Which regression problem needs which response?

> Variance problem: covariance correction or a justified weighting model.

> Measurement and selection problems: revisit the information and assumptions supporting the comparison.

Unequal error spread is heteroskedasticity

It concerns the error variance at different regressor values.

> $E(u_i \mid X_i) = 0$ concerns the mean

> $\mathrm{Var}(u_i \mid X_i)$ concerns spread

Robust covariance replaces a constant-variance calculation

The same OLS weights can propagate different observation variances.

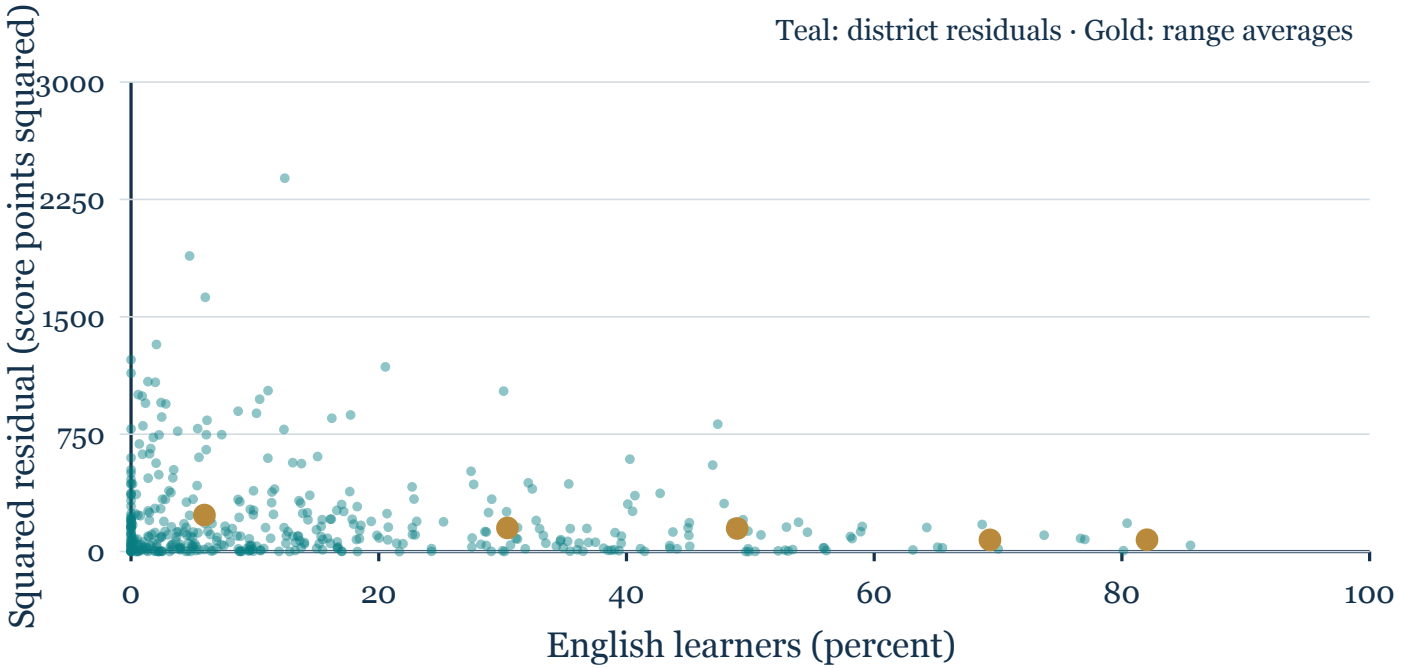
> With slope weights w_i, $\operatorname{Var}(\widehat{\beta}_1 \mid \mathcal{X}) = \sum_i w_i^2 \sigma_i^2$.

> HC0 estimates this by $\sum_i w_i^2 \widehat{u}_i^2$.

> HC1 multiplies HC0 by $n/(n-p)$. The same OLS fit supplies the residuals.

