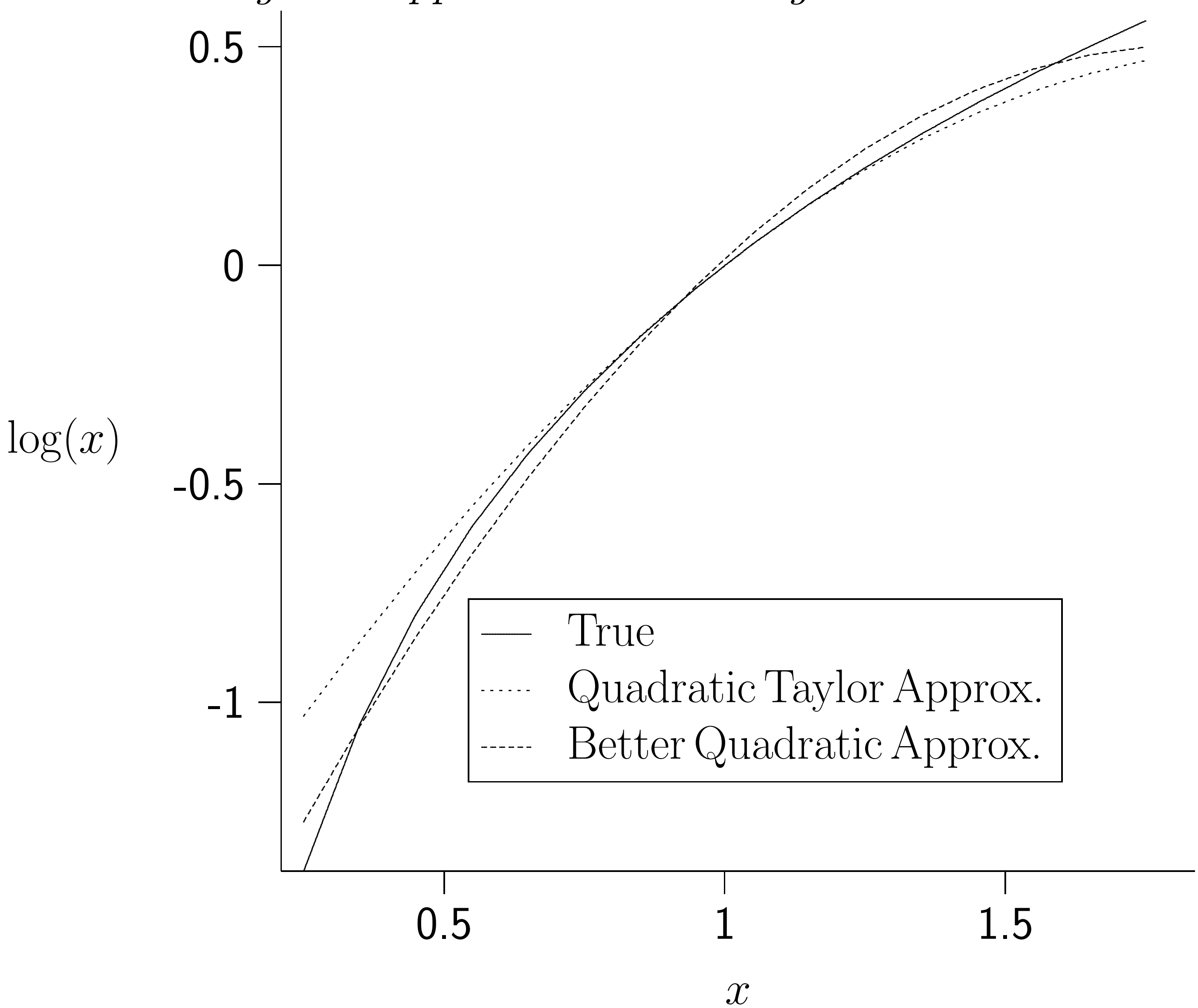
MPV: 7.1-7.2

9. Some explanatory variables can be transformations of existing variables.
   a. $\log$, $\exp$, $\sin$, $\cos$ can exist in a model along side the original.
   b. More immediately, $x^2$, $x^3$, etc.
      i. Result of having $1$ (that is, the intercept), $x, x^2, x^3, \ldots, x^k$ is called a polynomial of order $k$
      ii. and so the model with no higher-order terms is a first order polynomial.
      iii. Useful because well-behaved functions of the explanatory variable can be expressed as a Taylor approximation about the mean.
      iv. Stone-Weierstrass theorem says that any continuous function on a bounded range can be approximated arbitrarily well by a polynomial.
      v. Useful polynomials are of relatively small order.
   c. Ex., $\log(X)$ near $\mu$ is approximately $\log(\mu) + \frac{x-\mu}{\mu} - \frac{(x-\mu)^2}{2\mu^2} + O\left((x-\mu)^3\right)$.
      i. See Fig. 6.

*Fig. 6: Approximation to Log Function*




10. Disadvantages:

   a. Quadratic approximation uses 2 parameters to represent something that might be represented with 1 parameter times a transformation

      i. This gets worse if you add more powers of the variable.

b. Adding parameters allows overfitting

i. For certain $x$ configurations, one can fit any $n$ $Y$ values exactly using $1, x, x^2, \ldots, x^{n-1}$,

ii. Design matrix $\boldsymbol{X} = \begin{pmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^{n-1} \\ 1 & x_2 & x_2^2 & \cdots & x_2^{n-1} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_n & x_n^2 & \cdots & x_n^{n-1} \end{pmatrix}$

iii. $\boldsymbol{X}$ is often (but not always) non-singular.

- Will be singular if $x_i$ are repeated.

iv. $(\boldsymbol{X}^\top \boldsymbol{X})^{-1} = \boldsymbol{X}^{-1} \boldsymbol{X}^{-1\top}$

v. $\hat{\boldsymbol{Y}} = \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{Y} = \boldsymbol{Y}$.

vi. This is to some extent a strawman argument, since no practical statistician adds this many terms.

c. $\boldsymbol{X}^\top \boldsymbol{X}$ has number of rows and columns equal to the number of parameters in the model.

