EPSE 596: Correlational Designs and Analysis

Ed Kroc

University of British Columbia

ed.kroc@ubc.ca

Return to the idea of "best fit"

- Recall in when we first wrote down the simple linear regression model:

$$Y = \beta_0 + \beta_1 X + \varepsilon,$$

where we assume $\varepsilon \sim N(0, \sigma^2)$ for some unknown but fixed $\sigma > 0$.

- We said: In practice, we observe n sample data points $(x_1, y_1), \ldots, (x_n, y_n)$, which we could then plug into our proposed regression model to get a sequence of n equations:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,$$

where $1 \leqslant i \leqslant n$.

Return to the idea of "best fit"

- These allow us to find the *errors* we make by using the proposed model for the observed data:

$$\varepsilon_i = y_i - \beta_0 - \beta_1 x_i$$

- But β_0, β_1 are unknown. So makes sense to choose estimators $\widehat{\beta}_0$, $\widehat{\beta}_1$ for these unknowns that *make the observed errors (i.e., residuals, e_i) small*.
- That is, choose $\widehat{\beta}_0$, $\widehat{\beta}_1$ so that

$$e_i = y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i$$

  are small (in some sense).
- **Big Question:** How should we choose the numbers $\widehat{\beta}_0$ and $\widehat{\beta}_1$ to make the errors given by the the proposed regression model small?
- **Realize: there is no one, unique, objective notion of what is small**; i.e., there are many ways to answer this question, and no one, objectively best or correct answer.

Return to the idea of "best fit"

- In Week 4 (and since then), we settled on the *ordinary least squares (OLS)* definition of "smallness".
- That is, we decided to choose the estimators $\widehat{\beta}_0$, $\widehat{\beta}_1$ so that

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i)^2 \text{ is minimized.}$$

  Recall: We worked with squares rather than absolute values just because they are algebraically and computationally more convenient.

- This yielded the OLS estimators, designed to ensure the sum of the squares of the residuals are minimized:

$$\widehat{\beta}_0^{\,OLS} := \overline{Y} - \widehat{\beta}_1^{\,OLS} \overline{X}$$
$$\widehat{\beta}_1^{\,OLS} := \frac{s_{XY}}{s_X^2}$$

Return to the idea of "best fit"

- But now let's consider some other common criteria for "make the observed errors small;" i.e., some different ways to find the "best fit" estimators.
- There are *infinitely many ways to do this.* E.g.:
  - Ordinary least squares
  - Maximum likelihood
  - Method of moments, weighted least squares, robust estimation
  - Bayesian estimation
- Realize: none of these different methods change anything about the *structure/specification* of the proposed regression model, they simply change what the proposed *estimators* of the unknown model parameters are.
- Some other approaches (e.g., restricted or quasi maximum likelihood) impose extra or less structure to the model, and then derive "best" estimators.

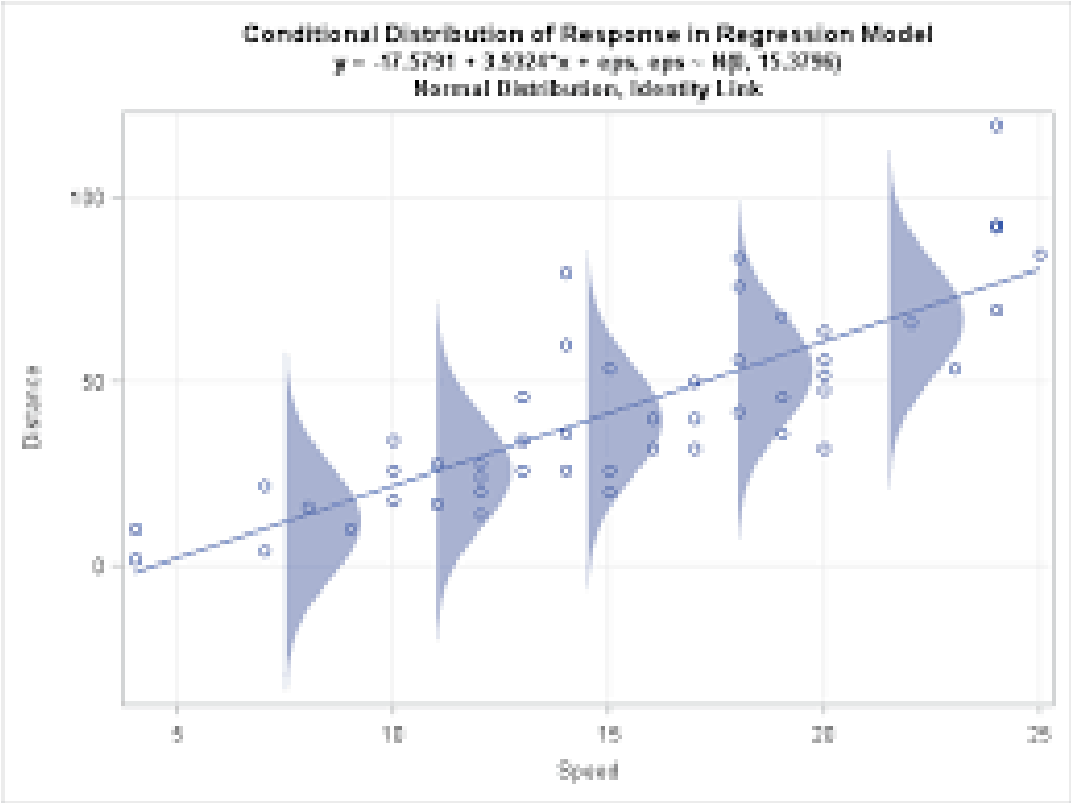
Return to the idea of "best fit"

