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).

```
## Installing package into 'C:/Users/Ann-Kristin/Documents/R/win-library/4.0'
## (as 'lib' is unspecified)
```

```
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.

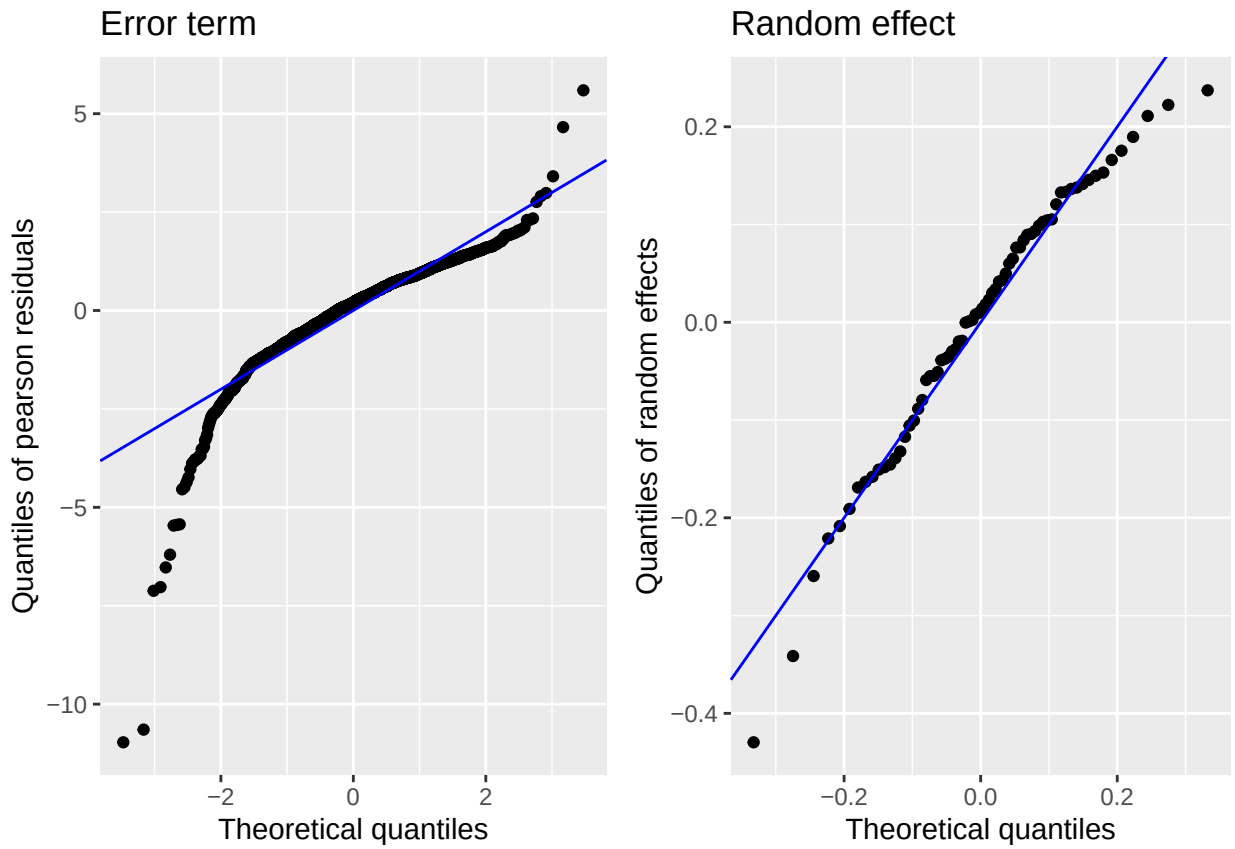
```
ebp_log <- ebp(fixed = eqIncome ~ gender + eqsize + cash + self_empl +
                   unempl_ben + age_ben + surv_ben + sick_ben + dis_ben + rent +
                   fam_allow + house_allow + cap_inv + tax_adj,
               pop_data = eusilcA_pop, pop_domains = "district",
```

