Computing and reporting confidence metrics

Key concepts & study plan

Experimental design

Data collection & processing

Model specification & estimation

Interpretation & application

Computing and reporting confidence metrics

Topics covered

- Standard errors for derived measures
- Signs of over-specification
- Reporting measures of confidence

The Delta method

Key concepts
& study plan

Experimental
design

Data collection
& processing

Model specification
& estimation

**Interpretation
& application**

The Delta method

Standard errors for derived measures

- Obtain estimates and standard errors for β
- Key interest is in functions of individual elements of β
  - MRS and WTP
  - difference between two parameters
  - demand forecasts and elasticities
  - welfare measures
  - moments of distributions
  - correlation between randomly distributed coefficients
- Need standard errors for derived quantities

```
Estimates:
        Estimate        s.e.   t.rat.(0)    Rob.s.e. Rob.t.rat.(0)
b_tt    -0.05977    0.004257     -14.040    0.006742        -8.865
b_tc    -0.13182    0.013506      -9.760    0.023638        -5.576
b_hw    -0.03745    0.001848     -20.269    0.002317       -16.161
b_ch    -1.15207    0.043419     -26.534    0.061373       -18.772

> diff_tt_hw=-0.05977-(-0.03745)
> diff_tt_hw
[1] -0.02232
> VTT_per_hour=60*(-0.05977/-0.13182)
> VTT_per_hour
[1] 27.20528
```

The Delta method

The Delta method

- Delta method is a first-derivative calculation
- Often described as an approximation
- Shown to be exact rather than an approximation by Daly et al. (2012)

Delta method calculations

- Let Φ be a function of β
- Estimates $\hat{\beta}$ and AVC matrix Ω
- $cov\left(\Phi\right) = \Phi'^{\,T}\Omega\Phi'$
- Φ' gives first derivatives of Φ against β

Key reference: *Daly, A.J., Hess, S. & de Jong, G. (2012), Calculating errors for measures derived from choice modelling estimates, Transportation Research Part B 46(2), pp. 333-341.*

The Delta method

Examples: difference and ratio

- Difference: $\Phi = \beta_1 - \beta_2$
  - i.e. $\phi'_1 = 1$ and $\phi'_2 = -1$
  - and $var\left(\beta_1 - \beta_2\right) = \omega_{11} + \omega_{22} - 2\,\omega_{12}$
- Ratio: $\Phi = \frac{\beta_1}{\beta_2}$
  - i.e. $\phi'_1 = \frac{1}{\beta_2}$ and $\phi'_2 = -\frac{\beta_1}{\beta_2^2}$
  - and

$$var\left(\frac{\beta_1}{\beta_2}\right) = \left(\frac{\beta_1}{\beta_2}\right)^2 \left(\frac{\omega_{11}}{\beta_1^2} + \frac{\omega_{22}}{\beta_2^2} - 2\,\frac{\omega_{12}}{\beta_1\beta_2}\right)$$

```
Estimates:
        Estimate        s.e.   t.rat.(0)    Rob.s.e. Rob.t.rat.(0)
b_tt    -0.05977    0.004257     -14.040    0.006742        -8.865
b_tc    -0.13182    0.013506      -9.760    0.023638        -5.576
b_hw    -0.03745    0.001848     -20.269    0.002317       -16.161
b_ch    -1.15207    0.043419     -26.534    0.061373       -18.772
```

```
> model$robvarcov
             b_tt         b_tc         b_hw         b_ch
b_tt 4.545565e-05 1.176310e-04 2.920627e-06 1.379524e-04
b_tc 1.176310e-04 5.587382e-04 6.516358e-06 3.154629e-04
b_hw 2.920627e-06 6.516358e-06 5.370123e-06 4.961005e-05
b_ch 1.379524e-04 3.154629e-04 4.961005e-05 3.766619e-03
```

```
> diff_tt_hw=-0.05977-(-0.03745)
> se_diff_tt_hw=sqrt(4.545565e-05+5.370123e-06-2*2.920627e-06)
> diff_tt_hw
[1] -0.02232
> se_diff_tt_hw
[1] 0.00670705
> VTT_per_hour=60*(-0.05977/-0.13182)
> se_VTT_per_hour=sqrt((VTT_per_hour)^2*(4.545565e-05/((-0.05977)^2)+5.58
7382e-04/((-0.13182)^2)-2*1.176310e-04/((-0.05977)*(-0.13182))))
> VTT_per_hour
[1] 27.20528
> se_VTT_per_hour
[1] 3.334052
```

```
> deltaMethod_settings=list(expression=c(diff_tt_hw="b_tt-b_hw",
+                                        VTT_per_hour="60*b_tt/b_tc"))
> apollo_deltaMethod(model, deltaMethod_settings)
Running Delta method computation for user-defined function:

   Expression   Value Robust s.e. Rob t-ratio (0)
   diff_tt_hw -0.0223      0.0067           -3.33
 VTT_per_hour 27.2065      3.3343            8.16
```

