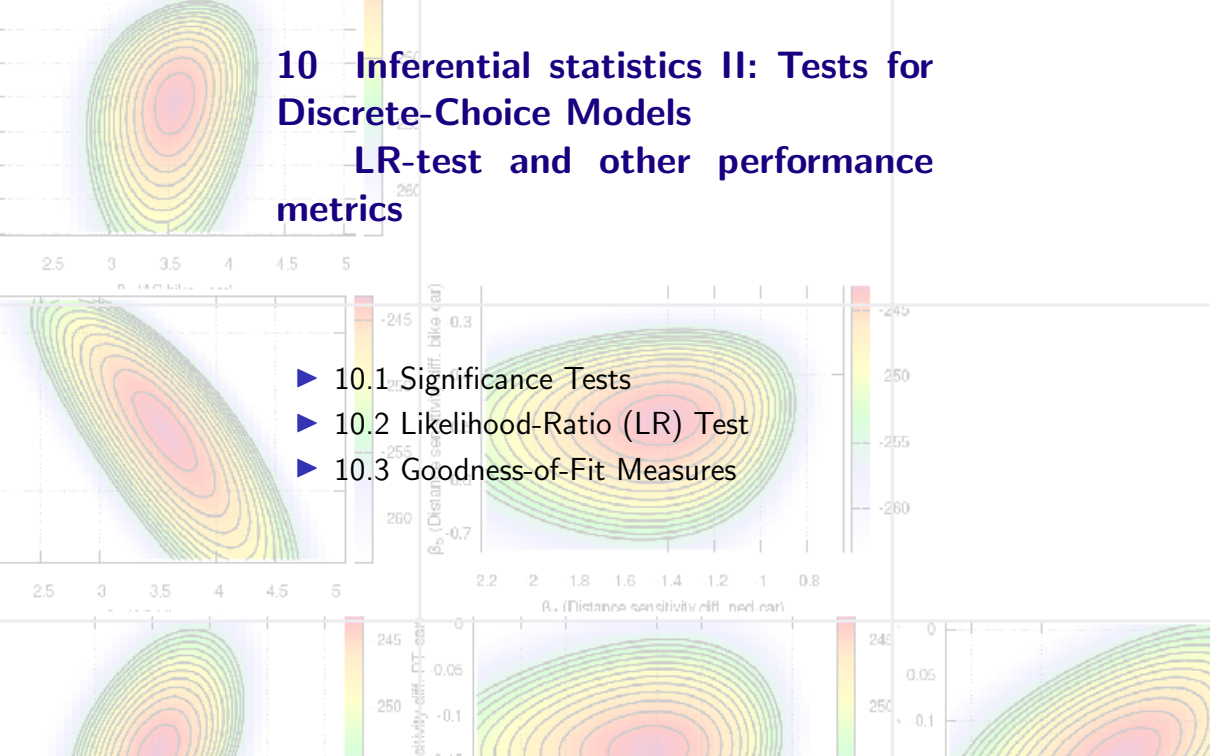


10 Inferential statistics II: Tests for Discrete-Choice Models

LR-test and other performance metrics

- ▶ 10.1 Significance Tests
- ▶ 10.2 Likelihood-Ratio (LR) Test
- ▶ 10.3 Goodness-of-Fit Measures



10.1 Significance Tests for Discrete-Choice Models

- ▶ The parameter test procedures are exactly the same as that of regression models. Because we only consider the asymptotic limit, the test statistic is always Gaussian:
- ▶ Confidence interval of a parameter β_m :

$$CI_{\alpha}(\beta_m) = [\hat{\beta}_m - \Delta_{\alpha}, \hat{\beta}_m + \Delta_{\alpha}], \quad \Delta_{\alpha} = z_{1-\alpha/2} \sqrt{V_{mm}}$$

- ▶ Test of a parameter β_m for $H_0 : \beta_j = \beta_{j0}, \geq \beta_{j0},$ or $\leq \beta_{j0}$:

$$T = \frac{\hat{\beta}_j - \hat{\beta}_{j0}}{\sqrt{V_{jj}}} \sim N(0, 1) \mid H_0^*$$

- ▶ p -values for $H_0 : \beta_j = \beta_{j0}, \geq \beta_{j0},$ or $\leq \beta_{j0},$ respectively:

$$p_{=} = 2(1 - \Phi(|t_{\text{data}}|)), \quad p_{\leq} = 1 - \Phi(t_{\text{data}}), \quad p_{\geq} = \Phi(t_{\text{data}})$$

- ▶ As in regression, a factor 4 of more data halves the error

10.1 Significance Tests for Discrete-Choice Models

- ▶ The parameter test procedures are exactly the same as that of regression models. Because we only consider the asymptotic limit, the test statistic is always Gaussian:
- ▶ Confidence interval of a parameter β_m :

$$CI_{\alpha}(\beta_m) = [\hat{\beta}_m - \Delta_{\alpha}, \hat{\beta}_m + \Delta_{\alpha}], \quad \Delta_{\alpha} = z_{1-\alpha/2} \sqrt{V_{mm}}$$

- ▶ Test of a parameter β_m for $H_0 : \beta_j = \beta_{j0}, \geq \beta_{j0},$ or $\leq \beta_{j0}$:

$$T = \frac{\hat{\beta}_j - \hat{\beta}_{j0}}{\sqrt{V_{jj}}} \sim N(0, 1) \mid H_0^*$$

- ▶ p -values for $H_0 : \beta_j = \beta_{j0}, \geq \beta_{j0},$ or $\leq \beta_{j0},$ respectively:

$$p_{=} = 2(1 - \Phi(|t_{\text{data}}|)), \quad p_{\leq} = 1 - \Phi(t_{\text{data}}), \quad p_{\geq} = \Phi(t_{\text{data}})$$

- ▶ As in regression, a factor 4 of more data halves the error

10.1 Significance Tests for Discrete-Choice Models

- ▶ The parameter test procedures are exactly the same as that of regression models. Because we only consider the asymptotic limit, the test statistic is always Gaussian:
- ▶ Confidence interval of a parameter β_m :

$$CI_{\alpha}(\beta_m) = [\hat{\beta}_m - \Delta_{\alpha}, \hat{\beta}_m + \Delta_{\alpha}], \quad \Delta_{\alpha} = z_{1-\alpha/2} \sqrt{V_{mm}}$$

- ▶ Test of a parameter β_m for $H_0 : \beta_j = \beta_{j0}, \geq \beta_{j0},$ or $\leq \beta_{j0}$:

$$T = \frac{\hat{\beta}_j - \hat{\beta}_{j0}}{\sqrt{V_{jj}}} \sim N(0, 1) \mid H_0^*$$

