Reproducibility Notebook for 'Performance evaluation of Google Docs'

Quang - Vinh DANG

June 19, 2015

Contents

1 Project Overview

1.1 Purpose of the experiment

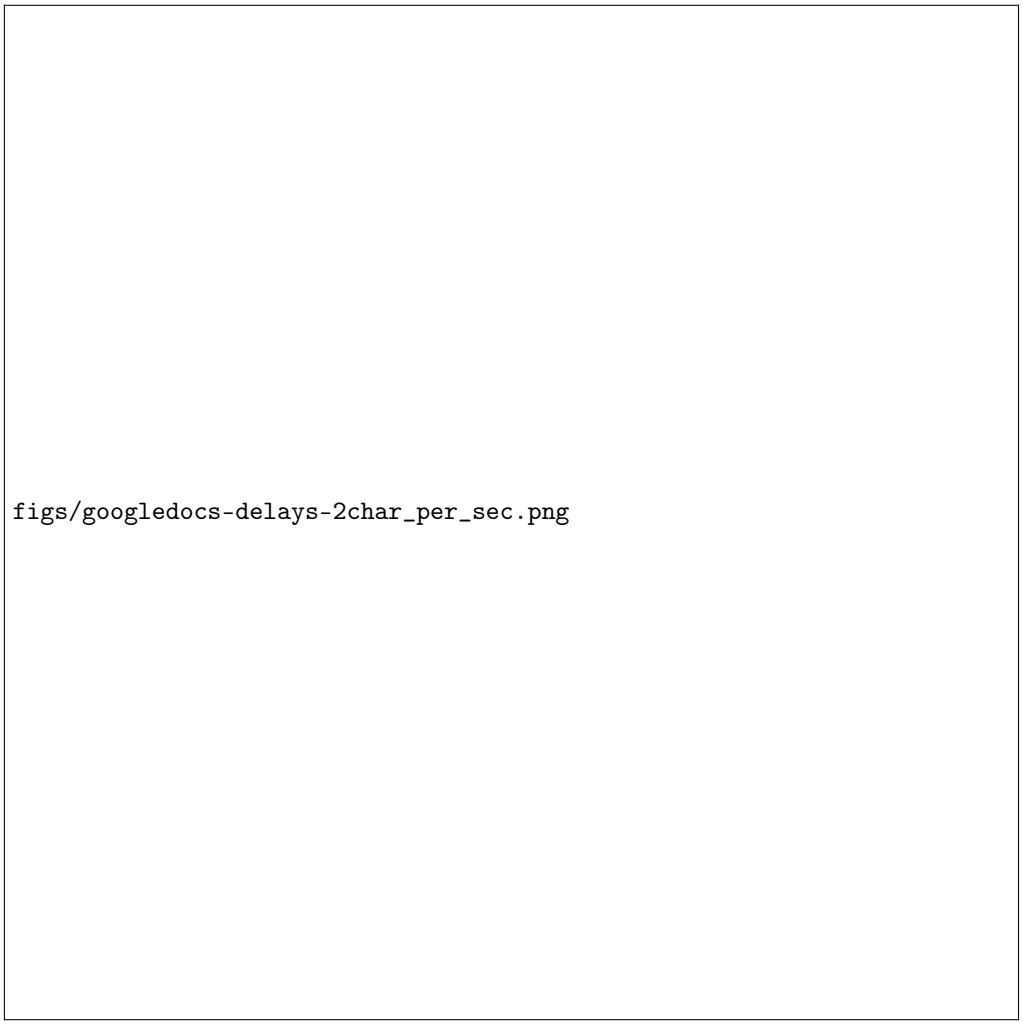
Perform the evaluation of Google Docs's performance in collaborative editing large scale settings.

2 Data Analysis

2.1 Delay Measurements in GoogleDocs

```
TYPING_SPEED=2 # 2 chars/sec

df <- read.table('googledocs-delays.txt', header=TRUE)
df$delay <- df$delay / 1000   # convert delay in seconds
df <- df[df$speed == TYPING_SPEED,] # filter observation for a specific typing speed
df$speed <- NULL   # suppress speed column
df <- df[df$delay<50,]   # remove (22 51641) outlier

# add missing row
for (newrow in c(3, 7, 9, 31, 33, 35, 37)) {
    df <- rbind(df, c(newrow, NA))
}

tgc <- summarySE(df, measurevar="delay", groupvars=c("user"))
is.nan.data.frame <- function(x) do.call(cbind, lapply(x, is.na))
tgc[is.nan(tgc)] <- 0
```

```
plot <- ggplot(df,aes(factor(user), delay)) +
    coord_map(ylim = c(0,18)) + # cropping y-axis
    geom_point(color="royalblue4", alpha=.4,shape=16,size=2) +
    stat_smooth(color="black", data=df, linetype="dashed", aes(group=1,x=factor(user), y=delay), method='lm', formula=y~x, s
    scale_x_discrete(breaks=c(0,5,10,15,20,25,30,35,40)) +
    labs(x = "Number of Users", y = "Delay (sec)") +
    theme_bw() +
    theme(plot.margin = unit(c(0, 0, 0, 0), "cm"))
print(plot)
```



2.2 Performance Measures

Data is presented in the following table whose columns are :

- Group: name of the user group
- Condition: delay condition is seconds
- WC15mn: words count in the document at 15 minutes
- ErrorNum15mn: errors count in the document at 15 minutes
- Keywords: number of keywords being in the document at 15 minutes.
- Resolution: Binary metric indicating if all redundancies have been solved at 15mn

- SumRedundancy: number of word redundancies in the document at 12mn
- Editor4Notes: rating on how much the editor/tool help the group of users (from questionnaire)
- RedAwareness: rating on how users were aware of redundancy (from questionnaire)
- NewCEExp: rating on previous experience in collaborative editing (from questionnaire)

Group	Condition	WC15mn	ErrorNum15mn	Keywords	Resolution	SumRedundancy	Editor4Notes	RedAwareness	NewCEExp
G4	4	605	58	67	0	14	7.5	2	0.5
G5	8	536	64	55	1	15	5.75	1	0.75
G6	0	422	15	57	0	6	7.25	1	0.75
G7	6	571	47	53	1	12	7	0	1
G8	6	540	36	59	0	11	6.5	1	0.75
G9	10	565	51	67	1	14	4.75	2	1
G10	10	499	49	47	1	12	3	1	0.75
G12	6	521	45	57	1	9	5.75	2	1
G13	10	571	59	68	1	11	4.5	2	0.75
G15	4	391	17	58	0	8	6.33	1	1
G16	8	393	32	45	0	9	5.75	3	1
G17	0	352	13	55	0	6	7.67	1	1
G18	4	530	46	61	1	10	6.75	2	1
G19	0	355	15	51	0	5	3.75	1	1
G20	8	731	95	63	1	11	5.5	1	0.75
G21	6	404	37	55	0	10	4.5	0	0.75
G25	10	465	28	60	0	10	6.5	1	0.5

Chat data is presented in the following table whose columns are :

- Group: name of the user group
- Condition: delay condition is seconds
- TotalChatWords: words count in the chat log
- Bdef:
- Ddef:
- Adef:
- Baccord:
- Daccord:
- Aaccord:

Group	Condition	TotalChatWords	Bdef	Ddef	Adef	Baccord	Daccord	Aaccord
G4	4	43	2	0	2	0	0	1
G5	8	52	0	0	0	0	1	2
G6	0	110	3	1	1	5	1	3
G7	6	42	1	0	4	0	0	1
G8	6	85	0	1	8	0	0	3
G9	10	187	0	0	7	2	2	1
G10	10	110	0	2	5	0	0	3
G12	6	73	0	2	1	0	1	0
G13	10	118	1	6	0	2	3	0
G15	4	40	0	2	1	0	1	0
G16	8	213	0	10	6	2	8	3
G17	0	77	0	2	4	0	3	1
G18	4	38	0	3	0	0	1	1
G19	0	128	2	7	0	0	1	2
G20	8	99	0	4	6	0	0	2
G21	6	79	0	0	6	0	0	3
G25	10	52	0	0	6	0	0	5

2.2.1 Document Content

Text base is larger for the high delay groups at 15 minutes

```
lm1 <- lm(data=mydata, WC15mn~Condition)
summary(lm1)

Call:
lm(formula = WC15mn ~ Condition, data = mydata)

Residuals:
     Min       1Q   Median       3Q      Max
-135.361  -58.868    7.639   41.147  202.639

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  410.332     43.945   9.337 1.22e-07 ***
Condition     14.754      6.471   2.280   0.0377 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 90.54 on 15 degrees of freedom
Multiple R-squared:  0.2574,Adjusted R-squared:  0.2078
F-statistic: 5.198 on 1 and 15 DF,  p-value: 0.03765

paste("Beta=", lm.beta(lm1))

Beta= 0.507305546770752
```

Proportion of keywords is negatively related to delay condition / Quality content decreases with delay condition First we compute the proportion of keywords that are in the document at 15 min and the arcsin transformation of these values.

```
mydata[,"PropKeywords"] <- with(mydata, Keywords / WC15mn)
mydata[,"TransPropKeywords"] <- with(mydata, asin(sqrt(PropKeywords)))

lm2 <-lm(data=mydata, TransPropKeywords~Condition)
summary(lm2)

Call:
lm(formula = TransPropKeywords ~ Condition, data = mydata)

Residuals:
      Min        1Q    Median        3Q       Max
-0.041992 -0.014604 -0.005166  0.022775  0.037859

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.381507   0.012581  30.324  7.1e-15 ***
Condition   -0.005194   0.001853  -2.804   0.0134 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.02592 on 15 degrees of freedom
Multiple R-squared:  0.3439,Adjusted R-squared:  0.3001
F-statistic: 7.861 on 1 and 15 DF,  p-value: 0.01336

paste("Beta=", lm.beta(lm2))

Beta= -0.586408363704111
```

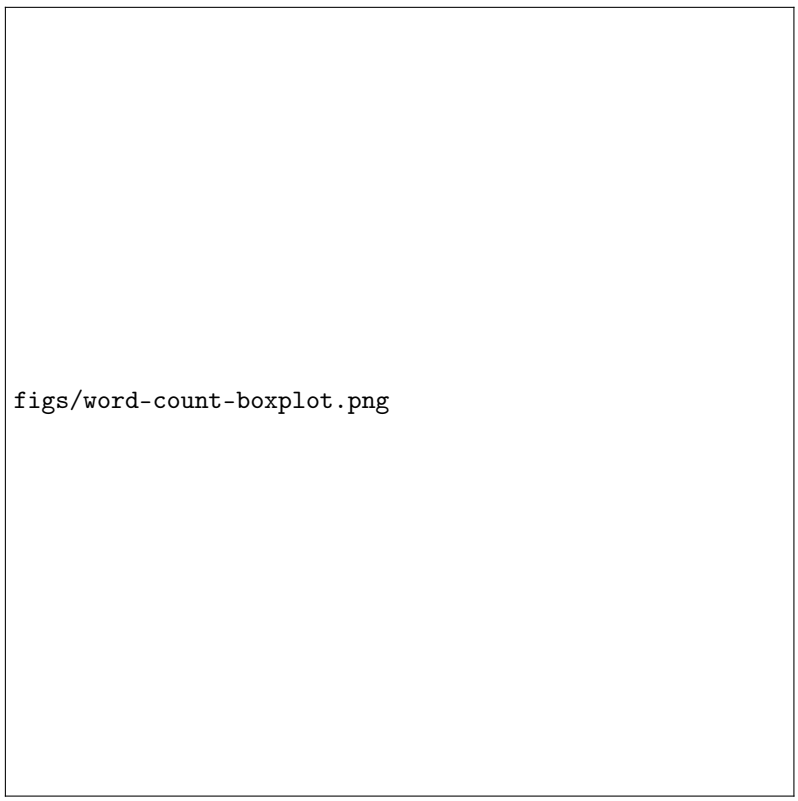
Document redundancy at 12 minutes is a function of delay condition.

```
lmr <- lm(data=mydata, SumRedundancy~Condition)
summary(lmr)
```

```
Call:
lm(formula = SumRedundancy ~ Condition, data = mydata)

Residuals:
    Min      1Q  Median      3Q     Max
-2.5397 -1.2440 -0.5397  0.9038  4.9038

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   6.8005     1.0180   6.680 7.35e-06 ***
Condition     0.5739     0.1499   3.828  0.00164 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.097 on 15 degrees of freedom
Multiple R-squared:  0.4942,Adjusted R-squared:  0.4605
F-statistic: 14.66 on 1 and 15 DF,  p-value: 0.001645

paste("Beta=", lm.beta(lmr))

Beta= 0.703009085129074
```

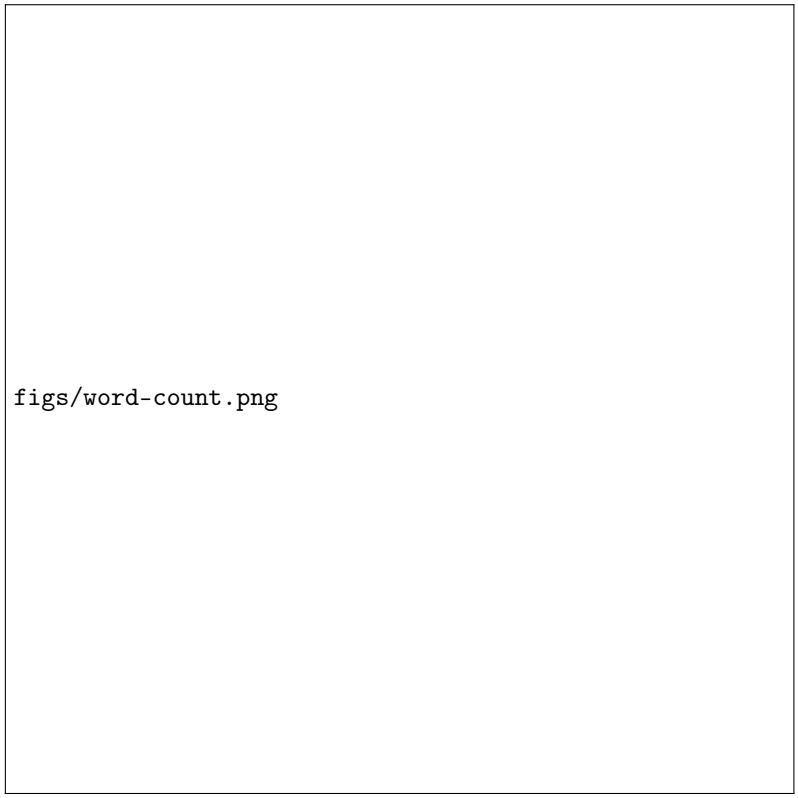
Words count as a function Delay condition

```
wc_15mn <- ggplot(data=mydata, aes(x=factor(Condition), y=WC15mn)) +
     geom_point() +
     geom_boxplot(color="black", outlier.shape=1, outlier.color="grey70") +
     labs(x="Delay Condition (sec)", y="Number of Words") +
     theme_bw()
     #ggtitle("Number of words as a function of delay condition")

print(wc_15mn)
```



