# MATH630 Replication Project Analysis

*Joshua Burkhart*

*November 22, 2015*

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```
library(plyr)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
##
## The following objects are masked from 'package:plyr':
##
##     arrange, count, desc, failwith, id, mutate, rename, summarise,
##     summarize
##
## The following objects are masked from 'package:stats':
##
##     filter, lag
##
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
library(broom)
library(ggfortify)
```

```
## Loading required package: grid
## Loading required package: scales
## Loading required package: proto
```

```
library(GGally)
```

```
##
## Attaching package: 'GGally'
##
## The following object is masked from 'package:dplyr':
##
##     nasa
```

```
library(car)
library(MBESS)
library(ggplot2)
library(magrittr)
```

```r
full <- read.csv("/Users/joshuaburkhart/SoftwareProjects/Probability/final_proj/data.csv", row.names = 1

## Generate centered and transformed variables
full$lnAs <- log(full$Ascounty) - mean(na.omit(log(full$Ascounty)))
full$lnInc <- log(full$MedIncome) - mean(na.omit(log(full$MedIncome)))
full$Population <- as.numeric(as.character(full$Population))
full$lnsmk <- full$smkrate
full$lnar <- log(full$AdjRate)
```

```r
## Bivariate, Untransformed
## Arsenic Levels and Lung Cancer Incidence, weighted
glm1 <- glm(full$AdjRate ~ full$Ascounty, family=poisson, weights=as.numeric(full$Population))
summary(glm1)
```

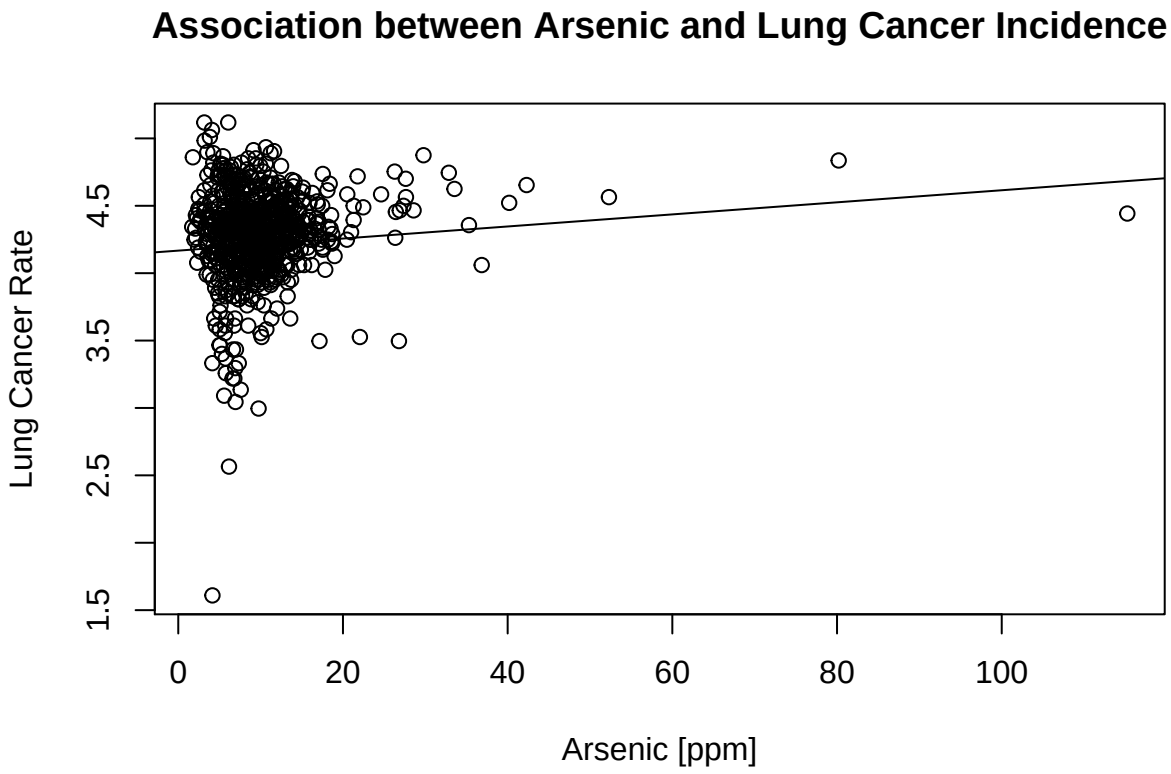
```
##
## Call:
## glm(formula = full$AdjRate ~ full$Ascounty, family = poisson,
##     weights = as.numeric(full$Population))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -6257.0   -103.2    161.1    444.0   2459.2
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)   4.167e+00  2.370e-05  175836   <2e-16 ***
## full$Ascounty 4.479e-03  1.937e-06    2312   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 317986702  on 741  degrees of freedom
## Residual deviance: 312986732  on 740  degrees of freedom
##   (15 observations deleted due to missingness)
## AIC: 835273995
##
## Number of Fisher Scoring iterations: 4
```

