Assignment-7.R

Purbasha Chatterjee

Fri Nov 17 18:45:58 2017

##Q1-Ex:7.20

We have 10 samples initially with standard error as 0.0344 and the estimated standard deviation remains same throughout.

σ = 0.08226, $s_x{}^2$=0.6344 as given in Display 7.10.

Equation for standard error is given by,

$$\text{SE}(\beta_o') = \sigma\sqrt{\frac{1}{n} + \frac{X^2}{(n-1)s_x{}^2}}$$

After putting the values, we get:

$$0.0344 = 0.08226\sqrt{\frac{1}{n} + \frac{X^2}{(n-1)s_x{}^2}}$$

$$\frac{1}{10} + \frac{X^2}{(n-1)s_x{}^2} = \left(\frac{0.0344}{0.08226}\right)^2$$

$$\frac{X^2}{9s_x{}^2} = \left(\frac{0.0344}{0.08226}\right)^2 + \frac{1}{10}$$

$$\frac{X^2}{9*0.6344} = 2.1259+0.1$$

$$X^2 = 12.71$$

The objective is to compute the required sample size n in order for the SE of the estimated slope is 0.01.

Now , the given standard error is 0.01, that is SE(β_o') = 0.01 for which population size be n,

$$\text{SE}(\beta_o') = \sigma\sqrt{\frac{1}{n} + \frac{X^2}{(n-1)s_x{}^2}}$$

$$0.01 = 0.08226\sqrt{\frac{1}{n} + \left(\frac{12.71}{(n-1)*0.6344}\right)}$$

$$\frac{0.0001}{0.0068}=\frac{1}{n}+\left(\frac{12.71}{(n-1)*0.6344}\right)$$

Solving for n, we get

n=17

Thus, the new population size is 17.

```
library(ggplot2)

## Warning: package 'ggplot2' was built under R version 3.4.2

library(Sleuth3)

## Warning: package 'Sleuth3' was built under R version 3.4.2

##Q2-Ex:7.24
mod_den <- lm(Denmark ~ Year, data = ex0724)
summary(mod_den)

##
## Call:
## lm(formula = Denmark ~ Year, data = ex0724)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.003225 -0.001339  0.000089  0.001119  0.003790
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.987e-01  4.080e-02  14.673   <2e-16 ***
## Year        -4.289e-05  2.069e-05  -2.073   0.0442 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.001803 on 43 degrees of freedom
## Multiple R-squared:  0.09083,    Adjusted R-squared:  0.06968
## F-statistic: 4.296 on 1 and 43 DF,  p-value: 0.04424

mod_neth <- lm(Netherlands ~ Year, data = ex0724)
summary(mod_neth)

##
## Call:
## lm(formula = Netherlands ~ Year, data = ex0724)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -0.0031437 -0.0008246  0.0002819  0.0009287  0.0021478
##
```

```
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  6.724e-01  2.792e-02   24.08  < 2e-16 ***
## Year        -8.084e-05  1.416e-05   -5.71 9.64e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.001233 on 43 degrees of freedom
## Multiple R-squared:  0.4313, Adjusted R-squared:  0.418 
## F-statistic: 32.61 on 1 and 43 DF,  p-value: 9.637e-07
```

