Lecture 1: Discrete Models and Optimization

Jean-François Houde
UW-Madison

November 1, 2021

Discrete Choice Models

- **Random utility framework:**
  - Choice-set: $j \in \{0, 1, \ldots, J\}$
  - Payoff function:

$$V_{ij} = u_j(X_{ij}; \beta) + \epsilon_{ij}$$

  - Choice-probabilities: Utility maximization

$$\begin{aligned} \Pr(y_i = j | X_i, \beta) &= \Pr(V_{ij} > V_{ij'}, \forall j' \neq j | X_i, \beta) \\ &= \Pr(u_{ij} - u_{ij'} > \epsilon_{ij} - \epsilon_{ij'}, \forall j' \neq j | X_i, \beta) \end{aligned}$$

  - **Goal:** Infer the shape of the payoff function from chosen actions (revealed-preference)
  - Likelihood function (panel setting):

$$l(Y, X | \beta) = \sum_i \ln \Pr(y_{i1}, \ldots, y_{iT} | X_i, \beta)$$

Discrete Choice Models

- **Random utility framework:**
  - Choice-set: $j \in \{0, 1, \dots, J\}$
  - Payoff function:
$$V_{ij} = u_j(X_{ij}; \beta) + \epsilon_{ij}$$
  - Choice-probabilities: Utility maximization
$$\begin{aligned}
\Pr(y_i = j | X_i, \beta) &= \Pr(V_{ij} > V_{ij'}, \forall j' \neq j | X_i, \beta) \\
&= \Pr(u_{ij} - u_{ij'} > \epsilon_{ij} - \epsilon_{ij'}, \forall j' \neq j | X_i, \beta)
\end{aligned}$$
  - **Goal:** Infer the shape of the payoff function from chosen actions (revealed-preference)
  - Likelihood function (panel setting):
$$l(Y, X | \beta) = \sum_i \ln \Pr(y_{i1}, \dots, y_{iT} | X_i, \beta)$$
- **Identification:**
  - *Scale invariant:* Choices are unaffected by multiplying V by any $c > 0$
  - *Normalization:* Choices are determined by differences in payoffs

Discrete Choice Models

- **Assumptions:**
  - Normalizations: $u_0(X_{i0}) = 0$ and $E(\epsilon_{ij}) = 0$
  - Linearity (not very restrictive): $u_j(X_{ij}; \beta) = X_{ij}\beta$
  - Conditional independence (relaxed later): $F(\epsilon_{ij}|X_i) = F(\epsilon_{ij})$

Discrete Choice Models

- **Assumptions:**
  - Normalizations: $u_0(X_{i0}) = 0$ and $E(\epsilon_{ij}) = 0$
  - Linearity (not very restrictive): $u_j(X_{ij};\beta) = X_{ij}\beta$
  - Conditional independence (relaxed later): $F(\epsilon_{ij}|X_i) = F(\epsilon_{ij})$
- **Scale restrictions:**
  - *IID errors:* $\text{Var}(\epsilon_{ij}) = 1$ and slopes are measured relative to s.d. of ϵ (β/σ_ϵ)
  - *Correlated errors:* Example $j = 1, 2, 3$.

$$\text{Var}(\epsilon_i) = \Omega = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} \\ . & \sigma_2^2 & \sigma_{23} \\ . & . & \sigma_3^2 \end{pmatrix}$$

  - After normalizing $(\epsilon_{ij} - \epsilon_{i0})$ and standardizing $(m = var(\epsilon_{i1} - \epsilon_{i0}) = 1)$:

$$\tilde{\Omega} = \begin{pmatrix} 1 & (\sigma_1^2 + \sigma_{23} - \sigma_{12} - \sigma_{13})/m \\ . & (\sigma_1^2 + \sigma_3^2 - \sigma_{13})/m \end{pmatrix} = \begin{pmatrix} 1 & \rho \\ . & \sigma \end{pmatrix}$$

Binary Choices

- *Probit:* $\epsilon_{i1} - \epsilon_{i0} \sim N(0,1)$

$$\Pr(y_i = 1|X_i,\beta) = 1 - \Phi(-X_i\beta) = \text{Normal CDF}$$

- *Logit:* $\epsilon_{ij} \sim \text{T1EV}(0,1)$

$$\Pr(y_i = 1|X_i,\beta) = \frac{\exp(X_i\beta)}{1+\exp(X_i\beta)} = \Lambda(X_i\beta)$$

