---

title: "HW 4"

author: "Gaea Daniel"

date: "10/18/2017"

output:

html\_document: default

pdf\_document: default

word\_document: default

---

```{r setup, include=FALSE}

knitr::opts\_chunk$set(echo = TRUE)

library(tableone)

library(Matrix)

library(survey)

library(survival)

library(testthat)

library(dummies)

library(grid)

library(Matching)

library(reshape2)

library(ggplot2)

library(knitr)

library(geepack)

library(lme4)

library(tidyverse)

library(haven)

library(lattice)

#Load data

help.spss <- read\_csv("helpmkh.csv")

data(help.spss)

```

```{r}

#To access R-commander

library(Rcmdr)

library(splines)

library(RcmdrMisc)

library(car)

```

### Perform a Simple Linear Regression for:

OUTCOME variable indtot: "Inventory of Drug Use Consequences (InDue) total score - Baseline"

PREDICTOR variable mcs: "SF36 Mental Composite Score - Baseline"

```{r}

#Simple linear regression

RegModel.1 <- lm(indtot~mcs, data=helpmkh)

summary(RegModel.1)

```

Call:

lm(formula = indtot ~ mcs, data = helpmkh)

Residuals:

Min 1Q Median 3Q Max

-28.607 -3.388 1.422 4.448 11.427

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 42.44479 0.82895 51.203 <2e-16 \*\*\*

mcs -0.21203 0.02426 -8.741 <2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.621 on 451 degrees of freedom

Multiple R-squared: 0.1449, Adjusted R-squared: 0.143

F-statistic: 76.41 on 1 and 451 DF, p-value: < 2.2e-16

Decide if you want to transform either variable indtot or mcs and if so, what transformation you applied and why.

