Regression Overview

Paul E. Johnson[1]

[1]Departments of Political Science and Psychology (by courtesy), University of Kansas

August 24, 2020

The Big Overview

- Regression Examples
- Trouble
- Various data types
- Various relationships between input and output

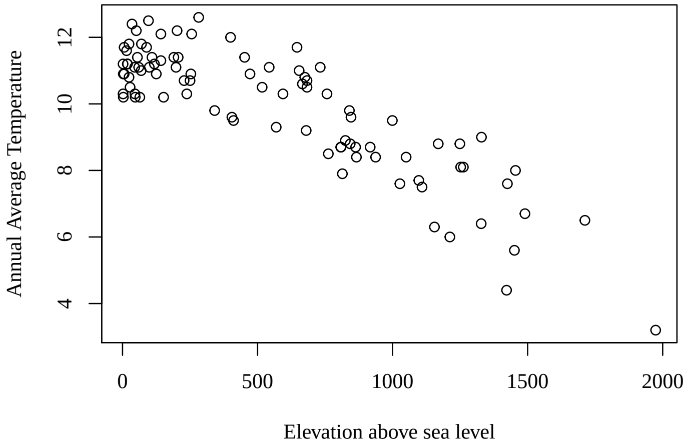
What is Regression?

- definition: predicting outcomes using a formula
- Predicted value of y, called $\hat{y}$ ("y hat")

$$\widehat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x1_i \quad (1)$$

  depends on a predictor variables $x1_i$ with two estimated parameters, $\hat{\beta}_0$ $\hat{\beta}_1$
- The data "comes from" a *data-generating process* in which "true" "unknown" values of β_0 and β_1 exist. We estimate with $\hat{\beta}_0$ and $\hat{\beta}_1$.

Example: The Temperature Across Oregon



You get a ruler, draw a line.

Choose your best line

- We might disagree about the “best line”: find objective criteria
- Afterwords, summarize our uncertainty
- Jargon to come: “standard errors” – “hypothesis test” – “confidence intervals”

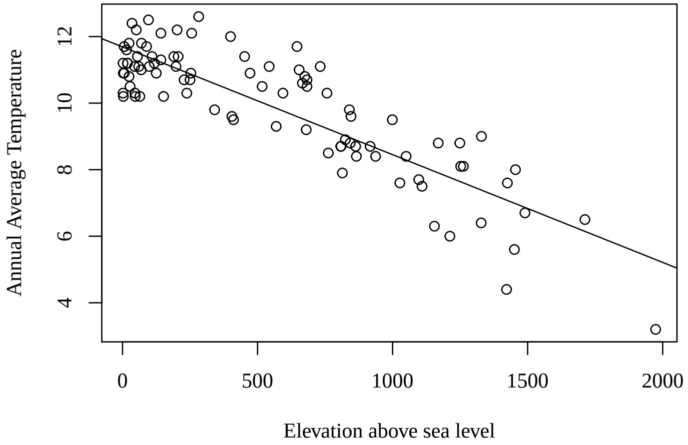
The Ordinary Least Squares Estimate

- OLS invented by Gauss more than 200 years ago
- The predicted value, AKA "line of best fit" is

$$\widehat{temperature}_i = 11.69 - 0.0032 \cdot elevation_i \tag{2}$$

- At sea level, the predicted temperature is 11.69
- For each additional foot of elevation, temperature declines by -0.0032.
- Maybe we'd re-scale, discuss 1000s of feet in elevation, so the effect would become -3.2 per 1000 feet.

Plot That


Annual Average Temperature
4
6
8
10
12
0
500
1000
1500
2000
Elevation above sea level

What looks ”about right?”

