Multiple Regression
Multiple Regression

• I have a hypothesis about the effect of X on Y.
• Why might we need additional variables?
  – Confounding variables
  – Conditional independence
  – Reduce/eliminate bias in estimates
  – Improve hypothesis testing
Multiple Regression Analysis

- We can make a simple extension of the two variable, or bivariate, model to the multivariate case
- Instead of a two dimensional space we move into a multi-dimensional space
- If we have two X variables, then we are fitting a two dimensional plane through points in a space
Basic Equation

• The basic equation for multiple regression is:

\[ Y_i = B_1 + B_2 X_{2i} + B_3 X_{3i} + \ldots + B_k X_{ki} + u_i \]

– Where \( Y_i \) is the outcome we seek to explain

– \( B_1 \) is the intercept

– Each \( X \) is a different variable, and the \( B \)'s that correspond with the \( X \)'s are the impact of the variable on \( Y \) (holding other factors constant)

– \( u \) is the error term
The Multiple Regression Equation

- The vectors and matrices in $y = X\hat{\beta} + \hat{u}$ are represented by

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{ik} \\ x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix}, \quad \hat{\beta} = \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_k \end{bmatrix}, \quad \hat{u} = \begin{bmatrix} \hat{u}_1 \\ \hat{u}_2 \\ \vdots \\ \hat{u}_n \end{bmatrix}$$
Assumptions of Multiple Regression

• They are similar to the Two Variable Case
  – Zero mean of \( u_i \): \( E(u_i | X_{2i}, X_{3i}) = 0 \)
    • Note the slight change in this assumption
  – No serial correlation: \( \text{cov}(u_i, u_j) = 0 \)
  – Homoskedasticity: \( \text{var}(u_i) = \sigma^2 \)
  – Zero covariance between \( u_i \) and each \( X \)
    • \( \text{Cov}(u_i, X_{2i}) = \text{Cov}(u_i, X_{3i}) = 0 \)
  – No specification bias
  – Linear in parameters (\( B_1, B_2, B_3, \ldots B_k \))
We add one more assumption

• No exact collinearity
• Informal Definition
  – None of the independent variables (X variables) can be written as exact linear combinations of the remaining independent variables in the model
Graphical Illustration

- Illustration
Graphical Illustration

- Illustration

\[ \hat{\beta}_1 \]

\[ \hat{\beta}_2 \]

Covariance of X1 and Y1 (orange area)

Discard (remains in the error term)
Hypothesis Tests

• Hypothesis tests of the individual coefficients are almost identical to the two variable case, except we state the null as:
  – With $X_3$ held constant, $X_2$ has no influence on $Y$

• Again, we need the normality assumption of $u_i \sim N(0, \sigma^2)$

• Thus, our estimates of $B_1, B_2 \ldots B_k$ are distributed normally with the true mean of $B$ and variances from previous section
T statistic

- The t-statistic is now the same, except for the degrees of freedom
- If calculated t is greater than critical value, we can reject the null.

\[ t_{n-k} = \frac{\hat{B}_k - B_k}{se(\hat{B}_k)} \]
Confidence Interval

- Again we use the t value for n-k degrees of freedom
- If zero is not in our CI, then we can reject the null hypothesis.

\[ B_k = \hat{B}_k \pm t_{.025}SE \]
P-Value

• Once we look up the critical t value, we will get the probability under the curve, or the p value

• This can be interpreted as follows: If the null were true, the probability of obtaining a t value as large as we do or greater is (insert the p value).
Multicollinearity

- Wooldridge, p. 866: “A term that refers to correlation among the independent variables in a multiple regression model; it is usually invoked when some correlations are ‘large,’ but an actual magnitude is not well-defined.”
- KKV, p. 122. “Any situation where we can perfectly predict one explanatory variable from one or more of the remaining explanatory variables.”
- UCLA On-line Regression Course: “The primary concern is that as the degree of multicollinearity increases, the regression model estimates of the coefficients become unstable and the standard errors for the coefficients can get wildly inflated.”
Causes of Multicollinearity

• x’s are causally related to one another and you put them both in your model. Proximate versus ultimate causality.
  (i.e. legacy v. political institutions and economic reform)

• Poorly constructed sampling design causes correlation among x’s.
  (i.e. You survey 50 districts in Indonesia, where the richest districts are all ethnically homogenous, so you cannot distinguish between ethnic tension and wealth on propensity to violence.).

• Poorly constructed measures over aggregate information can make cases correlate.
  (i.e. Freedom House – Political Liberties, Global Competitiveness Index)
Causes of Multicollinearity

• **Statistical model specification: adding polynomial terms or trend indicators.**
  (i.e. Time since the last independent election correlates with Eastern Europe or Former Soviet Union).