```
               smp_data = eusilcA_smp, smp_domains = "district",
               threshold = 10885.33, MSE = FALSE,
               transformation = 'log')
summary(ebp_log)
```

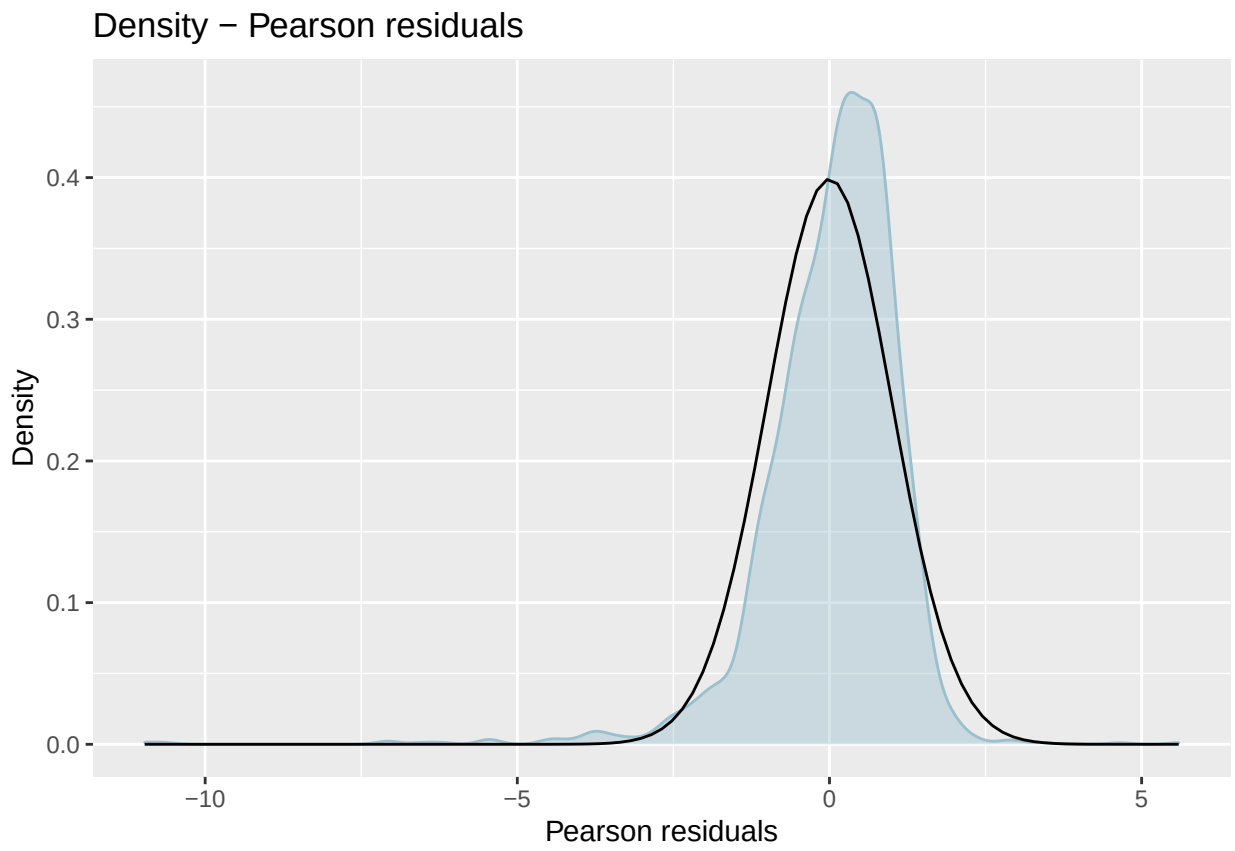
```
## Empirical Best Prediction
##
## Call:
##  ebp(fixed = eqIncome ~ gender + eqsize + cash + self_empl + unempl_ben +
##     age_ben + surv_ben + sick_ben + dis_ben + rent + fam_allow +
##     house_allow + cap_inv + tax_adj, pop_data = eusilcA_pop,
##     pop_domains = "district", smp_data = eusilcA_smp, smp_domains = "district",
##     threshold = 10885.33, transformation = "log", 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
```

The transformation is log with a zero shift since there are no negative values in the dependent variable in this example. The Shapiro-Wilk test rejects normality for both error terms. Additional to the tests shown in the summary, the plot method can be used to assess the normality of the error terms.

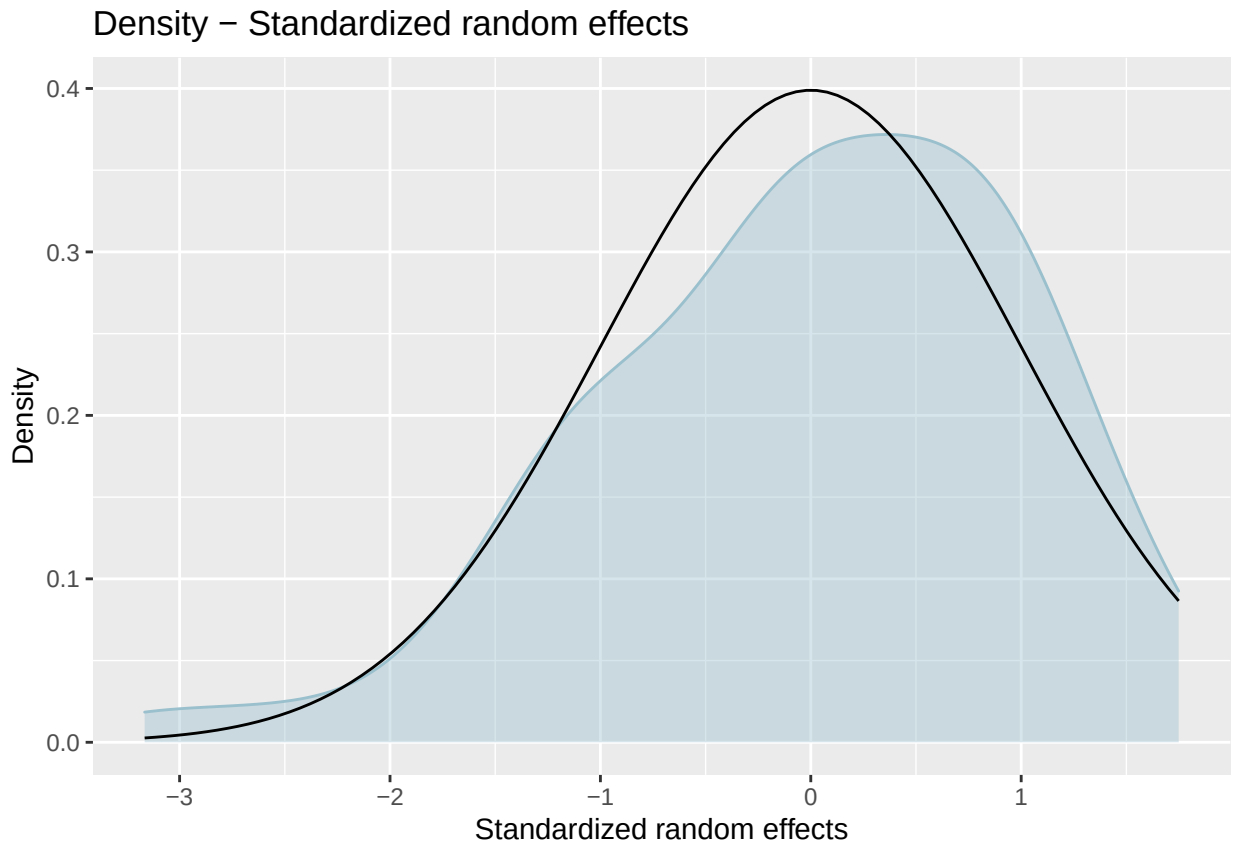
```
plot(ebp_log)
```



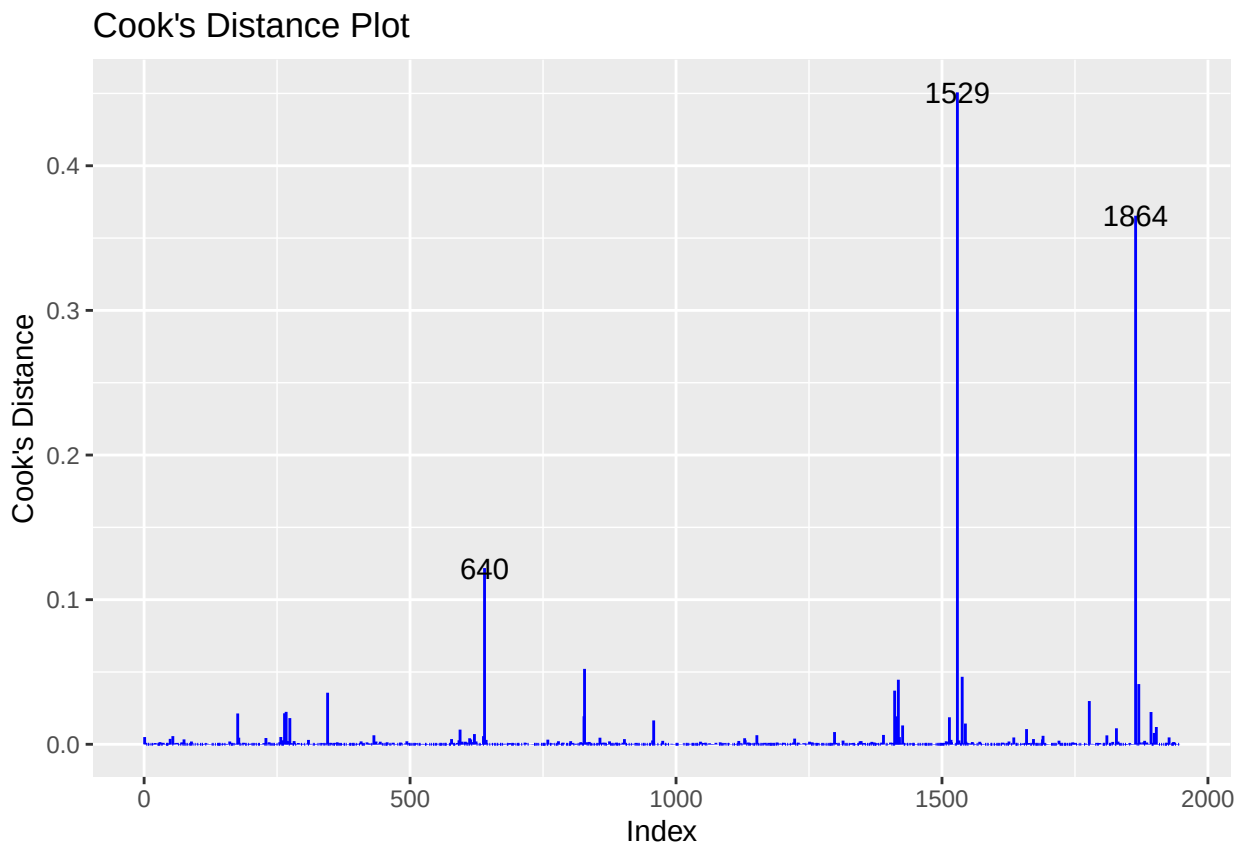

```
## Press [enter] to continue
```




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




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



Transformation with transformation parameter

Box-Cox transformation

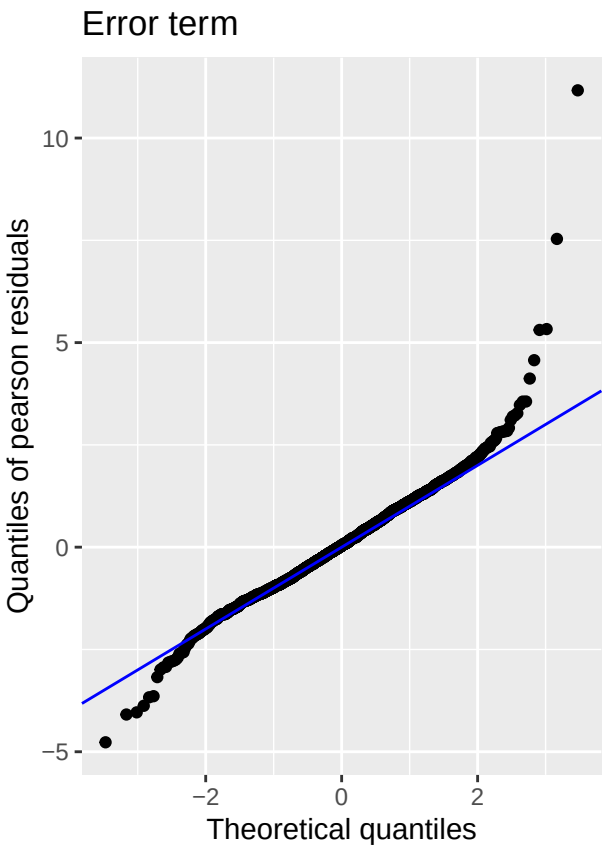
The Box-Cox transformation depends on one transformation parameter and is only defined for positive y. Therefore, a deterministic shift first shifts the dependent variable to the positive range and then the transformation is applied. For the estimation of the transformation parameter, an interval for the optimization needs to be defined. In emdi, a default option can be chosen which equals an interval between -1 and 2. This interval will be reasonable for many applications but it can happen that the interval needs to be adjusted. This can be done with a numeric vector of length two defining the lower and upper limit of the interval, e.g. c(-1, 2) for the default interval.

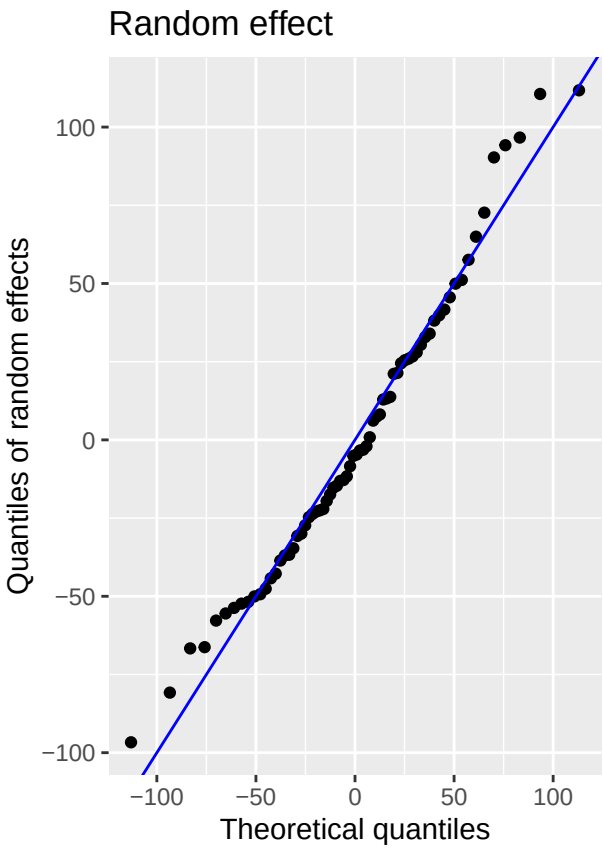
```
ebp_bc <- ebp(fixed = eqIncome ~ gender + eqsize + cash + self_empl +
                    unempl_ben + age_ben + surv_ben + sick_ben + dis_ben + rent +
                    fam_allow + house_allow + cap_inv + tax_adj,
              pop_data = eusilcA_pop, pop_domains = "district",
              smp_data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

```
## Empirical Best Prediction
##
## Call:
##  ebp(fixed = eqIncome ~ gender + eqsize + cash + self_empl + unempl_ben +
```