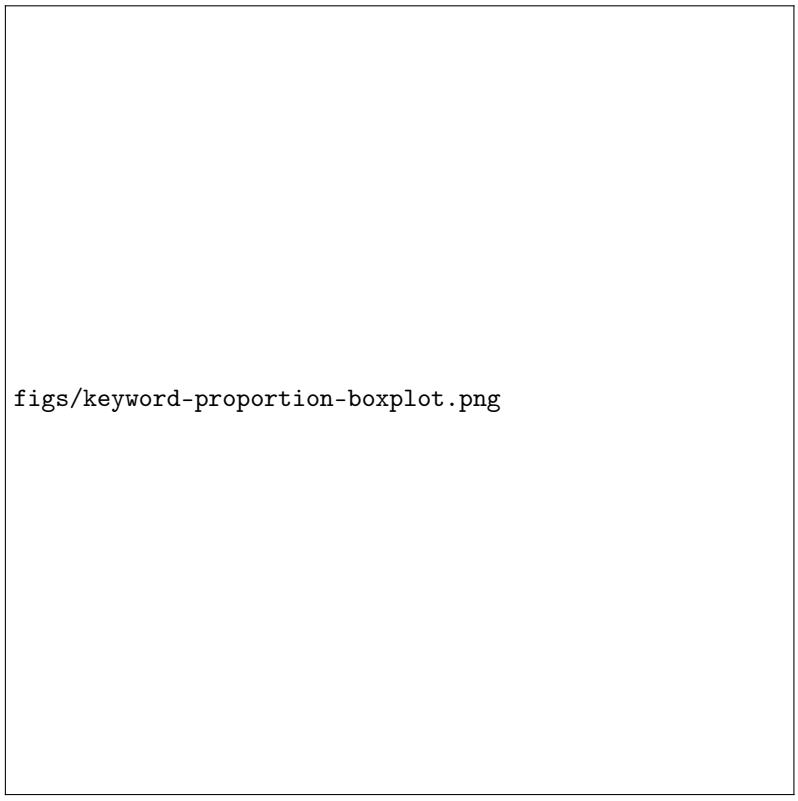

```
tg <- mydata[c("Condition", "WC15mn")]
tgc <- summarySE(tg, measurevar="WC15mn", groupvars=c("Condition"))
```

```
wc_15mn <- ggplot(data=tg, aes(x=Condition, y=WC15mn)) +
     geom_point(color="blue",shape=18,size=3) +
     geom_errorbar(size=.3,width=.5, data=tgc, aes(ymin=WC15mn-sd, ymax=WC15mn+sd)) +
     stat_smooth(linetype="dashed", color="grey40", data=mydata, aes(x=Condition, y=WC15mn), method='lm', formula=y~x, se=F
     labs(x="Delay Condition (sec)", y="Number of Words") +
     scale_x_discrete(limits=c(0,2,4,6,8,10)) +
     expand_limits(x=c(-1,11)) +
     ggtitle("Word Count")

print(wc_15mn)
```

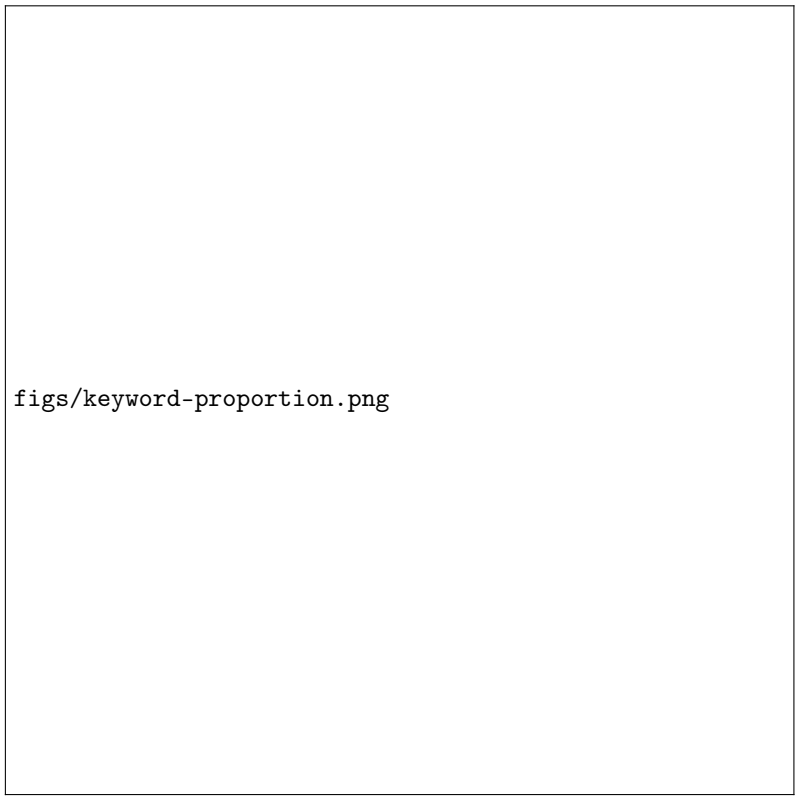


Proportion of keywords as a function of Delay condition

```
kw_proportion <- ggplot(data=mydata, aes(x=factor(Condition), y=PropKeywords)) +
     geom_point() +
     geom_boxplot(color="black", outlier.shape=1, outlier.color="grey70") +
     labs(x="Delay Condition (sec)", y="Keyword Proportion") +
     theme_bw()
     #ggtitle("Proportion of keywords as a function of delay condition")
print(kw_proportion)
```




```
tg <- mydata[c("Condition", "PropKeywords")]
tgc <- summarySE(tg, measurevar="PropKeywords", groupvars=c("Condition"))

wc_15mn <- ggplot(data=tg, aes(x=Condition, y=PropKeywords)) +
     geom_point(color="blue",shape=18,size=3) +
     geom_errorbar(size=.3,width=.5, data=tgc, aes(ymin=PropKeywords-sd, ymax=PropKeywords+sd)) +
     stat_smooth(linetype="dashed", color="grey40", data=mydata, aes(x=Condition, y=PropKeywords), method='lm', formula=y~x
     labs(x="Delay Condition (sec)", y="Keyword Proportion") +
     scale_x_discrete(limits=c(0,2,4,6,8,10)) +
     expand_limits(x=c(-1,11)) +
     ggtitle("Keyword Proportion")

print(wc_15mn)
```