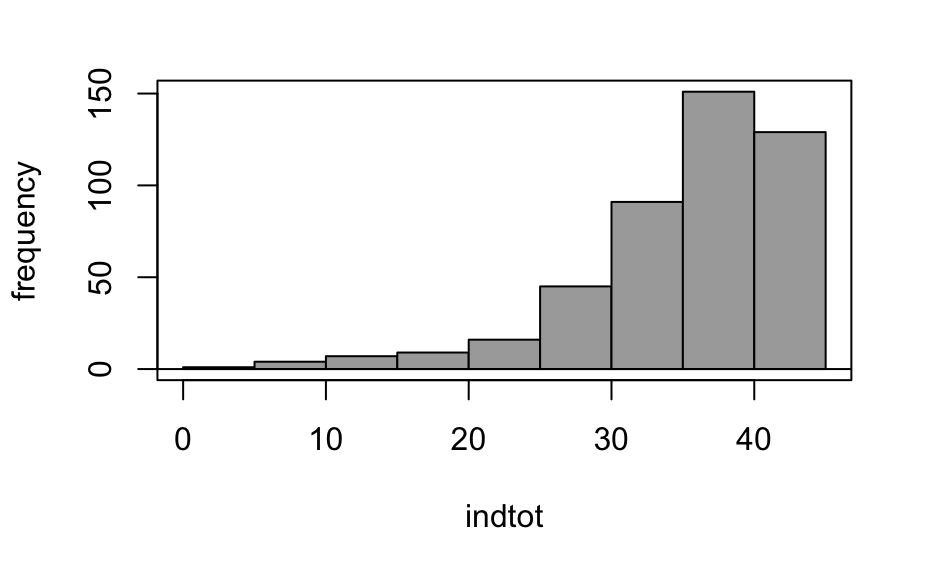
Histograms were used to check the distribution of indtot and mcs. Mcs had a fairly normal distribution, but indtot was clearly left-skewed. Transformations (log, natural log, the mathematical constant, and square root) of indtot were created to normalize the distribution.

```{r}

#Histogram for indtot to check distribution

with(helpmkh, Hist(indtot, scale="frequency", breaks="Sturges",

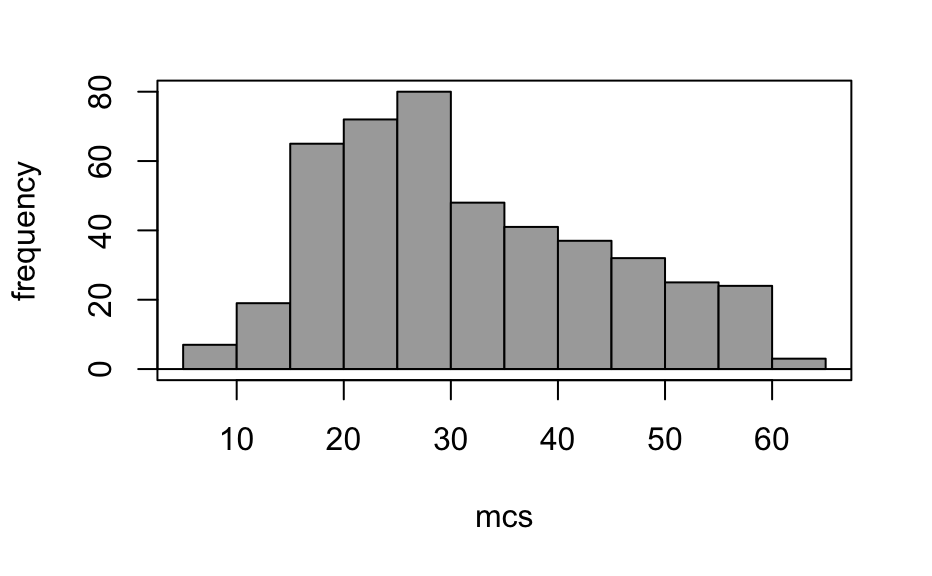
col="darkgray"))



#Histogram for mcs to check for distribution

with(helpmkh, Hist(mcs, scale="frequency", breaks="Sturges",

col="darkgray"))

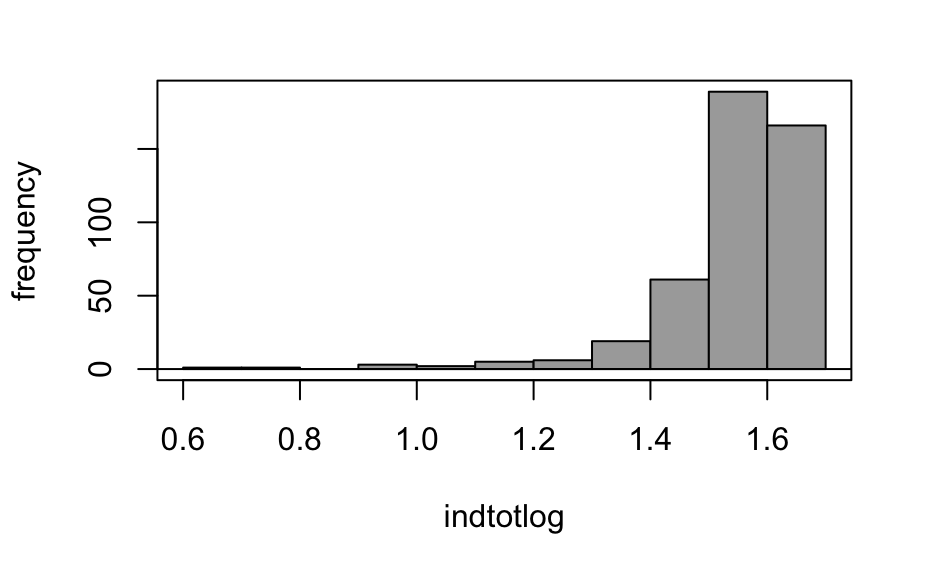


#Log transformation and histogram for indtot

helpmkh$indtotlog <- with(helpmkh, log10(indtot))