```r
mod_can <- lm(Canada ~ Year, data = ex0724)
summary(mod_can)
```

```
## 
## Call:
## lm(formula = Canada ~ Year, data = ex0724)
## 
## Residuals:
##        Min         1Q     Median         3Q        Max 
## -1.494e-03 -6.161e-04 -8.312e-05  4.951e-04  1.284e-03 
## 
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  7.338e-01  5.480e-02  13.390 3.98e-11 ***
## Year        -1.112e-04  2.768e-05  -4.017 0.000738 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.000768 on 19 degrees of freedom
##   (24 observations deleted due to missingness)
## Multiple R-squared:  0.4592, Adjusted R-squared:  0.4307 
## F-statistic: 16.13 on 1 and 19 DF,  p-value: 0.0007376
```

```r
mod_usa <- lm(USA ~ Year, data = ex0724)
summary(mod_usa)
```

```
## 
## Call:
## lm(formula = USA ~ Year, data = ex0724)
## 
## Residuals:
##        Min         1Q     Median         3Q        Max 
## -5.343e-04 -1.800e-04 -1.714e-05  2.571e-04  3.743e-04 
## 
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  6.201e-01  1.860e-02  33.340  < 2e-16 ***
## Year        -5.429e-05  9.393e-06  -5.779 1.44e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 0.0002607 on 19 degrees of freedom
##   (24 observations deleted due to missingness)
## Multiple R-squared:  0.6374, Adjusted R-squared:  0.6183
## F-statistic:  33.4 on 1 and 19 DF,  p-value: 1.439e-05
```

```
confint(mod_den)
```

```
##                     2.5 %        97.5 %
## (Intercept)  5.164327e-01  6.810139e-01
## Year        -8.461396e-05 -1.156787e-06
```

```
confint(mod_neth)
```

```
##                     2.5 %        97.5 %
## (Intercept)  0.6160931738  0.7287035763
## Year        -0.0001093949 -0.0000522915
```

```
confint(mod_can)
```

```
##                     2.5 %        97.5 %
## (Intercept)  0.6190865670  8.484849e-01
## Year        -0.0001690974 -5.324024e-05
```

```
confint(mod_usa)
```

```
##                     2.5 %        97.5 %
## (Intercept)  5.811580e-01  6.590134e-01
## Year        -7.394606e-05 -3.462537e-05
```

```
mod_den2 <- lm(Denmark ~ Year-1, data = ex0724)
summary(mod_den2)
```

```
##
## Call:
## lm(formula = Denmark ~ Year - 1, data = ex0724)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -0.0082631 -0.0037131  0.0003941  0.0033155  0.0087476
##
## Coefficients:
##       Estimate Std. Error t value Pr(>|t|)
## Year 2.607e-04  3.302e-07   789.7   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.004368 on 44 degrees of freedom
## Multiple R-squared:  0.9999, Adjusted R-squared:  0.9999
## F-statistic: 6.236e+05 on 1 and 44 DF,  p-value: < 2.2e-16
```