- *Ordinary least squares* (OLS) and *maximum likelihood* (ML) estimation are, by far, the two most classical and common methods of model parameter estimation.
- And for simple modelling frameworks, like all of the regression models we have considered in this class, it turns out that OLS and ML estimation *is the same thing*.
- In more complex frameworks (e.g., random/mixed effects modelling: EPSE 594, 683), these methods of estimation do *not* necessarily give the same answers/estimators.
- Moreover, ML estimation has been tweaked and generalized in so many other useful ways (e.g., restricted and quasi-ML) that it is critical to understand the estimation procedure itself.

Maximum likelihood (ML) estimation

- Recall: A compact way to write down a simple regression model with the assumptions on the errors encoded is:

$$Y \sim N(\beta_0 + \beta_1 X, \sigma^2).$$



Maximum likelihood (ML) estimation

- We can equivalently express the model

$$Y \sim N(\beta_0 + \beta_1 X, \sigma^2)$$

as a more explicit probability statement because we are literally saying that *"we propose Y follows a normal distribution with a particular mean (dependent on X) and a particular variance*.

- So we can restate this model as the equation of the normal distribution being proposed. Easiest way to do this is to write down the *probability density function* of the normal random variable; i.e., the equation of the particular proposed "bell curve" for Y given X:

$$f(y|x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y-\beta_0-\beta_X x)^2}{2\sigma^2}}$$

Maximum likelihood (ML) estimation

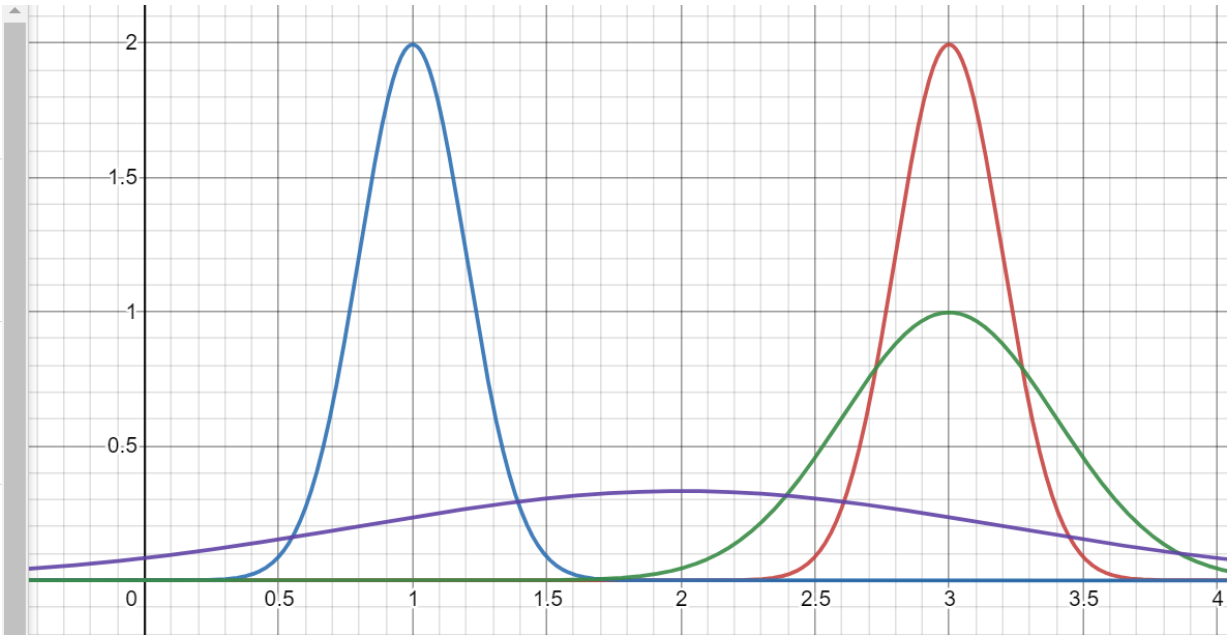
- Different bell curves:

$$y = \frac{1}{\sqrt{2 \cdot \pi \cdot .2^2}} \cdot e^{-\frac{(x-3)^2}{2 \cdot .2^2}}$$

$$y = \frac{1}{\sqrt{2 \cdot \pi \cdot .2^2}} \cdot e^{-\frac{(x-1)^2}{2 \cdot .2^2}}$$

$$y = \frac{1}{\sqrt{2 \cdot \pi \cdot .4^2}} \cdot e^{-\frac{(x-3)^2}{2 \cdot .4^2}}$$

$$y = \frac{1}{\sqrt{2 \cdot \pi \cdot 1.2^2}} \cdot e^{-\frac{(x-2)^2}{2 \cdot 1.2^2}}$$



- Notice: For "tight" curves (i.e., when σ is very small), the values of the function can be greater than 1.

- Hence, $f(y)$ is *not* the probability that the normal r.v. equals y. But it *is* the "likelihood" that the normal r.v. equals y.

Maximum likelihood (ML) estimation

Probability vs. likelihood:

- General mathematical fact: intuition about *continuous* objects is less accurate/useful than intuition about *discrete/finite* objects. [General mathematical irony: analyzing continuous objects is usually easier than analyzing discrete/finite objects.]
- *Probability* encodes how likely/believable some random event is to occur on a standardized [0,1] scale.
- *Likelihood* encodes how likely/believable some random event is to occur *relative* to the other, alternative outcomes.
- Analogous to the difference between *correlation* and *covariance*. Both measure the same thing, i.e., the strength of the linear relationship between two phenomena. But *correlation* is measured on a standardized scale of [-1,1], whereas *covariance* can assume any value, so only meaningful *relative* to something else.

Maximum likelihood (ML) estimation

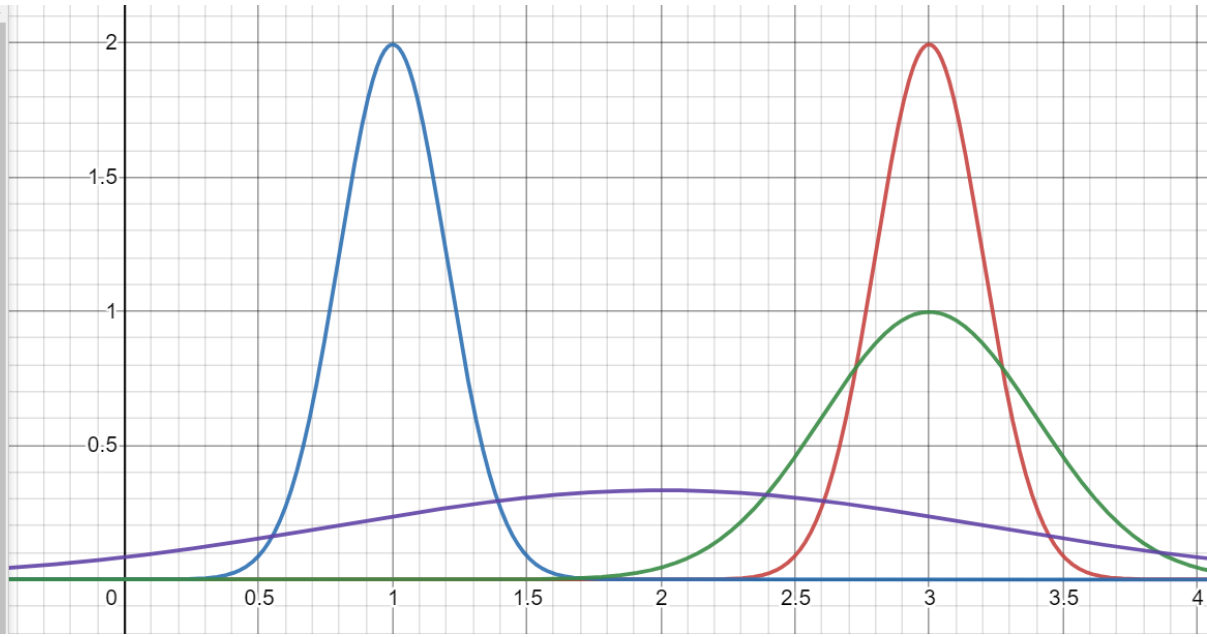
- Different bell curves:

$$y = \frac{1}{\sqrt{2 \cdot \pi \cdot .2^2}} \cdot e^{-\frac{(x-3)^2}{2 \cdot .2^2}}$$

$$y = \frac{1}{\sqrt{2 \cdot \pi \cdot .2^2}} \cdot e^{-\frac{(x-1)^2}{2 \cdot .2^2}}$$

$$y = \frac{1}{\sqrt{2 \cdot \pi \cdot .4^2}} \cdot e^{-\frac{(x-3)^2}{2 \cdot .4^2}}$$

$$y = \frac{1}{\sqrt{2 \cdot \pi \cdot 1.2^2}} \cdot e^{-\frac{(x-2)^2}{2 \cdot 1.2^2}}$$



- Notice: for the green curve say, it's clear that an outcome near 3 is more likely than an outcome near 2.5.
- Compare with the red curve: the same statement is true, but the different shape of the curve shows that an outcome near 3 is *far more likely* than an outcome near 2.5.

Maximum likelihood (ML) estimation

- All this to say, if we observe a sample datum (x, y), then under our proposed (simple) regression model, the *likelihood* of observing the sample datum is:

$$\ell(\beta_0, \beta_1, \sigma \mid x, y) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y-\beta_0-\beta_1 x)^2}{2\sigma^2}}$$

- If we observe a bunch of sample data $(x_1, y_1), \ldots, (x_n, y_n)$ *independently*, then the *likelihood* of observing this entire sample, assuming our proposed regression model, is:

$$\ell(\beta_0, \beta_1, \sigma \mid (x_1, y_1), \ldots, (x_n, y_n)) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y_i-\beta_0-\beta_1 x_i)^2}{2\sigma^2}}$$

- This is called the *likelihood function* for the sample data under the proposed model.