- ▶ p -values for $H_0 : \beta_j = \beta_{j0}, \geq \beta_{j0},$ or $\leq \beta_{j0},$ respectively:

$$p_{=} = 2(1 - \Phi(|t_{\text{data}}|)), \quad p_{\leq} = 1 - \Phi(t_{\text{data}}), \quad p_{\geq} = \Phi(t_{\text{data}})$$

- ▶ As in regression, a factor 4 of more data halves the error

10.1 Significance Tests for Discrete-Choice Models

- ▶ The parameter test procedures are exactly the same as that of regression models. Because we only consider the asymptotic limit, the test statistic is always Gaussian:
- ▶ Confidence interval of a parameter β_m :

$$CI_{\alpha}(\beta_m) = [\hat{\beta}_m - \Delta_{\alpha}, \hat{\beta}_m + \Delta_{\alpha}], \quad \Delta_{\alpha} = z_{1-\alpha/2} \sqrt{V_{mm}}$$

- ▶ Test of a parameter β_m for $H_0 : \beta_j = \beta_{j0}, \geq \beta_{j0},$ or $\leq \beta_{j0}$:

$$T = \frac{\hat{\beta}_j - \hat{\beta}_{j0}}{\sqrt{V_{jj}}} \sim N(0, 1) \mid H_0^*$$

- ▶ p -values for $H_0 : \beta_j = \beta_{j0}, \geq \beta_{j0},$ or $\leq \beta_{j0},$ respectively:

$$p_{=} = 2(1 - \Phi(|t_{\text{data}}|)), \quad p_{\leq} = 1 - \Phi(t_{\text{data}}), \quad p_{\geq} = \Phi(t_{\text{data}})$$

- ▶ As in regression, a factor 4 of more data halves the error

10.1 Significance Tests for Discrete-Choice Models

- ▶ The parameter test procedures are exactly the same as that of regression models. Because we only consider the asymptotic limit, the test statistic is always Gaussian:
- ▶ Confidence interval of a parameter β_m :

$$CI_{\alpha}(\beta_m) = [\hat{\beta}_m - \Delta_{\alpha}, \hat{\beta}_m + \Delta_{\alpha}], \quad \Delta_{\alpha} = z_{1-\alpha/2} \sqrt{V_{mm}}$$

- ▶ Test of a parameter β_m for $H_0 : \beta_j = \beta_{j0}, \geq \beta_{j0},$ or $\leq \beta_{j0}$:

$$T = \frac{\hat{\beta}_j - \hat{\beta}_{j0}}{\sqrt{V_{jj}}} \sim N(0, 1) \mid H_0^*$$

- ▶ p -values for $H_0 : \beta_j = \beta_{j0}, \geq \beta_{j0},$ or $\leq \beta_{j0},$ respectively:

$$p_{=} = 2(1 - \Phi(|t_{\text{data}}|)), \quad p_{\leq} = 1 - \Phi(t_{\text{data}}), \quad p_{\geq} = \Phi(t_{\text{data}})$$

- ▶ As in regression, a factor 4 of more data halves the error

10.2. Likelihood-Ratio (LR) Test

Like in regression (F-test), one sometimes wants to test null hypotheses fixing several parameters *simultaneously* to given values, i.e., H_0 corresponds to a **restraint model**

- ▶ H_0 : The restraint model with some fixed parameters and M_r remaining parameters describes the data as well as the full model with M parameters
- ▶ Test statistics:

$$\lambda^{\text{LR}} = 2 \ln \left(\frac{L(\hat{\beta})}{L^r(\hat{\beta}^r)} \right) = 2 \left[\tilde{L}(\hat{\beta}) - \tilde{L}^r(\hat{\beta}^r) \right] \sim \chi^2(M - M_r) \text{ if } H_0$$

- ▶ Data realization: calibrate both M and M_r and evaluate $\lambda_{\text{data}}^{\text{LR}}$
- ▶ Result: reject H_0 at α based on the $1 - \alpha$ quantile:

$$\lambda_{\text{data}}^{\text{LR}} > \chi_{1-\alpha, M-M_r}^2$$

$$p\text{-value: } p = 1 - F_{\chi^2(M-M_r)}(\lambda_{\text{data}}^{\text{LR}})$$

10.2. Likelihood-Ratio (LR) Test

Like in regression (F-test), one sometimes wants to test null hypotheses fixing several parameters *simultaneously* to given values, i.e., H_0 corresponds to a **restraint model**

- ▶ H_0 : The restraint model with some fixed parameters and M_r remaining parameters describes the data as well as the full model with M parameters
- ▶ Test statistics:

$$\lambda^{\text{LR}} = 2 \ln \left(\frac{L(\hat{\beta})}{L^r(\hat{\beta}^r)} \right) = 2 \left[\tilde{L}(\hat{\beta}) - \tilde{L}^r(\hat{\beta}^r) \right] \sim \chi^2(M - M_r) \text{ if } H_0$$

- ▶ Data realization: calibrate both M and M_r and evaluate $\lambda_{\text{data}}^{\text{LR}}$
- ▶ Result: reject H_0 at α based on the $1 - \alpha$ quantile:

$$\lambda_{\text{data}}^{\text{LR}} > \chi_{1-\alpha, M-M_r}^2$$

$$p\text{-value: } p = 1 - F_{\chi^2(M-M_r)}(\lambda_{\text{data}}^{\text{LR}})$$

10.2. Likelihood-Ratio (LR) Test

Like in regression (F-test), one sometimes wants to test null hypotheses fixing several parameters *simultaneously* to given values, i.e., H_0 corresponds to a **restraint model**

- ▶ H_0 : The restraint model with some fixed parameters and M_r remaining parameters describes the data as well as the full model with M parameters
- ▶ Test statistics:

$$\lambda^{\text{LR}} = 2 \ln \left(\frac{L(\hat{\beta})}{L^r(\hat{\beta}^r)} \right) = 2 \left[\tilde{L}(\hat{\beta}) - \tilde{L}^r(\hat{\beta}^r) \right] \sim \chi^2(M - M_r) \text{ if } H_0$$

- ▶ Data realization: calibrate both M and M_r and evaluate $\lambda_{\text{data}}^{\text{LR}}$
- ▶ Result: reject H_0 at α based on the $1 - \alpha$ quantile:

$$\lambda_{\text{data}}^{\text{LR}} > \chi_{1-\alpha, M-M_r}^2$$

$$p\text{-value: } p = 1 - F_{\chi^2(M-M_r)}(\lambda_{\text{data}}^{\text{LR}})$$

10.2. Likelihood-Ratio (LR) Test

Like in regression (F-test), one sometimes wants to test null hypotheses fixing several parameters *simultaneously* to given values, i.e., H_0 corresponds to a **restraint model**

