STAT/SOC/CSSS 221
Statistical Concepts and Methods for the Social Sciences

Introduction to Mulitple Regression

Christopher Adolph

Department of Political Science
and
Center for Statistics and the Social Sciences
University of Washington, Seattle

Motivating Example: Cross-national determinants of fertility

We have cross-national data from several sources:

Fertility The average number of children born per adult female, in 2000 (United Nations)

Education Ratio The ratio of girls to boys in primary and secondary education, in 2000 (Word Bank Development Indicators)

GDP per capita Economic activity in thousands of dollars, purchasing power parity in 2000 (Penn World Tables)

Agricultural Labor Percentage of the labor force working in agriculture in 2000 (International Labor Organization)

Note the addition of a fourth variable

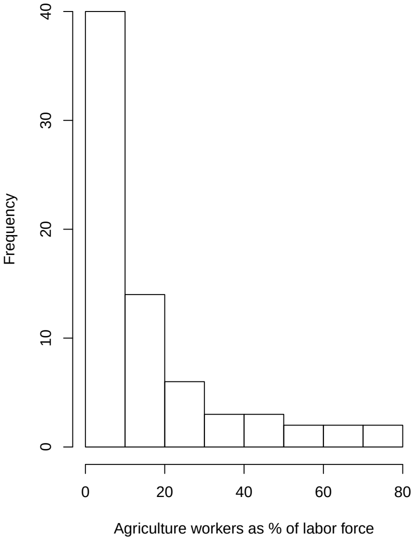
Motivating Example: Cross-national determinants of fertility

All three independent variables might cause the fertility rate

More agricultural nations may have more children to bolster the labor force on family farms

Let's look at the univariate summaries & bivariate regression results for this new covariate

Summary of Univariate Distribution: Agricultural Labor

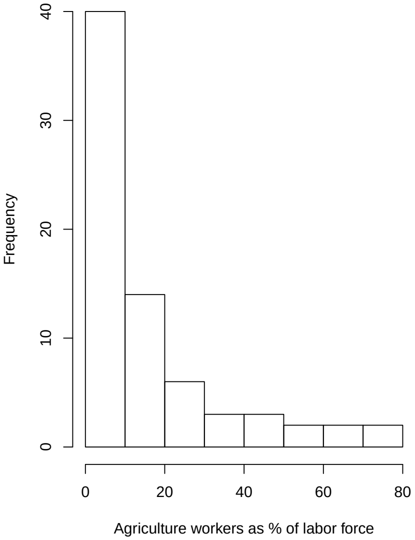


Median = 8.1%

Mean = 16.0 %

std dev = 17.9%

Summary of Univariate Distribution: Agricultural Labor



Median = 8.1%

Mean = 16.0 %

std dev = 17.9%

How would you describe this distribution?

Regression of Fertility on Agricultural Labor

Variable	Estimates	se	t-stat	p-value
Intercept	1.83	(0.15)	12.34	< 0.001
Agricultural Labor	0.02	(0.01)	3.52	< 0.001
N	72			
R^2	0.15			
RMSE	0.93			

How do we read this table?

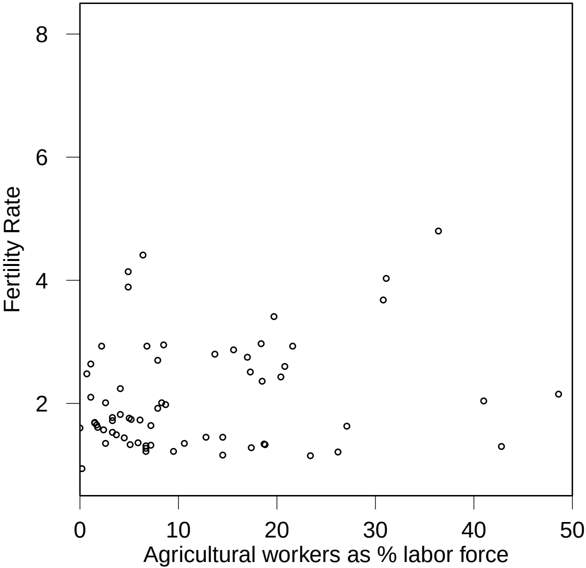
Regression of Fertility on Agricultural Labor

Variable	Estimates	se	t-stat	p-value
Intercept	1.83	(0.15)	12.34	< 0.001
Agricultural Labor	0.02	(0.01)	3.52	< 0.001
N	72			
R^2	0.15			
RMSE	0.93			

