9.6 Transformation of response

- possible remedy means, especially in situations when heteroscedasticity and/or non-normality detected

heteroscedasticity / non-normality

- serious problem especially when prediction is the main aim of modelling

---

often $Y|X \sim N_n(X\beta, \sigma^2 I_n)$

but for (Y cannot be assumed)

$Y^* = (t(Y_1), \dots, t(Y_n))^T$ model

$Y^*|X \sim N_n(X\beta, \sigma^2 I_n)$ is OK

$t: \mathbb{R} \to \mathbb{R}$, chosen (non-linear) transformat.

→ normal linear model for transformed response

WARNING: model for transformed response is not always suitable, especially when effect of covariates on response expectation is of interest

$$E(t(Y)|X=x) \neq t(E(Y|X=x))$$

(but see an exception of log-normal LogN)

9.6.1 Prediction based on a model with transformed response

If prediction is the main aim of modelling then no problems at all.

AIM: predict Y_{new}, given $X_{new} = x_{new}$

ASSUME: t is strictly increasing $t(Y_i)$

MODEL: $M: \underline{Y}^* | \mathbb{X} \sim N_n(\mathbb{X}\beta, \sigma^2 I)$ $\quad Y_i^* = X_i^T\beta + \varepsilon_i$,

$\underline{Y}^* = (t(Y_1), \ldots, t(Y_n))^T$ $\quad \varepsilon_i | \mathbb{X} \sim N(0, \sigma^2)$

$Y_{new}^* = t(Y_{new})$

1. $\hat{Y}_{new}^*$ and $(\hat{Y}_{new}^{*L}, \hat{Y}_{new}^{*U})$:

point and interval prediction for Y_{new}^* based on model M, i.e.

$$1-\alpha = P\left((\hat{Y}_{new}^{*L}, \hat{Y}_{new}^{*U}) \ni Y_{new}^*; \text{ parm. values}\right)$$

$$= P\left((\hat{Y}_{new}^{*L}, \hat{Y}_{new}^{*U}) \ni t(Y_{new}); \text{ parm. values}\right)$$

2.
$$= P\left(\left(\underbrace{t^{-1}(\hat{Y}_{new}^{*L})}_{\hat{Y}_{new}^{L}}, \underbrace{t^{-1}(\hat{Y}_{new}^{*U})}_{\hat{Y}_{new}^{U}}\right) \ni Y_{new}; \text{ parm. values}\right)$$

3. $\hat{Y}_{new} := t^{-1}(\hat{Y}_{new}^*)$: point prediction

(not necessarily in the middle of the prediction interval but who cares)

9.6.2 Log-normal model

→ Reasonable interpretation of regression coefficients can still be derived

Let the following (log-normal linear) model hold:

$$\log(Y_i) = X_i^T \beta + \varepsilon_i \quad , i = 1, \dots, n$$

$$\varepsilon_i | X \overset{\text{indep}}{\sim} N(0, \sigma^2)$$

≡ normal linear model for log-transformed response

⇒ multiplicative model for the original response

$$Y_i = \exp(X_i^T \beta) \cdot \eta_i \quad , i = 1, \dots, n$$

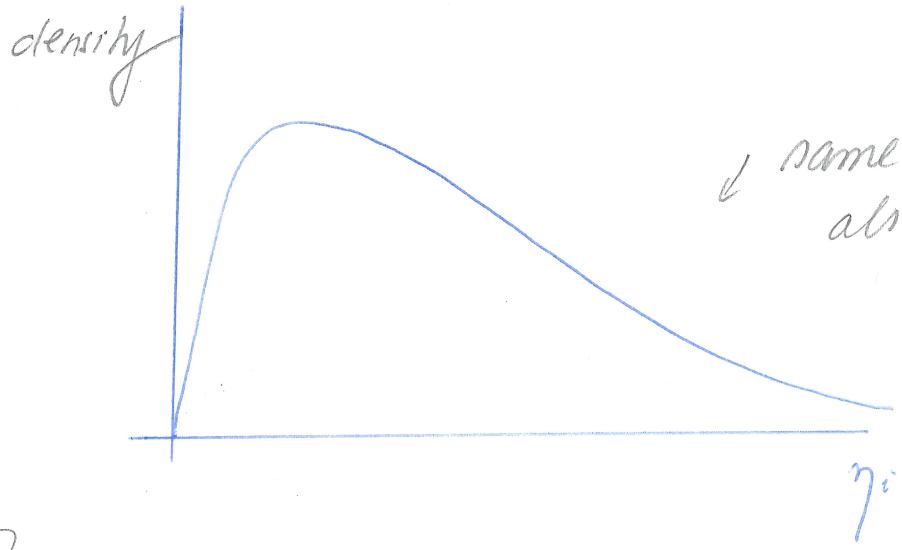
$$\eta_i = \exp(\varepsilon_i), \quad \eta_i | X \overset{\text{indep}}{\sim} LN(0, \sigma^2)$$

log-normal distrib.

