Homework 4 - Answer Key

Melinda Higgins

October 30, 2018

1. Perform a Simple Linear Regression for:

- OUTCOME variable `cesd`: "Center for Epidemiological Studies-Depression (CESD) total score - Baseline"
- PREDICTOR variable `indtot`: "Inventory of Drug Use Consequences (InDue) total score - Baseline"
- decide if you want to transform either variable `cesd` or `indtot` and if so, what transformation you applied and why - you can also decide not to transform (i.e. tradeoffs between model fit and interpretability of your results) - discuss your reasoning.

1. Answer

Here is the code and output after running a simple linear regression using `lm()` function.

```
library(tidyverse)
library(haven)

helpdat <- haven::read_spss("helpmkh.sav")

# create subset
# select indtot, cesd and racegrp

h1 <- helpdat %>%
  select(indtot, cesd, racegrp)

# run simple linear regression
# using the lm
# save the results in the fit1 object
fit1 <- lm(cesd ~ indtot, data=h1)
```

Model summary

```
# look at a summary() of the model
summary(fit1)
```

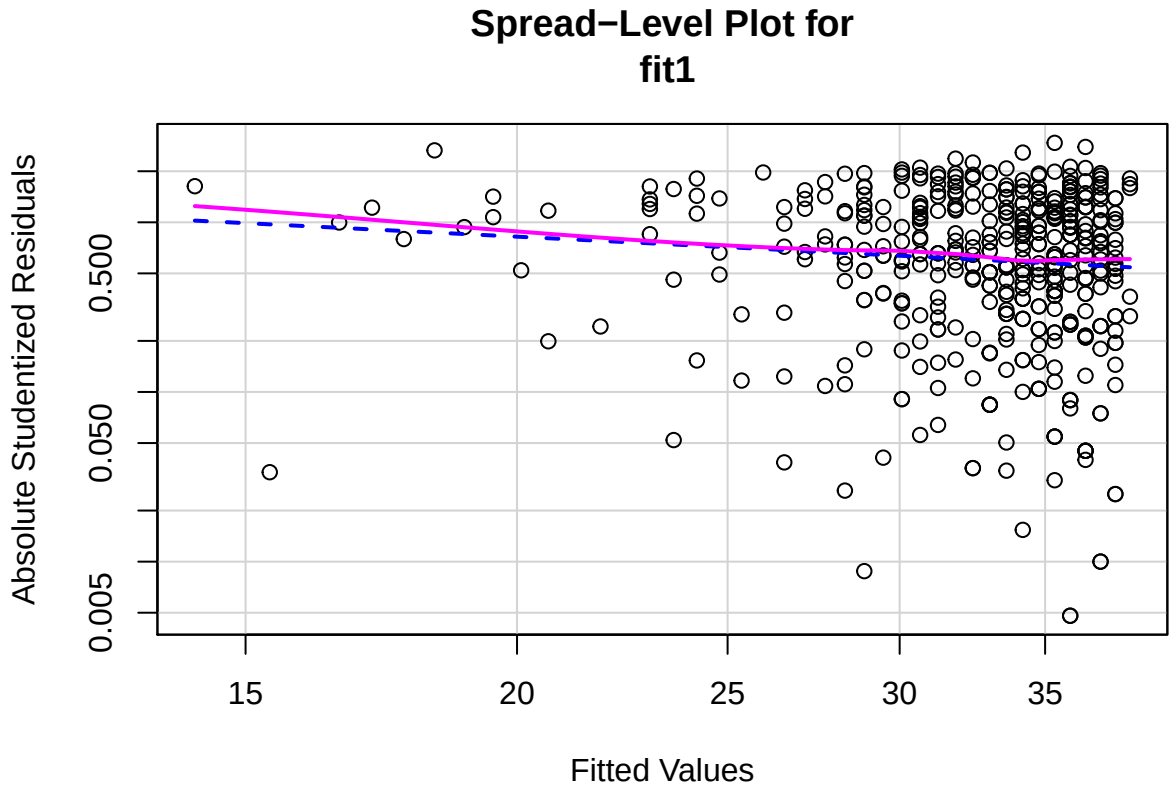
```
##
## Call:
## lm(formula = cesd ~ indtot, data = h1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -34.356  -7.658   0.644   8.057  30.674
##
## Coefficients:
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 11.86597    2.82788   4.196 3.27e-05 ***
## indtot       0.58725    0.07761   7.566 2.18e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.8 on 451 degrees of freedom
## Multiple R-squared:  0.1126, Adjusted R-squared:  0.1107
## F-statistic: 57.25 on 1 and 451 DF,  p-value: 2.176e-13
```

