Model estimation and outputs

Key concepts & study plan

Experimental design

Data collection & processing

Model specification & estimation

Interpretation & application

Maximum likelihood estimation

Estimation of model parameters

- Observe explanatory data, say x
  - general notation, so x can also include characteristics of decision-maker (z) and choice situation (w)
- Observe choices, say Y
  - Y_{nt} is the chosen alternative for person n in choice situation t
- Specify utility functions and select model type
- Find parameter values β that best explain the choices

Overview reference MMNL: *Train, K.A. (2009), 'Discrete Choice Methods with Simulation', ch. 8, Cambridge University Press, Cambridge, MA.*

Maximum likelihood estimation

Likelihood and log-likelihood

- N people, T_n observations for n
- Y_{nt} chosen by n in situation t
- $y_{jnt} = \begin{cases} 1 & \text{if } Y_{nt} = j \\ 0 & \text{if } Y_{nt} \neq j \end{cases}$
- β groups together all model parameters
- Likelihood given by $L(\beta)$
- Even with modest N and J, $L(\beta) \rightarrow 0$
- Instead work with log-likelihood $LL(\beta)$

$$
\begin{aligned}
L(\beta) &= \prod_{n=1}^{N} \prod_{t=1}^{T_n} \prod_{j=1}^{J} \left(P_{jnt}(\beta, x_{nt})\right)^{y_{jnt}} \\
LL(\beta) &= log\left(L(\beta)\right) \\
&= \sum_{n=1}^{N} \sum_{t=1}^{T_n} \sum_{j=1}^{J} y_{jnt} \cdot log\left(P_{jnt}(\beta, x_{nt})\right) \\
\widehat{\beta} &= \arg\max_{\beta} L(\beta) \\
&= \arg\max_{\beta} LL(\beta)
\end{aligned}
$$

Maximum likelihood estimation

Implementation

Theoretical notation

- Sum over probabilities of all alternatives, but only one has a non-zero y_{jnt}

$$\boxed{LL\left(\beta\right) = \sum_{n=1}^{N} \sum_{t=1}^{T_n} \sum_{j=1}^{J} y_{jnt} \cdot log\left(P_{jnt}\left(\beta, x_{nt}\right)\right)}$$

Implementation

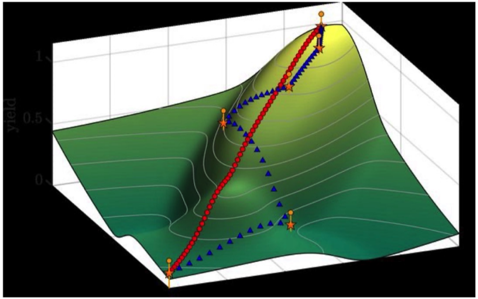
- Only need probability of chosen alternative for each observation, i.e. Y_{nt}

$$\boxed{LL\left(\beta\right) = \sum_{n=1}^{N} \sum_{t=1}^{T_n} log\left(P_{Y_{nt}}\left(\beta, x_{nt}\right)\right)}$$

Maximum likelihood estimation

Global optimum

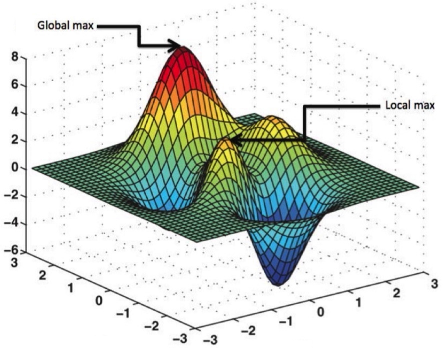
- $LL(\beta)$ for linear in parameters MNL is globally concave
- If a solution exists, it is unique



Maximum likelihood estimation

Local optima

- In more advanced models or with utility specifications that are not linear in parameters, no longer have a single global optimum
- Numerous local optima
- Starting in a *bad* location may get us trapped in one of these local optima
- Closed form choice probabilities