2.2.2 Error Proportions at 15 Minutes

Error rate is a function of delay First we compute the ratio between the number of errors and the number of words in the document at 15 min. Then, we compute the arcsin transformation of this metric.

```
mydata[,"Ratio15mn"] <- with(mydata, ErrorNum15mn / WC15mn)
mydata[,"Trans15mn"] <- with(mydata, asin(sqrt(Ratio15mn)))

lm3 <-lm(data=mydata, Trans15mn~Condition)
summary(lm3)

Call:
lm(formula = Trans15mn ~ Condition, data = mydata)

Residuals:
      Min        1Q    Median        3Q       Max
-0.079488 -0.022029 -0.008376  0.024842  0.063794

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.215408   0.019049  11.308 9.69e-09 ***
Condition   0.011200   0.002805   3.993  0.00118 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.03925 on 15 degrees of freedom
Multiple R-squared:  0.5152,Adjusted R-squared:  0.4829
F-statistic: 15.94 on 1 and 15 DF,  p-value: 0.001176

paste("Beta=", lm.beta(lm3))

Beta= 0.71779806011064
```

Error proportion metric is negatively correlated with the proportion of keywords

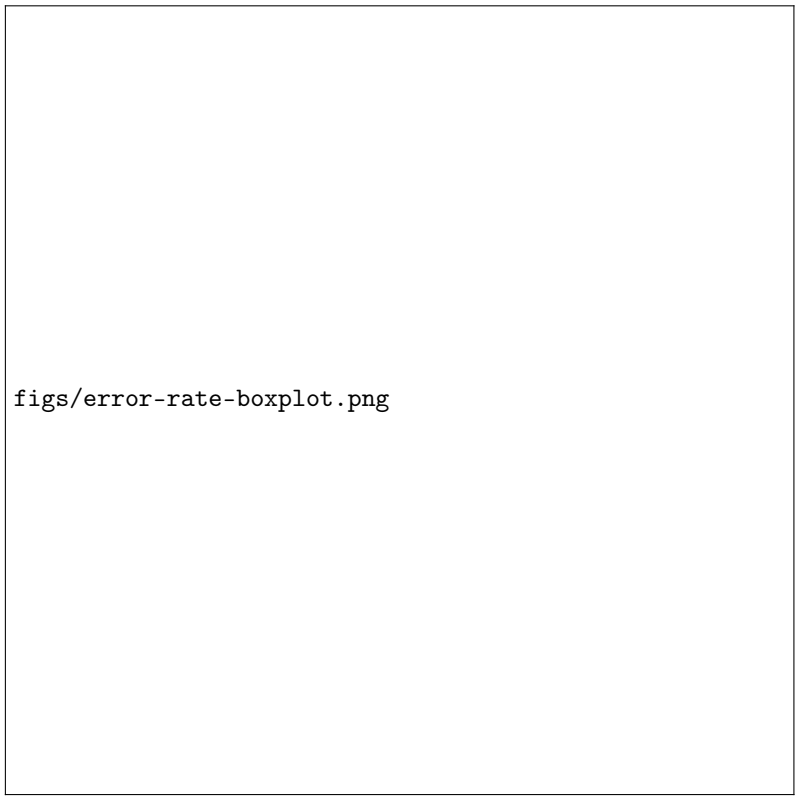
```
lm4t <- lm(data=mydata, Trans15mn~TransPropKeywords)
summary(lm4t)

Call:
lm(formula = Trans15mn ~ TransPropKeywords, data = mydata)

Residuals:
      Min        1Q    Median        3Q       Max
-0.055769 -0.017652 -0.002535  0.016988  0.064213

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)        0.77686    0.09577   8.112 7.25e-07 ***
TransPropKeywords -1.41209    0.27188  -5.194 0.000109 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.0337 on 15 degrees of freedom
Multiple R-squared:  0.6426,Adjusted R-squared:  0.6188
F-statistic: 26.98 on 1 and 15 DF,  p-value: 0.000109

paste("Beta=", lm.beta(lm4t))

Beta= -0.801653184062439
```

Redundancy and error rate are correlated

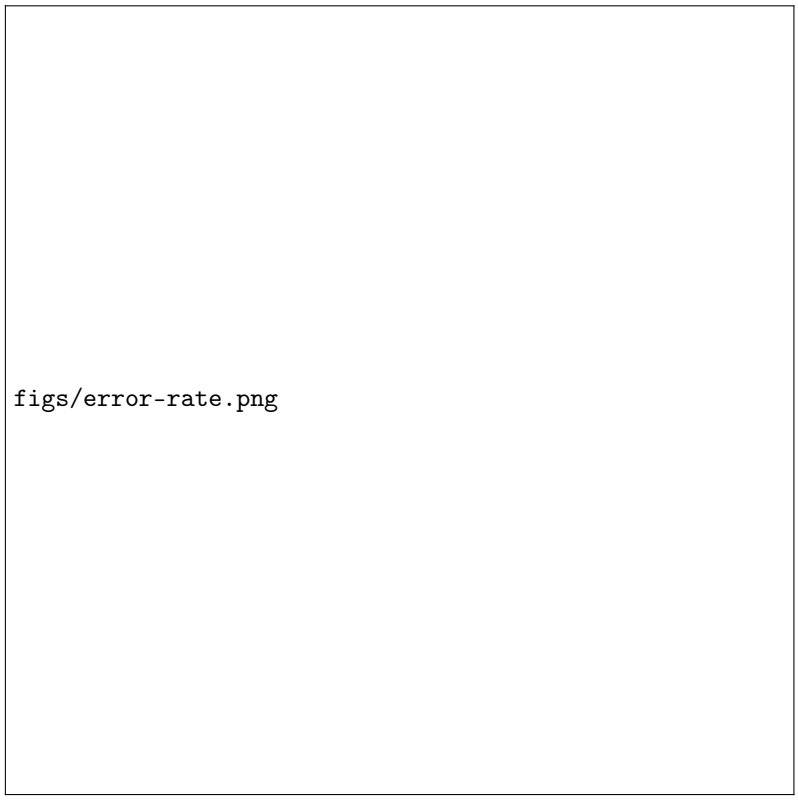
```
lmrr <- lm(data=mydata, SumRedundancy~Trans15mn)
summary(lmrr)

[1] "Beta= -0.801653184062439"

Call:
lm(formula = SumRedundancy ~ Trans15mn, data = mydata)

Residuals:
    Min      1Q  Median      3Q     Max
-2.8517 -1.2751 -0.3287  1.4162  2.8212

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   -1.637      2.306  -0.710 0.488766
Trans15mn     41.997      8.057   5.212 0.000105 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.759 on 15 degrees of freedom
Multiple R-squared:  0.6443,Adjusted R-squared:  0.6206
F-statistic: 27.17 on 1 and 15 DF,  p-value: 0.0001052

paste("Beta=", lm.beta(lmrr))

Beta= 0.802671237202611
```

Error Rate as a function of Delay condition

```
kw_proportion <- ggplot(data=mydata, aes(x=factor(Condition), y=Trans15mn)) +
     geom_point() +
     geom_boxplot(color="black", outlier.shape=1, outlier.color="grey70") +
     labs(x="Delay Condition (sec)", y="Error Rate") +
     theme_bw()
     #ggtitle("Error Rate as a function of delay condition")
print(kw_proportion)
```