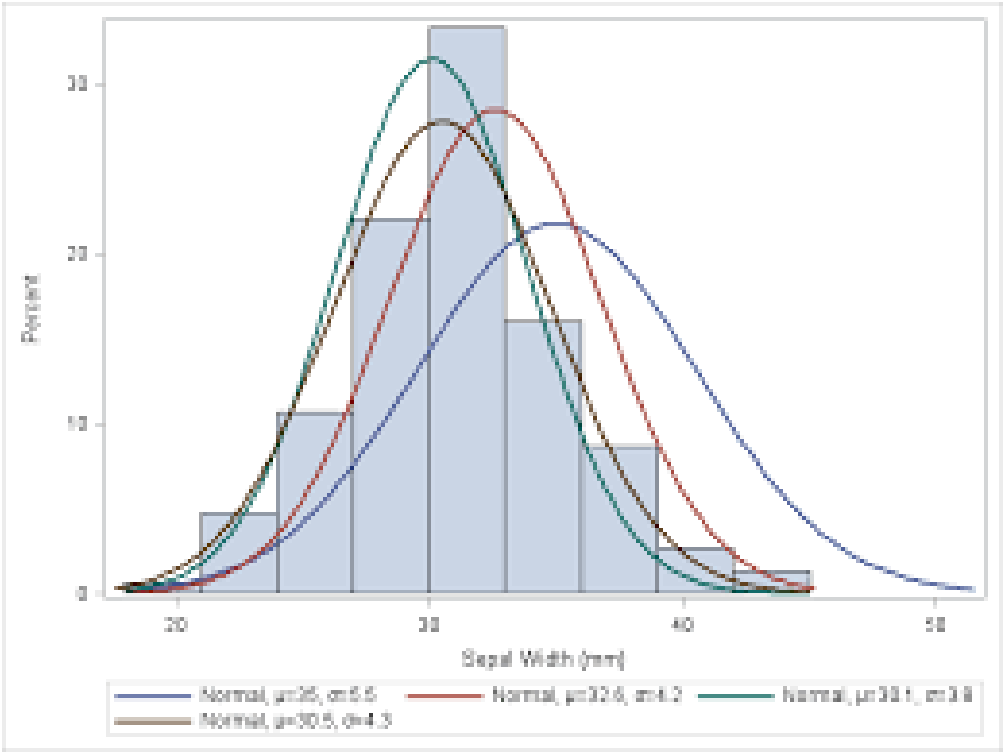
Maximum likelihood (ML) estimation

- For convenience, use the common vector notation: $\mathbf{x} = \{x_1, \ldots, x_n\}$, $\mathbf{y} = \{y_1, \ldots, y_n\}$.
- Notice that this likelihood function depends on the unknown model parameters β_0, β_1, and σ:

$$\ell(\beta_0, \beta_1, \sigma \mid \mathbf{x}, \mathbf{y}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma^2}}$$

- So a reasonable way to choose "best estimates" for these unknowns is to choose numbers $\widehat{\beta}_0$, $\widehat{\beta}_1$, σ that *maximize* the likelihood function.
- This is another calculus exercise.
- Maximizing the likelihood function is equivalent to finding the model parameters that are *most likely* to have generated the observed sample data, assuming our model is the correct one.

Maximum likelihood (ML) estimation



- Plotted are normal likelihood functions with four different means and variances (i.e. four different collections of values for β_0, β_X, and σ).
- The histogram corresponds to our *observed* sample data.
- Clearly, the blue curve is less likely to have generated these sample data than the other curves; i.e. the parameters that specify the blue curve do *not* maximize the sample likelihood.

Maximum likelihood (ML) estimation

- Interestingly, maximizing the (normal) likelihood function

$$\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y_i-\beta_0-\beta_1 x_i)^2}{2\sigma^2}} = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n e^{-\frac{\sum_{i=1}^{n}(y_i-\beta_0-\beta_1 x_i)^2}{2\sigma^2}}$$

simultaneously in β_0, β_1 yields the *same solutions* as if we minimize

$$\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$$

- That is, for ordinary linear regression, ML and OLS estimation give the *same estimators*:

$$\hat{\beta}_0^{MLE} := \hat{\beta}_0^{OLS}$$
$$\hat{\beta}_1^{MLE} := \hat{\beta}_1^{OLS}$$

- This holds regardless of how many predictors are in the model; this is a unique feature of ordinary linear regression. When we consider more complex models and/or data structures, these two estimation methods can produce *different* sample estimators of the model parameters.

Estimating regression model parameters

So if the OLS and ML estimators are the same, why should we bother with the distinction?

- For ordinary linear regression models (i.e., all modelling *fixed effects*, all assuming *normal errors*, all assuming *independent observations*, all assuming *no measurement error*, etc.), OLS and ML are equivalent.
- But once you start looking at less basic models (i.e., relax any of those simplifying assumptions), OLS and ML can produce different answers.
- Moreover, ML is mathematically more convenient to generalize and adapt to novel situations.

Restricted maximum likelihood estimation

Restricted maximum likelihood (REML):

- REML is very common when estimating random/mixed effects models (e.g., models containing data with repeated measures, i.e., stochastic dependence structure)
- For a general mixed effects regression model, one has the pseudo-equation:

$$\text{response} = \text{fixed effects} + \text{random effects} + \text{error},$$

where the fixed effects are the usual deterministic portion of our model (i.e., the function that describes how the predictors relate to the response on average), and the random part of the model is now decomposed into two pieces: a *random effects part* that captures some kind of stochastic dependence structure in the response *after* accounting for the deterministic dependence structure of the fixed effects (e.g., think repeated measures), and a *random error* which is unique to each data point given the model, as before.

Restricted maximum likelihood estimation

- With fixed effects models (i.e., the ones we have always considered), the *random effects* part of the general model above is zero, and we have to estimate the unknown coefficients on the terms in the *fixed effects* part of the above model (i.e., the β coefficients), and the residual variation that is captured by the *error* part, σ.
- In the more general mixed effects framework, we still need to estimate these quantities in addition to the variance and/or dependence components of the *random effects* part of the model. Usually, one assumes these random effects are mean zero (like the *error*), but that they have some kind of unique variance/covariance structure between observations.
    - For example, student responses are likely to be more correlated when the students all come from the same class or school.

