6338_Proj1

Wenxiong Lu

3/16/2021

```
## [1] "/Users/wenxionglu/Documents/GraduateCourse/Stats  6338/data"
```

```
## Loading required package: carData
```

```
## [1] TRUE
```

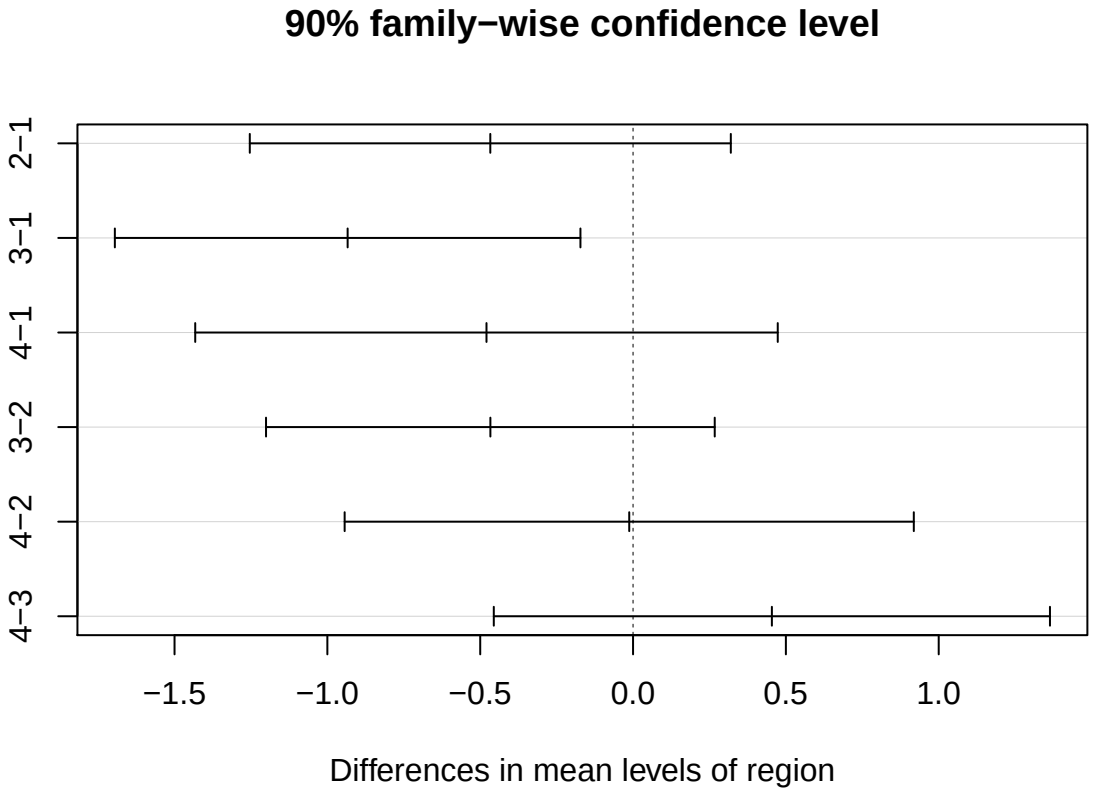
#####1 (a)

```
## Analysis of Variance Table
##
## Response: infprob
##            Df  Sum Sq Mean Sq F value  Pr(>F)
## region      3  13.997  4.6656   2.714 0.04839 *
## Residuals 109 187.383  1.7191
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

H0: all mean infection risks for different regions are equal. H1: at least one infection risk of different regions is not equal p: $0.04839 < 0.05$ conclude: At least one infection risk of different regions is not equal.

(b) confidnece interval:

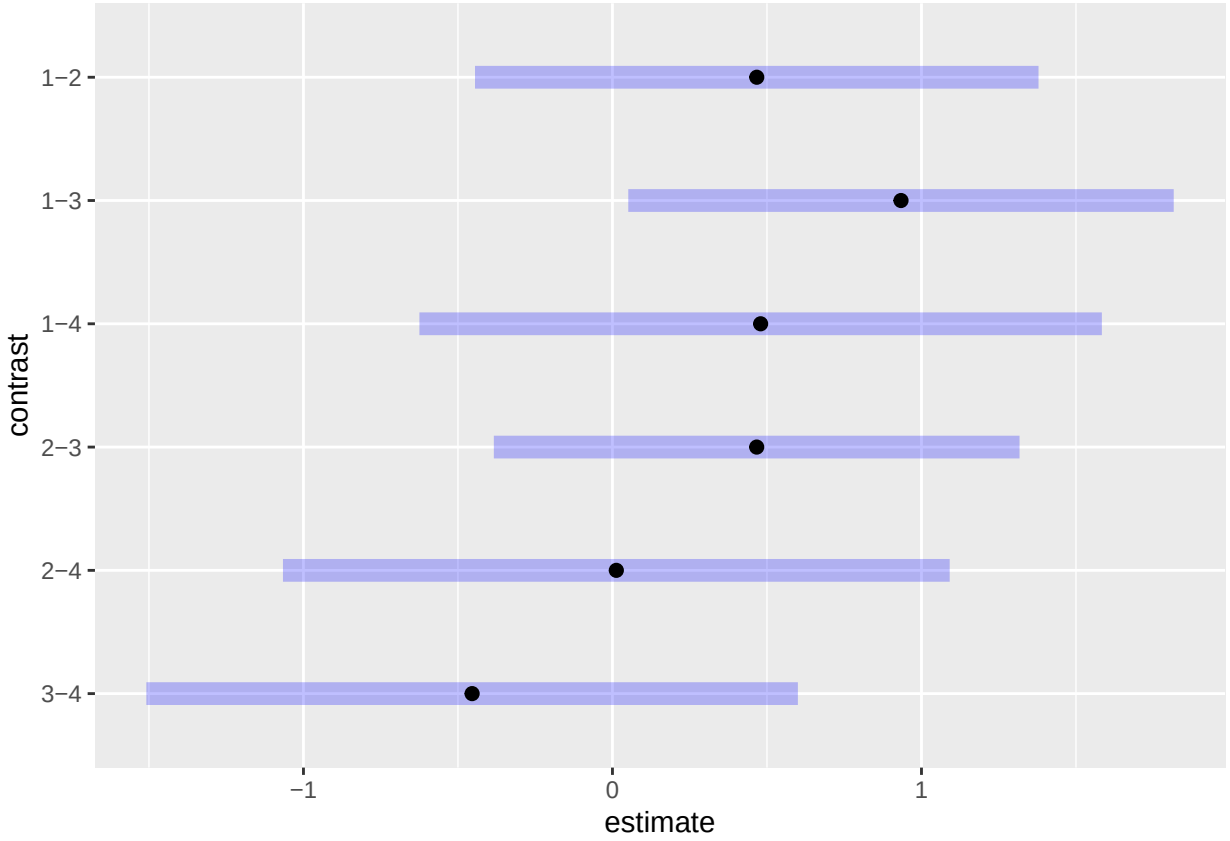
```
##   Tukey multiple comparisons of means
##     90% family-wise confidence level
##
## Fit: aov(formula = infprob ~ region)
##
## $region
##           diff        lwr        upr     p adj
## 2-1 -0.4669643 -1.2537701  0.3198415 0.5168684
## 3-1 -0.9336873 -1.6952799 -0.1720946 0.0269952
## 4-1 -0.4794643 -1.4323334  0.4734048 0.6489580
## 3-2 -0.4667230 -1.2007205  0.2672746 0.4563333
## 4-2 -0.0125000 -0.9434612  0.9184612 0.9999891
## 4-3  0.4542230 -0.4555290  1.3639749 0.6545991
```



Under Tukey comparison procedure, with alpha=0.1 we can find that region 3 vs 1 is siginificant. Comparison between region 2 vs 1, 4 vs 1, 3 vs 2, 4 vs 2, 4 vs 3 are insiginificant.

(c) Use Bonferroni Method

