Linear Regression CpG & Obesity

Guannan Shen

October 2, 2018

Contents


```
## set up workspace
library(knitr)
library(tidyverse)
options(stringsAsFactors = F)
options(dplyr.width = Inf)
getwd()
```

```
## [1] "/home/guanshim/Documents/gitlab/ECCHO_github/Code"
```

1 ECCHO_Guannan

ECCHO grant (chemicals and methylation project of the Healthy Start cohort).

1.1 9/27/2018

linear regression model for 9 outcomes on obesity and CpG
$ y = CpG + maternal age + race + CellTypes(7) (All Male)$
The linear models, each has 10 covariates in the
Outcomes: "birth_weight", "ipv3_pp_fm_pct", "Chol_IPV3", "FFA_IPV3", "Gluc_IPV3", "HDL_IPV3", "Insu_IPV3", "Trig_IPV3", "Leptin_actual_*ng_ml*".
Although we discussed include the variable "infant sex" as a covariate, we actually did the chemicals-methylation analysis stratified by sex. So the DMRs were only detected among male offspring. Therefore we should probably restrict our analysis to individuals with infant_sex = 2 (males), at least to begin with.
There will be missing data for some outcomes, I think it is OK if the sample size differs for each model.

1.2 Methods

1.2.1 Data Pre-processing

The dataset to model the association between obesity and CpG M-value was generated from three dataset: cordblood cell counts (celltype), obesity related outcomes (pfas) and top300 CpG list (m300). The final dataset was the inner join product (the intersection) of these 3 datasets and was joint by "pid" (participants ID). The "celltype" and "m300" datasets share the same collection of participants with sample size 600. The outcome data set contains 589 subjects, and 588 out of 589 subjects are shared across all 3 dataset. The subject with pid 30568 is only in the "pfas" dataset.
In this way, the final dataset obs_cpg contains 588 rows (subjects) and 320 columns. After the dataset was filtered by "infant_sex == 2 (males)", 308 participants left.