Possible Transformation

I probably would not do a transformation since the residuals look fairly normal (see diagnostic plots below), but you could run the `spreadLevelPlot()` function from the `car` package to see if a power transformation is suggested. This suggests a power transformation of 1.639 which could be rounded up to 2. This is optional and not needed for this data.

```
library(car)
# look at the spreadLevelPlot
# this also provides a suggestion of
# possible power transformation
car::spreadLevelPlot(fit1)
```




```
##
## Suggested power transformation:  1.63886
```

Plus the `gvlma()` function from the `gvlma` package can be run to check model assumptions, which also all look ok.

```
# global test of linear model assumptions
# install gvlma package
library(gvlma)
gvmodel <- gvlma::gvlma(fit1)
summary(gvmodel)
```

```
##
## Call:
## lm(formula = cesd ~ indtot, data = h1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -34.356  -7.658   0.644   8.057  30.674
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 11.86597    2.82788   4.196 3.27e-05 ***
## indtot       0.58725    0.07761   7.566 2.18e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.8 on 451 degrees of freedom
## Multiple R-squared:  0.1126, Adjusted R-squared:  0.1107
## F-statistic: 57.25 on 1 and 451 DF,  p-value: 2.176e-13
##
##
## ASSESSMENT OF THE LINEAR MODEL ASSUMPTIONS
## USING THE GLOBAL TEST ON 4 DEGREES-OF-FREEDOM:
## Level of Significance =  0.05
##
## Call:
##  gvlma::gvlma(x = fit1)
##
##                      Value p-value                Decision
## Global Stat        6.65298  0.1554 Assumptions acceptable.
## Skewness           2.17129  0.1406 Assumptions acceptable.
## Kurtosis           2.89772  0.0887 Assumptions acceptable.
## Link Function      1.52631  0.2167 Assumptions acceptable.
## Heteroscedasticity 0.05766  0.8102 Assumptions acceptable.
```

2. Perform regression diagnostics:

- check the normality of the residuals (histogram and Q-Q plots)
- check for linearity - is there any systematic relationship between the residuals and the predicted (or fitted) values?
- homoscedasticity - plot of standardized residuals versus fitted values - this is known as a "Scale-Location" graph.
- check for outliers and data points with high leverage or influence: outliers are often identified with standardized residuals > 3 (or <-3) and influential observations are often identified using Cook's D