Moments of the log-normal distributions

$$M_i := E(\eta_i) = E(\eta_i | X) = \exp\left(\frac{\sigma^2}{2}\right) > 1 \ (\text{if } \sigma^2 > 0)$$

$$V_i := \text{var}(\eta_i) = \text{var}(\eta_i | X) = \{\exp(\sigma^2) - 1\} \cdot \exp(\sigma^2)$$



↓ same shape has also assumed distribut. of Y_i

MODEL: $Y_i = \exp(X_i^T\beta) \cdot \eta_i$

$$E(Y|X=x) = \exp(x^T\beta) \cdot E(\eta) = \exp(x^T\beta) \cdot \underbrace{\exp\left(\frac{\sigma^2}{2}\right)}_{M>1}$$

$$\mathrm{var}(Y|X=x) = \exp(2x^T\beta) \cdot V = V \cdot \left(\frac{E(Y|X=x)}{M}\right)^2$$

$\mathrm{var}(Y|X=x)$ increases with $|E(Y|X=x)|$

log-normal linear model captures situations when

1. response (conditional) distribution is skewed
2. response (conditional) variance increases with the (absolute value) of expectation

Interpretation of regression coefficients in case of log-normal linear model

MODEL: $\log(Y) = X^T\beta + \varepsilon$, $\varepsilon \sim N(0, \sigma^2)$

$$E(Y \mid X = x) = \exp(x^T\beta) \cdot M, \quad M = \exp\left(\frac{\sigma^2}{2}\right)$$

$$x = (x_0, \dots, x_j, \dots, x_{k-1})^T \in \mathcal{X}$$

$$x^{j(+1)} = (x_0, \dots, x_{j+1}, \dots, x_{k-1})^T \in \mathcal{X}$$

$$\beta = (\beta_0, \dots, \beta_j, \dots, \beta_{k-1})^T$$

$$\Rightarrow \quad \frac{E(Y \mid X = x^{j(+1)})}{E(Y \mid X = x)} = \frac{M \exp(x^{j(+1)T}\beta)}{M \exp(x^T\beta)} = e^{\beta_j}$$

→ in output, report $e^{\hat{\beta}_j}$ rather than $\hat{\beta}_j$

→ analogously, if $(\hat{\beta}_j^L, \hat{\beta}_j^U)$ is confidence interval for β_j then $(e^{\hat{\beta}_j^L}, e^{\hat{\beta}_j^U})$ is confidence interval for e^{β_j} (and should be reported)

Special case, one-way classification & log-normal linear model

- one categorical covariate parameterized by chosen (pseudo) contrasts $\mathbb{C} = \begin{pmatrix} \mathbb{C}_1^T \\ \vdots \\ \mathbb{C}_G^T \end{pmatrix}$

MODEL:

$$E(\log(Y) \mid Z = z) = \beta_0 + \mathbb{C}_g^T \beta^Z$$

$$\Rightarrow \frac{E(Y \mid Z=g)}{E(Y \mid Z=h)} = \frac{M \cdot \exp(\beta_0 + \mathbb{C}_g^T \beta^Z)}{M \cdot \exp(\beta_0 + \mathbb{C}_h^T \beta^Z)} =$$

$$= \exp\{(\mathbb{C}_g^T - \mathbb{C}_h^T)\beta^Z\}$$

For example: reference group (pseudo) contrasts (dummy variables)

$$\mathbb{C} = \begin{pmatrix} 0 & \cdots & 0 \\ 1 & \cdots & 0 \\ & \ddots & \\ 0 & \cdots & 1 \end{pmatrix}$$

$$E(\log(Y) \mid Z=g) = \beta_0 + \beta_g^Z, \quad \beta_1^Z = 0$$

$$\Rightarrow \frac{E(Y \mid Z=g)}{E(Y \mid Z=1)} = \exp(\beta_g^Z), \quad g = 2, \dots, G$$

(1 ← reference)

$$\frac{E(Y \mid Z=g)}{E(Y \mid Z=h)} = \exp(\beta_g^Z - \beta_h^Z), \quad g, h = 2, \dots, G$$