University of Waterloo W. H. Cherry

RESPONSE MODELS IN STAT 231: An Introductory Overview

This Statistical Highlight #71 summarizes the four common STAT 231 response models and ideas associated with them.

A useful general idea is: *statistical model = structure + noise*;
for us, this becomes specifically: *response = structural component + residual*;
elsewhere, the residual may be called the *error*, a useage of 'error' *entirely different* from that in STAT 231.

We use response models to quantify, for estimates of respondent population attributes, the uncertainty arising from sample error and measurement error; we do so by using these models to describe the outcome of *conceptual* repetition of the sampling and measuring processes without doing *actual* repetition. [The acronym EPS used below denotes *equiprobable selecting*.]

Question aspect, Plan:	**Descriptive**	**Causative, Blocked**	**Causative, *Un*blocked**
Model:	$Y_j = \mu + R_j$, $j = 1, 2, ..., n$, $R_j \sim G(0, \sigma)$, independent, EPS.	$Y_{ij} = \mu_i + \gamma_j + R_{ij}$, $i = 1, 2,$ $j = 1, 2, ..., n$, $R_{ij} \sim G(0, \sigma)$, independent, EPS; ⇩ $Y_j = \mu_d + R_j$, $j = 1, 2, ..., n$, $R_j \sim G(0, \sigma_d)$, independent, EPS.	$Y_{ij} = \mu_i + R_{ij}$, $i = 1, 2,$ $j = 1, 2, ..., n_i$, $R_{ij} \sim G(0, \sigma)$, independent, EPS.
Parameters:	μ, σ	μ_d, σ_d	$\mu_1 - \mu_2$, σ
Estimates:	$\hat{\mu} = \bar{y}$, $\hat{\sigma} = \sqrt{\frac{\sum_{j=1}^{n}(y_j - \bar{y})^2}{n-1}}$	$\hat{\mu}_d = \bar{y}$, $\hat{\sigma}_d = \sqrt{\frac{\sum_{j=1}^{n}(y_j - \bar{y})^2}{n-1}}$	$\hat{\mu}_1 - \hat{\mu}_2 = \bar{y}_1 - \bar{y}_2$, $\hat{\sigma} = \sqrt{\frac{(n_1-1)\hat{\sigma}_1^2 + (n_2-1)\hat{\sigma}_2^2}{n_1+n_2-2}}$
Estimators:	$\tilde{\mu} \sim G(\mu, \sigma\sqrt{\frac{1}{n}})$, $\frac{\tilde{\sigma}}{\sigma} \sim K_{n-1}$	$\tilde{\mu}_d \sim G(\mu_d, \sigma_d\sqrt{\frac{1}{n}})$, $\frac{\tilde{\sigma}_d}{\sigma_d} \sim K_{n-1}$	$\tilde{\mu}_1 - \tilde{\mu}_2 \sim G(\mu_1 - \mu_2, \sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}})$, $\frac{\tilde{\sigma}}{\sigma} \sim K_{n_1+n_2-2}$
CI for mean:	$\hat{\mu} \pm t^*_{n-1}\hat{\sigma}\sqrt{\frac{1}{n}}$	$\hat{\mu}_d \pm t^*_{n-1}\hat{\sigma}_d\sqrt{\frac{1}{n}}$	$\hat{\mu}_1 - \hat{\mu}_2 \pm t^*_{n_1+n_2-2}\hat{\sigma}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$
CI for σ:	$(\frac{\hat{\sigma}}{k_2^*}, \frac{\hat{\sigma}}{k_1^*})$	$(\frac{\hat{\sigma}_d}{k_2^*}, \frac{\hat{\sigma}_d}{k_1^*})$	$(\frac{\hat{\sigma}}{k_2^*}, \frac{\hat{\sigma}}{k_1^*})$
Discrepancy Measure:	$\frac{\tilde{\mu} - \mu_0}{\tilde{\sigma}\sqrt{\frac{1}{n}}} \sim t_{n-1}$	$\frac{\tilde{\mu}_d - \mu_{d0}}{\tilde{\sigma}_d\sqrt{\frac{1}{n}}} \sim t_{n-1}$	$\frac{(\tilde{\mu}_1 - \tilde{\mu}_2) - (\mu_1 - \mu_2)_0}{\tilde{\sigma}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2}$