Look for a variance pattern before choosing a remedy

A squared residual is an observable clue about error spread.



Actual OLS residuals from 420 districts. Gold points average squared residuals in five language-share ranges; these are diagnostic summaries, not a known variance curve.

- Residual: observed outcome minus fitted outcome.
- Squared residual: $\widehat{u}_i^2$.
- Gold: average squared residual within each of five language-share ranges.

Breusch–Pagan turns the variance question into a test

The null is constant conditional error variance, given the chosen mean model.

```
. estat hettest str el, iid

Breusch–Pagan/Cook–Weisberg test for heteroskedasticity
Assumption: i.i.d. error terms
Variables: str el

H0: Constant variance

    chi2(2) =  29.50
Prob > chi2 = 0.0000
```

Actual StataMP output; reproduce with variance.do in the Stata download.

- $H_0 : \mathrm{Var}(u_i \mid X_i) = \sigma^2$: constant variance; X_i denotes the included regressors.
- Regress $\widehat{u}_i^2$ on a constant, staffing and language share.
- Test statistic: nR^2_{aux}, compared with χ^2_2.

White allows a broader variance pattern

Squares and products expand the auxiliary variance regression.

```
. estat imtest, white

White's test
H0: Homoskedasticity
Ha: Unrestricted heteroskedasticity

    chi2(5) =  42.46
Prob > chi2 = 0.0000
```

Actual White-test block. The complete diagnostic output is in variance_output.txt.

- Auxiliary columns: staffing, language share, their squares and their product.
- White statistic: nR^2_{aux}, compared with χ^2_5.
- Distinguish a variance diagnostic from a test of the causal comparison.

Practice

Why do BP and White have different degrees of freedom here?

Explanation: BP uses two auxiliary slopes; White adds two squares and one product, giving five.

Weighted least squares changes how the line is fitted

Give less weight to observations with greater conditional error variance.

> WLS minimises $\sum_i w_i (y_i - a - b_1 STR_i - b_2 EL_i)^2$.

> Candidate model: $\text{Var}(u_i \mid X_i, n_i) = \sigma^2 / n_i \Rightarrow w_i = n_i$.

> Multiply y_i, every regressor and the constant by $\sqrt{n_i}$.

Read a weighted fit without assuming it must be better

A variance model and an economic target both matter.

```
. regress score str el [aw=enrol], vce(robust)

------------------------------------------------------------------------------
             |               Robust
       score | Coefficient  std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
         str |   -1.32611   .6605944    -2.01   0.045     -2.62462   -.0276004
          el |  -.6390283   .0408085   -15.66   0.000    -.7192442   -.5588124
       _cons |   693.1913   13.26319    52.26   0.000     667.1203    719.2624
------------------------------------------------------------------------------
```

Actual StataMP output; reproduce with variance.do in the Stata download. Coefficient table shown; the download retains the full model output.

- OLS and WLS use the same 420 districts.
- The transformed regression reproduces the weighted coefficient.
- Weighting does not remove an omitted common cause.

Practice

Which column is easy to forget when transforming the weighted model?

Explanation: Multiply the constant column by the square root of enrolment as well.

Compare conventional and robust output

Same coefficients, different estimated uncertainty.

```
. regress score str el, vce(robust)

------------------------------------------------------------------------------
             |               Robust
       score | Coefficient  std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
         str |  -1.101296   .4328472    -2.54   0.011     -1.95213   -.2504616
          el |  -.6497768   .0310318   -20.94   0.000     -.710775   -.5887786
       _cons |   686.0322   8.728224    78.60   0.000     668.8754     703.189
------------------------------------------------------------------------------
```

After the robust fit, use test str el for the joint null of two zero slopes.

Use the covariance from the fitted model for joint inference.