2. Answer

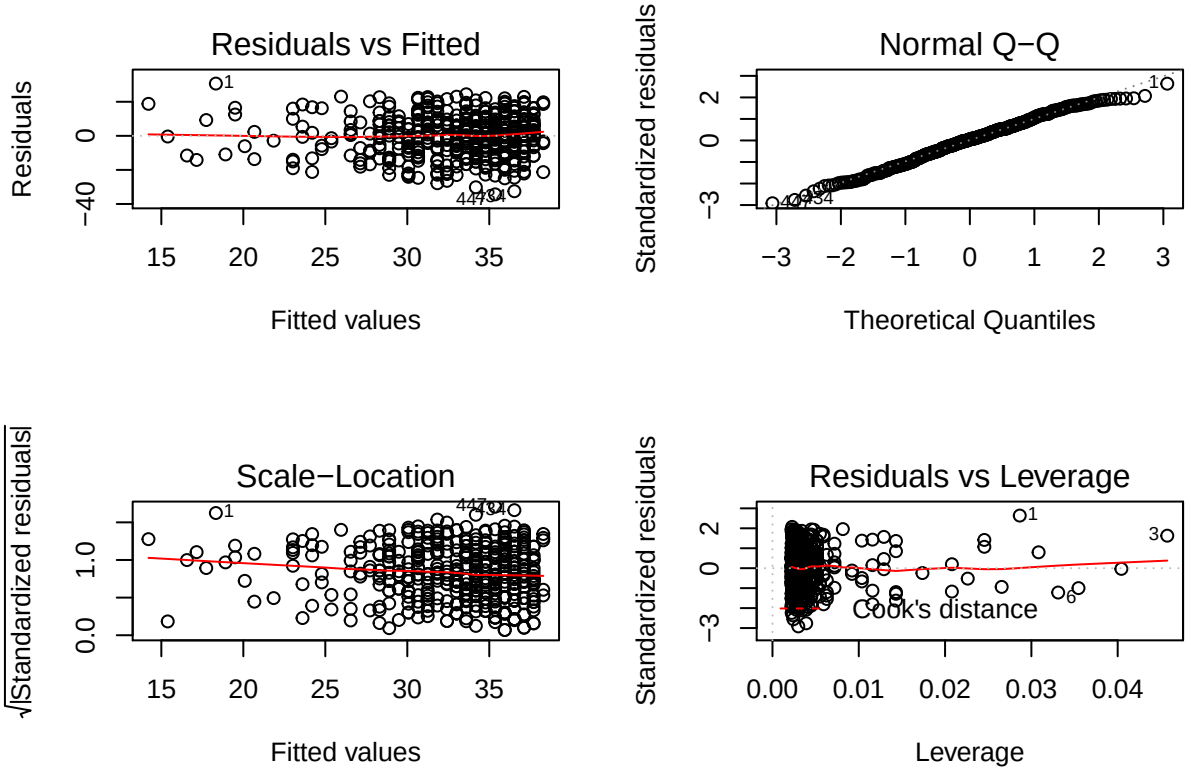
A good set of 4 diagnostic plots can be obtained using the `plot()` function for the fitted model output.

Diagnostic plots

These 4 diagnostic plots show:

- residuals vs fitted values - the line here is flat and shows no obvious trend, but the data do cluster on the higher end of the fitted values than the lower end, indicating some skewness
- normal Q-Q plot of the residuals - this plot looks fairly linear indicating a close to normal distribution
- scale-location plot - there is a slight trend downwards, but this slope is minor - and in general the variability looks pretty consistent across all of the fitted values (no obvious heteroscedasticity)
- the last plot of Cook's distance does highlight a few possible outliers - cases 1, 3 and 6, in the but these appear to be minor as they are not obvious in the Q-Q plot nor in the histogram

```
# get diagnostic plots
par(mfrow=c(2,2))
plot(fit1)
```

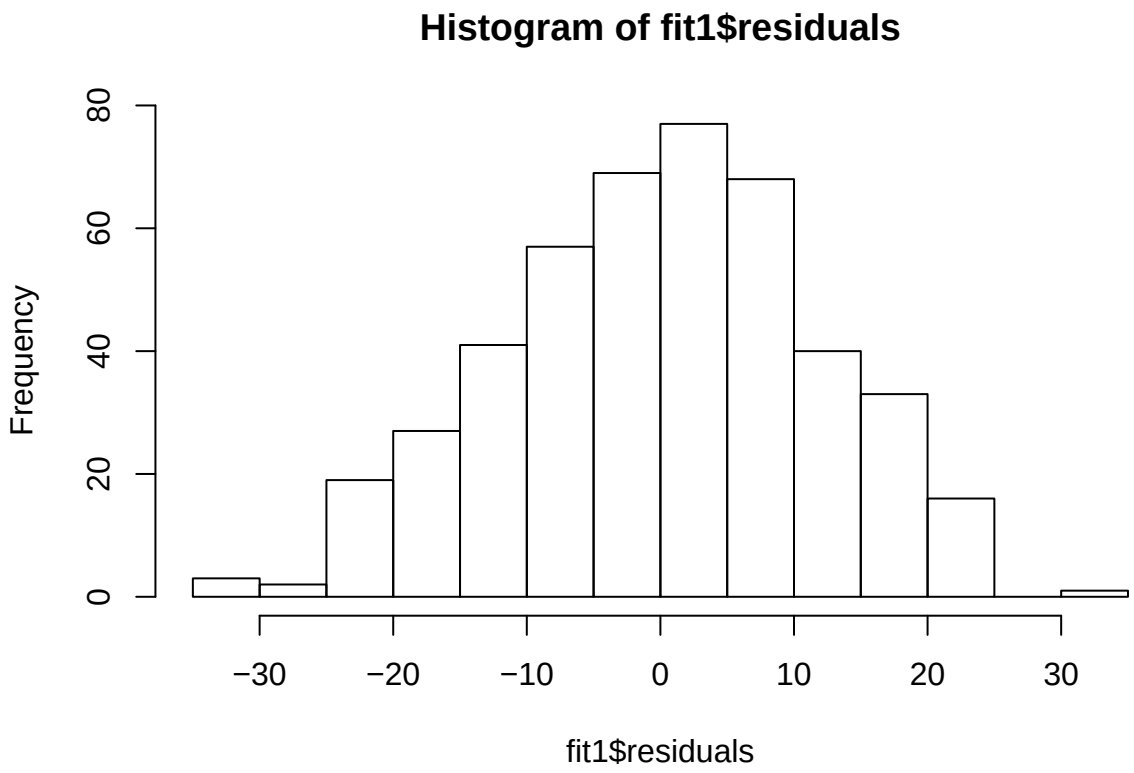



```
# reset par
par(mfrow=c(1,1))
```