with(helpmkh, Hist(indtotlog, scale="frequency",

breaks="Sturges", col="darkgray"))

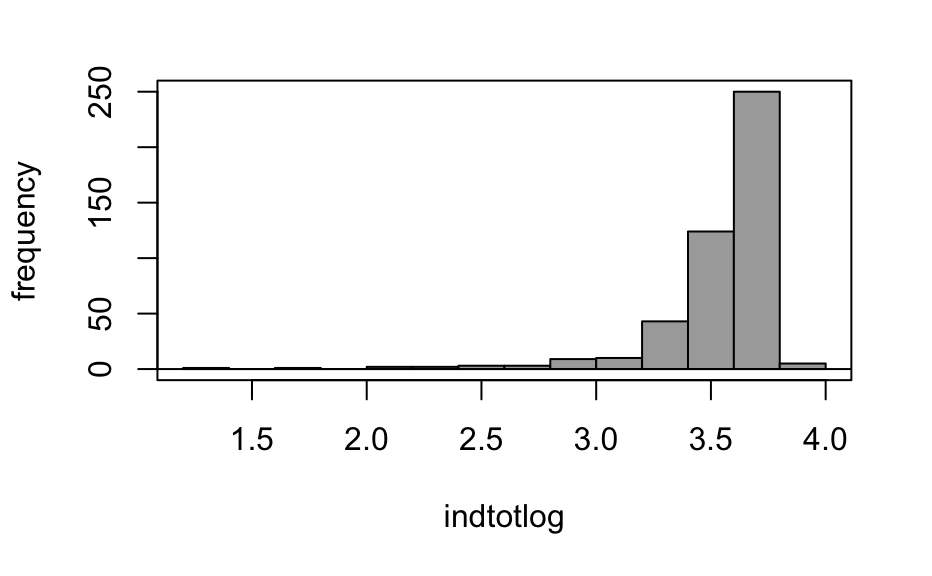


#Natural log transformation and histogram for indtot

helpmkh$indtotlog <- with(helpmkh, log(indtot))

with(helpmkh, Hist(indtotlog, scale="frequency",

breaks="Sturges", col="darkgray"))

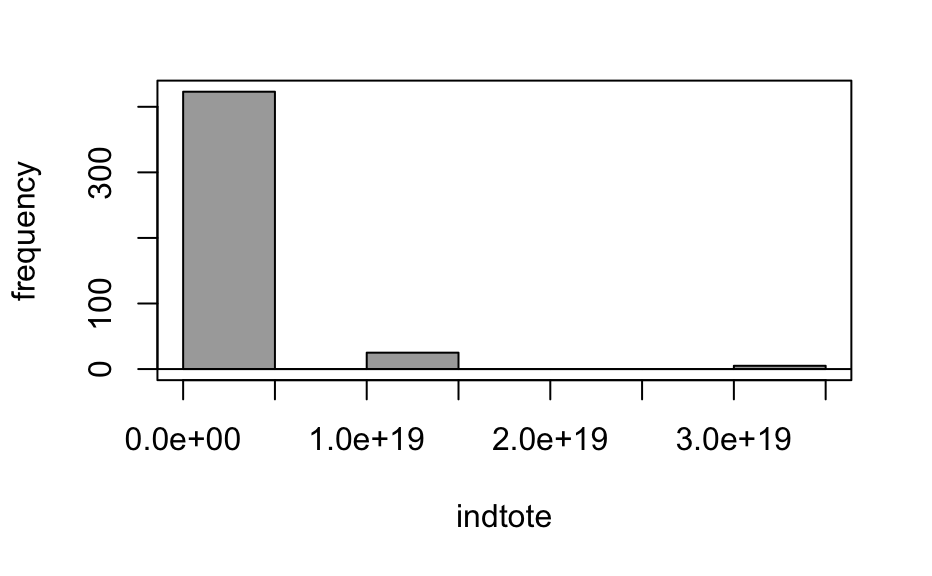


#e transformation and histogram for indtot

helpmkh$indtote <- with(helpmkh, exp(indtot))

with(helpmkh, Hist(indtote, scale="frequency", breaks="Sturges",

col="darkgray"))



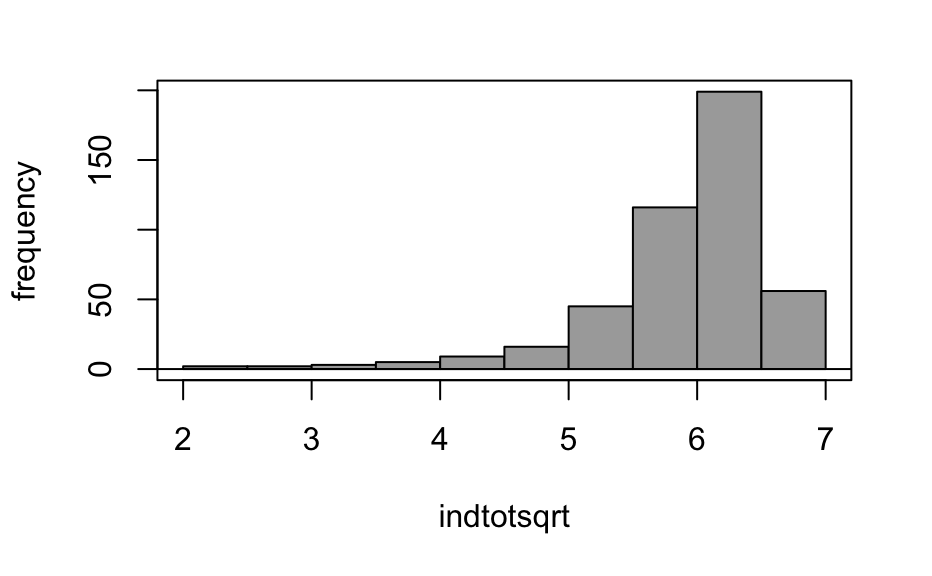
#Square root transformation and histogram for indtot

helpmkh$indtotsqrt <- with(helpmkh, sqrt(indtot))

with(helpmkh, Hist(indtotsqrt, scale="frequency",

breaks="Sturges", col="darkgray"))