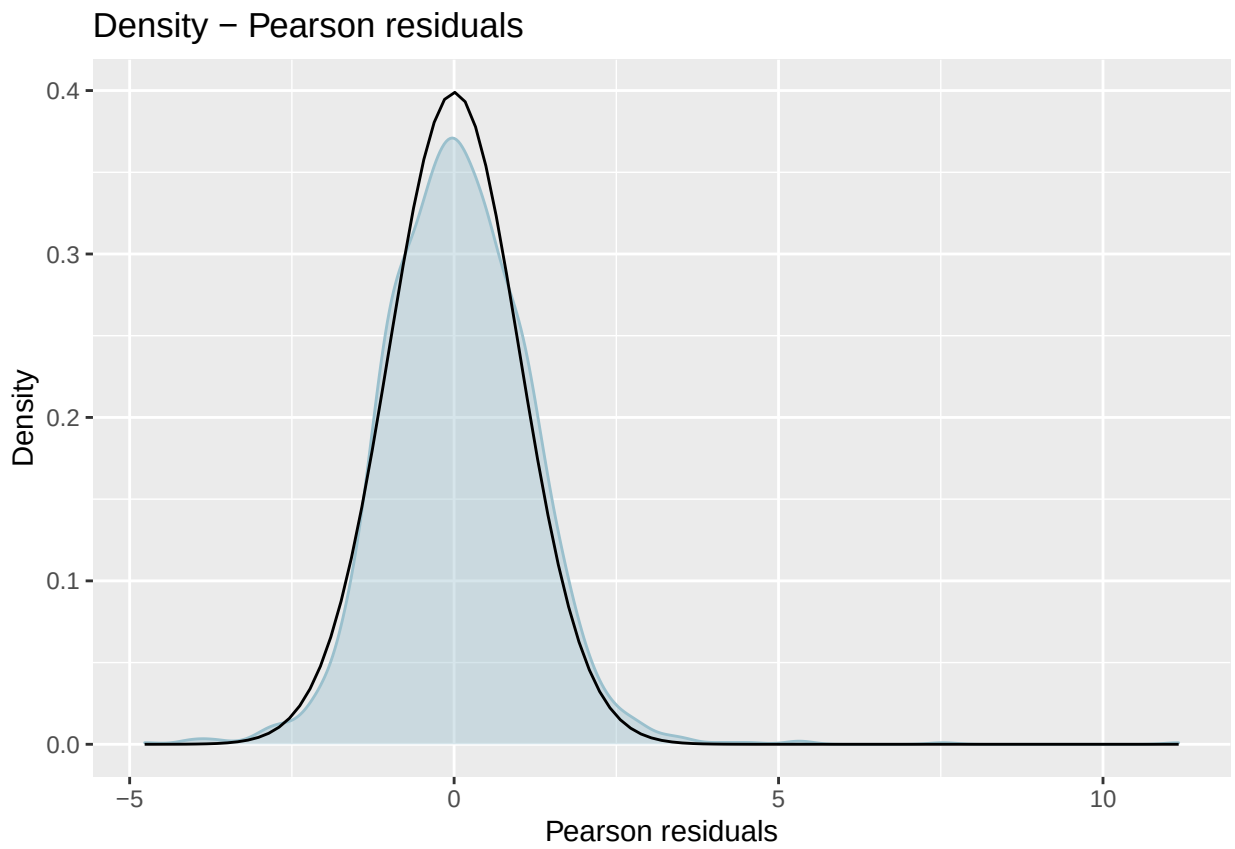
```
##     age_ben + surv_ben + sick_ben + dis_ben + rent + fam_allow +
##     house_allow + cap_inv + tax_adj, pop_data = eusilcA_pop,
##     pop_domains = "district", smp_data = eusilcA_smp, smp_domains = "district",
##     threshold = 10885.33, transformation = "box.cox", interval = "default",
##     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
##    0.6325942       0.709266
##
## Residual diagnostics:
##                Skewness Kurtosis Shapiro_W    Shapiro_p
## Error         0.7523871 9.646993 0.9619824 3.492626e-22
## Random_effect 0.4655324 2.837176 0.9760574 1.995328e-01
##
## ICC:  0.2086841
##
## Transformation:
##  Transformation Method Optimal_lambda Shift_parameter
##         box.cox   reml      0.6046901               0
```

```
plot(ebp_bc)
```