i. Large matrices, even if invertible, may be close to singular

ii. Loosely,

- distance from singularity is referred to as the matrix's condition, and
- matrices close to singular are called ill conditioned.

iii. Ill-conditioned matries are bad:

- Exact inverse leads to highly-variable responses.
- Numerical inverse harder to compute exactly.
- Centering variable before raising to power can help this.

d. Because these higher-order terms are highly variable, extrapolation in this case is a bigger problem than in the linear case.

11. Fits are invariant to affine transformations of regressors used in polynomials

a. Suppose $\hat{Y}_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2$

b. $x_i = \gamma_1 z_i + \gamma_0$

c. Then

$$\hat{Y}_i = \beta_0 + \beta_1(\gamma_1 z_i + \gamma_0) + \beta_2(\gamma_1^2 z_i^2 + 2\gamma 1 \gamma_0 z_i + \gamma_0^2)$$
$$= (\beta_0 + \beta_1\gamma_0 + \beta_2\gamma_0^2) + (\beta_1\gamma_1 + 2\beta_2\gamma_0\gamma_1)z_i + \beta_2\gamma_2^2 z_i^2$$
$$= \alpha_0 + \alpha_1 z_i + \alpha_2 z_i^2$$

i. For $\alpha_0 = \beta_0 + \beta_1\gamma_0 + \beta_2\gamma_0^2$, $\alpha_1 = \beta_1\gamma_1 + 2\beta_2\gamma_0\gamma_1$, $\alpha_2 = \beta_2\gamma_2^2$.

d. Hence model using quadratic in $x_i$ and model using quadratic in $z_i$

i. give same fits.

   ii. can convert back and forth without refitting.

e. Works only if you don't skip powers.

f. Text calls such models hierarchical

g. If the range of the transformed variable is small relative to the curvature of the transformation,

   i. the higher-order terms may be almost colinear with the linear terms.

   ii. Can mask significance of lower-order terms.

12. Changes interpretation of parameter estimates

   a. Columns of design matrix cannot be treated as changable independently.

   b. Hence in the case of polynomial terms, coefficients no longer represent the change in response associated with a unit response in the explanatory variable.

   c. In model $\mathrm{E}\left[Y_j\right] = \beta_0 + \beta_1 x_j + \beta_2 x_j^2$, $d\mathrm{E}\left[Y\right]/dx = \beta_1 + 2\beta_2 x$, and is hence dependent on $x$.