```



The square root transformation of indtot appeared to normalize the indtot variable the best and was used for the simple linear regression.

```{r}

#Simple linear regression with square root transformation of indtot and mcs

RegModel.5 <- lm(indtotsqrt~mcs, data=helpmkh)

summary(RegModel.5)

```

Call:

lm(formula = indtotsqrt ~ mcs, data = helpmkh)

Residuals:

Min 1Q Median 3Q Max

-3.5150 -0.2476 0.1480 0.3932 1.0262

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 6.568675 0.079266 82.868 <2e-16 \*\*\*

mcs -0.019898 0.002319 -8.579 <2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6331 on 451 degrees of freedom

Multiple R-squared: 0.1403, Adjusted R-squared: 0.1384

F-statistic: 73.59 on 1 and 451 DF, p-value: < 2.2e-16

### Perform regression diagnostics:

Check the normality of the residuals (histogram and Q-Q plots)

```{r}

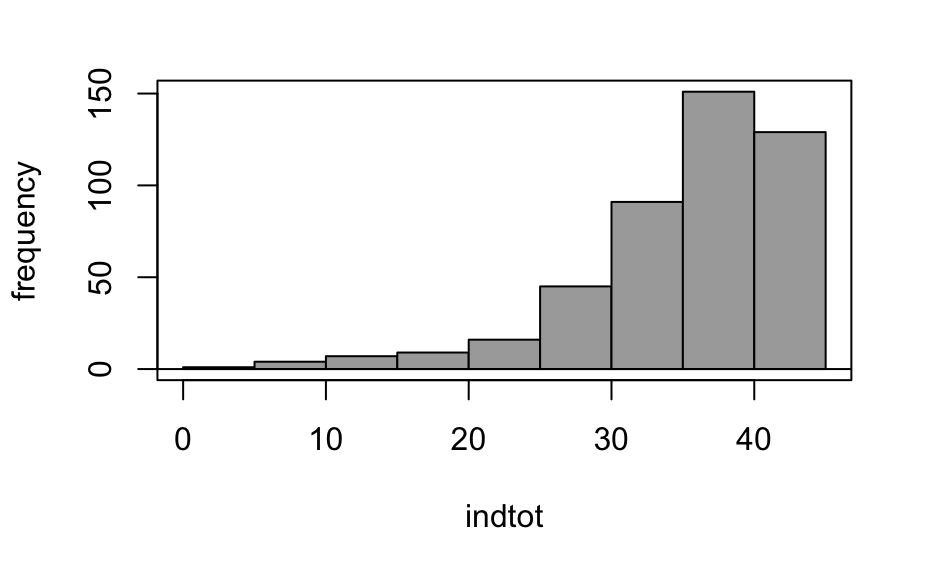
#Histogram and Q-Q plots for indtot

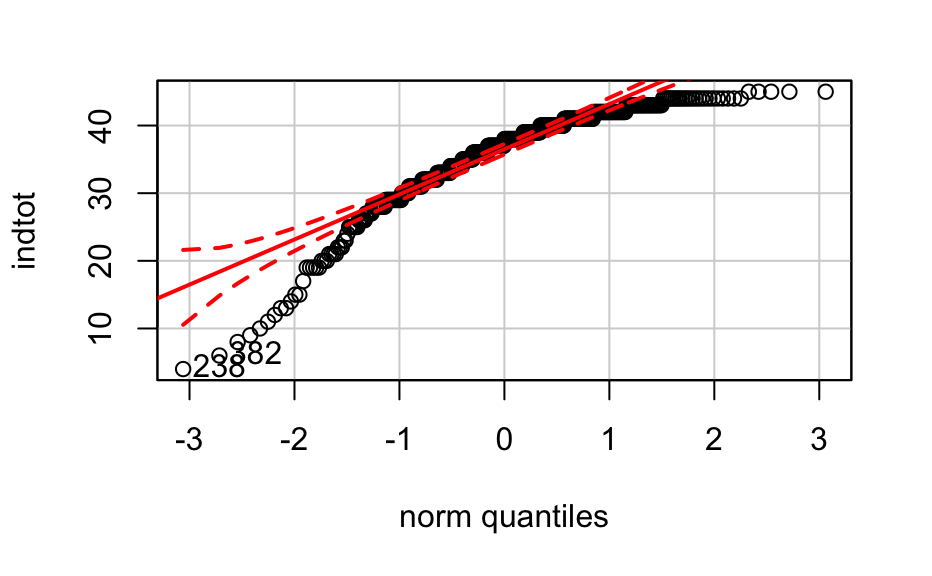
with(helpmkh, Hist(indtot, scale="frequency", breaks="Sturges",

col="darkgray"))

with(helpmkh, qqPlot(indtot, dist="norm", id.method="y", id.n=2,

labels=rownames(helpmkh)))





#Histogram and Q-Q plots for mcs

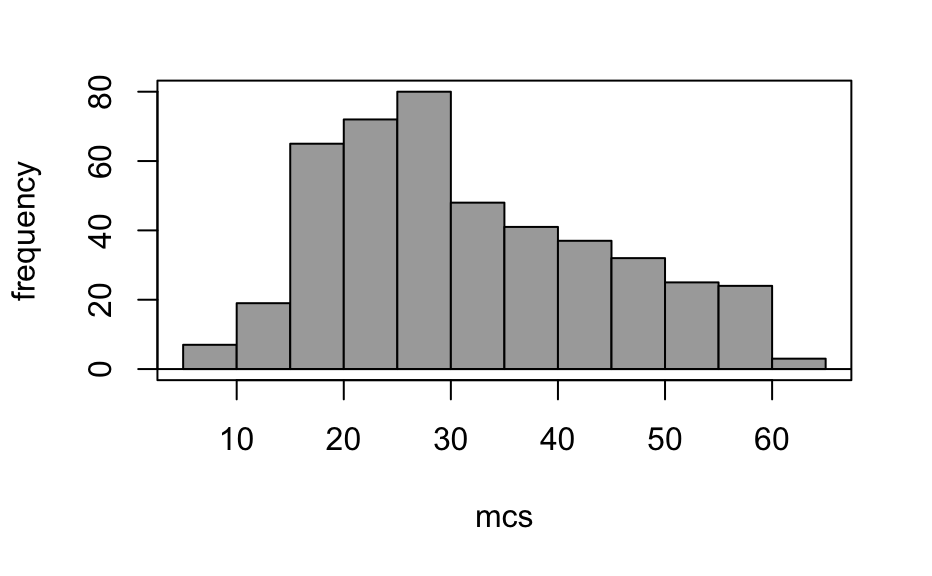
with(helpmkh, Hist(mcs, scale="frequency", breaks="Sturges",

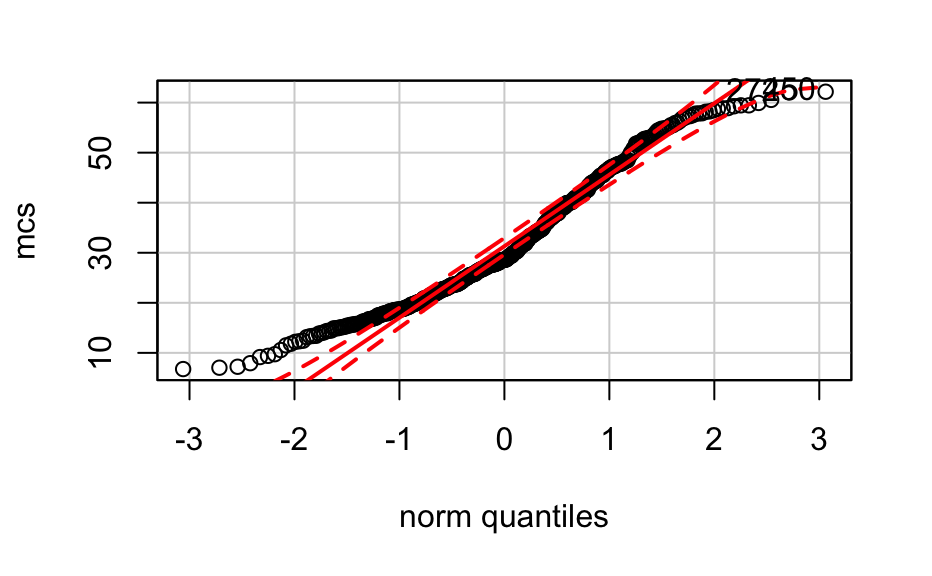
col="darkgray"))

with(helpmkh, qqPlot(mcs, dist="norm", id.method="y", id.n=2,

labels=rownames(helpmkh)))