```
mod_neth2 <- lm(Netherlands ~ Year-1, data = ex0724)
summary(mod_neth2)
```

```
##
## Call:
## lm(formula = Netherlands ~ Year - 1, data = ex0724)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -0.0078880 -0.0032481 -0.0003059  0.0034345  0.0087761
##
## Coefficients:
##       Estimate Std. Error t value Pr(>|t|)
## Year 2.601e-04  3.508e-07   741.4   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.004641 on 44 degrees of freedom
## Multiple R-squared:  0.9999, Adjusted R-squared:  0.9999
## F-statistic: 5.497e+05 on 1 and 44 DF,  p-value: < 2.2e-16
```

```
mod_usa2 <- lm(USA ~ Year-1, data = ex0724)
summary(mod_usa2)
```

```
##
## Call:
## lm(formula = USA ~ Year - 1, data = ex0724)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -0.0031831 -0.0014297  0.0003058  0.0019002  0.0033947
##
## Coefficients:
##       Estimate Std. Error t value Pr(>|t|)
## Year 2.589e-04  2.160e-07    1199   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.00196 on 20 degrees of freedom
##   (24 observations deleted due to missingness)
## Multiple R-squared:      1,  Adjusted R-squared:      1
## F-statistic: 1.437e+06 on 1 and 20 DF,  p-value: < 2.2e-16
```

```
mod_can2 <- lm(Canada ~ Year-1, data = ex0724)
summary(mod_can2)
```

```
##
## Call:
## lm(formula = Canada ~ Year - 1, data = ex0724)
##
## Residuals:
```

```
##        Min         1Q     Median         3Q        Max
## -0.0036994 -0.0013834 -0.0000646  0.0016137  0.0039703
##
## Coefficients:
##       Estimate Std. Error t value Pr(>|t|)
## Year 2.594e-04  2.665e-07   973.4   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.002418 on 20 degrees of freedom
##   (24 observations deleted due to missingness)
## Multiple R-squared:      1,  Adjusted R-squared:      1
## F-statistic: 9.475e+05 on 1 and 20 DF,  p-value: < 2.2e-16
```

```
anova(mod_den2, mod_den)
```

```
## Analysis of Variance Table
##
## Model 1: Denmark ~ Year - 1
## Model 2: Denmark ~ Year
##   Res.Df        RSS Df  Sum of Sq      F    Pr(>F)
## 1     44 0.00083935
## 2     43 0.00013973  1 0.00069962 215.29 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(mod_usa2, mod_usa)
```

```
## Analysis of Variance Table
##
## Model 1: USA ~ Year - 1
## Model 2: USA ~ Year
##   Res.Df       RSS Df Sum of Sq      F    Pr(>F)
## 1     20 7.681e-05
## 2     19 1.291e-06  1 7.552e-05 1111.6 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

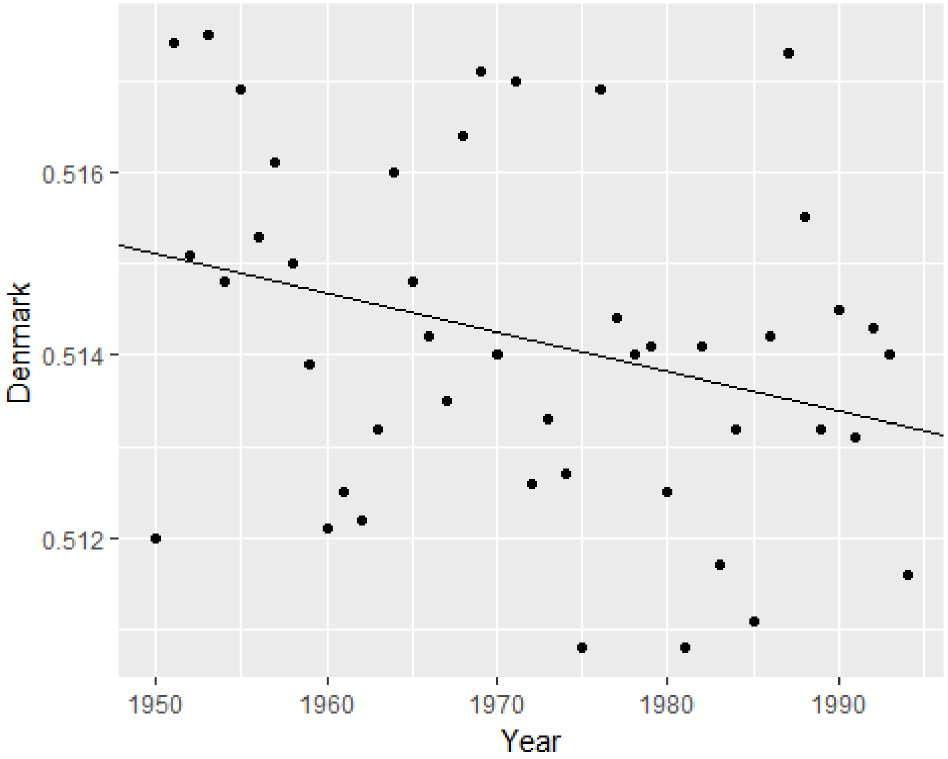
```
anova(mod_neth2, mod_neth)
```