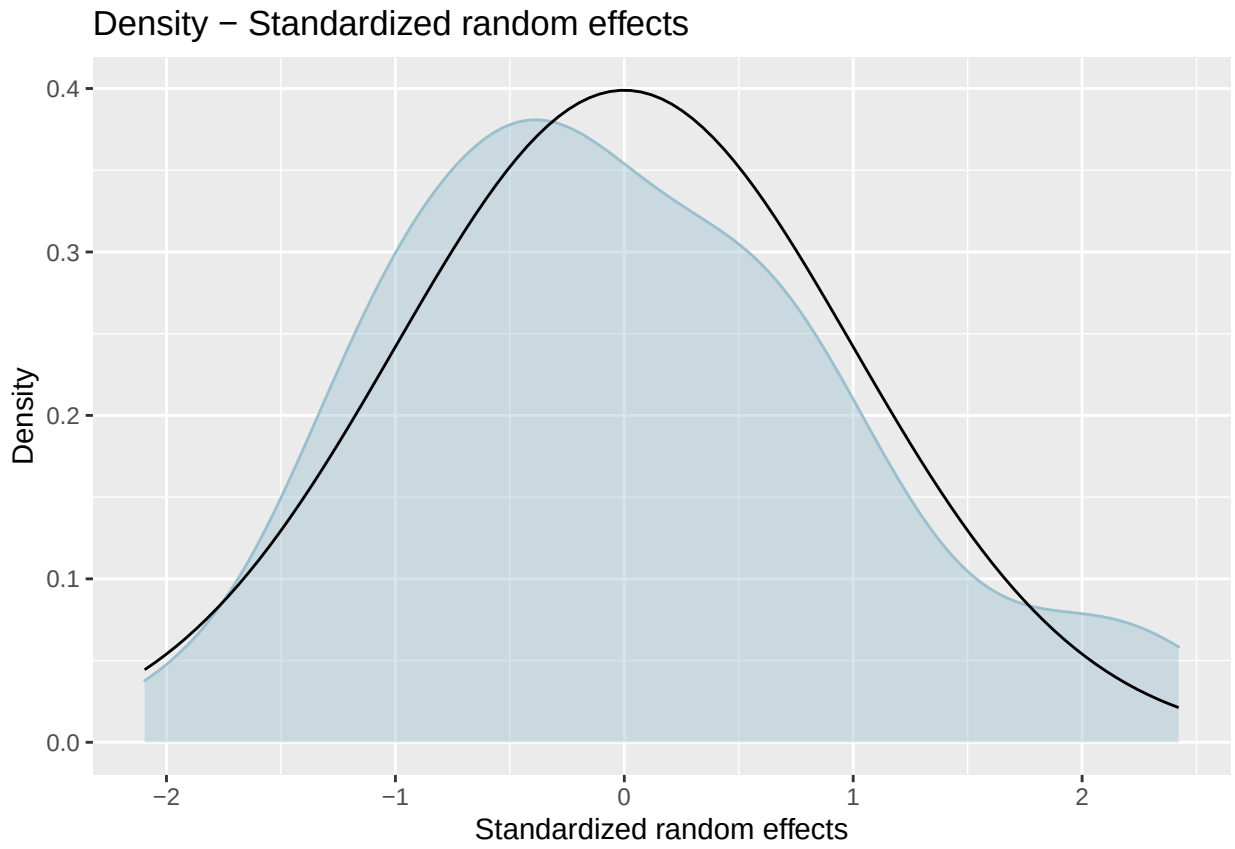




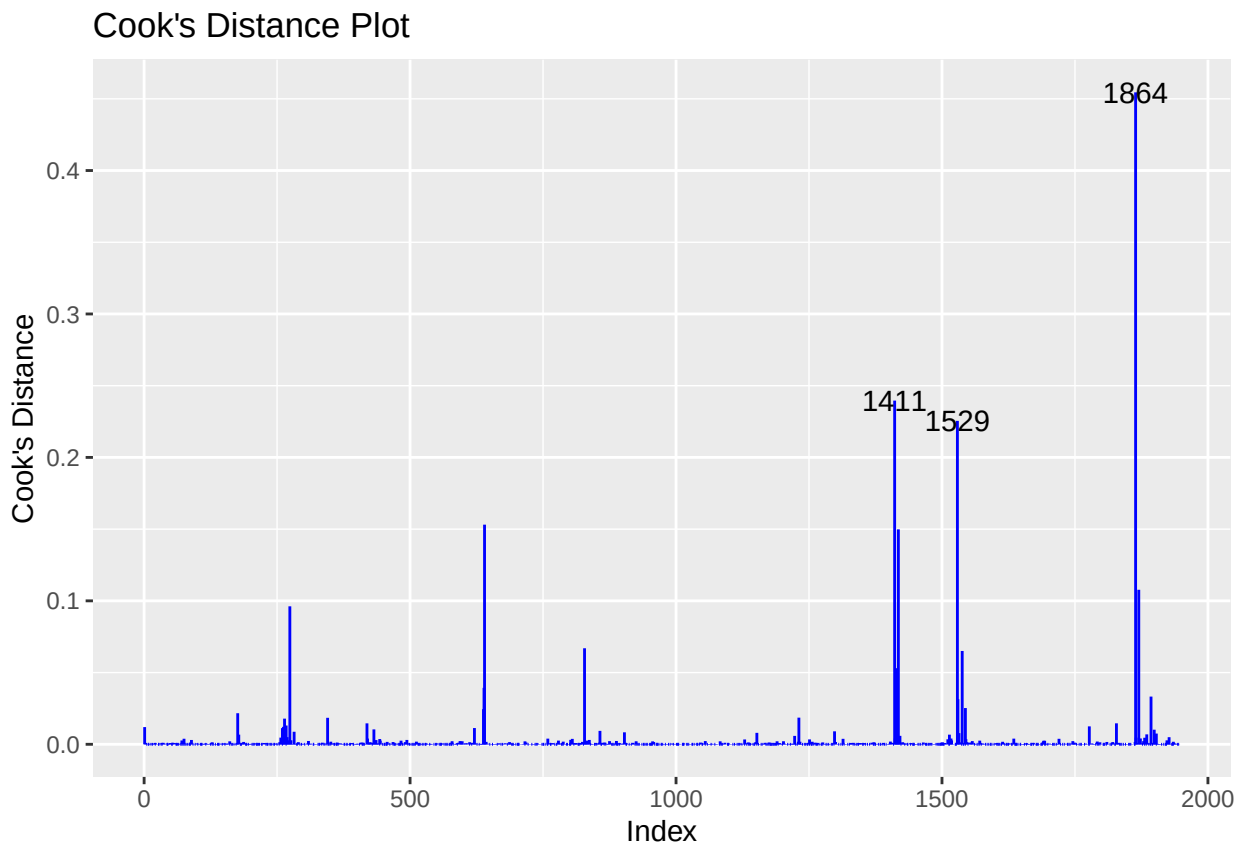

```
## Press [enter] to continue
```



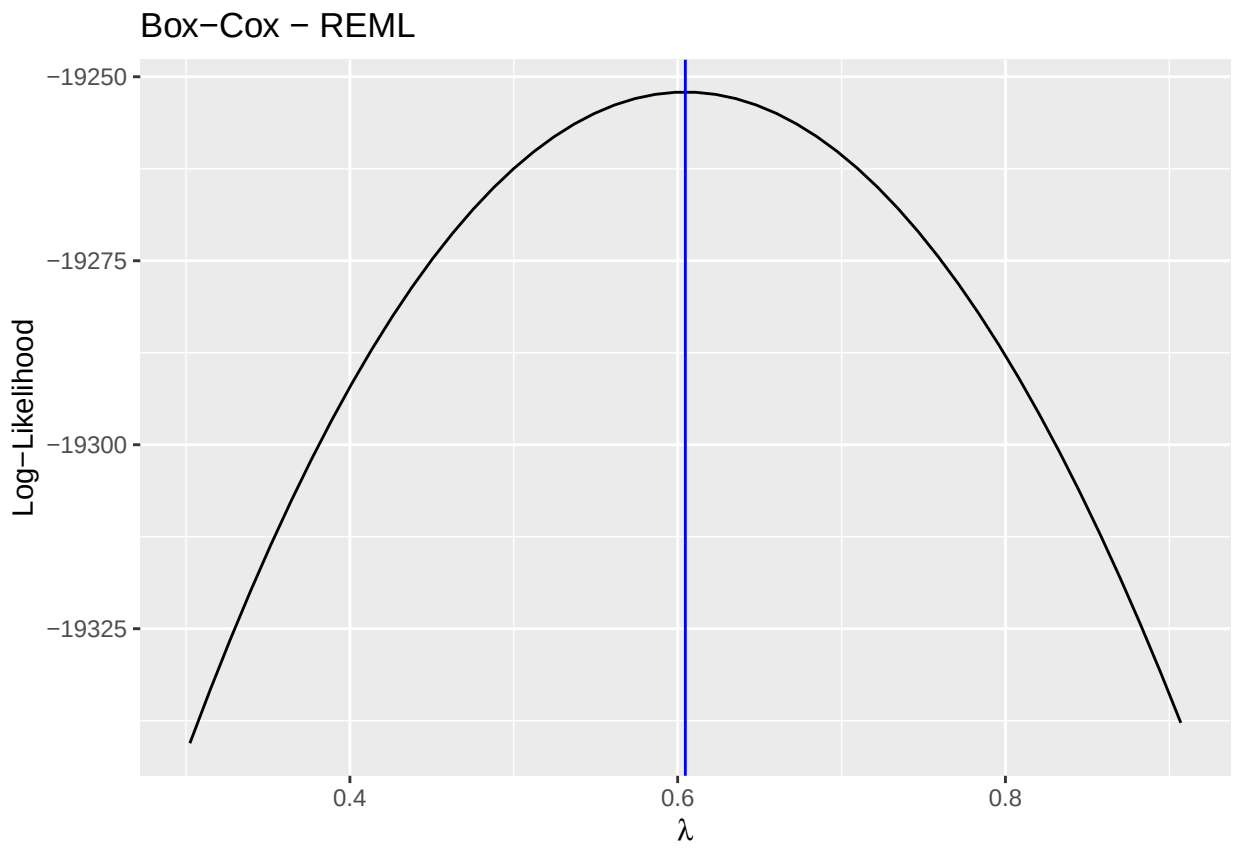

```
## Press [enter] to continue
```




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




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



Dual transformation

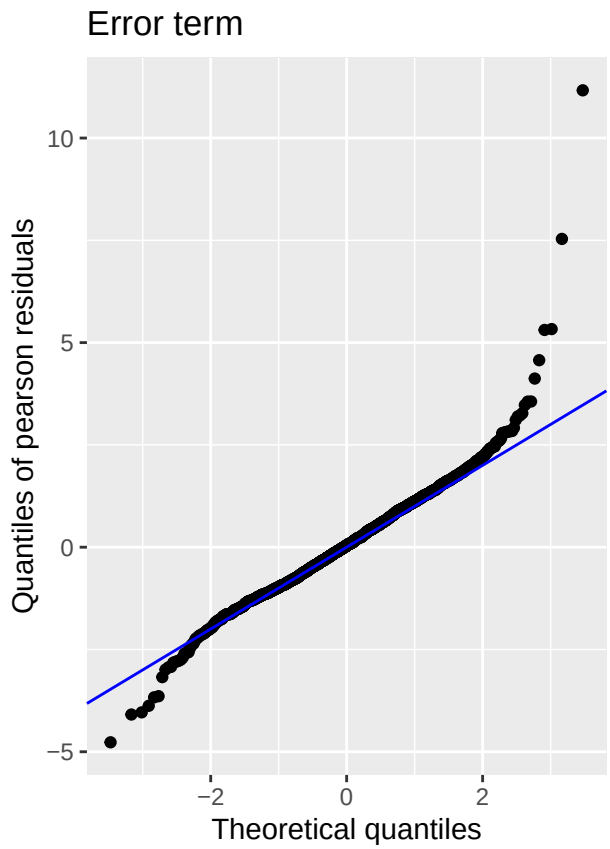
The Dual transformation depends on one transformation parameter and is only defined for positive y. Therefore, a deterministic shift first shifts the dependent variable to the positive range and then the transformation is applied. For the estimation of the transformation parameter, an interval for the optimization needs to be defined. In emdi, a default option can be chosen which equals an interval between 0 and 2 since the Dual transformation does not allow for negative transformation parameter. The interval will be reasonable for many applications but it can happen that the interval needs to be adjusted. This can be done with a numeric vector of length two defining the lower and upper limit of the interval, e.g. c(0, 2) for the default interval.