Regarding the male only dataset, "cpg_male", there are 308 rows and 320 columns. Missing values only occur in outcomes, thus there are different sample sizes for different outcomes. The first 2 columns, pid and infant_sex would not be included in the linear regression model. The 3rd and 4th columns are "maternal_age" and categorical variable "race_4", which would always be included in the model. From column 5 to 13, these are 9 outcomes. From column 14 to 20, these are 7 celltypes, "Bcell", "CD4T", "CD8T", "Gran", "Mono", "NK" and "nRBC", which would always be included in the model. From column 21 to 320, these are 300 CpG M values. Each linear regression model would contain 1 out of 300 CpGs.

```
# readin data in ASUS largest dataset is 3Mb celltype <-
# read_csv('C:/Users/hithr/Documents/Stats/gitlab/ECCHO_Guannan/DataProcessed/healthy_start_cordblood_c
# m300 <-
# read_csv('C:/Users/hithr/Documents/Stats/gitlab/ECCHO_Guannan/DataRaw/HS_450K_CB_Mval_normbatch_Starl
# pfas <-
# read_csv('C:/Users/hithr/Documents/Stats/gitlab/ECCHO_Guannan/DataRaw/pfas_methyl_di.csv')

## readin data in Ubuntu
celltype <- read_csv("~/Documents/gitlab/ECCHO_github/DataProcessed/healthy_start_cordblood_cellcounts_

m300 <- read_csv("~/Documents/gitlab/ECCHO_github/DataRaw/HS_450K_CB_Mval_normbatch_StarlingSubset_10-0
pfas <- read_csv("~/Documents/gitlab/ECCHO_github/DataRaw/pfas_methyl_di.csv")

# Exploratory Data Analysis
dim(celltype)
```

```
## [1] 600   8
```

```
colnames(celltype)
```

```
## [1] "pid"   "Bcell" "CD4T"  "CD8T"  "Gran"  "Mono"  "NK"    "nRBC"
```

```
##
dim(m300)
```

```
## [1] 600 301
```

```
##
summary(pfas)
```

```
##       pid         birth_weight    infant_sex     maternal_age
##  Min.   :10002   Min.   :1750   Min.   :1.000   Min.   :16.00
##  1st Qu.:10525   1st Qu.:2990   1st Qu.:1.000   1st Qu.:22.00
##  Median :10943   Median :3250   Median :2.000   Median :28.00
##  Mean   :11583   Mean   :3269   Mean   :1.525   Mean   :27.59
##  3rd Qu.:11315   3rd Qu.:3550   3rd Qu.:2.000   3rd Qu.:32.00
##  Max.   :31069   Max.   :4635   Max.   :2.000   Max.   :44.00
##
##      race_4      ipv3_pp_fm_pct     Chol_IPV3         FFA_IPV3
##  Min.   :1.000   Min.   : 1.200   Min.   : 26.00   Min.   : 34.0
##  1st Qu.:2.000   1st Qu.: 6.350   1st Qu.: 47.00   1st Qu.:174.0
##  Median :2.000   Median : 8.800   Median : 56.00   Median :251.0
##  Mean   :2.039   Mean   : 8.953   Mean   : 58.62   Mean   :280.7
##  3rd Qu.:2.000   3rd Qu.:11.600   3rd Qu.: 67.00   3rd Qu.:361.5
##  Max.   :4.000   Max.   :24.850   Max.   :280.00   Max.   :961.0
##                  NA's   :20       NA's   :39       NA's   :82
##    Gluc_IPV3         HDL_IPV3       Insu_IPV3         Trig_IPV3
##  Min.   : 34.00   Min.   : 9.00   Min.   :  2.000   Min.   : 11.00
##  1st Qu.: 70.75   1st Qu.:21.00   1st Qu.:  5.000   1st Qu.: 30.00
```

```
##  Median : 80.00   Median :25.00   Median :  7.000   Median : 39.00
##  Mean   : 82.93   Mean   :25.86   Mean   :  9.306   Mean   : 46.74
##  3rd Qu.: 91.25   3rd Qu.:29.00   3rd Qu.: 11.000   3rd Qu.: 55.00
##  Max.   :253.00   Max.   :98.00   Max.   :115.000   Max.   :649.00
##  NA's   :25       NA's   :81      NA's   :46        NA's   :47
##  Leptin_actual__ng_ml_
##  Min.   :  0.325
##  1st Qu.:  5.904
##  Median : 10.879
##  Mean   : 16.404
##  3rd Qu.: 22.070
##  Max.   :129.676
##  NA's   :106
```

```
dim(pfas)
```

```
## [1] 589  13
```

```
outcomes <- colnames(pfas)[-c(1, 3:5)]
paste(outcomes)
```

```
## [1] "birth_weight"          "ipv3_pp_fm_pct"        "Chol_IPV3"
## [4] "FFA_IPV3"              "Gluc_IPV3"             "HDL_IPV3"
## [7] "Insu_IPV3"             "Trig_IPV3"             "Leptin_actual__ng_ml_"
```

```
# compare m300 and celltype, the subjects
m300$pid[!(m300$pid == celltype$pid)]
```

```
## integer(0)
```

```
sum(m300$pid %in% celltype$pid)
```

```
## [1] 600
```

```
## merge dataset by ID, inner join
pfas_cell <- merge(pfas, celltype, by = "pid")
obs_cpg <- merge(pfas_cell, m300, by = "pid")
dim(obs_cpg)
```

```
## [1] 588 320
```

```
# check final sujects with outcomes
pfas$pid[!(obs_cpg$pid == pfas$pid)]
```

```
## [1] 30568 30604 30635 30645 30664 30671 30756 30808 31069
```

```
#
pfas$pid[!(pfas$pid %in% m300$pid)]
```

```
## [1] 30568
```

```
message("The subject with PID 30568 is in Outcomes but not in celltype and top 300 CpG")


# filter based on gender, make race_4 as catrgorical
# variable.
cpg_male <- obs_cpg %>% filter(infant_sex == 2) %>% mutate(race_4 = as.factor(race_4)) %>%
    select(pid, infant_sex, maternal_age, race_4, everything())

dim(cpg_male)
```

```
## [1] 308 320
```

```
############### missing data ####################
kable(apply(cpg_male[, 5:13], 2, function(x) {
    sum(is.na(x))
}), caption = "Missing Data Summary", col.names = "# Missing Values")
```

