Regression Assumptions

Jacob M. Montgomery

Quantitative Political Methodology

Regression Diagnostics

Poster questions?

Road map

Where we have been:

- Regression
- Interpreting regression
- Causal inference in regression
- Model Fit

Road map

Where we have been:

- Regression
- Interpreting regression
- Causal inference in regression
- Model Fit

Today:

- Regression assumptions
- How they can be broken
- Simple tricks to solve them

Class Business

- PS Due today. One more to go.
- Poster files will be due before the final lecture period.
- Optional poster session on Monday, December 11
- 1/% extra credit for the best poster as chosen by faculty
- Cookies
- Be proud of your work

Assumptions of regression

1. Observations are independent

Assumptions of regression

1. Observations are independent
2. For any given value of x, the response y varies around the regression line according to a normal distribution

Assumptions of regression

1. Observations are independent
2. For any given value of x, the response y varies around the regression line according to a normal distribution
3. There is a linear relationship between x and y

Assumptions of regression

1. Observations are independent
2. For any given value of x, the response y varies around the regression line according to a normal distribution
3. There is a linear relationship between x and y
4. The standard deviation of y (σ) is the same for all values of x

Assumptions of regression

1. Observations are independent
2. For any given value of x, the response y varies around the regression line according to a normal distribution
3. There is a linear relationship between x and y
4. The standard deviation of y (σ) is the same for all values of x

For each one:

- What can go wrong?
- What can we do about it?

Assumption 1: Observations are independent

Autocorrelation

- Time-series
- Repeated observations
- Space

Assumption 1: Observations are independent

Autocorrelation

- Time-series
- Repeated observations
- Space

Conditional standard deviation

$$\hat{\sigma} = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{\sum(Y_i - \hat{Y}_i)^2}{n-2}}$$

Assumption 1: Observations are independent

Autocorrelation

- Time-series
- Repeated observations
- Space

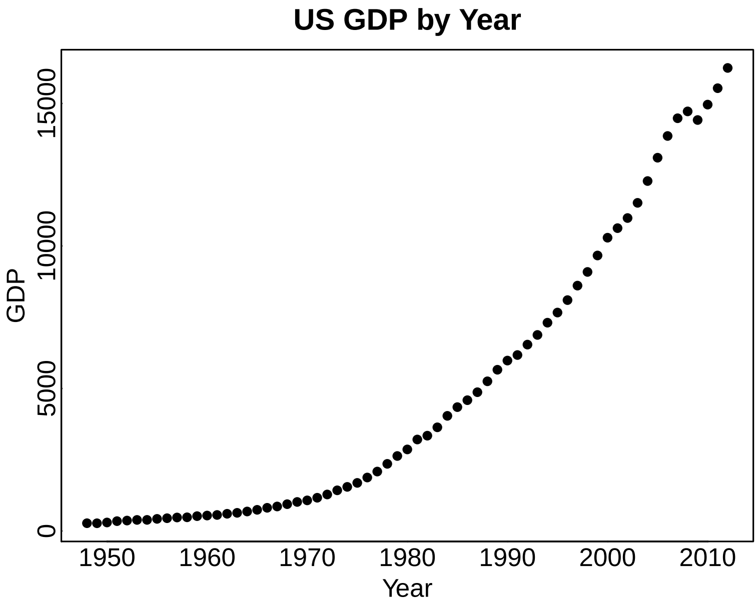
Conditional standard deviation

$$\hat{\sigma} = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{\sum(Y_i - \hat{Y}_i)^2}{n-2}}$$

