Kaitlin's EDA

Kaitlin Maciejewski

12/1/2018

Exploratory

```
library(dplyr)
data <- read.csv("../final_data/frmgham2.csv") %>% janitor::clean_names()
attach(data)

library(psych)
knitr::kable(describe(data)[,c(2,3,4,5,8,9,10,13)], digits = 3)
```

	n	mean	sd	median	min	max	range	se
randid	11627	5004740.917	2900877.440	5006008.00	2448.00	9999312.0	9996864.00	26902.680
sex	11627	1.568	0.495	2.00	1.00	2.0	1.00	0.005
totchol	11218	241.162	45.368	238.00	107.00	696.0	589.00	0.428
age	11627	54.793	9.564	54.00	32.00	81.0	49.00	0.089
sysbp	11627	136.324	22.799	132.00	83.50	295.0	211.50	0.211
diabp	11627	83.038	11.660	82.00	30.00	150.0	120.00	0.108
cursmoke	11627	0.433	0.495	0.00	0.00	1.0	1.00	0.005
cigpday	11548	8.250	12.187	0.00	0.00	90.0	90.00	0.113
bmi	11575	25.877	4.103	25.48	14.43	56.8	42.37	0.038
diabetes	11627	0.046	0.209	0.00	0.00	1.0	1.00	0.002
bpmeds	11034	0.086	0.280	0.00	0.00	1.0	1.00	0.003
heartrte	11621	76.782	12.463	75.00	37.00	220.0	183.00	0.116
glucose	10187	84.125	24.994	80.00	39.00	478.0	439.00	0.248
educ	11332	1.990	1.027	2.00	1.00	4.0	3.00	0.010
prevchd	11627	0.072	0.259	0.00	0.00	1.0	1.00	0.002
prevap	11627	0.054	0.226	0.00	0.00	1.0	1.00	0.002
prevmi	11627	0.032	0.176	0.00	0.00	1.0	1.00	0.002
prevstrk	11627	0.013	0.114	0.00	0.00	1.0	1.00	0.001
prevhyp	11627	0.460	0.498	0.00	0.00	1.0	1.00	0.005
time	11627	1957.019	1758.777	2156.00	0.00	4854.0	4854.00	16.311
period	11627	1.899	0.807	2.00	1.00	3.0	2.00	0.007
hdlc	3027	49.365	15.627	48.00	10.00	189.0	179.00	0.284
ldlc	3026	176.467	46.863	173.00	20.00	565.0	545.00	0.852
death	11627	0.303	0.460	0.00	0.00	1.0	1.00	0.004
angina	11627	0.164	0.370	0.00	0.00	1.0	1.00	0.003
hospmi	11627	0.099	0.299	0.00	0.00	1.0	1.00	0.003
mi_fchd	11627	0.154	0.361	0.00	0.00	1.0	1.00	0.003
anychd	11627	0.272	0.445	0.00	0.00	1.0	1.00	0.004
stroke	11627	0.091	0.288	0.00	0.00	1.0	1.00	0.003
cvd	11627	0.249	0.433	0.00	0.00	1.0	1.00	0.004
hyperten	11627	0.743	0.437	1.00	0.00	1.0	1.00	0.004
timeap	11627	7241.557	2477.780	8766.00	0.00	8766.0	8766.00	22.979
timemi	11627	7593.847	2136.730	8766.00	0.00	8766.0	8766.00	19.816
timemifc	11627	7543.037	2192.120	8766.00	0.00	8766.0	8766.00	20.330
timechd	11627	7008.154	2641.345	8766.00	0.00	8766.0	8766.00	24.496

