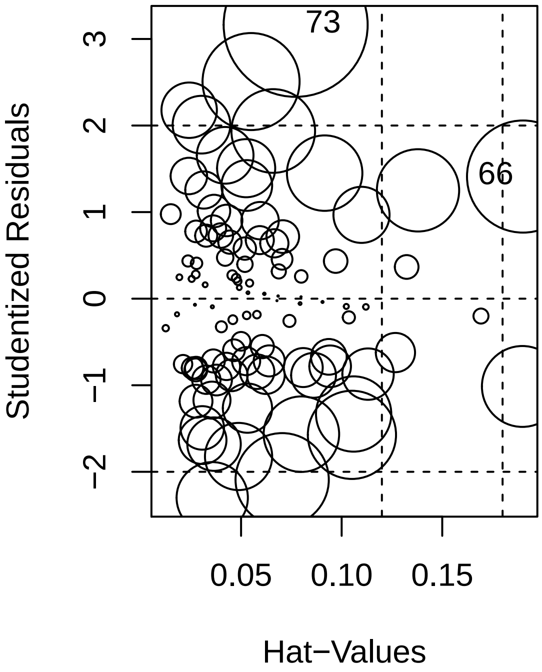
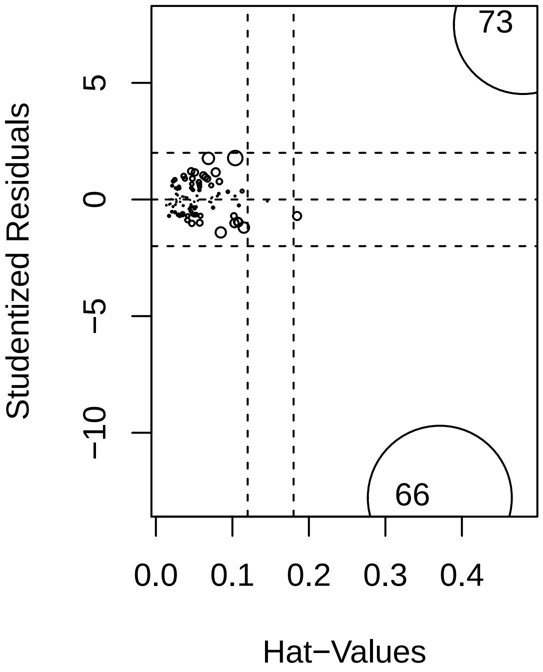


Fig. Modified Influence Plot for Normal Error OLS Regression



We observe that the OLS model still explains much of the overall variation in the data when fitting on the modified data with influenctual points. However, it no longer shows the "x1" covariate as significant. In this case, 2% of the data being influential has caused us to lose a significant covariate.

Table 5: Coefficient Summary for OLS

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-1.115	1.018	-1.095	0.276
x1	1.990	0.066	30.216	0.000
x2	-4.860	0.189	-25.717	0.000
x3	3.009	0.012	249.608	0.000
x4	0.094	0.062	1.517	0.133
x5	-0.010	0.120	-0.081	0.936

Table 6: Coefficient Summary for OLS (Modified Data)