Restricted maximum likelihood estimation

- REML works as follows:

  (1) First estimate the *average response* with ML as before, i.e., ignoring the *random effects* part of the model.

  (2) Then estimate the *dependence/covariance* between related/repeated observations using ML for every fitted (i.e., estimated average) response. That is, perform a regression on the fixed effects model's *residuals* to estimate the variance/covariance parameters, including the leftover individual *error* variance.

- Thus, when we want to account for a stochastic dependence structure to the data, we first *restrict* the ML estimation to a particular cross-section of the data (via the fitted values), and then use ML again to estimate the variance/covariance of the residuals across the fitted values.

Restricted maximum likelihood estimation

- REML usually produces "better" (i.e., *unbiased*) estimators than simple ML estimation for the random parts of the model.
- This is true even for the purely fixed effects models we consider in this class. Recall from Week 4, we estimated the *residual variance* as

$$\hat{\sigma}^2 = \frac{1}{n-p}\sum_{i=1}^{n} e_i^2,$$

where n is the number of data points and p is the number of coefficients in our model that we need to estimate. This is an *unbiased* estimator of the residual variance, meaning: *on average, this estimator equals the population quantity it is trying to estimate*.

Restricted maximum likelihood estimation

- However, the ML estimate of the residual variance is

$$\widehat{\sigma}^2_{ML} = \frac{1}{n}\sum_{i=1}^{n} e_i^2$$

  This is a *biased* estimator of the target quantity (on average, it is *too small*), but notice that when n is large (relative to p), this estimator is very close to the other one. In fact, they converge to the same, true population quantity as the sample size increases; i.e., both estimators are *consistent*.

- REML fixes this *bias* problem; i.e., one gets

$$\widehat{\sigma}^2_{REML} = \frac{1}{n-p}\sum_{i=1}^{n} e_i^2$$

- Mixed effects models (and REML) are common in meta-analysis (EPSE 594), and when one wants to account for hierarchical structure, repeated measures, or stochastic temporal or spatial patterns (EPSE 683).

Quasi-maximum likelihood (qML) estimation

- qML is very common when estimating certain *generalized linear models* (GLMs) that are adaptations of our ordinary linear regression framework to handle *non-continuous response data* and/or *non-normal error structures*.
- Basic idea: For non-continuous data, the errors cannot (by definition) be normally distributed.
  - But normal distributions are super convenient because the *mean* parameter is always totally separate from the *variance* parameter: $N(\mu, \sigma^2)$.
  - Most other probability distribution do *not* have this feature; i.e., μ is a function of σ^2 or vice-versa.
  - qML relaxes this condition by adding an extra *dispersion parameter* to the regression model (which is then estimated via ML).

Quasi-maximum likelihood (qML) estimation

Example: Binary (0/1) response data:

- Suppose for any value of your predictor X, your response Y can only be either 0 or 1; e.g., X represents a student's GPA in high school and Y represents whether or not they pass their intro stats course.
- If we propose a model like

$$Y \sim N(\beta_0 + \beta_1 X, \sigma^2),$$

this will get us into a lot of trouble because for any particular value of X (i.e., any particular GPA), regardless of what β_0 and β_1 actually are, Y can only be one of two quantities. Thus, there's no way for the response to "vary" around the mean value, $\beta_0 + \beta_1 X$, like a normal curve does (in particular, there's no continuum of values for Y to vary over!).

- So instead, we specify a different random error structure for the response; i.e., we propose a *Binomial* model for the response variable.

Quasi-maximum likelihood (qML) estimation

- Binomial random variables simply count the number of "successes" in a sequence of "trials." Here, that means for every different incoming GPA value, we count the number of students with that GPA that pass the course.
- One can show that for Binomial random variables, the mean and variance are given by:

$$\mu = mp, \quad \sigma^2 = mp(1-p),$$

where m is the number of "trials" and p is the (unknown, to be estimated) probability of "success."

- Notice then that we have

$$\sigma^2 = \mu(1-p) = \mu\left(1 - \frac{\mu}{m}\right),$$

that is, the variance is an explicit function of the mean!