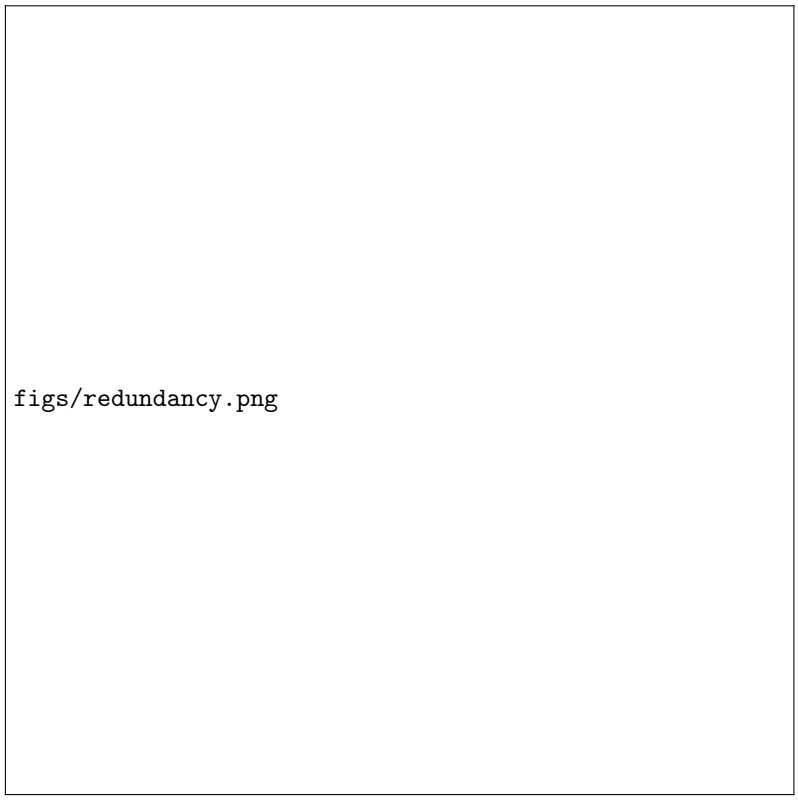

```
tg <- mydata[c("Condition", "Trans15mn")]
tgc <- summarySE(tg, measurevar="Trans15mn", groupvars=c("Condition"))

wc_15mn <- ggplot(data=tg, aes(x=Condition, y=Trans15mn)) +
     geom_point(color="blue",shape=18,size=3) +
     geom_errorbar(size=.3,width=.5, data=tgc, aes(ymin=Trans15mn-sd, ymax=Trans15mn+sd)) +
     stat_smooth(linetype="dashed", color="grey40", data=mydata, aes(x=Condition, y=Trans15mn), method='lm', formula=y~x, se
     labs(x="Delay Condition (sec)", y="Error Rate") +
     scale_x_discrete(limits=c(0,2,4,6,8,10)) +
     expand_limits(x=c(-1,11)) +
     ggtitle("Error Rate")
print(wc_15mn)
```



Redundancy as a function of Delay condition

```
kw_proportion <- ggplot(data=mydata, aes(x=factor(Condition), y=SumRedundancy)) +
     geom_point() +
     geom_boxplot(color="black", outlier.shape=1, outlier.color="grey70") +
     labs(x="Delay Condition (sec)", y="Redundancies") +
     theme_bw()
     #ggtitle("Redundancy as a function of delay condition")
print(kw_proportion)
```


```
tg <- mydata[c("Condition", "SumRedundancy")]
tgc <- summarySE(tg, measurevar="SumRedundancy", groupvars=c("Condition"))
tgc

wc_15mn <- ggplot(data=tg, aes(x=Condition, y=SumRedundancy)) +
     geom_point(color="blue",shape=18,size=3) +
     geom_errorbar(size=.3,width=.5, data=tgc, aes(ymin=SumRedundancy-sd, ymax=SumRedundancy+sd)) +
     stat_smooth(linetype="dashed", color="grey40", data=mydata, aes(x=Condition, y=SumRedundancy), method='lm', formula=y~x
     labs(x="Delay Condition (sec)", y="Redundancies") +
     scale_x_discrete(limits=c(0,2,4,6,8,10)) +
     expand_limits(x=c(-1,11)) +
     ggtitle("Redundancy")
print(wc_15mn)
```



2.2.3 Subjective Difficulty Ratings

Editor difficulty ratings are not related to delay condition

```
lmd <- lm(data=mydata, Editor4Notes~Condition)
summary(lmd)

[1] "Beta= 0.802671237202611"
 null device
           1
 null device
           1
 null device
           1

 Condition N SumRedundancy        sd        se       ci
1         0 3      5.666667 0.5773503 0.3333333 1.434218
2         4 3     10.666667 3.0550505 1.7638342 7.589166
3         6 4     10.500000 1.2909944 0.6454972 2.054260
4         8 3     11.666667 3.0550505 1.7638342 7.589166
5        10 4     11.750000 1.7078251 0.8539126 2.717531
null device
           1

Call:
lm(formula = Editor4Notes ~ Condition, data = mydata)

Residuals:
    Min      1Q  Median      3Q     Max
-3.0373 -0.3739  0.2934  0.7107  1.3781

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.78726    0.60489  11.221 1.08e-08 ***
Condition   -0.16633    0.08907  -1.867   0.0815 .
```

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

Residual standard error: 1.246 on 15 degrees of freedom
Multiple R-squared:  0.1886,Adjusted R-squared:  0.1345
F-statistic: 3.487 on 1 and 15 DF,  p-value: 0.08151
```

Editor difficulty ratings do not correlate with any of the performance measures

Editor difficulty ratings do not correlate with Transformed error rate

```
lmd2 <- lm(data=mydata, Editor4Notes~Trans15mn)
summary(lmd2)

Call:
lm(formula = Editor4Notes ~ Trans15mn, data = mydata)

Residuals:
    Min      1Q  Median      3Q     Max
-2.6657 -0.8638  0.4063  0.6924  1.9650

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    8.108      1.710   4.740 0.000263 ***
Trans15mn     -8.172      5.976  -1.368 0.191600
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.305 on 15 degrees of freedom
Multiple R-squared:  0.1109,Adjusted R-squared:  0.05158
F-statistic:  1.87 on 1 and 15 DF,  p-value: 0.1916
```

Editor difficulty ratings do not correlate with Redundancy at 12 minutes

```
lmd3 <- lm(data=mydata, Editor4Notes~SumRedundancy)
summary(lmd3)

Call:
lm(formula = Editor4Notes ~ SumRedundancy, data = mydata)

Residuals:
    Min      1Q  Median      3Q     Max
-2.7282 -0.8898  0.1544  0.9334  1.8602

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)     6.2586     1.2718   4.921 0.000185 ***
SumRedundancy  -0.0442     0.1206  -0.367 0.719101
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.377 on 15 degrees of freedom
Multiple R-squared:  0.008876,Adjusted R-squared:  -0.0572
F-statistic: 0.1343 on 1 and 15 DF,  p-value: 0.7191
```

Editor difficulty ratings do not correlate with Proportion of keywords

```
lmd4t <- lm(data=mydata, Editor4Notes~TransPropKeywords)
summary(lmd4t)

Call:
lm(formula = Editor4Notes ~ TransPropKeywords, data = mydata)

Residuals:
    Min      1Q  Median      3Q     Max
```

```
-2.5447 -1.0629  0.1095  0.9738  1.7704

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)          3.433      3.884   0.884    0.391
TransPropKeywords    6.769     11.026   0.614    0.548

Residual standard error: 1.367 on 15 degrees of freedom
Multiple R-squared:  0.02451,Adjusted R-squared:  -0.04052
F-statistic: 0.377 on 1 and 15 DF,  p-value: 0.5484
```

Editor difficulty ratings do not correlate with Word count

