

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

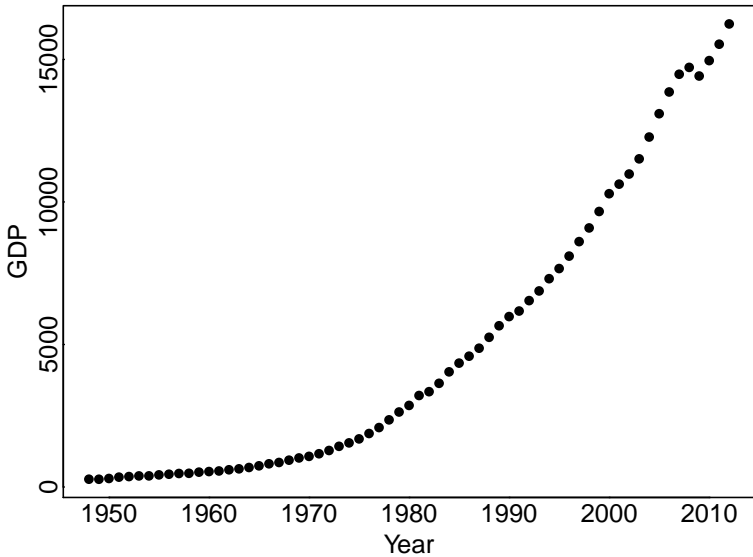
$$\hat{\sigma} = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{\sum (Y_i - \hat{Y}_i)^2}{n-2}}$$

Standard error for β

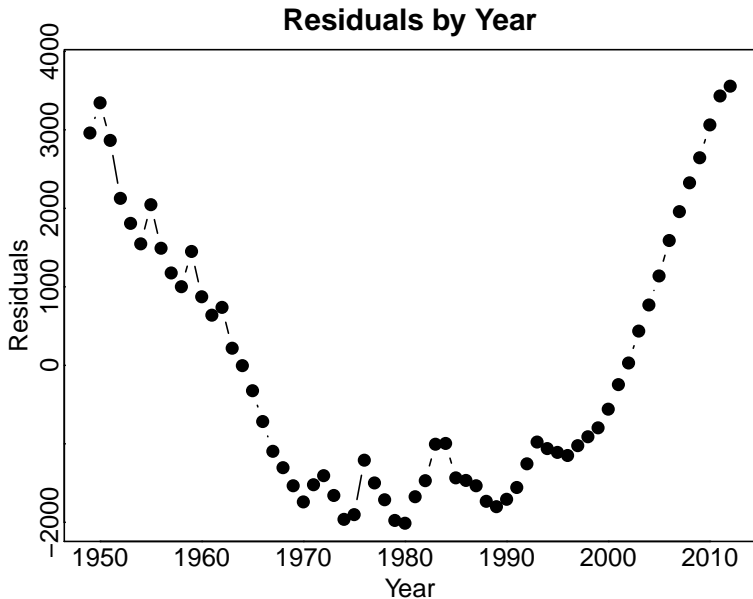
$$\hat{\sigma}_{\hat{\beta}} = \frac{\hat{\sigma}}{\sqrt{\sum (X_i - \bar{X})^2}}$$

Your standard errors will be too small!