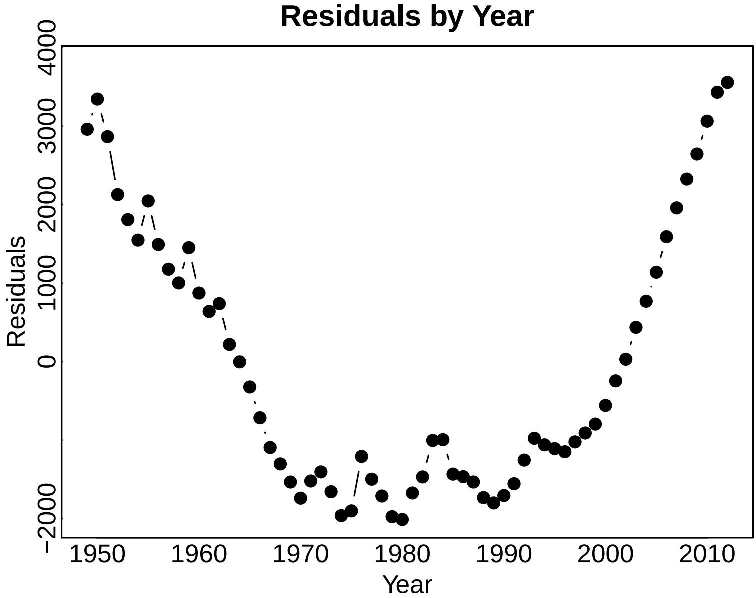
Standard error for β

$$\hat{\sigma}_{\hat{\beta}} = \frac{\hat{\sigma}}{\sqrt{\sum(X_i - \bar{X})^2}}$$

Your standard errors will be too small!

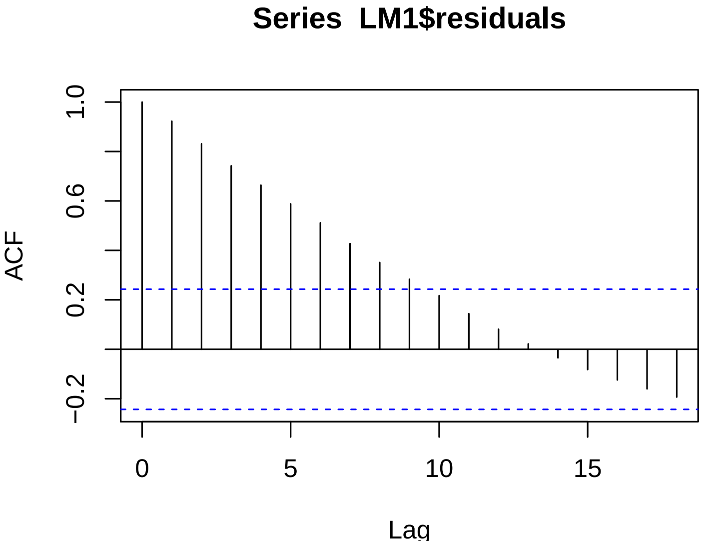

US GDP by Year
GDP
0
5000
10000
15000
1950
1960
1970
1980
1990
2000
2010
Year

GDP $\sim$ Year + Unemployment



GDP ∼ Year + Unemployment

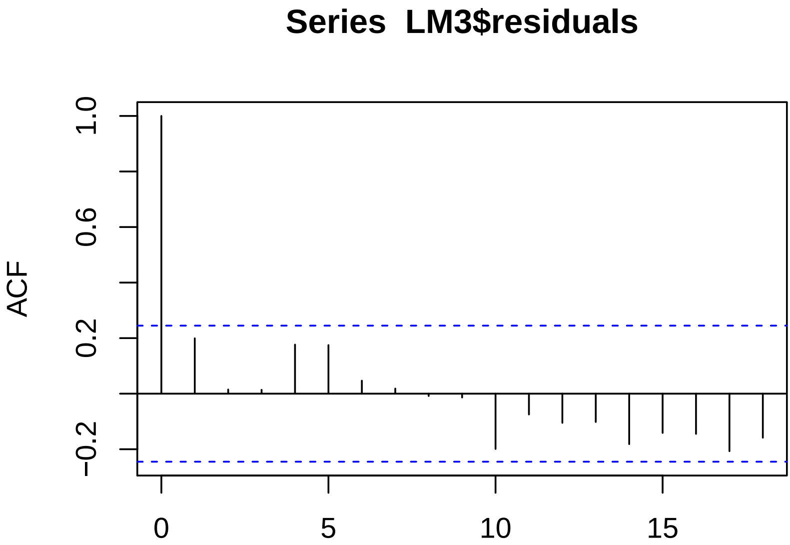
```
acf(LM1$residuals)
```



Autocorrelation: Solution

- Lagged dependent variable
- "Differencing" dependent and/or independent variables
- Some combination thereof
- Also . . . fixed effects