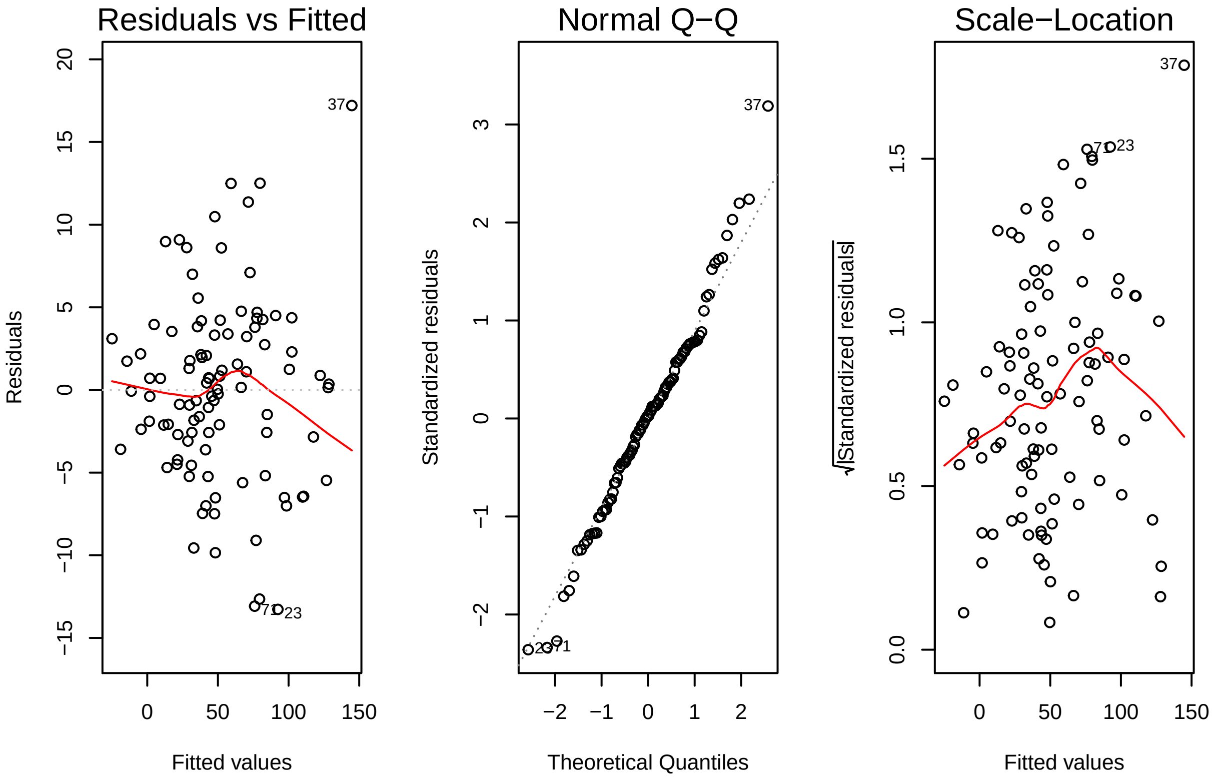
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	21.899	6.804	3.218	0.002
x1	-0.110	0.408	-0.270	0.788
x2	-4.917	1.239	-3.967	0.000
x3	2.818	0.081	34.782	0.000
x4	0.178	0.398	0.447	0.656
x5	-0.362	0.802	-0.452	0.653

Table 7: OLS Model R-squared and F-statistic (df = 5, 94)

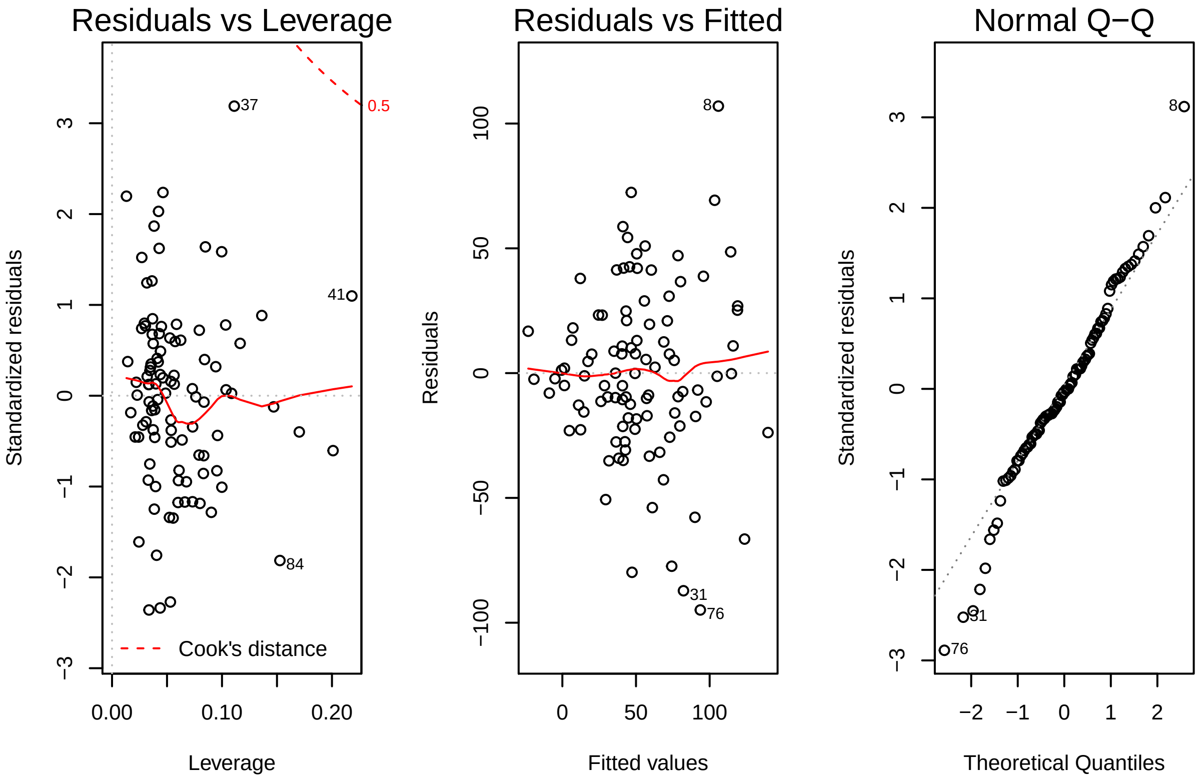
	Adj..R.squared	F.statistic
Normal Data	0.999	15712.288
Modified Data	0.941	318.466