US GDP by Year

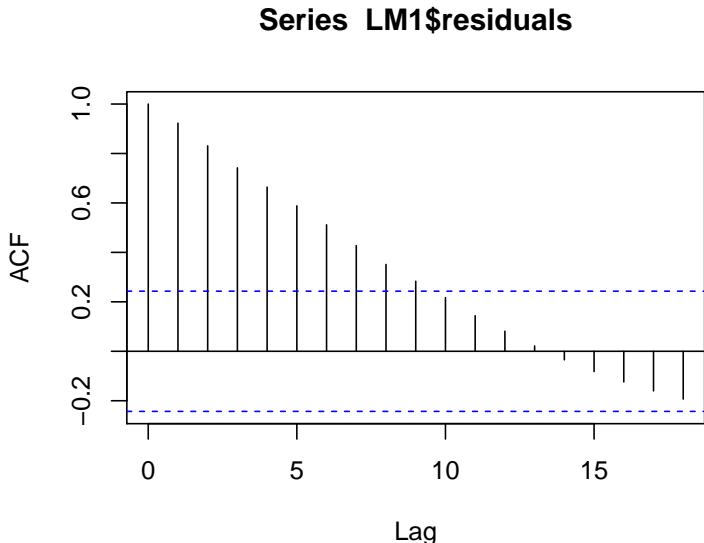


$$\text{GDP} \sim \text{Year} + \text{Unemployment}$$



$GDP \sim Year + Unemployment$

`acf(LM1$residuals)`

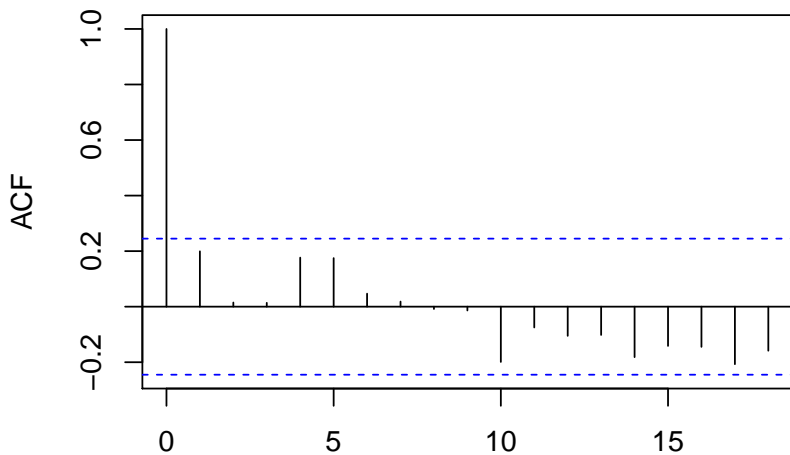


Autocorrelation: Solution

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

$$\Delta\text{GDP} \sim \text{Year} + \text{Unemployment} + \text{Lagged.GDP}$$

Series LM3\$residuals

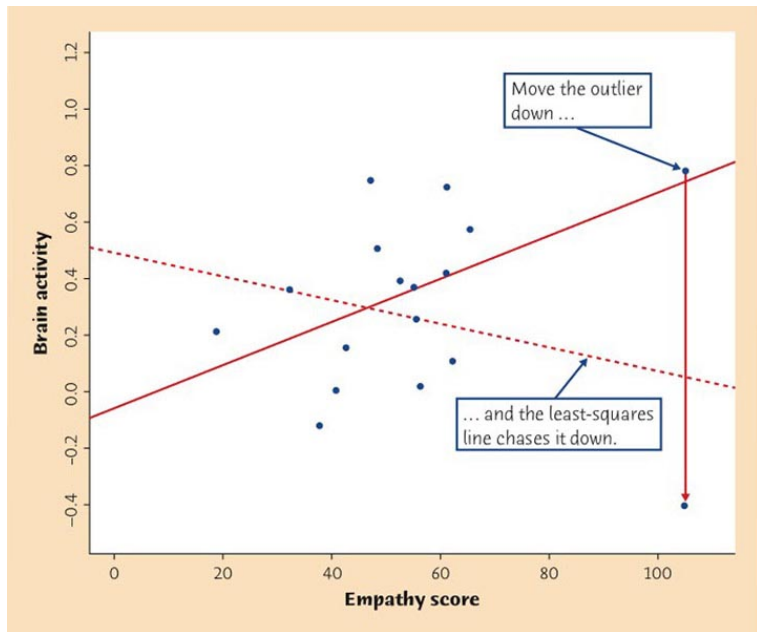


Assumption 2: Response varies around line according to a normal

What can go wrong?