Maximum likelihood estimation

Outputs from model estimation

- Model estimation produces three key outputs
  - Model fit
  - Parameter estimates
  - Covariance matrix
- We now look at the interpretation of these outputs

Model fit

Key concepts & study plan

Experimental design

Data collection & processing

Model specification & estimation

Interpretation & application

Model fit

Log-likelihoood at convergence

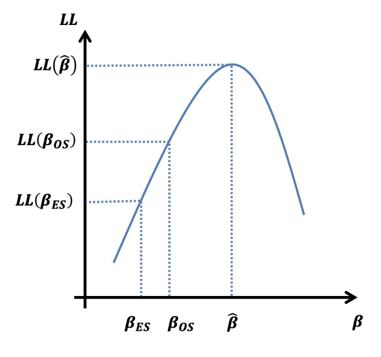
- Likelihood $L(\beta)$ shows how likely choices in our data are
  - conditional on chosen model and at parameter values β
- Classical estimation maximises log-likelihood
  - Log-likelihood $LL(\beta) = log\left[L(\beta)\right]$
- At convergence, obtain parameters values $\hat{\beta}$
- $LL\left(\hat{\beta}\right)$ used extensively during specification search

Model fit

Obtain metrics other than just $LL(\hat{\beta})$

- ❑ $LL\left(\hat{\beta}\right)$: LL at convergence (MLE)
- ❑ $LL\left(\beta_0\right)$: LL at starting values
- ❑ $LL\left(\beta_{ES}\right)$: LL at equal shares
  - random model, $P_j = \frac{1}{J}\ \forall j$
  - often written as $LL\left(0\right)$
  - often same as $LL\left(\beta_0\right)$
- ❑ $LL\left(\beta_{OS}\right)$: LL at observed shares
  - replicates aggregate shares in the data
  - helps understand how much other parameters contribute to understanding choices especially in labelled settings

Different LL measures



Model fit

Outlier detection

- $LL\left(\hat{\beta} \mid x, Y\right)$ gives LL for entire sample
- After estimation, can compute P_{nt} for each observed choice for each person
- And average across choices, say $\bar{P}_{nt} = \frac{1}{T_n}\sum_{t=1}^{T_n} P_{nt}$
- After estimation, expect $\bar{P}_{nt} > \frac{1}{J_n}$ for most people, where J_n is number of alternatives
- But clearly not for all people (that's why we have an error term)
- Outlier detection looks for people where P_n is very low
  - no specific guidance of what *very low* means

Model fit

Using information from outlier detection

- ❑ Bad idea to remove outliers from model
- ❑ They tell us that our model is struggling to explain their choices
- ❑ Use this to improve the model
- ❑ Possible findings
  - Coding or measurement errors in the data
    - Look for signs of data errors
    - Correct or remove the observation
  - Model misspecification
    - Seek clues of missing variables from the observation
    - Keep the observation and improve the model

Model fit

Model performance

- In regression, evaluate performance using e.g. R^2
- In choice modelling, focus is more on relative performance
  - how well does one model on a dataset do compared to another
- Or look at out-of-sample validation
  - compare fit on validation data to estimation data (often 20% – 80% split)
  - checks for overfitting
  - limited insight if validation data from same sample

Model fit

Hit rate (and why not to use it)

Assigns choice to highest P_j

		Model 1		Model 2	
Person	Choice	P_A	P_B	P_A	P_B
1	A	0.8	0.2	0.55	0.45
2	A	0.75	0.25	0.53	0.47
3	B	0.3	0.7	0.48	0.52
4	A	0.85	0.15	0.54	0.46
5	B	0.25	0.75	0.49	0.51
6	B	0.2	0.8	0.49	0.51
7	B	0.6	0.4	0.44	0.56
8	A	0.45	0.55	0.54	0.46
9	A	0.85	0.15	0.51	0.49
10	A	0.35	0.65	0.55	0.45
	LL	**-4.47**		**-6.32**	
	Av prob choice	**0.67**		**0.53**	
	Hit rate	**0.7**		**1**	

Why is this a bad idea?