```
lmd5 <- lm(data=mydata, Editor4Notes~WC15mn)
summary(lmd5)

Call:
lm(formula = Editor4Notes ~ WC15mn, data = mydata)

Residuals:
    Min      1Q  Median      3Q     Max
-2.8093 -1.0777 -0.0299  0.9320  1.9015

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 5.6707063  1.7229059   3.291  0.00495 **
WC15mn      0.0002778  0.0033994   0.082  0.93594
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.383 on 15 degrees of freedom
Multiple R-squared:  0.0004451,Adjusted R-squared:  -0.06619
F-statistic: 0.00668 on 1 and 15 DF,  p-value: 0.9359
```

2.3 Mediation Analyses

```
compute_r_rsquare_beta <- function(data, formula) {
  lmf <- lm(data=data, formula)
  rsquare <- summary(lmf)$adj.r.squared
  paste("R=", sqrt(rsquare), "   ", "adj-R^2=", rsquare, "   ", "Beta=", lm.beta(lmf), sep="")
}

compute_r_rsquare_beta(data=mydata, Trans15mn~Condition)

[1] "R=0.694921812464182   adj-R^2=0.482916325438504   Beta=0.71779806011064"

compute_r_rsquare_beta(data=mydata, SumRedundancy~Condition)

[1] "R=0.678603879563734   adj-R^2=0.460503225358951   Beta=0.703009085129074"

compute_r_rsquare_beta(data=mydata, Trans15mn~SumRedundancy)

[1] "R=0.787760447535414   adj-R^2=0.620566522701195   Beta=0.802671237202612"

compute_r_rsquare_beta(data=mydata, Trans15mn~Condition+SumRedundancy)

[1] "R=0.8041855278394   adj-R^2=0.646714363186335   Beta=0.303518182526151"
[2] "R=0.8041855278394   adj-R^2=0.646714363186335   Beta=0.589295197384863"
```

2.4 Redundancy Management Analyses

2.4.1 Redundancy Awareness

```
lr <- lm(data=mydata, RedAwareness~SumRedundancy)
summary(lr)
```

```
Call:
lm(formula = RedAwareness ~ SumRedundancy, data = mydata)

Residuals:
    Min      1Q  Median      3Q     Max
-1.3377 -0.3138 -0.1943  0.6145  1.7340

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    1.05095    0.73307   1.434    0.172
SumRedundancy  0.02390    0.06951   0.344    0.736

Residual standard error: 0.7939 on 15 degrees of freedom
Multiple R-squared:  0.007818,Adjusted R-squared:  -0.05833
F-statistic: 0.1182 on 1 and 15 DF,  p-value: 0.7358

lr <- lm(data=mydata, RedAwareness~Resolution)
summary(lr)

Call:
lm(formula = RedAwareness ~ Resolution, data = mydata)

Residuals:
    Min      1Q  Median      3Q     Max
-1.3750 -0.3750 -0.2222  0.6250  1.7778

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.2222     0.2643   4.624 0.000331 ***
Resolution    0.1528     0.3853   0.397 0.697296
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7929 on 15 degrees of freedom
Multiple R-squared:  0.01037,Adjusted R-squared:  -0.0556
F-statistic: 0.1572 on 1 and 15 DF,  p-value: 0.6973
```

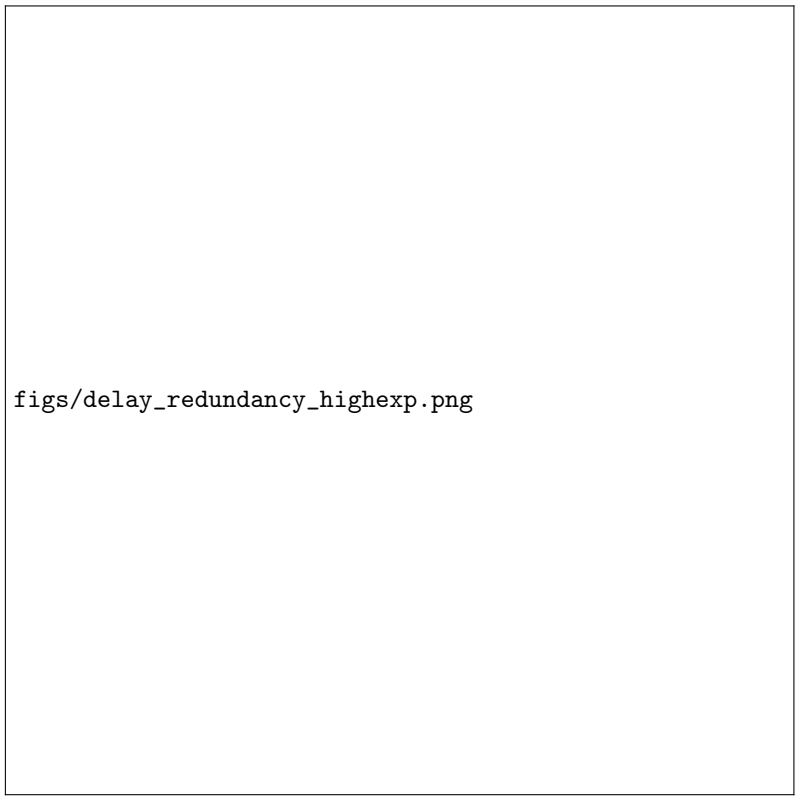
2.4.2 Experience

```
lme <- lm(data=mydata, SumRedundancy~Condition+NewCEExp+Condition*NewCEExp)
summary(lme)

Call:
lm(formula = SumRedundancy ~ Condition + NewCEExp + Condition *
    NewCEExp, data = mydata)

Residuals:
    Min      1Q  Median      3Q     Max
-2.8129 -1.0170 -0.1886  0.8326  3.4865

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)         19.7494     5.1652   3.824  0.00211 **
Condition           -1.0745     0.7145  -1.504  0.15654
NewCEExp           -14.5820     5.7229  -2.548  0.02428 *
Condition:NewCEExp   1.8828     0.8231   2.287  0.03958 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.839 on 13 degrees of freedom
Multiple R-squared:  0.6629,Adjusted R-squared:  0.5851
F-statistic:  8.52 on 3 and 13 DF,  p-value: 0.002179

df <- mydata[c("Condition", "NewCEExp", "SumRedundancy")]
df_highexp <- df[df["NewCEExp"] == 1,]
df_lowexp <- df[df["NewCEExp"] <= .75,]
```

Delay predicts Redundancy for High-experienced Group

```
lexp_split1 <- lm(data=mydata, SumRedundancy~Condition, NewCEExp==1)
summary(lexp_split1)

Call:
lm(formula = SumRedundancy ~ Condition, data = mydata, subset = NewCEExp ==
    1)

Residuals:
    Min      1Q  Median      3Q     Max
-2.4743 -0.7757 -0.1371  1.1643  1.9714

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   5.6914     0.9834   5.788  0.00116 **
Condition     0.7229     0.1699   4.255  0.00535 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.589 on 6 degrees of freedom
Multiple R-squared:  0.7511,Adjusted R-squared:  0.7096
F-statistic:  18.1 on 1 and 6 DF,  p-value: 0.005353

paste("Beta=", lm.beta(lexp_split1))