- ▶ H_0 : The restraint model with some fixed parameters and M_r remaining parameters describes the data as well as the full model with M parameters
- ▶ Test statistics:

$$\lambda^{\text{LR}} = 2 \ln \left(\frac{L(\hat{\beta})}{L^r(\hat{\beta}^r)} \right) = 2 \left[\tilde{L}(\hat{\beta}) - \tilde{L}^r(\hat{\beta}^r) \right] \sim \chi^2(M - M_r) \text{ if } H_0$$

- ▶ Data realization: calibrate both M and M_r and evaluate $\lambda_{\text{data}}^{\text{LR}}$
- ▶ Result: reject H_0 at α based on the $1 - \alpha$ quantile:

$$\lambda_{\text{data}}^{\text{LR}} > \chi_{1-\alpha, M-M_r}^2$$

$$p\text{-value: } p = 1 - F_{\chi^2(M-M_r)}(\lambda_{\text{data}}^{\text{LR}})$$

10.2. Likelihood-Ratio (LR) Test

Like in regression (F-test), one sometimes wants to test null hypotheses fixing several parameters *simultaneously* to given values, i.e., H_0 corresponds to a **restraint model**

- ▶ H_0 : The restraint model with some fixed parameters and M_r remaining parameters describes the data as well as the full model with M parameters
- ▶ Test statistics:

$$\lambda^{\text{LR}} = 2 \ln \left(\frac{L(\hat{\beta})}{L^r(\hat{\beta}^r)} \right) = 2 \left[\tilde{L}(\hat{\beta}) - \tilde{L}^r(\hat{\beta}^r) \right] \sim \chi^2(M - M_r) \text{ if } H_0$$

- ▶ Data realization: calibrate both M and M_r and evaluate $\lambda_{\text{data}}^{\text{LR}}$
- ▶ Result: reject H_0 at α based on the $1 - \alpha$ quantile:

$$\lambda_{\text{data}}^{\text{LR}} > \chi_{1-\alpha, M-M_r}^2$$

$$p\text{-value: } p = 1 - F_{\chi^2(M-M_r)}(\lambda_{\text{data}}^{\text{LR}})$$

Example: Mode choice for the route to this lecture

Distance class n	Distance r_n	$i = 1$ (ped/bike)	$i = 2$ (PT/car)
$n = 1$: 0-1 km	0.5 km	7	1
$n = 2$: 1-2 km	1.5 km	6	4
$n = 3$: 2-5 km	3.5 km	6	12
$n = 4$: 5-10 km	7.5 km	1	10
$n = 5$: 10-20 km	15.0 km	0	5

$$V_{n1}(\beta_1, \beta_2) = \beta_1 r_n + \beta_2,$$

$$V_{n2}(\beta_1, \beta_2) = 0$$

- ▶ β_1 : Difference in distance sensitivity (utility/km) for choosing ped/bike over PT/car (expected < 0)
- ▶ β_2 : Utility difference ped/bike over PT/car at zero distance (> 0)

Do the data allow to distinguish this model from the trivial model $V_{ni} = 0$?

Example: Mode choice for the route to this lecture

Distance class n	Distance r_n	$i = 1$ (ped/bike)	$i = 2$ (PT/car)
$n = 1$: 0-1 km	0.5 km	7	1
$n = 2$: 1-2 km	1.5 km	6	4
$n = 3$: 2-5 km	3.5 km	6	12
$n = 4$: 5-10 km	7.5 km	1	10
$n = 5$: 10-20 km	15.0 km	0	5

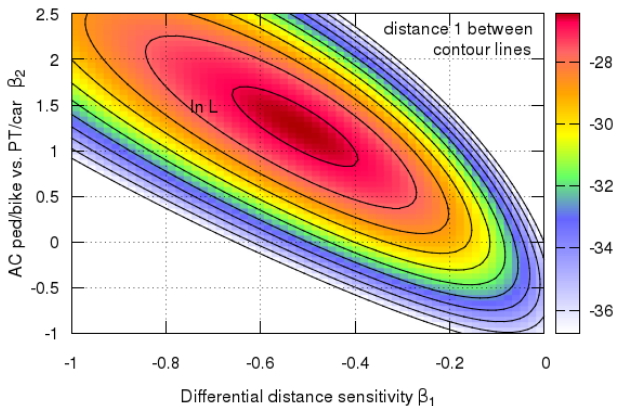
$$V_{n1}(\beta_1, \beta_2) = \beta_1 r_n + \beta_2,$$

$$V_{n2}(\beta_1, \beta_2) = 0$$

- ▶ β_1 : Difference in distance sensitivity (utility/km) for choosing ped/bike over PT/car (expected < 0)
- ▶ β_2 : Utility difference ped/bike over PT/car at zero distance (> 0)

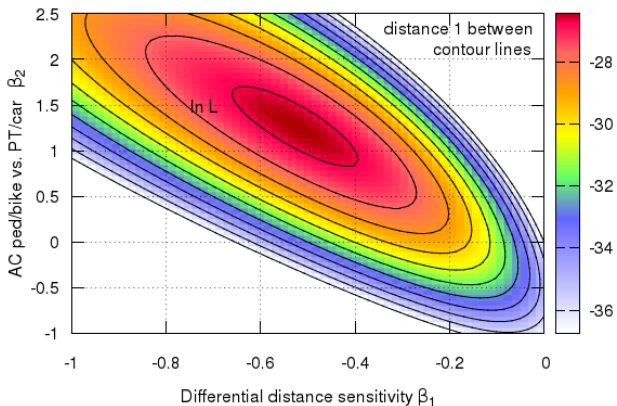
Do the data allow to distinguish this model from the trivial model $V_{ni} = 0$?

LR test for the corresponding Logit models



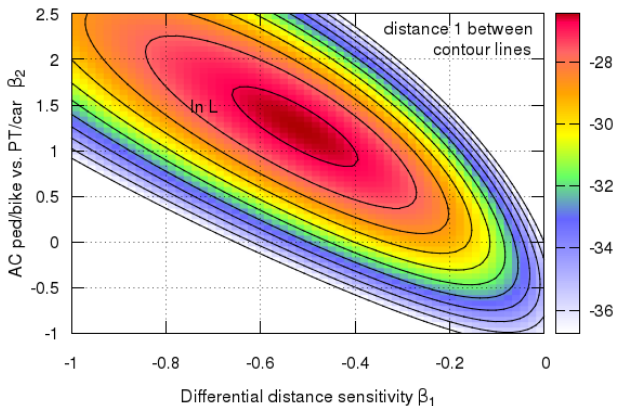
- ▶ H_0 : The trivial model $V_{ni} = 0$ describes the data as well as the full model
 $V_{n1}(\beta_1, \beta_2) = (\beta_1 r_n + \beta_2) \delta_{i1}$
- ▶ Test statistics: $\lambda^{\text{LR}} = 2 \left[\tilde{L}(\hat{\beta}_1, \hat{\beta}_2) - \tilde{L}(0, 0) \right] \sim \chi^2(2) | H_0$
- ▶ Data realization (1 \tilde{L} -unit per contour): $\lambda_{\text{data}}^{\text{LR}} = 2(-26.5 + 35.5) = 18$
- ▶ Decision: Rejection range $\lambda^{\text{LR}} > \chi_{2,0.95}^2 = 5.99 \Rightarrow H_0$ rejected.