Measurement error creates a different source of inconsistency

A noisy regressor is not just a noisier outcome.

$$W = X + M, \quad Y = a + \beta W + (u - \beta M)$$

$$\operatorname{plim} \hat{b}_W = \beta \frac{\operatorname{Var}(X)}{\operatorname{Var}(X) + \operatorname{Var}(M)}$$

Check attenuation in a separate generated comparison

Use a clean outcome equation and known noise variance.

```
. regress clean_y noisy_x, vce(robust)

------------------------------------------------------------------------------
             |               Robust
     clean_y | Coefficient  std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
     noisy_x |   .5177554   .0112879    45.87   0.000     .4956249    .5398859
       _cons |   1.045476   .0221565    47.19   0.000     1.002037    1.088915
------------------------------------------------------------------------------
```

Missing observations can change the population represented

Complete-case analysis needs a selection argument.

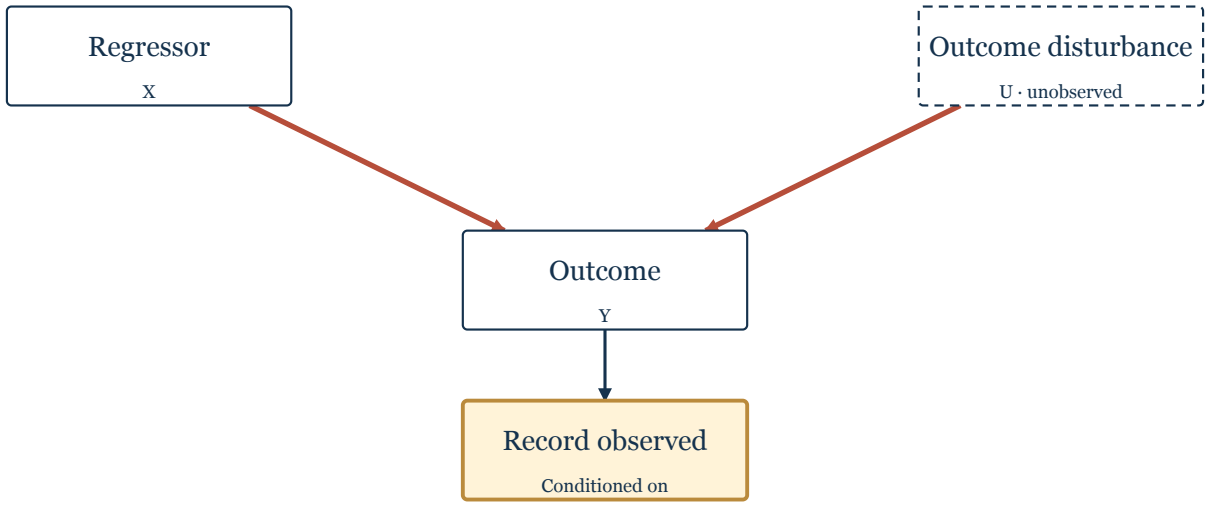


1 open directed route · 0 open alternatives · gold: conditioned on

Assumed causal structure · dashed boxes: unobserved · gold boxes: conditioned on

A missing-data assumption should be stated in context

Explain what makes the analysed cases informative.



0 open directed routes · 1 open alternative · gold: conditioned on

Assumed causal structure · dashed boxes: unobserved · gold boxes: conditioned on

Selection on observed inputs needs a conditional-mean argument

A nonrandom sample can preserve a coefficient while changing its target population.

> $S = 1$: included observation; key mean condition: $E[u \mid X, S = 1] = 0$.

> One sufficient restriction: $P(S = 1 \mid X, u) = P(S = 1 \mid X)$, with $E[u \mid X] = 0$.

> Preserved conditional coefficients do not imply population representativeness.

Leverage and influence are different

An unusual regressor value need not be a bad observation.

> Leverage: unusual regressor values.

> Large residual: unusual outcome given the fitted predictors.

> Influence: material change in the fitted result when the case changes.