Histogram of the Residuals

The histogram of the residuals look normal. No skewness and no obvious outliers.

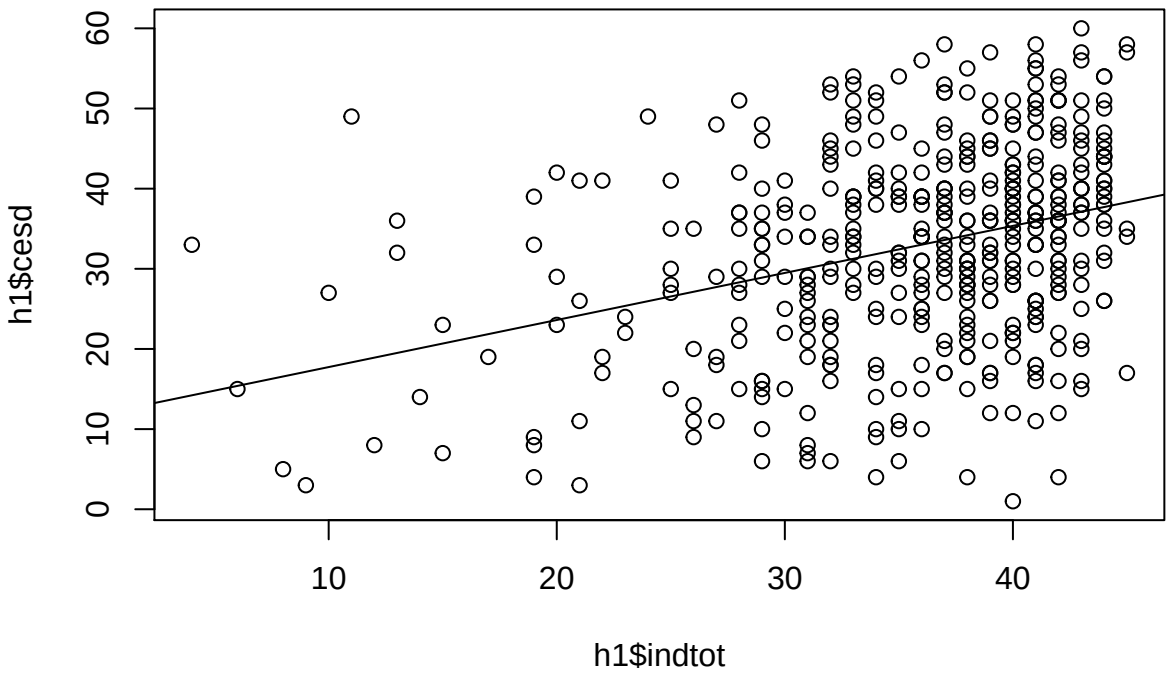
```
# histogram of the residuals
hist(fit1$residuals)
```



Scatterplot of Data with Fitted Line

There is quite a bit of scatter in the plot with a lot of variability in both `indtot` and `cesd`, so a linear fit line looks as good as any - there is no obvious curvature to the data.

```
plot(h1$indtot, h1$cesd)
abline(lm(h1$cesd ~ h1$indtot))
```