NOTES: 1. Definitions of the symbols (see also Statistical Highlight #72) in the model for a Question with a descriptive aspect are:

- Y_j: a random variable whose distribution represents the possible values of the measured response variate **Y** for the jth unit in the sample of n units selected equiprobably from the respondent population of **N** units (usually $n \ll N$), if the selecting and measuring processes were to be repeated over and over.
- μ: a model parameter representing the *average* ($\bar{\mathbf{Y}}$) of the measured response variate of the **N** respondent population units.
- R_j: a random variable (called the *residual*) whose distribution represents the possible *differences*, from the structural component of the model, of the value of the measured response variate **Y** for the jth unit in the sample of n units selected equiprobably from the respondent population of **N** units, under repetition of the selecting and measuring processes. [*We* are usually concerned with the extra-structural variation due *only* to sampling and/or measuring.]
- σ: the (probabilistic) *standard deviation* of the Gaussian model for the distribution of the residual, is a model parameter representing the (data) standard deviation (**S**) of the measured response variate **Y** of the **N** respondent population units; this (data) standard deviation (and, hence, σ) *quantifies* the *variation* of **Y** over the **N** respondent population units – as this variation increases, so does **S** (and, hence, so does σ).

2. We are careful to distinguish:

- μ: a model parameter representing the *average* ($\bar{\mathbf{Y}}$) of the measured response variate of the **N** respondent population units.
- $\hat{\mu}$: the least squares *estimate* for μ – a *number* whose value is derived from an appropriate set of *data*;
- $\tilde{\mu}$: the least squares *estimator* for μ – a *random variable* whose distribution represents the possible values of the estimate $\hat{\mu}$ if the selecting, measuring and estimating processes were to be repeated over and over.

3. There are also noteworthy distinctions among *residuals*, as summarized by the four equations below.

- Equation HL71.1 is the response model for a Question with a descriptive aspect – it involves two *random variables* Y_j and R_j and the *constant* μ (whose *value* we want to determine). [We may also want to determine the value of σ in the probabilistic Gaussian (sub)model for R_j.]
- In equation HL71.2, y_j and r_j are numbers – *values* of the corresponding random

$$Y_j = \mu + R_j \quad \text{-----(HL71.1)}$$
$$y_j = \mu + r_j \quad \text{-----(HL71.2)}$$
$$y_j = \hat{\mu} + \hat{r}_j \quad \text{-----(HL71.3)}$$
$$Y_j = \tilde{\mu} + \tilde{r}_j \quad \text{-----(HL71.4)}$$

(*continued overleaf*)

University of Waterloo W. H. Cherry

NOTES: (cont.) 3. ● variables Y_j and R_j in equation HL71.1; r_j is the *realized* residual.
– Even though the value y_j is taken as the *known* y_j (see below), because μ is *un*known, so is the value r_j.
● Like equation HL71.2, the left-hand side of equation HL71.3 is a number – the (data) value y_j *observed* for the response variate of the jth unit selected for the sample; on the *right*-hand side, the value of the estimate $\hat{\mu}$ is *known* so the *estimated* residual $\hat{r}_j$ also has a *known* value.
– It is easy to think *mistakenly* that the estimated residual $\hat{r}_j$ is a value of the random variable R_j.
● Like equation HL71.1, equation HL71.4 has Y_j on its left-hand side; on its *right*-hand side are the *estimators* $\tilde{\mu}$ and $\tilde{r}_j$ so the terms in this equation are all random variables.
– Because $\tilde{r}_j = Y_j - \tilde{\mu}$ and the *co*variance of Y_j and $\tilde{\mu} \equiv \overline{Y}$ is σ^2/n, $\tilde{r}_j \sim G(0, \sigma\sqrt{1-\frac{1}{n}})$; with least squares estimating and a Gaussian model for Y_j, $\tilde{r}_j$ and $\tilde{\mu}$ are *independent* Gaussian random variables.

