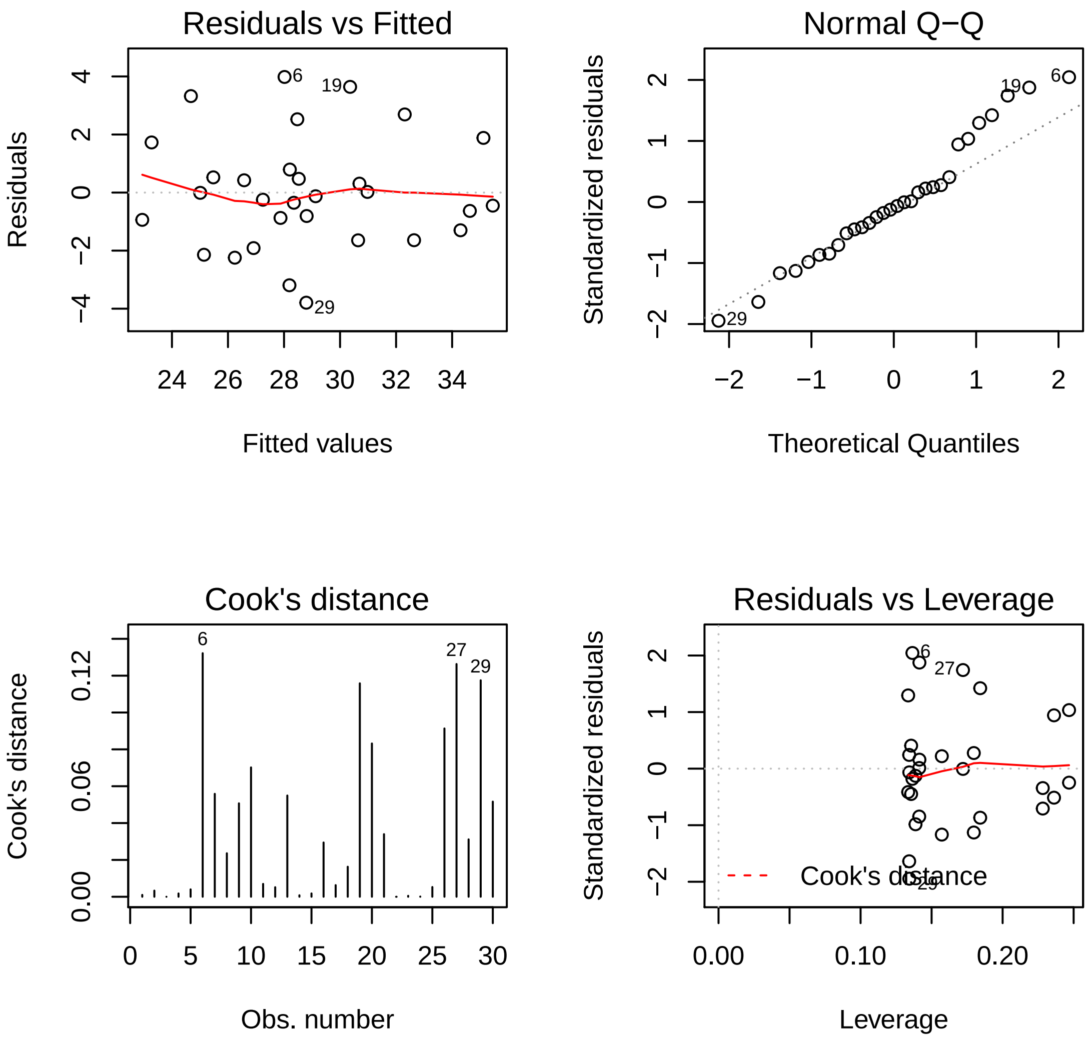
STAT7030 Assignment 1

Yijin Liu, Rui Qiu, Di Zhao

2017-08-28

Q1

(a)



The plot of the residuals against the fitted values shows that the variance seems to be relatively large in the middle. Generally, the scatter points tend to form a eclipse in the graph, rather than a rectangle. So our assumption of homoscedasticity is challenged.