- *Linear probability:* $\epsilon_{i1} - \epsilon_{i0} \sim U(-\sigma,\sigma)$

$$\Pr(y_i = 1|X_i,\beta) = 1 - \Pr(\epsilon_{i1} - \epsilon_{i0} < X_i\beta) = \frac{2\sigma + X_i\tilde{\beta}}{2\sigma} = X_i\beta$$

- *Semi-parametric:*

$$\Pr(y_i = 1|X_i,\beta) = G(X_i\beta)$$

where $G(\cdot)$ is the non-parametric CDF (e.g. Kernel or Mixture-of-Normals).

Estimation: Logit example

- ML estimator (cross-section):

$$\max_{\beta} \sum_{i} \ln \left[\Lambda(X_i\beta)_i^y \left(1 - \Lambda(X_i\beta)\right)^{1-y_i} \right] = \max_{\beta} L(\beta)$$

- The FOC is given by the Score of the log-likelihood:

$$g(\beta) = \sum_{i=1}^{n} (y_i - \Lambda_i) X_i = 0$$

  where $\Lambda_i \equiv \Lambda(X_i\beta)$ and $g(\beta)$ is a $1 \times K$ vector.
- The SOC is given by the Hessian:

$$H(\beta) = -\sum_{i=1}^{n} \Lambda_i (1 - \Lambda_i) X_i' X_i < 0$$

Numerical optimization: Newton's Method

- Second-order Taylor's expansion:

$$L(\beta) \approx L(\beta^0) + g(\beta^0)(\beta - \beta^0) + \frac{1}{2}(\beta - \beta^0)' H(\beta^0)(\beta - \beta^0)$$

- If $H(\beta)$ is negative definite, the approximation to the LLF is maximized at:

$$\beta = \beta^0 - H\left(\beta^0\right)^{-1} g(\beta^0)'$$

- This suggests an iterative approach to maximize $L(\beta)$:
  - Initial value: β^0
  - Iteration k: $\beta^k = \beta^{k-1} - H\left(\beta^{k-1}\right)^{-1} g(\beta^{k-1})'$
  - Repeat until: $||\beta^k - \beta^{k-1}|| < \eta$ (e.g. 10^{-12})
- Newton's algorithm converges quadratically when starting values are in a "neighborhood" of the solution. To facilitate convergence, the updating step is augmented by a "line search" parameter $s^k \in (0, 1)$:

$$\max_{s^k} L\left(\beta^{k-1} - s_k H\left(\beta^{k-1}\right)^{-1} g(\beta^{k-1})'\right)$$

Alternatives: Quasi-Newton Methods

- Quasi-newton methods replace the Hessian with an approximation:

$$\beta^k = \beta^{k-1} - B^k g(\beta^{k-1})'$$

$$\beta^k = \beta^{k-1} - s^k$$

- Steepest descent method ($B^k = I$):

$$\beta^k = \beta^{k-1} - g(\beta^{k-1})$$

- BFGS (default method): Update Hessian approximation as follows

$$B^{k+1} = B^k - \frac{B^k z^k z^{k'} B^k}{z^{k'} B^k z^k} + \frac{y^k y^{k'}}{y^{k'} z^k}$$

where $y^k = (g(\beta^k) - g(\beta^{k-1}))'$ and $z^k = \beta^k - \beta^{k-1}$.

Alternatives: Simplex method (Nelder-Mead)

- Nelder-Mead is a commonly used derivative-free optimization method for multivariate problems (matlab: fminsearch)
- Two-dimension example: $\beta = (\beta_1, \beta_2)$
- Step 1: Simplex calculation

$$\begin{pmatrix} \beta^0 \\ \beta^1 \\ \beta^2 \end{pmatrix} = \begin{pmatrix} (\beta_1, \beta_2) \\ (\beta_1 + s_1, \beta_2) \\ (\beta_1, \beta_2 + s_2) \end{pmatrix}$$

  where $s_k = \beta_k \delta$ (e.g. $\delta = .05$)
  - Notation: Simplex centroid (M)

$$M = \sum_{l=1}^{3} \beta^l \frac{1}{3}$$

  - Evaluate and re-order function: $L(\beta^{(1)}) > L(\beta^{(2)}) > L(\beta^{(3)})$
- Important: The algorithm updates the parameter by repeatedly replacing the worst point in the simplex ($\beta^{(3)}$)