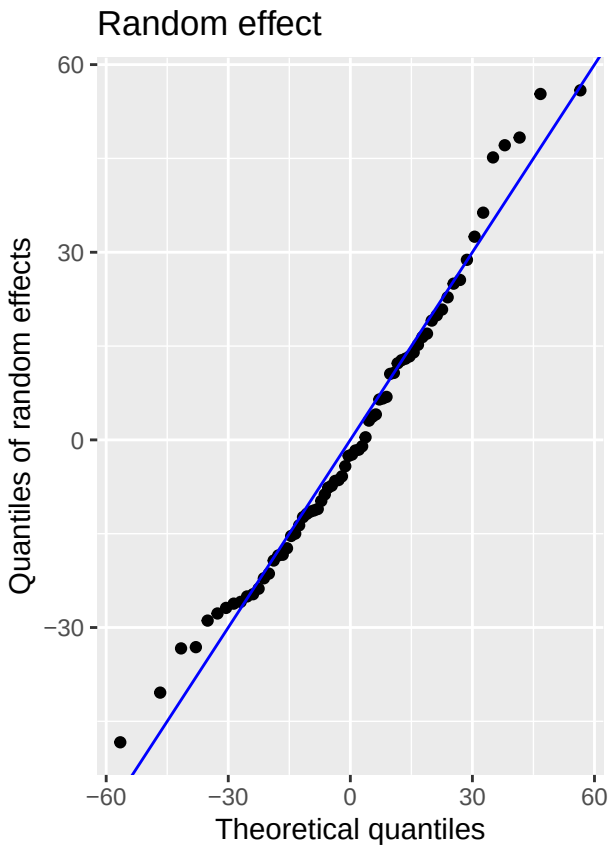
```
ebp_dual <- ebp(fixed = eqIncome ~ gender + eqsize + cash + self_empl +
                  unempl_ben + age_ben + surv_ben + sick_ben + dis_ben + rent +
                  fam_allow + house_allow + cap_inv + tax_adj,
                pop_data = eusilcA_pop, pop_domains = "district",
               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

```
## 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",
```

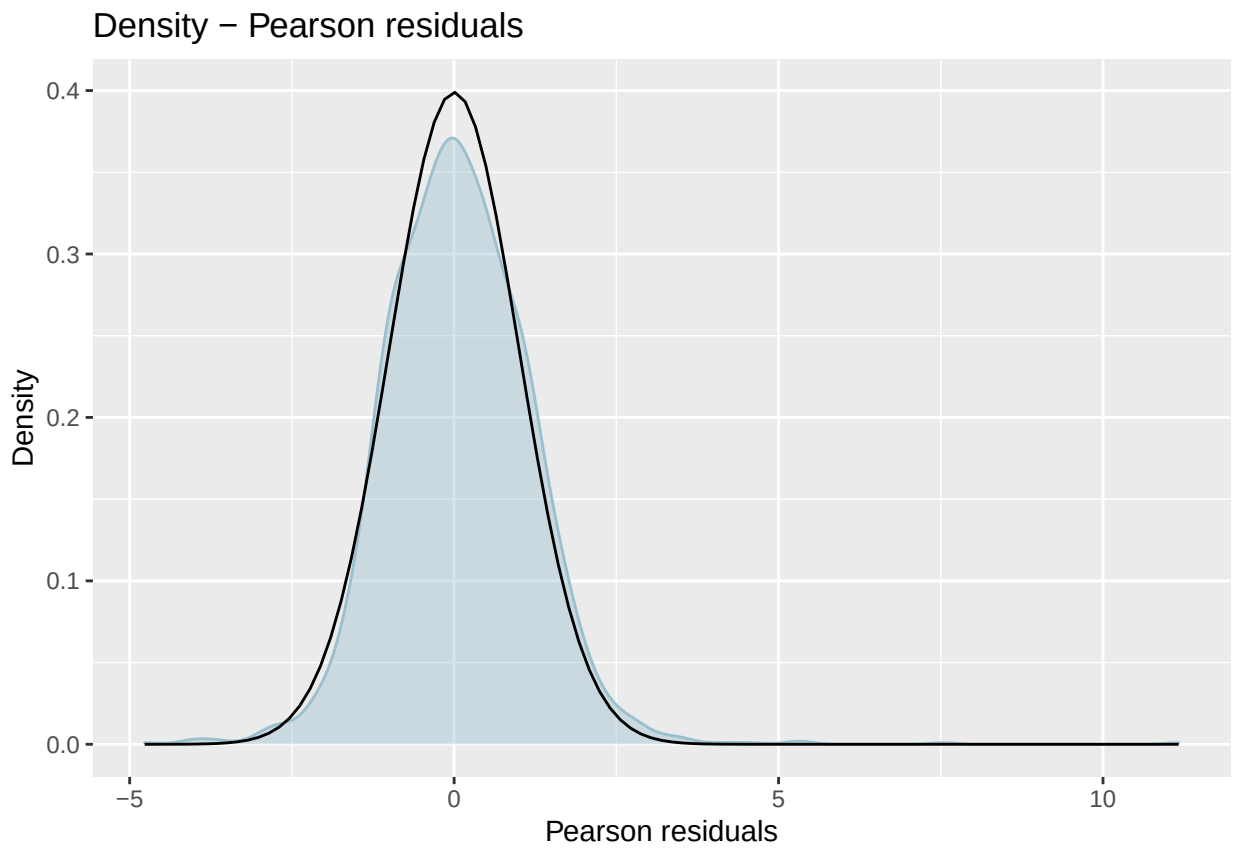
```
##     threshold = 10885.33, transformation = "dual", interval = "default",
##     MSE = FALSE)
##
## Out-of-sample domains:  24
## In-sample domains:  70
##
## Sample sizes:
## Units in sample:  1945
## Units in population:  25000
##                     Min. 1st Qu. Median      Mean 3rd Qu. Max.
## 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
##    0.6325965      0.7092674
##
## Residual diagnostics:
##               Skewness Kurtosis Shapiro_W    Shapiro_p
## Error         0.752435 9.647438 0.9619800 3.487023e-22
## Random_effect 0.465552 2.837214 0.9760562 1.995026e-01
##
## ICC:  0.2086831
##
## Transformation:
##  Transformation Method Optimal_lambda Shift_parameter
##            dual   reml      0.6047161               0
```