As for Q-Q plot, we notice that several observations on the top right are a little far from the line and might be a problem, but most of points are along the diagonal line. This issue is worth checking in further study.

The Cook's distances of 4 observations appear relatively large to others. However, the vertical scale on this plot only goes to just around 0.12, which is not large at all for Cook's distance. So we claim there is no obvious problem.

The plot of the standardized residuals against the leverages also has not detected any suspicious points (as no observations appear be "outside" the line of Cook's distance).

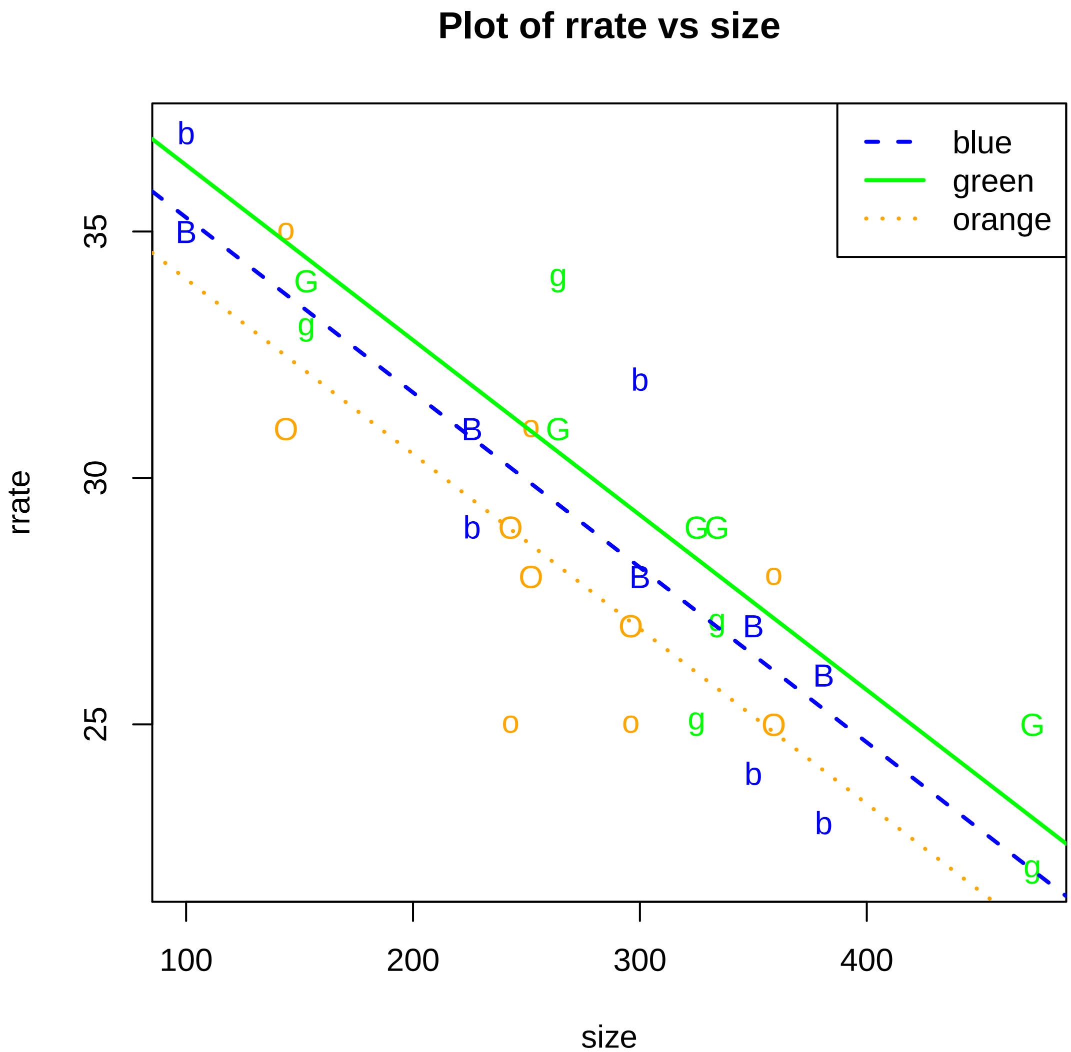
(b)

```
## Analysis of Variance Table
##
## Response: rrate
##           Df Sum Sq Mean Sq F value    Pr(>F)
## colours    2   3.27    1.63  0.3720    0.6931
## weeks      1   0.83    0.83  0.1898    0.6668
## size       1 326.29  326.29 74.3075 5.866e-09 ***
## Residuals 25 109.78    4.39
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The term `week`'s p-value is $0.6668 > 0.05$, so we **do not reject** H_0 in favor of H_A and conclude that the term `week` does not significantly increase the proportion of the variance explained by the model and it is **not a significant addition** to the model.

Then we visualize our data with required symbols in the plot below:

```
##   (Intercept)  coloursgreen coloursorange          size
##   38.83371820    1.06306109   -1.24725442   -0.03549638
```



After refitting the model without term `week`, the regression lines for each colour level are listed below:

$$\text{Blue: rrate} = 38.8337 - 0.0355 \times \text{size}$$
$$\text{Orange: rrate} = 37.5864 - 0.0355 \times \text{size}$$
$$\text{Green: rrate} = 39.8968 - 0.0355 \times \text{size}$$

(c)

Although we detected some minor problems in the diagnostic plots from part (a) where non-nomality and heteroscedasticity occur, no simple transformations (e.g. log transformation, square-root transformation) would solve the problem completely. So we decide not to apply any redundant transformations.

Algebraically, our model can expressed as below:

$$\text{rrate}_{ij} = \beta_0 + \tau_j + \beta_1 \text{size}_{ij} + \epsilon_{ij}, \ \epsilon_{ij} \overset{i.i.d.}{\sim} N(0, \sigma^2)$$

Here j represents our 3 different levels of `colour` (blue, green, orange), and i corresponds to $1, 2, \ldots, 10$ observations within each of the three `colour` group (in fact, 5 from week A, and the other 5 from week B). Accordingly, τ_j is the jth level effect, and the error term ϵ_{ij} is normally distributed.

When it comes to discuss if our contrasts used in this model is a good choice, we will expand our investigation from two persepctives.

1. **Using treatment constraint $\tau_j = 0$ is totally fine in this case**, since we are studying the colours' influence on response rate, what we really want to compare is one colour versus another colour and try to find out if there is a difference between. In this case, switching to zero-sum constraint loses our focus on comparison between each colour, as it cares more about the difference between mean and each level.
2. **Using colour "blue" as reference group is fine as well.** The default treatment contrasts in R used colour group "blue" because it follows the alphabetical order, so the constraint applied is $\tau_{\text{blue}} = 0$ such that we are comparing group "blue" vs group "green" and group "blue" vs group "orange". But group "green" and group "orange" are not ever compared. Still, switching our baseline won't make a difference in terms of model predictions or in terms of measures such as R^2 etc.

To conclude the default contrasts used in this model is a decent choice.

```
##
## Call:
## lm(formula = rrate ~ colours + size)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.9608 -1.3761 -0.0202  0.8919  3.8152
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   38.833718   1.278859  30.366  < 2e-16 ***
## coloursgreen   1.063061   0.935453   1.136    0.266
## coloursorange -1.247254   0.923827  -1.350    0.189
## size          -0.035496   0.004053  -8.758 3.11e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.063 on 26 degrees of freedom
## Multiple R-squared:  0.7487, Adjusted R-squared:  0.7197
## F-statistic: 25.82 on 3 and 26 DF,  p-value: 5.838e-08
```

The summary table shows clearly that both `colourgreen`'s and `colourorange`'s p-values are greater than 0.05, hence they are not significant. In other words, we don't see significant differences between `colourblue` and the other two.

(d)

For the reduced model in part (b), the estimated response rate with `size`= 250 are displayed as the `fit` column in the matrix below. And the `lower` and `upper` columns are the corresponding upper and lower bounds of the required 95% confidence interval.