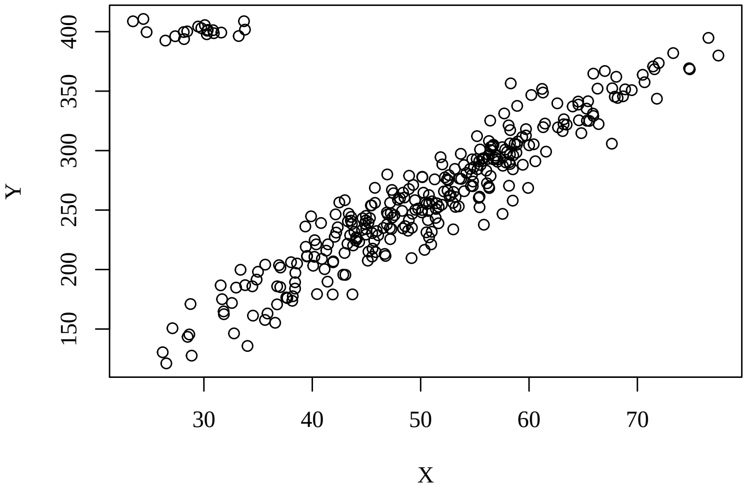
- The data cloud appears to be evenly dispersed above and below the line
- We’ll show you how to do that later in the semester
- Additional diagnostics can be done

We'll make nice looking tables

	M1	
	Estimate	(S.E.)
(Intercept)	11.688***	(0.150)
elevation (1000s feet)	-3.238***	(0.202)
N	92	
RMSE	0.958	
R^2	0.741	

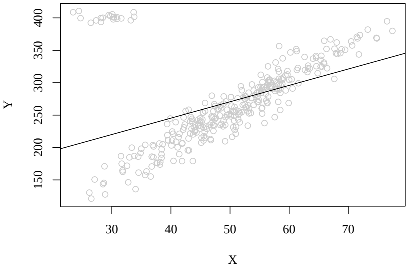
$*p \leq 0.05 ** \ p \leq 0.01 *** p \leq 0.001$

Trouble 1: Outlying observations

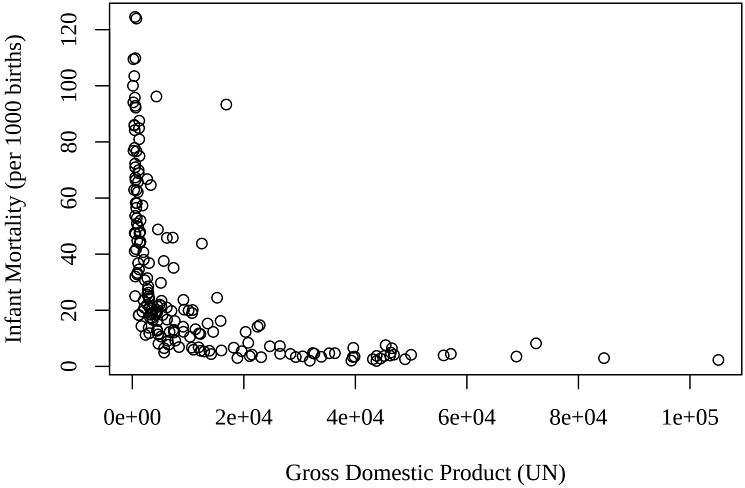


Checklist

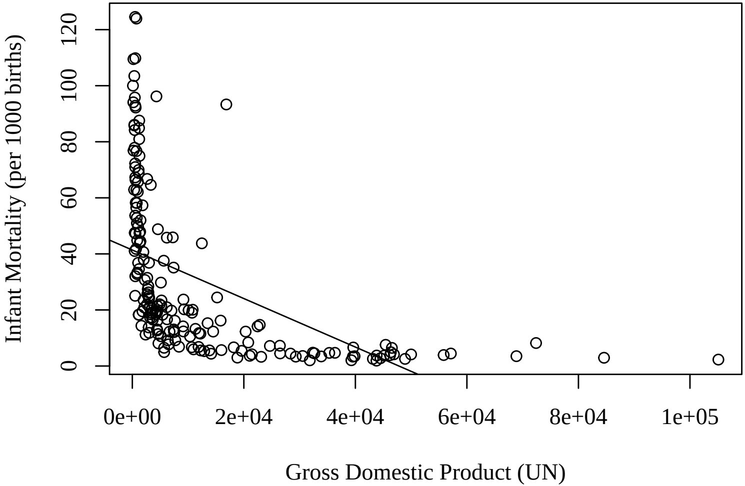
- Fit the line to the data "as is"
- Data is not "symmetrically dispersed" above and below
- "Ill-fitting cases" should be investigated
- Later we diagnose "influential" points.