Alternatives: Simplex method (Nelder-Mead)

- Step 2: Evaluate the function at new reflection point β^R

$$\beta^R = M + \alpha(M - \beta^{(3)}), \quad \alpha = 1$$

- Step 3: Update simplex
  - Case 1 (Improvement) If $L(\beta^R) > L(\beta^1)$, **expand** the simplex in the same direction

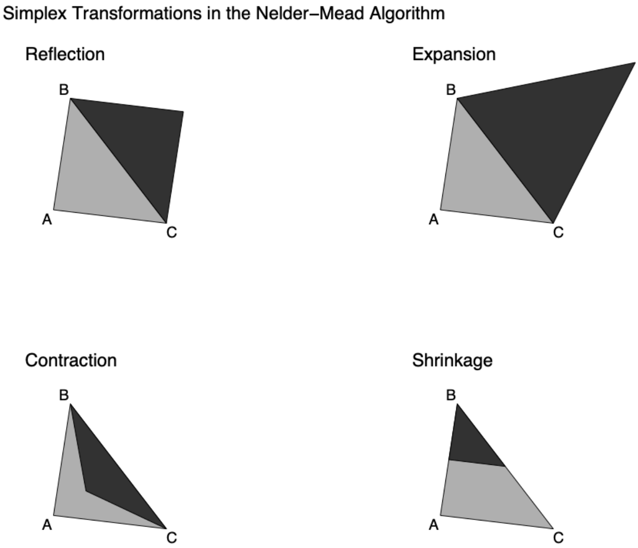
    $$\beta^E = \beta^R + \gamma(\beta^R - M), \quad \gamma = 1$$

    If $L(\beta^E) > L(\beta^0)$, replace $\beta^{(3)}$ with β^E, otherwise use β^R. Repeat 2.
  - Case 2: If $L(\beta^{(2)}) < L(\beta^R) < L(\beta^1)$, replace $\beta^{(3)}$ by β^R. Repeat 2.
  - Case 3 (Contraction): If $L(\beta^{(2)}) > L(\beta^R)$ we contract the simplex

    $$\beta^C = \begin{cases} M + \beta(\beta^R - M) & L(\beta^R) > L(\beta^{(3)}) \\ M + \beta(\beta^{(3)} - M) & L(\beta^R) < L(\beta^{(3)}) \end{cases}$$

    where $\beta \leq 1/2$. If β^C is an improvement over $\beta^{(3)}$, replace and go to 2.
  - Case 4: Otherwise **shrink** the simplex towards the best direction.

Simplex: Visual example



Additional options

- Other derivative-free algorithm:
    - Golden search method
    - Simulated Annealing
    - Genetics algorithm
    - ...
- Advices:
    - Know your function: Start with a (univariate) grid searches to verify that the function is smooth
    - If the function is not smooth... Try to smooth the model!
    - Before using Newton-based methods, find good starting values using grid-search and/or derivative free methods (e.g. Simplex)
    - If the model is inherently non-smooth use derivative-free methods and start from different points (and be patient!)
- More advices: Pay attention to the scale of the parameters and of the objective function
    - Good practice: $\beta^* \approx 1$
    - Good practice: $\max_\beta L(\beta)/N$

Multinomial Choice Models

- **McFadden:** Multinomial Logit Model
  - Random-utility:
    $$V_{ij} = X_{ij}\beta + \epsilon_{ij}$$
  - Utility shocks: $\epsilon_{ij} \sim \text{T1EV}(0,1)$.
  - Choice-probabilities under utility maximization:
    $$\begin{aligned} \Pr(y_i = 1|X_i,\beta) &= \Pr(V_{i1} > V_{i0},\dots,V_{i1} > V_{iJ}) \\ &= \frac{\exp(X_{i1}\beta)}{\sum_{j=0}^{J}\exp(X_{ij}\beta)} \end{aligned}$$
- Likelihood function:
  $$\max_{\beta} \sum_{i} \sum_{j=0}^{J} \mathbf{1}(y_i = j) \ln\left[\Pr(y_i = 1|X_i,\beta)\right]$$

Common Specifications

- *Multinomial Logit Model:* Choice depends chooser's characteristics

$$\Pr(y_i = j|x_i) = \frac{\exp(x_i\beta_j)}{1 + \sum_{j'=1}^{J} \exp(x_i\beta_{j'})}$$

  - Interpretation: β_j measures the marginal effect of individual attribute x_i on the payoff of option j relative to option 0 (scaled by sd of ϵ_{ij}).