```
## Analysis of Variance Table
##
## Model 1: Netherlands ~ Year - 1
## Model 2: Netherlands ~ Year
##   Res.Df        RSS Df  Sum of Sq      F    Pr(>F)
## 1     44 0.00094781
## 2     43 0.00006542  1 0.00088239 580.01 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(mod_can2, mod_can)
```

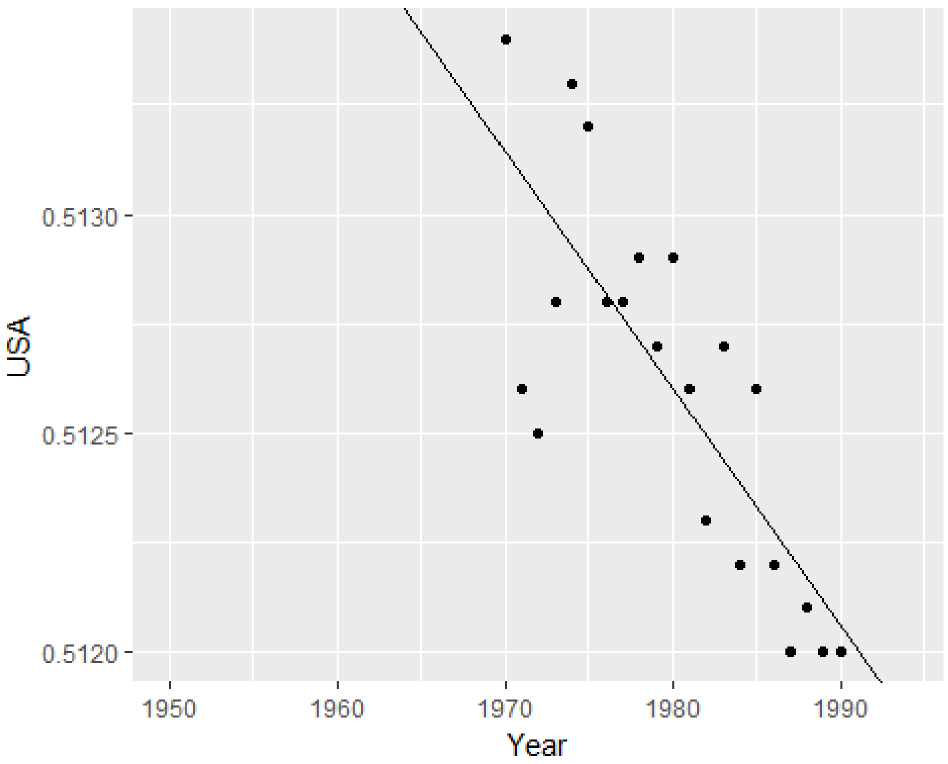
```
## Analysis of Variance Table
## 
## Model 1: Canada ~ Year - 1
## Model 2: Canada ~ Year
##   Res.Df        RSS Df  Sum of Sq      F    Pr(>F)    
## 1     20 1.1696e-04                                   
## 2     19 1.1207e-05  1 0.00010575 179.29 3.984e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
bhat_den <- coefficients(mod_den)
ggplot(data = ex0724, aes(Year, Denmark)) + geom_point()  +
  geom_abline(slope = bhat_den[2], intercept = bhat_den[1])
```



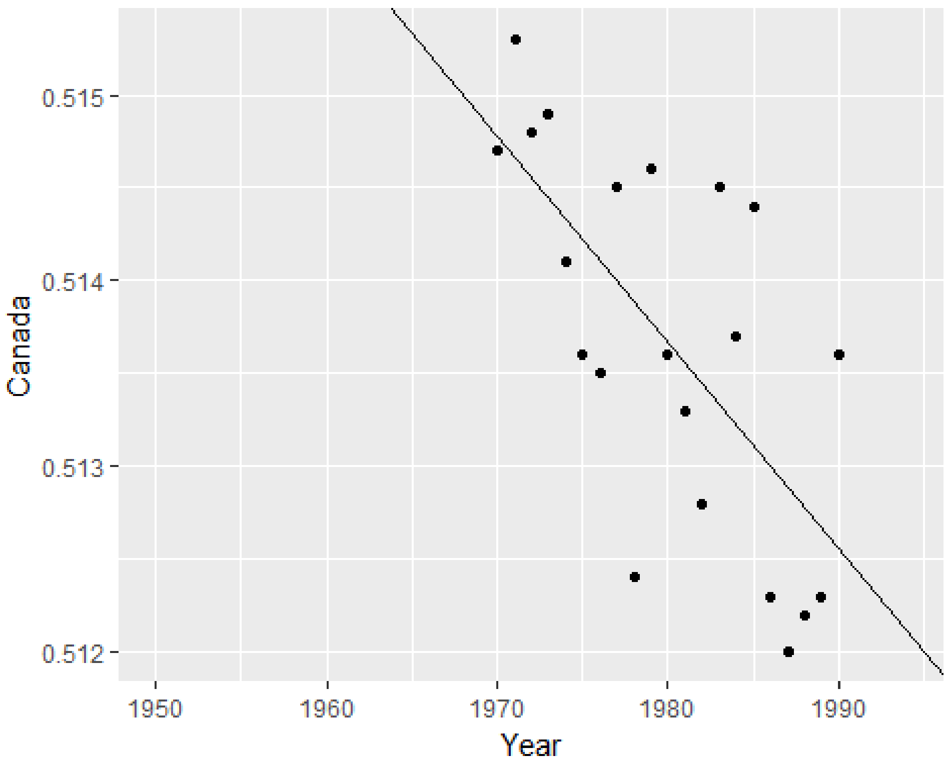
```
bhat_usa <- coefficients(mod_usa)
ggplot(data = ex0724, aes(Year, USA)) + geom_point()  +
  geom_abline(slope = bhat_usa[2], intercept = bhat_usa[1])