	n	mean	sd	median	min	max	range	se
timestrk	11627	7660.880	2011.077	8766.00	0.00	8766.0	8766.00	18.651
timecvd	11627	7166.083	2541.668	8766.00	0.00	8766.0	8766.00	23.571
timedth	11627	7854.103	1788.370	8766.00	26.00	8766.0	8740.00	16.585
timehyp	11627	3598.956	3464.165	2429.00	0.00	8766.0	8766.00	32.127

```
table(period) #missing by wave
```

```
## period
##    1    2    3
## 4434 3930 3263
```

```
# library(caret)
# data_2 <- select(data, -c(randid, time, period, timeap:timehyp))
# featurePlot(data_2,as.factor(cursmoke),"box")

# library(GGally)
# ggpairs(data_2)
```

```
# spaghetti
# lorellogram
#

library(ggplot2)

ggplot(data= data, aes(age, cigpday)) +
  geom_point(alpha = .6) +
  geom_smooth(method = 'lm', col = 'red') +
  geom_smooth(method = 'loess') +
  ggtitle("cigpday vs Age Scatterplot")

ggplot(data=data, aes(age, cigpday, group = randid, color = as.factor(sex))) +
  geom_path(alpha = .6) +
  geom_smooth(aes(group = NULL), method = 'lm', col = 'red') +
  geom_smooth(aes(group = NULL), method = 'loess') +
  ggtitle("cigpday vs Age Spaghetti Plot")

ggplot(data=data, aes(as.factor(period), cigpday, group = randid, color = as.factor(sex))) +
  geom_path(alpha = .6) +
  geom_smooth(aes(group = NULL), method = 'lm', col = 'red') +
  geom_smooth(aes(group = NULL), method = 'loess') +
  ggtitle("cigpday vs time Spaghetti Plot")

ggplot(data, aes(y = cursmoke, x = age)) + geom_jitter(height = 0.1) +
stat_summary(fun.y = 'mean', geom="line", col = 'red')

select(data, -c(randid, time,timeap:timehyp)) %>% GGally::ggcorr(.)


corr <- data[,c(-1,-21)] %>% cor(., use = "complete.obs")

library(corrplot)
corrplot(corr, type = "upper", order = "hclust",
         tl.col = "black", tl.srt = 75, tl.offset = 1, tl.cex = .8, method = "ellipse")
```

```r
corr2 <- data[,c(2,3,4,5,6,7,8,10,11,12,13,14)] %>% cor(., use = "complete.obs")
corrplot(corr2, type = "upper", order = "hclust",
         tl.col = "black", tl.srt = 45)


ggplot(data = data,
aes(x = as.factor(cursmoke), y = age)) +
geom_violin() +
  facet_wrap(~sex)

ggplot(data = data,
aes(x = cigpday, y = age)) +
geom_point() +
  facet_wrap(~sex) +
  geom_smooth()


resp.long_2 <- data
theme_set(theme_bw(base_size = 10))
```

Models GLMER

(1) Is there a relationship between age and smoking status? Does this relationship differ by sex?

```r
# possible confounders of smoking status...
# education
# disease status

library(lme4)

smoke_stat_age <- glmer(cursmoke~age+as.factor(educ) + (1|randid), family=binomial, na.action = "na.omi

knitr::kable(summary(smoke_stat_age)$coefficients,digits = 3)
```

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	9.438	0.486	19.405	0.000
age	-0.195	0.009	-22.054	0.000
as.factor(educ)2	0.552	0.265	2.081	0.037
as.factor(educ)3	-0.783	0.304	-2.575	0.010
as.factor(educ)4	-0.253	0.347	-0.729	0.466

```r
smoke_stat_sex <- glmer(cursmoke~age+as.factor(educ) + as.factor(sex) + (1|randid), family=binomial, na

knitr::kable(summary(smoke_stat_sex)$coefficients,digits = 3)
```

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	11.842	0.586	20.214	0.000
age	-0.205	0.009	-22.325	0.000
as.factor(educ)2	0.882	0.287	3.070	0.002
as.factor(educ)3	-0.384	0.328	-1.171	0.241

	Estimate	Std. Error	z value	Pr(>|z|)
as.factor(educ)4	-0.699	0.379	-1.848	0.065
as.factor(sex)2	-3.822	0.336	-11.368	0.000

(2) Is there a relationship between the number of cigarettes smoked per day and age? Does this relationship differ by sex?

```
cigpday_age <- glmer(cigpday~age+as.factor(educ) +(1|randid), family=poisson, na.action = "na.omit")

knitr::kable(summary(cigpday_age)$coefficients, digits = 3)
```

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.132	0.111	1.187	0.235
age	-0.018	0.001	-25.495	0.000
as.factor(educ)2	0.517	0.142	3.628	0.000
as.factor(educ)3	-0.232	0.176	-1.319	0.187
as.factor(educ)4	0.242	0.197	1.224	0.221

```
ncig_gee <- glmer(cigpday~age+as.factor(sex) + as.factor(educ)+ (1|randid), family=poisson, na.action =

knitr::kable(summary(ncig_gee)$coefficients,digits = 3)
```

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.202	0.118	10.229	0.000
age	-0.018	0.001	-25.527	0.000
as.factor(sex)2	-2.032	0.119	-17.024	0.000
as.factor(educ)2	0.673	0.138	4.881	0.000
as.factor(educ)3	0.088	0.170	0.517	0.605
as.factor(educ)4	0.027	0.190	0.143	0.887

(1) The relationship between current smoking status and systolic blood pressure.

```
smoke_sys<-glmer(cursmoke~sysbp+ as.factor(educ) + (1|randid), family=binomial, na.action = "na.omit")

knitr::kable(summary(smoke_sys)$coefficients,digits = 3)
```

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	3.552	0.394	9.022	0.000
sysbp	-0.035	0.003	-12.949	0.000
as.factor(educ)2	0.999	0.223	4.483	0.000
as.factor(educ)3	-0.391	0.263	-1.488	0.137
as.factor(educ)4	0.045	0.297	0.152	0.879

```
smoke_sys<-glmer(cursmoke~sysbp + as.factor(sex) + as.factor(educ)+ (1|randid), family=binomial, na.acti

knitr::kable(summary(smoke_sys)$coefficients,digits = 3)
```