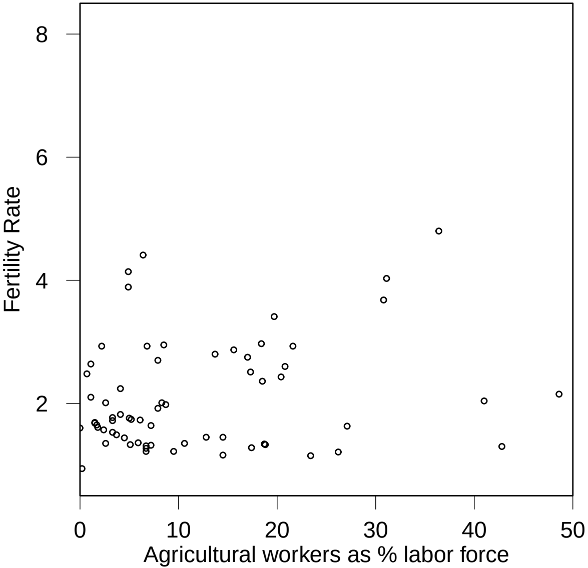
How do we read this table?

Note the reduction in N: lots of cases are missing data on agricultural labor

Any cases missing *any* covariates need to be deleted from the data before using regression (listwise deletion)

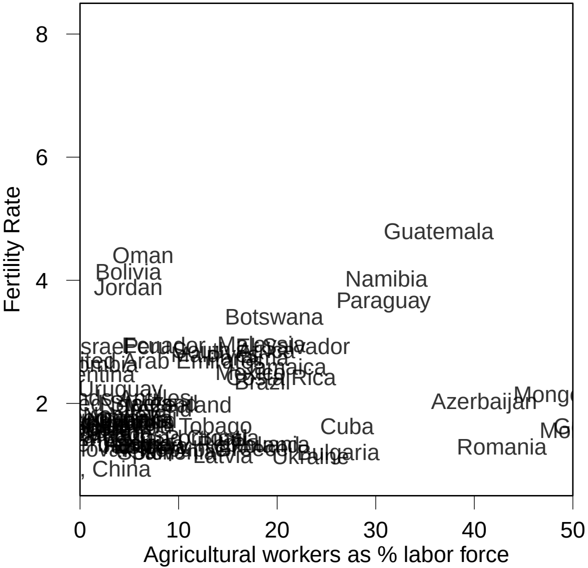


What looks different about this scatter-plot?

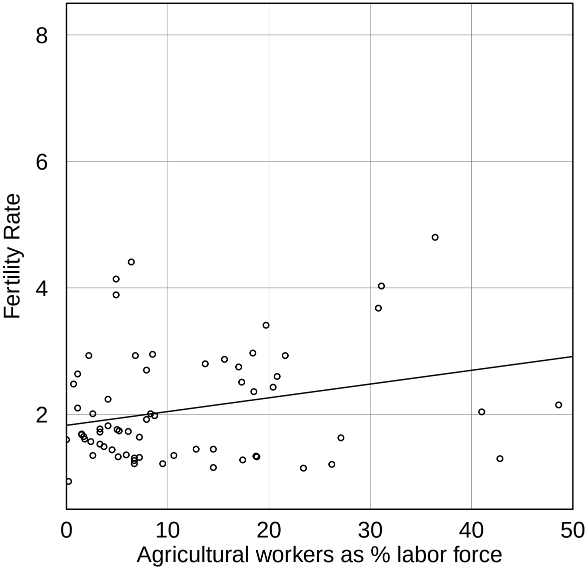


What looks different about this scatter-plot?

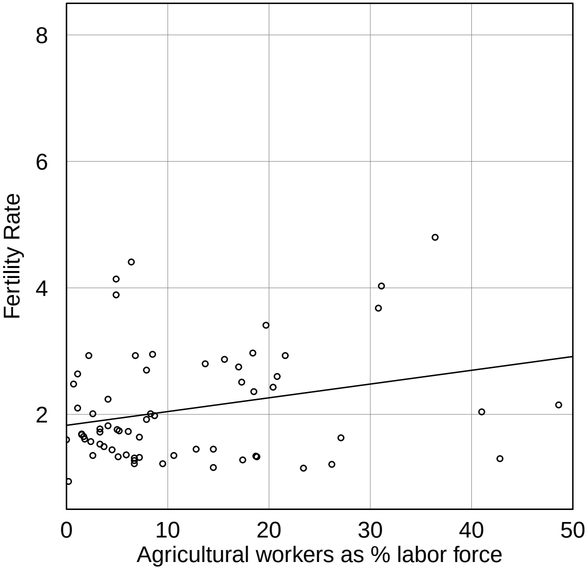
The high fertility cases seem to be missing (deleted due to missing data)



Fertility Rate
Agricultural workers as % labor force
0
10
20
30
40
50
2
4
6
8
Guatemala
Oman
Bolivia
Jordan
Namibia
Paraguay
Botswana
El Salvador
Jamaica
Rica
Brazil
Uruguay
Tobago
Cuba
Azerbaijan
Romania
Bulgaria
Ukraine
Latvia
China