- ▶ Outliers
- ▶ Exacerbated by high leverage points

Influential observations



Leverage (for bivariate regression)

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

Leverage (for bivariate regression)

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

Cooks distance – a measure of influence

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

Leverage (for bivariate regression)

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

Cooks distance – a measure of influence

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

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

Easy solutions

- ▶ Transforming variables (X_s and Y_s)

Easy solutions

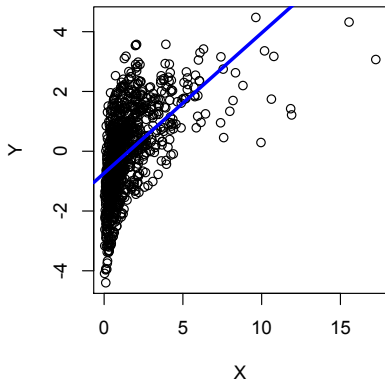
- ▶ Transforming variables (X s and Y s)
- ▶ Remove high-leverage outliers

Easy solutions

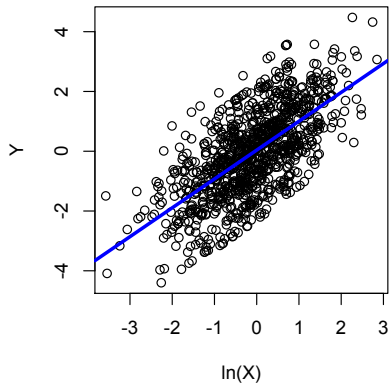
- ▶ Transforming variables (X s and Y s)
- ▶ Remove high-leverage outliers
- ▶ Trim the variable

Visualizing different Xs

Untransformed skewed X

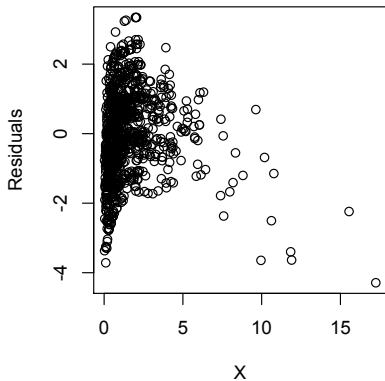


Log-transformed X

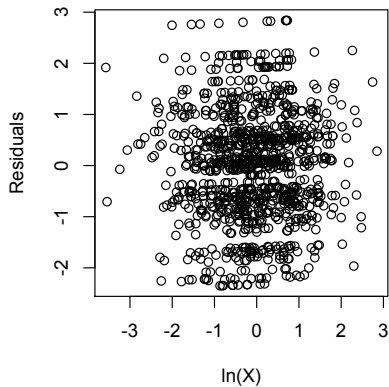


Visualizing different Xs

Untransformed skewed X

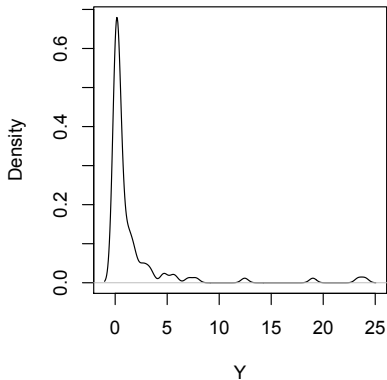


Log-transformed X

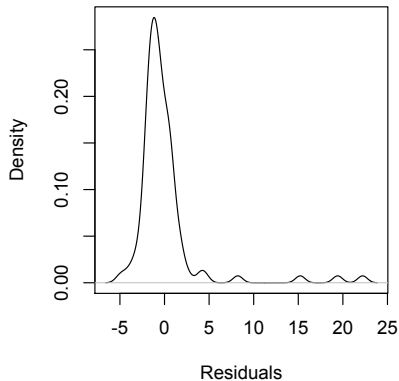


Transforming Y's

A skewed Y..