LR test for the corresponding Logit models



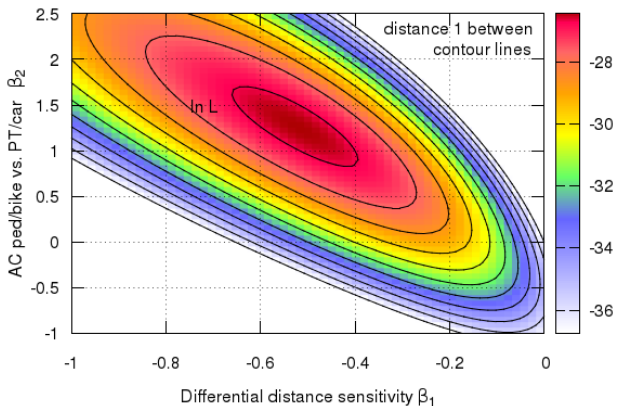
- ▶ H_0 : The trivial model $V_{ni} = 0$ describes the data as well as the full model $V_{n1}(\beta_1, \beta_2) = (\beta_1 r_n + \beta_2) \delta_{i1}$
- ▶ Test statistics: $\lambda^{\text{LR}} = 2 \left[\tilde{L}(\hat{\beta}_1, \hat{\beta}_2) - \tilde{L}(0, 0) \right] \sim \chi^2(2) | H_0$
- ▶ Data realization (1 \tilde{L} -unit per contour): $\lambda_{\text{data}}^{\text{LR}} = 2(-26.5 + 35.5) = 18$
- ▶ Decision: Rejection range $\lambda^{\text{LR}} > \chi_{2,0.95}^2 = 5.99 \Rightarrow H_0$ rejected.

LR test for the corresponding Logit models



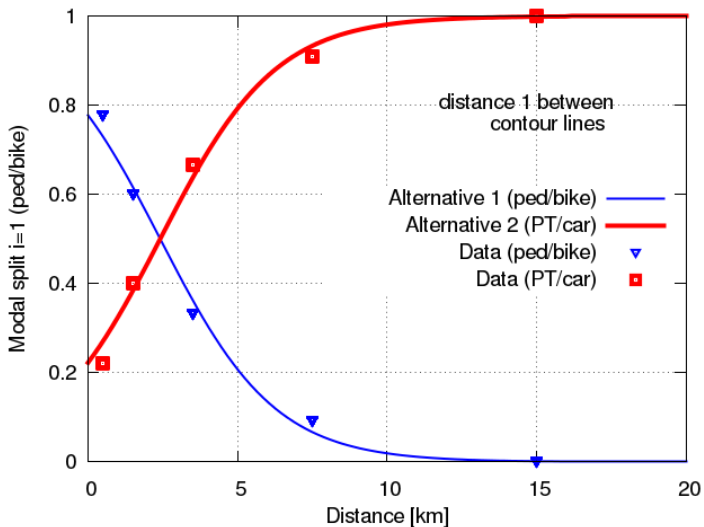
- ▶ H_0 : The trivial model $V_{ni} = 0$ describes the data as well as the full model $V_{n1}(\beta_1, \beta_2) = (\beta_1 r_n + \beta_2) \delta_{i1}$
- ▶ Test statistics: $\lambda^{\text{LR}} = 2 \left[\tilde{L}(\hat{\beta}_1, \hat{\beta}_2) - \tilde{L}(0, 0) \right] \sim \chi^2(2) | H_0$
- ▶ Data realization (1 \tilde{L} -unit per contour): $\lambda_{\text{data}}^{\text{LR}} = 2(-26.5 + 35.5) = 18$
- ▶ Decision: Rejection range $\lambda^{\text{LR}} > \chi_{2,0.95}^2 = 5.99 \Rightarrow H_0$ rejected.

LR test for the corresponding Logit models



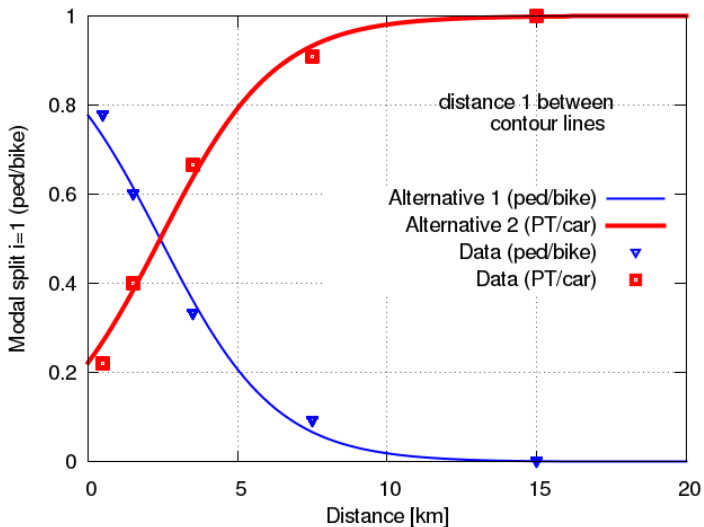
- ▶ H_0 : The trivial model $V_{ni} = 0$ describes the data as well as the full model $V_{n1}(\beta_1, \beta_2) = (\beta_1 r_n + \beta_2) \delta_{i1}$
- ▶ Test statistics: $\lambda^{\text{LR}} = 2 \left[\tilde{L}(\hat{\beta}_1, \hat{\beta}_2) - \tilde{L}(0, 0) \right] \sim \chi^2(2) | H_0$
- ▶ Data realization (1 \tilde{L} -unit per contour): $\lambda_{\text{data}}^{\text{LR}} = 2(-26.5 + 35.5) = 18$
- ▶ Decision: Rejection range $\lambda^{\text{LR}} > \chi_{2,0.95}^2 = 5.99 \Rightarrow H_0$ rejected.