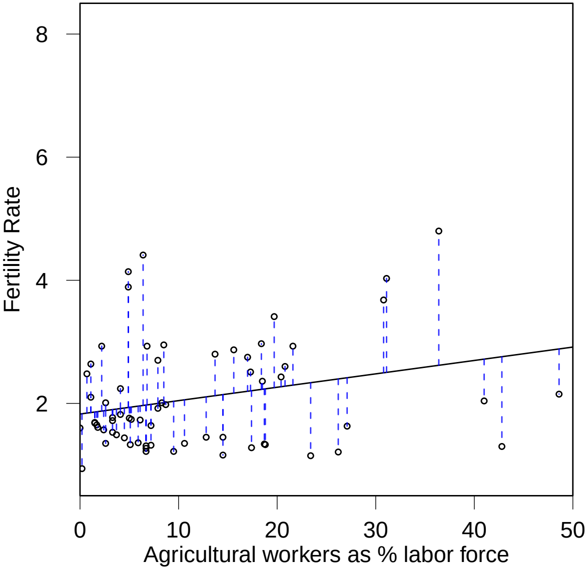


Is this a strong relationship?

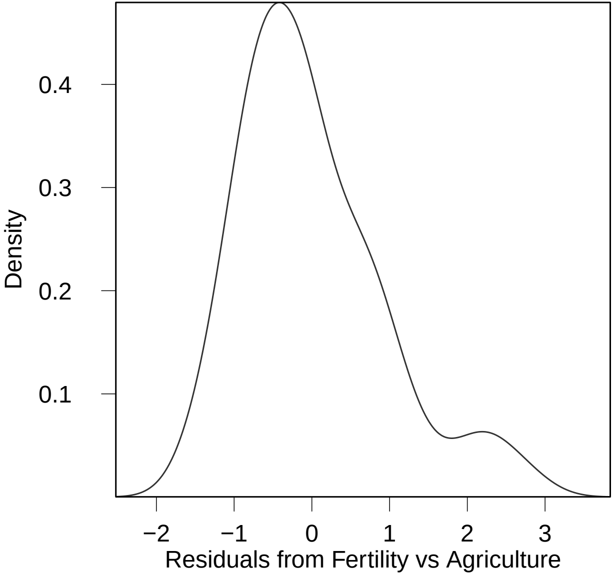


Is this a strong relationship?

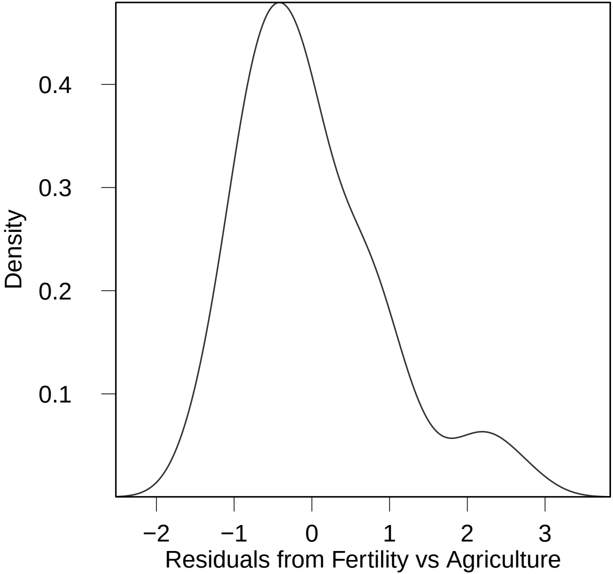
How many datapoints would have to move to reduce the slope to 0?



Which are larger, the residuals or the explained variance?

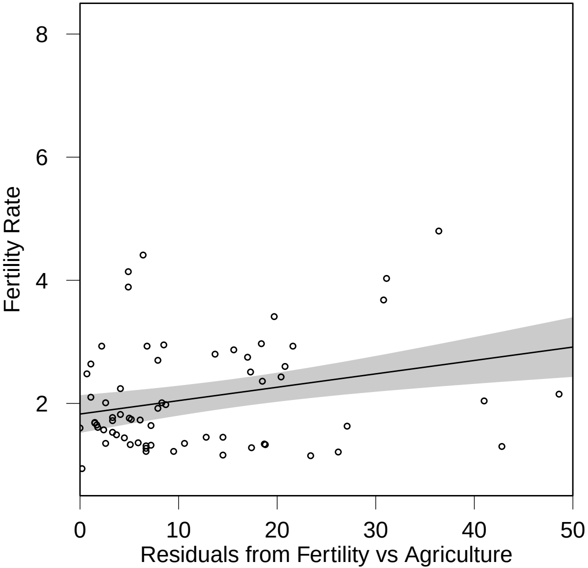


What is the standard deviation of this distribution called?

