> completedData$F1\_D <- F1\_D$.data\_2

> completedData$F1\_D <- F1\_D$.data\_1

> completedData$F1\_D.1 <- F1\_D$.data\_1

> completedData$F1\_D.2 <- F1\_D$.data\_2

> q2.model <- lm(completedData$ESS.C1 ~ completedData$F1\_D.2, data=completedData)

> summary(q2.model)

Call:

lm(formula = completedData$ESS.C1 ~ completedData$F1\_D.2, data = completedData)

Residuals:

Min 1Q Median 3Q Max

-451.8 -427.8 -63.4 -41.4 30547.2

Coefficients:

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

(Intercept) 452.8 150.8 3.004 0.00283 \*\*

completedData$F1\_D.2 -386.4 200.6 -1.926 0.05478 .

---

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

Residual standard error: 2034 on 416 degrees of freedom

Multiple R-squared: 0.008839, Adjusted R-squared: 0.006456

F-statistic: 3.71 on 1 and 416 DF, p-value: 0.05478

> correlations <- cor(completedData[,1:9])

> corrplot(correlations, method="circle")

> pairs(completedData, col=completedData$F1\_D.2)

> install.packages('caret')

> library('caret')

Loading required package: lattice

> x <- completedData[,1:6]

> y <- completedData[,9]

> scales <- list(x=list(relation='free'), y=list(relation='free'))

> featurePlot(x=x, y=y, plot='density', scales=scales)

NULL

> featurePlot(x=x, y=y, plot='density', scales=scales)

NULL

> glm.fit.F1.C1.3 <- glm(completedData$F1\_D.2 ~ completedData$ESS.C1 + completedData$ESS.C3)

> summary(glm.fit.F1.C1.3)

Call:

glm(formula = completedData$F1\_D.2 ~ completedData$ESS.C1 + completedData$ESS.C3)

Deviance Residuals:

Min 1Q Median 3Q Max

-0.5866 -0.5810 0.4139 0.4165 0.7976

Coefficients:

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

(Intercept) 5.868e-01 2.530e-02 23.199 <2e-16 \*\*\*

completedData$ESS.C1 -1.373e-05 1.245e-05 -1.103 0.2706

completedData$ESS.C3 -1.853e-04 7.970e-05 -2.326 0.0205 \*

---

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

(Dispersion parameter for gaussian family taken to be 0.2422593)

Null deviance: 102.76 on 417 degrees of freedom

Residual deviance: 100.54 on 415 degrees of freedom

AIC: 598.6

Number of Fisher Scoring iterations: 2

> glm.probs <- predict(glm.fit.F1.C1.3, type = "response")

> glm.probs[1:3]

1 2 3

0.5861100 0.5846613 0.5860552

> glm.pred <- ifelse(glm.probs > 0.5, '1' and '0')

Error: unexpected symbol in "glm.pred <- ifelse(glm.probs > 0.5, '1' and"

> glm.pred <- ifelse(glm.probs > 0.5, 1 and 0)

Error: unexpected symbol in "glm.pred <- ifelse(glm.probs > 0.5, 1 and"

> glm.pred <- ifelse(glm.probs > 0.5, "Yes" and "No")

Error: unexpected symbol in "glm.pred <- ifelse(glm.probs > 0.5, "Yes" and"

> glm.pred <- ifelse(glm.probs > 0.5, "Yes", "No")

> attach(completedData)

The following object is masked \_by\_ .GlobalEnv:

F1\_D

> table(glm.pred, completedData$F1\_D.2)

glm.pred 0 1

No 19 6

Yes 163 230

> mean(glm.pred == completedData$F1\_D.2)

[1] 0

> # trying a model between C1, C3 and F1\_D.2 - but using the log vars for C1 and C3

> save.image("C:/Users/JonathanW/Desktop/EES data exploring\_reg\_model\_C1\_C3\_F1.RData")

> glm.fit.F1.C1.log <- glm(completedData$F1\_D.2 ~ completedData$log\_C1 + completedData$log\_C3)

Error in glm.fit(x = c(1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, :

NA/NaN/Inf in 'x'

> hist(completedData$log\_C1)

> hist(completedData$log\_C3)

> install.packages('ResourceSelection')

> library("ResourceSelection")

ResourceSelection 0.3-4 2019-01-08

> hoslem.test(completedData$F1\_D.2m fitted(glm.fit.F1.C1.3))

Error: unexpected symbol in "hoslem.test(completedData$F1\_D.2m fitted"

> hoslem.test(completedData$F1\_D.2, fitted(glm.fit.F1.C1.3))

Hosmer and Lemeshow goodness of fit (GOF) test

data: completedData$F1\_D.2, fitted(glm.fit.F1.C1.3)

X-squared = 6.753, df = 8, p-value = 0.5635

> model\_C1 <- glm(completedData$F1\_D.2 ~ completedData$ESS.C1, fafamily = binominal)

Error in do.call("glm.control", control) : object 'binominal' not found

> model\_C1 <- glm(completedData$F1\_D.2 ~ completedData$ESS.C1, family = binominal)

Error in glm(completedData$F1\_D.2 ~ completedData$ESS.C1, family = binominal) :

object 'binominal' not found

> model\_C1 <- glm(completedData$F1\_D.2 ~ completedData$ESS.C1, family = binomial)

> summary(model\_C1)

Call:

glm(formula = completedData$F1\_D.2 ~ completedData$ESS.C1, family = binomial)

Deviance Residuals:

Min 1Q Median 3Q Max

-1.302 -1.300 1.058 1.059 1.258

Coefficients:

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

(Intercept) 0.2886989 0.1000393 2.886 0.0039 \*\*

completedData$ESS.C1 -0.0001901 0.0001688 -1.126 0.2601

---

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

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 572.48 on 417 degrees of freedom

Residual deviance: 567.80 on 416 degrees of freedom

AIC: 571.8

Number of Fisher Scoring iterations: 6

> model\_C3 <- glm(completedData$F1\_D.2 ~ completedData$ESS.C3, family = binomial)

> summary(model\_C3)

Call:

glm(formula = completedData$F1\_D.2 ~ completedData$ESS.C3, family = binomial)

Deviance Residuals:

Min 1Q Median 3Q Max

-1.331 -1.318 1.032 1.037 1.888

Coefficients:

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

(Intercept) 0.3546897 0.1048055 3.384 0.000714 \*\*\*

completedData$ESS.C3 -0.0009767 0.0003909 -2.499 0.012463 \*

---

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

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 572.48 on 417 degrees of freedom

Residual deviance: 564.33 on 416 degrees of freedom

AIC: 568.33

Number of Fisher Scoring iterations: 4

> range(completedData$ESS.C1)

[1] 1 31000

> range(completedData$ESS.C3)

[1] 0 2000

> xC1 <- seq(0, 31000, 50)

> yC1 <- predict(model\_C1, list(completedData$ESS.C1=xC1), type="response")

Error: unexpected '=' in "yC1 <- predict(model\_C1, list(completedData$ESS.C1="

> yC1 <- predict(model\_C1, list(C1=xC1), type="response")

> plot(completedData$ESS.C1, completedData$F1\_D.2,pch=16, xlab= "Number of Payroll Employees", ylab= "Skills Gaps in Workforce" )

> lines(xC1, yC1, col ='red', 1wd= 2)