Fit quality of the full model



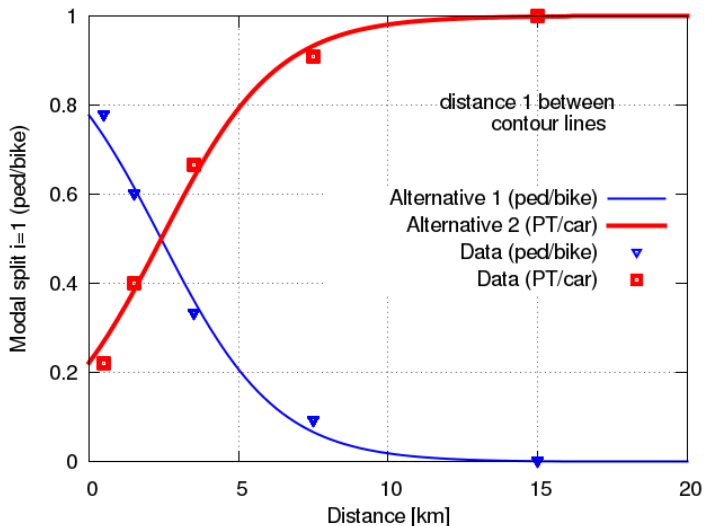
- ? What would be the modeled ped/bike modal split for the null model $V_{ni} = 0$? 50:50
- ? Read off from the \tilde{L} contour plot the parameter of the AC-only model $V_{ni} = \beta_2 \delta_{i1}$ and give the modeled modal split $\hat{\beta}_2 = \ln(P_1/P_2) = -0.5$, OK with $P_1/P_2 = e^{\hat{\beta}_2} \approx N_1/N_2 = 20/32$
- ? Motivate the negative correlation between the parameter errors. This makes at least sure that, in case of correlated errors, about the same fraction chooses alternative 2 as for the calibrated model.

Fit quality of the full model



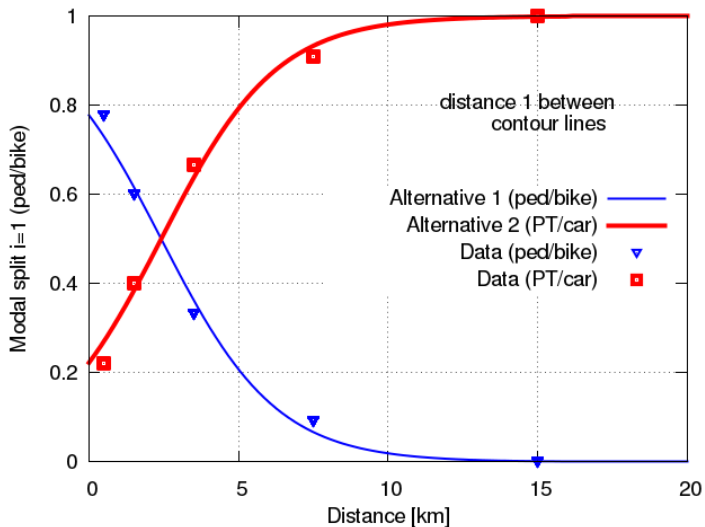
- ? What would be the modeled ped/bike modal split for the null model $V_{ni} = 0$? 50:50
- ? Read off from the \tilde{L} contour plot the parameter of the AC-only model $V_{ni} = \beta_2 \delta_{i1}$ and give the modeled modal split $\hat{\beta}_2 = \ln(P_1/P_2) = -0.5$, OK with $P_1/P_2 = e^{\hat{\beta}_2} \approx N_1/N_2 = 20/32$
- ? Motivate the negative correlation between the parameter errors. This makes at least sure that, in case of correlated errors, about the same fraction chooses alternative 2 as for the calibrated model.

Fit quality of the full model



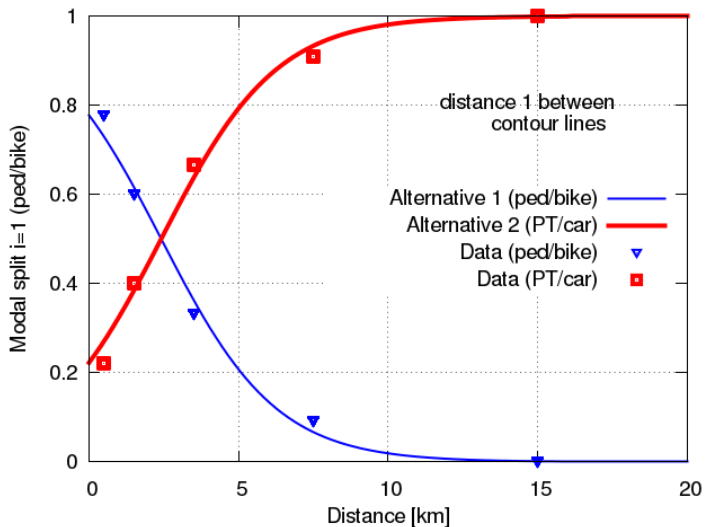
- ? What would be the modeled ped/bike modal split for the null model $V_{ni} = 0$? 50:50
- ? Read off from the \tilde{L} contour plot the parameter of the AC-only model $V_{ni} = \beta_2 \delta_{i1}$ and give the modeled modal split
 $\hat{\beta}_2 = \ln(P_1/P_2) = -0.5$, OK with $P_1/P_2 = e^{\hat{\beta}_2} \approx N_1/N_2 = 20/32$
- ? Motivate the negative correlation between the parameter errors This makes at least sure that, in case of correlated errors, about the same fraction chooses alternative 2 as for the calibrated model

Fit quality of the full model



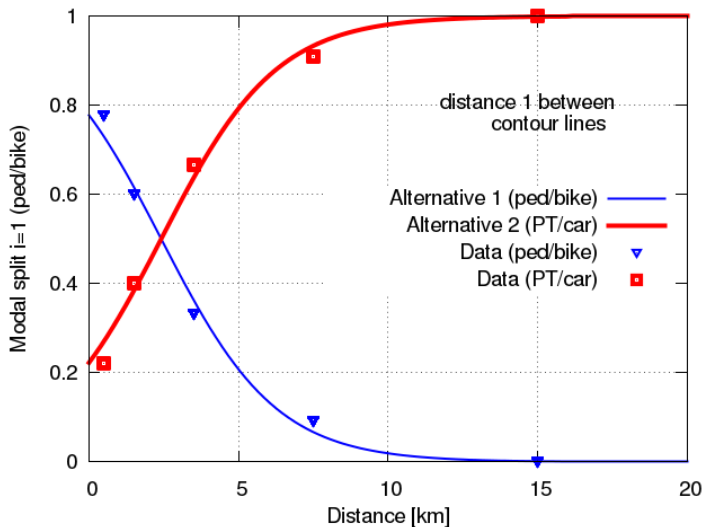
- ? What would be the modeled ped/bike modal split for the null model $V_{ni} = 0$? 50:50
- ? Read off from the \tilde{L} contour plot the parameter of the AC-only model $V_{ni} = \beta_2 \delta_{i1}$ and give the modeled modal split $\hat{\beta}_2 = \ln(P_1/P_2) = -0.5$, OK with $P_1/P_2 = e^{\hat{\beta}_2} \approx N_1/N_2 = 20/32$
- ? Motivate the negative correlation between the parameter errors This makes at least sure that, in case of correlated errors, about the same fraction chooses alternative 2 as for the calibrated model

