

Figure 7.1. RESPONSE MODELS

This Figure 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 usage of ‘error’ *different* from that in STAT 231.

We use response models to quantify, for estimates of study 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, Unblocked
Model:	$Y_j = \mu + R_j, \quad j=1, 2, \dots, n,$ $R_j \sim G(0, \sigma),$ independent, EPS.	$Y_{ij} = \mu_i + \gamma_j + R_{ij}, \quad i=1, 2,$ $j=1, 2, \dots, n,$ $R_{ij} \sim G(0, \sigma),$ independent, EPS; $\Downarrow$ $Y_j = \mu_d + R_j, \quad j=1, 2, \dots, n,$ $R_j \sim G(0, \sigma_d),$ independent, EPS.	$Y_{ij} = \mu_i + R_{ij}, \quad i=1, 2,$ $j=1, 2, \dots, n_i,$ $R_{ij} \sim G(0, \sigma),$ independent, EPS.
Parameters:	μ, σ	μ_d, σ_d	$\mu_1 - \mu_2, \sigma$
Estimates:	$\hat{\mu} = \bar{y}, \quad \hat{\sigma} = \sqrt{\frac{\sum_{j=1}^n (y_j - \bar{y})^2}{n-1}}$	$\hat{\mu}_d = \bar{y}, \quad \hat{\sigma}_d = \sqrt{\frac{\sum_{j=1}^n (y_j - \bar{y})^2}{n-1}}$	$\hat{\mu}_1 - \hat{\mu}_2 = \bar{y}_1 - \bar{y}_2,$ $\hat{\sigma} = \sqrt{\frac{(n_1-1)\hat{\sigma}_1^2 + (n_2-1)\hat{\sigma}_2^2}{n_1 + n_2 - 2}}$
Estimators:	$\tilde{\mu} \sim G(\mu, \sigma\sqrt{\frac{1}{n}}), \quad \tilde{\sigma} \sim K_{n-1}$	$\tilde{\mu}_d \sim G(\mu_d, \sigma_d\sqrt{\frac{1}{n}}), \quad \tilde{\sigma}_d \sim K_{n-1}$	$\tilde{\mu}_1 - \tilde{\mu}_2 \sim G(\mu_1 - \mu_2, \sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}),$ $\tilde{\sigma} \sim K_{n_1 + n_2 - 2}$
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 Figure 5.9) 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 $\mathbf{Y}$ for the j th unit in the sample of n units selected equiprobably from the respondent population of $\mathbf{N}$ units (usually $n \ll \mathbf{N}$), 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 $\mathbf{N}$ 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 $\mathbf{Y}$ for the j th unit in the sample of n units selected equiprobably from the respondent population of $\mathbf{N}$ 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 ($\mathbf{S}$) of the measured response variate $\mathbf{Y}$ of the $\mathbf{N}$ respondent population units; this (data) standard deviation (and, hence, σ) *quantifies* the *variation* of $\mathbf{Y}$ over the $\mathbf{N}$ respondent population units – as this variation increases, so does $\mathbf{S}$ (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 $\mathbf{N}$ 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 7.1 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 7.2, y_j and r_j are numbers – *values* of the corresponding random vari-

$$Y_j = \mu + R_j \quad \text{-----(7.1)}$$

$$y_j = \mu + r_j \quad \text{-----(7.2)}$$

$$y_j = \hat{\mu} + \hat{r}_j \quad \text{-----(7.3)}$$

$$Y_j = \tilde{\mu} + \tilde{r}_j \quad \text{-----(7.4)}$$

NOTES: 3. • ables Y_j and R_j in equation 7.1; r_j is the *realized* residual.

(cont.)

- Even though the value y_j is taken as the *known* y_j (see below), because μ is *unknown*, so is the value r_j .
- Like equation 7.2, the left-hand side of equation 7.3 is a number – the (data) value 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 $\hat{r}_j$ 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 7.1, equation 7.4 has Y_j on its left-hand side; on its *right*-hand side are the *estimators* $\hat{\mu}$ and $\hat{r}_j$ so the terms in this equation are all random variables.
- Because $\tilde{r}_j = Y_j - \hat{\mu}$ and the covariance of Y_j and $\hat{\mu} \equiv \bar{Y}$ 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 $\hat{\mu}$ 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.