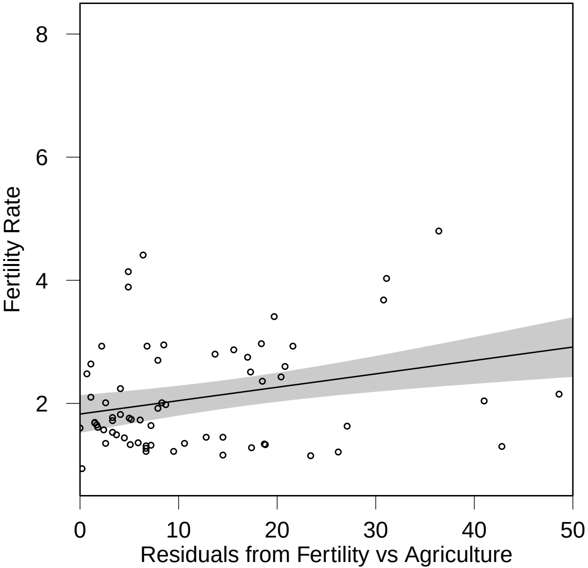


What is the standard deviation of this distribution called?

The RMSE, or standard error of the regression: how much predictions from this model tend to miss by



How confident are we that this line has a positive slope?



How confident are we that this line has a positive slope?

Are we as confident as we were for the other models?

Confounders and Omitted Variable Bias

Which (if any) of the three models we've looked at are right?

Do Education, GDP, and Ag Labor all affect Fertility?

Confounders and Omitted Variable Bias

Which (if any) of the three models we've looked at are right?

Do Education, GDP, and Ag Labor all affect Fertility?

What if Education, GDP, and Ag Labor are *correlated*?

Confounders and Omitted Variable Bias

Which (if any) of the three models we've looked at are right?

Do Education, GDP, and Ag Labor all affect Fertility?

What if Education, GDP, and Ag Labor are *correlated*?

If we regress Fertility on Education, and Education is correlated with GDP and Ag, might it proxy all three variables?

Confounders and Omitted Variable Bias

Which (if any) of the three models we've looked at are right?

Do Education, GDP, and Ag Labor all affect Fertility?

What if Education, GDP, and Ag Labor are *correlated*?

If we regress Fertility on Education, and Education is correlated with GDP and Ag, might it proxy all three variables?

Yes: if countries which educate women also tend to be rich and have few ag workers, then the bivariate results will blur all three relationships

Confounders and Omitted Variable Bias

Should we be worried?

Correlation between:

Education & GDP is 0.46

Confounders and Omitted Variable Bias

Should we be worried?

Correlation between:

Education & GDP is 0.46

Correlation between GDP & Ag is -0.64

Confounders and Omitted Variable Bias

Should we be worried?

Correlation between:

Education & GDP is 0.46

Correlation between GDP & Ag is -0.64

Correlation between Education & Ag is -0.41

(What do these numbers mean?)

Confounders and Omitted Variable Bias

Should we be worried?

Correlation between:

Education & GDP is 0.46

Correlation between GDP & Ag is -0.64

Correlation between Education & Ag is -0.41

(What do these numbers mean?)

Confounders and Omitted Variable Bias

Should we be worried?

Correlation between:

Education & GDP is 0.46

Correlation between GDP & Ag is -0.64

Correlation between Education & Ag is -0.41

(What do these numbers mean?)

Omitted variable bias: Leaving any of these variables out of our model could lead to misleading estimates of the effects of any variables we do include

The linear regression model, redux

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_k x_{ki} + \varepsilon_i$$

Our dependent variable likely depends on many covariates

For example,
x_1 might be the education ratio,
x_2 might be GDP per capita,
and so on for as many covariates as we have, up to our kth covariate

This leads to the above model, with multiple *partial* slopes β_1, β_2, $\ldots \beta_3$

The linear regression model, redux

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_k x_{ki} + \varepsilon_i$$

This model is still a linear regression model.

Sometimes called a "multiple" regression model to distinguish from a "bivariate" regression, but mathematically, they are equivalent

Henceforth, we will assume a linear regression can have many covariates

How many covariates are allowed?
Up to $N - 1$, where N is the number of observations

Each covariate added uses up a degree of freedom; once they are gone, there is nothing left for an additional covariate to explain

The linear regression model, redux

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_k x_{ki} + \varepsilon_i$$

How do we interpret the β's? Just as before.

The linear regression model, redux

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_k x_{ki} + \varepsilon_i$$

How do we interpret the β's? Just as before.

The β's are still slopes, or the amount y_i changes on average for a 1 unit increase in x, *all else held equal*

The linear regression model, redux

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_k x_{ki} + \varepsilon_i$$

How do we interpret the β's? Just as before.

The β's are still slopes, or the amount y_i changes on average for a 1 unit increase in x, *all else held equal*

If we increase x_1 by 1 unit,
and hold x_2 fixed at its present level,
then y goes up by β_1

The linear regression model, redux

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_k x_{ki} + \varepsilon_i$$

How do we interpret the β's? Just as before.