$\Delta GDP \sim Year + Unemployment + Lagged.GDP$

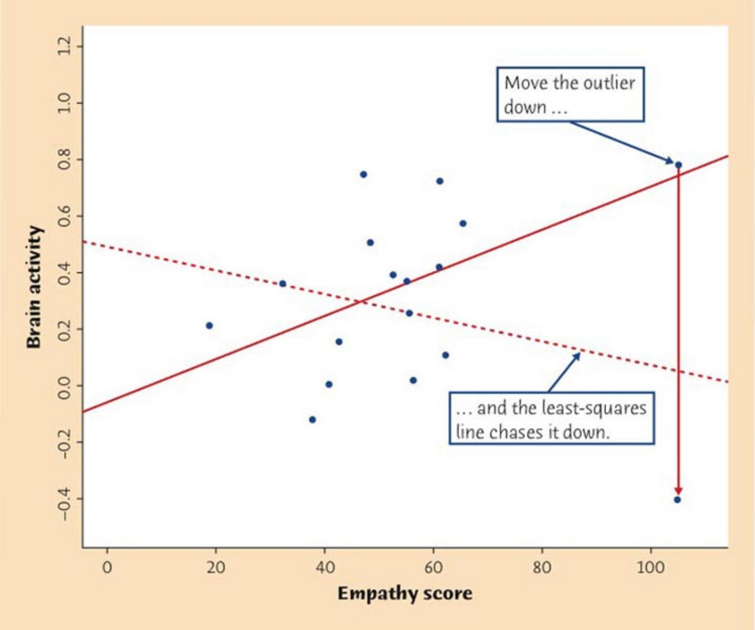


Assumption 2: Response varies around line according to a normal

What can go wrong?

- Outliers
- Exacerbated by high leverage points

Influential observations



Leverage (for bivariate regression)

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum(x - \bar{x})}$$

Leverage (for bivariate regression)

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum(x - \bar{x})}$$

Cook's distance – a measure of influence

$$D_i = \frac{e_i^2}{p\ MSE}\left(\frac{h_i}{(1 - h_i)^2}\right)$$

Leverage (for bivariate regression)

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum(x - \bar{x})}$$

Cook's distance – a measure of influence

$$D_i = \frac{e_i^2}{p\ MSE}\left(\frac{h_i}{(1 - h_i)^2}\right)$$

- $e_i = \hat{y}_i$ - $\hat{y}_{(i)}$
- p is the number of variables in the model
- MSE is the mean square error of the regression

Easy solutions

- Transforming variables (Xs and Ys)

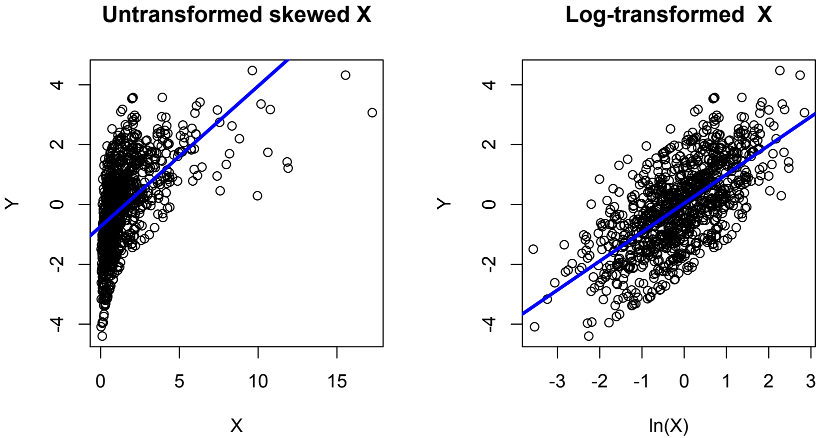
Easy solutions

- Transforming variables (Xs and Ys)
- Removing high-leverage outliers

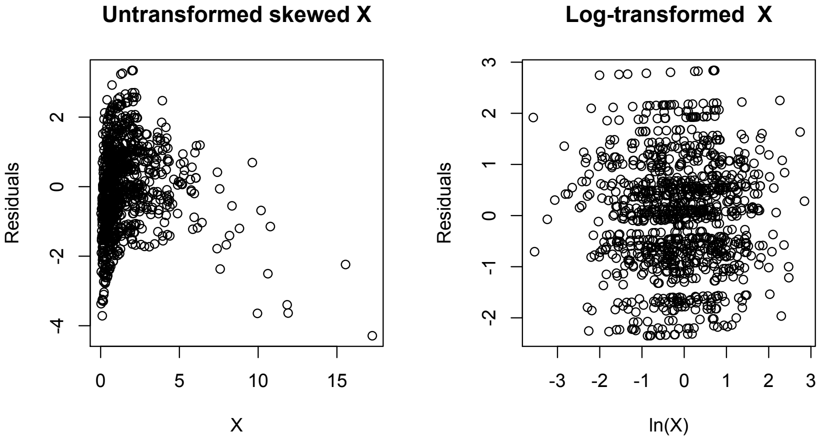
Easy solutions

- Transforming variables (Xs and Ys)
- Removing high-leverage outliers
- Trimming the variable

