PDE PROVIDER
Professional Development for the Expert

Course Title

Choosing Variables for Regression Analysis

Instructor

Alan Fata, DBA

Credit	Questions
1 PDU	10

Professional Development for the Expert, Inc.
(571) 619-5550, info@pdeprovider.com

Choosing Variables for Regression Analysis in a Research Paper, an Introduction

Continuing Education Course

A practical guide for researchers in business, social sciences, education, and applied research

I. Course Overview

Regression analysis is often easy to run in statistical software but much harder to design correctly. The most important decisions are made before opening SPSS, Stata, R, or another program: What is the outcome? Which variables are theoretically expected to explain it? Which variables should be controls? Which variables should not be included? This course provides a research-oriented framework for making those decisions and defending them in a paper.

By the end of the course, the researcher should be able to:

- Distinguish dependent, independent, control, mediator, and moderator variables.
- Translate a research question and conceptual framework into a regression model.
- Choose predictors using theory and prior evidence rather than statistical convenience.
- Recognize common problems such as multicollinearity, over-control, omitted variables, and post-treatment bias.
- Decide how variables should be measured, coded, transformed, and entered into a model.
- Interpret model specification, coefficients, R^2, adjusted R^2, significance, and practical importance.
- Write a defensible regression specification and justify variable selection in a research paper.

II. Start With the Research Question, Not the Software

A regression model should answer a substantive research question. It should not be created by placing every available variable into a statistical package and keeping the variables that produce significant p-values. A good starting point is a sentence such as: "Does perceived remote-work effectiveness predict employee performance after accounting for age, tenure, and job level?" That sentence immediately identifies the outcome, focal predictor, and plausible controls.

Table 1: The basic roles of variables

Role	Purpose	Example
Dependent variable (Y)	The outcome the study seeks to explain or predict.	Employee performance
Independent variable (X)	The main explanatory variable(s) tied to the hypothesis.	Remote-work flexibility
Control variable	A variable included to account for alternative explanations or improve comparability.	Age, tenure
Mediator	A mechanism through which X may influence Y.	Job satisfaction
Moderator	A variable that changes the strength/direction of the X–Y relationship.	Managerial support

The distinction matters because these variables should not automatically be treated in the same way. For example, a mediator may be inappropriate as a control if the research question concerns the total effect of X on Y. Similarly, a moderator is not simply another control: it requires an interaction term and a corresponding hypothesis.

A useful first question: Before selecting predictors, write the causal or explanatory story in plain language: "X is expected to influence Y because…". Then identify the variables needed to test that story. This prevents the statistical model from becoming detached from the research design.

III. Build the Conceptual Framework Before the Regression

Variable selection should normally follow the conceptual framework, literature review, and hypotheses. Suppose a study examines whether digital banking capability affects bank performance. The researcher might argue that digital capability improves service efficiency and customer reach, which can contribute to performance. Bank size, capital adequacy, and age may be included as controls because they can also be related to performance.

A simple model can be represented as:

$$\text{Performance} = \beta_0 + \beta_1 \text{ Digital Capability} + \beta_2 \text{ Size} + \beta_3 \text{ Capital Adequacy} + \beta_4 \text{ Bank Age} + \varepsilon$$

Each coefficient corresponds to a research proposition. β_1 represents the expected association between digital capability and performance, conditional on the other variables in the model. The error term represents influences on performance not captured by the included predictors.

IV. The literature review is part of variable selection

Prior empirical studies help establish which variables have been used, how they have been measured, and why they may matter. However, copying the variables from one previous paper is not sufficient. Different samples, contexts, theories, and outcomes can require different specifications. The goal is a theoretically justified model that is informed by prior evidence.

Choosing Independent Variables: The strongest predictor candidates are variables for which there is a clear theoretical reason to expect a relationship with the dependent variable. A researcher should be able to answer three questions for every focal independent variable: Why should it affect Y? What does prior research suggest? How will it be measured?

Use hypotheses to guide the focal predictors: If the study has H1: “Training intensity is positively associated with employee performance,” then training intensity belongs in the model as a focal predictor. If H2 states that job autonomy is positively associated with performance, job autonomy is another focal predictor. The regression should therefore map visibly onto the hypotheses.