- Under the model overleaf on page 7.1 for a Question with a descriptive aspect, a *prediction interval* (PI) for the Y value of an equiprobably-selected *individual* unit is: $\hat{\mu} \pm t_{n-1}^* \hat{\sigma} \sqrt{\frac{1}{n} + 1}$.
- The model for a Question with a descriptive aspect is also used for a *calibrated* measuring process, usually with τ replacing μ ; for an *uncalibrated* process, we use: ${}_MY_j = \tau + \delta + R_j$, $j=1, 2, \dots, n$, [or: $Y_j = \delta + R_j$, $R_j \sim G(0, \sigma)$, ($Y_j = {}_MY_j - \tau$)] independent, EPS.
- In the model overleaf on page 7.1 for a Question with a causative aspect answered using a blocked Plan, we work with *differences within* the matched pairs (or blocks): $Y_j = Y_{1j} - Y_{2j}$, $\mu_d = \mu_1 - \mu_2$.
- It is of interest to interpret the prediction interval 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 $Y_{\text{new}}(x)$:
 - 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 unknown or uncontrolled explanatory variates;
 - S_{xx} – the uncertainty from estimating the model parameter β_1 ;
 - $(x - \bar{x})^2$ – the effect of how far x is from $\bar{x}$ (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:	$Y_j = \alpha + \beta_1(x_j - \bar{x}) + R_j$, $j=1, 2, \dots, n$, $R_j \sim G(0, \sigma)$, independent, EPS.
Parameters:	$\alpha, \beta_1, \sigma, \beta_0 = \alpha - \beta_1\bar{x}, \mu(x) = \alpha + \beta_1(x - \bar{x})$
Estimates:	$\hat{\alpha} = \bar{y}, \hat{\beta}_1 = \frac{\sum_{j=1}^n y_j(x_j - \bar{x})}{\sum_{j=1}^n (x_j - \bar{x})^2} \equiv \frac{\sum_{j=1}^n (y_j - \bar{y})(x_j - \bar{x})}{S_{xx}},$ $\hat{\sigma} = \sqrt{\frac{\sum_{j=1}^n [y_j - \bar{y} - \hat{\beta}_1(x_j - \bar{x})]^2}{n-2}},$ $\hat{\beta}_0 = \hat{\alpha} - \hat{\beta}_1\bar{x}, \hat{\mu}(x) = \hat{\alpha} + \hat{\beta}_1(x - \bar{x})$
Estimators:	$\tilde{\alpha} \sim G(\alpha, \sigma\sqrt{\frac{1}{n}}), \tilde{\beta}_1 \sim G(\beta_1, \sigma\sqrt{\frac{1}{S_{xx}}}), \tilde{\sigma} \sim K_{n-2}$ $\tilde{\beta}_0 = \tilde{\alpha} - \tilde{\beta}_1\bar{x}, \tilde{\mu}(x) = \tilde{\alpha} + \tilde{\beta}_1(x - \bar{x})$
CIs for β_1, σ:	$\hat{\beta}_1 \pm t_{n-2}^* \hat{\sigma} \sqrt{\frac{1}{S_{xx}}}, \left(\frac{\hat{\sigma}}{k_2^*}, \frac{\hat{\sigma}}{k_1^*} \right)$
CI for $\mu(x)$:	$\hat{\mu}(x) \pm t_{n-2}^* \hat{\sigma} \sqrt{\frac{(x - \bar{x})^2}{S_{xx}} + \frac{1}{n}}$
Discrepancy Measures:	$\frac{\tilde{\beta}_1 - \beta_1}{\hat{\sigma} \sqrt{\frac{1}{S_{xx}}}} \sim t_{n-2}, \frac{\tilde{\mu}(x) - \mu(x)_0}{\hat{\sigma} \sqrt{\frac{(x - \bar{x})^2}{S_{xx}} + \frac{1}{n}}} \sim t_{n-2}$
PI for $Y_{\text{new}}(x)$:	$\hat{\mu}(x) \pm t_{n-2}^* \hat{\sigma} \sqrt{\frac{(x - \bar{x})^2}{S_{xx}} + \frac{1}{n} + 1}$

General Response Model

Model:	$Y_j = \mu_j + R_j$, $j=1, 2, \dots, n$, $R_j \sim G(0, \sigma)$, independent, EPS;
or:	$\underline{Y} = \underline{X}\underline{\Theta} + \underline{R}$
where:	$\mu_j = \theta_1 x_{j1} + \dots + \theta_k x_{jk} + \dots + \theta_q x_{jq}$ [explanatory variates $X_1, \dots, X_k, \dots, X_q$]
Parameters:	$\underline{\Theta}, \sigma$ [$\underline{\Theta}$ is a vector of length q ; $\underline{Y}$ and $\underline{R}$ are of length $n \gg q$]
Estimates:	$\hat{\underline{\Theta}} = (\underline{X}^T \underline{X})^{-1} \underline{X}^T \underline{Y} = \underline{A} \underline{y}, \hat{\theta}_k = \sum_{j=1}^n a_{jk} y_j,$ [and: $\underline{\Theta} = \underline{A} \underline{\mu}, \theta_k = \sum_{j=1}^n a_{jk} \mu_j]$ $\hat{\sigma} = \sqrt{\frac{\sum_{j=1}^n \hat{r}_j^2}{n-q}}$
Estimators:	$\tilde{\underline{\Theta}} = (\underline{X}^T \underline{X})^{-1} \underline{X}^T \underline{Y} = \underline{A} \underline{y} = \underline{\Theta} + \underline{A} \underline{R},$ $\tilde{\theta}_k = \sum_{j=1}^n a_{jk} Y_j \sim G(\theta_k, \sigma \sqrt{\sum_{j=1}^n a_{jk}^2}),$ $\tilde{\sigma} = \sqrt{\frac{\sum_{j=1}^n \tilde{r}_j^2}{n-q}}, \tilde{\sigma} \sim K_{n-q}$
CIs for θ_k, σ:	$\hat{\theta}_k \pm t_{n-q}^* \hat{\sigma} \sqrt{\sum_{j=1}^n a_{jk}^2}, \left(\frac{\hat{\sigma}}{k_2^*}, \frac{\hat{\sigma}}{k_1^*} \right)$
Discrepancy Measure:	$\frac{\tilde{\theta}_k - \theta_{k0}}{\hat{\sigma} \sqrt{\sum_{j=1}^n a_{jk}^2}} \sim t_{n-q}$
PI for Y_{new}:	$E(\hat{Y}_{\text{new}}) \pm t_{n-q}^* \hat{\sigma} \sqrt{\sum_{j=1}^n a_{jk}^2 + 1}$