Heteroskedasicity

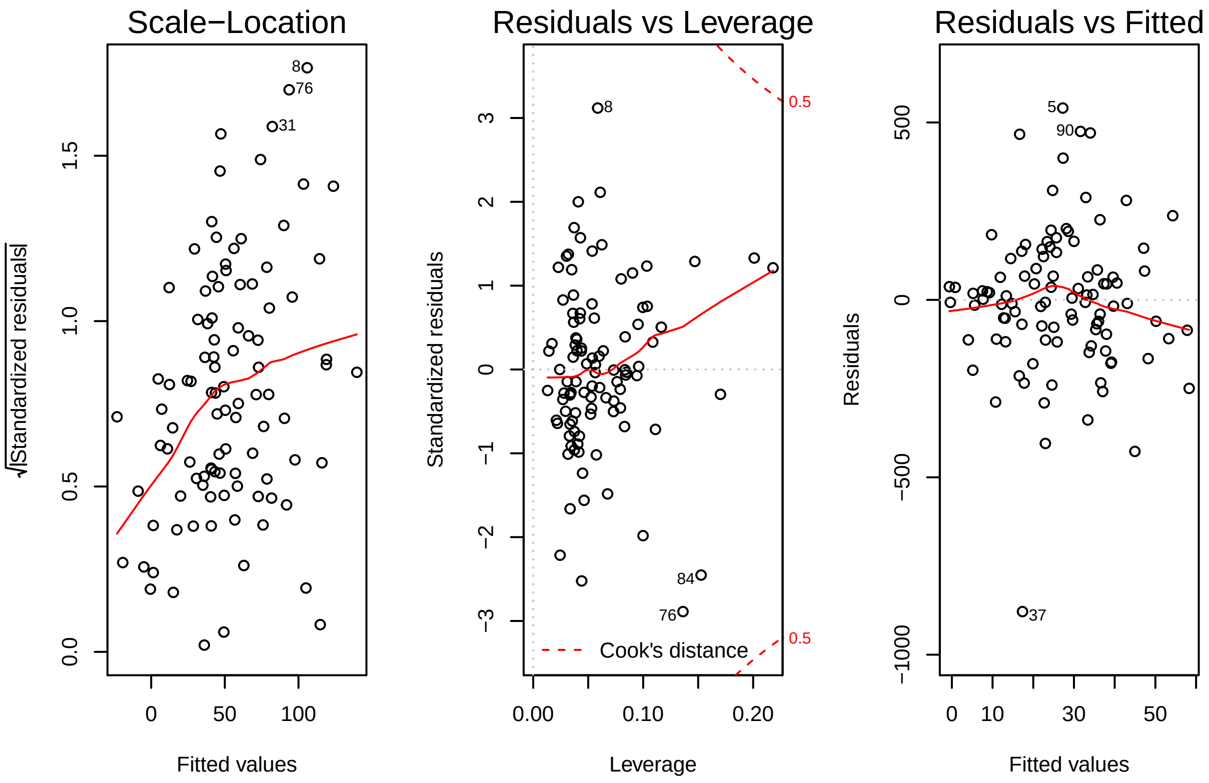
Data that shows non-constant variance over the error term means our OLS estimates are no longer the most optimal estimates of the coefficients. OLS will still give us unbiased point estimates of the coefficients, however, the standard errors, confidence intervals, and significance tests maybe be severly biased and unnecessarily conservative. In the simulated data below, the variance of the errors are linearly related to the the scale of the predictors. The error of the data becomes more disperse as we move toward higher multidimensional values in the covariate domain. Heteroskedasticy often occurs when the error of model increase with time as a system becomes more complex, or the error is linked to the scale of one of the covariates.