Fit quality of the full model



- ? What would be the modeled ped/bike modal split for the null model $V_{ni} = 0$? 50:50
- ? Read off from the \tilde{L} contour plot the parameter of the AC-only model $V_{ni} = \beta_2 \delta_{i1}$ and give the modeled modal split $\hat{\beta}_2 = \ln(P_1/P_2) = -0.5$, OK with $P_1/P_2 = e^{\hat{\beta}_2} \approx N_1/N_2 = 20/32$
- ? Motivate the negative correlation between the parameter errors This makes at least sure that, in case of correlated errors, about the same fraction chooses alternative 2 as for the calibrated model

Fit quality of the full model



- ? What would be the modeled ped/bike modal split for the null model $V_{ni} = 0$? 50:50
- ? Read off from the \tilde{L} contour plot the parameter of the AC-only model $V_{ni} = \beta_2 \delta_{i1}$ and give the modeled modal split $\hat{\beta}_2 = \ln(P_1/P_2) = -0.5$, OK with $P_1/P_2 = e^{\hat{\beta}_2} \approx N_1/N_2 = 20/32$
- ? Motivate the negative correlation between the parameter errors This makes at least sure that, in case of correlated errors, about the same fraction chooses alternative 2 as for the calibrated model

10.3 Goodness-of-Fit Measures

- ▶ The parameter tests for equality and the LR test are related to **significance**: Is the more complicated of two nested models significantly better in describing the data?
- ▶ This can be used to find the best model using the **top-down ansatz**:



Make is as simple as possible but not simpler!

- ▶ Problem: For very big samples, nearly any new parameter gives significance and the top-down ansatz fails
- ▶ More importantly: Significance/LR tests cannot give evidence for missing but relevant factors
- ▶ A further problem: We cannot compare non-nested models
- ▶ Finally, in reality, one often is interested in **effect strength** (difference in the fit and validation quality), not significance

⇒ we need measures for absolute fit quality

10.3 Goodness-of-Fit Measures

- ▶ The parameter tests for equality and the LR test are related to **significance**: Is the more complicated of two nested models significantly better in describing the data?
- ▶ This can be used to find the best model using the **top-down ansatz**:



Make is as simple as possible but not simpler!

- ▶ Problem: For very big samples, nearly any new parameter gives significance and the top-down ansatz fails
- ▶ More importantly: Significance/LR tests cannot give evidence for missing but relevant factors
- ▶ A further problem: We cannot compare non-nested models
- ▶ Finally, in reality, one often is interested in **effect strength** (difference in the fit and validation quality), not significance

⇒ we need measures for absolute fit quality

10.3 Goodness-of-Fit Measures

- ▶ The parameter tests for equality and the LR test are related to **significance**: Is the more complicated of two nested models significantly better in describing the data?
- ▶ This can be used to find the best model using the **top-down ansatz**:




Make it as simple as possible but not simpler!

- ▶ Problem: For very big samples, nearly any new parameter gives significance and the top-down ansatz fails
- ▶ More importantly: Significance/LR tests cannot give evidence for missing but relevant factors
- ▶ A further problem: We cannot compare non-nested models
- ▶ Finally, in reality, one often is interested in **effect strength** (difference in the fit and validation quality), not significance


⇒ we need measures for absolute fit quality

10.3 Goodness-of-Fit Measures

- ▶ The parameter tests for equality and the LR test are related to **significance**: Is the more complicated of two nested models significantly better in describing the data?
- ▶ This can be used to find the best model using the **top-down ansatz**:
 *Make is as simple as possible but not simpler!*
- ▶ Problem: For very big samples, nearly any new parameter gives significance and the top-down ansatz fails
- ▶ More importantly: Significance/LR tests cannot give evidence for missing but relevant factors
- ▶ A further problem: We cannot compare non-nested models
- ▶ Finally, in reality, one often is interested in **effect strength** (difference in the fit and validation quality), not significance


⇒ we need measures for absolute fit quality

10.3 Goodness-of-Fit Measures

- ▶ The parameter tests for equality and the LR test are related to **significance**: Is the more complicated of two nested models significantly better in describing the data?
- ▶ This can be used to find the best model using the **top-down ansatz**:
 *Make is as simple as possible but not simpler!*
- ▶ Problem: For very big samples, nearly any new parameter gives significance and the top-down ansatz fails
- ▶ More importantly: Significance/LR tests cannot give evidence for missing but relevant factors
- ▶ A further problem: We cannot compare non-nested models
- ▶ Finally, in reality, one often is interested in **effect strength** (difference in the fit and validation quality), not significance

⇒ we need measures for absolute fit quality

10.3 Goodness-of-Fit Measures

- ▶ The parameter tests for equality and the LR test are related to **significance**: Is the more complicated of two nested models significantly better in describing the data?
- ▶ This can be used to find the best model using the **top-down ansatz**:
 *Make is as simple as possible but not simpler!*
- ▶ Problem: For very big samples, nearly any new parameter gives significance and the top-down ansatz fails
- ▶ More importantly: Significance/LR tests cannot give evidence for missing but relevant factors
- ▶ A further problem: We cannot compare non-nested models
- ▶ Finally, in reality, one often is interested in **effect strength** (difference in the fit and validation quality), not significance

⇒ **we need measures for absolute fit quality**

Information-based goodness-of-fit (GoF) measures

▶ **Akaike's information criterion:**

$$\text{AIC} = -2\tilde{L} + 2M \frac{N}{N - (M + 1)}$$

▶ **Bayesian information criterion:**

$$\text{BIC} = -2\tilde{L} + M \ln N$$

N : number of decisions; M : number of parameters

- ▶ Both criteria give the needed additional information (in bit) to obtain the actual micro-data from the model's prediction, including an over-fitting penalty: the lower, the better.
- ▶ Both the AIC and BIC are equivalent to the corresponding GoF measures of regression.
- ▶ the BIC focuses more on parsimonious models (low M).
- ▶ For nested models satisfying the null hypothesis of the LR test and $N \gg M$, the expected AIC is the same (verify!). However, since the AIC is an absolute measure, it allows comparing non-nested models.