```





Check for linearity - is there any systematic relationship between the residuals and the predicted (or fitted) values? No

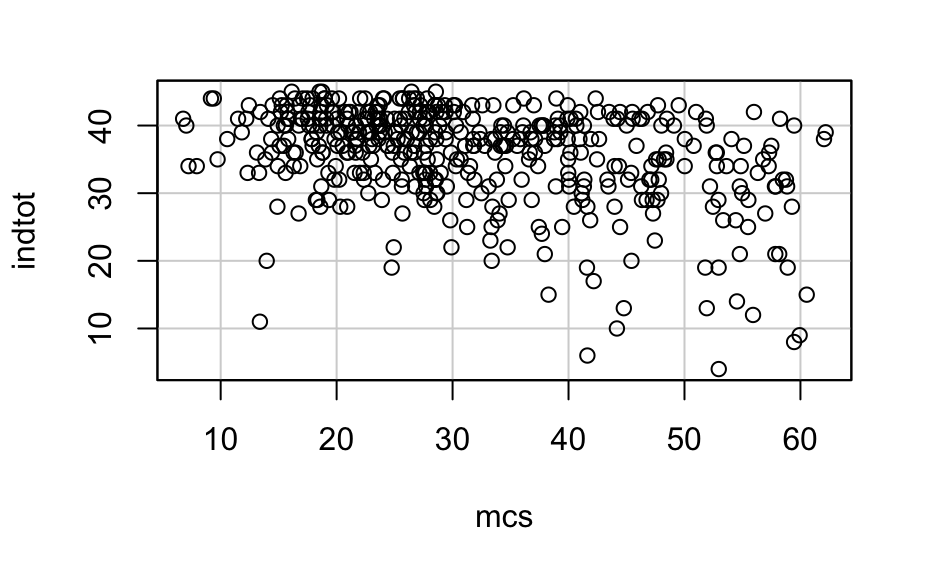
```{r}

#Whitout the line

scatterplot(indtot~mcs, reg.line=FALSE, smooth=FALSE,

spread=FALSE, boxplots=FALSE, span=0.5, ellipse=FALSE,

levels=c(.5, .9), data=helpmkh)



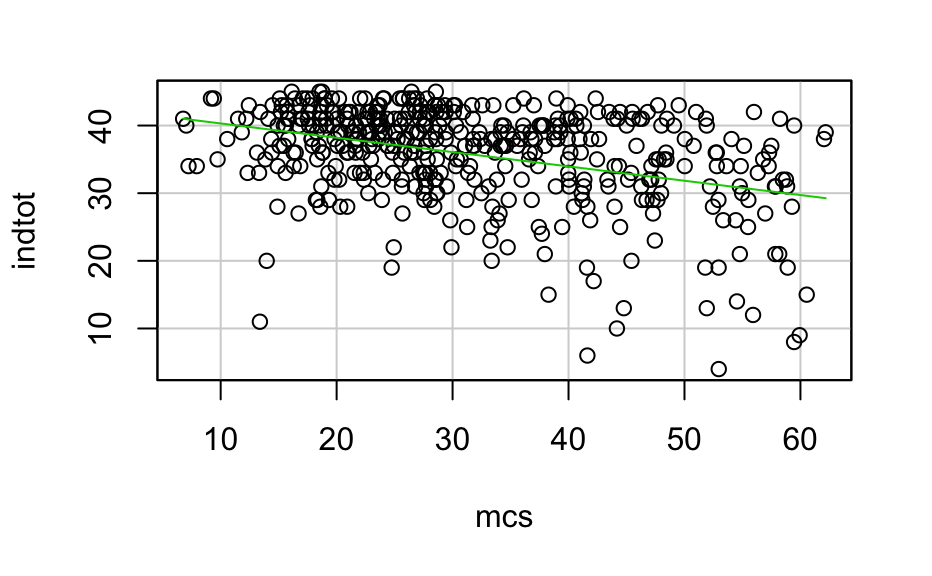
#With the line

scatterplot(indtot~mcs, reg.line=lm, smooth=FALSE, spread=FALSE,

boxplots=FALSE, span=0.5, ellipse=FALSE, levels=c(.5, .9),

data=helpmkh)