```
plot(ebp_dual)
```

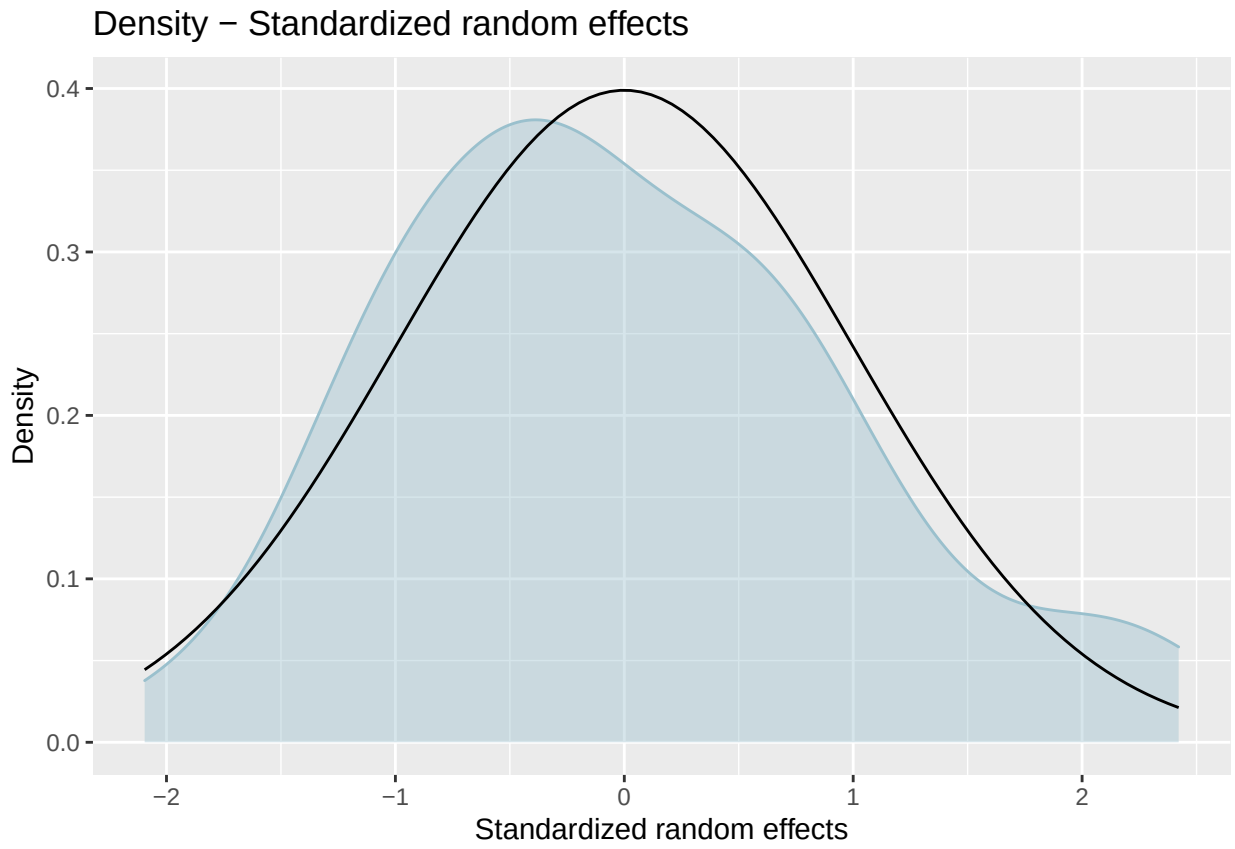




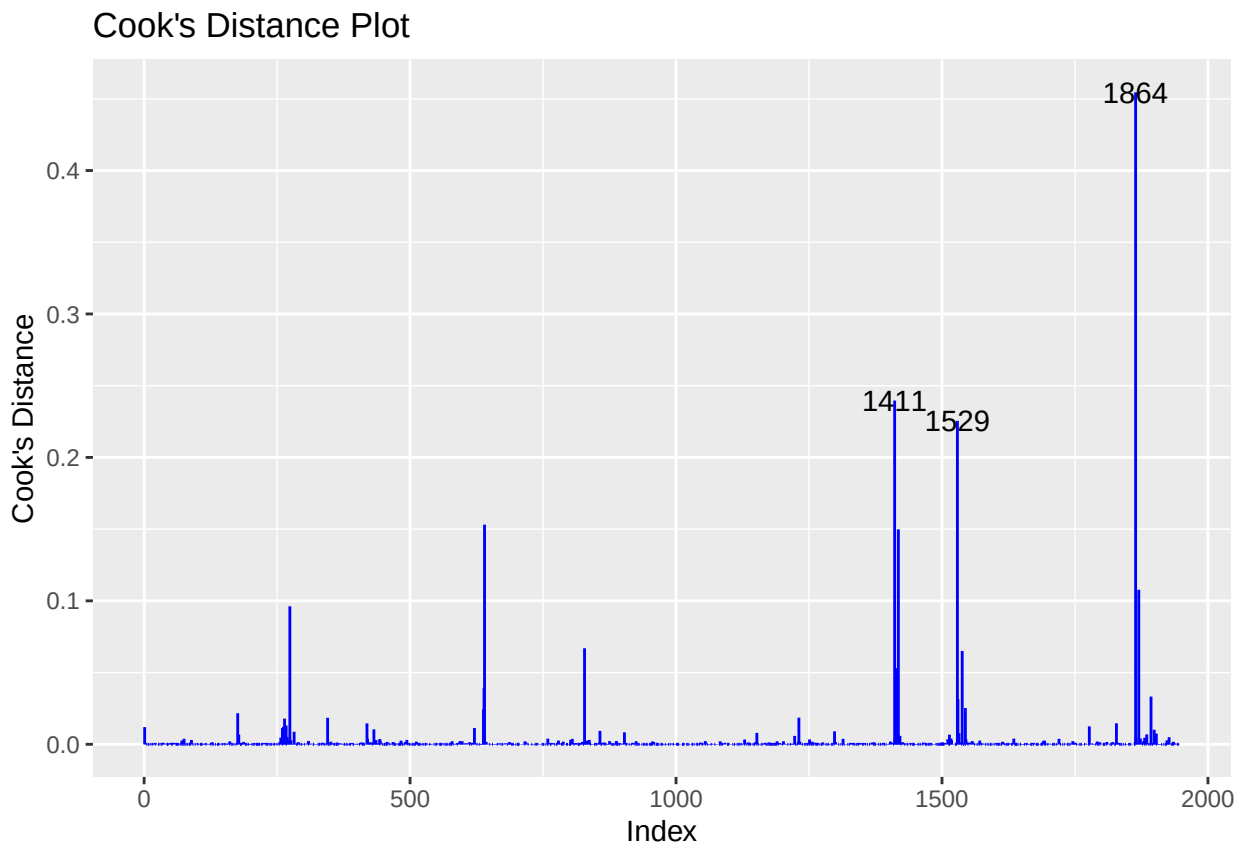