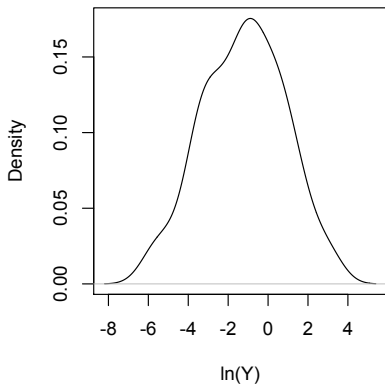


...leads to skewed residuals

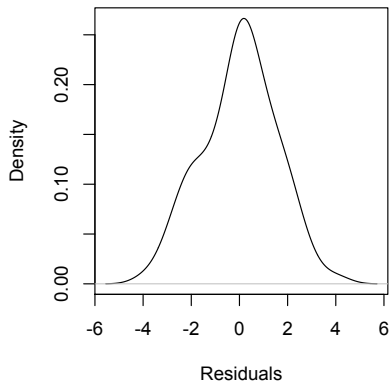


Transforming Y's

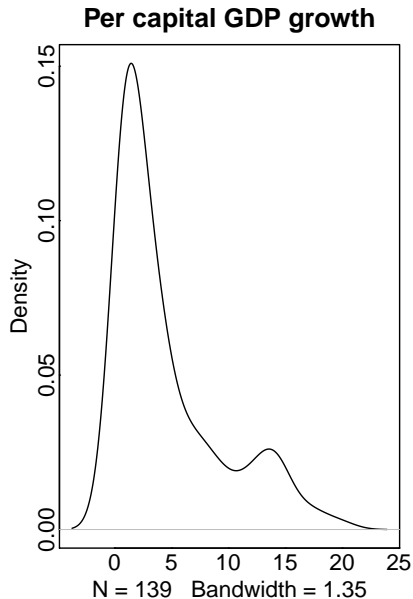
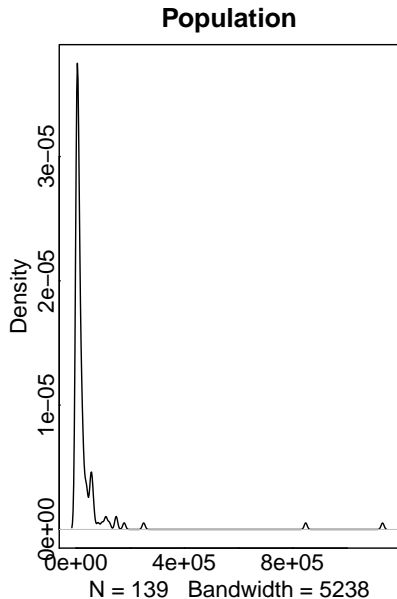
An appropriately transformed Y..



