Multiple Regression Model

Model: $y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u$;

Population regression function: $E(y|x_1, \ldots, x_k) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$;

OLS regression line: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k$;

Sampling variance:

$$Var(\hat{\beta}_j) = \frac{\sigma^2}{SST_j(1 - R_j^2)}, j = 1, \ldots, k,$$

where $SST_j = \sum_{i=1}^n (x_{ij} - \bar{x}_j)^2$, and R_j^2 is the R-squared from regressing x_j on all other independent variables;

Estimation of σ^2:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^n \hat{u}_i^2}{n - k - 1}$$

Sums of Squares: $SST = \sum_{i=1}^n (y_i - \bar{y})^2$, $SSE = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$, $SSR = \sum_{i=1}^n \hat{u}_i^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2$;

R-squared:

$$R^2 = \frac{SSE}{SST} = 1 - \frac{SSR}{SST}$$

Adjusted R-squared:

$$\bar{R}^2 = 1 - \frac{SSR/(n-k-1)}{SST/(n-1)}$$

Omitted variable bias: the true model is $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$, and we estimate the model without variable x_2, which gives $\tilde{y} = \tilde{\beta}_0 + \tilde{\beta}_1 x_1$. Then $Bias(\tilde{\beta}_1) = \beta_2 \tilde{\delta}_1$, where $\tilde{\delta}_1$ comes from the regression $x_2 = \tilde{\delta}_0 + \tilde{\delta}_1 x_1$;

AIC: $n \ln(SSR/n) + 2k$; BIC: $n \ln(SSR/n) + k \ln(n)$; AICc: AIC $+ \frac{2(k+2)(k+3)}{n-k-3}$;

VIF$_j = \frac{1}{1-R_j^2}$, where R_j^2 is the R-squared from regressing x_j on all other independent variables;

For a model $y = \beta_0 + \beta_1 x + \beta_2 x^2 + u$, $\Delta\hat{y} \approx (\hat{\beta}_1 + 2\hat{\beta}_2 x)\Delta x$;

Predicting y in a log model: $\hat{y} = n^{-1} \sum_{i=1}^n \exp(\hat{u}_i) \exp(\widehat{logy})$;

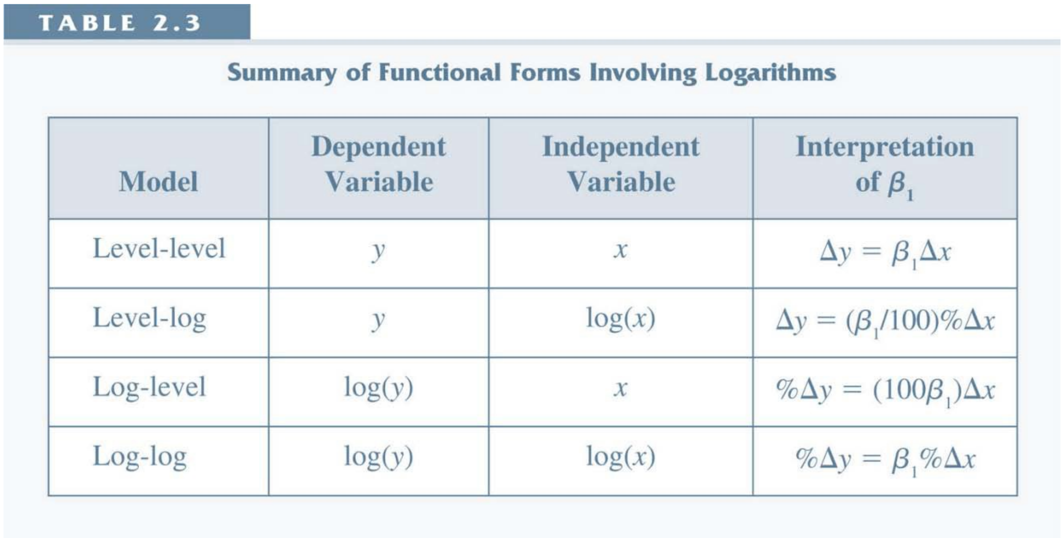

TABLE 2.3

Summary of Functional Forms Involving Logarithms

Model	Dependent Variable	Independent Variable	Interpretation of β_1
Level-level	y	x	$\Delta y = \beta_1 \Delta x$
Level-log	y	$\log(x)$	$\Delta y = (\beta_1/100)\%\Delta x$
Log-level	$\log(y)$	x	$\%\Delta y = (100\beta_1)\Delta x$
Log-log	$\log(y)$	$\log(x)$	$\%\Delta y = \beta_1 \%\Delta x$

Inference in Regression Models

Standard error of $\hat{\beta}_j$:

$$se(\hat{\beta}_j) = \frac{\hat{\sigma}}{\sqrt{SST_j(1 - R_j^2)}}$$

Sampling distribution:

$$\frac{\hat{\beta}_j - \beta_j}{se(\hat{\beta}_j)} \sim t_{n-k-1},$$

where $k+1$ is the number of unknown parameters in the population model, and $n-k-1$ is the degrees of freedom;

A $(1-\alpha)\%$ confidence interval is defined as $\left[\hat{\beta}_j - c \cdot se(\hat{\beta}_j), \hat{\beta}_j + c \cdot se(\hat{\beta}_j)\right]$, where c is the $(1-\alpha/2)$ percentile in a t_{n-k-1} distribution;

Testing q restrictions:

$$F = \frac{(SSR_r - SSR_{ur})/q}{SSR_{ur}/(n-k-1)} = \frac{(R_{ur}^2 - R_r^2)/q}{(1 - R_{ur}^2)/(n-k-1)} \sim F_{q,n-k-1}$$

Binary Dependent Variables

Linear probability model: $P(y=1|\mathbf{x}) = E(y|\mathbf{x}) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$;

Logit model:

$$G(z) = \frac{\exp(z)}{1 + \exp(z)}$$

Probit model:

$$G(z) = \Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{v^2}{2}\right) dv$$

Likelihood ratio statistics:

$$LR = 2(\mathsf{L}_{ur} - \mathsf{L}_r) \sim \chi_q^2$$

Pseudo R-squared:

$$1 - \mathsf{L}_{ur}/\mathsf{L}_r$$

AIC: $2k - 2\mathsf{L}$; BIC: $\ln(n)k - 2\mathsf{L}$; AICc: AIC $+\frac{2(k+2)(k+3)}{n-k-3}$;