```
## Press [enter] to continue
```



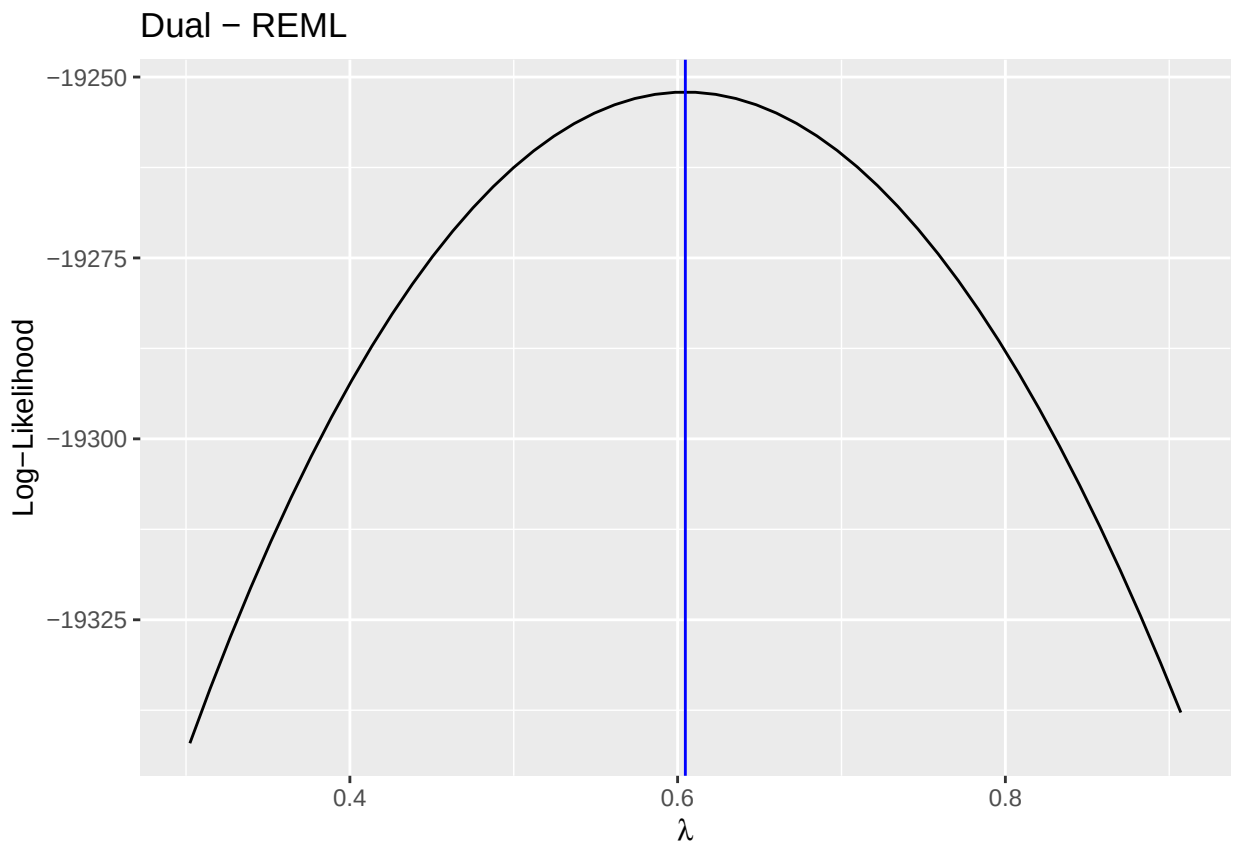

```
## Press [enter] to continue
```




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




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



Log-Shift transformation

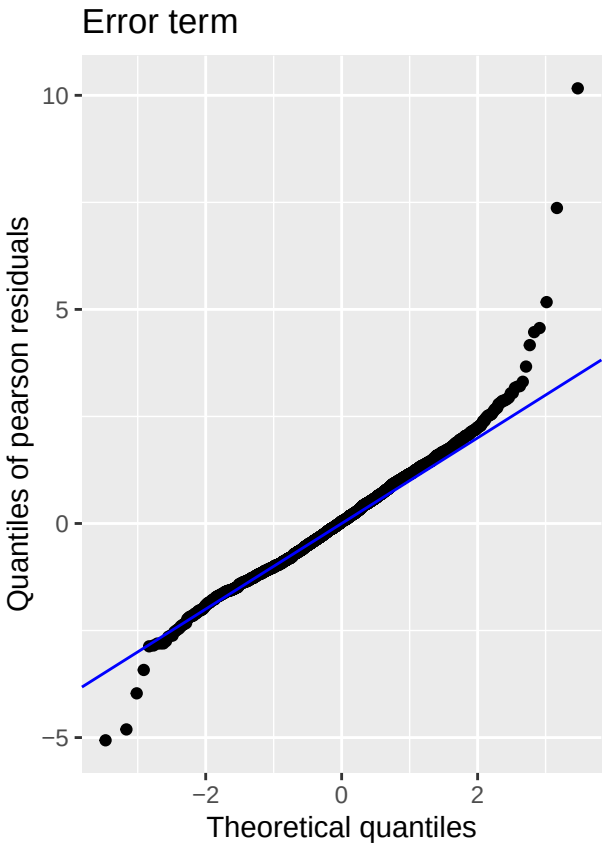
The Log-Shift transformation depends on one transformation parameter and is only defined for positive y. The transformation parameter is the shift such that there is no extra deterministic shift even though the positive scale of y is ensured internally.For the estimation of the transformation parameter, an interval for the optimization needs to be defined. In emdi, a default option can be chosen where the interval is based on the total range of y. If $min(y) + 1 <= 1$, the lower limit of the interval is $|min(y)| + 1$, otherwise the lower limit is 0. The upper limit is $\frac{max(y)-min(y)}{2}$. This interval will be reasonable for many applications but it can happen that the interval needs to be adjusted. This can be done with a numeric vector of length two defining the lower and upper limit of the interval, e.g. c(20000, 30000) for the default interval.

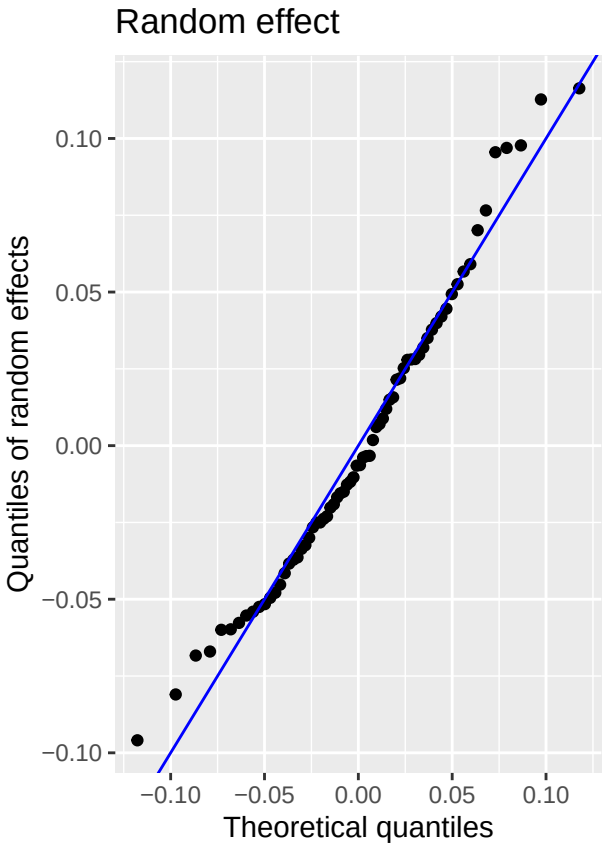
```
ebp_logShift <- ebp(fixed = eqIncome ~ gender + eqsize + cash + self_empl +
                    unempl_ben + age_ben + surv_ben + sick_ben + dis_ben + rent +
                    fam_allow + house_allow + cap_inv + tax_adj,
                pop_data = eusilcA_pop, pop_domains = "district",
       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

```
## 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,
```

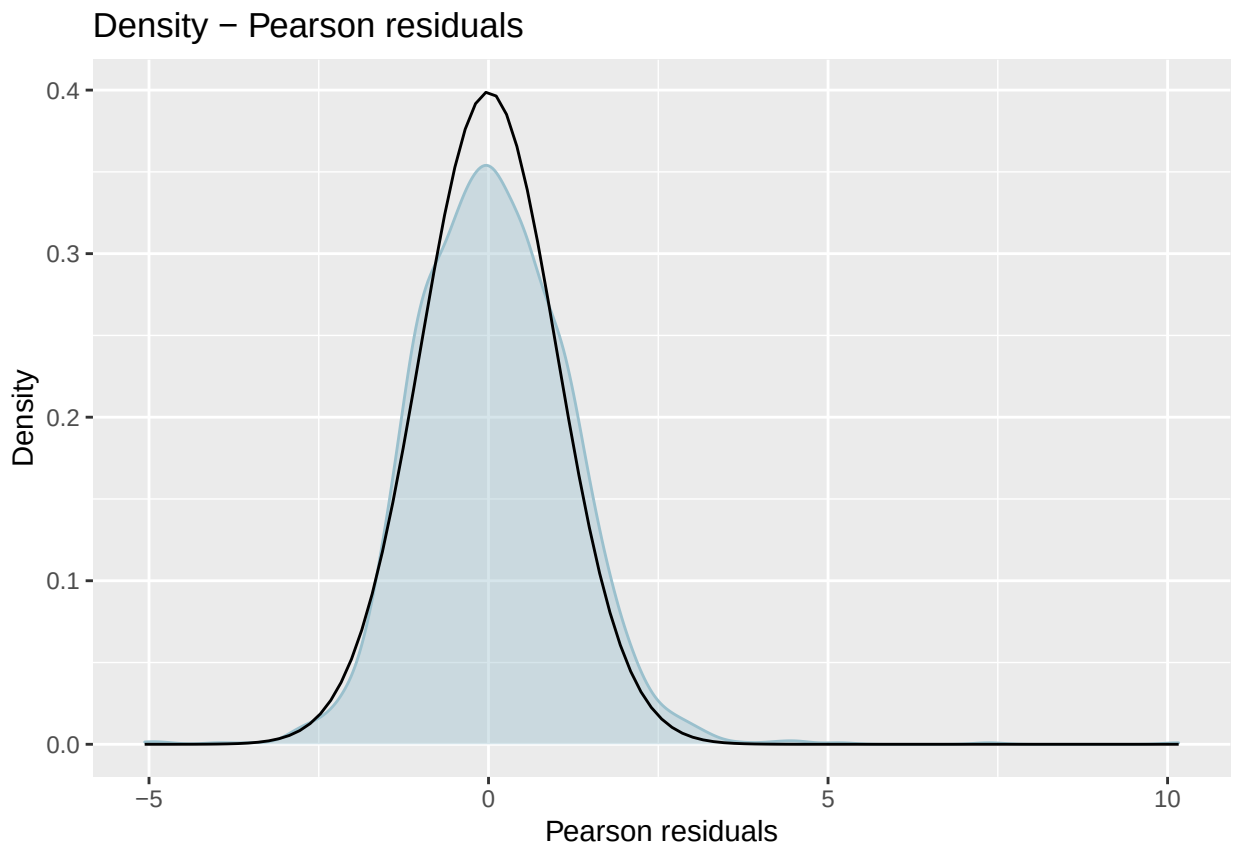
```
##     pop_domains = "district", smp_data = eusilcA_smp, smp_domains = "district",
##     threshold = 10885.33, transformation = "log.shift", interval = "default",
##     MSE = FALSE)
##
## Out-of-sample domains:  24
## In-sample domains:  70
##
## Sample sizes:
## Units in sample:  1945
## Units 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
##    0.6233538      0.7054886
##
## Residual diagnostics:
##                Skewness Kurtosis Shapiro_W    Shapiro_p
## Error         0.6222910 7.607189 0.9706711 1.705890e-19
## Random_effect 0.4788713 2.726898 0.9737695 1.487627e-01
##
## ICC:  0.2180689
##
## Transformation:
##  Transformation Method Optimal_lambda
##       log.shift   reml       27907.57
```

```
plot(ebp_logShift)
```

