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:	$Y_j = \mu + R_j$, $j = 1, 2, ..., \mathrm{n}$, $R_j \sim G(0, \sigma)$, independent, EPS.	$Y_{ij} = \mu_i + \gamma_j + R_{ij}$, $i = 1, 2$, $j = 1, 2, ..., \mathrm{n}$, $R_{ij} \sim G(0, \sigma)$, independent, EPS; $\Downarrow$ $Y_j = \mu_d + R_j$, $j = 1, 2, ..., \mathrm{n}$, $R_j \sim G(0, \sigma_d)$, independent, EPS.	$Y_{ij} = \mu_i + R_{ij}$, $i = 1, 2$, $j = 1, 2, ..., \mathrm{n}_i$, $R_{ij} \sim G(0, \sigma)$, independent, EPS.
Parameters:	μ, σ	μ_d, σ_d	$\mu_1 - \mu_2$, σ
Estimates:	$\hat{\mu} = \bar{y}$, $\hat{\sigma} = \sqrt{\dfrac{\sum_{j=1}^{n}(y_j - \bar{y})^2}{\mathrm{n}-1}}$	$\hat{\mu}_d = \bar{y}$, $\hat{\sigma}_d = \sqrt{\dfrac{\sum_{j=1}^{n}(y_j - \bar{y})^2}{\mathrm{n}-1}}$	$\hat{\mu}_1 - \hat{\mu}_2 = \bar{y}_1 - \bar{y}_2$, $\hat{\sigma} = \sqrt{\dfrac{(\mathrm{n}_1 - 1)\hat{\sigma}_1^2 + (\mathrm{n}_2 - 1)\hat{\sigma}_2^2}{\mathrm{n}_1 + \mathrm{n}_2 - 2}}$
Estimators:	$\tilde{\mu} \sim G(\mu, \sigma\sqrt{\tfrac{1}{\mathrm{n}}})$, $\dfrac{\tilde{\sigma}}{\sigma} \sim K_{\mathrm{n}-1}$	$\tilde{\mu}_d \sim G(\mu_d, \sigma_d\sqrt{\tfrac{1}{\mathrm{n}}})$, $\dfrac{\tilde{\sigma}_d}{\sigma_d} \sim K_{\mathrm{n}-1}$	$\tilde{\mu}_1 - \tilde{\mu}_2 \sim G(\mu_1 - \mu_2, \sigma\sqrt{\tfrac{1}{\mathrm{n}_1} + \tfrac{1}{\mathrm{n}_2}})$, $\dfrac{\tilde{\sigma}}{\sigma} \sim K_{\mathrm{n}_1 + \mathrm{n}_2 - 2}$
CI for mean:	$\hat{\mu} \pm t^*_{\mathrm{n}-1}\hat{\sigma}\sqrt{\tfrac{1}{\mathrm{n}}}$	$\hat{\mu}_d \pm t^*_{\mathrm{n}-1}\hat{\sigma}_d\sqrt{\tfrac{1}{\mathrm{n}}}$	$\hat{\mu}_1 - \hat{\mu}_2 \pm t^*_{\mathrm{n}_1 + \mathrm{n}_2 - 2}\hat{\sigma}\sqrt{\tfrac{1}{\mathrm{n}_1} + \tfrac{1}{\mathrm{n}_2}}$
CI for $\boldsymbol{\sigma}$:	$(\dfrac{\hat{\sigma}}{k_2^*}, \dfrac{\hat{\sigma}}{k_1^*})$	$(\dfrac{\hat{\sigma}_d}{k_2^*}, \dfrac{\hat{\sigma}_d}{k_1^*})$	$(\dfrac{\hat{\sigma}}{k_2^*}, \dfrac{\hat{\sigma}}{k_1^*})$
Discrepancy Measure:	$\dfrac{\tilde{\mu} - \mu_0}{\tilde{\sigma}\sqrt{\tfrac{1}{\mathrm{n}}}} \sim t_{\mathrm{n}-1}$	$\dfrac{\tilde{\mu}_d - \mu_{d0}}{\tilde{\sigma}_d\sqrt{\tfrac{1}{\mathrm{n}}}} \sim t_{\mathrm{n}-1}$	$\dfrac{(\tilde{\mu}_1 - \tilde{\mu}_2) - (\mu_1 - \mu_2)_0}{\tilde{\sigma}\sqrt{\tfrac{1}{\mathrm{n}_1} + \tfrac{1}{\mathrm{n}_2}}} \sim t_{\mathrm{n}_1 + \mathrm{n}_2 - 2}$

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 ≪ **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.

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

- 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

(*continued overleaf*)

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* $\mathrm{y_j}$ (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 $\mathrm{y_j}$ *observed* for the response variate of the *j*th unit selected for the sample; on the *right*-hand side, the value of the estimate $\hat{\mu}$ is *known* so the *estimated* residual $\hat{\mathrm{r}}_j$ also has a *known* value.
   – It is easy to think *mistakenly* that the estimated residual $\hat{\mathrm{r}}_j$ 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, $\tilde{r}_j \sim G(0, \sigma\sqrt{1-\frac{1}{\mathrm{n}}})$; 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 $\mathrm{y_j}$ are taken as *random variable* values y_j; it is investigators' responsibility to ensure that this assumed equivalence of $\mathrm{y_j}$ 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: $\hat{\mu} \pm t^*_{\mathrm{n}-1}\hat{\sigma}\sqrt{\frac{1}{\mathrm{n}}+1}$.