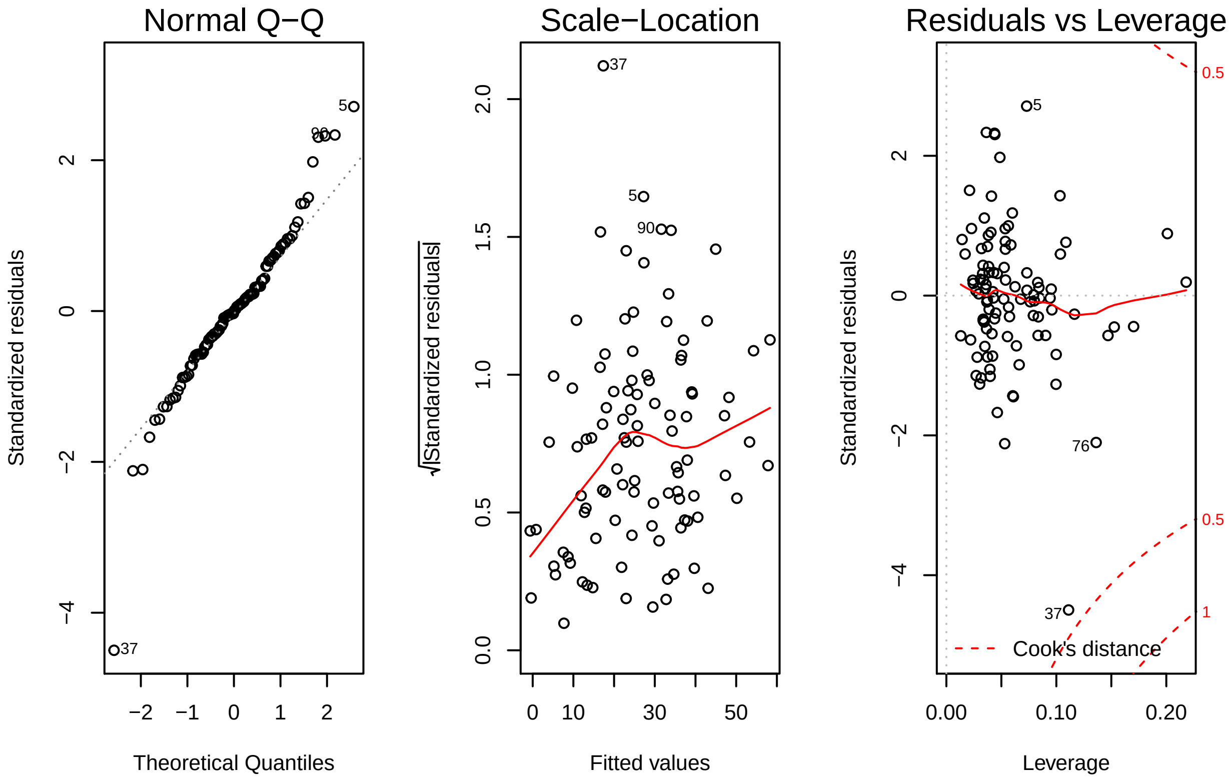


```
## NULL
```


Residuals vs Leverage
Standardized residuals
Leverage
0.00
0.10
0.20
−3
−2
−1
0
1
2
3
37
41
84
0.5
Cook's distance
Residuals vs Fitted
Residuals
Fitted values
0
50
100
−100
−50
0
50
100
8
31
76
Normal Q−Q
Standardized residuals
Theoretical Quantiles
−2
−1
0
1
2
−3
−2
−1
0
1
2
3
8
31
76

```
## NULL
```