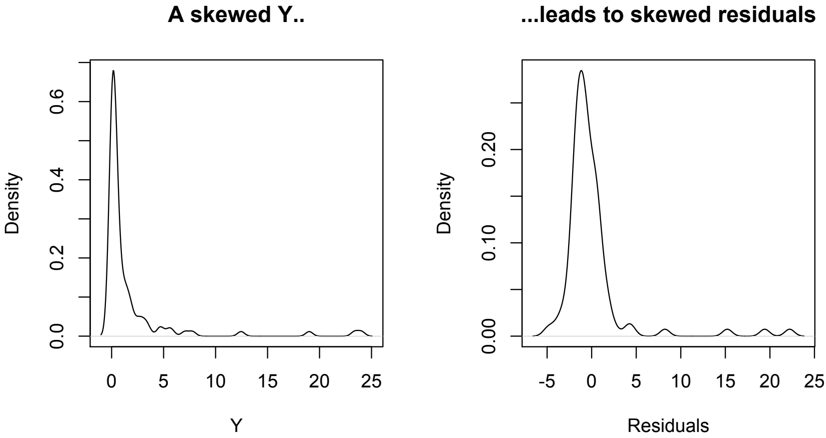
Visualizing different Xs



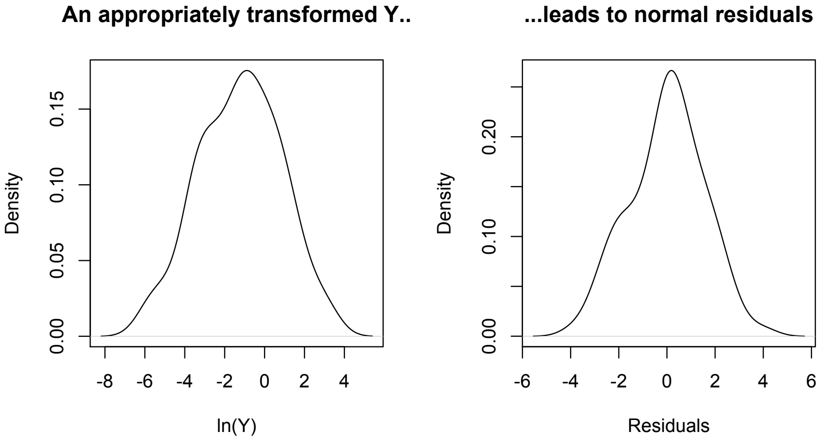
Visualizing different Xs



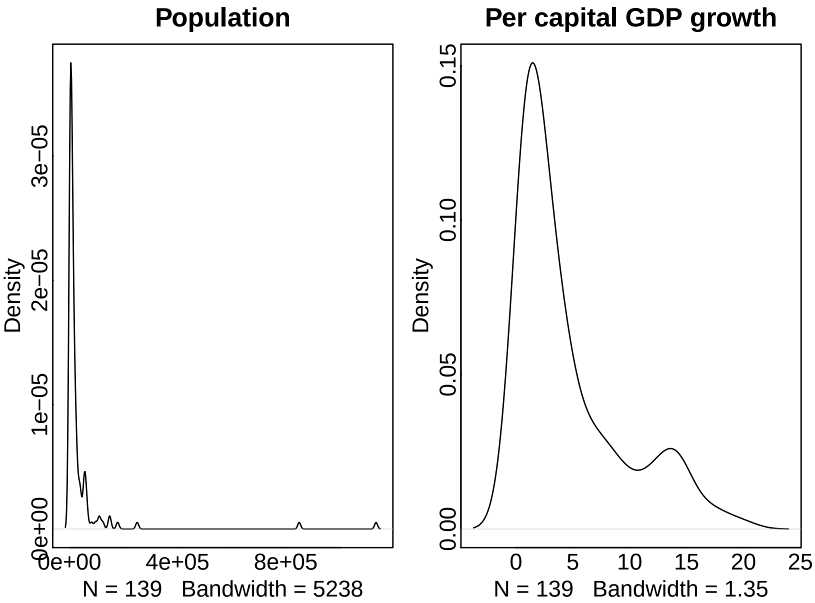
Transforming Y's



