Transformations for the EBP in emdi

Transformations for the EBP in emdi

The R package emdi allows a range of data transformations for the function ebp to get domain specific indicators obtained by Empirical Best Prediction (EBP). Since the relies on the normality assumption for the error terms transformations may help to achieve the normality.

With emdi version XX, the following options for the transformation argument in function ebp will be available:

- `no`: No transformation
- `log`: Log transformation with a deterministic shift
- `box.cox`: Box-Cox transformation with a deterministic shift
- `dual`: Dual transformation with a deterministic shift
- `log.shift`: Log transformation with an optimized shift

While the log transformation does not rely on a transformation parameter, the Box-Cox, Dual and Log-shift transformation depend on a transformation parameter lambda that can be estimated from the data to find the optimal transformation parameter. The estimation approach provided in emdi is the restricted maximum likelihood following Gurka (2006).

A comparison of the various data-driven transformations in the EBP, can be found in Rojas et al (2019).

```
# Install the package
library(emdi)
```

```
##
## Attaching package: 'emdi'
```

```
## The following object is masked from 'package:stats':
##
##     step
```

```
# Load sample data set
data("eusilcA_smp")
data('eusilcA_pop')
```

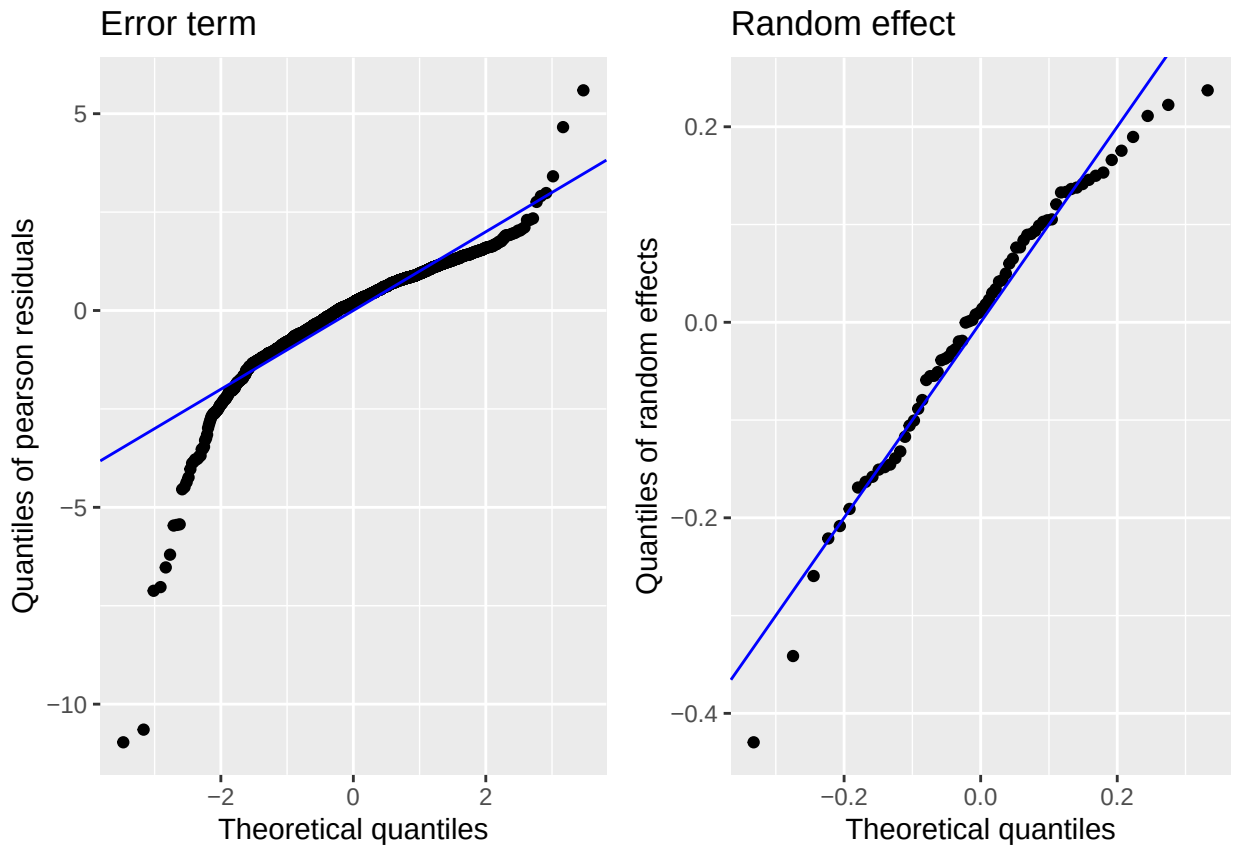
Transformation without transformation parameter - Log transformation

The log transformation does not depend on a transformation parameter but the vector of the dependent variable is shifted to the positive range by a deterministic shift.