Beta= 0.866636579872041

delay_highexp <- ggplot() +
    geom_boxplot(color="tomato4", fill="mistyrose", data=df_highexp, aes(x=factor(Condition), y=SumRedundancy), outlier.shap
    geom_point(data=df_highexp, aes(x=factor(Condition), y=SumRedundancy), color="tomato4") +
    stat_smooth(color="red", data=df_highexp, aes(group=1,x=factor(Condition), y=SumRedundancy), method='lm', formula=y~x, s
    labs(x="Delay Condition (sec)", y="Redundancies") +
    theme_bw()
    #ggtitle("Delay condition predicts redundancy for high-experienced group")
print(delay_highexp)
```

```
tg <- df_highexp[c("Condition", "SumRedundancy")]
tgc <- summarySE(tg, measurevar="SumRedundancy", groupvars=c("Condition"))

is.nan.data.frame <- function(x) do.call(cbind, lapply(x, is.na))
tgc[is.nan(tgc)] <- 0

wc_15mn <- ggplot(data=tg, aes(x=Condition, y=SumRedundancy)) +
     geom_point(color="blue",shape=18,size=3) +
     geom_errorbar(size=.3,width=.5, data=tgc, aes(ymin=SumRedundancy-sd, ymax=SumRedundancy+sd)) +
     stat_smooth(linetype="dashed", color="grey40", data=df_highexp, aes(x=Condition, y=SumRedundancy), method='lm', formula
     labs(x="Delay Condition (sec)", y="Redundancies") +
     scale_x_discrete(limits=c(0,2,4,6,8,10)) +
     expand_limits(x=c(-1,11)) +
     ggtitle("Redundancy for High-Experienced Groups")
print(wc_15mn)
```



Delay does not predict Redundancy for Low-experienced Group

```
lexp_split2 <-lm(data=mydata, SumRedundancy~Condition, NewCEExp<=.75)
summary(lexp_split2)

null device
          1
Warning message:
In qt(conf.interval/2 + 0.5, datac$N - 1) : NaNs produced

Call:
lm(formula = SumRedundancy ~ Condition, data = mydata, subset = NewCEExp <=
    0.75)

Residuals:
   Min     1Q Median     3Q    Max
  -2.7   -1.2   -0.5    0.2    3.9

Coefficients:
```

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   8.7000     1.9672   4.422  0.00307 **
Condition     0.3500     0.2598   1.347  0.21991
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.449 on 7 degrees of freedom
Multiple R-squared:  0.2059,Adjusted R-squared:  0.09244
F-statistic: 1.815 on 1 and 7 DF,  p-value: 0.2199
```

```
delay_highexp <- ggplot() +
    geom_boxplot(color="royalblue4", Fill="lightskyblue1", data=df_lowexp, aes(x=factor(Condition), y=SumRedundancy)) +
    geom_point(data=df_lowexp, aes(x=factor(Condition), y=SumRedundancy), color="royalblue4") +
    stat_smooth(data=df_lowexp, aes(group=1,x=factor(Condition), y=SumRedundancy), method='lm', formula=y~x, se=FALSE, fullr
    labs(x="Delay Condition (sec)", y="Redundancies") +
    theme_bw()
    #ggtitle("Delay condition does not predict redundancy for low-experienced group")
print(delay_highexp)
```


```
tg <- df_lowexp[c("Condition", "SumRedundancy")]
tgc <- summarySE(tg, measurevar="SumRedundancy", groupvars=c("Condition"))

is.nan.data.frame <- function(x) do.call(cbind, lapply(x, is.na))
tgc[is.nan(tgc)] <- 0

wc_15mn <- ggplot(data=tg, aes(x=Condition, y=SumRedundancy)) +
     geom_point(color="blue",shape=18,size=3) +
     geom_errorbar(size=.3,width=.5, data=tgc, aes(ymin=SumRedundancy-sd, ymax=SumRedundancy+sd)) +
     labs(x="Delay Condition (sec)", y="Redundancies") +
     scale_x_discrete(limits=c(0,2,4,6,8,10)) +
     expand_limits(x=c(-1,11)) +
     ggtitle("Redundancy for Low-Experienced Groups")
print(wc_15mn)
```

2.4.3 Chat Behavior

```
chat[,"TotalDef"] <- with(chat, Bdef + Ddef + Adef)
chat[,"TotalAccord"] <- with(chat, Baccord + Daccord + Aaccord)
```

Definite determiners and agreement are highly correlated

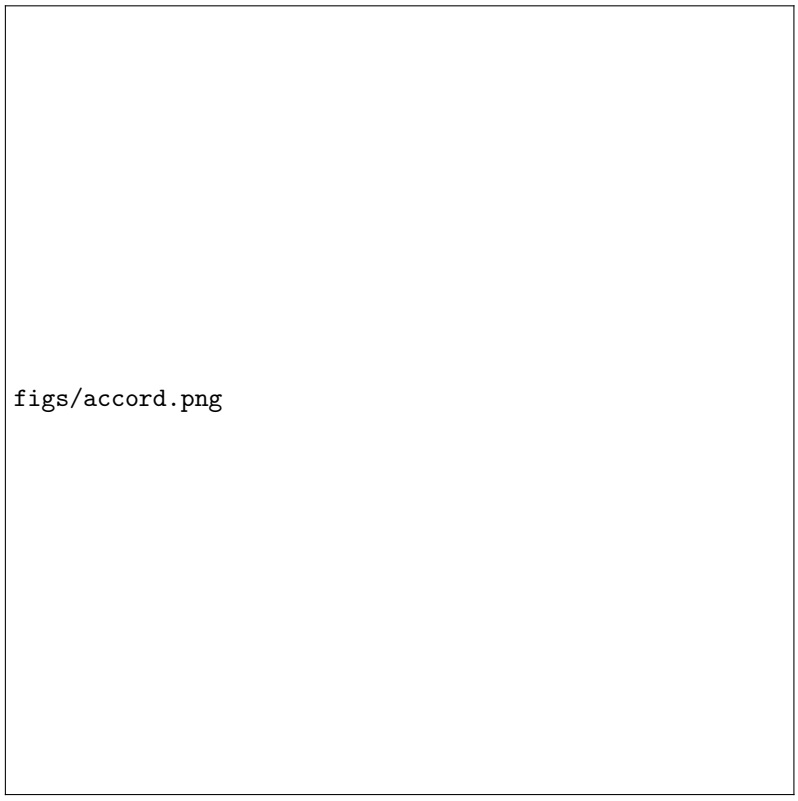
```
lmdd <- lm(data=chat, TotalDef~TotalAccord)
summary(lmdd)

Call:
lm(formula = TotalDef ~ TotalAccord, data = chat)

Residuals:
    Min      1Q  Median      3Q     Max
-5.6903 -1.2648 -0.1157  1.3097  5.0225

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.5521     1.1158   3.184  0.00617 **
TotalAccord   0.7127     0.2306   3.091  0.00746 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.89 on 15 degrees of freedom
Multiple R-squared:  0.3891,Adjusted R-squared:  0.3484
F-statistic: 9.553 on 1 and 15 DF,  p-value: 0.007456

paste("Beta=", lm.beta(lmdd))

Beta= 0.623762793540248

tg <- chat[c("TotalDef", "TotalAccord", "Condition")]
tgc <- summarySE(tg, measurevar="TotalAccord", groupvars=c("Condition"))
```

```
tgc2 <- summarySE(tg, measurevar="TotalDef", groupvars=c("Condition"))

tga <- tgc[c("Condition")]
tga[, c("TotalAccord", "sd_accord")] <- tgc[c("TotalAccord", "sd")]
tga[, c("TotalDef", "sd_def")] <- tgc2[c("TotalDef", "sd")]

[1] "Beta= 0.623762793540248"

d <- ggplot(data=tg,aes(Condition, TotalAccord)) +
    geom_point(shape=5,size=3,color="tomato4", data=tg, aes(x=Condition, y=TotalAccord)) +
    geom_errorbar(size=.5,width=.5,color="tomato", data=tga, aes(ymin=TotalAccord-sd_accord, ymax=TotalAccord+sd_accord)) +

    geom_point(shape=4,size=3,color="royalblue4", data=tga, aes(x=Condition, y=TotalDef)) +
    geom_errorbar(size=.5,width=.5,color="lightskyblue2", data=tga, aes(ymin=TotalDef-sd_def, ymax=TotalDef+sd_def)) +

    labs(x="Delay Condition (sec)", y="Number of Words") +
    scale_x_discrete(limits=c(0,2,4,6,8,10)) +
    expand_limits(x=c(-1,11))
    #theme_bw()
    #ggtitle("Definite determiners and agreement by condition")
print(d)
```