Table 1: Missing Data Summary

	# Missing Values
birth_weight	0
ipv3_pp_fm_pct	13
Chol_IPV3	18
FFA_IPV3	40
Gluc_IPV3	10
HDL_IPV3	45
Insu_IPV3	23
Trig_IPV3	21
Leptin_actual_*ng_ml*	54

```
################################################## to build up the lm model, get the variables' name

## 'maternal_age', 'race_4'
colnames(cpg_male)[3:4]
```

```
## [1] "maternal_age" "race_4"
```

```
## outcomes 'birth_weight', 'ipv3_pp_fm_pct', 'Chol_IPV3',
## 'FFA_IPV3', 'Gluc_IPV3', 'HDL_IPV3', 'Insu_IPV3'
## 'Trig_IPV3', 'Leptin_actual__ng_ml_'
colnames(cpg_male)[5:13]
```

```
## [1] "birth_weight"          "ipv3_pp_fm_pct"        "Chol_IPV3"
## [4] "FFA_IPV3"              "Gluc_IPV3"             "HDL_IPV3"
## [7] "Insu_IPV3"             "Trig_IPV3"             "Leptin_actual__ng_ml_"
```

```
## cell types names 'Bcell', 'CD4T', 'CD8T', 'Gran', 'Mono',
## 'NK' and 'nRBC'
colnames(cpg_male)[14:20]
```

```
## [1] "Bcell" "CD4T"  "CD8T"  "Gran"  "Mono"  "NK"    "nRBC"
```

```
message("loop over 21 to 320")
```

Table 2: Top10 FDR birth_weight n = 308

	names	Estimate	Std.Error	t.statistic	p.value	FDR	FWER
49	cg16725984	-223.9782	62.8628	-3.5630	0.0004	0.1200000	0.12
67	cg25195288	531.2117	180.0618	2.9502	0.0034	0.4725000	1.00
167	cg16495448	-321.1247	116.1410	-2.7650	0.0061	0.4725000	1.00
184	cg25137968	338.3636	123.0875	2.7490	0.0063	0.4725000	1.00
204	cg15045292	158.0504	59.4286	2.6595	0.0083	0.4980000	1.00
22	cg00784263	319.4127	131.6031	2.4271	0.0158	0.5925000	1.00
71	cg16672637	646.5242	263.2745	2.4557	0.0146	0.5925000	1.00
83	cg20741567	505.6544	207.0949	2.4417	0.0152	0.5925000	1.00

	names	Estimate	Std.Error	t.statistic	p.value	FDR	FWER
57	cg23206463	-116.4047	55.4694	-2.0985	0.0367	0.6804545	1.00
79	cg23629795	172.6436	85.8507	2.0110	0.0452	0.6804545	1.00

Table 3: Top10 FDR ipv3_pp_fm_pct n = 295

	names	Estimate	Std.Error	t.statistic	p.value	FDR	FWER
112	cg24366087	-2.8077	1.0239	-2.7421	0.0065	0.8400000	1
139	cg08743751	2.3959	0.9020	2.6561	0.0084	0.8400000	1
190	cg10832304	-1.3755	0.4869	-2.8251	0.0051	0.8400000	1
4	cg21853587	3.1614	1.5731	2.0097	0.0454	0.8626829	1
7	cg27354586	-2.1132	1.2972	-1.6290	0.1044	0.8626829	1
22	cg00784263	1.9467	1.1015	1.7673	0.0783	0.8626829	1
23	cg22305268	-2.7010	1.5456	-1.7475	0.0816	0.8626829	1
35	cg16251579	-1.0009	0.6074	-1.6479	0.1005	0.8626829	1
39	cg16472896	-2.9514	1.8646	-1.5829	0.1146	0.8626829	1
41	cg04772025	0.9450	0.5649	1.6729	0.0955	0.8626829	1

Table 4: Top10 FDR Chol_IPV3 n = 290

	names	Estimate	Std.Error	t.statistic	p.value	FDR	FWER
95	cg17850055	-26.7278	9.9250	-2.6930	0.0075	0.5040000	1.00
112	cg24366087	-12.9362	4.7274	-2.7364	0.0066	0.5040000	1.00
254	cg22692511	7.2223	2.3323	3.0966	0.0022	0.5040000	0.66
266	cg12857407	10.9809	4.1342	2.6561	0.0084	0.5040000	1.00
271	cg08162803	15.1963	5.2370	2.9017	0.0040	0.5040000	1.00
49	cg16725984	5.9979	2.3273	2.5772	0.0105	0.5250000	1.00
279	cg17132124	8.4670	3.5930	2.3565	0.0191	0.8185714	1.00
28	cg12872489	-6.5322	2.9982	-2.1787	0.0302	0.8584615	1.00
58	cg09887862	4.7067	2.2466	2.0950	0.0371	0.8584615	1.00
69	cg04168590	20.7551	9.0932	2.2825	0.0232	0.8584615	1.00