In expressions for estimates (like $\hat{\mu}$ and $\hat{\sigma}$), *data* values y_j are taken as *random variable* values y_j; it is investigators' responsibility to ensure that this assumed equivalence of y_j and y_j is reasonable under the Plan as executed.

4. Under the model overleaf on page HL71.1 for a Question with a descriptive aspect, a *prediction interval* (PI) for the **Y** value of an equiprobably-selected *individual* unit is: $\hat{\mu} \pm t^*_{n-1}\hat{\sigma}\sqrt{\frac{1}{n}+1}$.

5. The model for a Question with a descriptive aspect is also used for a *calibrated* measuring process, usually with τ replacing μ; for an *un*calibrated process, we use: $_M Y_j = \tau + \delta + R_j$, $j = 1, 2, ..., n$, $R_j \sim G(0, \sigma)$, independent, EPS. [or: $Y_j = \delta + R_j$, $(Y_j = {}_M Y_j - \tau)$]

6. In the model overleaf on page HL71.1 for a Question with a causative aspect answered using a blocked Plan, we work with *differences* with*in* the matched pairs (or blocks): $Y_j = Y_{1j} - Y_{2j}$, $\mu_d = \mu_1 - \mu_2$.

7. It is of interest to interpret the prediction interval (PI) from the regression model (the *last* expression on the left below) in terms of how its components contribute to quantifying the *overall* uncertainty in predicting $Y_{new}(x)$:
● t^*_{n-2} – determined by n and by the confidence level chosen (the higher the level, the *greater* the uncertainty);
● $\hat{\sigma}$ – the uncertainty from estimating σ, which quantifies variation about the straight-line structural component of the model due to uncontrolled, unmeasured and unknown explanatory variates;
● S_{xx} – the uncertainty from estimating the model parameter β_1;
● $(x-\overline{x})^2$ – the effect of how far x is from $\overline{x}$ (the further it is, the *greater* the uncertainty);
● 1/n – the uncertainty from estimating the centred model parameter α;
● 1 – the uncertainty from the *individual* whose response (value) is being predicted.

Regression (A model with one or more explanatory variates in its structural component.)

Model: $Y_j = \alpha + \beta_1(x_j - \overline{x}) + R_j$, $j = 1, 2, ..., n$, $R_j \sim G(0, \sigma)$, independent, EPS.

Parameters: $\alpha,\ \beta_1,\ \sigma,\ \beta_0 = \alpha - \beta_1\overline{x},\ \mu(x) = \alpha + \beta_1(x - \overline{x})$

Estimates: $\hat{\alpha} = \overline{y},\quad \hat{\beta}_1 = \dfrac{\sum_{j=1}^{n} y_j(x_j - \overline{x})}{\sum_{j=1}^{n}(x_j - \overline{x})^2} \equiv \dfrac{\sum_{j=1}^{n}(y_j - \overline{y})(x_j - \overline{x})}{S_{xx}}$,

$\hat{\sigma} = \sqrt{\dfrac{\sum_{j=1}^{n}[y_j - \overline{y} - \hat{\beta}_1(x_j - \overline{x})]^2}{n-2}}$,

$\hat{\beta}_0 = \hat{\alpha} - \hat{\beta}_1\overline{x},\quad \hat{\mu}(x) = \hat{\alpha} + \hat{\beta}_1(x - \overline{x})$

Estimators: $\tilde{\alpha} \sim G(\alpha, \sigma\sqrt{\frac{1}{n}}),\ \tilde{\beta}_1 \sim G(\beta_1, \sigma\sqrt{\frac{1}{S_{xx}}}),\ \frac{\tilde{\sigma}}{\sigma} \sim K_{n-2}$