Information-based goodness-of-fit (GoF) measures

▶ **Akaike's information criterion:**

$$\text{AIC} = -2\tilde{L} + 2M \frac{N}{N - (M + 1)}$$

▶ **Bayesian information criterion:**

$$\text{BIC} = -2\tilde{L} + M \ln N$$

N : number of decisions; M : number of parameters

- ▶ Both criteria give the needed additional information (in bit) to obtain the actual micro-data from the model's prediction, including an over-fitting penalty: the lower, the better.
- ▶ Both the AIC and BIC are equivalent to the corresponding GoF measures of regression.
- ▶ the BIC focuses more on parsimonious models (low M).
- ▶ For nested models satisfying the null hypothesis of the LR test and $N \gg M$, the expected AIC is the same (verify!). However, since the AIC is an absolute measure, it allows comparing non-nested models.

Information-based goodness-of-fit (GoF) measures

▶ **Akaike's information criterion:**

$$\text{AIC} = -2\tilde{L} + 2M \frac{N}{N - (M + 1)}$$

▶ **Bayesian information criterion:**

$$\text{BIC} = -2\tilde{L} + M \ln N$$

N : number of decisions; M : number of parameters

- ▶ Both criteria give the needed additional information (in bit) to obtain the actual micro-data from the model's prediction, including an over-fitting penalty: the lower, the better.
- ▶ Both the AIC and BIC are equivalent to the corresponding GoF measures of regression.
- ▶ the BIC focuses more on parsimonious models (low M).
- ▶ For nested models satisfying the null hypothesis of the LR test and $N \gg M$, the expected AIC is the same (**verify!**). However, since the AIC is an absolute measure, it allows comparing non-nested models.

Information-based goodness-of-fit (GoF) measures

▶ **Akaike's information criterion:**

$$\text{AIC} = -2\tilde{L} + 2M \frac{N}{N - (M + 1)}$$

▶ **Bayesian information criterion:**

$$\text{BIC} = -2\tilde{L} + M \ln N$$

N : number of decisions; M : number of parameters

- ▶ Both criteria give the needed additional information (in bit) to obtain the actual micro-data from the model's prediction, including an over-fitting penalty: the lower, the better.
- ▶ Both the AIC and BIC are equivalent to the corresponding GoF measures of regression.
- ▶ the BIC focuses more on parsimonious models (low M).
- ▶ For nested models satisfying the null hypothesis of the LR test and $N \gg M$, the expected AIC is the same (**verify!**). However, since the AIC is an absolute measure, it allows comparing non-nested models.

Information-based goodness-of-fit (GoF) measures

▶ **Akaike's information criterion:**

$$\text{AIC} = -2\tilde{L} + 2M \frac{N}{N - (M + 1)}$$

▶ **Bayesian information criterion:**

$$\text{BIC} = -2\tilde{L} + M \ln N$$

N : number of decisions; M : number of parameters

- ▶ Both criteria give the needed additional information (in bit) to obtain the actual micro-data from the model's prediction, including an over-fitting penalty: the lower, the better.
- ▶ Both the AIC and BIC are equivalent to the corresponding GoF measures of regression.
- ▶ the BIC focuses more on parsimonious models (low M).
- ▶ For nested models satisfying the null hypothesis of the LR test and $N \gg M$, the expected AIC is the same (**verify!**). However, since the AIC is an absolute measure, it allows comparing non-nested models.

GoF measures corresponding to the coefficient of determination R^2 of linear models (\tilde{L}^0 : log-likelihood of the estimated AC-only or trivial model)

- ▶ **LR-Index** resp. **McFadden's R^2** :

$$\rho^2 = 1 - \frac{\tilde{L}}{\tilde{L}^0}$$

- ▶ **Adjusted LR-Index/McFadden's R^2** :

$$\bar{\rho}^2 = 1 - \frac{\tilde{L} - M}{\tilde{L}^0}$$

- ▶ The LR-Index ρ^2 and the adjusted LR-Index $\bar{\rho}^2$ correspond to the coefficient of determination R^2 and the adjusted coefficient \bar{R}^2 of regression models, respectively: The higher, the better.
- ▶ In contrast to regression models, even the best-fitting model has ρ^2 and $\bar{\rho}^2$ values far from 1. Values as low as 0.3 may characterize a good model, see [the Example 9.2.1](#), while $R^2 = 0.3$ means a really bad fit for a regression model.
- ▶ An over-fitted model with M parameters fitting $N = M$ decisions reaches the "ideal" LR-index value $\rho^2 = 1$ while $\bar{\rho}^2$ is near zero.

GoF measures corresponding to the coefficient of determination R^2 of linear models (\tilde{L}^0 : log-likelihood of the estimated AC-only or trivial model)

- ▶ **LR-Index** resp. **McFadden's R^2** :

$$\rho^2 = 1 - \frac{\tilde{L}}{\tilde{L}^0}$$

- ▶ **Adjusted LR-Index/McFadden's R^2** :

$$\bar{\rho}^2 = 1 - \frac{\tilde{L} - M}{\tilde{L}^0}$$

- ▶ The LR-Index ρ^2 and the adjusted LR-Index $\bar{\rho}^2$ correspond to the coefficient of determination R^2 and the adjusted coefficient \tilde{R}^2 of regression models, respectively: The higher, the better.
- ▶ In contrast to regression models, even the best-fitting model has ρ^2 and $\bar{\rho}^2$ values far from 1. Values as low as 0.3 may characterize a good model, see [the Example 9.2.1](#), while $R^2 = 0.3$ means a really bad fit for a regression model.
- ▶ An over-fitted model with M parameters fitting $N = M$ decisions reaches the "ideal" LR-index value $\rho^2 = 1$ while $\bar{\rho}^2$ is near zero.

GoF measures corresponding to the coefficient of determination R^2 of linear models (\tilde{L}^0 : log-likelihood of the estimated AC-only or trivial model)

- ▶ **LR-Index** resp. **McFadden's R^2** :

$$\rho^2 = 1 - \frac{\tilde{L}}{\tilde{L}^0}$$

- ▶ **Adjusted LR-Index/McFadden's R^2** :

$$\bar{\rho}^2 = 1 - \frac{\tilde{L} - M}{\tilde{L}^0}$$

- ▶ The LR-Index ρ^2 and the adjusted LR-Index $\bar{\rho}^2$ correspond to the coefficient of determination R^2 and the adjusted coefficient \tilde{R}^2 of regression models, respectively: The higher, the better.
- ▶ In contrast to regression models, even the best-fitting model has ρ^2 and $\bar{\rho}^2$ values far from 1. Values as low as 0.3 may characterize a good model, see [the Example 9.2.1](#), while $R^2 = 0.3$ means a really bad fit for a regression model.
- ▶ An over-fitted model with M parameters fitting $N = M$ decisions reaches the "ideal" LR-index value $\rho^2 = 1$ while $\bar{\rho}^2$ is near zero.