Error: unexpected symbol in "lines(xC1, yC1, col ='red', 1wd"

> lines(xC1, yC1, col = "red", 1wd = 2)

Error: unexpected symbol in "lines(xC1, yC1, col = "red", 1wd"

> model\_C1 <- glm(completedData$F1\_D.2 ~ completedData$log\_C1, family = binomial)

> summary(model\_C1)

Call:

glm(formula = completedData$F1\_D.2 ~ completedData$log\_C1, family = binomial)

Deviance Residuals:

Min 1Q Median 3Q Max

-1.367 -1.279 1.026 1.078 1.198

Coefficients:

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

(Intercept) 0.43522 0.20263 2.148 0.0317 \*

completedData$log\_C1 -0.06165 0.06203 -0.994 0.3202

---

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

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 572.48 on 417 degrees of freedom

Residual deviance: 571.48 on 416 degrees of freedom

AIC: 575.48

Number of Fisher Scoring iterations: 4

> range(completedData$log\_C1)

[1] 0.00000 10.34174

> xC1 <- seq(0, 15, 0.01)

> yC1 <- predict(model\_C1, list(log\_C1=xC1), type="response")

> plot(completedData$log\_C1, completedData$F1\_D.2,pch=16, xlab= "Log of Number of Payroll Employees", ylab= "Skills Gaps in Workforce" )

> lines(xC1, yC1, col = "red", 1wd = 2)

Error: unexpected symbol in "lines(xC1, yC1, col = "red", 1wd"

> lines(xC1, yC1, col = "red", lwd=2)

Error in xy.coords(x, y) : 'x' and 'y' lengths differ

> xC1 <- seq(0, 15, 0.1)

> lines(xC1, yC1, col = "red", lwd=2)

Error in xy.coords(x, y) : 'x' and 'y' lengths differ

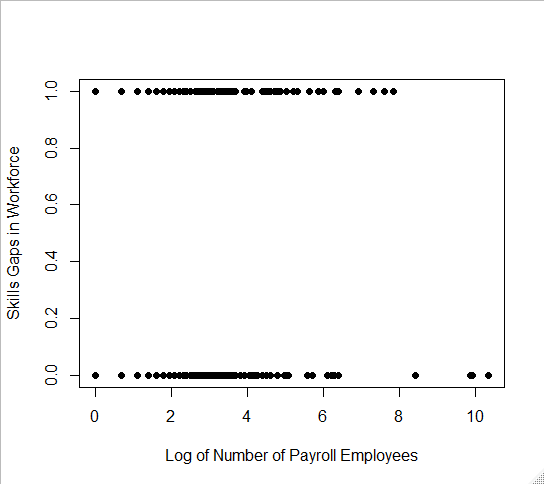
> xC1 <- seq(0, 12, 0.01)

> xC1 <- seq(0, 15, 0.01)

> xC1 <- seq(0, 12, 0.1)

> plot(completedData$log\_C1, completedData$F1\_D.2,pch=16, xlab= "Log of Number of Payroll Employees", ylab= "Skills Gaps in Workforce" )

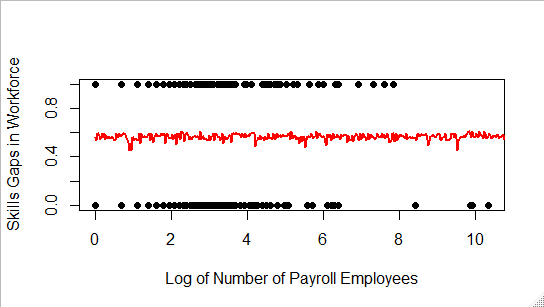
|  |  |
| --- | --- |
|  | |
|  | *Dependent variable:* |
|  |  |
|  | F1\_D.2 |
|  | *normal* |
|  | |
| Constant | 0.59\*\*\* (0.54, 0.64) |
| Number of payroll staff | -0.0000 (-0.0000, 0.0000) |
| Number of freelance staff | -0.0002\*\* (-0.0003, -0.0000) |
|  | |
| Observations | 418 |
| Log Likelihood | -296.30 |
| Akaike Inf. Crit. | 598.60 |
|  | |
| *Note:* | \*p\*\*p\*\*\*p<0.01 |



# code to length vector – therefore enabling a line to be fitted to a regression plot

XC3 <- seq(0, 2500,by=((2500-0)/(50+367)))

# trying to fit the line again



# What I think went wrong here is that there is simply no relationship between the variables; as such it is impossible to fit a sigmoid curve line to the output

# Re fitting the line code

> plot(completedData$log\_C1, completedData$F1\_D.2,pch=16, xlab= "Log of Number of Payroll Employees", ylab= "Skills Gaps in Workforce" )

> lines(xC1, yC1, col="red", lwd=2)

> summary(completedData$log\_C3)

Min. 1st Qu. Median Mean 3rd Qu. Max.

-Inf 1.609 2.708 -Inf 3.912 7.601

# fixing the issue of infity values in Log\_C3

> do.call(completedData$log\_C3, lapply(DT, function(0) replace(0, is.infinite(0), 0.000000001)))

Error: unexpected numeric constant in "do.call(completedData$log\_C3, lapply(DT, function(0"

> do.call(completedData$ESS.C3, lapply(DT, function(0) replace(0, is.infinite(0), 0.000000001)))

Error: unexpected numeric constant in "do.call(completedData$ESS.C3, lapply(DT, function(0"

> is.na(completedData$log\_C3) <- do.call(cbind,lapply(completedData$log\_C3, is.infinite))

> range(completedData$log\_C3)

[1] NA NA

> summary(completedData$log\_C3)

Min. 1st Qu. Median Mean 3rd Qu. Max. NA's

0.000 1.609 2.708 2.912 3.912 7.601 5

> completedData$log\_C3[is.na(completedData$log\_C3)] <- 0.0000000000000001

# Log\_C3 model ready for plotting

> model\_C3\_log <- glm(completedData$F1\_D.2 ~ completedData$log\_C3, family = binomial)

> summary(model\_C3\_log)

Call:

glm(formula = completedData$F1\_D.2 ~ completedData$log\_C3, family = binomial)

Deviance Residuals:

Min 1Q Median 3Q Max

-1.5607 -1.2444 0.8898 1.0511 1.4951

Coefficients:

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

(Intercept) 0.86706 0.20034 4.328 1.51e-05 \*\*\*

completedData$log\_C3 -0.20900 0.05936 -3.521 0.00043 \*\*\*

---

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

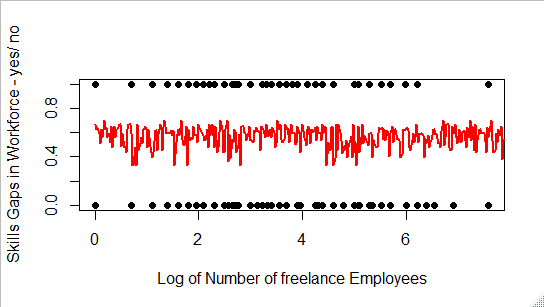
(Dispersion parameter for binomial family taken to be 1)

Null deviance: 572.48 on 417 degrees of freedom

Residual deviance: 559.61 on 416 degrees of freedom

AIC: 563.61

Number of Fisher Scoring iterations: 4



Again I believe this shows that there is limited correlation between the variables, as such it is impossible to fit a sigmoid line to the data and offer some form of prediction between the variables. The code can be viewed below.