• **Too many variables in the model – x’s measure the same conceptual variable.**
  (i.e. Two causal variables essentially pick-up up on the same underlying variable).
Multi-collinearity Warning Signs

• F-test of joint-significance of several variables is significant but coefficients are not.

• Coefficients are substantively large but statistically insignificant.

• Standard errors of $\beta$’s change when other variables included or removed, but estimated value of $\beta$ does not.
Multicollinearity: The Diagnostic

• Diagnostic of multicollinearity is the auxiliary R-squared

• Regress each $x$ on all other $x$’s in the model

• R-squared will show you linear correlation between each $x$ and all other $x$’s in the model
Multicollinearity: The Diagnostic

• There is no definitive threshold when the auxiliary R-squared is too high.
  – Depends on whether $\beta$ is significant

• Tolerance for multicollinearity depends on $n$
  – Larger $n$ means more information

• If auxiliary R-squared is close to 1 and $\hat{\beta}$ is insignificant, you should be concerned
  – If $n$ is small then .7 or .8 may be too high
Multicollinearity: Remedies

- Increase sample size to get more information
- Change sampling mechanism to allow greater variation in x’s
- Change unit of analysis to allow more cases and more variation in x’s (districts instead of states)
- Look at bivariate correlations prior to modeling
Multicollinearity: Remedies

• Disaggregate measures to capture independent variation

• Create a composite scale or index if variables measure the same concept

• Construct measures to avoid correlations
Omitted Variable Bias
What is Model Specification?

• Model Specification is two sets of choices:
  – The set of variables that we include in a model
  – The functional form of the relationships we specify

• These are central *theoretical* choices

• Can’t get the right answer if we ask the wrong question
What is the “Right” Model?

• In truth, all our models are misspecified to some extent.

• Our theories are always a simplification of reality, and all our measures are imperfect.

• Our task is to seek models that are reasonably well specified – keeps our errors relatively modest
Model Specification: Causes

- There are four basic types of model misspecification:
  - Inclusion of an irrelevant variable
  - Exclusion of a relevant variable
  - Measurement error
  - Erroneous functional form for the relationship
Omitted Variable Bias: Excluding Relevant Variables

- Imagine that the true causes of $y$ can be represented as:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$$

- But we estimate the model:

$$y = \beta_0 + \beta_1 x_1 + u^*$$

**Where:**

$$u^* = \beta_2 x_2 + u$$
Omitted Variable Bias: Consequences

• Clearly this model violates our assumption that $E(u)=0$
  - $E({\beta_2}x_2+u)$ is not 0

• We required this assumption in order to show that $\hat{\beta}$ is an unbiased estimator of $\beta$

• Thus by excluding $x_2$, we risk biasing our coefficients for $x_1$

• If $\hat{\beta}$ is biased, then $\sigma_{\hat{\beta}}$ is also biased
Omitted Variable Bias: Consequences

\[ y = \beta_0 + \beta_1 x_1 + u^* \]

*Where:*

\[ u^* = \beta_2 x_2 + u \]

- This equation indicates that our estimate of \( \beta_1 \) will be biased by two factors
- The extent of the correlation between \( x_1 \) and \( x_2 \)
- The extent of \( x_2 \)'s impact on \( y \)
- Often difficult to know direction of bias
Omitted Variable Bias: Consequences

• Remember, if either of these sources of bias is 0, then the overall bias is 0

• Thus $E(\hat{\beta}_1) = \beta_1$ if:
  
  – Correlation of $x_1$ & $x_2=0$; note this is a characteristic of the sample
  OR
  
  – $\beta_2=0$; note this is not a statement about the sample, but a theoretical assertion
Summary of Bias in $\beta_1^{\text{hat}}$, the Estimator when $x_2$ is omitted

<table>
<thead>
<tr>
<th>Relationship</th>
<th>Corr ($x_1, x_2$)$&gt;0$</th>
<th>Corr ($x_1, x_2$)$&lt;0$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta_2&gt;0$</td>
<td>Positive Bias: $\beta_1^{\text{hat}}$ will appear to have a strong positive relationship with $y$. (Also called upward bias or biased to the right)</td>
<td>Negative Bias: $\beta_1^{\text{hat}}$ will appear to have a strong negative relationship with $y$. (Also called downward bias or biased to the left)</td>
</tr>
<tr>
<td>$\beta_2&lt;0$</td>
<td>Negative Bias</td>
<td>Positive Bias</td>
</tr>
</tbody>
</table>
### Examples of Omitted Variable Bias