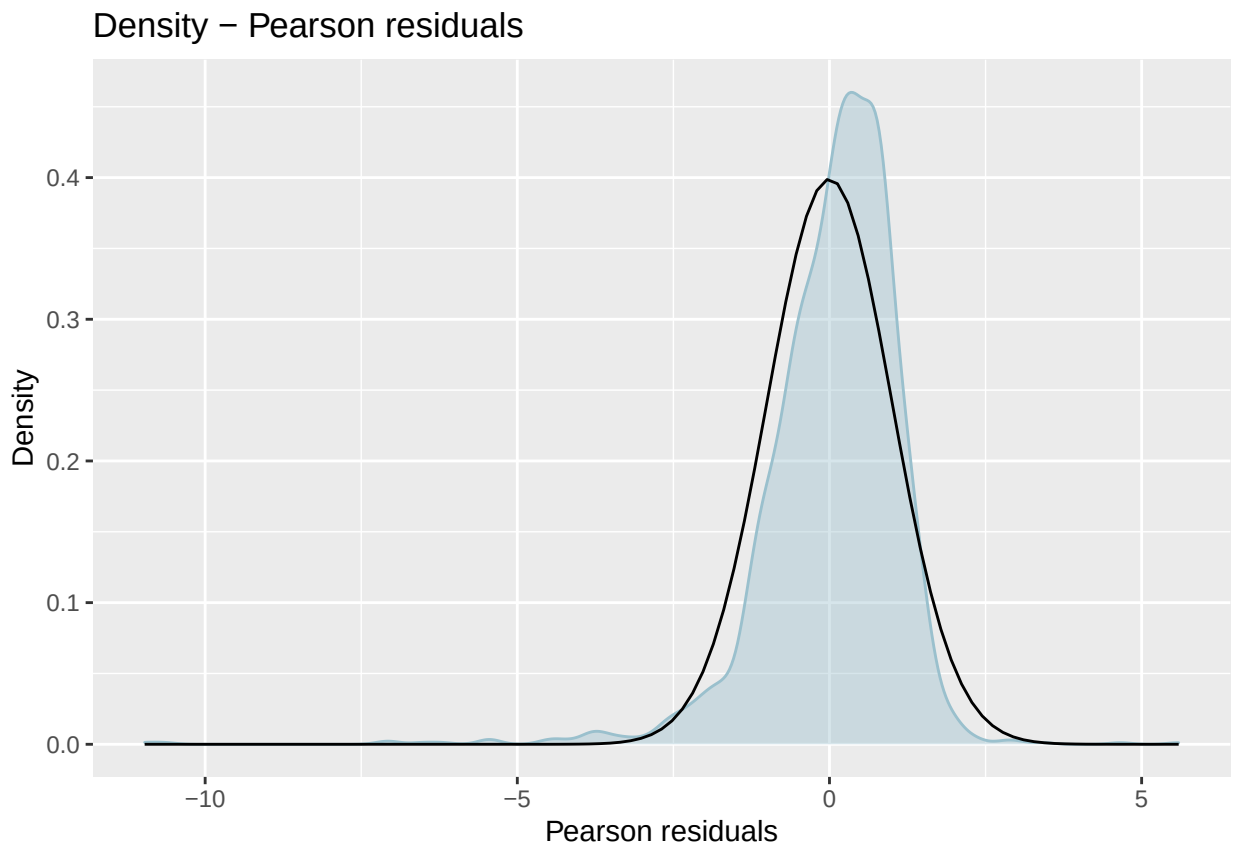
```
ebp_log <- ebp(fixed = eqIncome ~ gender + eqsize + cash + self_empl +
                    unempl_ben + age_ben + surv_ben + sick_ben + dis_ben + rent +
                    fam_allow + house_allow + cap_inv + tax_adj,
               pop_data = eusilcA_pop, pop_domains = "district",
               smp_data = eusilcA_smp, smp_domains = "district",
               threshold = 10885.33, MSE = FALSE,
               transformation = 'log', interval = 'default')
summary(ebp_log)
```

```
## Empirical Best Prediction
##
## Call:
##  ebp(fixed = eqIncome ~ gender + eqsize + cash + self_empl + unempl_ben +
##     age_ben + surv_ben + sick_ben + dis_ben + rent + fam_allow +
##     house_allow + cap_inv + tax_adj, pop_data = eusilcA_pop,
##     pop_domains = "district", smp_data = eusilcA_smp, smp_domains = "district",
##     threshold = 10885.33, transformation = "log", interval = "default",
##     MSE = FALSE)
##
## Out-of-sample domains:  24
## In-sample domains:  70
##
## Sample sizes:
## Units in sample:  1945
## Units in population:  25000
##                     Min. 1st Qu. Median      Mean 3rd Qu. Max.
## Sample_domains       14    17.0   22.5  27.78571   29.00  200
## Population_domains    5   126.5  181.5 265.95745  265.75 5857
##
## Explanatory measures:
##  Marginal_R2 Conditional_R2
##    0.5022296      0.5909727
##
## Residual diagnostics:
##                 Skewness  Kurtosis Shapiro_W    Shapiro_p
## Error         -2.1828119 17.863231 0.8670156 8.641339e-38
## Random_effect -0.6609709  3.361441 0.9682563 7.261244e-02
##
## ICC:  0.1782811
##
## Transformation:
##  Transformation Shift_parameter
##             log               0
```