3. Provide a summary of the regression results.

- provide a **FIGURE** of the model, in this case a scatterplot with the fitted line overlaid and 95% confidence intervals if you can
- Make a **TABLE** presenting the fitted regression model (coefficients and tests of significance for those coefficients)
- describe the variance explained by the model (based on r2)
- describe the model itself based on the y-intercept and slope terms
- note any limitations or issues with the model fit or interpretation of the model

Model fit with 95% confidence intervals for fitted line

```
# using a ggplot2 approach
ggplot(h1, aes(indtot, cesd)) +
  geom_point() +
  stat_smooth(method = lm)
```

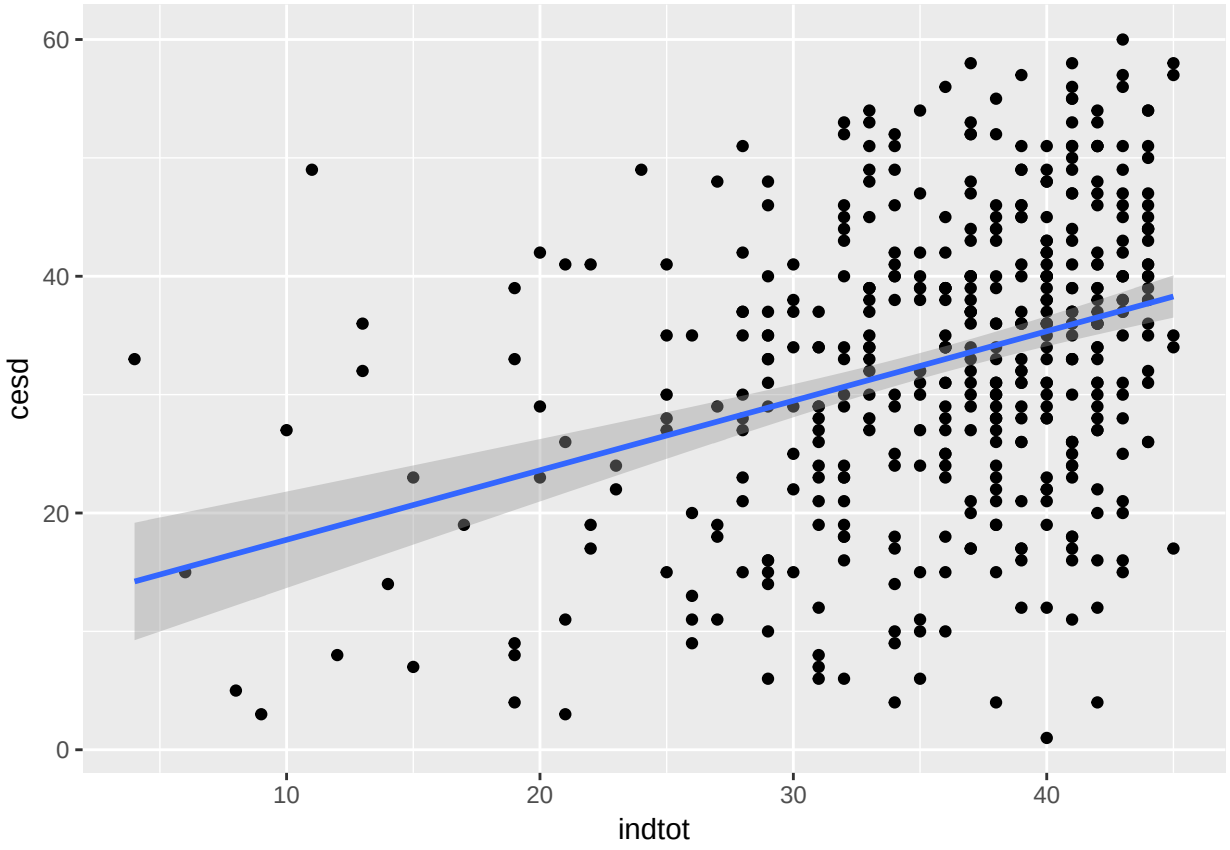


Table of the model fit

```
library(knitr)
library(xtable)
print(xtable(summary(fit1)), type = "html")
```

Estimate

Std. Error

t value

Pr(>|t|)

(Intercept)

11.8660

2.8279

4.20

0.0000

indtot

0.5873

0.0776

7.57

0.0000

Variance explained

The variablity explained by this model is 11.0676446%, which is the adjusted r2 captured as `summary(fit1)$adj.r.squared`.