```



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

```{r}

# Change the layout to 2x2 to accommodate all plots

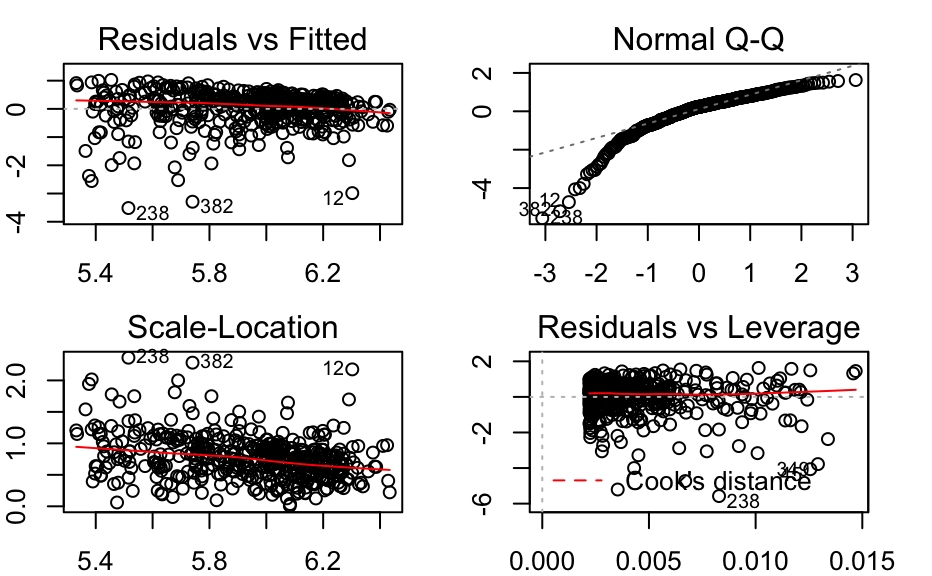
par(mfrow=c(2,2))

par(mar = rep(2, 4))

# Diagnostic Plots

plot(RegModel.5)

```



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

See above

Perform a One-way ANOVA for:

OUTCOME variable indtot: "Inventory of Drug Use Consequences (InDue) total score - Baseline"

GROUP variable racegrp: "Racial Group of Respondent"

I would suggest merging "other" and "hispanic" together and create a 3-group variable for race, since the "other" category is only about 6% of the sample.

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

Bonferroni is used because the significant difference may not exist between all conditions.

```{r}

#One-way ANOVA

AnovaModel.7 <- aov(indtot ~ racegrp, data=helpmkh)

summary(AnovaModel.7)

with(helpmkh, numSummary(indtot, groups=racegrp,

statistics=c("mean", "sd")))

#Bonferroni

outlierTest(AnovaModel.7)

```

Df Sum Sq Mean Sq F value Pr(>F)

racegrp 3 1069 356.3 7.254 0.0000923 \*\*\*

Residuals 449 22053 49.1

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

mean sd data:n

black 34.27014 7.716279 211

hispanic 35.14000 7.119992 50

other 36.88462 6.476585 26

white 37.57831 6.038504 166

rstudent unadjusted p-value Bonferonni p

238 -4.417877 0.000012514 0.0056689

382 -4.114543 0.000046177 0.0209180

438 -4.012875 0.000070294 0.0318430

Perform model diagnostics:

homoscedasticity - look at a test for equal variance (Levene's test or Bartlett's test).