> range(completedData$log\_C3)

[1] 0.000000 7.600902

> xC3 <- seq(-0.0001, 10, by=((10-0)/(10+407)))

> yC3 <- predict(model\_C3\_log, list(F1\_D.2=xC3), type='response')

> plot(completedData$log\_C3, completedData$F1\_D.2,pch=16, xlab= "Log of Number of freelance Employees", ylab= "Skills Gaps in Workforce - yes/ no" )

> abline(model\_C3\_log)

> lines(xC3, yC3, col="red", lwd=2)

# Regression and correlation test between C1 and C3

> cor(completedData$log\_C1, completedData$log\_C3)

[1] 0.2667513

> plot(completedData$log\_C1, completedData$log\_C3)

> plot(completedData$log\_C3, completedData$log\_C1)

> plot(completedData$log\_C1, completedData$log\_C3)

> plot(completedData$log\_C1 ~ completedData$log\_C3, data=completedData)

> mod\_C1\_C3\_pred <- lm(completedData$log\_C1 ~ completedData$log\_C3, data=completedData)

> summary(mod\_C1\_C3\_pred)

Call:

lm(formula = completedData$log\_C1 ~ completedData$log\_C3, data = completedData)

Residuals:

Min 1Q Median 3Q Max

-3.2633 -1.1738 -0.0912 0.9473 7.9455

Coefficients:

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

(Intercept) 2.1246 0.1467 14.479 < 2e-16 \*\*\*

completedData$log\_C3 0.2473 0.0438 5.645 3.06e-08 \*\*\*

---

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

Residual standard error: 1.537 on 416 degrees of freedom

Multiple R-squared: 0.07116, Adjusted R-squared: 0.06892

F-statistic: 31.87 on 1 and 416 DF, p-value: 3.058e-08

> abline(mod\_C1\_C3\_pred)

> save.image("C:/Users/JonathanW/Desktop/EES data exploring\_reg\_model\_C1\_C3\_F1 V1.RData")

> mod\_C3\_C1\_pred <- lm(completedData$log\_C3 ~ completedData$log\_C1, data=completedData)

> summary(mod\_C3\_C1\_pred)

Call:

lm(formula = completedData$log\_C3 ~ completedData$log\_C1, data = completedData)

Residuals:

Min 1Q Median 3Q Max

-3.9382 -1.2069 -0.0711 1.1024 4.0491

Coefficients:

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

(Intercept) 2.06078 0.16577 12.432 < 2e-16 \*\*\*

completedData$log\_C1 0.28777 0.05098 5.645 3.06e-08 \*\*\*

---

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

Residual standard error: 1.658 on 416 degrees of freedom

Multiple R-squared: 0.07116, Adjusted R-squared: 0.06892

F-statistic: 31.87 on 1 and 416 DF, p-value: 3.058e-08

> plot(completedData$log\_C3 ~ completedData$log\_C1, data=completedData)

> abline(mod\_C3\_C1\_pred)

> library("sjPlot")

Registered S3 methods overwritten by 'ggplot2':

method from

[.quosures rlang

c.quosures rlang

print.quosures rlang

> library("stargazer")

Please cite as:

Hlavac, Marek (2018). stargazer: Well-Formatted Regression and Summary Statistics Tables.

R package version 5.2.2. https://CRAN.R-project.org/package=stargazer

> stargazer(mod\_C1\_C3\_pred, mod\_C3\_C1\_pred, type="html", out = "lm.reg\_models\_payroll\_freelance.doc")

<table style="text-align:center"><tr><td colspan="3" style="border-bottom: 1px solid black"></td></tr><tr><td style="text-align:left"></td><td colspan="2"><em>Dependent variable:</em></td></tr>

<tr><td></td><td colspan="2" style="border-bottom: 1px solid black"></td></tr>

<tr><td style="text-align:left"></td><td>log\_C1</td><td>log\_C3</td></tr>

<tr><td style="text-align:left"></td><td>(1)</td><td>(2)</td></tr>

<tr><td colspan="3" style="border-bottom: 1px solid black"></td></tr><tr><td style="text-align:left">log\_C3</td><td>0.247<sup>\*\*\*</sup></td><td></td></tr>

<tr><td style="text-align:left"></td><td>(0.044)</td><td></td></tr>

<tr><td style="text-align:left"></td><td></td><td></td></tr>

<tr><td style="text-align:left">log\_C1</td><td></td><td>0.288<sup>\*\*\*</sup></td></tr>

<tr><td style="text-align:left"></td><td></td><td>(0.051)</td></tr>

<tr><td style="text-align:left"></td><td></td><td></td></tr>

<tr><td style="text-align:left">Constant</td><td>2.125<sup>\*\*\*</sup></td><td>2.061<sup>\*\*\*</sup></td></tr>

<tr><td style="text-align:left"></td><td>(0.147)</td><td>(0.166)</td></tr>

<tr><td style="text-align:left"></td><td></td><td></td></tr>

<tr><td colspan="3" style="border-bottom: 1px solid black"></td></tr><tr><td style="text-align:left">Observations</td><td>418</td><td>418</td></tr>

<tr><td style="text-align:left">R<sup>2</sup></td><td>0.071</td><td>0.071</td></tr>

<tr><td style="text-align:left">Adjusted R<sup>2</sup></td><td>0.069</td><td>0.069</td></tr>

<tr><td style="text-align:left">Residual Std. Error (df = 416)</td><td>1.537</td><td>1.658</td></tr>

<tr><td style="text-align:left">F Statistic (df = 1; 416)</td><td>31.869<sup>\*\*\*</sup></td><td>31.869<sup>\*\*\*</sup></td></tr>

<tr><td colspan="3" style="border-bottom: 1px solid black"></td></tr><tr><td style="text-align:left"><em>Note:</em></td><td colspan="2" style="text-align:right"><sup>\*</sup>p<0.1; <sup>\*\*</sup>p<0.05; <sup>\*\*\*</sup>p<0.01</td></tr>

</table>

> median(completedData$ratio2)

[1] 0.8416667

> summary(completedData$ratio2)

Min. 1st Qu. Median Mean 3rd Qu. Max.

0.0100 0.2000 0.8417 Inf 3.9375 Inf

> summary(completedData$ESS.C1)

Min. 1st Qu. Median Mean 3rd Qu. Max.

1.0 5.0 17.0 234.6 40.0 31000.0

> summary(completedData$ESS.C3)

Min. 1st Qu. Median Mean 3rd Qu. Max.

0.0 5.0 15.0 102.6 50.0 2000.0

> 234/102.6

[1] 2.280702

> ratio\_pr\_fr <- 2/28

> ratio\_pr\_fr

[1] 0.07142857

> 17/15

[1] 1.133333

> 15/17

[1] 0.8823529

> summary(completedData$ratio)

Min. 1st Qu. Median Mean 3rd Qu. Max.