GoF measures corresponding to the coefficient of determination R^2 of linear models (\tilde{L}^0 : log-likelihood of the estimated AC-only or trivial model)

- ▶ **LR-Index** resp. **McFadden's R^2** :

$$\rho^2 = 1 - \frac{\tilde{L}}{\tilde{L}^0}$$

- ▶ **Adjusted LR-Index/McFadden's R^2** :

$$\bar{\rho}^2 = 1 - \frac{\tilde{L} - M}{\tilde{L}^0}$$

- ▶ The LR-Index ρ^2 and the adjusted LR-Index $\bar{\rho}^2$ correspond to the coefficient of determination R^2 and the adjusted coefficient \tilde{R}^2 of regression models, respectively: The higher, the better.
- ▶ In contrast to regression models, even the best-fitting model has ρ^2 and $\bar{\rho}^2$ values far from 1. Values as low as 0.3 may characterize a good model, see [the Example 9.2.1](#), while $R^2 = 0.3$ means a really bad fit for a regression model.
- ▶ An over-fitted model with M parameters fitting $N = M$ decisions reaches the “ideal” LR-index value $\rho^2 = 1$ while $\bar{\rho}^2$ is near zero.

Questions on GoF metrics

- ? Discuss the model to be tested, the AC-only model, and the trivial model in the context of weather forecasts

Full forecast info, info from climate table, 50:50

- ? Give the log-likelihood of the AC-only and trivial models if there are I alternatives and N_i decisions for alternative i (total number of decisions $N = \sum_{i=1}^I N_i$)

Trivial model: $P_{ni} = 1/I$, $\tilde{L} = \sum_n \ln P_{i_n} = \sum_i N_i \ln P_i = -N \ln I$;

AC-only model: $P_{ni} = N_i/N$, $\tilde{L} = \sum_i N_i \ln P_i = N \ln N - \sum_i N_i \ln N_i$

- ? Consider a binary choice situation where the $N/2$ persons with short trips chose the pedestrian/bike option with a probability of 3/4, and the PT/car option with 1/4. The other $N/2$ persons with long trips had the reverse modal split with a ped/bike usage of 25%, only.

What would be the LR-index for the "perfect" model exactly reproducing the observed 3:1 and 1:3 modal splits for the short and long trips, respectively?
(less than 0.18)

Questions on GoF metrics

- ? Discuss the model to be tested, the AC-only model, and the trivial model in the context of weather forecasts

Full forecast info, info from climate table, 50:50

- ? Give the log-likelihood of the AC-only and trivial models if there are I alternatives and N_i decisions for alternative i (total number of decisions $N = \sum_{i=1}^I N_i$)

Trivial model: $P_{ni} = 1/I$, $\tilde{L} = \sum_n \ln P_{i_n} = \sum_i N_i \ln P_i = -N \ln I$;

AC-only model: $P_{ni} = N_i/N$, $\tilde{L} = \sum_i N_i \ln P_i = N \ln N - \sum_i N_i \ln N_i$

- ? Consider a binary choice situation where the $N/2$ persons with short trips chose the pedestrian/bike option with a probability of $3/4$, and the PT/car option with $1/4$. The other $N/2$ persons with long trips had the reverse modal split with a ped/bike usage of 25%, only.

What would be the LR-index for the "perfect" model exactly reproducing the observed 3:1 and 1:3 modal splits for the short and long trips, respectively?
(less than 0.18)

Questions on GoF metrics

- ? Discuss the model to be tested, the AC-only model, and the trivial model in the context of weather forecasts

Full forecast info, info from climate table, 50:50

- ? Give the log-likelihood of the AC-only and trivial models if there are I alternatives and N_i decisions for alternative i (total number of decisions $N = \sum_{i=1}^I N_i$)

Trivial model: $P_{ni} = 1/I$, $\tilde{L} = \sum_n \ln P_{i_n} = \sum_i N_i \ln P_i = -N \ln I$;

AC-only model: $P_{ni} = N_i/N$, $\tilde{L} = \sum_i N_i \ln P_i = N \ln N - \sum_i N_i \ln N_i$

- ? Consider a binary choice situation where the $N/2$ persons with short trips chose the pedestrian/bike option with a probability of $3/4$, and the PT/car option with $1/4$. The other $N/2$ persons with long trips had the reverse modal split with a ped/bike usage of 25%, only.

What would be the LR-index for the "perfect" model exactly reproducing the observed 3:1 and 1:3 modal splits for the short and long trips, respectively?

(less than 0.18)

Questions on GoF metrics

- ? Discuss the model to be tested, the AC-only model, and the trivial model in the context of weather forecasts

Full forecast info, info from climate table, 50:50

- ? Give the log-likelihood of the AC-only and trivial models if there are I alternatives and N_i decisions for alternative i (total number of decisions $N = \sum_{i=1}^I N_i$)

Trivial model: $P_{ni} = 1/I$, $\tilde{L} = \sum_n \ln P_{i_n} = \sum_i N_i \ln P_i = -N \ln I$;

AC-only model: $P_{ni} = N_i/N$, $\tilde{L} = \sum_i N_i \ln P_i = N \ln N - \sum_i N_i \ln N_i$

- ? Consider a binary choice situation where the $N/2$ persons with short trips chose the pedestrian/bike option with a probability of $3/4$, and the PT/car option with $1/4$. The other $N/2$ persons with long trips had the reverse modal split with a ped/bike usage of 25%, only.

What would be the LR-index for the "perfect" model exactly reproducing the observed 3:1 and 1:3 modal splits for the short and long trips, respectively?

(less than 0.18)

Questions on GoF metrics

- ? Discuss the model to be tested, the AC-only model, and the trivial model in the context of weather forecasts

Full forecast info, info from climate table, 50:50

- ? Give the log-likelihood of the AC-only and trivial models if there are I alternatives and N_i decisions for alternative i (total number of decisions $N = \sum_{i=1}^I N_i$)

Trivial model: $P_{ni} = 1/I$, $\tilde{L} = \sum_n \ln P_{i_n} = \sum_i N_i \ln P_i = -N \ln I$;

AC-only model: $P_{ni} = N_i/N$, $\tilde{L} = \sum_i N_i \ln P_i = N \ln N - \sum_i N_i \ln N_i$

- ? Consider a binary choice situation where the $N/2$ persons with short trips chose the pedestrian/bike option with a probability of $3/4$, and the PT/car option with $1/4$. The other $N/2$ persons with long trips had the reverse modal split with a ped/bike usage of 25%, only.

What would be the LR-index for the “perfect” model exactly reproducing the observed 3:1 and 1:3 modal splits for the short and long trips, respectively?

(less than 0.18)