Trouble 2: Nonlinearity



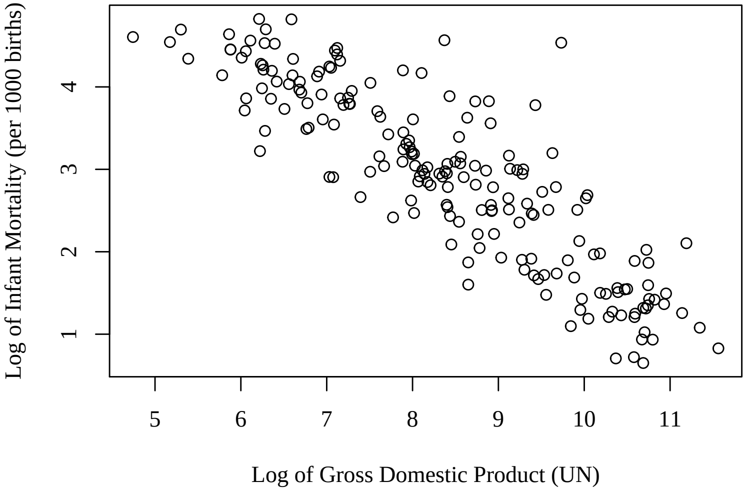
Linear Model Fits Worse than Last Year's Skinny Jeans



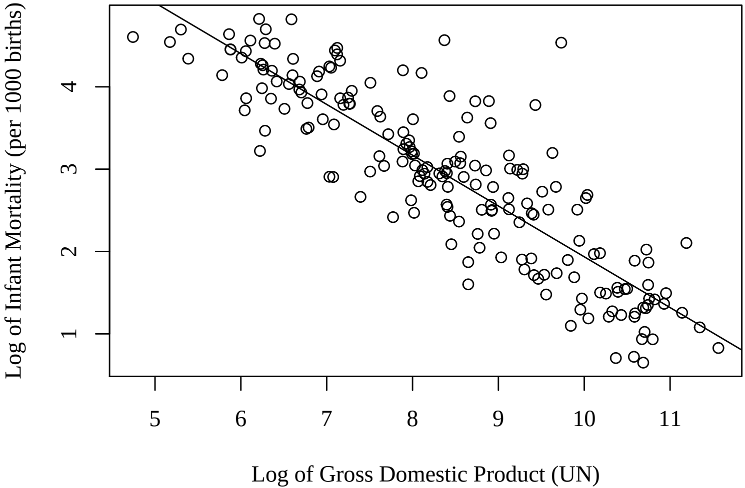
Two Modeling Strategies

- Transform the data to fit a straight line, or
- Transform the line to fit the curved data