Signs of over-specification

Key concepts & study plan

Experimental design

Data collection & processing

Model specification & estimation

Interpretation & application

Signs of over-specification

Why am I getting `Inf` or `NaN` for standard errors?

- Theoretical identification issues
  - e.g. missing normalisation for ASCs
- Empirical identification issues
  - e.g. parameters going towards $-$inf or $+$inf, if one group of people never or always chooses a given option
- Calculation of numerical derivatives could lead to some zero probabilities
  - use analytical derivatives, and if not possible, use bootstrapping

```
Estimates:
               Estimate        s.e.   t.rat.(0)    Rob.s.e. Rob.t.rat.(0)
asc_car         0.73984          NA          NA          NA            NA
asc_bus        -1.30304          NA          NA          NA            NA
asc_air         0.41869          NA          NA          NA            NA
asc_rail        0.14451          NA          NA          NA            NA
b_tt           -0.01205          NA          NA          NA            NA
b_access       -0.01992          NA          NA          NA            NA
b_cost         -0.05870          NA          NA          NA            NA
b_no_frills    -0.26666          NA          NA          NA            NA
b_wifi          0.68484          NA          NA          NA            NA
b_food          0.14502          NA          NA          NA            NA
```

```
Unconstrained optimisation.

These outputs have had the scaling used in estimation applied to them.
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
```

Reporting

Key concepts
& study plan

Experimental
design

Data collection
& processing

Model specification
& estimation

**Interpretation
& application**

Reporting

How should significance be reported?

- Minimum output should be estimates and standard errors, or estimates and t-ratios, as s.e. can be calculated from t-ratios
- Common practice in some fields to report estimates and p-values only
  - This is bad practice, for two reasons
  - p-values imply an analyst decision on whether a one-sided or two-sided test is used, and this is often not reported
  - p-values are often reported with a numerical precision that prevents an analyst from recovering standard errors (e.g. $p < 0.001$)
- Even worse is the reliance on $*$ measures in some fields, e.g. using * for 90% confidence, ** for 95% confidence and *** for 99% confidence
  - The same issues apply as for p-values, but they are further compounded by the fact that e.g. $***$ could mean a t-ratio of 4 or 40
- p-values and $*$ measures should never replace s.e. or t-ratios

Reporting

Recommendations

- Wasserstein et al. (2019) conclude *"that it is time to stop using the term 'statistically significant' entirely. Nor should variants such as 'significantly different', 'p<0.05', and 'nonsignificant' survive, whether expressed in words, by asterisks in a table, or in some other way."*
- And *"[analysts should not] believe that an association or effect exists just because it was statistically significant [or] that an association or effect is absent just because it was not statistically significant."*

Wasserstein, R.L., Schirm, A.L., Lazar, N.A. (2019), Moving to a world beyond "p<0.05". The American Statistician 73, 1–19.

Reporting

Recommendations (continued)

- In health, "*clinical significance*" measures whether a treatment has noticeable effect on health outcomes. Choice modellers may wish to consider "*behavioural significance*", i.e. does a parameter change predictions and "*policy significance*", and does it have a significant impact on outcome of any process using the results
- Finally, note that removing a parameter that is "*not significant*" may have undesirable impact on other parameters
  - useful approximation to say that removal of parameter 1 will change parameter 2 by $-t_1 * \frac{r_{12}}{t_2}$, where t are the respective t-ratios and r_{12} is the correlation