```
##  contrast estimate    SE  df lower.CL upper.CL
##  3-4       -0.4542 0.392 109  -1.5085     0.60
##  2-4        0.0125 0.401 109  -1.0663     1.09
##  2-3        0.4667 0.317 109  -0.3838     1.32
##  1-4        0.4795 0.411 109  -0.6247     1.58
##  1-3        0.9337 0.328 109   0.0511     1.82
##  1-2        0.4670 0.339 109  -0.4448     1.38
##
## Confidence level used: 0.95
## Conf-level adjustment: bonferroni method for 6 estimates
```



The result shows that only comparison between Region 1 vs Region 3 is significantly different. The rest comparison does not indicate difference.

#####2

```
## Registered S3 method overwritten by 'sets':
##   method        from
##   print.element ggplot2
```

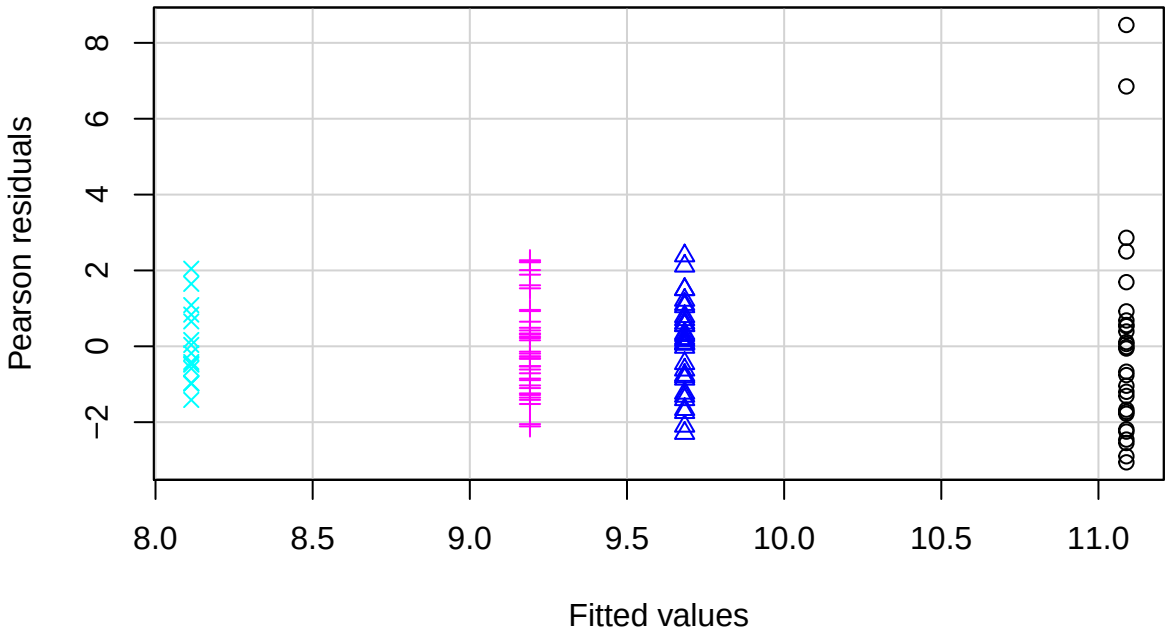
```
## Analysis of Variance Table
##
## Response: infprob
##            Df  Sum Sq Mean Sq F value Pr(>F)
## age         3   3.068  1.0226   0.562 0.6412
## Residuals 109 198.312  1.8194
```

H0: The mean infection risk are equal for the four age groups. H1: Not all four age group have the same mean infection risk. if p-value $>$ alpha $= 0.1$ conclude H0, O.W. conclude H1. The p-value: $0.6412 > 0.1$. Conclude H0: The mean infection risk are qual for the four age groups.

#####3

```
## Anova Table (Type II tests)
##
## Response: stay
##           Sum Sq  Df F value    Pr(>F)
## region    103.55   3  12.309 5.376e-07 ***
## Residuals 305.66 109
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As the Anova table shows, the mean length of stay is different across differently regions.



(a)

There are a few outliers exist in region 1. No serious departure from the Anova result.

(b)

```
## Levene's Test for Homogeneity of Variance (center = "median")
##        Df F value  Pr(>F)
## group   3  4.3262 0.00637 **
##       109
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The BF test shows the constancy of variance is violated (with P-value=0.00637 < 0.05)

(c)

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 11.088929  0.3164640 35.040093 3.709652e-61
## region2     -1.405491  0.4333362 -3.243419 1.567964e-03
## region3     -1.897577  0.4194500 -4.523966 1.553030e-05
## region4     -2.975179  0.5247962 -5.669207 1.188403e-07
```

The table above shows Ybar (Estimates) and si (std.Error).

```
##               [,1]        [,2]        [,3]
## [1,]  0.009031483  0.02853874 0.002573624
## [2,] -0.133604737 -0.30831658 0.219365735
## [3,] -0.092717316 -0.22104500 0.116488013
## [4,] -0.092569588 -0.17639150 0.059287701
```

each columns (left to right) represents ratio: 'si$^{2/yi}$','si/yi','si/yi^{2}'. each row (top to bottom) represents region 1, region 2, region 3.

```
## [1] 0.003698031
```

```
## [1] 0.02039787
```

```
## [1] 0.008556039
```

By checking the variance for each type of ration, we can find that ratio: si/yi is the most stable relationship So we may choose transformation: y'=log(y). Conclusion: the constant variance is violated,log transformation: y'=log(y) can be used for remedification.

(d)

```
##  [1] -1.0 -0.9 -0.8 -0.7 -0.6 -0.5 -0.4 -0.3 -0.2 -0.1  0.0  0.1  0.2  0.3  0.4
## [16]  0.5  0.6  0.7  0.8  0.9  1.0
```

```
##  [1] 46.80196 47.09448 47.45772 47.89413 48.40622 48.99657 49.66781 50.42263
##  [9] 51.26373 52.19386 53.21577 54.33220 55.54584 56.85938 58.27541 59.79643
## [17] 61.42485 63.16295 65.01282 66.97640 69.05545
```

```
## [1] -1
```

The best lambda = -1 which has lowest SSE = 46.80196. The Box-cox procedure shows the reciprocal transformation is reasonable.

(e) Fit the reciprocal transformation y'=1/y and obtain the ANOVA result.

```
## Anova Table (Type II tests)
##
## Response: reciprol_Y
##             Sum Sq  Df F value    Pr(>F)
## region    0.010349   3  14.788 3.815e-08 ***
## Residuals 0.025428 109
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The residual is 0.025428.

(f)

```
## Levene's Test for Homogeneity of Variance (center = "median")
##        Df F value Pr(>F)
## group   3  0.9694   0.41
##       109
```

H0: The geographic region variance for the reciprocal transformed length of stay are equal. H1: Not all region variance are the same. alpha=0.01 The resulting p-value=0.41>0.1. Conclude H0.