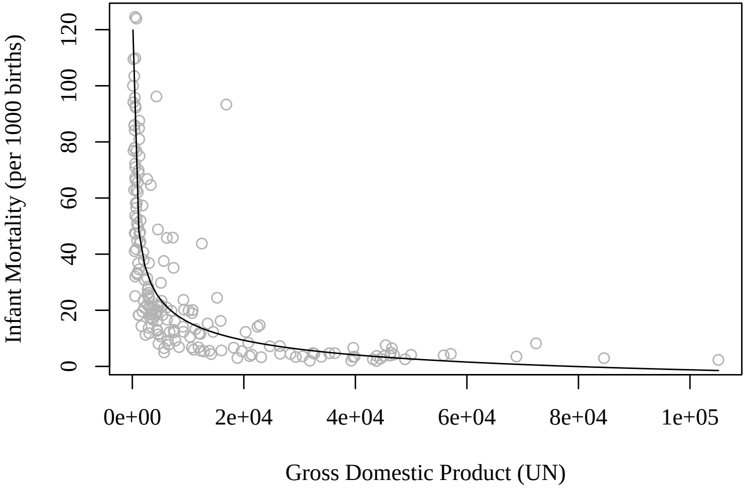
Transform The Data: Log both variables



Fit Linear model to the logged data



Bend the line into the data!

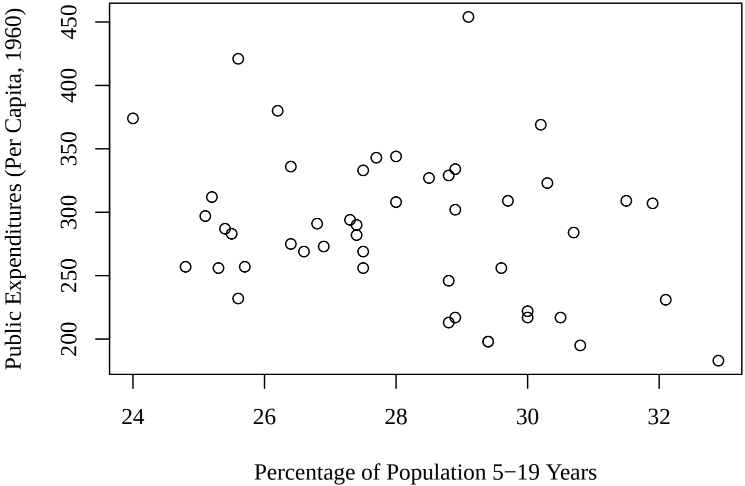


The Magic Trick is called Nonlinear Least Squares

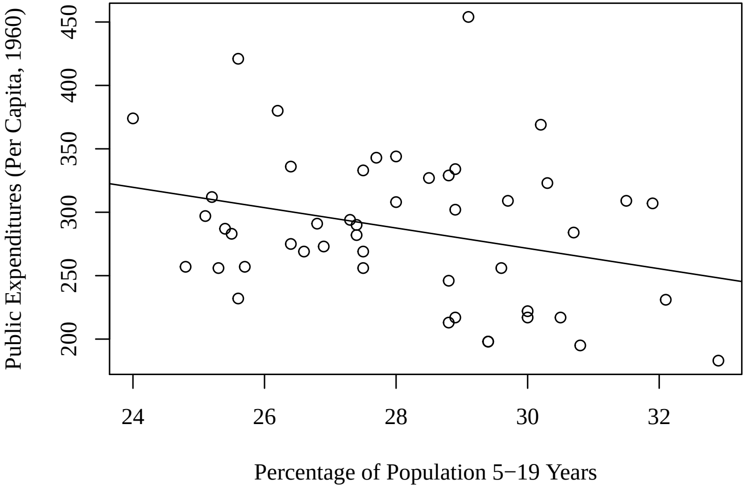
- Assume the "true relationship" is some formula
- Adjust the coefficient estimates to make the bending curve as close to the data.
- Fitted model I end up with is like this

$$inf.\widehat{mortality}_i = -90.23 + 336.6 \cdot \left(\frac{1}{x_i^{\frac{1}{8}}}\right)$$