Table 5: Top10 FDR FFA_IPV3 n = 268

	names	Estimate	Std.Error	t.statistic	p.value	FDR	FWER
4	cg21853587	-169.3376	63.0974	-2.6837	0.0078	0.7016667	1
7	cg27354586	-110.6213	51.8697	-2.1327	0.0339	0.7016667	1
9	cg20510724	172.4090	72.6345	2.3737	0.0184	0.7016667	1
14	cg09473264	73.8047	35.9716	2.0517	0.0412	0.7016667	1
28	cg12872489	-54.8344	26.8486	-2.0424	0.0421	0.7016667	1
54	cg19529074	-97.4207	44.1996	-2.2041	0.0284	0.7016667	1
96	cg21215576	82.6143	30.8336	2.6794	0.0079	0.7016667	1
119	cg00438284	-77.5458	36.8574	-2.1039	0.0364	0.7016667	1
126	cg05390685	-69.3031	31.9189	-2.1712	0.0308	0.7016667	1
148	cg13598480	98.5534	41.1924	2.3925	0.0175	0.7016667	1

Table 6: Top10 FDR Gluc_IPV3 n = 298

	names	Estimate	Std.Error	t.statistic	p.value	FDR	FWER
145	cg06404838	27.8481	8.4948	3.2783	0.0012	0.360	0.36
248	cg11196848	-15.3248	5.5395	-2.7665	0.0060	0.900	1.00
16	cg06873590	-32.3360	14.4880	-2.2319	0.0264	0.948	1.00
27	cg17519749	11.7052	4.5215	2.5888	0.0101	0.948	1.00
59	cg20324199	11.6527	5.0313	2.3160	0.0213	0.948	1.00
77	cg23478547	7.9711	3.6905	2.1599	0.0316	0.948	1.00
135	cg17171260	-15.1541	6.8867	-2.2005	0.0286	0.948	1.00
150	cg14163408	11.2035	4.8768	2.2973	0.0223	0.948	1.00
217	cg01816336	-18.4503	8.3155	-2.2188	0.0273	0.948	1.00
287	cg26781129	11.1141	5.1032	2.1779	0.0302	0.948	1.00

Table 7: Top10 FDR HDL_IPV3 n = 263

	names	Estimate	Std.Error	t.statistic	p.value	FDR	FWER
42	cg15355952	-6.1054	2.0229	-3.0181	0.0028	0.4000	0.84
49	cg16725984	3.2542	1.0221	3.1839	0.0016	0.4000	0.48
236	cg04061372	1.9260	0.6637	2.9022	0.0040	0.4000	1.00
271	cg08162803	6.2363	2.3198	2.6883	0.0077	0.5775	1.00
26	cg03452190	6.5512	2.8815	2.2736	0.0238	0.7380	1.00
145	cg06404838	-6.5192	2.8836	-2.2608	0.0246	0.7380	1.00
211	cg00893875	1.7387	0.7042	2.4690	0.0142	0.7380	1.00
281	cg22946159	-7.9944	3.3218	-2.4067	0.0168	0.7380	1.00
286	cg03989507	4.3550	1.8841	2.3115	0.0216	0.7380	1.00
290	cg00798281	-3.7706	1.5943	-2.3651	0.0188	0.7380	1.00

Table 8: Top10 FDR Insu_IPV3 n = 285

	names	Estimate	Std.Error	t.statistic	p.value	FDR	FWER
169	cg17501712	9.6030	2.9405	3.2658	0.0012	0.3600000	0.36
1	cg02779535	5.4132	5.8329	0.9281	0.3542	0.9769751	1.00
3	cg07551200	9.7645	5.8388	1.6723	0.0956	0.9769751	1.00
4	cg21853587	-1.7637	4.3767	-0.4030	0.6873	0.9769751	1.00
5	cg12657739	-0.9770	2.8429	-0.3436	0.7314	0.9769751	1.00
6	cg26724375	1.1364	2.8714	0.3958	0.6926	0.9769751	1.00
8	cg00637826	-0.9823	2.4423	-0.4022	0.6879	0.9769751	1.00
9	cg20510724	2.9743	4.9811	0.5971	0.5509	0.9769751	1.00
10	cg15977816	0.7458	2.0905	0.3568	0.7216	0.9769751	1.00
11	cg02233835	2.3929	3.0380	0.7876	0.4316	0.9769751	1.00