```
##           lower      fit    upper
## blue   28.61891 29.95962 31.30033
## green  29.12663 31.02268 32.91873
## orange 26.81632 28.71237 30.60842
```

(e)

The multiplicative model includes an interaction term which allows different slopes as well as different intercepts for three different colour groups.

$$\text{rrate}_{ij} = \beta_0 + \tau_j + \beta_1 \text{size}_{ij} + \gamma_j \text{size}_{ij} + \epsilon_{ij},\ \epsilon_{ij} \overset{i.i.d.}{\sim} N(0, \sigma^2)$$
$$\gamma_{\text{blue}} = 0$$

We actually produced the `summary()` and `anova()` table of this multiplicative model, together with the `anova()` table for both models.

```
##
## Call:
## lm(formula = rrate ~ colours + size + colours * size)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.9033 -1.1464 -0.1204  1.0202  3.9253
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        39.878528   1.933193  20.628  < 2e-16 ***
## coloursgreen       -0.573994   2.859837  -0.201    0.843
## coloursorange      -3.234987   3.200501  -1.011    0.322
## size               -0.039346   0.006680  -5.890 4.47e-06 ***
## coloursgreen:size   0.005761   0.009285   0.621    0.541
## coloursorange:size  0.007493   0.011621   0.645    0.525
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.122 on 24 degrees of freedom
## Multiple R-squared:  0.7545, Adjusted R-squared:  0.7034
## F-statistic: 14.75 on 5 and 24 DF,  p-value: 1.182e-06
```

```
## Analysis of Variance Table
##
## Response: rrate
##              Df Sum Sq Mean Sq F value    Pr(>F)
## colours       2   3.27    1.63  0.3628    0.6995
## size          1 326.29  326.29 72.4698 1.032e-08 ***
## colours:size  2   2.55    1.28  0.2834    0.7557
## Residuals    24 108.06    4.50
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## Analysis of Variance Table
##
## Model 1: rrate ~ colours + size
## Model 2: rrate ~ colours + size + colours * size
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     26 110.61
## 2     24 108.06  2     2.552 0.2834 0.7557
```

The F-test associated with the additional interaction term `colour:size` tests:

$$H_0: \ \frac{\sigma^2_{\text{addition}}}{\sigma^2_{\text{Error}}} = 1,\ H_A: \ \frac{\sigma^2_{\text{addition}}}{\sigma^2_{\text{Error}}} > 1$$

or equivalently,

$$H_0:\ \tau_{\text{blue}} = \tau_{\text{green}} = \tau_{\text{orange}} = 0,\ H_A:\ \text{not all } \tau_j = 0.$$

Since $p = 0.7557 > 0.05$, we do not reject H_0 in favor of H_A, and conclude that the interaction term `colours:size` is not a significant addition to the model. Hence, separate slopes for different colour groups are NOT required.

What's more, the p-value for `colours` is greater than 0.05 as well, so it is also not a significant term. Only the p-value of `size` is less than 0.05, leaving it as the only significant explanatory variable against response rates.

(f)

```
##
## Call:
## lm(formula = qcolour.A$rrate ~ colours.A + qcolour.A$size)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.54735 -0.17479 -0.01275  0.18398  0.52896
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     37.4912250  0.3033120 123.606  < 2e-16 ***
## colours.Agreen   1.3448159  0.2218649   6.061 8.17e-05 ***
## colours.Aorange -1.7756427  0.2191075  -8.104 5.78e-06 ***
## qcolour.A$size  -0.0298129  0.0009613 -31.013 4.64e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3459 on 11 degrees of freedom
## Multiple R-squared:  0.9894, Adjusted R-squared:  0.9865
## F-statistic: 341.8 on 3 and 11 DF,  p-value: 3.9e-11
```

```
## Analysis of Variance Table
##
## Response: qcolour.A$rrate
##                Df  Sum Sq Mean Sq F value    Pr(>F)
## colours.A       2   7.600    3.80  31.758 2.693e-05 ***
## qcolour.A$size  1 115.084  115.08 961.809 4.645e-12 ***
## Residuals      11   1.316    0.12
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

All of the coefficients of the reduced model with only *Week A* are significantly different from zero, since the p-values are less than 0.05. In other words, the colour of questionnaires does seeem to have some influence on the response rate if there is only one round of experiment.