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:
   $${}_{\mathrm{M}}Y_j = \tau + \delta + R_j, \quad j = 1, 2, \ldots, \mathrm{n}, \quad R_j \sim G(0, \sigma), \text{ independent, EPS.} \quad [\text{or: } Y_j = \delta + R_j, \ (Y_j = {}_{\mathrm{M}}Y_j - \tau)]$$

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): $Y_j = Y_{1j} - Y_{2j}, \quad \mu_{\mathrm{d}} = \mu_1 - \mu_2$.

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 $Y_{\mathrm{new}}(\mathrm{x})$:
   - $t^*_{\mathrm{n}-2}$ – 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;
   - $\mathrm{S_{xx}}$ – the uncertainty from estimating the model parameter β_1;
   - $(\mathrm{x}-\overline{\mathrm{x}})^2$ – the effect of how far x is from $\overline{\mathrm{x}}$ (the further it is, the *greater* the uncertainty);
   - $1/\mathrm{n}$ – 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: $Y_j = \alpha + \beta_1(\mathrm{x}_j - \overline{\mathrm{x}}) + R_j, \quad j = 1, 2, \ldots, \mathrm{n}, \quad R_j \sim G(0, \sigma)$, independent, EPS.

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

Estimates:
$$\hat{\alpha} = \overline{y}, \quad \hat{\beta}_1 = \frac{\sum_{j=1}^{\mathrm{n}} y_j(\mathrm{x}_j - \overline{\mathrm{x}})}{\sum_{j=1}^{\mathrm{n}} (\mathrm{x}_j - \overline{\mathrm{x}})^2} \equiv \frac{\sum_{j=1}^{\mathrm{n}} (y_j - \overline{y})(\mathrm{x}_j - \overline{\mathrm{x}})}{\mathrm{S_{xx}}},$$
$$\hat{\sigma} = \sqrt{\frac{\sum_{j=1}^{\mathrm{n}} [y_j - \overline{y} - \hat{\beta}_1(\mathrm{x}_j - \overline{\mathrm{x}})]^2}{\mathrm{n}-2}},$$
$$\hat{\beta}_0 = \hat{\alpha} - \hat{\beta}_1\overline{\mathrm{x}}, \quad \hat{\mu}(\mathrm{x}) = \hat{\alpha} + \hat{\beta}_1(\mathrm{x}-\overline{\mathrm{x}})$$

Estimators:
$$\tilde{\alpha} \sim G(\alpha, \sigma\sqrt{\tfrac{1}{\mathrm{n}}}), \quad \tilde{\beta}_1 \sim G(\beta_1, \sigma\sqrt{\tfrac{1}{\mathrm{S_{xx}}}}), \quad \frac{\tilde{\sigma}}{\sigma} \sim K_{\mathrm{n}-2}$$
$$\tilde{\beta}_0 = \tilde{\alpha} - \tilde{\beta}_1\overline{\mathrm{x}}, \quad \tilde{\mu}(\mathrm{x}) = \tilde{\alpha} + \tilde{\beta}_1(\mathrm{x}-\overline{\mathrm{x}})$$

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

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

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

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

General Response Model

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

where: $\mu_j = \theta_1\mathrm{x}_{j1} + \ldots + \theta_{\mathrm{k}}\mathrm{x}_{j\mathrm{k}} + \ldots + \theta_{\mathrm{q}}\mathrm{x}_{j\mathrm{q}}$
[explanatory variates $\mathrm{X}_1, \ldots, \mathrm{X}_{\mathrm{k}}, \ldots, \mathrm{X}_{\mathrm{q}}$]

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

Estimates:
$$\hat{\underset{\sim}{\Theta}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\underset{\sim}{\boldsymbol{y}} = \mathbf{A}\underset{\sim}{\boldsymbol{y}}, \quad \hat{\theta}_{\mathrm{k}} = \sum_{j=1}^{\mathrm{n}} \mathrm{a}_{j\mathrm{k}} y_j,$$
[and: $\underset{\sim}{\Theta} = \mathbf{A}\underset{\sim}{\boldsymbol{\mu}}, \quad \theta_{\mathrm{k}} = \sum_{j=1}^{\mathrm{n}} \mathrm{a}_{j\mathrm{k}}\mu_j$]
$$\hat{\sigma} = \sqrt{\frac{\sum_{j=1}^{\mathrm{n}} \hat{r}_j^2}{\mathrm{n}-\mathrm{q}}}$$

Estimators:
$$\tilde{\underset{\sim}{\Theta}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\underset{\sim}{\boldsymbol{Y}} = \mathbf{A}\underset{\sim}{\boldsymbol{Y}} = \underset{\sim}{\Theta} + \mathbf{A}\underset{\sim}{\boldsymbol{R}},$$
$$\tilde{\theta}_{\mathrm{k}} = \sum_{j=1}^{\mathrm{n}} \mathrm{a}_{j\mathrm{k}} Y_j \sim G(\theta_{\mathrm{k}}, \sigma\sqrt{\textstyle\sum_{j=1}^{\mathrm{n}} \mathrm{a}_{j\mathrm{k}}^2}),$$
$$\tilde{\sigma} = \sqrt{\frac{\sum_{j=1}^{\mathrm{n}} \tilde{r}_j^2}{\mathrm{n}-\mathrm{q}}}, \quad \frac{\tilde{\sigma}}{\sigma} \sim K_{\mathrm{n}-\mathrm{q}}$$

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

Discrepancy Measure:
$$\frac{\tilde{\theta}_{\mathrm{k}} - \theta_{\mathrm{k}0}}{\tilde{\sigma}\sqrt{\sum_{j=1}^{\mathrm{n}} \mathrm{a}_{j\mathrm{k}}^2}} \sim t_{\mathrm{n}-\mathrm{q}}$$

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