MPV: 7.2.3

13. Trigonometric terms

   a. $\delta_1 \sin(x) + \gamma_1 \cos(x)$.

i. $x$ should be scaled to make period $2\pi$

ii. Angles are measured in radians.

b. Counterpart of higher-order terms for polynomials:

$\delta_j \sin(jx) + \gamma_j \cos(jx)$

i. There is a counterpart to the Stone-Weierstrass theorem demonstrating that one can approximate a bounded function arbitrarily closely with trig terms.

ii. Typically one uses only a few such terms.

c. Careful: if you have a time scale suggesting periodicity, you probably have dependence between temporally similar observations.

d. Terms can represent phase shift using Sum of Angles formula.

i. Let $\alpha = \sqrt{\delta_1^2 + \gamma_1^2}$, $\theta = \tan^{-1}(\gamma_1/\delta_1)$, ( $\theta \in (\pi/2, 3\pi/2)$ if $\delta_1 < 0$ ).

ii. Then $\delta_1 = \alpha\cos(\theta)$, $\gamma_1 = \alpha\cos(\theta)$.

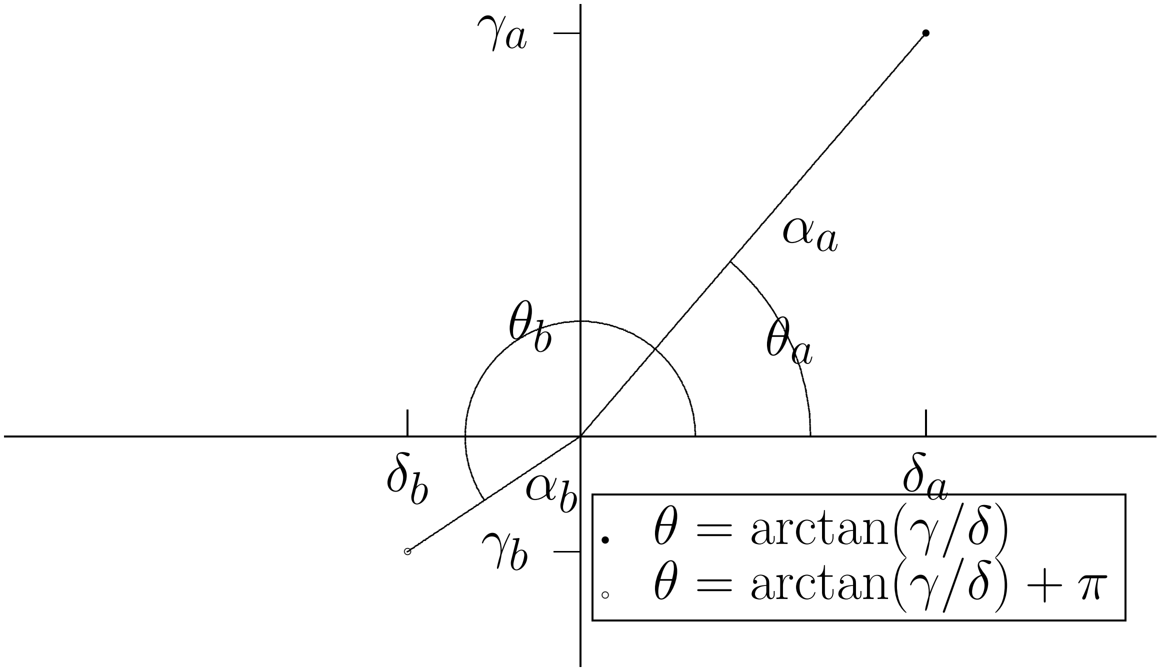
- See Fig. 7.

iii. Then $\delta_1 \sin(x) + \gamma_1 \cos(x) = \alpha \sin(x + \theta)$.

iv. $\theta$ is time shift.