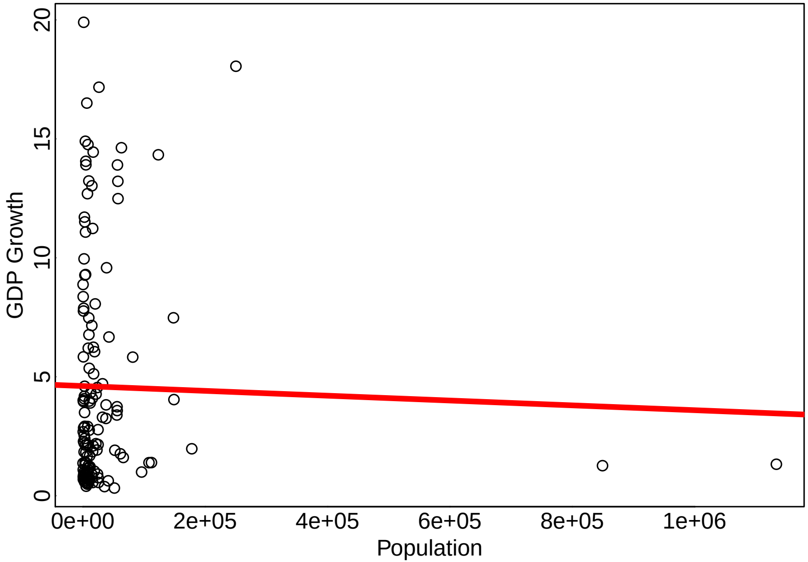
Transforming Y's



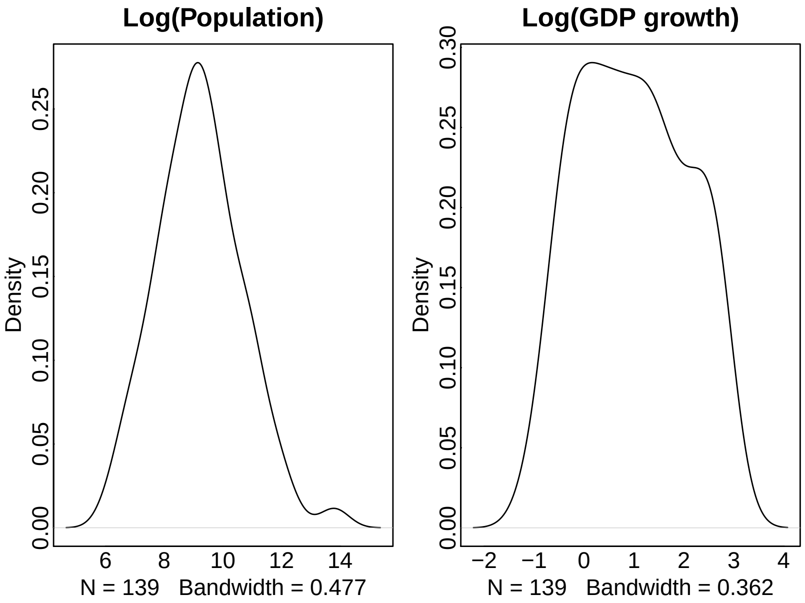
Example: GDP versus per capital GDP in 1990



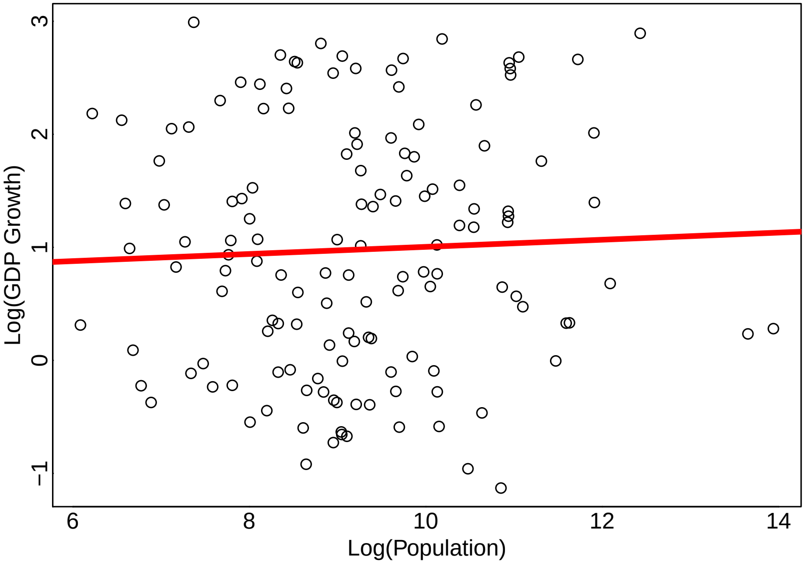
Example: GDP versus per capital GDP in 1990