Common Specifications

- *Multinomial Logit Model:* Choice depends chooser's characteristics

$$\Pr(y_i = j|x_i) = \frac{\exp(x_i\beta_j)}{1 + \sum_{j'=1}^{J} \exp(x_i\beta_{j'})}$$

  - Interpretation: β_j measures the marginal effect of individual attribute x_i on the payoff of option j relative to option 0 (scaled by sd of ϵ_{ij}).

- *Conditional Logit Model:* Choice depends option characteristics

$$\Pr(y_i = j|z_i, X_{i0}, \ldots, X_{iJ}) = \frac{\exp(X_{ij}\gamma)}{\sum_{j'=0}^{J} \exp(X_{ij}\gamma)}$$

  - Intercept and common variables normalization: $\gamma_0 = 0$ and $\gamma_z = 0$

Example: Major choice

- Return to education: Cobb-Douglas function for majors $j = 0, \ldots, J$

$$\ln \text{Wage}_{ij} = \gamma_{j0} + \sum_k x_{ik}\gamma_{jk} + \sigma_\epsilon \epsilon_{ij}$$

Option 0: No college.

- Behavioral assumption: Students maximize return to schooling

$$\Pr(\text{Major}_i = j | x_i) = \begin{cases} \frac{\exp(\beta_{j0} + \sum_k x_{ik}\beta_{jk})}{1 + \sum_{j'=1}^{J} \exp(\beta_{j'0} + \sum_k x_{ik}\beta_{j'k})} & \text{If } j \geq 1 \\ \frac{1}{1 + \sum_{j'=1}^{J} \exp(\beta_{j'0} + \sum_k x_{ik}\beta_{j'k})} & \text{If } j = 0 \end{cases}$$

Where $\beta_{jk} = (\gamma_{jk} - \gamma_{0k})/\sigma_\epsilon$

Example: Major choice

- Return to education: Cobb-Douglas function for majors $j = 0, \dots, J$

$$\ln \text{Wage}_{ij} = \gamma_{j0} + \sum_k x_{ik}\gamma_{jk} + \sigma_\epsilon \epsilon_{ij}$$

  Option 0: No college.

- Behavioral assumption: Students maximize return to schooling

$$\Pr(\text{Major}_i = j|x_i) = \begin{cases} \frac{\exp(\beta_{j0}+\sum_k x_{ik}\beta_{jk})}{1+\sum_{j'=1}^J \exp(\beta_{j'0}+\sum_k x_{ik}\beta_{j'k})} & \text{If } j \geq 1 \\ \frac{1}{1+\sum_{j'=1}^J \exp(\beta_{j'0}+\sum_k x_{ik}\beta_{j'k})} & \text{If } j = 0 \end{cases}$$

  Where $\beta_{jk} = (\gamma_{jk} - \gamma_{0k})/\sigma_\epsilon$

- Interpretation of β_k?

$$\text{Odds ratio}_j = \frac{\Pr(\text{Major}_i = j|x_i)}{\Pr(\text{Major}_i = 0|x_i)} = \exp(\beta_{j0} + \sum_k x_{ik}\beta_k)$$

$$\Rightarrow \frac{\text{Odds ratio}_j(x_{ik} + \Delta)}{\text{Odds ratio}_j(x_{ik})} = \exp(\beta_{jk}\Delta) = \Delta \text{ Odds of choosing } j \text{ over } 0$$

Consumer Surplus and Elasticities

- Random utility: Surplus

$$\text{Willingness to Pay}_{ij} - p_{ij} \Rightarrow V_{ij} = X_{ij}\beta - \alpha p_{ij} + \epsilon_{ij}, \quad \alpha = 1/\sigma_\epsilon$$

Consumer Surplus and Elasticities

- Random utility: Surplus

$$\text{Willingness to Pay}_{ij} - p_{ij} \Rightarrow V_{ij} = X_{ij}\beta - \alpha p_{ij} + \epsilon_{ij}, \quad \alpha = 1/\sigma_\epsilon$$

