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

Introduction to Multiple 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

<table>
<thead>
<tr>
<th>Variable</th>
<th>Estimates</th>
<th>se</th>
<th>t-stat</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>1.83</td>
<td>(0.15)</td>
<td>12.34</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Agricultural Labor</td>
<td>0.02</td>
<td>(0.01)</td>
<td>3.52</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>$N$</td>
<td>72</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$R^2$</td>
<td>0.15</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RMSE</td>
<td>0.93</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

How do we read this table?
### Regression of Fertility on Agricultural Labor

<table>
<thead>
<tr>
<th>Variable</th>
<th>Estimates</th>
<th>se</th>
<th>t-stat</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>1.83</td>
<td>0.15</td>
<td>12.34</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Agricultural Labor</td>
<td>0.02</td>
<td>0.01</td>
<td>3.52</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>$N$</td>
<td>72</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$R^2$</td>
<td>0.15</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RMSE</td>
<td>0.93</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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)
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?

The RMSE, or standard error of the regression: how much predictions from this model tend to miss by.
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?
Which (if any) of the three models we’ve looked at are right?

Do Education, GDP, and Ag Labor all affect Fertility?
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?
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 \( k \)th covariate

This leads to the above model, with multiple *partial* slopes \( \beta_1, \beta_2, \ldots \beta_k \)
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 \( \beta \)'s? Just as before.
The linear regression model, redux

\[ y_i = \beta_0 + \beta_1x_{1i} + \beta_2x_{2i} + \ldots + \beta_kx_{ki} + \varepsilon_i \]

How do we interpret the \( \beta \)'s? Just as before.

The \( \beta \)'s are still slopes, or the amount \( y_i \) changes on average for a 1 unit increase in \( x \), \textit{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 \( \beta \)'s? Just as before.

The \( \beta \)'s are still slopes, or the amount \( y_i \) changes on average for a 1 unit increase in \( x \), \textit{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 \( \beta_1 \).
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 $\beta$’s? Just as before.

The $\beta$’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 $\beta_1$

→ Finally found a way to control for confounders using observational data!

Aside for calculus-users: The $\beta$’s are partial derivatives with respect to the $x$ they multiply
Our estimates, $\hat{\beta}_k$, are the $\beta_k$’s that minimize the sum of the squared residuals (least squares).
Multiple regression: just like bivariate

1. Our estimates, $\hat{\beta}_k$, are the $\beta_k$'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 \( \beta_k \)'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 $\beta_k$’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 \( \beta_k \)'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:

\[
\hat{Fertility}_i = \hat{\beta}_0 + \hat{\beta}_1 EduRatio_i + \hat{\beta}_2 GDPpc_i
\]

\[
\hat{Fertility}_i = 11.24 - 0.08 \times EduRatio_i - 0.05 \times GDPpc_i
\]

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

<table>
<thead>
<tr>
<th>Variable</th>
<th>Estimates</th>
<th>se</th>
<th>t-stat</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>11.25</td>
<td>(0.73)</td>
<td>15.46</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Education Ratio</td>
<td>−0.08</td>
<td>(0.01)</td>
<td>−9.93</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>GDP per capita ($k)</td>
<td>−0.05</td>
<td>(0.01)</td>
<td>−5.32</td>
<td>&lt; 0.001</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>N</th>
<th>130</th>
</tr>
</thead>
<tbody>
<tr>
<td>$R^2$</td>
<td>0.64</td>
</tr>
<tr>
<td>RMSE</td>
<td>1.01</td>
</tr>
</tbody>
</table>

How do we interpret the above?
### Three regression models of fertility

<table>
<thead>
<tr>
<th>Variable</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>12.59</td>
<td>4.13</td>
<td>11.25</td>
</tr>
<tr>
<td></td>
<td>(0.75)</td>
<td>(0.17)</td>
<td>(0.73)</td>
</tr>
<tr>
<td>Education Ratio</td>
<td>−0.10</td>
<td>−0.08</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.01)</td>
<td>(0.01)</td>
<td></td>
</tr>
<tr>
<td>GDP per capita</td>
<td></td>
<td>−0.10</td>
<td>−0.05</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.01)</td>
<td>(0.01)</td>
</tr>
<tr>
<td><strong>N</strong></td>
<td>130</td>
<td>130</td>
<td>130</td>
</tr>
<tr>
<td><strong>$R^2$</strong></td>
<td>0.55</td>
<td>0.35</td>
<td>0.64</td>
</tr>
<tr>
<td><strong>RMSE</strong></td>
<td>1.12</td>
<td>1.35</td>
<td>1.01</td>
</tr>
</tbody>
</table>

Standard errors in parentheses

**How do we interpret the above table?**
### Three regression models of fertility

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

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

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

Standard errors in parentheses
## Regression models including Agricultural Labor

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

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 additional 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.