$\tilde{\beta}_0 = \tilde{\alpha} - \tilde{\beta}_1\overline{x},\quad \tilde{\mu}(x) = \tilde{\alpha} + \tilde{\beta}_1(x - \overline{x})$

CIs for β_1, σ: $\hat{\beta}_1 \pm t^*_{n-2}\hat{\sigma}\sqrt{\frac{1}{S_{xx}}},\quad (\frac{\hat{\sigma}}{k^*_2},\ \frac{\hat{\sigma}}{k^*_1})$

CI for $\mu(x)$: $\hat{\mu}(x) \pm t^*_{n-2}\hat{\sigma}\sqrt{\frac{(x-\overline{x})^2}{S_{xx}} + \frac{1}{n}}$

Discrepancy Measures: $\dfrac{\tilde{\beta}_1 - \beta_1}{\tilde{\sigma}\sqrt{\frac{1}{S_{xx}}}} \sim t_{n-2},\quad \dfrac{\tilde{\mu}(x) - \mu(x)_0}{\tilde{\sigma}\sqrt{\frac{(x-\overline{x})^2}{S_{xx}} + \frac{1}{n}}} \sim t_{n-2}$

PI for $Y_{new}(x)$: $\hat{\mu}(x) \pm t^*_{n-2}\hat{\sigma}\sqrt{\frac{(x-\overline{x})^2}{S_{xx}} + \frac{1}{n} + 1}$

General Response Model

Model: $Y_j = \mu_j + R_j$, $j = 1, 2, ..., n$, $R_j \sim G(0, \sigma)$, independent, EPS;
or: $\underset{\sim}{Y} = \mathbf{X}\underset{\sim}{\Theta} + \underset{\sim}{R}$

where: $\mu_j = \theta_1 x_{j1} + ... + \theta_k x_{jk} + ... + \theta_q x_{jq}$
[explanatory variates $X_1, ..., X_k, ..., X_q$]

Parameters: $\underset{\sim}{\Theta},\ \sigma$ [Θ is a vector of length q; $\underset{\sim}{Y}$ and $\underset{\sim}{R}$ are of length $n \gg q$]

Estimates: $\hat{\underset{\sim}{\Theta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\underset{\sim}{y} = \mathbf{A}\underset{\sim}{y},\quad \hat{\theta}_k = \sum_{j=1}^{n} a_{jk} y_j$,

[and: $\underset{\sim}{\Theta} = \mathbf{A}\underset{\sim}{\mu},\quad \theta_k = \sum_{j=1}^{n} a_{jk}\mu_j$] $\hat{\sigma} = \sqrt{\dfrac{\sum_{j=1}^{n}\hat{r}_j^2}{n-q}}$

Estimators: $\tilde{\underset{\sim}{\Theta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\underset{\sim}{Y} = \mathbf{A}\underset{\sim}{Y} = \underset{\sim}{\Theta} + \mathbf{A}\underset{\sim}{R}$,

$\tilde{\theta}_k = \sum_{j=1}^{n} a_{jk} Y_j \sim G(\theta_k, \sigma\sqrt{\sum_{j=1}^{n} a_{jk}^2})$,

$\tilde{\sigma} = \sqrt{\dfrac{\sum_{j=1}^{n}\tilde{r}_j^2}{n-q}},\quad \dfrac{\tilde{\sigma}}{\sigma} \sim K_{n-q}$

CIs for θ_k, σ: $\hat{\theta}_k \pm t^*_{n-q}\hat{\sigma}\sqrt{\sum_{j=1}^{n} a_{jk}^2},\quad (\frac{\hat{\sigma}}{k^*_2},\ \frac{\hat{\sigma}}{k^*_1})$

Discrepancy Measure: $\dfrac{\tilde{\theta}_k - \theta_{k0}}{\tilde{\sigma}\sqrt{\sum_{j=1}^{n} a_{jk}^2}} \sim t_{n-q}$

PI for Y_{new}: $E(\hat{Y}_{new}) \pm t^*_{n-q}\hat{\sigma}\sqrt{\sum_{j=1}^{n} a_{jk}^2 + 1}$