Choose the covariance and estimator for the stated problem

Try the comparison before opening the explanation.

Practice

For each case, name what changes and justify it: unequal district error variances under zero conditional mean; measurement error in staffing; selection related to an unobserved cause of scores. Does a small robust p-value establish a valid causal comparison?

Choose the covariance and estimator for the stated problem: explanation

Follow the calculation and interpret its units.

> Independent-data robust covariance addresses unequal variance without changing OLS coefficients. Measurement error and endogenous selection may alter identification and require additional information or design assumptions. A robust p-value does not establish those assumptions.

Choose a remedy for the named problem

More observations, better measurement and better design do different jobs.

> If uncertainty is large under a credible comparison, more informative observations may help.

> State the problem before proposing a remedy.

From data problems to ordered observations

> From data problems to ordered observations

A robust Wald test uses the covariance of the restrictions

Account for co-movement when testing several equalities.

> Matrix HC0: $(X\prime X)^{-1}[\sum_i x_i x_i\prime \widehat{u}_i^2](X\prime X)^{-1}$.

> $H_0 : R\beta = r$; $g = R\widehat{\beta} - r$; $R : q \times p$.

> $W = g\prime(R\widehat{V}R\prime)^{-1}g \xrightarrow{d} \chi^2_q$.

> For one restriction, $W = t^2$.

Generalised least squares includes variance and dependence

WLS is one special case of a broader transformation idea.

> GLS: transform using the assumed error variances and covariances.

> Feasible GLS: estimate that error structure before transforming.

> Choose separately: mean specification, estimator and uncertainty calculation.

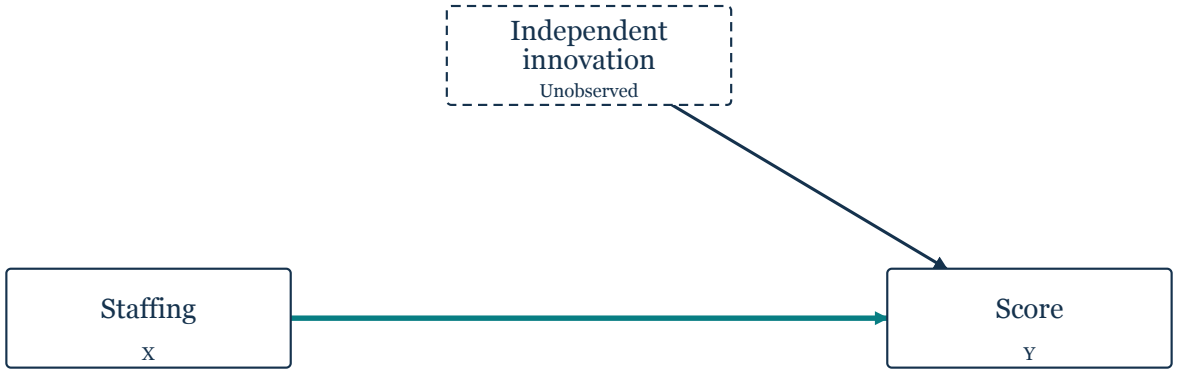
Practice

Would weighting remove an omitted cause related to staffing and scores?

Explanation: No. Justify the causal comparison separately from the error-variance model.

A DAG does not encode the size of the noise

Use a mean-zero innovation with a regressor-dependent scale.