Avoid the “everything significant” strategy: A common but weak approach is to test dozens of variables and retain only those with $p < .05$. This can capitalize on chance, produce unstable models, and make the final model difficult to defend theoretically. Variable selection should be determined primarily by the research question and design, with statistical diagnostics used to evaluate whether the proposed model is credible.

Prudence matters: A model should contain enough variables to represent the research question and address important confounding, but not so many that the model becomes unnecessarily complex. Every added predictor consumes information and, especially in smaller samples, can make coefficient estimates less precise.

V. Selecting Control Variables Correctly

Controls are often the most misunderstood part of regression. A control should have a substantive reason for inclusion. Typical examples include demographic characteristics, organizational characteristics, baseline measures, or other variables known to be associated with both the predictor and outcome.

Think about confounding: Suppose X is training and Y is performance. Employees with longer tenure may receive more training and may also perform differently because of accumulated experience. Tenure could therefore be a reasonable control if supported by the research design and literature.

Do not control for everything that correlates with Y: A variable can correlate strongly with the dependent variable and still be inappropriate as a control. The key question is not merely "Does this variable predict Y?" but "Does including this variable help isolate the relationship I am trying to estimate without blocking part of the relationship or introducing bias?" In particular, researchers should be cautious about controlling for mediators or variables that occur after the focal exposure. If X affects M and M affects Y, controlling for M can remove part of the pathway the researcher intends to study. The correct choice depends on whether the research question concerns a total effect, direct effect, or another estimate.

VI. Measurement, Coding, and Transformation

Choosing the right construct is only the first step. The researcher must also choose a defensible operational measure. Regression results are only as meaningful as the variables entered into the model.

Continuous variables

Age, income, years of experience, sales, and many scale scores can be represented as continuous variables. Keeping meaningful variation is usually preferable to unnecessarily dividing a continuous variable into categories, because categorization can discard information and reduce statistical power.

Categorical variables

Variables such as gender, industry, region, education category, or organizational type may require dummy or indicator coding. For a variable with k categories, a model typically uses $k-1$ indicators and treats one category as the reference group.

Likert-scale constructs

Many social-science studies measure constructs with several Likert items and create composite scores after assessing reliability and construct validity. The researcher should explain how the score was created and why the resulting measure represents the construct.

Transformations

Highly skewed variables such as income, firm size, or transaction volume may sometimes be transformed, for example using a logarithm. The transformation should have a statistical or substantive justification and should be reported clearly.

Missing data

Missing observations can change the effective sample size and potentially introduce bias. Before interpreting a regression, report how missing data were handled and whether the final analytic sample differs meaningfully from the original sample.

VII. Check Multicollinearity Before Interpreting Coefficients

Multicollinearity occurs when predictors contain substantial overlapping information. For example, a study might include annual revenue, number of employees, total assets, and market size. If these variables are strongly related, individual coefficients can become unstable and standard errors can increase.

Researchers commonly inspect correlations and variance inflation factors (VIF), but there is no single universal cutoff that should be treated as an automatic pass/fail rule. Context, theoretical redundancy, sample size, and the purpose of the model all matter.

If two variables measure nearly the same construct, the researcher should consider whether both are conceptually necessary. Possible solutions include selecting the theoretically superior measure, creating a theoretically justified index, or specifying separate models when the research question permits it.

VIII. A Step-by-Step Variable Selection Procedure

1. **Step 1 - Define Y:** State exactly what outcome the paper seeks to explain. Identify its unit, measurement, and time frame.
2. **Step 2 - Identify X:** List the main predictors derived from the research questions and hypotheses.
3. **Step 3 - Draw the conceptual model:** Show expected relationships among predictors, outcome, mediators, moderators, and controls.
4. **Step 4 - Review the literature:** Document theoretical and empirical reasons for each important predictor and control.
5. **Step 5 - Define operational measures:** Specify how every construct becomes an observable variable.
6. **Step 6 - Examine the data:** Inspect distributions, outliers, coding errors, and descriptive statistics.
7. **Step 7 - Check relationships among predictors:** Assess correlations and multicollinearity and reconsider redundant variables.
8. **Step 8 - Estimate the planned model:** Run the regression that corresponds to the hypotheses and research design.
9. **Step 9 - Conduct diagnostics and robustness checks:** Examine residuals, influential observations, specification concerns, and alternative reasonable specifications.
10. **Step 10 - Report the rationale:** Explain why variables were included or excluded; do not report only the final p-values.