## Warning: Removed 24 rows containing missing values (geom_point).
```

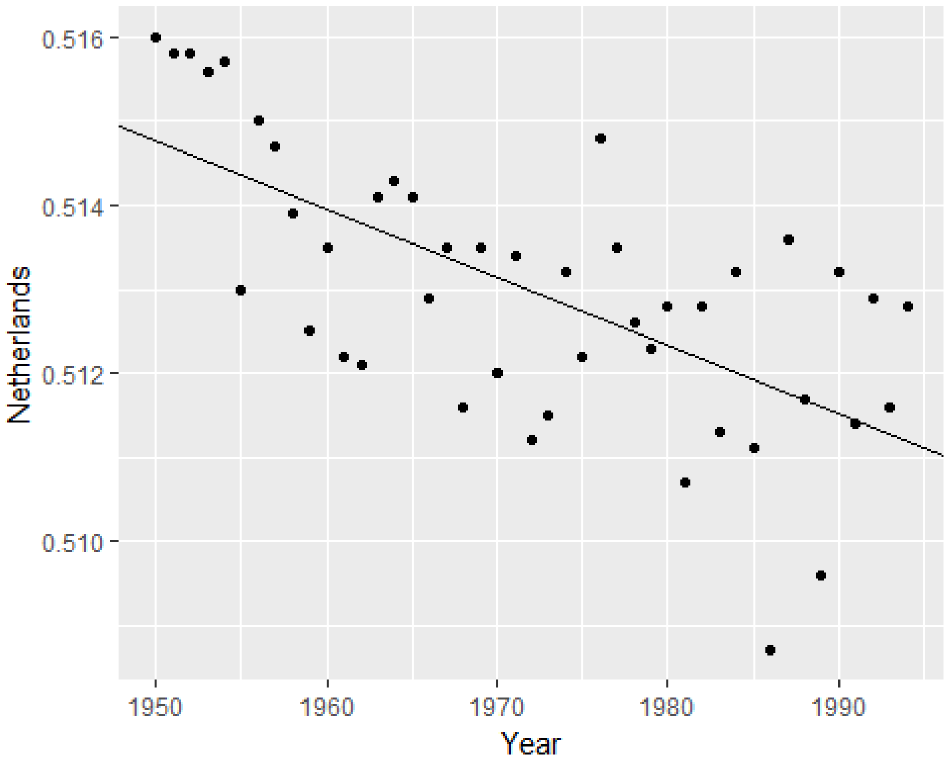


```
bhat_can <- coefficients(mod_can)
ggplot(data = ex0724, aes(Year, Canada)) + geom_point()  +
  geom_abline(slope = bhat_can[2], intercept = bhat_can[1])