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
0
50
100
Fitted values
8
76
31
Residuals vs Leverage
Standardized residuals
−3
−2
−1
0
1
2
3
0.00
0.10
0.20
Leverage
8
84
76
0.5
0.5
Cook's distance
Residuals vs Fitted
Residuals
−1000
−500
0
500
0
10
30
50
Fitted values
5
90
37



```
## NULL
```

Table 8: Percentage Capture of 95% CI

	Hetero.1	Hetero.2	Hetero.3
Intercept	0.90	0.96	0.93
x1	0.89	0.97	0.91
x2	0.95	0.97	0.93
x3	0.96	0.99	0.85
x4	0.96	0.97	0.93
x5	0.93	0.91	0.94

Table 9: Percentage Capture of P-values < 0.05

	Hetero.1	Hetero.2	Hetero.3
Intercept	0.10	0.04	0.07
x1	1.00	0.20	0.10
x2	1.00	0.18	0.09
x3	1.00	1.00	0.30
x4	0.04	0.03	0.07
x5	0.07	0.09	0.06

Robust Regression Models

Several modifications and alternative models have been developed to correct failures in the OLS model assumptions.

Least Absolute Deviations Regression

Least absolute deviations regression (LAD) finds the coefficient estimates that minimize the absolute value of the difference between the specified model and the responses. In this model we simply replace the squared objective function of OLS with the absolute value function.

$$\arg\min_{\theta} \Sigma_{i=1}^{n} |y - X^t \beta|$$

Distributionally, the errors of the model are charaterized by the Laplace distribution, which has fatter tails compared to the normal distribution. The importance of resdiuals in the LAD model does not scale with the size of the residual. In contrast, larger residuals have a greater impact on the estimates of the model as they scale quadratically. One key drawback of LAD is that data points that are symmetrical about the horizontal axis can cause the estimates to be unstables. In fact, there maybe infinitely many solutions in the given scenario. *how to know when unstable?*

```
##   Intercept           X        <NA>        <NA>        <NA>        <NA> 
##  0.19803073  1.87479973 -4.70601082  3.00258446  0.05417321  0.20898959
```

Comparison http://article.sapub.org/10.5923.j.statistics.20150503.02.html#Sec3

https://stats.stackexchange.com/questions/277823/least-absolute-deviation-regressions-coefficient-significance-levels?utm_medium=organic&utm_source=google_rich_qa&utm_campaign=google_rich_qa **There's not actually a t-test, because the estimate divided by its standard error doesn't have a t-distribution. Similar for an F-test. Being an MLE, there would be an asymptotic z-test, or an asymptotic chi-square test. [There's the possibility of using some resampling-based tests as well, permutation tests or bootstrapping. You could also use L1pack's ability to simulate from L1 models to do a parametric bootstrap.]

+Iteratively reweighted least squares (MASS)

- Least trimmed squares (robustbase)
- M regression(robustreg) (maybe not because old)
- fast-S algorithms and heteroscedasticity and autocorrelation corrected erros (robustbase)
- MM-estimation (MASS)
- (quantreg)
- median-based Theil-Sen (mblm)
- Robust Bayesian

Comparison of Robust Regression Models

Using the simulated data set for each departure from the OLS assumptions, we compare the performance of these robust regression models for each scenario. The final comparison is a dataset that combines each of these scenarios.

2) Compare regression model performance in the four data cases 1) non-normal errors, 2) reponse outliers, 3) influential points 4) and heteroskedasticty

3) Combine data issues into one data set and compare all models again