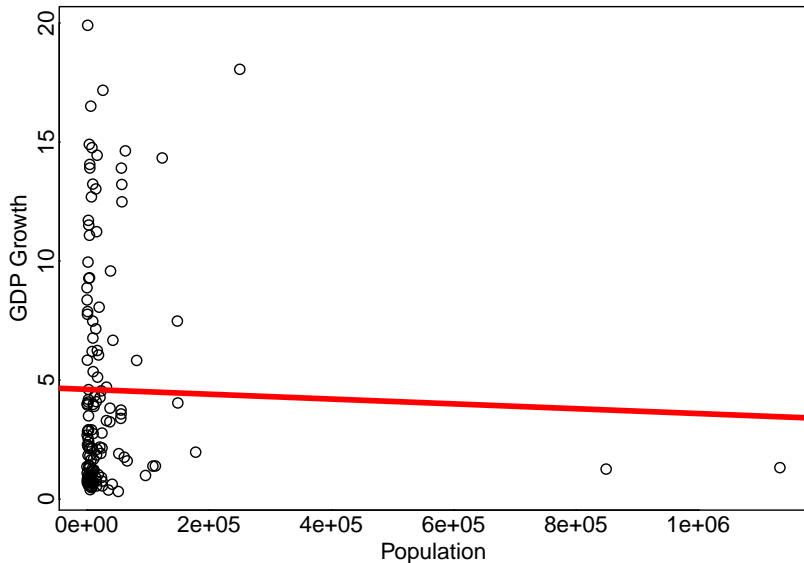
...leads to normal residuals



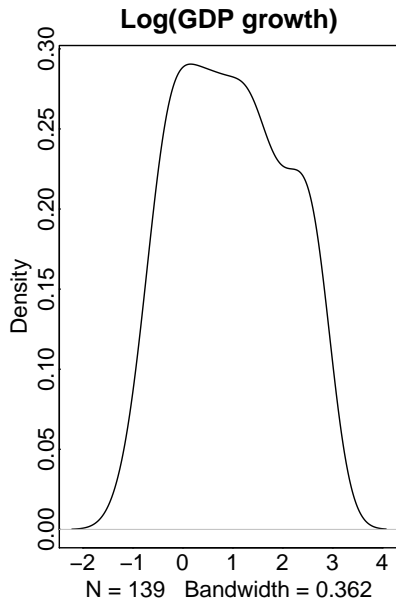
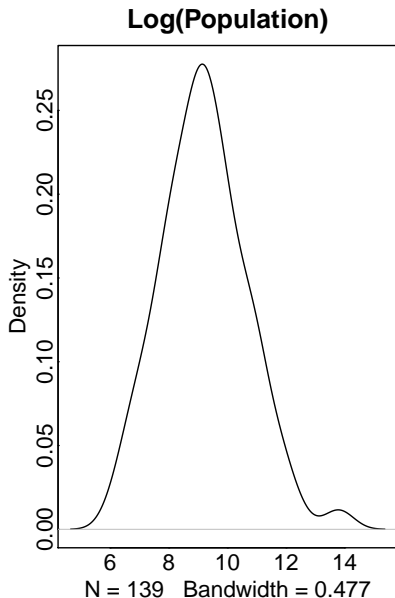
Example: GDP versus per capital GDP in 1990



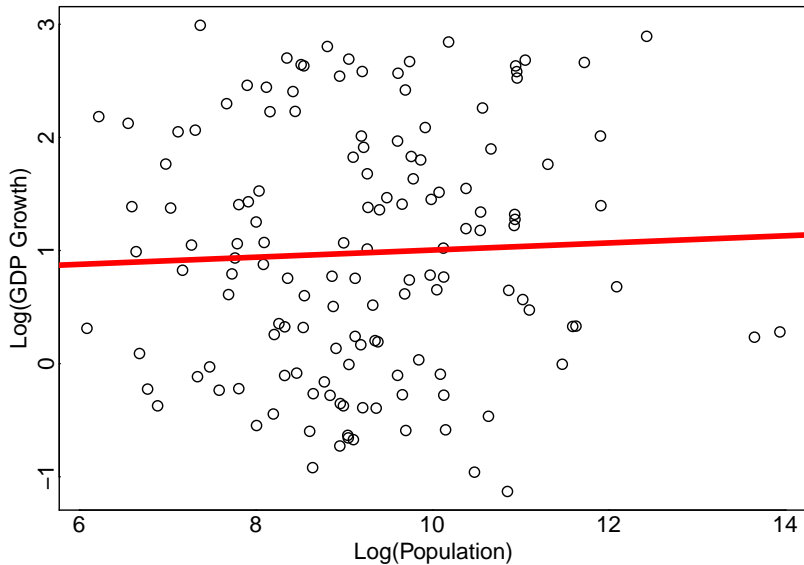
Example: GDP versus per capital GDP in 1990