Table 9: Top10 FDR Trig_IPV3 n = 287

	names	Estimate	Std.Error	t.statistic	p.value	FDR	FWER
291	cg09630142	-28.1212	8.7979	-3.1963	0.0016	0.4800000	0.48
1	cg02779535	13.3008	25.5877	0.5198	0.6036	0.9521495	1.00
5	cg12657739	10.4746	12.4564	0.8409	0.4011	0.9521495	1.00
6	cg26724375	-14.5740	12.4430	-1.1713	0.2425	0.9521495	1.00

	names	Estimate	Std.Error	t.statistic	p.value	FDR	FWER
7	cg27354586	14.8697	15.8276	0.9395	0.3483	0.9521495	1.00
8	cg00637826	-4.7679	10.6850	-0.4462	0.6558	0.9521495	1.00
9	cg20510724	10.5021	21.1497	0.4966	0.6199	0.9521495	1.00
10	cg15977816	-4.7967	9.0155	-0.5321	0.5951	0.9521495	1.00
12	cg21870229	-23.0559	16.4747	-1.3995	0.1628	0.9521495	1.00
14	cg09473264	11.8505	10.3567	1.1442	0.2535	0.9521495	1.00

Table 10: Top10 FDR Leptin_actual___ng_ml n = 254

	names	Estimate	Std.Error	t.statistic	p.value	FDR	FWER
22	cg00784263	12.8922	4.0607	3.1749	0.0017	0.2550000	0.51
49	cg16725984	-6.6381	1.9729	-3.3647	0.0009	0.2550000	0.27
19	cg00128386	-16.7190	6.2751	-2.6643	0.0082	0.5233333	1.00
85	cg23572459	-15.7889	6.4071	-2.4643	0.0144	0.5233333	1.00
104	cg10119082	-5.3893	2.2145	-2.4336	0.0157	0.5233333	1.00
116	cg21183455	5.2322	2.1486	2.4351	0.0156	0.5233333	1.00
134	cg05906144	7.7035	2.7746	2.7765	0.0059	0.5233333	1.00
135	cg17171260	-10.7783	4.4176	-2.4399	0.0154	0.5233333	1.00
209	cg24280832	9.1099	3.5388	2.5742	0.0106	0.5233333	1.00
260	cg17284440	-18.5351	7.9487	-2.3319	0.0205	0.6150000	1.00

```
##
## Call:
## lm(formula = Leptin_actual__ng_ml_ ~ cpg_male[, 42] + maternal_age +
##     race_4 + Bcell + CD4T + CD8T + Gran + Mono + NK + nRBC, data = cpg_male)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -19.312  -7.439  -2.423   4.076  98.677
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     204.3828    77.9850   2.621  0.00933 **
## cpg_male[, 42]   12.8922     4.0607   3.175  0.00169 **
## maternal_age      0.2367     0.1371   1.726  0.08555 .
## race_42          -0.7539     1.9590  -0.385  0.70068
## race_43           0.3732     2.5227   0.148  0.88253
## race_44          -2.3800     3.8262  -0.622  0.53451
## Bcell          -129.9695    77.1407  -1.685  0.09331 .
## CD4T           -180.5702    74.5936  -2.421  0.01623 *
## CD8T           -166.9154    78.6523  -2.122  0.03484 *
## Gran           -160.5740    75.1731  -2.136  0.03368 *
## Mono           -169.4660    77.7576  -2.179  0.03027 *
## NK             -171.3660    77.1968  -2.220  0.02736 *
## nRBC           -128.8336    75.2295  -1.713  0.08808 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.97 on 241 degrees of freedom
##   (54 observations deleted due to missingness)
```

```
## Multiple R-squared:  0.1361, Adjusted R-squared:  0.09306
## F-statistic: 3.163 on 12 and 241 DF,  p-value: 0.0003275
```

```
################## loop for birth_weight ###################################
birth_weight_lm <- lapply(21:320, function(i) {
    lm = lm(birth_weight ~ cpg_male[, i] + maternal_age + race_4 +
        Bcell + CD4T + CD8T + Gran + Mono + NK + nRBC, data = cpg_male)