Describe the model

For `indtot` scores equal to 0, the model estimates that a subject would have a `cesd` score of 11.8659738 given the y-intercept. For each 1 point increase in `indtot` scores, the `cesd` score will increase on average by 0.5872545 based on the slope estimate.

Model fit and Any Other issues

It is worth noting that given the wide variability in both the `indtot` and `cesd` a linear fit line indicates a weak positive correlation between these 2 variables, but the residuals do show wide variation about the best fitted line, indicating that the linear trend is weak at best.

4. Perform a One-way ANOVA for:

- OUTCOME variable `cesd`: "Center for Epidemiological Studies-Depression (CESD) total score - Baseline"
- GROUP variable `racegrp`: "Racial Group of Respondent"
- options - you can use either an ANOVA or GLM modeling approach
- if the GROUP variable is significant, also perform *post hoc* tests - use some kind of pairwise error rate adjustment (i.e. bonferroni, sidak, Tukey's HSD, etc) - be sure to report which one you used and why

ANOVA Model Results with dummy coding

```
# one-way ANOVA
# we can use the lm() function
# it does "dummy" coding on the fly
# run racegrp as either the character
# type or as a factor - either will work
fit2.lm <- lm(cesd ~ racegrp, data=h1)
summary(fit2.lm)
```

```
##
## Call:
## lm(formula = cesd ~ racegrp, data = h1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -31.3012  -8.3012   0.8199   8.6400  27.8199
##
## Coefficients:
```

```
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      30.1801     0.8468  35.642  < 2e-16 ***
## racegrphispanic   4.1799     1.9346   2.161   0.0313 *
## racegrpother      5.7430     2.5565   2.246   0.0252 *
## racegrpwhite      5.1211     1.2761   4.013 7.02e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.3 on 449 degrees of freedom
## Multiple R-squared:  0.04042,    Adjusted R-squared:  0.03401
## F-statistic: 6.304 on 3 and 449 DF,  p-value: 0.0003396
```

ANOVA results for group effect overall

```
# the aov() function
# gives the global test for the "group" effect
fit2.aov <- aov(cesd ~ racegrp, data=h1)
summary(fit2.aov)
```

```
##              Df Sum Sq Mean Sq F value  Pr(>F)
## racegrp       3   2861   953.7   6.304 0.00034 ***
## Residuals   449  67927   151.3
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Yes, `racegrp` is significant - there are significant differences between the 4 races.

Post hoc tests

Since `racegrp` was significant, let's run all of the pairwise comparisons. The code below will show the options for adjusting the error-rate due to multiple pairwise comparisons including: Bonferroni, Holm, and Tukey HSD.

```
# post hoc tests
# Tukey HSD
TukeyHSD(fit2.aov)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = cesd ~ racegrp, data = h1)
##
## $racegrp
##                       diff        lwr       upr     p adj
## hispanic-black  4.1799052 -0.8088317  9.168642 0.1359780
## other-black     5.7429821 -0.8494072 12.335372 0.1125093
## white-black     5.1211100  1.8305336  8.411686 0.0004071
## other-hispanic  1.5630769 -6.1057988  9.231953 0.9528809
## white-hispanic  0.9412048 -4.1754298  6.057839 0.9647000
## white-other    -0.6218721 -7.3115695  6.067825 0.9951504
```

```r
# using Bonferroni error-rate correction
pairwise.t.test(h1$cesd, h1$racegrp, p.adj = "bonf")
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  h1$cesd and h1$racegrp
##
##          black   hispanic other
## hispanic 0.18753 -        -
## other    0.15097 1.00000  -
## white    0.00042 1.00000  1.00000
##
## P value adjustment method: bonferroni
```