## Warning: Removed 24 rows containing missing values (geom_point).
```



```
bhat_neth <- coefficients(mod_neth)
ggplot(data = ex0724, aes(Year, Netherlands)) + geom_point()  +
  geom_abline(slope = bhat_neth[2], intercept = bhat_neth[1])
```



As it can be observed that by applying simple linear regression model with intercept, following results are obtained for four countries:

Country		Estimate	Standard Error	t-value	Pr(>|t|)
Denmark	Intercept	0.5987	0.04080	14.673	<2e-16
	Slope	-0.000043	0.00002069	-2.073	0.0442
Netherlands	Intercept	0.6724	0.02792	24.08	< 2e-16
	Slope	-0.000081	0.00001416	-5.71	9.64e-07
Canada	Intercept	0.7338	0.05480	13.390	3.98e-11
	Slope	-0.0001112	0.00002768	-4.017	0.000738
USA	Intercept	0.6201	0.01860	33.340	< 2e-16
	Slope	-0.00005429	0.000009393	-5.779	1.44e-05

a) As it can be observed from the above table that the slope and intercept of each countries found by using simple linear regression model confirms the estimates and standard errors displayed in the fig. Display 7.17

b)

Country	Estimate	Standard Error	t-value	Pr(>|t|)
Denmark	2.607e-04	3.302e-07	789.7	<2e-16
Netherlands	2.601e-04	3.508e-07	741.4	<2e-16
Canada	2.594e-04	2.665e-07	973.4	<2e-16
USA	2.589e-04	2.160e-07	1199	<2e-16

From the above table, t-values has been recorded for the test that the slopes of the regressions are zero. We can observe that p-value in each of these cases are less than 0.05 giving us sufficient evidence to reject the null hypothesis.
It can be observed from the previous table that there exists a negative relation between the year and the birth rate of each country.

c) From the above table, it can be observed that USA has the maximum t-value, even though it has the third largest slope in the model with intercept. This is because the standard error is lowest in case of USA and t-statistics is defined as the ratio of slope by standard error.

d) The standard error for USA is the lowest even though Canada and USA has the same sample size. The standard deviation that is the residual error for USA is observed to be the lowest 0.0002607 as compared to Canada with residual error as 0.000768. Additionally, the proportion of variance is highest for USA with value 0.6374 whereas Canada has 0.4592.

e) The standard deviations about the regression line might be different for four countries because the proportion of males that are born that is the average number of male that are born are different. Each of these countries have different population and standard deviation is the residual error which is dependent on degrees of freedom. Each of these samples might not have equal degree of freedom, producing different standard of error.
From the above analysis, following points have also been observed:

- The confidence interval for USA is -7.394606e-05 to -3.462537e-05, whereas for Canada it is -0.0001690974 to -5.324024e-05, Netherlands is -0.0001093949 to -0.0000522915 and Denmark is -8.461396e-05 to -1.156787e-06.
- The p-value is less than 0.05 by applying anova over the two lm models, giving convincing evidence of non-zero y-intercept.

```
##Q3-Ex:7.30
out <- lm(Refusal~Age, data=ex0730)
summary(out)
```

```
##
## Call:
## lm(formula = Refusal ~ Age, data = ex0730)
##
## Residuals:
##          1          2          3          4          5          6
## -0.0028021  0.0059720 -0.0005604 -0.0070928  0.0063748 -0.0018914
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.4444308  0.0073312   60.62 4.43e-07 ***
## Age         -0.0023468  0.0001557  -15.08 0.000113 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.005881 on 4 degrees of freedom
## Multiple R-squared:  0.9827, Adjusted R-squared:  0.9784
## F-statistic: 227.3 on 1 and 4 DF,  p-value: 0.0001128
```

```
confint(out)
```

```
                   2.5 %       97.5 %
(Intercept)  0.424076174  0.464785472
Age         -0.002778942 -0.001914578
```

Here, it is observed that the constant estimate value is 0.444 and age variable estimate is -0.0023. This implies there exist a negative relation between the age of the interviewer and the refusal rate.

Moreover, a confidence interval of -0.002778942 to -0.001914578 was obtained. The p-value is less than 0.05, allowing to reject the null hypothesis. The estimate of the standard deviation of the population is 0.005881 and the proportion of variance is 0.9827. The variable standard error is 0.00016 and absolute t-statistics value is 15.08 .

Manual proof:

$\mu\{Y|X\} = \beta_1 X + \beta_0$

$$\beta_1 = \frac{\sum_{i=1}^{n}(X_i - X\prime)(Y_i \quad - Y\prime)}{\sum_{i=1}^{n}(X_i - X\prime)\hat{}2} \text{ --------->(1)}$$

$$\beta_0 = Y' - \beta_1 X' \quad \text{---------->(2)}$$

total age(X_i) = 267, total refusal(Y_i)= 2.04

Average age (X') = 44.5, average refusal (Y')= 0.34

- ⇨ $\beta_1 = \frac{-3.35}{1427.5}$ = -0.002
- ⇨ β_0 = 0.34 – ((-0.002) x 44.5) =0.444

Hence, giving a linear equation $\mu\{Y|X\}$ = -0.002 X + 0.444