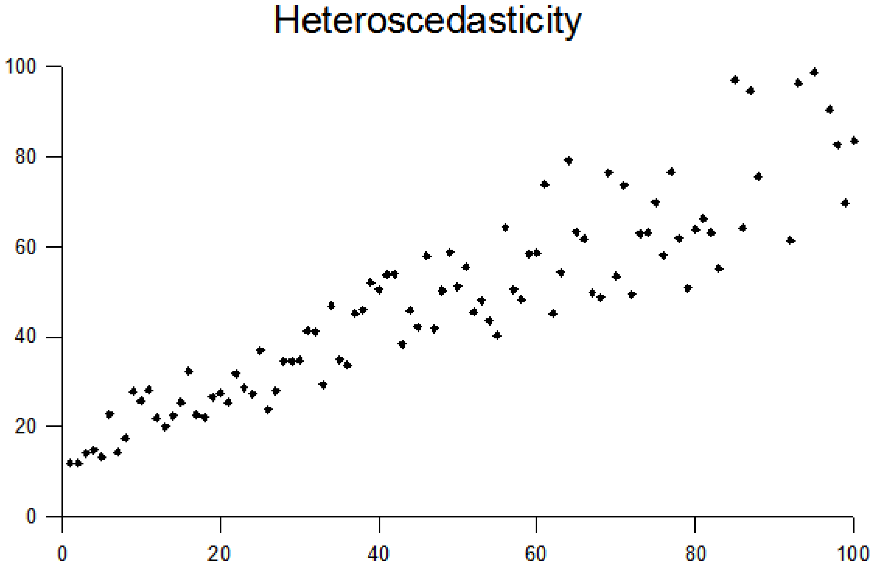
Example: GDP versus per capital GDP in 1990



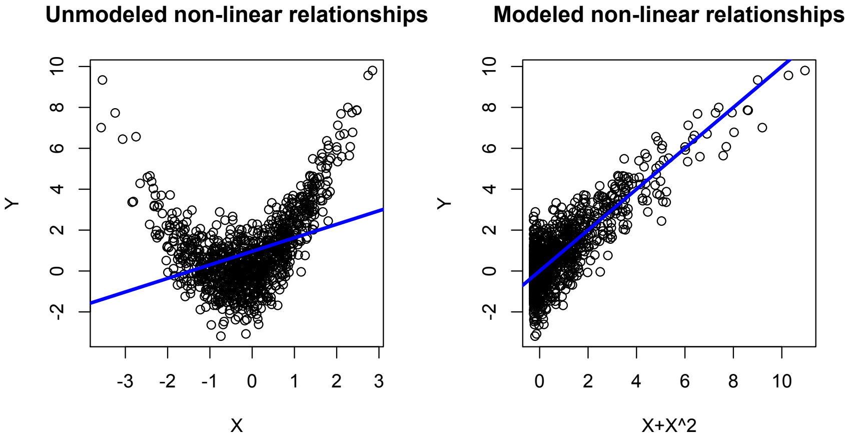
Example: GDP versus per capital GDP in 1990



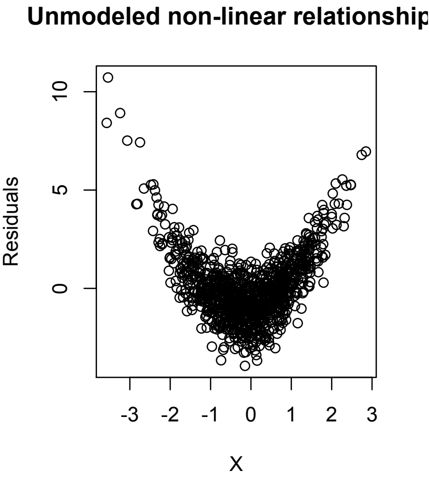
Heteroscedasticity: Variance changes based on value of x