figs/definite_determiners_accord.png

```
d <- ggplot(data=tg,aes(Condition, TotalDef)) +
    geom_point(color="blue",shape=18,size=3, data=tga, aes(x=Condition, y=TotalDef)) +
    geom_errorbar(size=.3,width=.5, data=tga, aes(ymin=TotalDef-sd_def, ymax=TotalDef+sd_def)) +
    stat_smooth(linetype="dashed", color="grey40", data=tg, aes(x=Condition, y=TotalDef), method='lm', formula=y~x, se=FALSE

    labs(x="Delay Condition (sec)", y="Number of Words") +
    scale_x_discrete(limits=c(0,2,4,6,8,10)) +
    expand_limits(x=c(-1,11)) +
    expand_limits(y=c(0,18)) +
    ggtitle("Definite Determiners")
print(d)
```


```
d <- ggplot(data=tg,aes(Condition, TotalAccord)) +
    geom_point(color="blue",shape=18,size=3, data=tg, aes(x=Condition, y=TotalAccord)) +
    geom_errorbar(size=.3,width=.5, data=tga, aes(ymin=TotalAccord-sd_accord, ymax=TotalAccord+sd_accord)) +
    stat_smooth(linetype="dashed", color="grey40", data=tg, aes(x=Condition, y=TotalAccord), method='lm', formula=y~x, se=F

    labs(x="Delay Condition (sec)", y="Number of Words") +
    scale_x_discrete(limits=c(0,2,4,6,8,10)) +
    expand_limits(x=c(-1,11)) +
    expand_limits(y=c(0,18)) +
    ggtitle("Accord Words")
print(d)
```



Redundancy with delay + common ground reveals significant effect on both delay condition and common ground Common ground opposes the effect of delay condition on redundancy.

```
chat[,"CommonGround"] <- with(chat, TotalDef + TotalAccord)

chat[,"SumRedundancy"] <- mydata["SumRedundancy"]
lmddc <- lm(data=chat, SumRedundancy~Condition+CommonGround)
summary(lmddc)

Call:
lm(formula = SumRedundancy ~ Condition + CommonGround, data = chat)

Residuals:
   Min     1Q Median     3Q    Max
-2.527 -1.217 -0.124  1.048  4.099

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    8.3256     1.1229   7.414 3.28e-06 ***
Condition      0.6134     0.1337   4.587 0.000423 ***
CommonGround  -0.1757     0.0773  -2.274 0.039262 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.855 on 14 degrees of freedom
Multiple R-squared:  0.6306,Adjusted R-squared:  0.5778
F-statistic: 11.95 on 2 and 14 DF,  p-value: 0.0009383

paste("Beta=", lm.beta(lmddc))
```

Beta= 0.751396081293287
Beta= -0.372470594268601

Total word count in not significant in a model with condition

```
lmddwc <- lm(data=chat, SumRedundancy~Condition+TotalChatWords)
summary(lmddwc)

[1] "Beta= 0.751396081293287"  "Beta= -0.372470594268601"

Call:
lm(formula = SumRedundancy ~ Condition + TotalChatWords, data = chat)

Residuals:
    Min      1Q  Median      3Q     Max
-3.3694 -0.9085 -0.4388  0.6560  4.2446

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)      7.9305     1.2164   6.519 1.36e-05 ***
Condition        0.6260     0.1473   4.251 0.000806 ***
TotalChatWords  -0.0158     0.0102  -1.549 0.143570
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.006 on 14 degrees of freedom
Multiple R-squared:  0.5683,Adjusted R-squared:  0.5066
F-statistic: 9.214 on 2 and 14 DF,  p-value: 0.002796
```

Effect of common ground and delay condition for both high and low -experienced groups

```
chat[,"NewCEExp"] <- mydata["NewCEExp"]

chat_highexp <- chat[chat["NewCEExp"] == 1,]
chat_lowexp <- chat[chat["NewCEExp"] <= .75,]

std <- function(x) sd(x)/sqrt(length(x))

paste("M=", mean(chat_highexp$CommonGround), "SE=", std(chat_highexp$CommonGround))
paste("M=", mean(chat_lowexp$CommonGround), "SE=", std(chat_lowexp$CommonGround))

[1] "M= 10.25 SE= 2.932271182158"
[1] "M= 9.77777777777778 SE= 1.19927961916238"
```

For high-experienced group delay condition is significant but not common ground

```
lexp_gc1 <- lm(data=chat, SumRedundancy~CommonGround+Condition, NewCEExp==1)
summary(lexp_gc1)

Call:
lm(formula = SumRedundancy ~ CommonGround + Condition, data = chat,
    subset = NewCEExp == 1)

Residuals:
      4       6       8      10      11      12      13      14
 1.3790  0.8750 -1.8540 -1.2526 -0.5431  0.6494  0.8639 -0.1176

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   6.51556    0.95272   6.839  0.00102 **
CommonGround -0.11650    0.06432  -1.811  0.12986
Condition     0.80074    0.15088   5.307  0.00317 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.353 on 5 degrees of freedom
Multiple R-squared:  0.8497,Adjusted R-squared:  0.7896
F-statistic: 14.13 on 2 and 5 DF,  p-value: 0.00876
```

For low-experienced group delay still misses significance but common ground is significant

```
lexp_gc2 <- lm(data=chat, SumRedundancy~CommonGround+Condition, NewCEExp<=.75)
summary(lexp_gc2)

Call:
lm(formula = SumRedundancy ~ CommonGround + Condition, data = chat,
    subset = NewCEExp <= 0.75)

Residuals:
    Min      1Q  Median      3Q     Max
-1.3089 -0.7166  0.1354  0.8589  1.4259

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   14.8120     1.4993   9.879 6.21e-05 ***
CommonGround  -0.5782     0.1125  -5.142  0.00213 **
Condition      0.2835     0.1214   2.336  0.05819 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.138 on 6 degrees of freedom
Multiple R-squared:  0.8531,Adjusted R-squared:  0.8042
F-statistic: 17.42 on 2 and 6 DF,  p-value: 0.003169
```