One possible explanation to this is that *Week B* is chrononically after *Week A* so that the customers interviewed could have overlap, thus leading to a meaningless response.

The results suggest that our model should not contain an interaction term between `colour` and `size`. Instead, a factor `colour` and a continuous explanatory variable `size` could be included.

(g)

```
##              numDF denDF  F-value p-value
## (Intercept)      1    25 5862.587  <.0001
## colours          2    25    0.384  0.6851
## size             1    25   76.698  <.0001
```

```
## Linear mixed-effects model fit by REML
##  Data: NULL
##        AIC      BIC   logLik
##   142.8028 150.3514 -65.4014
##
## Random effects:
##  Formula: ~1 | week
##         (Intercept) Residual
## StdDev: 7.935918e-05  2.06258
##
## Fixed effects: rrate ~ colours + size
##                   Value Std.Error DF   t-value p-value
## (Intercept)    38.83372 1.2788593 25 30.365904  0.0000
## coloursgreen    1.06306 0.9354526 25  1.136414  0.2666
## coloursorange  -1.24725 0.9238265 25 -1.350096  0.1891
## size           -0.03550 0.0040532 25 -8.757718  0.0000
##  Correlation:
##               (Intr) clrsgr clrsrn
## coloursgreen  -0.212
## coloursorange -0.408  0.483
## size          -0.860 -0.166  0.055
##
## Standardized Within-Group Residuals:
##          Min           Q1          Med           Q3          Max
## -1.920334074 -0.667184839 -0.009788086  0.432400283  1.849720430
##
## Number of Observations: 30
## Number of Groups: 2
```

There has been almost NO real change from the model in part (c), the residual standard error is unchanged at 2.063.

```
## [1] 1.480378e-09
```

The intra-class correlation coefficient is calculated as follows:

$$\frac{\hat{\sigma}_\delta^2}{\hat{\sigma}_\delta^2 + \hat{\sigma}_\epsilon^2} = \frac{(7.935918 \times 10^{-5})^2}{(7.935918 \times 10^{-5})^2 + 2.06258^2} = 1.480378 \times 10^{-9} \approx 0$$

Therefore, the inclusion of `week` as a random effect does not provide extra explanation to variability, we should just exclude it from the model.

(h)

- **Fit of the various models**
  - For this assignment, we fitted 5 different models for our data, namely, the original model containing `colour`,`week` and `size`, the reduced model with only `colour` and `size`, the multiplicative model with interaction `colour:size`, the reduced model but only with observations from week A, and

the model with `week` as random effect. Sadly, only the reduced model with week A data seems to be a good fit, the others are not so appropriate.
    - If we investigate into the response rate `rrate` again, it won't be hard to find out that it does have a limited range from 0 to 100. We know that linear regression is good at continuous response with infinite numbers of possible values. Nonetheless, if we insist to use linear regression here, not only we still find it hard to fit a proper model, but also large value of depende variable `size` could cause our response `rrate` to be negative. (Recall the plot from part (b), all of our 3 regression lines share the same negative slope.)
    - Moreover, in part (a) we demonstrated the violation of assumptions in the first two diagnostic plots, and no simple transformations would fix these minor problems immediately. Using generalized linear model (for example, logistic regression), could solve this.
- **Experimental design**
    - One of major problems of this experimental design comes from the following sentence:
      "*The entire experiment was repeated in a different week, with the same colours assigned to the same car parks.*"
    - It indicates our design is not fully randomized. We all agree that randomization could reduce confounding by eualizing those factors that have not been accounted for. In our case, these factors are the supermarkets we selected. Ideally, we should eliminate the effect of locations this survey conducted in, while sending out the questionnaires of the same colour is not helping at all. Thus, a complete randomization is suggested.