"This statistic incorporates a notion that is opposed to the meaning of probabilities and the purpose of specifying choice probabilities. The statistic is based on the idea that the decision maker is predicted by the researcher to choose the alternative for which the model gives the highest probability."
(Kenneth Train, Discrete Choice Methods with Simulation, Cambridge University Press)

Parameter estimates

Key concepts
& study plan

Experimental
design

Data collection
& processing

Model specification
& estimation

Interpretation
& application

Parameter estimates

Model outputs

- Model estimation returns maximum likelihood estimates, or MLE, given by $\hat{\beta}$
- These are the estimates that give us the log-likelihood at convergence
- Parameter types
  1. parameters capturing impact of changes in an attribute on utility
  2. parameters relating to model structure (e.g. nesting parameters)
  3. parameters capturing socio-demographic interactions
- Our focus for now is on the first of these

Parameter estimates

Illustrative example

- Can get initial insights from signs
  - Each £ in cost loses us 0.005 units in utility
  - Each GB in memory gains us 0.004 units in utility
  - Moving from 3G to 4G and 5G increases utility
  - Highest utility for Apple ahead of Samsung & Huawei
- But what about the size of the estimates?

Mobile phone choice

Parameters	Estimate ($\hat{\beta}_k$)
$\beta_{cost,£}$	-0.005
$\beta_{memory,GB}$	0.004
β_{3G}	0
β_{4G}	0.5
β_{5G}	0.75
β_{Huawei}	0
$\beta_{Samsung}$	1
β_{Apple}	1.75

Parameter estimates

Scale matters

- Different studies have different levels of noise
- Greater noise means smaller β, and vice versa
- We thus cannot say that cost matters more in our study than in a study where $\beta_{cost,£} = -0.002$

Mobile phone choice

Parameters	Estimate ($\hat{\beta}_k$)
$\beta_{cost,£}$	-0.005
$\beta_{memory,GB}$	0.004
β_{3G}	0
β_{4G}	0.5
β_{5G}	0.75
β_{Huawei}	0
$\beta_{Samsung}$	1
β_{Apple}	1.75

Parameter estimates

Units, levels and ranges

- Cannot say cost is more important than memory
  - with continuous attributes, units matter
- Cannot say that Apple is better than 5G
  - with categorical attributes, levels and ranges matter
  - remember that only differences in utility matter, in this case differences against the base
- Also incorrect to say that brand is *more important* than download speed
  - can at best say that with the specific levels used here, brand can influence choice more than download speed

Mobile phone choice

Parameters	Estimate ($\hat{\beta}_k$)
$\beta_{cost,£}$	-0.005
$\beta_{memory,GB}$	0.004
β_{3G}	0
β_{4G}	0.5
β_{5G}	0.75
β_{Huawei}	0
$\beta_{Samsung}$	1
β_{Apple}	1.75

Parameter estimates

Marginal rate of substitution (MRS)

- Absolute values of parameters have no meaning
- Can only look at relative impacts on utility of changes in attributes
    - most common example is willingness-to-pay (WTP)
    - e.g. how much does one unit in time matter compared to one unit in cost?
- In the simplest case, this is the ratio of two coefficients
- The computation of these measures depends on the type of attribute, the utility specification and the model type
- This will be discussed in detail later in the course

Parameter estimates

Continuous attributes with a linear specification

- $V_j = \sum_{k=1}^{K} \beta_k x_{j,k}$
- Impact on utility given by partial derivatives
  - $\frac{\partial V_j}{\partial x_{j,l}} = \beta_l$ is impact of a one unit change in attribute $x_{j,l}$
- MRS: relative impact on utility of unit changes in two attributes, say $x_{j,l}$ and $x_{j,m}$
  - $MRS_{x_{j,l},x_{j,m}} = \frac{\frac{\partial V_j}{\partial x_{j,l}}}{\frac{\partial V_j}{\partial x_{j,m}}} = \frac{\beta_l}{\beta_m}$