The β's are still slopes, or the amount y_i changes on average for a 1 unit increase in x, *all else held equal*

If we increase x_1 by 1 unit,
and hold x_2 fixed at its present level,
then y goes up by β_1

$\rightarrow$ Finally found a way to control for confounders using observational data!

Aside for calculus-users: The β's are partial derivatives with respect to the x they multiply

Multiple regression: just like bivariate

1. Our estimates, $\hat{\beta}_k$, are the β_k's that minimize the sum of the squared residuals (least squares)

Multiple regression: just like bivariate

1. Our estimates, $\hat{\beta}_k$, are the β_k's that minimize the sum of the squared residuals (least squares)

2. The uncertainty of each $\hat{\beta}_k$ is given by its standard error

Multiple regression: just like bivariate

1. Our estimates, $\hat{\beta}_k$, are the β_k's that minimize the sum of the squared residuals (least squares)
2. The uncertainty of each $\hat{\beta}_k$ is given by its standard error
3. We can still perform t-tests and calculate confidence intervals for each $\hat{\beta}_k$

Multiple regression: just like bivariate

1. Our estimates, $\hat{\beta}_k$, are the β_k's that minimize the sum of the squared residuals (least squares)
2. The uncertainty of each $\hat{\beta}_k$ is given by its standard error
3. We can still perform t-tests and calculate confidence intervals for each $\hat{\beta}_k$
4. We can still calculate the fitted value $\hat{y}_i$ of any observation i: this is the model prediction for that case

Multiple regression: just like bivariate

1. Our estimates, $\hat{\beta}_k$, are the β_k's that minimize the sum of the squared residuals (least squares)
2. The uncertainty of each $\hat{\beta}_k$ is given by its standard error
3. We can still perform t-tests and calculate confidence intervals for each $\hat{\beta}_k$
4. We can still calculate the fitted value $\hat{y}_i$ of any observation i: this is the model prediction for that case
5. We can still summarize goodness of fit using such measures as RMSE and R^2

Fertility as function of Education and GDP per capita

Let's start small: a model with two covariates:

$$\begin{aligned}\widehat{\text{Fertility}}_i &= \hat{\beta}_0 + \hat{\beta}_1 \text{EduRatio}_i + \hat{\beta}_2 \text{GDPpc}_i \\ \widehat{\text{Fertility}}_i &= 11.24 - 0.08 \times \text{EduRatio}_1 - 0.05 \times \text{GDPpc}_i\end{aligned}$$

We can present this result in several ways:

1. In a table by itself
2. In a table compared to other models
3. Through graphics

Regression of Fertility on Education Ratio & GDP

Variable	Estimates	se	t-stat	p-value
Intercept	11.25	(0.73)	15.46	< 0.001
Education Ratio	-0.08	(0.01)	-9.93	< 0.001
GDP per capita ($k)	-0.05	(0.01)	-5.32	< 0.001
N	130			
R^2	0.64			
RMSE	1.01			

How do we interpret the above?

Three regression models of fertility

	Model		
Variable	1	2	3
Intercept	12.59	4.13	11.25
	(0.75)	(0.17)	(0.73)
Education Ratio	−0.10		−0.08
	(0.01)		(0.01)
GDP per capita		−0.10	−0.05
		(0.01)	(0.01)
N	130	130	130
R^2	0.55	0.35	0.64
RMSE	1.12	1.35	1.01

Standard errors in parentheses

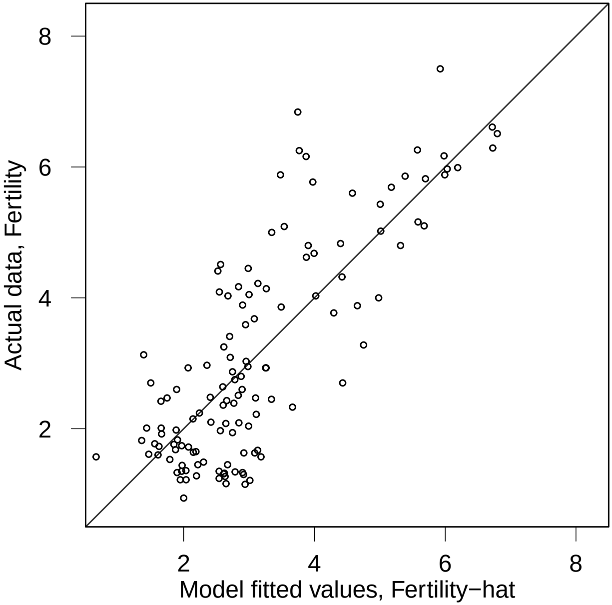
How do we interpret the above table?

