

Difficulties with Linear Least Squares

Recall standard assumptions of linear regression using **ordinary least squares (OLS)**:

- $E(Y_i) = \mathbf{x}'_i \boldsymbol{\beta}$, where each \mathbf{x}_i is a known constant vector, and $\boldsymbol{\beta}$ is entirely unknown
- $V(Y_i) = \sigma^2$ (unknown finite constant, not depending on i , and unrelated to $\boldsymbol{\beta}$)
- $\text{Cov}(Y_i, Y_{i'}) = 0$, $i \neq i'$
- For hypothesis tests and confidence intervals,

$$\mathbf{Y} = [Y_1 \cdots Y_n]'$$

has a multivariate normal distribution
(viz. $N(\mathbf{X}\boldsymbol{\beta}, \mathbf{I}\sigma^2)$, so Y_i s are independent).

What happens if these assumptions are not met?
Even if they are, when might the results of an OLS fit have undesirable properties?

Some conditions that can cause difficulties for ordinary least squares linear regression:

- Non-Normality of Y
- Non-Constant Variance of Y
- Correlation Among Y s
- Influential Data Points and Outliers
- Inadequacies in Mean Model
- Near-Collinearity Among X Columns
- Errors in Independent Variables

(These cause problems for estimation and testing. Even if they are remedied, still need to *validate* the fitted model, as discussed previously.)

Non-Normality

Problem: Y s (or, equivalently, their errors) are not normally distributed

Most problematic if responses are discrete (e.g. binary or small counts) or if distribution is highly skewed (e.g. distributions of maxima).

Hypothesis tests and confidence intervals/sets are affected — the true α may be slightly different than the nominal α . F -tests are not strongly affected, but some two-sided CIs may be misleading.

OLS still has good properties (e.g. unbiasedness, lowest variance among linear unbiased estimators) even if errors are not normal, provided the other assumptions are satisfied.

Usually not a serious problem if sample size n is sufficiently large or if data are from a randomized experiment.

Diagnostics: frequency plots, normal probability plots, and various other plots of residuals; skewness and kurtosis coefficients; normality tests

Remedies: transformations of Y (Ch. 12); generalized linear models (other courses)

Non-Constant (Heterogeneous) Variances

Problem: $V(Y_i)$ depends on i

Variance often related to mean, perhaps because of natural boundaries on the response (e.g. non-negative Y s may have smaller variance when $E(Y)$ is near zero), or because experimental treatments affect both mean and variance of response.

OLS has less precision than weighted estimators (as we will see later). Estimation of σ^2 is no longer meaningful, and standard errors, tests, and CIs may be incorrect.

Diagnostics: e -versus- \hat{Y} and e -versus- X plots (Sec. 11.1); tests for heterogeneous variances among groups or relative to predicted values (Bartlett, Levene, Hartley, and others)

Remedies: transformations of Y (Ch. 12); weighted least squares (Sec. 12.5.1)

Correlated Errors

Problem: $\text{Cov}(Y_i, Y_{i'}) \neq 0$ for some $i \neq i'$

Usually occurs when observations are related in time or space (e.g. time series or geographic data) or by association with the same experimental unit (in designed experiments) or a latent random variable.

OLS has less precision than estimators that account for covariance. Standard errors can be badly biased, and tests and CIs may be substantially incorrect.

Diagnostics: residual-versus-time/space plots; lagged-residual and autocorrelation plots (serial data); variograms (spatial data); various time-series and multivariate tests