Parameter estimates

Willingness-to-pay (WTP)

- MRS where the denominator is a monetary attribute
- Let us say $x_{j,c}$ is the cost attribute for alternative j
  - $WTP_{x_{j,l},x_{j,c}} = \frac{\frac{\partial V_j}{\partial x_{j,l}}}{\frac{\partial V_j}{\partial x_{j,c}}} = \frac{\beta_l}{\beta_c}$
- Presents monetary value of changes in attribute l
- Can relate to both willingness-to-pay (WTP) for an improvement in an attribute, or willingness-to-accept (WTA) a worse value in return for lower cost

Parameter estimates

Categorical attributes

- Can only look at changes in utility between levels, and compare this to impacts of other attributes
- Let us say $x_{j,q}$ and $x_{j,r}$ are categorical variables
  - Relative impact: $\frac{\beta_{q,\mathbf{3}}-\beta_{q,\mathbf{2}}}{\beta_{r,\mathbf{2}}-\beta_{r,\mathbf{1}}}$
  - how much does a change from the second to the third level of attribute q matter compared to a change from the first to the second level of attribute r
- Can of course combine with continuous attributes too

Parameter estimates

Results for our example

- $\frac{\beta_{memory,GB}}{\beta_{cost,£}} = -0.8$
  - increasing memory by 1GB has the same impact on utility as decreasing cost by £0.8
  - can interpret as a WTP of £0.8 per GB
- $\frac{\beta_{Apple}-\beta_{Samsung}}{\beta_{5G}-\beta_{3G}} = 1$
  - Going from Samsung to Apple is as valuable as going from 3G to 5G
- $\frac{\beta_{Apple}-\beta_{Huawei}}{\beta_{cost,£}} = -£350$
  - Going from Huawei to Apple is the same as reducing cost by £350
  - WTP of £350 for going from Huawei to Apple (or reverse need for compensation)

Parameters	Estimate ($\hat{\beta}_k$)
$\beta_{cost,£}$	-0.005
$\beta_{memory,GB}$	0.004
β_{3G}	0
β_{4G}	0.5
β_{5G}	0.75
β_{Huawei}	0
$\beta_{Samsung}$	1
β_{Apple}	1.75

Parameter estimates

Discussion

- MRS calculations are the key output for interpretation of random utility models
- Used universally across disciplines (unlike some other metrics)
- Computation here has looked only at linear-in-attributes specifications
  - non-linearity adds complexity
  - as does heterogeneity
- Meaning of MRS is different for non-RUM models as value functions (e.g. regret) are context dependent

Covariance matrix

Key concepts & study plan

Experimental design

Data collection & processing

Model specification & estimation

Interpretation & application

Covariance matrix

Covariance matrix and standard errors

- Estimation gives us estimates ($\hat{\beta}$) and asymptotic variance-covariance matrix Ω
- Our key interest is in standard errors (σ)
  - given by square root of diagonal elements of covariance matrix
  - expression of precision of parameter estimates
  - a smaller standard error means we can be surer about the estimated value
- Can also calculate standard errors for functions of parameters (e.g. MRS) using Delta method
- We use standard errors for asymptotic confidence intervals and for statistical tests

Covariance matrix and standard errors

$$\Omega = I\left(\beta\right)^{-1}$$

$$I\left(\beta\right) = -E\left(H\left(\beta\right)\right)$$

$$H\left(\beta\right) = \left(\frac{\partial^2 LL\left(\beta\right)}{\partial\beta\partial\beta'}\right)$$

$$\sigma_{\beta_k} = \sqrt{\Omega_{k,k}}$$

Covariance matrix

Standard errors and sample size

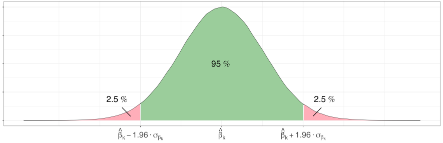
- β_k is one parameter in model
- True value given by β_k^*
- Maximum likelihood estimate (MLE): $\widehat{\beta}_k$
- Incomplete data leads to sampling error
- Asymptotic normality:

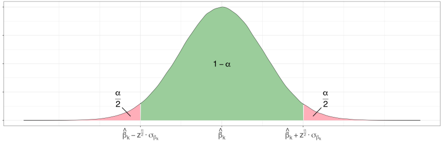
$$\boxed{\sqrt{N}\left(\widehat{\beta} - \beta^*\right) \rightarrow \mathcal{N}\left(0, \Omega\right)}$$

Covariance matrix

Confidence intervals

- ❑ With new data, estimates would change
- ❑ Often report 95% confidence intervals
  - 95% chance that the true value for a parameter lies in that range
  - with smaller standard errors, CIs will be narrower
- ❑ Calculation of CIs uses estimated value, standard error, and critical value from a $N(0,1)$ distribution
  - e.g. for 95%, we use $\boxed{\widehat{\beta}_k \pm 1.96\sigma_{\beta_k}}$





CI	α	$\frac{\alpha}{2}$	$z^{\frac{\alpha}{2}}$
99%	0.01	0.005	2.57
95%	0.05	0.025	1.96
90%	0.1	0.05	1.64

Covariance matrix

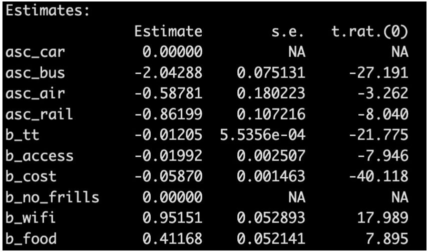
Statistical tests: t-ratios

- Used (typically) to test whether parameter is different from zero ($H_0 : \beta_k = 0$)

- Test statistic: $\boxed{t_{\hat{\beta}_k} = \frac{\hat{\beta}_k - 0}{\sigma_{\beta_k}}}$

- We compare the test statistic to a critical value and see if it exceeds it
  - e.g. 1.96 for a 95% two-sided test, or 1.64 for a one-sided test

confidence level	1 sided critical value	2 sided critical value
99%	2.33	2.58
95%	1.64	1.96
90%	1.28	1.64



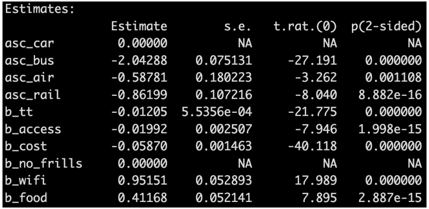
```
Estimates:
               Estimate        s.e.    t.rat.(0)
asc_car         0.00000          NA          NA
asc_bus        -2.04288    0.075131     -27.191
asc_air        -0.58781    0.180223      -3.262
asc_rail       -0.86199    0.107216      -8.040
b_tt           -0.01205  5.5356e-04     -21.775
b_access       -0.01992    0.002507      -7.946
b_cost         -0.05870    0.001463     -40.118
b_no_frills     0.00000          NA          NA
b_wifi          0.95151    0.052893      17.989
b_food          0.41168    0.052141       7.895
```

Covariance matrix

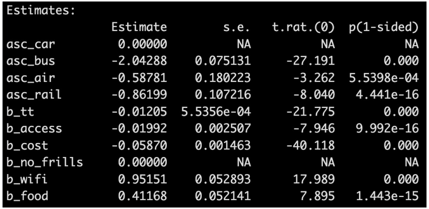
Statistical tests: t-ratios and p-values