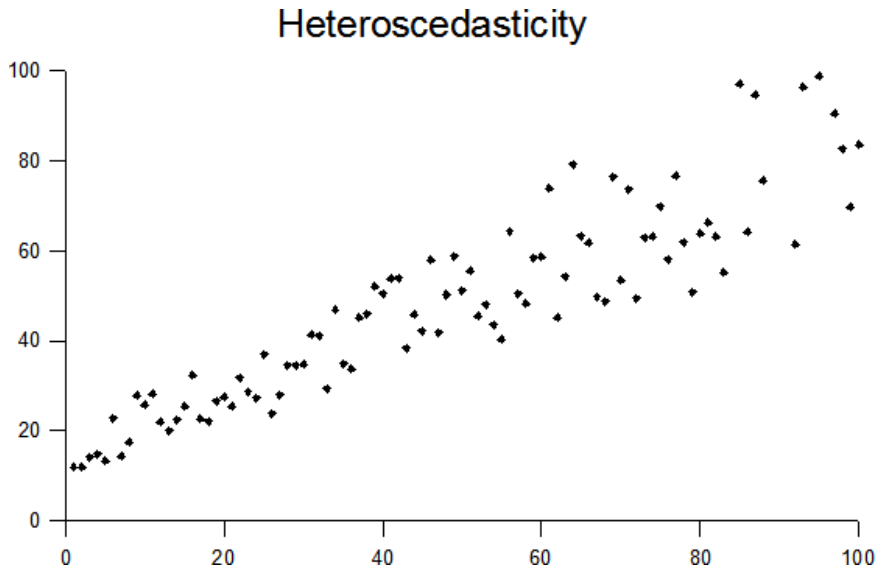
Example: GDP versus per capital GDP in 1990



Example: GDP versus per capital GDP in 1990

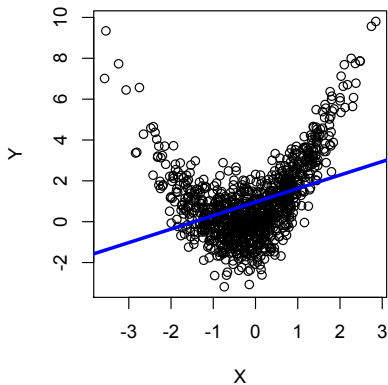


Heteroscedasticity: Variance changes based on value of x

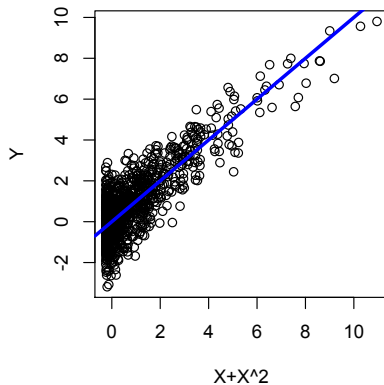


Problem 4: Nonlinear relationships

Unmodeled non-linear relationships

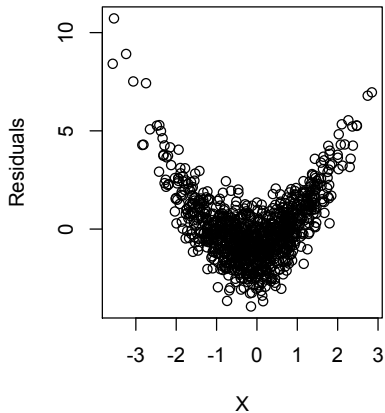


Modeled non-linear relationships

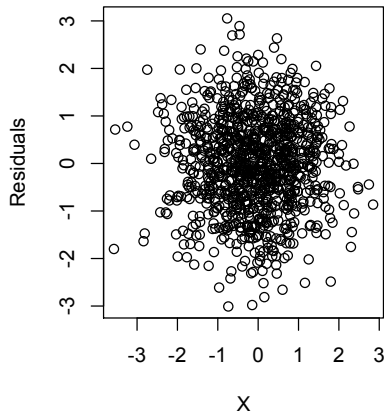


Solution: Model the relationship

Unmodeled non-linear relationship



Modeled non-linear relationships



Easy regression diagnostics in R

```
# The first regression
```

```
FL.LM1 <- lm(gdpen~pop, data=FL)
```