1 open directed route · 0 open alternatives · gold: conditioned on

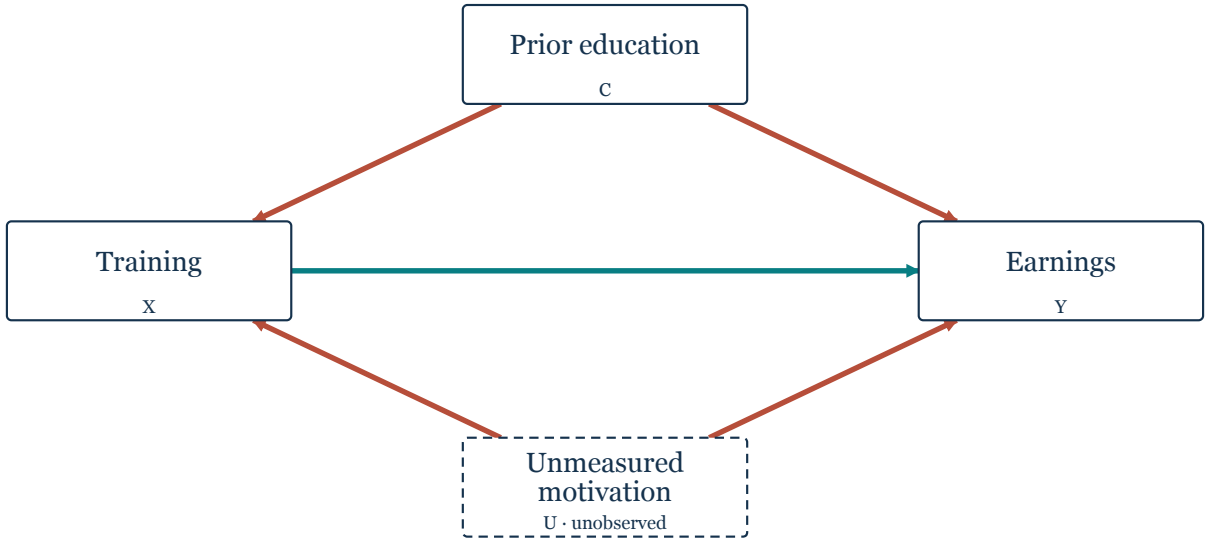
Assumed causal structure · dashed boxes: unobserved · gold boxes: conditioned on

$$Y = a + bX + \sigma(X)\varepsilon$$

$$E(u \mid X) = 0, \quad \mathrm{Var}(u \mid X) = \sigma(X)^2$$

Do not choose an SE by the resulting significance

Match it to the sampling and dependence structure.



1 open directed route · 2 open alternatives · gold: conditioned on

Assumed causal structure · dashed boxes: unobserved · gold boxes: conditioned on

Robust does what?

Try the comparison before opening the explanation.

Practice

After changing regress y x to regress y x, vce(robust), which point estimate changes? Does this solve omitted-variable bias?

Robust does what?: explanation

Follow the calculation and interpret its units.

> On the same sample the OLS coefficient does not change. The estimated covariance matrix changes, so SEs, intervals and tests can change. Heteroskedasticity-robust inference does not remove omitted-variable bias or automatically allow serial dependence.

Measurement error

Try the comparison before opening the explanation.

Practice

Observed x equals true X^* plus mean-zero noise independent of X^* and the outcome error. In a simple regression, what happens to the slope toward a positive true effect?

Measurement error: explanation

Follow the calculation and interpret its units.

> Under classical measurement-error assumptions, the probability limit is $\beta \dfrac{\mathrm{Var}(X^*)}{\mathrm{Var}(X^*) + \mathrm{Var}(noise)}$, which is between zero and β. More observations do not remove this attenuation. The result does not automatically extend to nonclassical error or every multiple-regression setting.

Outcome noise has another role

Mean-zero independent outcome error need not attenuate the slope.

> If the outcome is measured with an additional error unrelated to the regressors, that error enters the regression disturbance.

> The direction of measurement matters.

Residual scaling and Cook’s distance answer different diagnostic questions

A diagnostic threshold prompts investigation rather than automatic deletion.

> $h_i = [X(X\prime X)^{-1}X\prime]_{ii}$, $\sum_i h_i = p$.

> Internally studentised residual: $r_i = \widehat{u}_i / [s\sqrt{1-h_i}]$.