- Can compute p-value (probability that our result was obtained by chance if H_0 is true)
- Should always report either standard error or t-ratio alongside *p*-value
- And state whether one-sided or two-sided

t-ratio	p-value (1 sided)	p-value (2 sided)
2.58	0.005	0.01
2.33	0.01	0.02
1.96	0.025	0.05
1.64	0.05	0.1
1.28	0.1	0.2



```
Estimates:
              Estimate        s.e.   t.rat.(0)  p(2-sided)
asc_car        0.00000          NA          NA          NA
asc_bus       -2.04288    0.075131     -27.191    0.000000
asc_air       -0.58781    0.180223      -3.262    0.001108
asc_rail      -0.86199    0.107216      -8.040   8.882e-16
b_tt          -0.01205  5.5356e-04     -21.775    0.000000
b_access      -0.01992    0.002507      -7.946   1.998e-15
b_cost        -0.05870    0.001463     -40.118    0.000000
b_no_frills    0.00000          NA          NA          NA
b_wifi         0.95151    0.052893      17.989    0.000000
b_food         0.41168    0.052141       7.895   2.887e-15
```

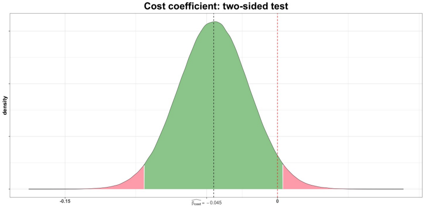


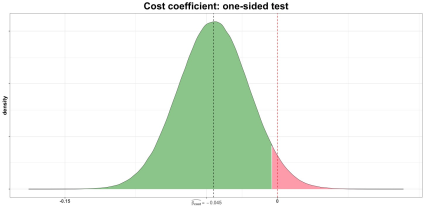
```
Estimates:
              Estimate        s.e.   t.rat.(0)  p(1-sided)
asc_car        0.00000          NA          NA          NA
asc_bus       -2.04288    0.075131     -27.191       0.000
asc_air       -0.58781    0.180223      -3.262  5.5398e-04
asc_rail      -0.86199    0.107216      -8.040   4.441e-16
b_tt          -0.01205  5.5356e-04     -21.775       0.000
b_access      -0.01992    0.002507      -7.946   9.992e-16
b_cost        -0.05870    0.001463     -40.118       0.000
b_no_frills    0.00000          NA          NA          NA
b_wifi         0.95151    0.052893      17.989       0.000
b_food         0.41168    0.052141       7.895   1.443e-15
```

Covariance matrix

One-sided or two-sided tests

- ❑ $\widehat{\beta}_{cost} = -0.045$
- ❑ $\sigma_{\beta_{cost}} = 0.025$
- ❑ $t_{\widehat{\beta}_{cost}} = -1.8$
- ❑ Typical critical value: 1.96
- ❑ Using 1.96 makes sense in absence of sign assumptions
- ❑ But often we know the sign
- ❑ Consider using one-sided tests
  - otherwise we would say that very negative values are also unacceptable





Covariance matrix

Recap: classical standard errors

- ❑ Classical covariance matrix Ω is a conservative estimate of sampling error
- ❑ Relies on number of assumptions
  - model is correct
  - no unmodelled correlation across choices

Classical covariance matrix

$$\Omega = I\left(\beta\right)^{-1}$$

$$I\left(\beta\right) = -E\left(H\left(\beta\right)\right)$$

$$H\left(\beta\right) = \left(\frac{\partial^2 LL\left(\beta\right)}{\partial\beta\partial\beta'}\right)$$

Covariance matrix

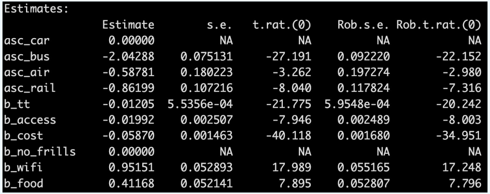
Robust covariance matrix

- Given by sandwich estimator involving the BHHH matrix (B)
- Robust se tend to be larger
- Using LL at person level in BHHH gives different robust standard errors from working at observation level
  - T observations each from N people is not the same as 1 observation each from NT people