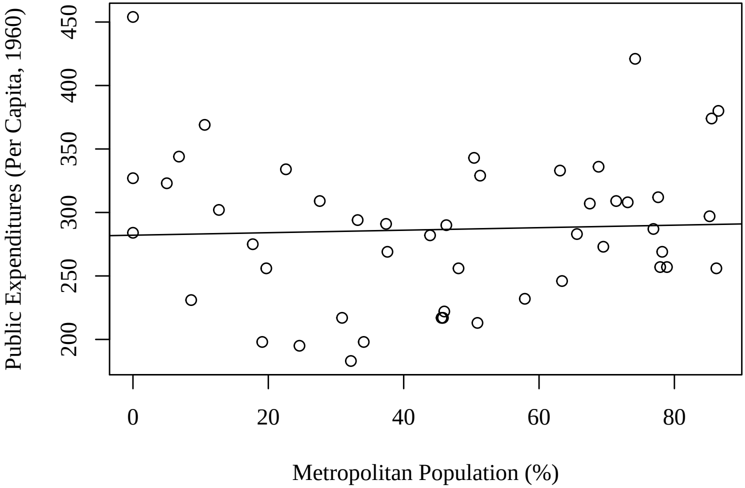
Another Example: Public Spending in 1960



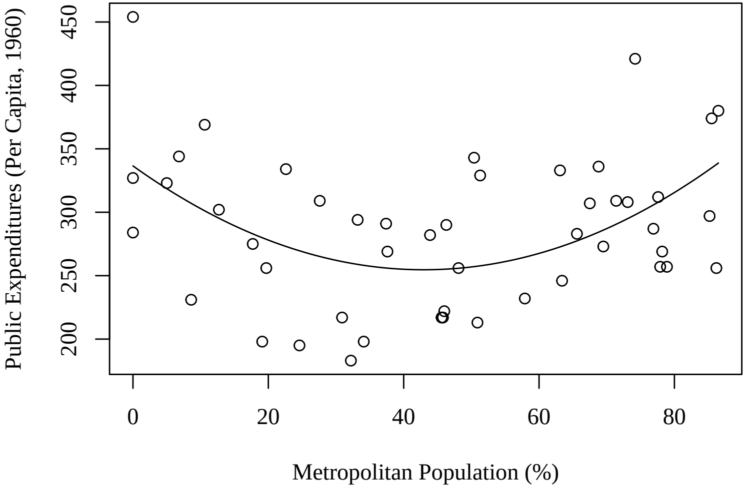
Public Expenditures: Maybe the Straight Line is OK



Metropolitan Population Effect: Linear?



Metropolitan Squared Makes Me Smile



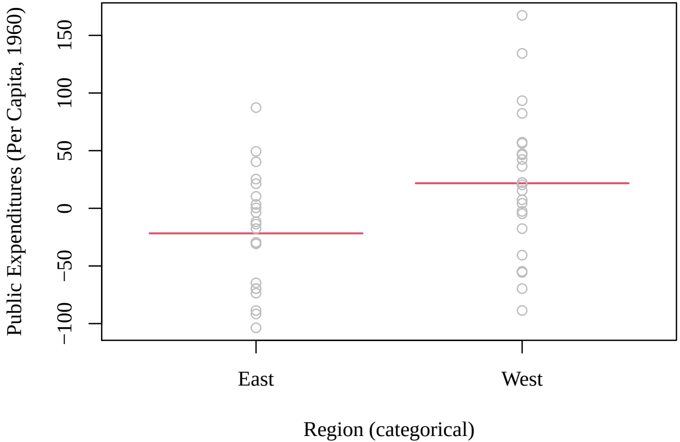
Trouble 3: Non-numeric predictors

Categorical variables

- "religious identification" {cath, prot, musl, jewi, hind, budi}
- Gender {male, female}
- Subjective scales {none, some, lots, plenty}
- We have ways to put those into regression models, usually by assigning them numerical scores and then interpreting them *VERY CAREFULLY*