- Consumer surplus: Integrating over the distribution of consumer shocks (ϵ_{ij})

$$CS_i(X_{ij}) = \frac{1}{\alpha} E[\max_j V_{ij}] = \frac{1}{\alpha} \ln \left[\sum_j \exp(X_{ij}\beta) \right] = \frac{1}{\alpha}[\text{Inclusive value}]$$

Consumer Surplus and Elasticities

- Random utility: Surplus

$$\text{Willingness to Pay}_{ij} - p_{ij} \Rightarrow V_{ij} = X_{ij}\beta - \alpha p_{ij} + \epsilon_{ij}, \quad \alpha = 1/\sigma_\epsilon$$

- Consumer surplus: Integrating over the distribution of consumer shocks (ϵ_{ij})

$$CS_i(X_{ij}) = \frac{1}{\alpha} E[\max_j V_{ij}] = \frac{1}{\alpha} \ln \left[\sum_j \exp(X_{ij}\beta) \right] = \frac{1}{\alpha}[\text{Inclusive value}]$$

- Substitution patterns. Let $D_j = \Pr(y = j)$:

$$\frac{\partial D_j}{\partial p_k} = \begin{cases} -\alpha D_j(1 - D_j) & \text{If } j = k \\ \alpha D_j D_k & \text{If } j \neq k \end{cases}$$

- This implies that the elasticity of substitution between j and k is proportional to demand for k

$$\text{Elasticity}_{j,k} = \alpha p_k D_k$$

Independence of Irrelevance Alternatives

- The relative odds of choosing alternative j over k is independent of the characteristics of other options:

$$\begin{aligned}
\frac{\Pr(y_i = j|X_i, \beta)}{\Pr(y_i = k|X_i, \beta)} &= \frac{\exp(X_{ij}\beta - \alpha p_{ij}) / \sum_{j'=0}^{J} \exp(X_{ij'}\beta - \alpha p_{ij'})}{\exp(X_{ik}\beta - \alpha p_{ik}) / \sum_{j'=0}^{J} \exp(X_{ij'}\beta - \alpha p_{ij'})} \\
&= \frac{\exp(X_{ij}\beta - \alpha p_{ij})}{\exp(X_{ik}\beta - \alpha p_{ik})}
\end{aligned}$$

Independence of Irrelevance Alternatives

- The relative odds of choosing alternative j over k is independent of the characteristics of other options:

$$\begin{aligned}\frac{\Pr(y_i = j|X_i, \beta)}{\Pr(y_i = k|X_i, \beta)} &= \frac{\exp(X_{ij}\beta - \alpha p_{ij}) / \sum_{j'=0}^{J} \exp(X_{ij'}\beta - \alpha p_{ij'})}{\exp(X_{ik}\beta - \alpha p_{ik}) / \sum_{j'=0}^{J} \exp(X_{ij'}\beta - \alpha p_{ij'})} \\ &= \frac{\exp(X_{ij}\beta - \alpha p_{ij})}{\exp(X_{ik}\beta - \alpha p_{ik})}\end{aligned}$$

- **Red-bus/Blue-bus:**
  - Two options: car (c) and blue-bus (bb)
  - Equal choice-probability: $P_c = P_{bb} = 1/2$ and $P_c/P_{bb} = 1/2$

Independence of Irrelevance Alternatives

- The relative odds of choosing alternative j over k is independent of the characteristics of other options:

$$\frac{\Pr(y_i = j|X_i, \beta)}{\Pr(y_i = k|X_i, \beta)} = \frac{\exp(X_{ij}\beta - \alpha p_{ij}) / \sum_{j'=0}^{J} \exp(X_{ij'}\beta - \alpha p_{ij'})}{\exp(X_{ik}\beta - \alpha p_{ik}) / \sum_{j'=0}^{J} \exp(X_{ij'}\beta - \alpha p_{ij'})}$$
$$= \frac{\exp(X_{ij}\beta - \alpha p_{ij})}{\exp(X_{ik}\beta - \alpha p_{ik})}$$

- **Red-bus/Blue-bus:**
  - Two options: car (c) and blue-bus (bb)
  - Equal choice-probability: $P_c = P_{bb} = 1/2$ and $P_c/P_{bb} = 1/2$
  - Identical new red-bus ($X_{i,bb}\beta = X_{i,rb}\beta$): $P_{bb}/P_{rb} = 1$.
  - If $P_b/P_{rb} = 1$ and $P_c/P_{bb} = 1$, the logit model predicts:

$$P_{bb} = P_c = P_{rb} = 1/3$$

  - Does it make sense?