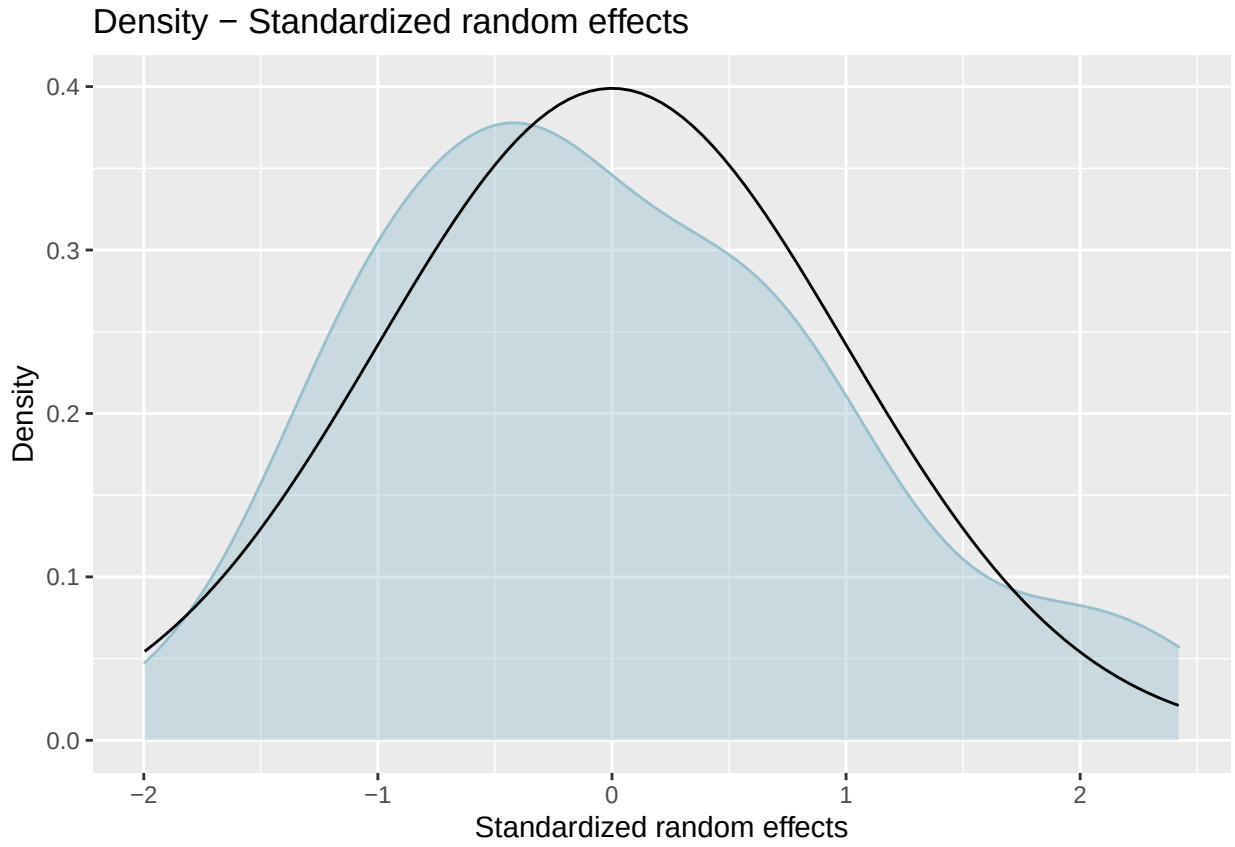




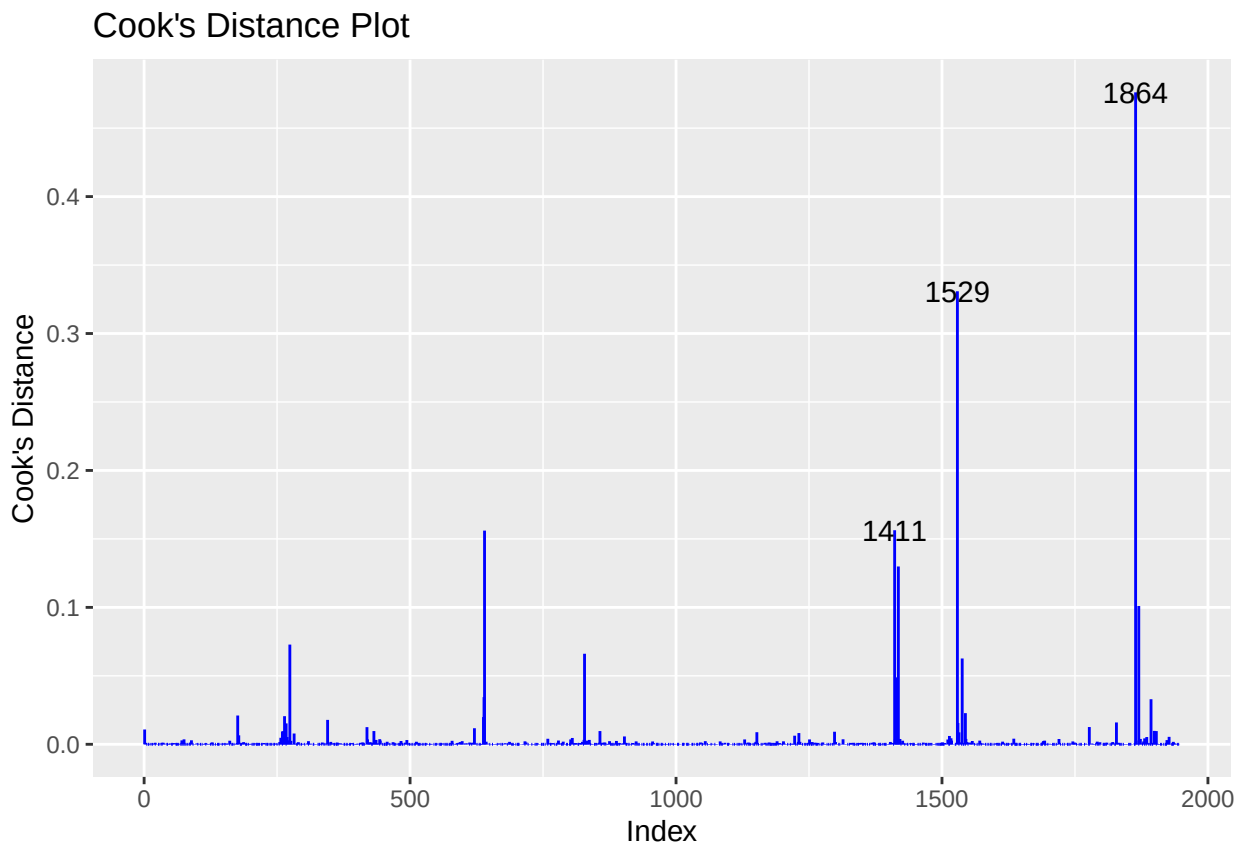

```
## Press [enter] to continue
```



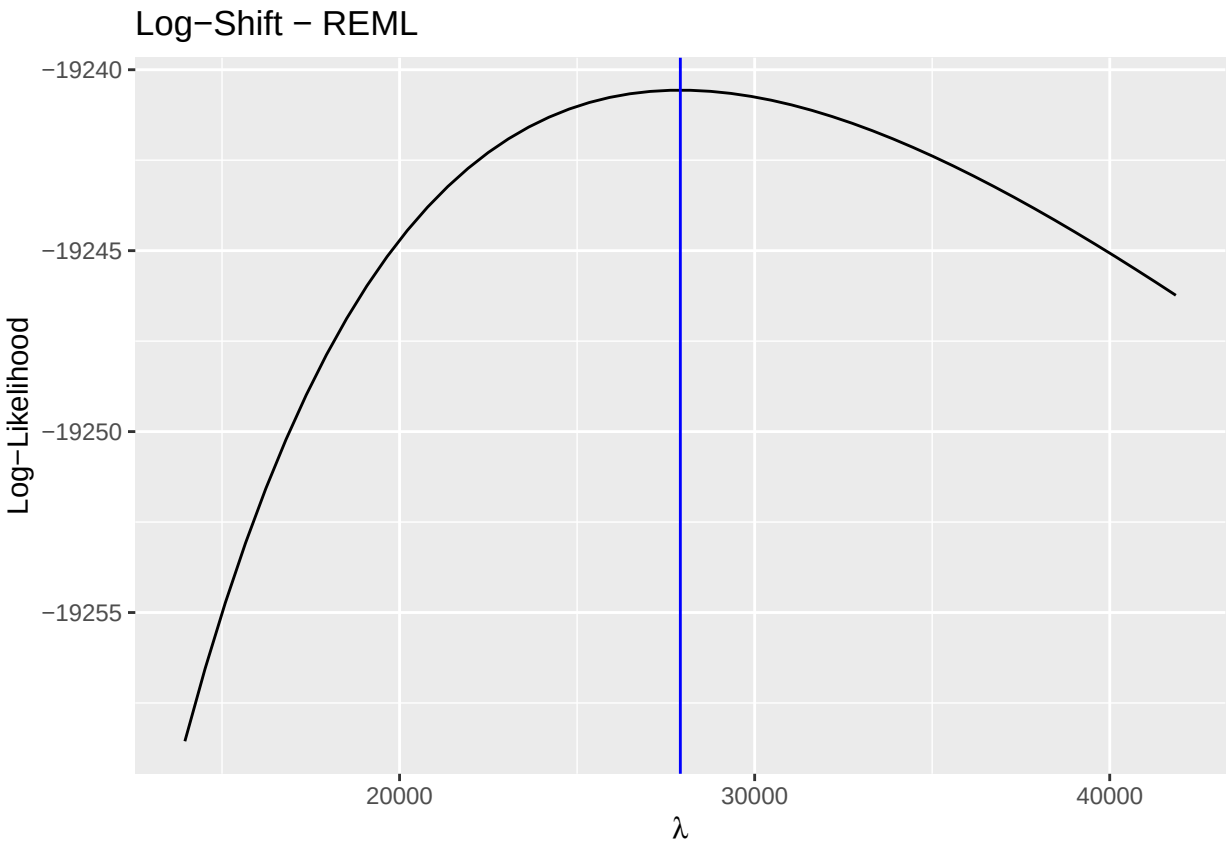

```
## Press [enter] to continue
```




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




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


Log−Shift − REML
−19240
−19245
−19250
−19255
Log−Likelihood
20000
30000
40000
λ