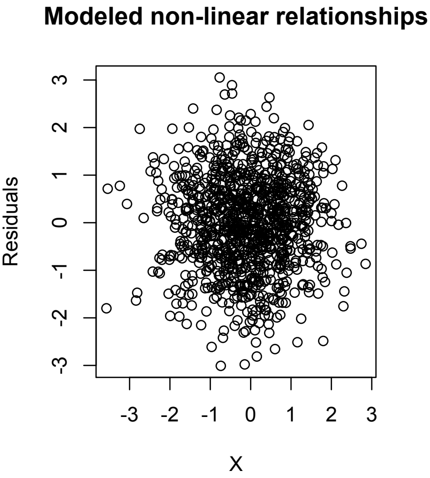


Problem 4: Nonlinear relationships



Solution: Model the relationship

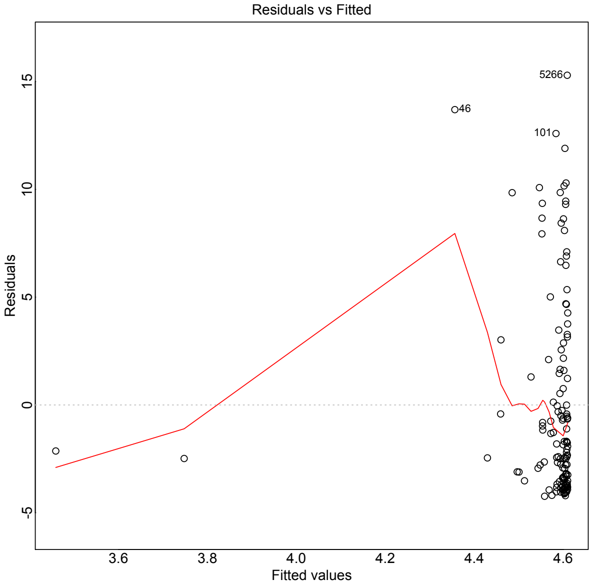




Easy regression diagnostics in R

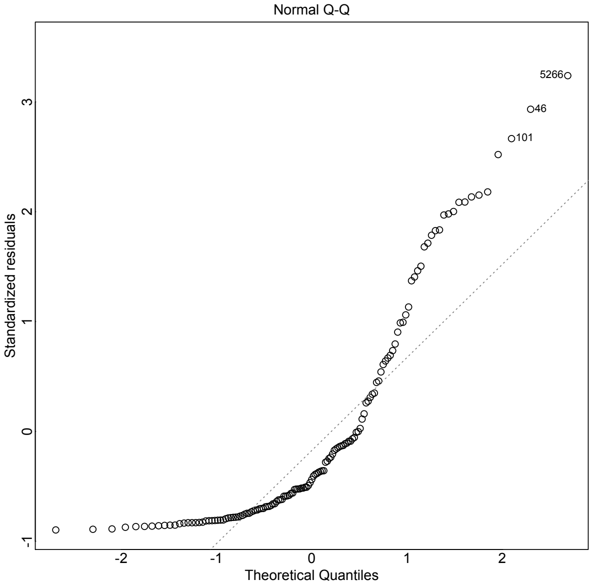
```
# The first regression
FL.LM1 <- lm(gdpen~pop, data=FL)
plot(FL.LM1)


# The second regression
FL.LM2 <- lm(log(gdpen)~log(pop), data=FL)
plot(FL.LM2)
```

Bad models

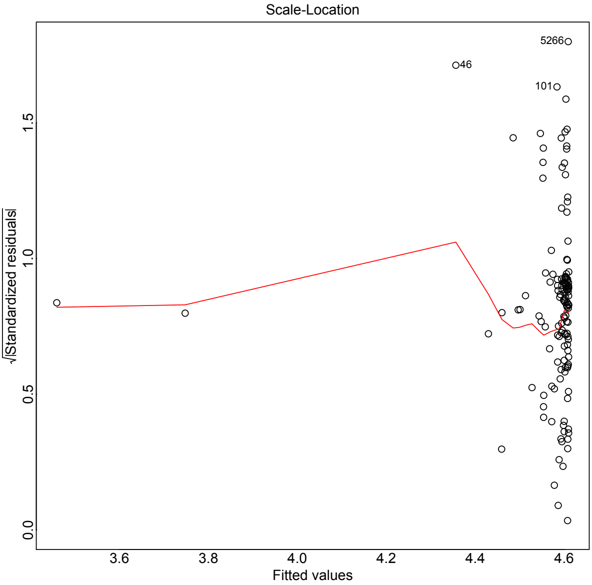


Residuals vs Fitted
5266
46
101
Residuals
Fitted values
15
10
5
0
-5
3.6
3.8
4.0
4.2
4.4
4.6