> Cook distance: $D_i = \dfrac{\widehat{u}_i^2}{ps^2}\dfrac{h_i}{(1-h_i)^2}$.

Influence

Try the comparison before opening the explanation.

Practice

A high-income district sits far from the remaining x values. Must it be deleted?

Influence: explanation

Follow the calculation and interpret its units.

> No. A distant x value can have high leverage; influence also depends on its residual and its effect on the fit. Verify the observation, explain substantive relevance and show defensible sensitivity analysis. Delete a confirmed error for a documented reason, not because it changes significance.

Sensitivity analysis answers a stated question

Change one defensible feature and explain the comparison.

> Sensitivity analysis examines how conclusions change under plausible alternative choices, such as a sample restriction, transformation or influential observation.

> For example, compare a fitted income relationship with and without one verified extreme-income record and show the range supported by each sample.

The data audit belongs before the policy sentence

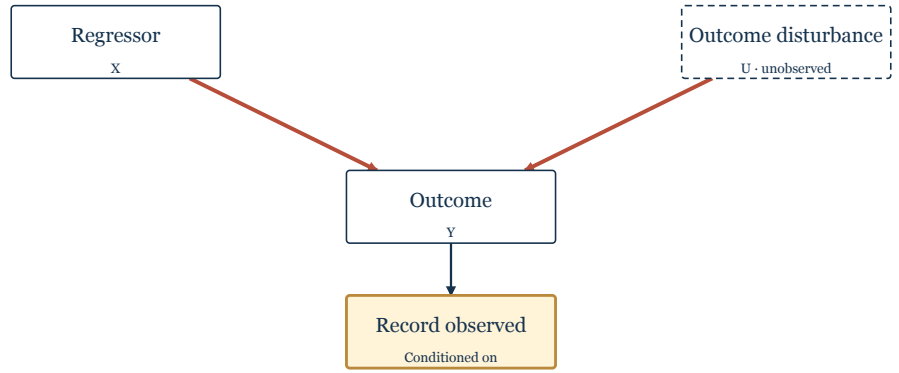
Units, ranges and recording rules affect interpretation.

> Check how variables were measured, whether their ranges are possible and whether units changed over time or across sources.

> Record sample restrictions and verify identifiers before merging data.

DAG practice: selection on an outcome descendant

Trace the routes before choosing a control.



0 open directed routes · 1 open alternative · gold: conditioned on

Assumed causal structure · dashed boxes: unobserved · gold boxes: conditioned on

Practice

Suppose X and the outcome disturbance U are independent before selection, both affect Y, and a record is observed depending on Y. Why can restricting attention to observed records create an X–U association? Compare this with selection depending only on included X.

DAG practice: selection on an outcome descendant: explanation

Explain what conditioning changes in this stated graph.

> Y is a collider on $X \to Y \leftarrow U$. The observation indicator S is its descendant. Conditioning on $S = 1$ can open the collider path, so the selected sample can violate the disturbance mean assumption even when the original sample satisfied it. If selection instead depends only on included X and carries no extra information about U conditional on X, this particular path is not opened. The conditional slope may remain valid under a correct model and adequate support, although precision and the represented population can change. The selection mechanism, rather than missingness alone, determines the concern.

Separate simulation truth from estimated evidence

Reproduce the evidence, then explain its economic meaning.

Practice

Then use the actual schools.csv data. Fit score on staffing, language share and income; identify the largest Cook distance and compare the staffing slope with and without that district. Cook distance summarises how much omitting a case changes the fitted regression. Report this as sensitivity, not an instruction to delete it. Keep your commands, log, output and interpretation.

Separate simulation truth from estimated evidence: explanation

A reproducible result needs both a calculation and an interpretation.

> Simulation: direct coefficient 1; omitted-factor probability limit 1.5.

> Real-data sensitivity: staffing slope $-0.068776 \rightarrow -0.093391$.

> 420 versus 419 districts; investigate an influential case before deciding whether to exclude it.