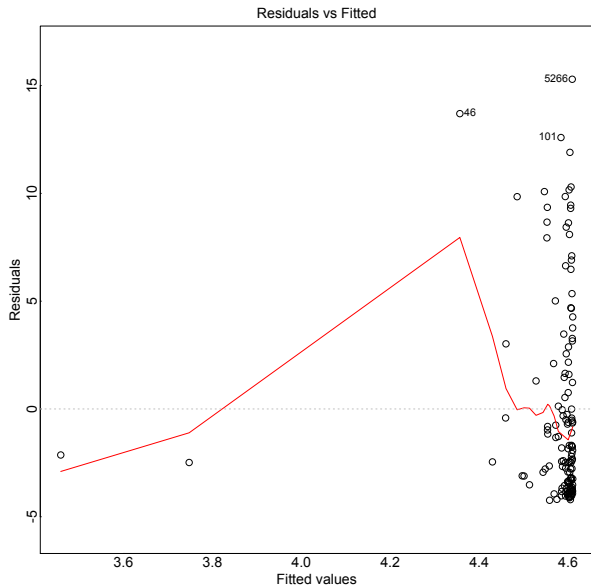
```
plot(FL.LM1)
```

```
# The second regression
```

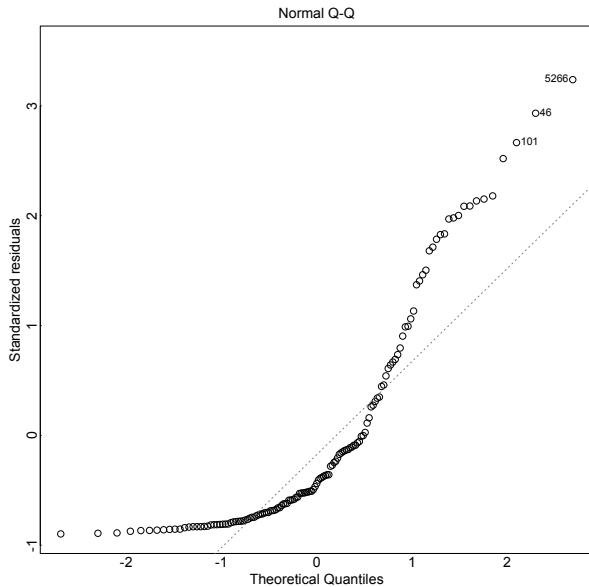
```
FL.LM2 <- lm(log(gdpen)~log(pop), data=FL)
```

```
plot(FL.LM2)
```