if this test of equal variances fails, you may want to report a modified F-test (e.g. Welch's test)

Present a summary of the ANOVA results.

```{r}

#Bartlett's test

with(helpmkh, tapply(indtot, racegrp, var, na.rm=TRUE))

bartlett.test(indtot ~ racegrp, data=helpmkh)

#Levene's test

with(helpmkh, tapply(indtot, racegrp, var, na.rm=TRUE))

leveneTest(indtot ~ racegrp, data=helpmkh, center="median")

#Two-variance F-test

with(helpmkh, tapply(indtot, treat, var, na.rm=TRUE))

var.test(indtot ~ treat, alternative='two.sided', conf.level=.95,

data=helpmkh)

``````

black hispanic other white

59.54096 50.69429 41.94615 36.46353

Bartlett test of homogeneity of variances

data: indtot by racegrp

Bartlett's K-squared = 11.032, df = 3, p-value =

0.01155

black hispanic other white

59.54096 50.69429 41.94615 36.46353

group coerced to factor.Levene's Test for Homogeneity of Variance (center = "median")

Df F value Pr(>F)

group 3 1.7579 0.1545

449

0 1

50.29098 52.21679

F test to compare two variances

data: indtot by treat

F = 0.96312, num df = 227, denom df = 224, p-value =

0.7778

alternative hypothesis: true ratio of variances is not equal to 1

95 percent confidence interval:

0.7412634 1.2510860

sample estimates:

ratio of variances

0.963119

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

```{r}

#Boxplot

with(helpmkh, tapply(indtot, treat, var, na.rm=TRUE))

leveneTest(indtot ~ treat, data=helpmkh, center="median")

Boxplot(indtot~racegrp, data=helpmkh, id.method="y")

```

0 1

50.29098 52.21679

group coerced to factor.Levene's Test for Homogeneity of Variance (center = "median")

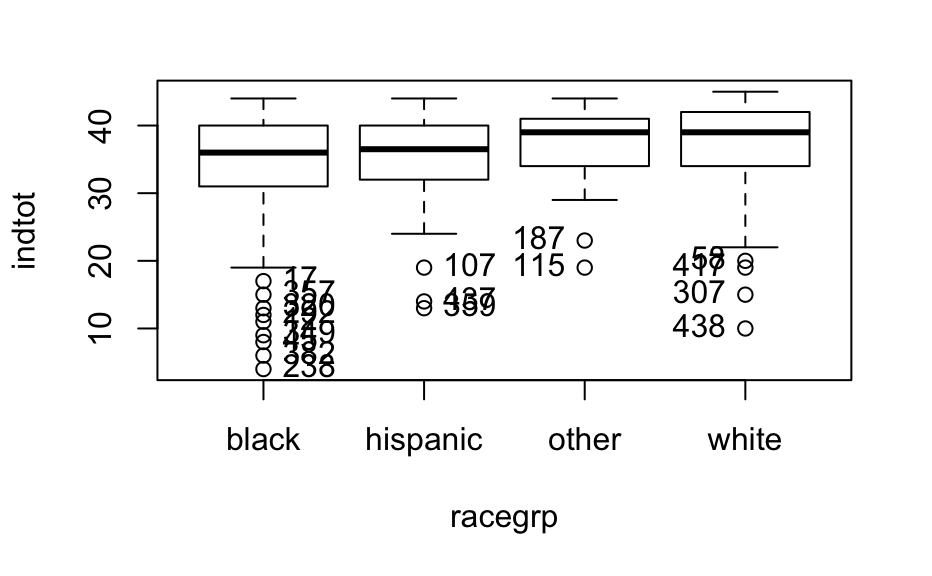
Df F value Pr(>F)

group 1 0.0011 0.9735

451

[1] "12" "17" "45" "238" "292" "320" "349" "357" "382"

[10] "107" "359" "437" "115" "187" "58" "307" "417" "438"



Make a TABLE presenting the ANOVA results

describe the model results - was the GROUP (racegrp) significant? Yes

If GROUP is significant, what did the post hoc tests reveal? The post hoc test (Bonferonni) revealed that it was not significant.