Three regression models of fertility

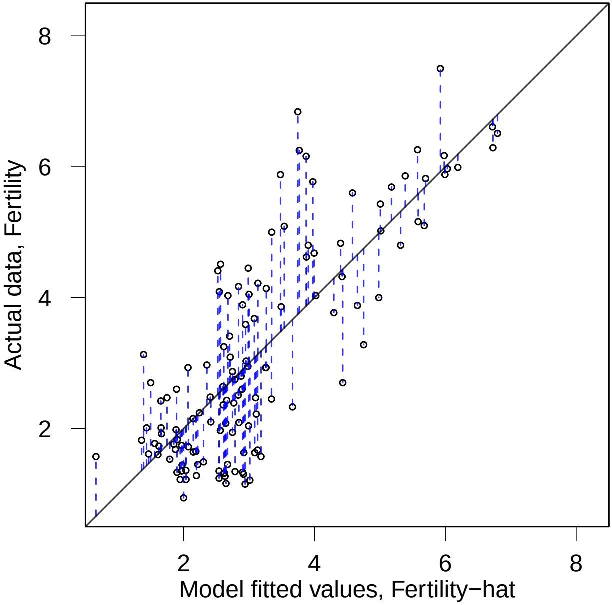
Variable	Model 1	2	3
Intercept	12.59	4.13	11.25
	[11.11, 14.08]	[3.80, 4.46]	[9.81, 12.69]
Education Ratio	-0.10		-0.08
	[-0.12, -0.08]		[-0.10, -0.06]
GDP per capita		-0.10	-0.05
		[-0.12, -0.08]	[-0.07, -0.03]
N	130	130	130
R^2	0.55	0.35	0.64
RMSE	1.12	1.35	1.01

95% confidence intervals in brackets

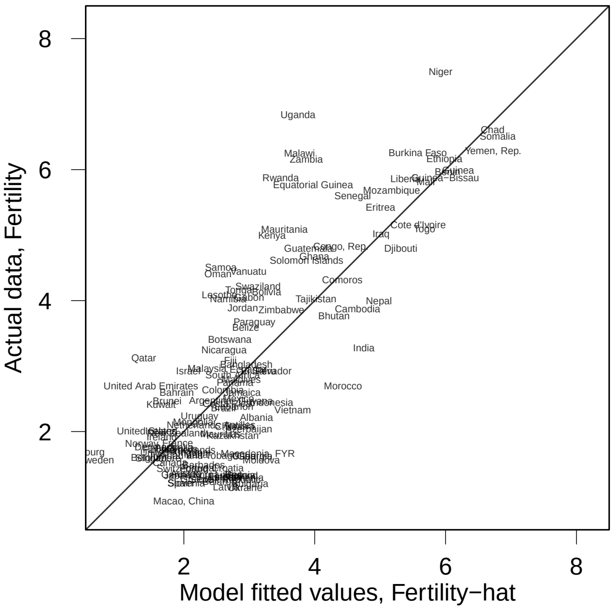
This table presents the same information, but is easier to digest



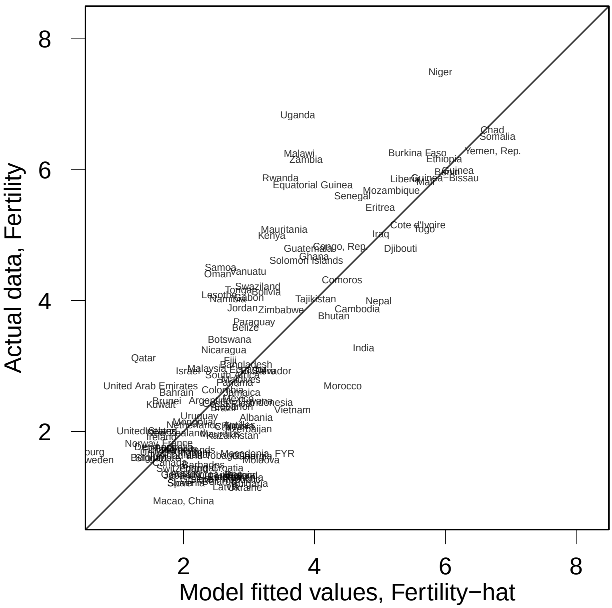
To see the residuals, compare the model fit with reality



Note that in the multivariate case, we need to plot against $\hat{y}_i$, not x_i, because there is more than one x_i

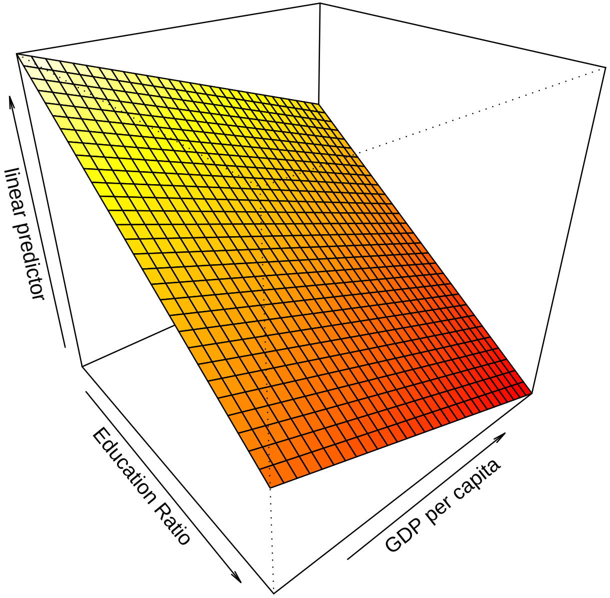


Examining which cases are big outliers may suggest additional variables to include as covariates