    coef = round(summary(lm)$coefficients[2, ], 4)
    return(coef)
})
birth_weight_lm <- data.frame(matrix(unlist(birth_weight_lm),
    ncol = 4, byrow = TRUE, dimnames = list(c(colnames(cpg_male)[21:320]),
        c("Estimate", "Std.Error", "t.statistic", "p.value"))))

# adjusted p-value
birth_weight_lm <- birth_weight_lm %>% mutate(FDR = p.adjust(p.value,
    "BH", 300), FWER = p.adjust(p.value, "bonferroni", 300),
    names = colnames(cpg_male)[21:320]) %>% select(names, everything())
birth_weight_lm <- birth_weight_lm[order(birth_weight_lm$FDR),
    ]

## sample size
size1 <- length(cpg_male$birth_weight) - sum(is.na(cpg_male$birth_weight))

## summary table
kable(head(birth_weight_lm, 10), caption = paste("Top10 FDR birth_weight ",
    "n = ", size1, sep = ""))
```

```
####################### ipv3_pp_fm_pct ####################
ipv3_pp_fm_pct_lm <- lapply(21:320, function(i) {
    lm = lm(ipv3_pp_fm_pct ~ cpg_male[, i] + maternal_age + race_4 +
        Bcell + CD4T + CD8T + Gran + Mono + NK + nRBC, data = cpg_male)
    coef = round(summary(lm)$coefficients[2, ], 4)
    return(coef)
})
ipv3_pp_fm_pct_lm <- data.frame(matrix(unlist(ipv3_pp_fm_pct_lm),
    ncol = 4, byrow = TRUE, dimnames = list(c(colnames(cpg_male)[21:320]),
        c("Estimate", "Std.Error", "t.statistic", "p.value"))))

# adjusted p-value
ipv3_pp_fm_pct_lm <- ipv3_pp_fm_pct_lm %>% mutate(FDR = p.adjust(p.value,
    "BH", 300), FWER = p.adjust(p.value, "bonferroni", 300),
    names = colnames(cpg_male)[21:320]) %>% select(names, everything())

ipv3_pp_fm_pct_lm <- ipv3_pp_fm_pct_lm[order(ipv3_pp_fm_pct_lm$FDR),
    ]

## sample size
size2 <- length(cpg_male$ipv3_pp_fm_pct) - sum(is.na(cpg_male$ipv3_pp_fm_pct))

## summary table
```

```
kable(head(ipv3_pp_fm_pct_lm, 10), caption = paste("Top10 FDR ipv3_pp_fm_pct ",
    "n = ", size2, sep = ""))

########################## Chol_IPV3 ##############################
Chol_IPV3_lm <- lapply(21:320, function(i) {
    lm = lm(Chol_IPV3 ~ cpg_male[, i] + maternal_age + race_4 +
        Bcell + CD4T + CD8T + Gran + Mono + NK + nRBC, data = cpg_male)
    coef = round(summary(lm)$coefficients[2, ], 4)
    return(coef)
})
Chol_IPV3_lm <- data.frame(matrix(unlist(Chol_IPV3_lm), ncol = 4,
    byrow = TRUE, dimnames = list(c(colnames(cpg_male)[21:320]),
        c("Estimate", "Std.Error", "t.statistic", "p.value"))))

# adjusted p-value
Chol_IPV3_lm <- Chol_IPV3_lm %>% mutate(FDR = p.adjust(p.value,
    "BH", 300), FWER = p.adjust(p.value, "bonferroni", 300),
    names = colnames(cpg_male)[21:320]) %>% select(names, everything())
Chol_IPV3_lm <- Chol_IPV3_lm[order(Chol_IPV3_lm$FDR), ]


## sample size
size3 <- length(cpg_male$Chol_IPV3) - sum(is.na(cpg_male$Chol_IPV3))

## summary table
kable(head(Chol_IPV3_lm, 10), caption = paste("Top10 FDR Chol_IPV3 ",
    "n = ", size3, sep = ""))

######################## FFA_IPV3 ###################################
FFA_IPV3_lm <- lapply(21:320, function(i) {
    lm = lm(FFA_IPV3 ~ cpg_male[, i] + maternal_age + race_4 +