0.0000 0.2542 1.1882 4.6817 5.0000 100.0000

> 15/17

[1] 0.8823529

> cor(completedData$log\_C1, completedData$log\_C3)

[1] 0.2667513

> cor(completedData$log\_C3, completedData$log\_C1)

[1] 0.2667513

> plot(mod\_C1\_C3\_pred, which = 1)

> plot(mod\_C3\_C1\_pred, which = 1)

> library("MASS")

> data(completedData)

Warning message:

In data(completedData) : data set ‘completedData’ not found

> plot(completedData$log\_C1 ~ completedData$log\_C3, data=completedData)

> abline(lm(completedData$log\_C1 ~ completedData$log\_C3, data = completedData), col="blue", lty=2, lwd=2)

> abline(lm(completedData$log\_C1 ~ completedData$log\_C3, data = completedData), col="red", lty=1, lwd=2)

> abline(lm(completedData$log\_C1 ~ completedData$log\_C3, data = completedData), col="blue", lty=2, lwd=2)

> plot(completedData$log\_C1 ~ completedData$log\_C3, data=completedData)

> abline(lm(completedData$log\_C1 ~ completedData$log\_C3, data = completedData), col="blue", lty=2, lwd=2)

> abline(rlm(completedData$log\_C1 ~ completedData$log\_C3, data = completedData), col="blue", lty=1, lwd=2)

# Residual plots

> par(mfrow = c(2, 2))

> plot(mod\_C3\_C1\_pred, which = 1)

> plot(mod\_C3\_C1\_pred)

> plot(mod\_C1\_C3\_pred)

# dummy var for B3

> Subsectors\_B3\_D <- fastDummies::dummy\_cols(ESS$B3)

> completedData$Film\_D <- Subsectors\_B3\_D$.data\_1

> completedData$HETV\_D <- Subsectors\_B3\_D$.data\_2

> completedData$Unscripted\_D <- Subsectors\_B3\_D$.data\_3

> completedData$Childrens\_D <- Subsectors\_B3\_D$.data\_4

> completedData$Animation\_D <- Subsectors\_B3\_D$.data\_5

> completedData$VFX\_D <- Subsectors\_B3\_D$.data\_6

> completedData$Games\_D <- Subsectors\_B3\_D$.data\_7

> completedData$Other\_Sector\_D <- Subsectors\_B3\_D$.data\_8

> completedData$Sectors <- Subsectors\_B3\_D$.data

# removing a column

> completedData$F1\_D <- NULL

# Dummy variable for B5

> B5\_D <- fastDummies::dummy\_cols(ESS$B5)

> completedData$Crewing\_Up\_Bring\_Own <- B5\_D$.data\_1

> completedData$Crewing\_Up\_Local <- B5\_D$.data\_2

> completedData$Crewing\_Up\_Both <- B5\_D$.data\_3

> completedData$Crewing\_Up\_DK <- B5\_D$.data\_5

> completedData$Crewing\_Up <- B5\_D$.data

# Dummy variable for D1\_d and assigning the dummy variable to the data frame completedData

> D1\_D <- fastDummies::dummy\_cols(ESS$D1)

> View(D1\_D)

> completedData$Div\_mon\_yes <- D1\_D$.data\_1

> completedData$Div\_mon\_no <- D1\_D$.data\_2

> View(E2\_D)

> completedData$Recruit\_Dif\_Yes <- E2\_D$.data\_1

> completedData$Recruit\_Dif\_No <- E2\_D$.data\_2

# Dummy Variable for G1

> G1\_D <- fastDummies::dummy\_cols(ESS$G1)

> completedData$Training\_Yes <- G1\_D$.data\_1

> completedData$Training\_No <- G1\_D$.data\_2

# Dummy variables for H3

> H3\_1\_D <- fastDummies::dummy\_cols(ESS$H1\_01)

> H3\_2\_D <- fastDummies::dummy\_cols(ESS$H1\_02)

> H3\_3\_D <- fastDummies::dummy\_cols(ESS$H1\_03)

> H3\_4\_D <- fastDummies::dummy\_cols(ESS$H1\_04)

> completedData$SGs\_Improve <- H3\_1\_D$.data\_1

> completedData$SGs\_Remain\_the\_same <- H3\_1\_D$.data\_2

> completedData$SGs\_get\_worse <- H3\_1\_D$.data\_3

> completedData$SGs\_DK <- H3\_1\_D$.data\_4

> completedData$RD\_improve <- H3\_2\_D$.data\_1

> completedData$RD\_remain\_same <- H3\_2\_D$.data\_2

> completedData$RD\_get\_worse <- H3\_2\_D$.data\_3

> completedData$RD\_DK <- H3\_2\_D$.data\_4

> completedData$Div\_improve <- H3\_3\_D$.data\_1

> completedData$RD\_remain\_same <- H3\_3\_D$.data\_2

> completedData$RD\_remain\_same <- H3\_2\_D$.data\_2

> completedData$Div\_remain\_same <- H3\_3\_D$.data\_2

> completedData$Div\_get\_worse <- H3\_3\_D$.data\_3

> completedData$Div\_DK <- H3\_3\_D$.data\_4

> completedData$london.divide\_improve <- H3\_4\_D$.data\_1

> completedData$london.divide\_remain\_same <- H3\_4\_D$.data\_2

> completedData$london.divide\_get\_worse <- H3\_4\_D$.data\_3

> completedData$london.divide\_DK <- H3\_4\_D$.data\_4

# looking for outliers

> boxplot(subset.log.C1$ESS.C1)

> boxplot(subset.log.C1$ESS.C3)

> boxplot(subset.log.C1$log\_C1)

> summary(completedData$log\_C1)

Min. 1st Qu. Median Mean 3rd Qu. Max.

0.000 1.609 2.833 2.836 3.689 10.342

> boxplot(completedData$ESS.C1)$out

[1] 400 20000 600 19000 121 120 500 350 530 1000 150 500 400 2500

[15] 550 160 4500 300 1500 100 280 98 145 260 182 200 100 2000

[29] 144 31000 100 110 100 150 600 130 450

> outliers.C1 <- boxplot(completedData$ESS.C1)$out

> outliers.C3 <- boxplot(completedData$ESS.C3)$out

> print(outliers.C3)

[1] 200 2000 2000 2000 162 200 500 200 163 162 2000 2000 212 200 200 1000 600

[18] 200 2000 162 200 2000 300 150 2000 600 150 250 150 500 500 2000 200 500

[35] 400 500 300 150 500 162 162 150 200 200 200 150 700 150 409 300 162

[52] 250 500 500 200 120 200 500 500 2000

> completedData.1 <- completedData[-which(completedData$ESS.C1 %in% outliers.C1),]

> summary(completedData.1$ESS.C1, completedData.1$ESS.C3)

Min. 1st Qu. Median Mean 3rd Qu. Max.