MPV: 7.2.2

*Fig. 7: Geometry behind Trigonometric Transformation*




14. Spline:

   a. A way to draw a smooth curve between two points $x_0$ and $x_N$:

      i. Pick $N-1$ intermediate points $x_1 < x_2 < \cdots < x_{N-2} < x_{N-1}$ (called knots).

      ii. Define a polynomial of degree $M$ between $x_{j-1}$ and $x_j$

      iii. Constrain so that the derivatives of order up to $M-1$ match up at knots.

   b. Use to fit pairs of points $(X_1, Y_1), \ldots, (X_n, Y_n)$.

      i. Taken to an extreme, if all $X_j$ are unique, then we can fit all $n$ points with a polynomial of degree $n-1$.

   c. Denote fit by $\hat{\mu}(x)$

d. Choose to minimize $\sum_{j=1}^{n}(Y_j - \hat{\mu}(X_j))^2$

i. Or penalize, to minimize $\sum_{j=1}^{n}(Y_j - \hat{\mu}(X_j))^2 + \lambda \int_{X_{(1)}}^{X_{(n)}} \hat{\mu}''(x)\, dx$

e. Alternative: The B-spline gives rescaled versions of the piecewise functions.

i. Divide by local product of knot spacing.

MPV: 7.4

15. Can insert polynomials with multiple explanatory variables.

a. Earlier ideas about hierarchical models hold here too.

b. If you want a model that gives the same fit under affine transformation of all regressors, you can include interaction terms

c. That is, the model including terms $x_1^2$ and $x_2^2$ might include $x_1 x_2$ as well.

d. This logic is less compelling that for one variable.

MPV: 7.5

16. One may use orthogonal polynomials to remove colinearity

a. Calculate the constant, linear, quadratic, *et cetera.* terms as before.

i. Let the result be $\boldsymbol{X}$

b. Use orthogonalization as we did earlier to give orthogonal regressors

i. Let the result be $\boldsymbol{Z}$

c. Normalize if desired, to make $\sum_i z_{ij}^2 = 1$ for all $i$

d. Just as before, column $j$ of $\boldsymbol{Z}$ is a linear combination of columns $1, \ldots, j$ of $\boldsymbol{X}$ .

e. Hence get same fitted values.

f. With multiple variables, orthonormalization is applied only to the portion of the matrix corresponding to powers of one variable.

MPV: 7.3

J. Nonparametric Regression

1. Kernel smoothing:

a. Get an expression that is explicit rather than implicit: $\hat{g}(x) = \sum_{j=1}^{n} Y_j w((x - X_j)/\Delta) / \sum_{j=1}^{n} w((x - X_j)/\Delta)$ .

b. Weight function can be

i. the same as above

ii. Often a normal density.

iii. Often uniform density centered at $0$.

c. Method is kernel smoothing, and specifically is Nadaraya-Watson smoothing.

2. A local regression smoother has smaller bias than kernel smoother.

a. $\hat{g}(x) = \sum_{\ell=0}^{L} \hat{\beta}_\ell x^\ell$, for $L = 1$

b. $\hat{\boldsymbol{\beta}} = \text{argmin}\left(\sum_{j=1}^{n}\left(Y_j - \sum_{\ell=0}^{L}\hat{\beta}_\ell X_j^\ell\right)^2 w\left(\frac{x-X_j}{\Delta_n}\right)\right)$.

3. LOESS

a. $f(x)$ fitted value at $x$ for low-degree (viz., linear or quadratic) regression of points with $X_j$ near $x$.

b. Specify the number of points $k$

c. Upweight points near $x$ and downweight them away from $x$

d. Weighting function scaled to make point in neighborhood farthest from $x$ have weight going down to zero.

i. This keeps the curve smooth as $x$ moves.

e. Common weight function is $w(x) = (1 - |x|^3)^3$.

f. So $\hat{f}(x) = \hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_2 x^2$ for

i. $\hat{\boldsymbol{\beta}} = \text{argmin}(\sum_{j \in N(x)}(Y_j - \beta_0 - \beta_1 X_j - \beta_2 X_j^2)^2 w((x - X_j)/\Delta))$ for

ii. $N(x) =$ indices of $k$ closest points to $x$, and

iii. $\Delta = \max\{\left|X_j - x\right| \, |j \in N(x)\}$.

g. Procedure formerly Lowess, Locally Weighted Sum of Squares.

h. Result can not be expressed as a simple formula.