Refining the Model

- Key Assumption: Model sufficiently describes the regression of Y on X
- misleading to omit important variables (bias)
- Based on Occam's Razor, we may want to eliminate non-essential variables to find a parsimonious model
- After resolving transformations of variables and outliers, analysis focuses on model simplification

Deforestation and Debt

- Theory: Developing countries cut down forests to pay foreign debt.
- Data from 11 Latin American countries
- Variables:
  - Country
  - Debt (millions of dollars)
  - Deforestation (thousands of hectares)
  - Population (thousands of people)

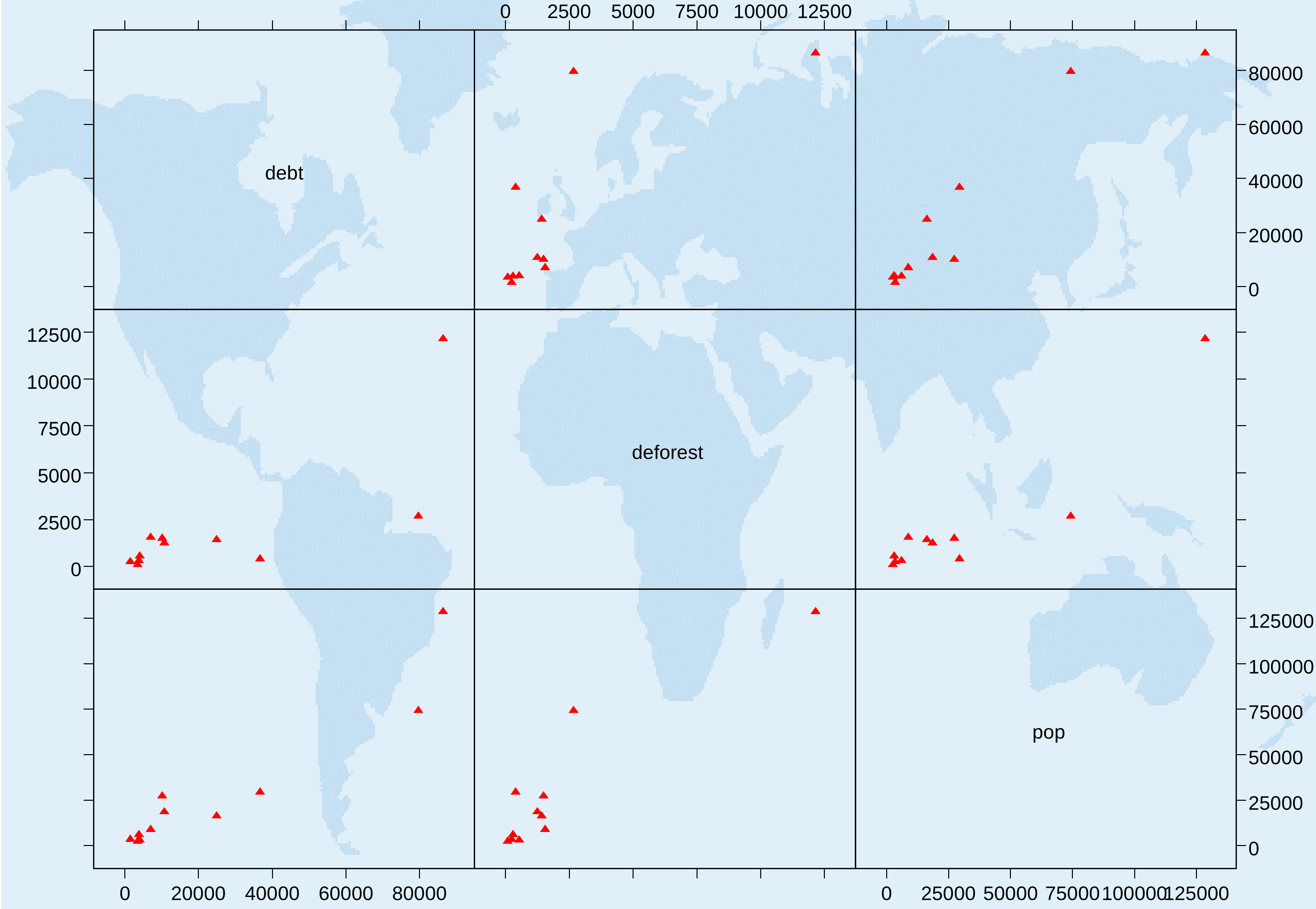
Questions

- Does the evidence significantly support the theory that debt causes deforestation?
  - can't answer because of observational data
  - may have confounding variables

- Does debt exert any effect after the effect of population size on deforestation is accounted for?

Use multiple regression to address

Deforestation Rates and Debt for 11 Latin American Countries



Transformations?

- Scatterplot matrix indicates nonlinear relationship between
  - debt and deforestation
  - deforestation and population
- Try log transformations of all variables
  - common procedure in economic models
  - relative ease of interpretation in terms of median
  - if model shows evidence of lack of fit, then try another transformation

Statistical Model

- $\log(\text{deforestation}) = \beta_0 + \beta_1 \log(\text{pop})_i + \beta_2 \log(\text{debt}_i) + \varepsilon_i$
- Hypothesis of interest is Ho: $\beta_2 = 0$
- Steps before testing?
  - Check residuals (constant variance & normality)
  - Check for influential cases Cook'sdistance
  - Check for outliers Studentized residuals

Estimates and P-values

Coefficients:

	Value	Std. Error	t value	Pr(>|t|)
(Intercept)	-1.3855532	1.8527158	-0.7478498	0.4759571
log.pop	1.0722430	0.5293857	2.0254476	0.0774019
log.debt	**-0.2208269**	**0.5155882**	**-0.4283009**	**0.6797300**

Residual standard error: 0.7966867 on 8 degrees of freedom

Multiple R-Squared: 0.7160206

F-statistic: 10.08553 on 2 and 8 degrees of freedom, the p-value is 0.006503506

First step is to look at the overall F-test. This implies that at least one of the regression coefficients (for logdebt or logpop) is nonzero

Now we can look at testin whether log(debt) exerts any influence after adjusting fo log(pop). P-value above implies that there is no significant effect.

Partial Residual Plots

- Negative coefficient may seem strange at first... because the scatterplot of debt and deforestation shows a positive association
- The coefficient in the multiple regression shows the effect of debt on deforestation, AFTER accounting for population size.
- Visually, a partial residual plot shows the relationship, after adjusting for population
- Can identify non-linear relationships

Partial Residuals

Suppose the model is

$$\mu(deforestation | pop, debt) = \beta_0 + \beta_1 \log(pop) + f(\log(debt))$$
$$f(\log(debt)) = \mu(deforestation | pop, debt) - \beta_0 - \beta_1 \log(pop)$$

where f(log(debt)) is an unknown function

would like to know what f(log(debt) is

Would like to plot

- log(deforestation) - β_0 - β_1 log(pop) versus
- log(debt) to explore

Partial Residuals

- Don't know the true coefficients
- Can use estimates, but which ones?
  - The ones from the regression on log(pop) alone are not correct, as they do not take into account debt
  - approximate f(log(debt)) by β_2 log(debt) and use the estimates of β_0 and β_1 from the multiple regression to construct the partial residual