1.00 5.00 16.00 22.95 27.00 90.00

> save.image("U:/Research/R data environments/EES data exploring\_reg\_model\_C1\_C3\_F1 V3.RData")

> mod\_C1\_C3\_outliers <- lm(completedData.1$ESS.C1 ~ completedData.1$ESS.C3, data = completedData.1)

> summary(mod\_C1\_C3\_outliers)

Call:

lm(formula = completedData.1$ESS.C1 ~ completedData.1$ESS.C3,

data = completedData.1)

Residuals:

Min 1Q Median 3Q Max

-33.831 -17.519 -9.376 3.652 68.681

Coefficients:

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

(Intercept) 21.31940 1.35910 15.686 < 2e-16 \*\*\*

completedData.1$ESS.C3 0.02851 0.01028 2.773 0.00582 \*\*

---

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

Residual standard error: 23.93 on 379 degrees of freedom

Multiple R-squared: 0.01989, Adjusted R-squared: 0.01731

F-statistic: 7.692 on 1 and 379 DF, p-value: 0.005821

>

> mod\_C3\_C1\_outliers <- lm(completedData.1$ESS.C3 ~ completedData.1$ESS.C1, data = completedData.1)

> summary(mod\_C3\_C1\_outliers)

Call:

lm(formula = completedData.1$ESS.C3 ~ completedData.1$ESS.C1,

data = completedData.1)

Residuals:

Min 1Q Median 3Q Max

-103.88 -49.95 -37.19 -10.05 947.74

Coefficients:

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

(Intercept) 41.0929 8.3718 4.909 1.37e-06 \*\*\*

completedData.1$ESS.C1 0.6977 0.2515 2.773 0.00582 \*\*

---

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

Residual standard error: 118.4 on 379 degrees of freedom

Multiple R-squared: 0.01989, Adjusted R-squared: 0.01731

F-statistic: 7.692 on 1 and 379 DF, p-value: 0.005821

> plot(completedData.1$ESS.C3, completedData.1$ESS.C1)

> plot(completedData.1$ESS.C1, completedData.1$ESS.C3)

> abline(mod\_C1\_C3\_outliers)

> plot(completedData.1$ESS.C1, completedData.1$ESS.C3)

> abline(mod\_C3\_C1\_outliers)

> hist(subset.CD.C3$ESS.C1)

> hist(subset.CD.C3$ESS.C3)

> summary(completedData$ESS.C1)

Min. 1st Qu. Median Mean 3rd Qu. Max.

1.0 5.0 17.0 234.6 40.0 31000.0

> summary(completedData.1$ESS.C1)

Min. 1st Qu. Median Mean 3rd Qu. Max.

1.00 5.00 16.00 22.95 27.00 90.00

> outliers.C1

[1] 400 20000 600 19000 121 120 500 350 530 1000 150 500 400 2500

[15] 550 160 4500 300 1500 100 280 98 145 260 182 200 100 2000

[29] 144 31000 100 110 100 150 600 130 450

> boxplot(completedData.1$)

Error: unexpected ')' in "boxplot(completedData.1$)"

> quantile(completedData$ESS.C1)

0% 25% 50% 75% 100%

1 5 17 40 31000

> quantile(completedData$ESS.C1, 0.95)

95%

209

> quantile(completedData$ESS.C1, 0.97)

97%

500

> hist(subset.CD.C3$ESS.C1)

> quantile(completedData$ESS.C1, 0.80)

80%

60

> quantile(completedData$ESS.C1, 0.85)

85%

67.25

> quantile(completedData$ESS.C1, 0.90)

90%

90

> summary(completedData.1$ESS.C3)

Min. 1st Qu. Median Mean 3rd Qu. Max.

0.0 5.0 12.0 57.1 40.0 1000.0

> sd(completedData$ESS.C1)

[1] 2040.442

> sd(completedData.1$ESS.C1)

[1] 24.13607

> sd(subset.CD.C3$ESS.C1)

[1] 24.04109

> hist(completedData.1$ESS.C1)

> boxplot.stats(completedData$ESS.C1)

$stats

[1] 1 5 17 40 90

$n

[1] 418

$conf

[1] 14.29519 19.70481

$out

[1] 400 20000 600 19000 121 120 500 350 530 1000 150 500 400 2500

[15] 550 160 4500 300 1500 100 280 98 145 260 182 200 100 2000

[29] 144 31000 100 110 100 150 600 130 450

> boxplot.stats(completedData.1$ESS.C1)

$stats

[1] 1 5 16 27 60

$n

[1] 381

$conf

[1] 14.21919 17.78081

$out

[1] 90 90 90 90 89 65 90 90 90 90 90 90 90 90 90 90 90 90 80 90 70 80 90 90 84 80 70 65

# R script for outlier function with automated plotting

outlierKD <- function(dt, var) {

var\_name <- eval(substitute(var),eval(dt))

na1 <- sum(is.na(var\_name))

m1 <- mean(var\_name, na.rm = T)

par(mfrow=c(2, 2), oma=c(0,0,3,0))

boxplot(var\_name, main="With outliers")

hist(var\_name, main="With outliers", xlab=NA, ylab=NA)

outlier <- boxplot.stats(var\_name)$out

mo <- mean(outlier)

var\_name <- ifelse(var\_name %in% outlier, NA, var\_name)

boxplot(var\_name, main="Without outliers")

hist(var\_name, main="Without outliers", xlab=NA, ylab=NA)

title("Outlier Check", outer=TRUE)

na2 <- sum(is.na(var\_name))

cat("Outliers identified:", na2 - na1, "n")

cat("Propotion (%) of outliers:", round((na2 - na1) / sum(!is.na(var\_name))\*100, 1), "n")

cat("Mean of the outliers:", round(mo, 2), "n")

m2 <- mean(var\_name, na.rm = T)

cat("Mean without removing outliers:", round(m1, 2), "n")

cat("Mean if we remove outliers:", round(m2, 2), "n")

response <- readline(prompt="Do you want to remove outliers and to replace with NA? [yes/no]: ")

if(response == "y" | response == "yes"){

dt[as.character(substitute(var))] <- invisible(var\_name)

assign(as.character(as.list(match.call())$dt), dt, envir = .GlobalEnv)

cat("Outliers successfully removed", "n")

return(invisible(dt))

} else{

cat("Nothing changed", "n")

return(invisible(var\_name))

}

}

# deciding on which vars to use in MV Logistic reg

> corr.test(completedData.1, use = "pairwise", method = "spearman", adjust = "none", alpha = .05)

Call:corr.test(x = completedData.1, use = "pairwise", method = "spearman",

adjust = "none", alpha = 0.05)