Independence of Irrelevance Alternatives

- The relative odds of choosing alternative j over k is independent of the characteristics of other options:

$$\begin{aligned}\frac{\Pr(y_i = j|X_i,\beta)}{\Pr(y_i = k|X_i,\beta)} &= \frac{\exp(X_{ij}\beta - \alpha p_{ij}) / \sum_{j'=0}^{J} \exp(X_{ij'}\beta - \alpha p_{ij'})}{\exp(X_{ik}\beta - \alpha p_{ik}) / \sum_{j'=0}^{J} \exp(X_{ij'}\beta - \alpha p_{ij'})} \\ &= \frac{\exp(X_{ij}\beta - \alpha p_{ij})}{\exp(X_{ik}\beta - \alpha p_{ik})}\end{aligned}$$

- **Red-bus/Blue-bus:**
  - Two options: car (c) and blue-bus (bb)
  - Equal choice-probability: $P_c = P_{bb} = 1/2$ and $P_c/P_{bb} = 1/2$
  - Identical new red-bus ($X_{i,bb}\beta = X_{i,rb}\beta$): $P_{bb}/P_{rb} = 1$.
  - If $P_b/P_{rb} = 1$ and $P_c/P_{bb} = 1$, the logit model predicts:

$$P_{bb} = P_c = P_{rb} = 1/3$$

  - Does it make sense?
  - Probably not... It would be more reasonable to expect $P_c = 1/2$ and $P_{bb} = P_{rb} = 1/4$.

Example: Demand for insurance

Source: Apesteguia and Ballester, JPE, 2018

- Two options:
  - Risky option:

$$u_{i1} = 0.9\frac{1^{1-\omega_i}}{1-\omega_i} + 0.1\frac{60^{1-\omega_i}}{1-\omega_i}$$

  - Risk-free option:

$$u_{i2} = \frac{5^{1-\omega_i}}{1-\omega_i}$$

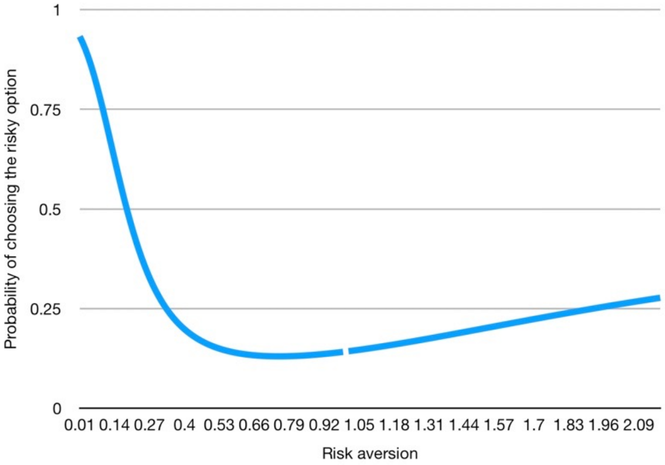
Example: Demand for insurance

Source: Apesteguia and Ballester, JPE, 2018

- Two options:
  - Risky option:
$$u_{i1} = 0.9\frac{1^{1-\omega_i}}{1-\omega_i} + 0.1\frac{60^{1-\omega_i}}{1-\omega_i}$$
  - Risk-free option:
$$u_{i2} = \frac{5^{1-\omega_i}}{1-\omega_i}$$
- Random-utility model:
$$\max\left\{u_{i1}/\sigma_\epsilon + \epsilon_{i1}, u_{i2}/\sigma_\epsilon + \epsilon_{i2}\right\}$$

  If $\epsilon_{ij} \sim T1EV(0,1)$, we get the following probability of choosing the risky option:
$$\Pr(y_i = 1) = \frac{\exp((u_{i1}-u_{i2})/\sigma_\epsilon)}{1+\exp((u_{i1}-u_{i2})/\sigma_\epsilon)}$$

Logit Probability: Demand for Insurance



- **Implications:** (i) non-monotonic relationship between gamble and risk-aversion, and (ii) Prob(risky) goes to 1/2 as $\omega \to \infty$ (instead of 0).