Quasi-maximum likelihood (qML) estimation

- This has all kinds of important implications, but in particular it means that correctly specifiying the deterministic part of the model cannot be separated from correctly specifying the stochastic part of the model.
- qML helps to (partially) solve this problem by introducing a *dispersion parameter* to essentially separate the model again into a deterministic (fixed effects of predictors) part and a stochastic (random error) part.
- We will talk more about regression with a binary response variable in the last week of class (via *logistic regression*).
- Much more about this and other *generalized linear models* (GLMs) in EPSE 682; e.g., what if your response data are counts: 0,1,2,...?

Bayesian estimation

- Bayesian methods are often talked about as super-fancy, complex, cutting-edge, and/or new techniques by applied practitioners.
- Realize: Bayesian methods are mostly just a slightly different way of estimating model parameters (I'm ignoring a bunch of philosophical stuff, but in terms of what difference they make in real world research scenarios, that is essentially it).
- Bayesian estimation makes direct use of the likelihood function, in addition to something else called a *prior distribution* and a mathematical fact called *Bayes' Theorem/Formula* to produce "best" estimates for unknown model parameters.
- In most applied research scenarios, Bayesian methods produce answers that are very close to the other *likelihood theory* based solutions like ML/REML/qML (when done correctly).
- Nonetheless, for certain data situations (e.g. mixed effects models with lots of parameters), Bayesian estimation can be considerably easier or more efficient than ML-based estimation.

Bayesian estimation

In brief, Bayesian estimation proceeds as follows:

(1) Propose your regression model as always, and write down its *likelihood function*, $\ell(\beta_1, \ldots, \beta_p, \sigma \mid \mathbf{x}, \mathbf{y})$

(2) For your model parameters, specify a *prior distribution*, $\pi(\beta_1, \ldots, \beta_p, \sigma)$, that describes how likely it is that the parameters assume any particular value *apriori*; i.e., *before you have looked at, collected, or otherwise used your data*.

- This sounds subjective, but is often not at all. People have argued about best approaches for decades. Nowadays, there is little problem with removing this subjectivity (when appropriate) as long as one knows what one is doing.
- The ability to specify *priors* can greatly increase inferential power when one has good information about where a model parameter can or cannot live.

Bayesian estimation

(3) Use Bayes' Theorem to combine the *prior* (i.e., apriori information about the phenomena) and *likelihood* (i.e., information about the phenomena from your sample data, assuming your model) into a *posterior distribution*, $f(\beta_1, \ldots, \beta_p, \sigma \mid \mathbf{x}, \mathbf{y})$, for the unknown model parameters:

$$f(\beta_1, \ldots, \beta_p, \sigma \mid \mathbf{x}, \mathbf{y}) = c \cdot \pi(\beta_1, \ldots, \beta_p, \sigma) \cdot \ell(\beta_1, \ldots, \beta_p, \sigma \mid \mathbf{x}, \mathbf{y}),$$

where c is a constant.

(4) Can now get sample estimates for your unknown model parameters by, say, finding the *mode* or the *mean* of this posterior distribution. Similarly, can get *credibility intervals* (the Bayesian analogue of *confidence intervals*) by computing percentiles of this posterior distribution for the different parameters.

Bayesian estimation

Some notes:

- The upshot is that Bayesian estimation gives you quantities that look and behave and are used in all the same essential ways as the sample quantities you are already used to (like sample means, sample variances, OLS or MLE estimates of regression coeffiicents, etc.).
- In particular, realize that there is no such thing as a *Bayesian model* or *Bayesian regression*; there is simply *regression modelling* and then one *chooses how to estimate the unknown model parameters* using Bayesian methods, or maximum likelihood, or OLS, etc.
- In fact, most of the time the Bayesian estimates and the ML estimates give nearly identical values. This is the result of some important mathematics that guarantees (under quite general conditions) that the Bayesian estimates will converge to the same thing as the maximum likelihood estimates as long as your sample size is big enough (a CLT also applies).

Estimating regression model parameters: In summary

- Lots of ways to "best estimate" model parameters, e.g., OLS, MLE, REML, qML, Bayesian estimation.
- Weighted least squares (WLS) and robust estimation are two other common ones you may encounter (WLS in EPSE 682).
- Sadly, not enough time to talk about all of these.
- But remember: None of these techniques actually change the *(deterministic) functional form of the model*; instead, the techniques propose different "best estimates" for the same model parameters. In some situations, one technique produces estimates with "better"/more desirable properties than other techniques (e.g., REML gives unbiased estimates of variance components while regular ML does not).
- When in doubt, *ask a statistician*.