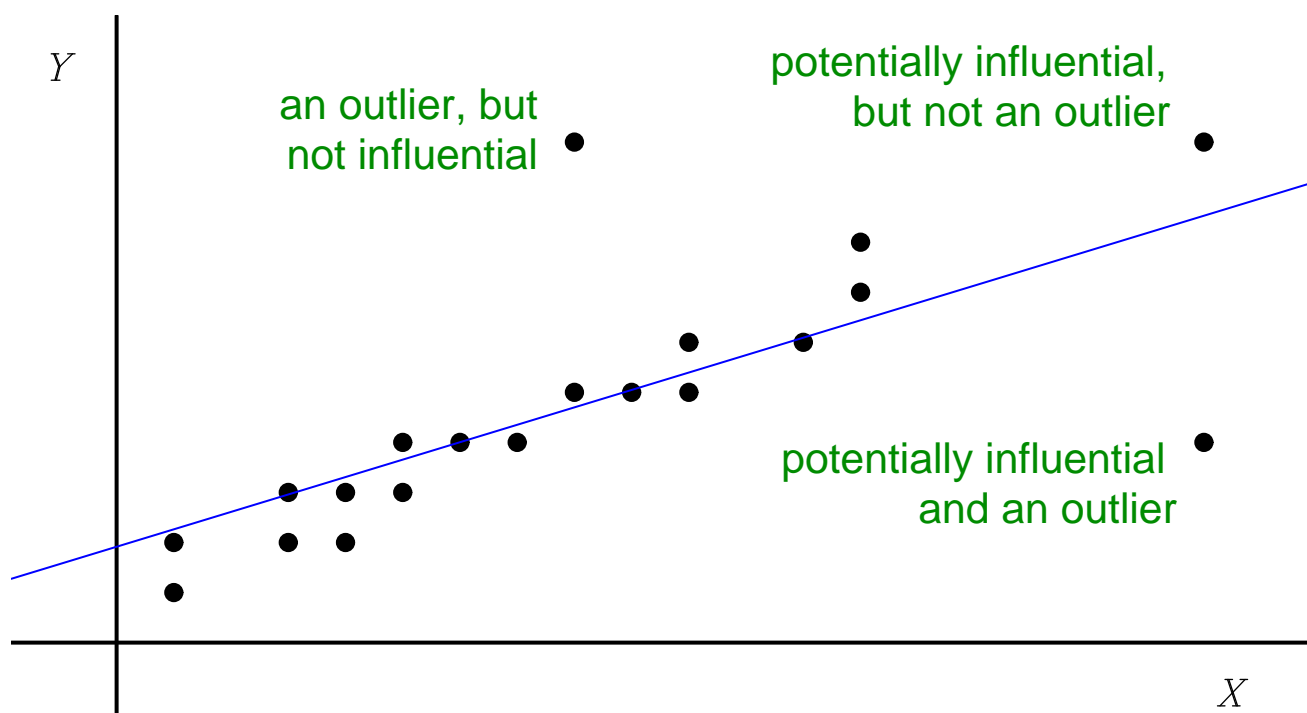
Remedies: covariance models (random effects, ARMA, ...) along with generalized least squares (Sec. 12.5.2)

Influential Data Points and Outliers

Problem: A few unusual observations have too much effect on the results.

Potentially influential (high-leverage) points are observations with *independent* variable values that (individually or collectively) are far outside of the range of the remaining observations.

Outliers are observations with unusually large or small *dependent* variable values, usually relative to what would be predicted based on the fitted model (or the other observations).



Either of these may result from mistakes in conducting an experiment, mistakes in recording the data, or unusual conditions or incidents that affect only a few observations.

OLS results can be very sensitive to outliers and influential points.

Problematic observations may either be useless, as when they are purely the result of mistakes, or highly useful, as when they provide unexpected information about the system that is not evident from the other observations.

Diagnostics: frequency plots of standardized or Studentized residuals; various residuals plots (Sec. 11.1); partial regression leverage plots (Sec. 11.1.6); influence statistics (Sec. 11.2)

Remedies: correction or deletion of faulty observations (if assigned a cause); robust regression procedures (other courses)

Inadequacies in the Mean Model

Problem: There is no β such that $E(Y) = x'\beta$ for all (observed or potential) data pairs (x, Y) .

Lack of fit is a special case in which there is no β such that $E(Y_i) = x'_i\beta$ for all observations i .

However, even if there is no lack of fit, the mean model may still be inadequate for extrapolation (and perhaps even for interpolation).

Could be due to unobserved but important independent variables, poor approximations (e.g. polynomials of insufficiently high degree), or unrealistic modeling (esp. for extrapolation).

Results in biased predictions and upwardly biased estimates of error variance. Parameter estimates may be biased and may even lose their relevance.

Models are rarely perfect for real data. In practice, adequate approximation for the given data, and for prediction, is sufficient.

Diagnostics: lack-of-fit tests (if lack of fit is the problem); e -versus- \hat{Y} and e -versus- X plots; partial regression leverage plots; various diagnostics using validation data

Remedies: larger models (with more derived variables); transformations of X_j s or Y (Ch. 12); nonlinear models (Ch. 15)

Collinearity

Problem: The \mathbf{X} matrix is exactly or nearly singular (for unintended reasons).