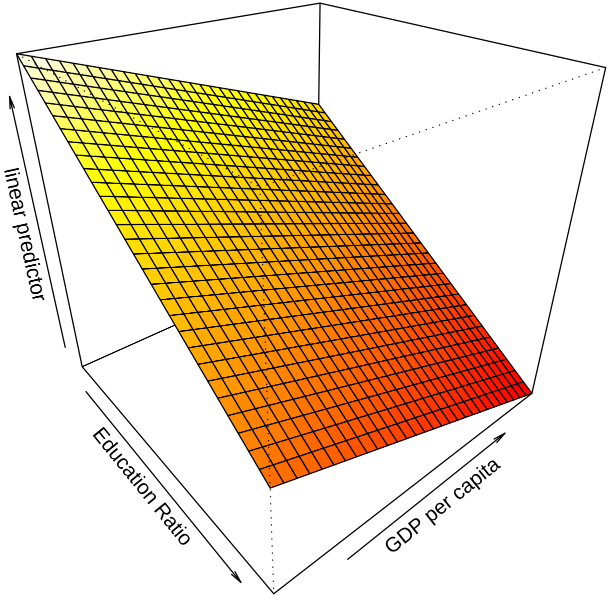


Examining which cases are big outliers may suggest additional variables to include as covariates

Think of what the missing cases have in common

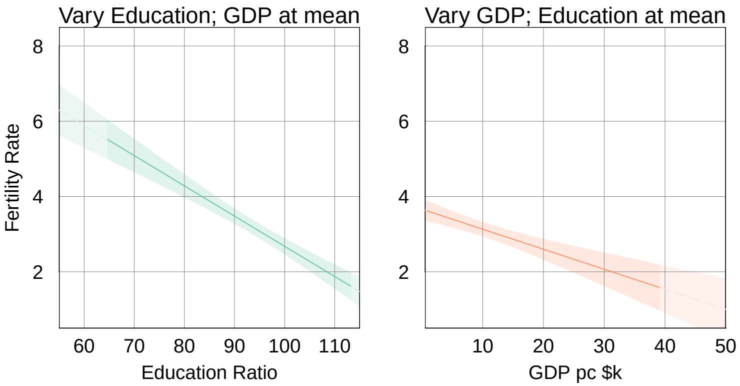


Visualizing the modelled relationship between many variables is tricky



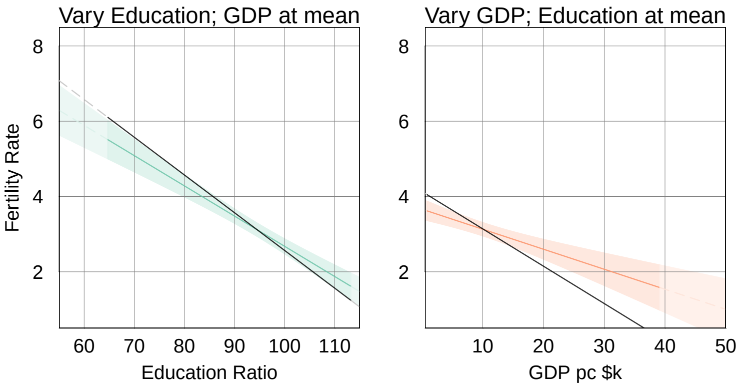
Visualizing the modelled relationship between many variables is tricky

We can do it with a 3D plot for 2 covariates, but not for 3 or more



An alternative that works for any number of covariates:
Plot out the model predictions as a function of each covariate, hold the other covariates fixed, e.g., at their means

Then predict what Fertility rate should happen on average if the country had average GDP but variable Education (or vice versa)



Let's compare the multiple regression estimates with the bivariate regression results (in black)

How are they different? Are the bivariate results affected by omitted variable bias?