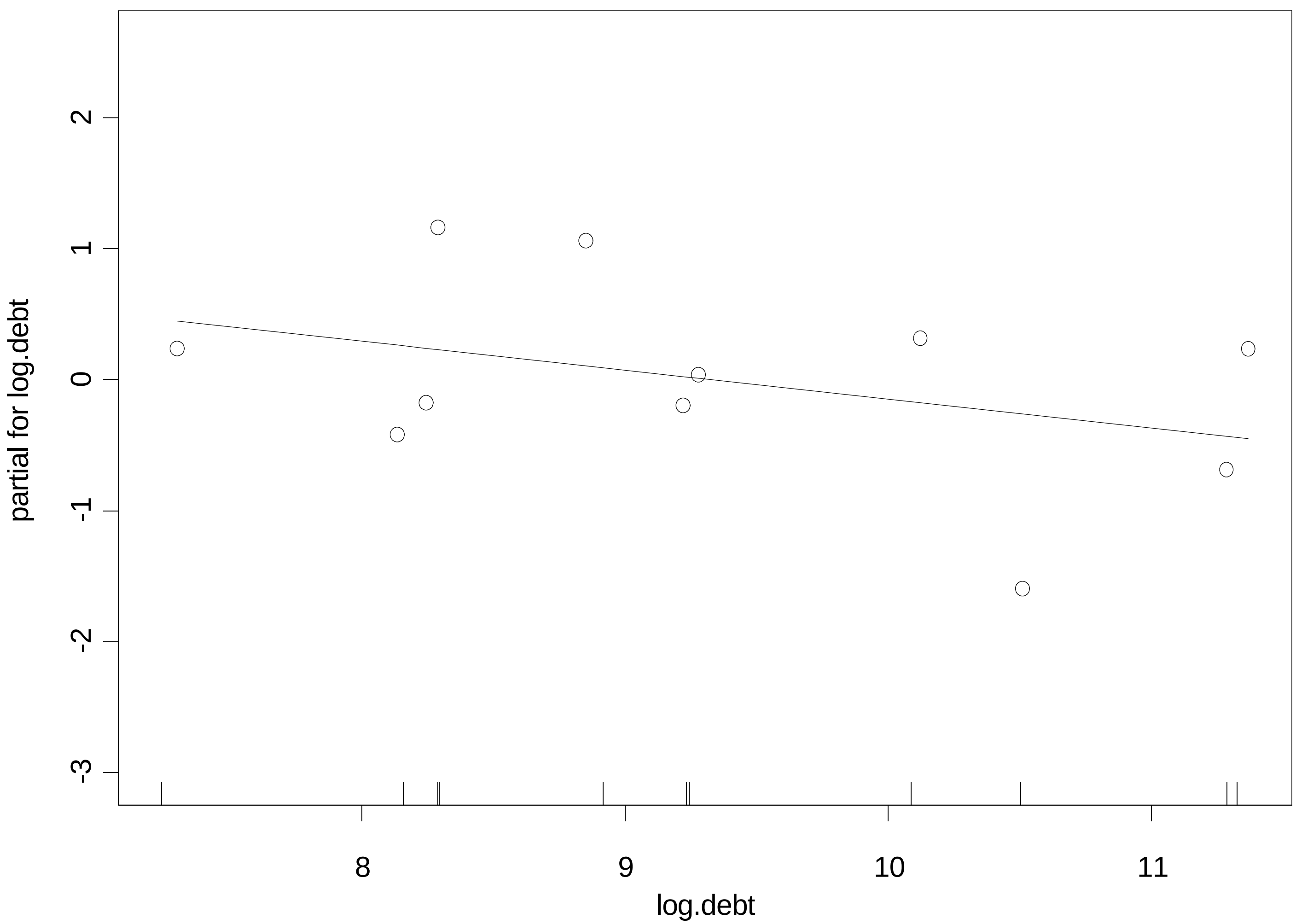
Partial Residual Plots

- Obtain the usual residuals, e_i, for fitting log(deforest) on log(pop) and log(debt)
- partial residual for debt is

$$pres_i = \log(deforestation_i) - \hat{\beta}_0 - \hat{\beta}_1 \log(pop_i)$$
$$pres_i = e_i + \hat{\beta}_2 \log(debt_i)$$

Plot partial residual versus log(debt)

Partial Residual Plot for log(debt)



Conclusions

- There is no evidence of an association beween debt and deforestation, after accounting for population (p-value = 0.68)
- No evidence of outliers or influential cases that would affect the conclusions