Occurs when independent variables happen to be highly correlated, but can also occur even if they are not.

Computation of $(\mathbf{X}'\mathbf{X})^{-1}$ is unstable. Parameter standard errors are very large, and many tests have low power. Estimates are also more sensitive to errors in the \mathbf{X} matrix itself.

In model selection, important variables might be mistakenly removed in favor of unimportant variables with which they are highly correlated.

Prediction within the joint range of the independent variables may still be safe, *but* the extent of this safe range is smaller under the condition of near-collinearity.

Diagnostics: correlation matrix; variance inflation factors (Sec. 11.3.2); eigen-analysis of standardized independent variables (multivariate analysis courses)

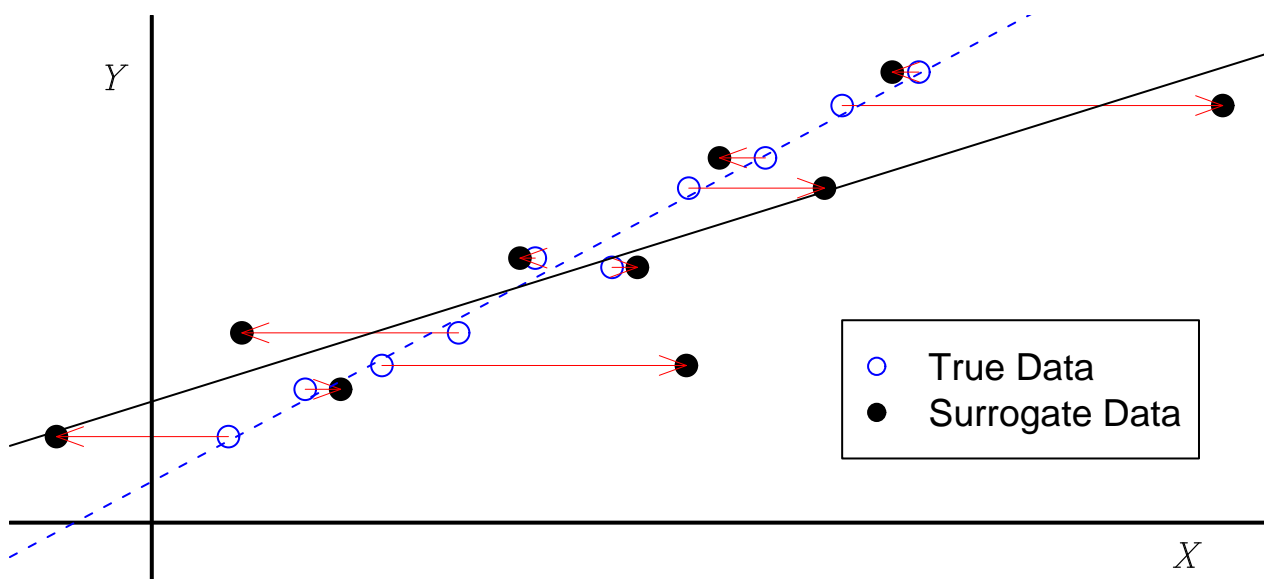
Remedies: additional observations (controlled to reduce correlations, if possible); standardization or orthogonalization of variables; biased regression methods (e.g. ridge regression)

Errors in Independent Variables

Problem: Instead of the “true” independent variables, you use “surrogate” independent variables that deviate randomly from the “true” variables.

Could be due to imprecise measurement of the intended independent variables or to imprecise application of treatment levels (in experiments).

Causes biased parameter estimates (even if surrogates are unbiased for the true independent variables). The bias is nontrivial to estimate, especially with multiple independent variables.



In SLR, errors in X tend to bias the slope estimate toward zero. But when there is more than one independent variable, biases in coefficient estimates can be either toward or away from zero.

Prediction is less precise, but generally still valid.

Diagnostics: additional studies to determine the nature and extent of the errors

Remedies: specialized measurement error models and methodologies (orthogonal regression, regression calibration, SIMEX)

Other Difficulties and Limitations

- Missing/Censored Data: Estimates can be biased if incomplete or unavailable observations are ignored (if the reasons for missingness or censoring depend on the values assumed by the observations).
- Can't make use of (imprecise) prior information about parameters.
- Can't penalize some types of deviations more severely than others.
- Requires additional methodology (like tests or variable selection criteria) to select a subset of terms.