Appendix

```r
library(nlme)

qcolour <- read.csv("qcolour.csv",header=T)
attach(qcolour)

colours <- as.factor(colour)
weeks <- as.factor(week)

lm.a <- lm(rrate~colours+weeks+size)
par(mfrow=c(2,2))
plot(lm.a,which = c(1,2,4,5))

anova(lm.a)

lm.b <- lm(rrate~colours+size)
par(mfrow=c(1,1))
plot(size, rrate, type="n")
title("Plot of rrate vs size")
points(qcolour[colour=='blue'&week=='A',]$size,
       qcolour[colour=='blue'&week=='A',]$rrate,pch='B',col='blue')
points(qcolour[colour=='blue'&week=='B',]$size,
       qcolour[colour=='blue'&week=='B',]$rrate,pch='b',col='blue')
points(qcolour[colour=='green'&week=='A',]$size,
       qcolour[colour=='green'&week=='A',]$rrate,pch='G',col='green')
points(qcolour[colour=='green'&week=='B',]$size,
       qcolour[colour=='green'&week=='B',]$rrate,pch='g',col='green')
points(qcolour[colour=='orange'&week=='A',]$size,
       qcolour[colour=='orange'&week=='A',]$rrate,pch='O',col='orange')
points(qcolour[colour=='orange'&week=='B',]$size,
```

```
       qcolour[colour=='orange'&week=='B',]$rrate,pch='o',col='orange')

coef(lm.b)
intercept <- coef(lm.b)[1]
coef.green <- coef(lm.b)[2]
coef.orange <- coef(lm.b)[3]
coef.size <- coef(lm.b)[4]

# regression lines
abline(intercept, coef.size, lty=2, col="blue", lwd=2) # blue
abline(intercept+coef.green, coef.size, lty=1, col="green", lwd=2) # green
abline(intercept+coef.orange, coef.size, lty=3, col="orange", lwd=2) # orange
legend("topright", c("blue","green","orange"),
       lty=c(2,1,3), col=c("blue","green","orange"), lwd=c(2,2,2))

summary.lm(lm.b)

lvl.mns <- tapply(rrate,colour,mean)
ni <- tapply(rrate,colour,length)
h.blue <- c(1,0,0)
h.green <- c(1,1,0)
h.orange <- c(1,0,1)

ci <- function(h) {
  h.extra <- h
  h.extra[length(h)+1] <- 250
  est <- t(h.extra)%*%coef(lm.b)
  MSE <- sum((rrate-fitted(lm.b))^2)/lm.b$df.residual
  sd <- sqrt(MSE)*sqrt(sum((h^2)/ni))
  upper <- est+qt(0.975,lm.b$df.residual)*sd
  lower <- est-qt(0.975,lm.b$df.residual)*sd
  c(lower,est,upper)
}
cis <- rbind(ci(h.blue),ci(h.green),ci(h.orange))
colnames(cis) <- c("lower","fit","upper")
rownames(cis) <- c("blue","green","orange")
cis

lm.e <- lm(rrate~colours+size+colours*size)
summary(lm.e)
anova(lm.e)
anova(lm.b,lm.e)

qcolour.A <- qcolour[qcolour$week=='A',]
colours.A <- as.factor(qcolour.A$colour)
lm.f <- lm(qcolour.A$rrate~colours.A+qcolour.A$size)
summary(lm.f)
anova(lm.f)

lm.g <- lme(rrate~colours+size, random=~1|week)
anova(lm.g)
summary(lm.g)
```

```r
# intra-class correlation
(icc <- (7.935918e-05)^2/((7.935918e-05)^2+2.06258^2))
```