IX. Hierarchical or blockwise regression

When theory suggests a sequence, researchers can enter variables in blocks. For example, Model 1 may contain demographic controls, Model 2 may add organizational controls, and Model 3 may add the focal predictors. This allows the researcher to examine whether the focal variables explain additional variance beyond the controls. The order should be theoretically justified rather than chosen solely because it produces a larger R^2.

X. Interactions and moderation

If theory predicts that the effect of X depends on Z, include an interaction term such as X × Z and interpret the conditional effect. A significant main effect of X alone does not establish moderation. Continuous predictors are often centered or standardized when useful for interpretation, while categorical variables require appropriate coding.

XI. What Researchers Should Avoid

- **Stepwise selection as the main theory:** Automated procedures can be useful exploratory tools, but they should not replace a theoretically grounded research model.

- **Adding variables because they are available:** Availability in a questionnaire or database is not a sufficient reason for inclusion.

- **Removing a theoretically important variable only because $p > .05$:** A non-significant coefficient does not prove that the variable is irrelevant to the research model.

- **Ignoring omitted-variable concerns:** A simple significant relationship may change after an important confounder is considered.

- **Over-controlling:** Controls can block causal pathways or create bias when they are consequences of the exposure.

- **Using highly correlated predictors without reflection:** The model may become difficult to interpret, and individual coefficients may be unstable.

- **Changing the model repeatedly until significance appears:** This increases the risk of data-driven findings that do not replicate.

- **Reporting only R^2 and p-values:** Researchers should also discuss coefficient direction, magnitude, uncertainty, model specification, diagnostics, and substantive meaning.

XII. Worked Example

Imagine a researcher wants to study whether remote-work characteristics influence employee performance. The conceptual framework identifies remote-work autonomy and communication quality as focal predictors. The literature also suggests that employee tenure, job level, and organizational size may influence performance.

A defensible specification might therefore be:
Employee Performance = $\beta_0 + \beta_1$ Autonomy + β_2 Communication Quality + β_3 Tenure + β_4 Job Level + β_5 Organization Size + ε

The researcher should explain why autonomy and communication quality are the main independent variables, why the other variables are controls, how each construct was measured, and why the chosen sample provides enough information for the proposed model.

If job satisfaction is hypothesized to explain how remote-work autonomy affects performance, the researcher should not automatically treat job satisfaction as an ordinary control. Instead, a mediation analysis may be more appropriate if that is the research question.

XIII. How to Write the Variable-Selection Section of the Paper

A strong methodology section does not merely list variables. It explains the logic connecting the research question, theory, measurement, and statistical model. A concise structure is:

- Define the dependent variable and explain why it represents the outcome of interest.
- Introduce each focal independent variable and connect it to the relevant hypothesis.
- Identify control variables and provide a theoretical or empirical reason for each.
- Explain measurement, coding, composite-score construction, and transformations.
- Present the regression equation and define each term.
- Describe diagnostic checks and any robustness analyses.
- State that variable selection was theory- and research-question-driven rather than based solely on statistical significance.

Model Specification Template

For a standard multiple linear regression, researchers can present a model in the following form:
$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + \varepsilon_i$

where Y is the dependent variable, X_1 through X_k are the predictors, β_0 is the intercept, β coefficients represent conditional associations, and ε is the error term. The equation should be adapted to the actual research design and outcome type.

Final Takeaway

The best regression model is not the model containing the largest number of significant variables. It is the model that most clearly answers the research question, reflects the conceptual framework, uses valid measures, accounts for important alternative explanations, and remains statistically and substantively defensible. Variable selection is therefore a research-design decision first and a software decision second.

Practical Researcher Checklist

- Can I state my dependent variable in one clear sentence?
- Can I explain why each focal predictor should be related to the outcome?
- Can I justify every control variable?
- Have I distinguished controls from mediators and moderators?
- Are my variables measured in a defensible way?

- Have I considered categorical coding and transformations where appropriate?
- Have I checked missing data and influential observations?
- Have I assessed multicollinearity?
- Is my sample adequate for the number of predictors?
- Does my final model correspond to my hypotheses and conceptual framework?
- Could I defend the inclusion or exclusion of every variable to a thesis committee or journal reviewer?

End of Course