Robust covariance matrix

$$\Omega_{robust} = \Omega\, B\, \Omega$$

$$B_{jk} = \sum_{n}^{N} \frac{\partial LL_n(\beta)}{\partial \beta_j} \frac{\partial LL_n(\beta)}{\partial \beta_k}$$



```
Estimates:
              Estimate         s.e.   t.rat.(0)    Rob.s.e. Rob.t.rat.(0)
asc_car        0.00000           NA          NA          NA            NA
asc_bus       -2.04288     0.075131     -27.191    0.092220       -22.152
asc_air       -0.58781     0.180223      -3.262    0.197274        -2.980
asc_rail      -0.86199     0.107216      -8.040    0.117824        -7.316
b_tt          -0.01205   5.5356e-04     -21.775  5.9548e-04       -20.242
b_access      -0.01992     0.002507      -7.946    0.002489        -8.003
b_cost        -0.05870     0.001463     -40.118    0.001680       -34.951
b_no_frills    0.00000           NA          NA          NA            NA
b_wifi         0.95151     0.052893      17.989    0.055165        17.248
b_food         0.41168     0.052141       7.895    0.052807         7.796
```

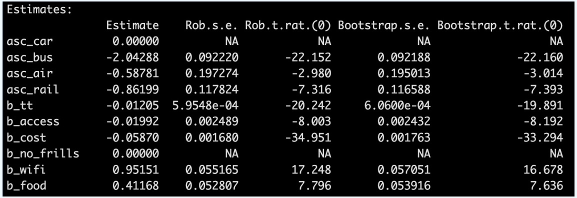
Covariance matrix

Bootstrapping uses fewest assumptions

- Works by repeated sampling
- Approach with least assumptions
- But computationally most demanding

Bootstrapped covariance matrix

1. Draw S versions of data with replacement
2. S sets of β: $\beta_s = \langle \beta_{1,s}, \ldots, \beta_{K,s} \rangle$, $s = 1, \ldots, S$
3. $\frac{\sum_{s=1}^{S} \beta_{k,s}}{S} \rightarrow \widehat{\beta}_k =$ with large S
4. $\Omega_{bootstrap} = var\left(\beta\right)$

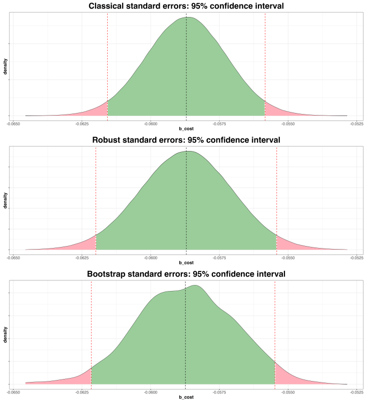


```
Estimates:
                 Estimate    Rob.s.e. Rob.t.rat.(0) Bootstrap.s.e. Bootstrap.t.rat.(0)
asc_car           0.00000          NA            NA             NA                  NA
asc_bus          -2.04288    0.092220       -22.152       0.092188             -22.160
asc_air          -0.58781    0.197274        -2.980       0.195013              -3.014
asc_rail         -0.86199    0.117824        -7.316       0.116588              -7.393
b_tt             -0.01205  5.9548e-04       -20.242     6.0600e-04             -19.891
b_access         -0.01992    0.002489        -8.003       0.002432              -8.192
b_cost           -0.05870    0.001680       -34.951       0.001763             -33.294
b_no_frills       0.00000          NA            NA             NA                  NA
b_wifi            0.95151    0.055165        17.248       0.057051              16.678
b_food            0.41168    0.052807         7.796       0.053916               7.636
```

Covariance matrix

Confidence interval comparison

- Asymptotic CI for classical and robust
  - symmetrical by definition
  - wider with robust se
- With bootstrapping, can look at empirical CI
  - No longer necessarily symmetrical
  - lower limit in this case 5.1% further from mean than upper limit


Classical standard errors: 95% confidence interval
density
b_cost
Robust standard errors: 95% confidence interval
density
b_cost
Bootstrap standard errors: 95% confidence interval
density
b_cost