Correlation matrix

ESS.C1 ESS.C3 ratio ratio2 log\_C1 log\_C3 F1\_D.1 F1\_D.2 Film\_D

ESS.C1 1.00 0.08 -0.52 0.52 1.00 0.08 -0.07 -0.02 0.03

ESS.C3 0.08 1.00 0.78 -0.78 0.08 1.00 0.13 -0.14 -0.16

ratio -0.52 0.78 1.00 -1.00 -0.52 0.78 0.16 -0.12 -0.17

ratio2 0.52 -0.78 -1.00 1.00 0.52 -0.78 -0.16 0.12 0.17

log\_C1 1.00 0.08 -0.52 0.52 1.00 0.08 -0.07 -0.02 0.03

log\_C3 0.08 1.00 0.78 -0.78 0.08 1.00 0.12 -0.14 -0.16

F1\_D.1 -0.07 0.13 0.16 -0.16 -0.07 0.12 1.00 -0.83 -0.03

F1\_D.2 -0.02 -0.14 -0.12 0.12 -0.02 -0.14 -0.83 1.00 0.05

Film\_D 0.03 -0.16 -0.17 0.17 0.03 -0.16 -0.03 0.05 1.00

HETV\_D 0.07 0.23 0.18 -0.18 0.07 0.23 0.01 -0.03 -0.38

Unscripted\_D 0.11 0.08 0.00 0.00 0.11 0.08 0.00 0.04 -0.29

Childrens\_D -0.04 0.05 0.07 -0.07 -0.04 0.05 -0.07 0.04 -0.18

Animation\_D -0.11 -0.03 0.01 -0.01 -0.11 -0.03 0.07 -0.02 -0.22

VFX\_D -0.03 -0.06 -0.02 0.02 -0.03 -0.06 0.03 -0.04 -0.12

Games\_D -0.06 -0.13 -0.06 0.06 -0.06 -0.13 0.03 0.00 -0.22

Other\_Sector\_D -0.08 0.00 0.04 -0.04 -0.08 0.00 -0.01 -0.10 -0.23

Sectors -0.10 0.03 0.09 -0.09 -0.10 0.03 0.04 -0.07 -0.87

Crewing\_Up\_Bring\_Own 0.01 -0.04 -0.04 0.04 0.01 -0.04 0.11 -0.12 0.00

Crewing\_Up\_Local 0.04 0.13 0.08 -0.08 0.04 0.13 0.16 -0.14 -0.05

Crewing\_Up\_Both -0.07 0.20 0.25 -0.25 -0.07 0.20 -0.01 0.04 -0.17

Crewing\_Up\_DK 0.16 0.02 -0.07 0.07 0.16 0.02 -0.06 -0.04 -0.03

Crewing\_Up 0.02 -0.03 -0.04 0.04 0.02 -0.02 -0.22 0.13 -0.05

HETV\_D Unscripted\_D Childrens\_D Animation\_D VFX\_D Games\_D

ESS.C1 0.07 0.11 -0.04 -0.11 -0.03 -0.06

ESS.C3 0.23 0.08 0.05 -0.03 -0.06 -0.13

ratio 0.18 0.00 0.07 0.01 -0.02 -0.06

ratio2 -0.18 0.00 -0.07 -0.01 0.02 0.06

log\_C1 0.07 0.11 -0.04 -0.11 -0.03 -0.06

log\_C3 0.23 0.08 0.05 -0.03 -0.06 -0.13

F1\_D.1 0.01 0.00 -0.07 0.07 0.03 0.03

F1\_D.2 -0.03 0.04 0.04 -0.02 -0.04 0.00

Film\_D -0.38 -0.29 -0.18 -0.22 -0.12 -0.22

HETV\_D 1.00 -0.18 -0.12 -0.14 -0.08 -0.14

Unscripted\_D -0.18 1.00 -0.09 -0.11 -0.06 -0.11

Childrens\_D -0.12 -0.09 1.00 -0.07 -0.04 -0.07

Animation\_D -0.14 -0.11 -0.07 1.00 -0.04 -0.08

VFX\_D -0.08 -0.06 -0.04 -0.04 1.00 -0.04

Games\_D -0.14 -0.11 -0.07 -0.08 -0.04 1.00

Other\_Sector\_D -0.14 -0.11 -0.07 -0.08 -0.05 -0.08

Sectors -0.04 0.18 0.19 0.29 0.19 0.39

Crewing\_Up\_Bring\_Own 0.00 0.12 -0.02 -0.04 -0.04 -0.04

Crewing\_Up\_Local 0.06 -0.02 0.05 0.06 0.04 -0.07

Crewing\_Up\_Both 0.25 0.18 0.10 -0.10 -0.05 -0.20

Crewing\_Up\_DK -0.02 0.07 0.06 -0.04 -0.02 -0.04

Crewing\_Up -0.03 -0.07 0.05 0.16 0.01 -0.08

Other\_Sector\_D Sectors Crewing\_Up\_Bring\_Own Crewing\_Up\_Local

ESS.C1 -0.08 -0.10 0.01 0.04

ESS.C3 0.00 0.03 -0.04 0.13

ratio 0.04 0.09 -0.04 0.08

ratio2 -0.04 -0.09 0.04 -0.08

log\_C1 -0.08 -0.10 0.01 0.04

log\_C3 0.00 0.03 -0.04 0.13

F1\_D.1 -0.01 0.04 0.11 0.16

F1\_D.2 -0.10 -0.07 -0.12 -0.14

Film\_D -0.23 -0.87 0.00 -0.05

HETV\_D -0.14 -0.04 0.00 0.06

Unscripted\_D -0.11 0.18 0.12 -0.02

Childrens\_D -0.07 0.19 -0.02 0.05

Animation\_D -0.08 0.29 -0.04 0.06

VFX\_D -0.05 0.19 -0.04 0.04

Games\_D -0.08 0.39 -0.04 -0.07

Other\_Sector\_D 1.00 0.48 -0.04 -0.03

Sectors 0.48 1.00 -0.03 0.02

Crewing\_Up\_Bring\_Own -0.04 -0.03 1.00 -0.07

Crewing\_Up\_Local -0.03 0.02 -0.07 1.00

Crewing\_Up\_Both -0.05 0.00 -0.21 -0.19

Crewing\_Up\_DK 0.03 0.02 -0.04 -0.03

Crewing\_Up 0.05 0.10 -0.63 -0.43

Crewing\_Up\_Both Crewing\_Up\_DK Crewing\_Up Div\_mon\_yes Div\_mon\_no

ESS.C1 -0.07 0.16 0.02 -0.06 -0.33

ESS.C3 0.20 0.02 -0.03 0.17 -0.02

ratio 0.25 -0.07 -0.04 0.18 0.19

ratio2 -0.25 0.07 0.04 -0.18 -0.19

log\_C1 -0.07 0.16 0.02 -0.06 -0.33

log\_C3 0.20 0.02 -0.02 0.17 -0.02

F1\_D.1 -0.01 -0.06 -0.22 0.05 0.02

F1\_D.2 0.04 -0.04 0.13 -0.03 0.05

Film\_D -0.17 -0.03 -0.05 -0.05 -0.07

HETV\_D 0.25 -0.02 -0.03 0.10 -0.05

Unscripted\_D 0.18 0.07 -0.07 0.07 -0.05

Childrens\_D 0.10 0.06 0.05 0.05 0.00

Animation\_D -0.10 -0.04 0.16 -0.02 0.02

VFX\_D -0.05 -0.02 0.01 -0.09 0.17

Games\_D -0.20 -0.04 -0.08 -0.10 0.17

Other\_Sector\_D -0.05 0.03 0.05 -0.02 -0.02

Sectors 0.00 0.02 0.10 -0.02 0.12

Crewing\_Up\_Bring\_Own -0.21 -0.04 -0.63 -0.01 -0.05

Crewing\_Up\_Local -0.19 -0.03 -0.43 0.04 -0.04

Crewing\_Up\_Both 1.00 -0.11 0.08 0.11 -0.09

Crewing\_Up\_DK -0.11 1.00 0.34 -0.07 -0.02

Crewing\_Up 0.08 0.34 1.00 -0.07 0.06

Recruit\_Dif\_Yes Recruit\_Dif\_No Training\_Yes Training\_No

ESS.C1 0.00 -0.08 0.09 -0.11

ESS.C3 0.10 0.05 0.20 -0.21

ratio 0.10 0.08 0.10 -0.09

ratio2 -0.10 -0.08 -0.10 0.09