West in the State Expenditures data

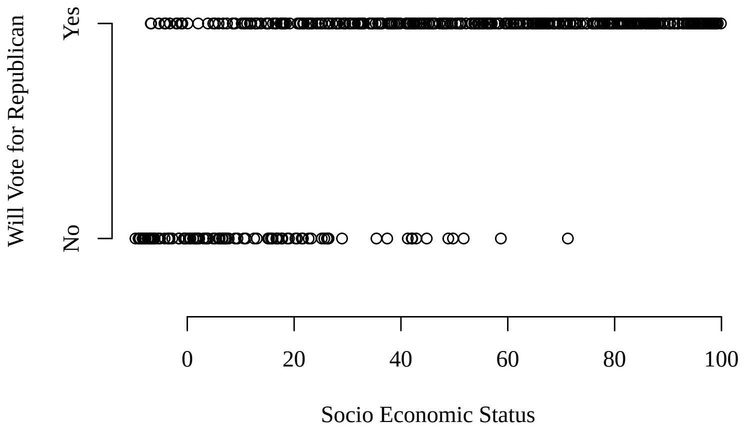
West coded 0="no" 1="yes". West coded as categorical variable (a.k.a "factor" in R, "class" in SAS)



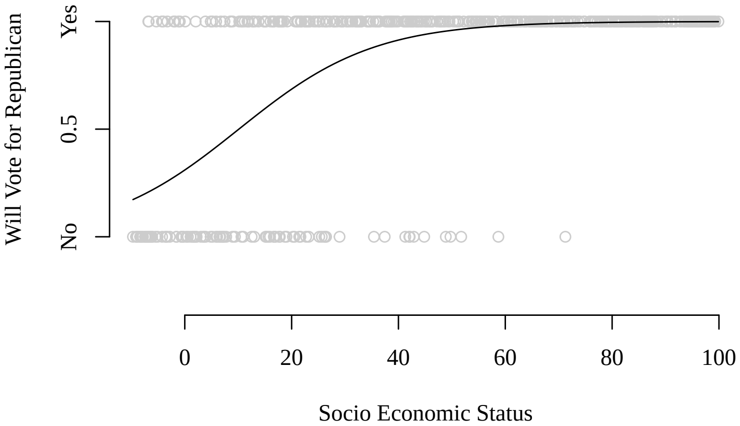
Trouble 4: Categorical Output

- This will be the very last topic we discuss in this semester.
- The output variable is a dichotomy
  - "yes" versus "no", "true" versus "false", "success" versus "failure"
- These are called "logistic regression models" (most common type)

Categorical Output



I'll Try to Sell You This S-shaped Probability Model



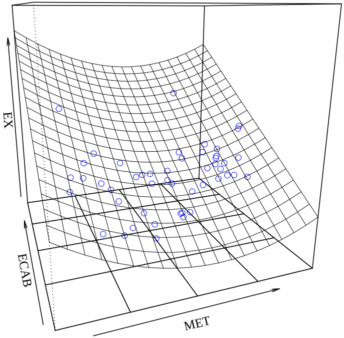
The Big Picture

- Wrestle back and forth between the data and the relationship you believe exists
- We are always trying to formulate predicted values and then re-evaluate the model that created them.
- Models explored in this case are a "first layer" of regression.
- After this, you'd want to study
  - categorical variables
  - random-effects models (Hierarchical models)
  - latent variable models (SEM)

rockchalk package

- Started developing that in 2010 for this class
- Includes essays “rockchalk” “RStyle” that I hope you might look at.
- Many graphing tools included, please run examples for plotSlopes and plotCurves and plotPlane

Basic 3-D plotting



Here's the table that goes with that, incidentally

	M1	
	Estimate	(S.E.)
(Intercept)	71.644	(130.027)
ECAB	1.813***	(0.333)
poly(MET, 2)1	-81.909	(52.684)
poly(MET, 2)2	123.039**	(42.723)
YOUNG	1.407	(4.020)
N	48	
RMSE	40.190	
R^2	0.573	
adj R^2	0.533	

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