        Bcell + CD4T + CD8T + Gran + Mono + NK + nRBC, data = cpg_male)
    coef = round(summary(lm)$coefficients[2, ], 4)
    return(coef)
})
FFA_IPV3_lm <- data.frame(matrix(unlist(FFA_IPV3_lm), ncol = 4,
    byrow = TRUE, dimnames = list(c(colnames(cpg_male)[21:320]),
        c("Estimate", "Std.Error", "t.statistic", "p.value"))))

# adjusted p-value
FFA_IPV3_lm <- FFA_IPV3_lm %>% mutate(FDR = p.adjust(p.value,
    "BH", 300), FWER = p.adjust(p.value, "bonferroni", 300),
    names = colnames(cpg_male)[21:320]) %>% select(names, everything())
FFA_IPV3_lm <- FFA_IPV3_lm[order(FFA_IPV3_lm$FDR), ]


## sample size
size4 <- length(cpg_male$FFA_IPV3) - sum(is.na(cpg_male$FFA_IPV3))

## summary table
kable(head(FFA_IPV3_lm, 10), caption = paste("Top10 FDR FFA_IPV3 ",
    "n = ", size4, sep = ""))
```

```
######################## Gluc_IPV3 ############################
Gluc_IPV3_lm <- lapply(21:320, function(i) {
    lm = lm(Gluc_IPV3 ~ cpg_male[, i] + maternal_age + race_4 +
        Bcell + CD4T + CD8T + Gran + Mono + NK + nRBC, data = cpg_male)
    coef = round(summary(lm)$coefficients[2, ], 4)
    return(coef)
})
Gluc_IPV3_lm <- data.frame(matrix(unlist(Gluc_IPV3_lm), ncol = 4,
    byrow = TRUE, dimnames = list(c(colnames(cpg_male)[21:320]),
        c("Estimate", "Std.Error", "t.statistic", "p.value"))))

# adjusted p-value
Gluc_IPV3_lm <- Gluc_IPV3_lm %>% mutate(FDR = p.adjust(p.value,
    "BH", 300), FWER = p.adjust(p.value, "bonferroni", 300),
    names = colnames(cpg_male)[21:320]) %>% select(names, everything())
Gluc_IPV3_lm <- Gluc_IPV3_lm[order(Gluc_IPV3_lm$FDR), ]

## sample size
size5 <- length(cpg_male$Gluc_IPV3) - sum(is.na(cpg_male$Gluc_IPV3))

## summary table
kable(head(Gluc_IPV3_lm, 10), caption = paste("Top10 FDR Gluc_IPV3 ",
    "n = ", size5, sep = ""))


######################## HDL_IPV3 ######################
HDL_IPV3_lm <- lapply(21:320, function(i) {
    lm = lm(HDL_IPV3 ~ cpg_male[, i] + maternal_age + race_4 +
        Bcell + CD4T + CD8T + Gran + Mono + NK + nRBC, data = cpg_male)
    coef = round(summary(lm)$coefficients[2, ], 4)
    return(coef)
})
HDL_IPV3_lm <- data.frame(matrix(unlist(HDL_IPV3_lm), ncol = 4,
    byrow = TRUE, dimnames = list(c(colnames(cpg_male)[21:320]),
        c("Estimate", "Std.Error", "t.statistic", "p.value"))))

# adjusted p-value
HDL_IPV3_lm <- HDL_IPV3_lm %>% mutate(FDR = p.adjust(p.value,
    "BH", 300), FWER = p.adjust(p.value, "bonferroni", 300),
    names = colnames(cpg_male)[21:320]) %>% select(names, everything())
HDL_IPV3_lm <- HDL_IPV3_lm[order(HDL_IPV3_lm$FDR), ]


## sample size
size6 <- length(cpg_male$HDL_IPV3) - sum(is.na(cpg_male$HDL_IPV3))

## summary table
kable(head(HDL_IPV3_lm, 10), caption = paste("Top10 FDR HDL_IPV3 ",
    "n = ", size6, sep = ""))

###################### Insu_IPV3 ###################################
Insu_IPV3_lm <- lapply(21:320, function(i) {
    lm = lm(Insu_IPV3 ~ cpg_male[, i] + maternal_age + race_4 +
```

```
        Bcell + CD4T + CD8T + Gran + Mono + NK + nRBC, data = cpg_male)
    coef = round(summary(lm)$coefficients[2, ], 4)