Regression models including Agricultural Labor

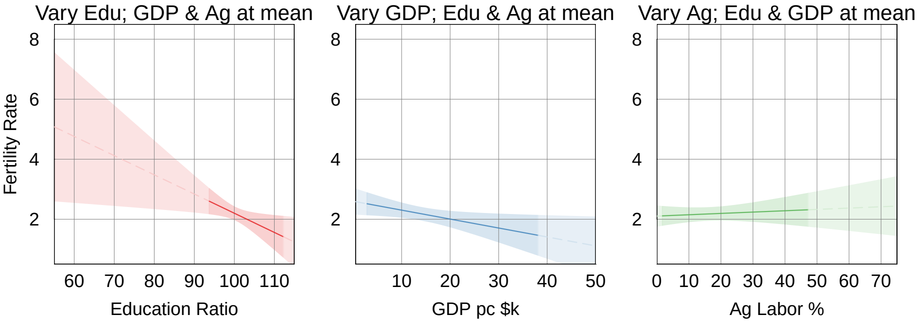
	Model			
Variable	1	2	3	4
Intercept	11.15	2.76	1.83	8.95
	(2.64)	(0.18)	(0.15)	(2.79)
Education Ratio	-0.09			-0.06
	(0.03)			(0.03)
GDP per capita ($k)		-0.04		-0.03
		(0.01)		(0.01)
Agriculture Labor			0.02	0.004
			(0.01)	(0.008)
N	72	72	72	72
R^2	0.13	0.17	0.14	0.26
RMSE	0.94	0.92	0.93	0.88

Standard errors in parentheses

Regression models including Agricultural Labor

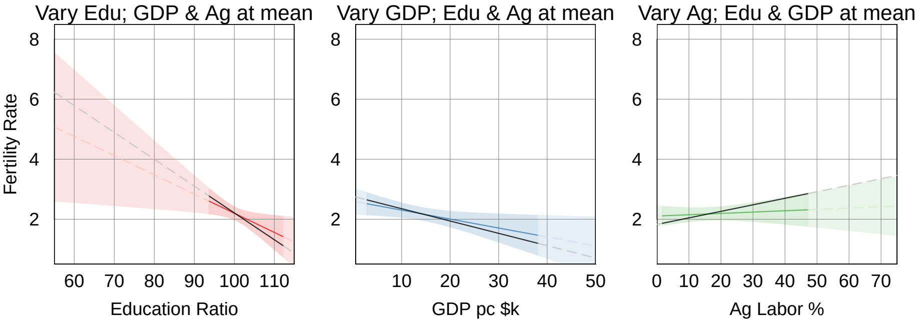
	Model			
Variable	1	2	3	4
Intercept	11.15	2.76	1.83	8.95
	[5.90, 16.41]	[2.39, 3.13]	[1.53, 2.12]	[3.38, 14.52]
Edu Ratio	-0.09			-0.06
	[-0.14, -0.04]			[-0.12, -0.01]
GDP pc		-0.04		-0.03
		[-0.06, -0.02]		[-0.06, -0.003]
Ag Labor			0.02	0 0.004
			[0.01, 0.03]	[-0.01, 0.02]
N	72	72	72	72
R^2	0.13	0.17	0.14	0.26
RMSE	0.94	0.92	0.93	0.88

95% confidence intervals in brackets

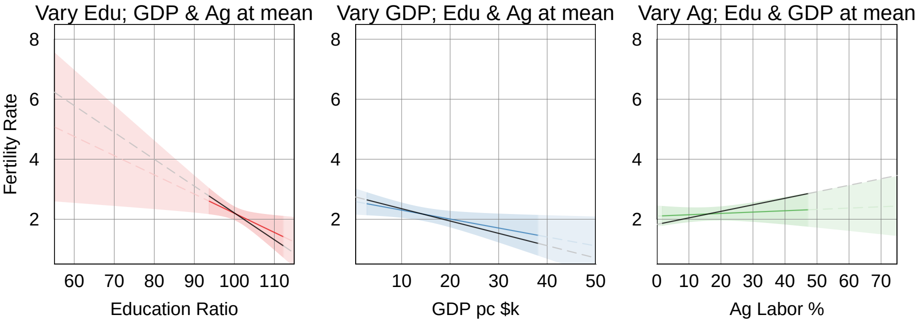


How do we interpret these plots?

The dashed lines indicate extrapolation: no observed data have these values for the covariates



The black lines show the bivariate results. Was there omitted variable bias?



The black lines show the bivariate results. Was there omitted variable bias?

YES. The apparent effect of Ag Labor was a mirage: just the omitted effect of GDP per capita. If we control for GDP, we see Ag Labor has no effect.

Warning!

Linear regression is powerful, but easy to misuse

We mentioned one assumption last time: That the error term is Normally distributed

To this we now add two additonal assumptions

Correct specification The model contains all the covariates that produce Y. If any omitted cause of Y is correlated with the included X's, then $\hat{\beta}$ can no longer be trusted.

No endogeneity of Y None of the included X's are caused by Y

Final thoughts on linear regression

Linear regression is a powerful tool for isolating conditional expectations of y given x after removing confounding variables

But vulnerable to many hazards:

- Outliers
- Reverse causation
- Selection bias
- Omitted variable bias

Advanced techniques can mitigate these problems, as well as deal with others

Topic of a sequence of required courses in most graduate social science programs