log\_C1 0.00 -0.08 0.09 -0.11

log\_C3 0.10 0.05 0.20 -0.21

F1\_D.1 0.39 -0.18 0.08 -0.07

F1\_D.2 -0.31 0.23 0.00 0.06

Film\_D -0.09 0.02 -0.03 0.08

HETV\_D 0.00 -0.01 0.08 -0.08

Unscripted\_D 0.01 0.12 0.09 -0.12

Childrens\_D 0.05 0.01 0.07 -0.09

Animation\_D 0.09 -0.08 -0.05 0.06

VFX\_D 0.05 0.00 0.00 0.02

Games\_D 0.07 -0.04 -0.09 0.10

Other\_Sector\_D -0.07 -0.06 -0.07 0.04

Sectors 0.09 -0.05 -0.04 0.00

Crewing\_Up\_Bring\_Own 0.13 -0.15 0.01 -0.05

Crewing\_Up\_Local 0.19 -0.10 0.10 -0.10

Crewing\_Up\_Both -0.05 0.15 0.10 -0.06

Crewing\_Up\_DK -0.05 0.04 -0.11 -0.03

Crewing\_Up -0.29 0.17 -0.22 0.18

SGs\_Improve SGs\_Remain\_the\_same SGs\_get\_worse SGs\_DK RD\_improve

ESS.C1 -0.05 -0.03 -0.05 0.17 -0.01

ESS.C3 0.00 0.01 -0.05 0.05 0.04

ratio 0.03 0.05 -0.02 -0.08 0.04

ratio2 -0.03 -0.05 0.02 0.08 -0.04

log\_C1 -0.05 -0.03 -0.05 0.17 -0.01

log\_C3 0.00 0.01 -0.05 0.04 0.04

F1\_D.1 0.08 -0.01 -0.11 0.02 0.03

F1\_D.2 -0.06 0.01 0.16 -0.11 -0.05

Film\_D 0.01 0.01 0.00 -0.03 0.05

HETV\_D -0.08 0.12 -0.05 0.01 0.02

Unscripted\_D 0.04 -0.07 0.05 -0.02 -0.06

Childrens\_D -0.01 -0.02 -0.02 0.07 -0.03

Animation\_D 0.01 -0.01 0.02 -0.03 -0.08

VFX\_D -0.01 -0.03 -0.03 0.08 -0.01

Games\_D 0.01 -0.01 0.10 -0.12 -0.02

Other\_Sector\_D 0.04 -0.06 -0.06 0.10 0.08

Sectors 0.02 -0.06 0.01 0.04 -0.04

Crewing\_Up\_Bring\_Own 0.07 -0.06 0.00 -0.02 0.01

Crewing\_Up\_Local 0.05 -0.03 0.00 -0.03 -0.03

Crewing\_Up\_Both 0.00 0.13 -0.10 -0.06 -0.05

Crewing\_Up\_DK -0.07 0.00 -0.07 0.17 -0.05

Crewing\_Up -0.12 0.05 -0.06 0.16 0.03

RD\_remain\_same RD\_get\_worse RD\_DK Div\_improve Div\_remain\_same

ESS.C1 -0.01 -0.05 0.08 -0.01 -0.08

ESS.C3 -0.01 -0.02 0.00 0.07 -0.04

ratio 0.00 0.01 -0.05 0.07 0.03

ratio2 0.00 -0.01 0.05 -0.07 -0.03

log\_C1 -0.01 -0.05 0.08 -0.01 -0.08

log\_C3 -0.01 -0.02 0.00 0.07 -0.04

F1\_D.1 0.06 -0.09 -0.03 0.04 0.14

F1\_D.2 -0.02 0.14 -0.09 -0.06 -0.07

Film\_D -0.08 0.06 0.00 -0.04 -0.15

HETV\_D 0.03 -0.06 0.01 -0.01 0.02

Unscripted\_D 0.09 -0.01 -0.06 -0.04 0.14

Childrens\_D 0.02 -0.08 0.10 0.00 -0.01

Animation\_D 0.02 0.08 -0.04 0.01 0.03

VFX\_D 0.10 -0.07 -0.06 0.06 0.01

Games\_D 0.00 0.03 -0.01 -0.06 0.11

Other\_Sector\_D -0.10 -0.01 0.07 0.16 -0.07

Sectors 0.04 -0.03 0.02 0.08 0.12

Crewing\_Up\_Bring\_Own 0.01 -0.02 0.00 0.12 -0.01

Crewing\_Up\_Local -0.02 0.04 0.02 -0.08 0.10

Crewing\_Up\_Both 0.09 -0.05 -0.04 0.02 0.14

Crewing\_Up\_DK 0.00 -0.01 0.07 -0.05 0.01

Crewing\_Up -0.03 -0.01 0.02 0.00 -0.03

Div\_get\_worse Div\_DK london.divide\_improve

ESS.C1 0.02 0.10 -0.08

ESS.C3 0.02 -0.03 0.03

ratio -0.01 -0.09 0.08

ratio2 0.01 0.09 -0.08

log\_C1 0.02 0.10 -0.08

log\_C3 0.02 -0.03 0.03

F1\_D.1 -0.15 -0.04 0.04

F1\_D.2 0.21 -0.12 0.00

Film\_D 0.21 -0.01 -0.05

HETV\_D -0.06 0.05 0.06

Unscripted\_D -0.10 -0.04 -0.06

Childrens\_D -0.04 0.06 0.00

Animation\_D -0.01 -0.04 0.08

VFX\_D -0.06 0.00 -0.01

Games\_D -0.06 -0.04 -0.01

Other\_Sector\_D -0.04 0.02 0.01

Sectors -0.19 -0.01 0.04

Crewing\_Up\_Bring\_Own -0.09 0.03 -0.07

Crewing\_Up\_Local -0.02 -0.05 0.04

Crewing\_Up\_Both -0.13 -0.07 0.06

Crewing\_Up\_DK -0.04 0.07 -0.12

Crewing\_Up 0.01 0.04 0.06

london.divide\_remain\_same london.divide\_get\_worse london.divide\_DK

ESS.C1 -0.01 -0.05 0.15

ESS.C3 -0.04 -0.06 0.03

ratio -0.01 -0.02 -0.07

ratio2 0.01 0.02 0.07

log\_C1 -0.01 -0.05 0.15

log\_C3 -0.04 -0.06 0.03

F1\_D.1 -0.08 -0.09 0.08

F1\_D.2 0.09 0.11 -0.15

Film\_D 0.03 0.05 0.00

HETV\_D -0.05 -0.10 0.04

Unscripted\_D 0.03 0.06 0.00

Childrens\_D -0.05 0.02 0.00

Animation\_D -0.02 -0.04 -0.06

VFX\_D 0.05 0.05 -0.06

Games\_D 0.08 0.02 -0.06

Other\_Sector\_D -0.07 -0.05 0.07

Sectors -0.01 -0.02 -0.02

Crewing\_Up\_Bring\_Own 0.09 0.06 -0.02

Crewing\_Up\_Local -0.05 -0.02 0.00

Crewing\_Up\_Both 0.00 -0.11 0.01

Crewing\_Up\_DK -0.03 0.01 0.17

Crewing\_Up -0.03 -0.11 0.03

[ reached getOption("max.print") -- omitted 22 rows ]

# Preparing to run logistic regression – writing function to test model

> accuracy <- function(predictions, answers){sum((predictions==answers)/ (length(answers)))}

> completedData.1$ratio2[is.na(completedData.1$ratio2)] <- 0.0000000000000001

# first new model F1.D.1(yes skill gaps) as X-var and crewing up dummy vars as predictor y-vars

> mod\_sgs\_crew <- glm(F1\_D.1 ~ Crewing\_Up\_Bring\_Own + Crewing\_Up\_Both + Crewing\_Up\_Local, data = completedData.1, family = binomial(link="logit"), control = list(maxit = 50))