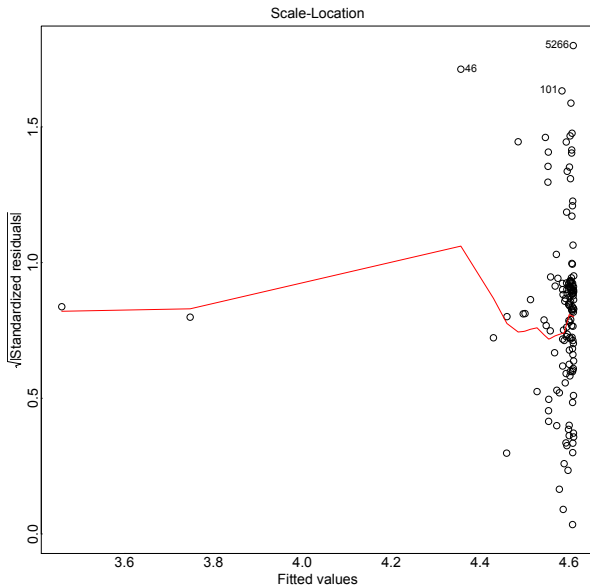
Bad models



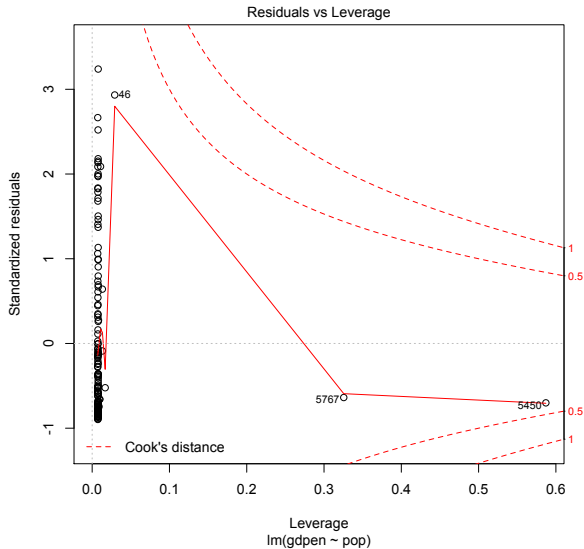
Bad models



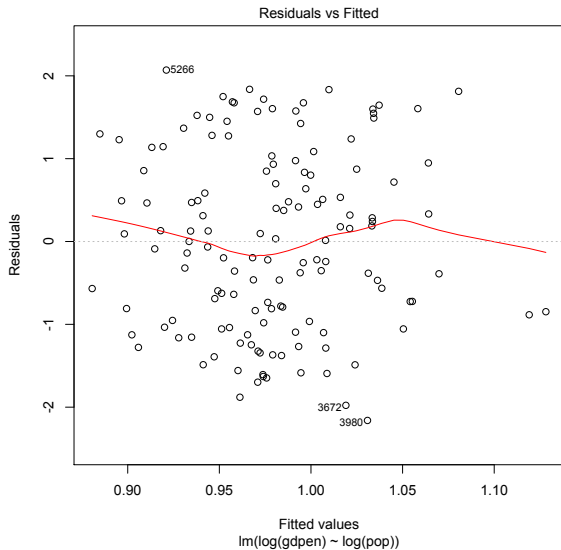
Bad models



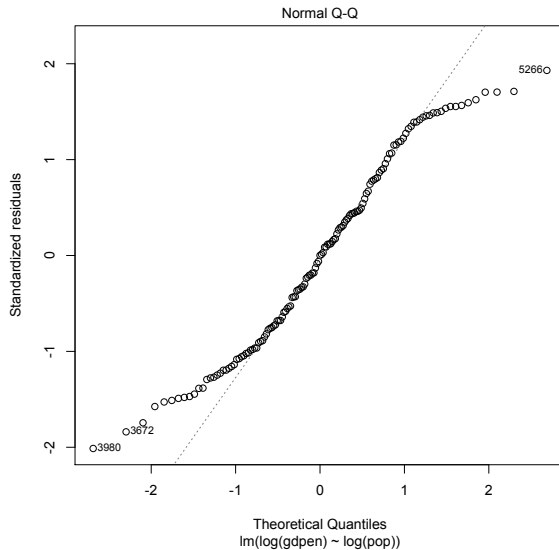
Bad models



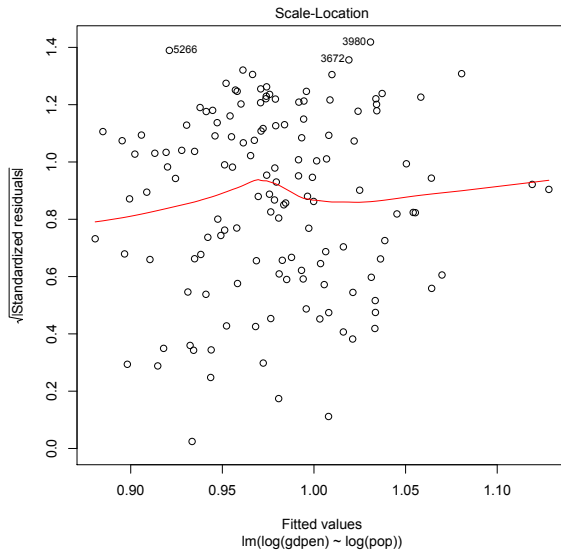
Better Model



Better Model



Better Model



Better Model