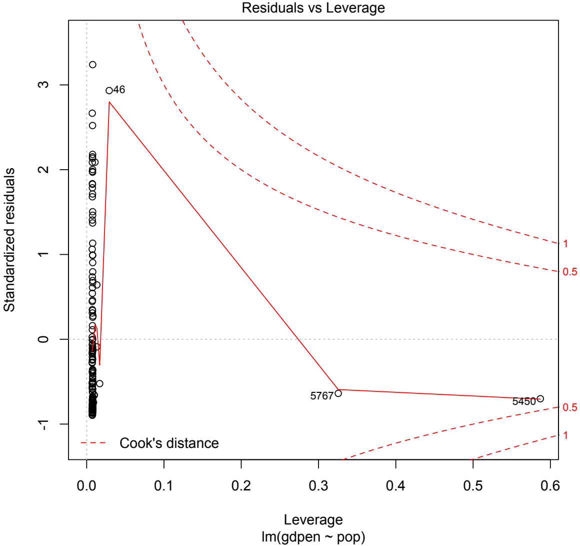
Bad models


Normal Q-Q
5266
46
101
Standardized residuals
Theoretical Quantiles
-1
0
1
2
3
-2
-1
0
1
2

Bad models

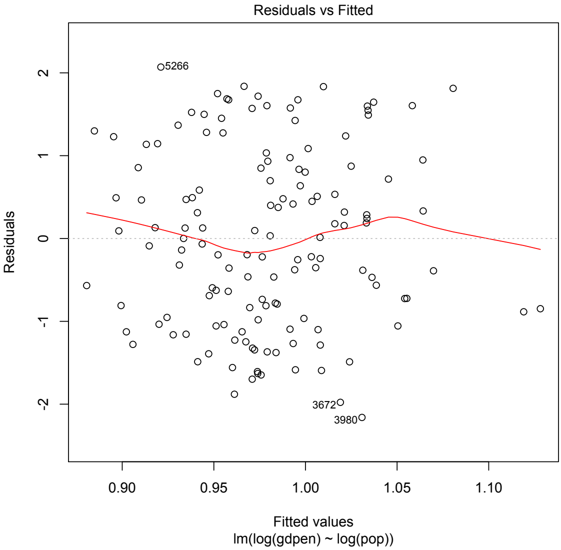

Scale-Location
5266
46
101
0.0
0.5
1.0
1.5
$\sqrt{|\text{Standardized residuals}|}$
3.6
3.8
4.0
4.2
4.4
4.6
Fitted values

Bad models

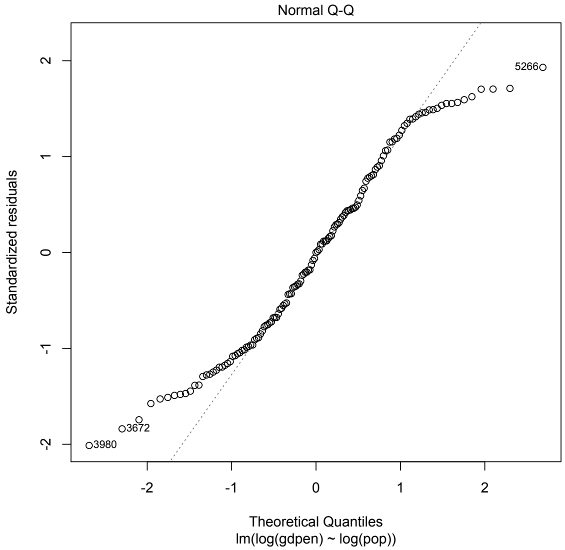


Residuals vs Leverage
Standardized residuals
Leverage
lm(gdpen ~ pop)
Cook's distance
46
5767
5450
1
0.5
0.5
1
-1
0
1
2
3
0.0
0.1
0.2
0.3
0.4
0.5
0.6

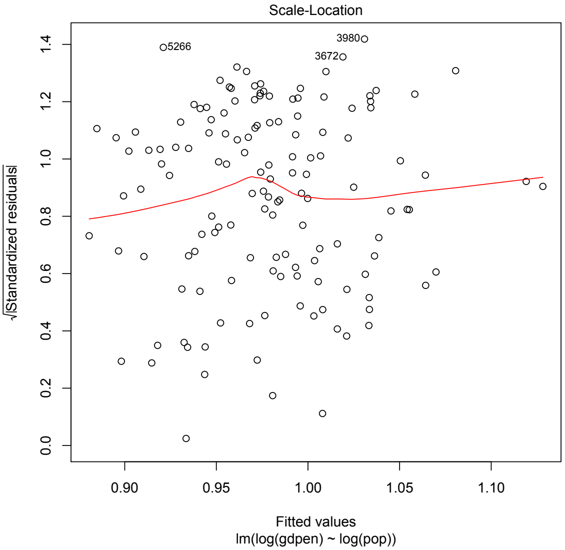
Better Model



Better Model



Better Model



Better Model