> summary(mod\_sgs\_crew)

Call:

glm(formula = F1\_D.1 ~ Crewing\_Up\_Bring\_Own + Crewing\_Up\_Both +

Crewing\_Up\_Local, family = binomial(link = "logit"), data = completedData.1,

control = list(maxit = 50))

Deviance Residuals:

Min 1Q Median 3Q Max

-1.4823 -0.9118 -0.8164 1.4686 1.5880

Coefficients:

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

(Intercept) -0.9275 0.1623 -5.716 1.09e-08 \*\*\*

Crewing\_Up\_Bring\_Own 1.0817 0.4255 2.542 0.011026 \*

Crewing\_Up\_Both 0.2649 0.2380 1.113 0.265752

Crewing\_Up\_Local 1.6207 0.4905 3.304 0.000953 \*\*\*

---

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

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 490.38 on 380 degrees of freedom

Residual deviance: 474.08 on 377 degrees of freedom

AIC: 482.08

Number of Fisher Scoring iterations: 4

# investigating F1\_D.1 and No training for a reg model

> table(completedData.1$Training\_No, completedData.1$F1\_D.1)

0 1

0 145 86

1 105 45

> chisq.test(completedData.1$F1\_D.1, completedData.1$Training\_No)

Pearson's Chi-squared test with Yates' continuity correction

data: completedData.1$F1\_D.1 and completedData.1$Training\_No

X-squared = 1.7986, df = 1, p-value = 0.1799

> cor(completedData.1$Training\_No, completedData.1$F1\_D.1)

[1] -0.0743619

# calculating pseudo R-squared value

> null.mod.crew <- glm(F1\_D.1 ~ 1, data = completedData.1, family = binomial(link = "logit"))

> 1-logLik(mod\_sgs\_crew)/ logLik(null.mod.crew)

'log Lik.' 0.03324793 (df=4)

# Wrangling data to workout ratio of freelance to payroll by subsectors

> boxplot(completedData$ESS.C1)$out

[1] 400 20000 600 19000 121 120 500 350 530 1000 150 500 400 2500 550

[16] 160 4500 300 1500 100 280 98 145 260 182 200 100 2000 144 31000

[31] 100 110 100 150 600 130 450

> table(ESS$B3)

1 2 3 4 5 6 7 8

158 80 53 22 29 12 33 31

> ratio.film <- subset(completedData, completedData$Sectors == 1)

> ratio.HETV <- subset(completedData, completedData$Sectors == 2)

> ratio.unscripted <- subset(completedData, completedData$Sectors == 3)

> ratio.ChildTV <- subset(completedData, completedData$Sectors == 4)

> ratio.Animation <- subset(completedData, completedData$Sectors == 5)

> ratio.VFX <- subset(completedData, completedData$Sectors == 6)

> ratio.Games <- subset(completedData, completedData$Sectors == 7)

> ratio.other <- subset(completedData, completedData$Sectors == 8)

# subsetted data sets now ready – ratios can now be calculated

# film ratio:

summary(ratio.film$ratio)

Min. 1st Qu. Median Mean 3rd Qu. Max.

0.0000 0.1667 0.6125 3.0157 2.7818 50.0000

> summary(ratio.film$ratio2)

Min. 1st Qu. Median Mean 3rd Qu. Max.

0.0200 0.3595 1.6333 Inf 6.0000 Inf

> summary(ratio.film$ESS.C1)

Min. 1st Qu. Median Mean 3rd Qu. Max.

1.00 6.00 16.50 45.00 37.75 1500.00

> summary(ratio.film$ESS.C3)

Min. 1st Qu. Median Mean 3rd Qu. Max.

0.00 5.00 10.00 34.25 30.00 600.00

> summary(ratio.HETV$ratio)

Min. 1st Qu. Median Mean 3rd Qu. Max.

0.0000 0.4167 3.3807 8.8333 12.7083 62.5000

> summary(ratio.HETV$ratio2)

Min. 1st Qu. Median Mean 3rd Qu. Max.

0.01600 0.07875 0.29750 Inf 2.40000 Inf

> summary(ratio.HETV$ESS.C1)

Min. 1st Qu. Median Mean 3rd Qu. Max.

2.00 8.75 20.00 117.36 32.00 4500.00

> summary(ratio.HETV$ESS.C3)

Min. 1st Qu. Median Mean 3rd Qu. Max.

0 10 40 181 200 2000

> summary(ratio.unscripted$ESS.C1)

Min. 1st Qu. Median Mean 3rd Qu. Max.

1 11 25 1010 60 31000

> summary(ratio.unscripted$ESS.C3)

Min. 1st Qu. Median Mean 3rd Qu. Max.

2 5 25 212 150 2000

> summary(ratio.ChildTV$ESS.C1)

Min. 1st Qu. Median Mean 3rd Qu. Max.

2.00 5.25 11.00 931.73 25.00 20000.00

> summary(ratio.ChildTV$ESS.C3)

Min. 1st Qu. Median Mean 3rd Qu. Max.

2.00 5.25 32.50 148.18 115.00 2000.00

> summary(ratio.Animation$ESS.C1)

Min. 1st Qu. Median Mean 3rd Qu. Max.

2.00 3.00 6.00 20.86 25.00 121.00

> summary(ratio.Animation$ESS.C3)

Min. 1st Qu. Median Mean 3rd Qu. Max.

0.00 5.00 12.00 24.66 30.00 163.00

> summary(ratio.VFX$ESS.C1)

Min. 1st Qu. Median Mean 3rd Qu. Max.

2.0 3.0 28.5 308.0 161.2 2500.0

> summary(ratio.VFX$ESS.C3)

Min. 1st Qu. Median Mean 3rd Qu. Max.

1.0 4.5 17.0 347.8 42.5 2000.0

> summary(ratio.Games$ESS.C1)

Min. 1st Qu. Median Mean 3rd Qu. Max.

2.00 3.00 17.00 65.73 40.00 600.00

> summary(ratio.Games$ESS.C3)

Min. 1st Qu. Median Mean 3rd Qu. Max.

0.00 3.00 6.00 82.45 20.00 2000.00

> summary(ratio.other$ESS.C1)

Min. 1st Qu. Median Mean 3rd Qu. Max.

1.00 2.00 6.00 35.16 39.50 400.00

> summary(ratio.other$ESS.C3)

Min. 1st Qu. Median Mean 3rd Qu. Max.

1.00 6.50 14.00 28.19 30.00 200.00

> ratio.film.PR.FL <- 16.5/10.0

> ratio.film.FL.PR <- 10.0/16.5

> ratio.HETV.FL.PR <- 40/20

> ratio.HETV.PR.FL <- 20/40

> ratio.unscripted.PR.FL <- 25/25

> ratio.unscripted.FL.PR <- 25/25

> ratio.CTV.PR.FL <- 11/32.5

> ratio.CTV.FL.PR <- 32.5/11

> ratio.animation.FL.PR <- 12/6

> ratio.animation.PR.FL <- 6/12

> ratio.VFX.PR.FL <- 28.5/17

> ratio.VFX.FL.PR <- 17/28.5

> ratio.Games.FL.PR <- 17/17

> ratio.other.FL.PR <- 14/6

> ratio.other.PR.FL <- 6/14