<table>
<thead>
<tr>
<th>Relationship</th>
<th>Corr ((x_1, x_2) &gt; 0)</th>
<th>Corr ((x_1, x_2) &lt; 0)</th>
</tr>
</thead>
</table>
| \(\beta_2 > 0\) | **DV**: Salary  
**\(X_1\)**: Years of Education  
**\(X_2\)**: Skill level  
*Sin of omission*: Years of education will appear to have a substantively greater effect on salary. | **DV**: Health  
**\(X_1\)**: Cigarettes smoked per day.  
**\(X_2\)**: Exercise per day  
*Sin of omission*: Effect of cigarettes will appear smaller. |
| \(\beta_2 < 0\) | **DV**: Civil War  
**\(X_1\)**: Ethnic homogeneity  
**\(X_2\)**: Economic equality  
*Sin of omission*: Ethnic homogeneity appears to have a greater effect on reducing civil war. | **DV**: Expropriation  
**\(X_1\)**: Secure land title  
**\(X_2\)**: Size of business  
*Sin of omission*: Leaving out size of business makes me think that land titles have a larger effect on reducing expropriation. |
Omitting Variables and Model Specification

• These criteria give us our conceptual standards for determining when a variable must be included

• A variable must be included in a regression equation IF:
  – The variable is correlated with other x’s AND
  – The variable is also a cause of y
Illustrating Omitted Variable Bias

• Imagine a true model where $x_1$ has a small effect on $y$ and is correlated with $x_2$ that has a large effect on $y$.

• Specifying both variables can distinguish these effects
Illustrating Omitted Variable Bias

• But when we run simple model excluding $x_2$, we attribute all causal influence to $x_1$

• *Coefficient is too big and variance of coefficient is too small*
Omitted Variable Bias: Causes

• This problem is primarily theoretical rather than “statistical”

• Though OVTEST or LINKTEST in STATA are designed to offer some indication, ultimately they are not satisfying. They may reveal that residuals are not correlated with the independent variables, but this may still not be theoretically satisfying.
The Classis Case of Omitted Variable Bias

% of Patients taking medicine

# of Patient Deaths

% of Patients Ill
Spuriousness and Omitted Variable Bias: An IR Example
Including Irrelevant Variables

• Do we see the same kinds of problems if we include irrelevant variables?
  • NO!
• Imagine we estimate the equation:

\[ \hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{u} \]
Including Irrelevant Variables

• In this case:

\[ E(\hat{\beta}) = \beta = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} \]

• If \( \hat{\beta}_2 = 0 \), our estimate of \( \hat{\beta}_1 \) is not affected
• If \( x_1'x_2 = 0 \), our estimate of \( \hat{\beta}_1 \) is not affected
• But including \( x_2 \) if it is not relevant does unnecessarily inflate the \( \sigma^2 \hat{\beta}_1 \)
Including Irrelevant Variables: Consequences

- $\sigma^2 \beta_{\text{hat}1}$ increases for two reasons:
  - Addition of parameter for $x_2$ reduces the degrees of freedom
    - Part of estimator for $\sigma_{\text{uhat}}^2$
  - If $\beta_{\text{hat}2} = 0$ but $x_1'x_2$ is not, then including $x_2$ unnecessarily reduces independent variation of $x_1$. (*Multicollinearity*)
- Thus parsimony remains a virtue
Rules for Model Specification

• Model specification is fundamentally a theoretical exercise. We build models to reflect our theories
• As much as we would like, the theorizing process cannot be replaced with statistical tests
• \textit{Avoid mechanistic rules for specification such as stepwise regression}
The Evils of Stepwise Regression

• Stepwise regression is a method of model specification that chooses variables on:
  – Significance of their t-scores
  – Their contribution to $R^2$

• Variables will be selected in or out depending on the order they are introduced into the model
The Evils of Stepwise

• Stepwise regression can fall victim to specification errors as a result of multicollinearity or spuriousness problems

• $R^2$ is not a generalizable property of the model

• Thus stepwise regression is a theoretical curve-fitting

• Our forecasting ability becomes limited
Choosing Your Variables

• Variable selection should be based on your theoretical understanding of the causal process underlying the dependent variable

• This is a theoretical rather than a statistical exercise

• There are rules of thumb for including or excluding variables from an analysis, but ultimately Mr. Theory needs to be your guide.
Reasons to Include a Variable

• $x_2$ is a potential Source of Omitted Variable Bias
  – Variable is a cause of $y$ and correlated with $x_1$

• Methodological Control variables
  – Selection parameters, temporal dependence, etc

• You must include these $x$’s as controls

• $x_2$ is a benchmark for evaluating the substantive impact of $x_1$
  – Selected at your discretion
Reasons to Exclude a Variable

• $x_2$ is NOT a cause of $y$, but it creates multicollinearity problems
• $x_2$ is an intervening variable in between $x_1$ and $y$ (a proximate cause of $y$, while $x_1$ remains the ultimate cause).
• With intervening variables, a good strategy is to run the model both with and without $x_2$
  – Examine change in coefficient for $x_1$