    return(coef)
})
Insu_IPV3_lm <- data.frame(matrix(unlist(Insu_IPV3_lm), ncol = 4,
    byrow = TRUE, dimnames = list(c(colnames(cpg_male)[21:320]),
        c("Estimate", "Std.Error", "t.statistic", "p.value"))))

# adjusted p-value
Insu_IPV3_lm <- Insu_IPV3_lm %>% mutate(FDR = p.adjust(p.value,
    "BH", 300), FWER = p.adjust(p.value, "bonferroni", 300),
    names = colnames(cpg_male)[21:320]) %>% select(names, everything())
Insu_IPV3_lm <- Insu_IPV3_lm[order(Insu_IPV3_lm$FDR), ]

## sample size
size7 <- length(cpg_male$Insu_IPV3) - sum(is.na(cpg_male$Insu_IPV3))

## summary table
kable(head(Insu_IPV3_lm, 10), caption = paste("Top10 FDR Insu_IPV3 ",
    "n = ", size7, sep = ""))

############## Trig_IPV3 ##########################
Trig_IPV3_lm <- lapply(21:320, function(i) {
    lm = lm(Trig_IPV3 ~ cpg_male[, i] + maternal_age + race_4 +
        Bcell + CD4T + CD8T + Gran + Mono + NK + nRBC, data = cpg_male)
    coef = round(summary(lm)$coefficients[2, ], 4)
    return(coef)
})
Trig_IPV3_lm <- data.frame(matrix(unlist(Trig_IPV3_lm), ncol = 4,
    byrow = TRUE, dimnames = list(c(colnames(cpg_male)[21:320]),
        c("Estimate", "Std.Error", "t.statistic", "p.value"))))

# adjusted p-value
Trig_IPV3_lm <- Trig_IPV3_lm %>% mutate(FDR = p.adjust(p.value,
    "BH", 300), FWER = p.adjust(p.value, "bonferroni", 300),
    names = colnames(cpg_male)[21:320]) %>% select(names, everything())
Trig_IPV3_lm <- Trig_IPV3_lm[order(Trig_IPV3_lm$FDR), ]


## sample size
size8 <- length(cpg_male$Trig_IPV3) - sum(is.na(cpg_male$Trig_IPV3))

## summary table
kable(head(Trig_IPV3_lm, 10), caption = paste("Top10 FDR Trig_IPV3 ",
    "n = ", size8, sep = ""))

############# Leptin_actual__ng_ml_ ####################
Leptin_actual__ng_ml__lm <- lapply(21:320, function(i) {
    lm = lm(Leptin_actual__ng_ml_ ~ cpg_male[, i] + maternal_age +
        race_4 + Bcell + CD4T + CD8T + Gran + Mono + NK + nRBC,
        data = cpg_male)
    coef = round(summary(lm)$coefficients[2, ], 4)
    return(coef)
```

```
})
Leptin_actual__ng_ml__lm <- data.frame(matrix(unlist(Leptin_actual__ng_ml__lm),
    ncol = 4, byrow = TRUE, dimnames = list(c(colnames(cpg_male)[21:320]),
        c("Estimate", "Std.Error", "t.statistic", "p.value"))))

# adjusted p-value
Leptin_actual__ng_ml__lm <- Leptin_actual__ng_ml__lm %>% mutate(FDR = p.adjust(p.value,
    "BH", 300), FWER = p.adjust(p.value, "bonferroni", 300),
    names = colnames(cpg_male)[21:320]) %>% select(names, everything())

Leptin_actual__ng_ml__lm <- Leptin_actual__ng_ml__lm[order(Leptin_actual__ng_ml__lm$FDR),
    ]


## sample size
size9 <- length(cpg_male$Leptin_actual__ng_ml_) - sum(is.na(cpg_male$Leptin_actual__ng_ml_))

## summary table
kable(head(Leptin_actual__ng_ml__lm, 10), caption = paste("Top10 FDR Leptin_actual__ng_ml ",
    "n = ", size9, sep = ""))

### Double check
check_lm <- lm(Leptin_actual__ng_ml_ ~ cpg_male[, 42] + maternal_age +
    race_4 + Bcell + CD4T + CD8T + Gran + Mono + NK + nRBC, data = cpg_male)
summary(check_lm)
```