```r
# using the Holm error-rate correction
pairwise.t.test(h1$cesd, h1$racegrp, p.adj = "holm")
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  h1$cesd and h1$racegrp
##
##          black   hispanic other
## hispanic 0.12581 -        -
## other    0.12581 1.00000  -
## white    0.00042 1.00000  1.00000
##
## P value adjustment method: holm
```

These pairwise comparisons show that there are significant differences between white-black, but none of the other pairwise comparisons were significant.

5. Perform model diagnostics:

- homoscedasticity - look at a test for equal variance (Levene's test or Bartlett's test or equivalent).
- if this test of equal variances fails, you may want to report a modified F-test (e.g. Welch's test)

Test of Equal Variances

```r
# barlett's test for homogenity of variances
# note: put the formula back in
bartlett.test(cesd ~ racegrp, data=h1)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  cesd by racegrp
## Bartlett's K-squared = 3.4367, df = 3, p-value = 0.3291
```

This was not significant, so we can report the usual F-statistic test for `racegrp` rather than the "robust" Welch's F-test.

6. Present a summary of the ANOVA results.

- Make a **FIGURE** of the group mean differences - either an error-bar plot or a series of boxplots one for each group to show the group differences in the outcome
- Make a **TABLE** presenting the ANOVA results
- describe the model results - was the GROUP (`racegrp`) significant?
- If GROUP is significant, what did the post hoc tests reveal?

Plot of Means and 95% Confidence Intervals

```
# get a means plot using
# plotmeans() from gplots package
library(gplots)
gplots::plotmeans(cesd ~ racegrp,
                  data=h1)
```

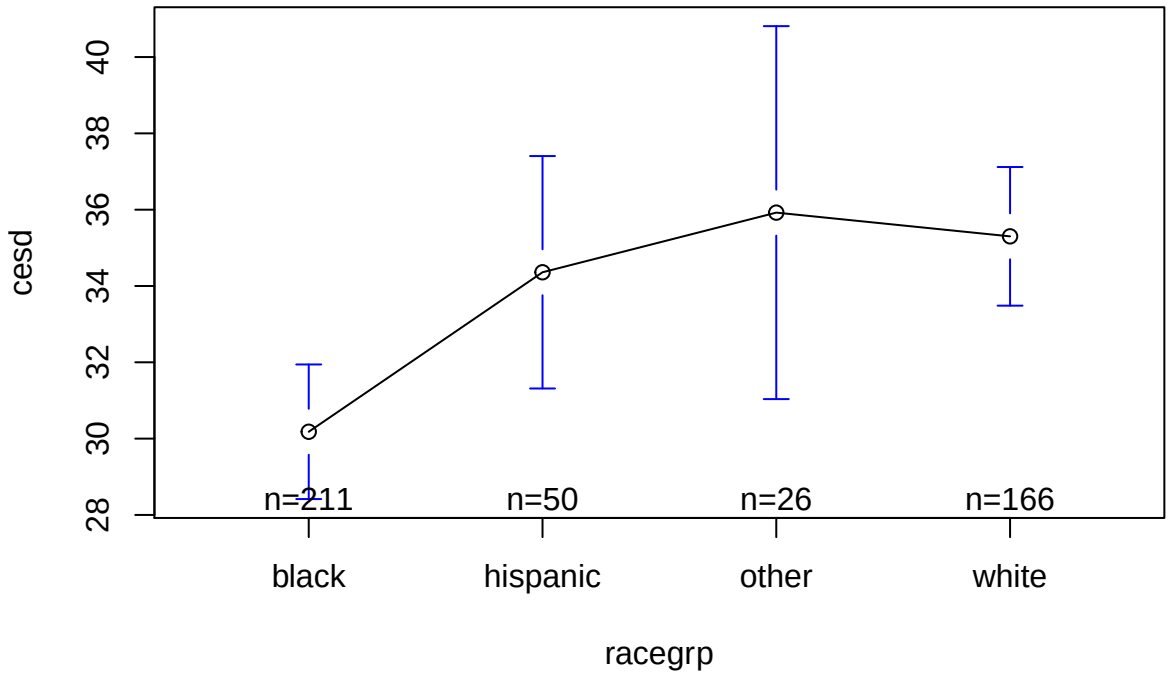


Table of the ANOVA results - overall group effect

`racegrp` was significant - no significant diff

```
print(xtable(summary(fit2.aov)),
      type = "html")
```

Df

Sum Sq

Mean Sq

F value

Pr(>F)

racegrp

3

2861.03

953.68

6.30

0.0003

Residuals

449

67927.46

151.29