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	5.097	0.448	11.381	0.000

	Estimate	Std. Error	z value	Pr(>|z|)
sysbp	-0.036	0.003	-12.453	0.000
as.factor(sex)2	-3.172	0.289	-10.992	0.000
as.factor(educ)2	1.389	0.258	5.387	0.000
as.factor(educ)3	-0.011	0.292	-0.039	0.969
as.factor(educ)4	-0.270	0.328	-0.824	0.410

(2) The relationship between current smoking status and diastolic blood pressure.

```
smoke_dias<-glmer(cursmoke~diabp +as.factor(educ)+ (1|randid), family=binomial,  na.action = "na.omit")

knitr::kable(summary(smoke_dias)$coefficients,digits = 3)
```

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.426	0.412	1.035	0.301
diabp	-0.020	0.005	-4.414	0.000
as.factor(educ)2	1.109	0.217	5.107	0.000
as.factor(educ)3	-0.230	0.257	-0.894	0.371
as.factor(educ)4	0.251	0.290	0.867	0.386

```
smoke_dias<-glmer(cursmoke~diabp + as.factor(sex) + as.factor(educ)+ (1|randid), family=binomial,  na.a

knitr::kable(summary(smoke_dias)$coefficients,digits = 3)
```

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	2.059	0.450	4.576	0.000
diabp	-0.023	0.005	-4.726	0.000
as.factor(sex)2	-3.172	0.286	-11.077	0.000
as.factor(educ)2	1.520	0.255	5.966	0.000
as.factor(educ)3	0.180	0.285	0.633	0.527
as.factor(educ)4	-0.049	0.319	-0.154	0.877

(3) The relationship between current smoking status and serum total cholesterol.

```
smoke_chol<-glmer(cursmoke~totchol + (1|randid) + as.factor(educ), family=binomial, na.action = "na.omi

knitr::kable(summary(smoke_chol)$coefficients,digits = 3)
```

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.819	0.347	-2.362	0.018
totchol	-0.002	0.001	-1.386	0.166
as.factor(educ)2	1.094	0.216	5.052	0.000
as.factor(educ)3	-0.225	0.256	-0.877	0.381
as.factor(educ)4	0.274	0.288	0.951	0.341

```
smoke_chol<-glmer(cursmoke~totchol + as.factor(sex) + as.factor(educ)+ (1|randid), family=binomial,  na

knitr::kable(summary(smoke_chol)$coefficients,digits = 3)
```

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.161	0.377	0.429	0.668
totchol	0.000	0.001	0.000	1.000
as.factor(sex)2	-3.240	0.305	-10.613	0.000
as.factor(educ)2	1.534	0.260	5.896	0.000
as.factor(educ)3	0.187	0.288	0.649	0.516
as.factor(educ)4	-0.009	0.323	-0.027	0.978

```
###

my.data.complete <- data %>%
  dplyr::select(-c(hdlc,ldlc)) %>%
  na.omit()

model.saturated <- geepack::geeglm(formula = cursmoke ~ as.factor(sex) +
    age + age * as.factor(sex) + sysbp + diabp + sysbp * diabp +
    bpmeds + as.factor(educ) + totchol + bmi + glucose + diabetes +
    heartrte + prevap + prevchd + prevmi + prevstrk + prevhyp,
    family = binomial, data = my.data.complete, id = randid,
    corstr = ("unstructured"))

model <- aov(model.saturated)
car::vif(model)
```

```
##                          GVIF Df GVIF^(1/(2*Df))
## as.factor(sex)     35.783769  1        5.981954
## age                 2.661847  1        1.631517
## sysbp              35.429075  1        5.952233
## diabp              25.680738  1        5.067617
## bpmeds              1.194381  1        1.092878
## as.factor(educ)     1.108310  3        1.017287
## totchol             1.107321  1        1.052293
## bmi                 1.183872  1        1.088059
## glucose             1.414456  1        1.189309
## diabetes            1.392703  1        1.180128
## heartrte            1.081176  1        1.039796
## prevap              4.787315  1        2.187993
## prevchd             7.474492  1        2.733952
## prevmi              2.402776  1        1.550089
## prevstrk            1.027812  1        1.013811
## prevhyp             2.071912  1        1.439414
## as.factor(sex):age 37.891150  1        6.155579
## sysbp:diabp        91.955361  1        9.589336
```

```
model<-glmer(cursmoke~age + as.factor(educ) + hyperten + totchol + as.factor(sex) + (1|randid), family=

knitr::kable(summary(model)$coefficients,digits = 3)
```

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	12.419	0.722	17.208	0.000
age	-0.201	0.009	-21.561	0.000
as.factor(educ)2	0.464	0.277	1.672	0.095
as.factor(educ)3	-0.671	0.327	-2.048	0.041

	Estimate	Std. Error	z value	Pr(>|z|)
as.factor(educ)4	-1.217	0.378	-3.219	0.001
hyperten	-2.349	0.310	-7.567	0.000
totchol	0.005	0.002	3.125	0.002
as.factor(sex)2	-3.831	0.317	-12.090	0.000