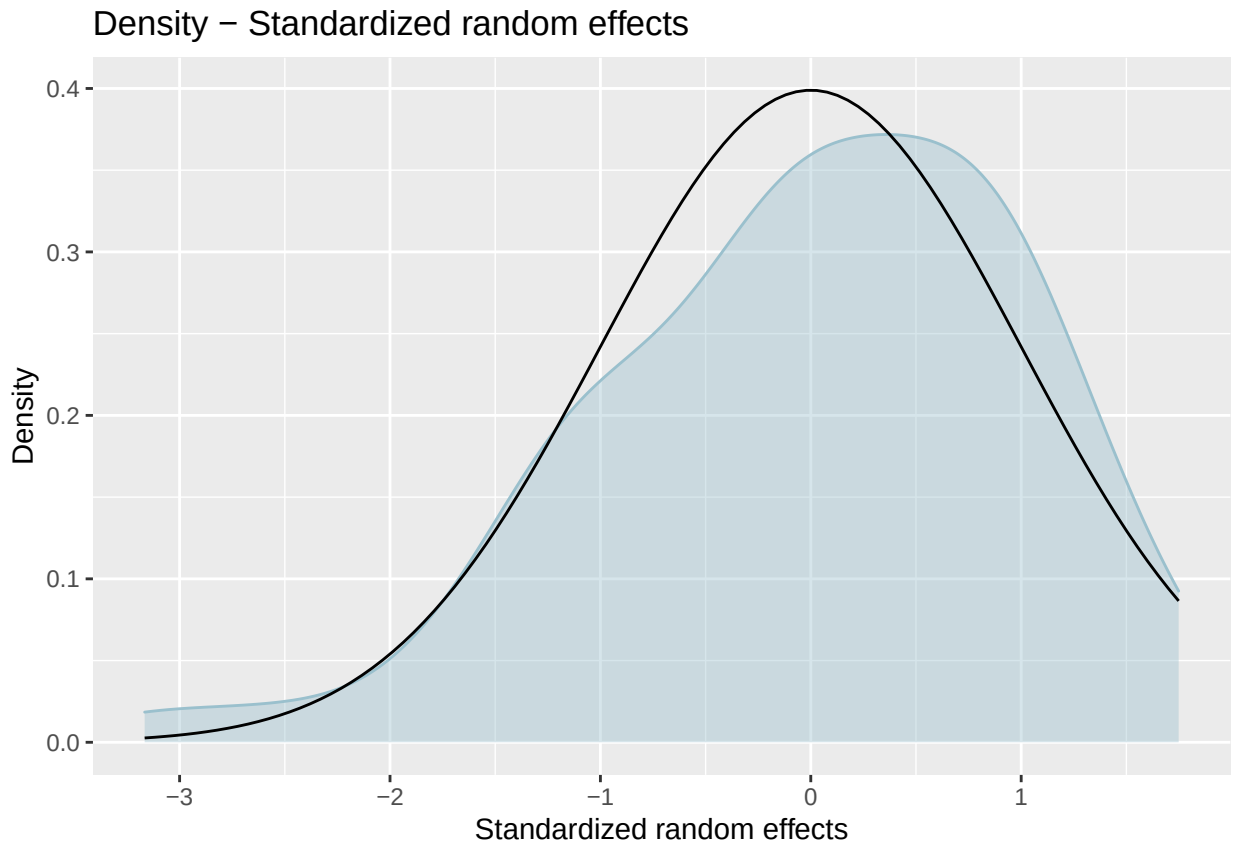
```
plot(ebp_log)
```



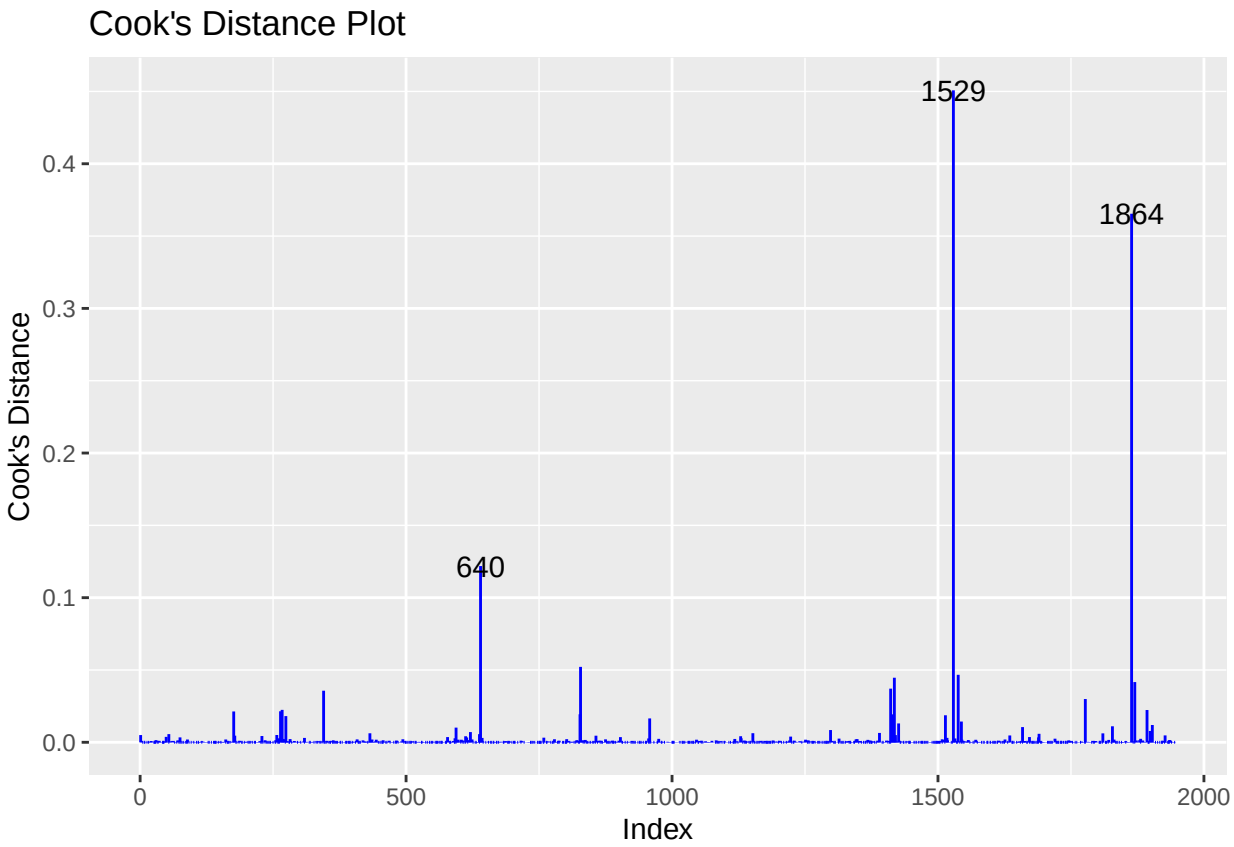

```
## Press [enter] to continue
```




```
## Press [enter] to continue
```




```
## Press [enter] to continue
```


Cook's Distance Plot
1529
1864
640
0.4
0.3
0.2
0.1
0.0
Cook's Distance
0
500
1000
1500
2000
Index