```r
lm1 <- lm(full$lnar ~ full$Ascounty, weights=as.numeric(full$Population))
summary(lm1)
```

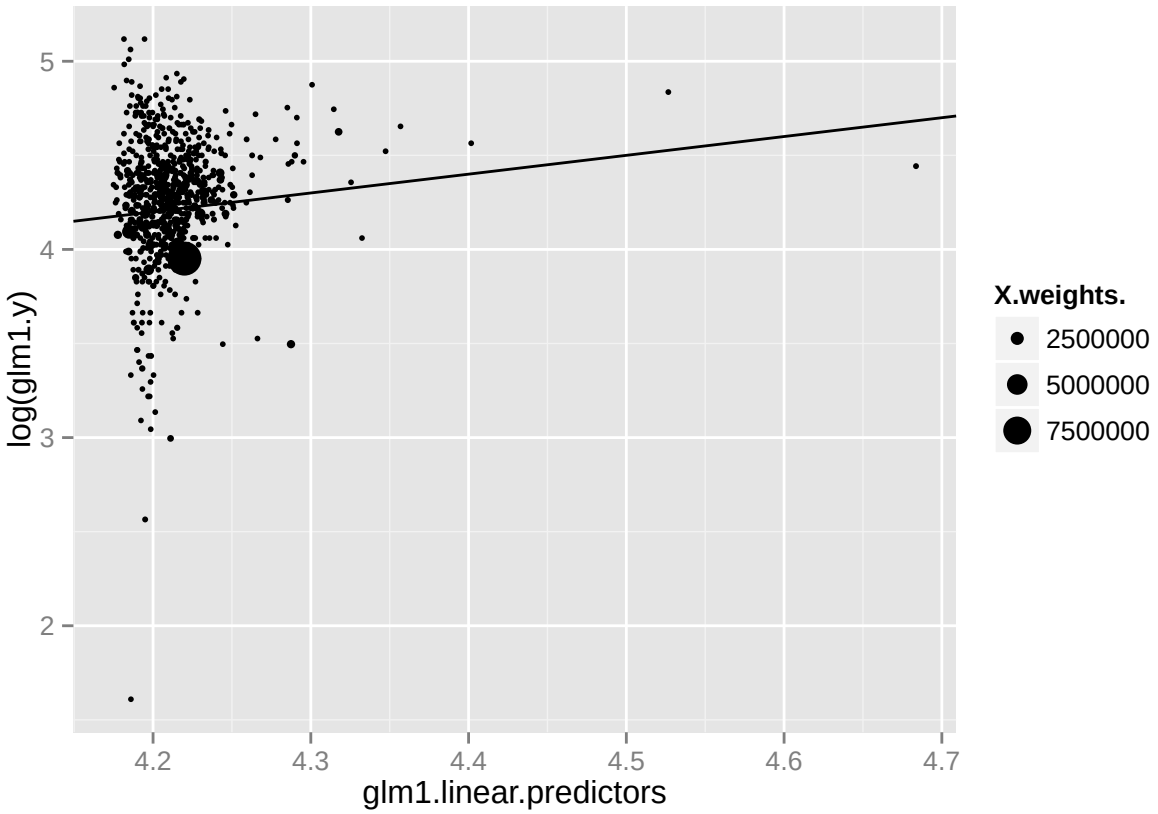
```
##
## Call:
## lm(formula = full$lnar ~ full$Ascounty, weights = as.numeric(full$Population))
##
## Weighted Residuals:
##     Min      1Q  Median      3Q     Max
## -738.43   -8.18   23.65   55.32  293.99
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)    4.14686    0.01727 240.168   <2e-16 ***
## full$Ascounty  0.00373    0.00147   2.537   0.0114 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 83.81 on 740 degrees of freedom
##   (15 observations deleted due to missingness)
## Multiple R-squared:  0.008621,   Adjusted R-squared:  0.007282
## F-statistic: 6.435 on 1 and 740 DF,  p-value: 0.01139
```

```
plot(y=log(full$AdjRate), x=full$Ascounty, ylab="Lung Cancer Rate", xlab="Arsenic [ppm]", main="Associa
abline(a=glm1$coef[1], b=glm1$coef[2], col=1)
```




```
glm1T <- data.frame(glm1$linear.predictors, glm1$y, glm1$model[3])
glm1gg <- ggplot(glm1T, aes(x=glm1.linear.predictors, y=log(glm1.y)))
glm1gg + geom_point(aes(size=X.weights.)) +
         geom_abline(coef(lm(log(full$AdjRate) ~ full$Ascounty,   weights=as.numeric(full$Population)))
```




```
## Adjusted, Untransformed
## Arsenic, Smoking, SES
SESassmk <- glm(full$AdjRate ~ full$smkrate + full$Ascounty + full$MedIncome, family=poisson, weights=as
summary(SESassmk)
```

```
##
## Call:
## glm(formula = full$AdjRate ~ full$smkrate + full$Ascounty + full$MedIncome,
##     family = poisson, weights = as.numeric(full$Population))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3559.7   -240.9     46.1    347.7   1949.6
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)     3.521e+00  1.213e-04   29017   <2e-16 ***
## full$smkrate    1.802e+00  1.913e-04    9419   <2e-16 ***
## full$Ascounty   3.931e-03  1.930e-06    2037   <2e-16 ***
## full$MedIncome -3.538e-06  1.308e-09   -2706   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 305018305  on 584  degrees of freedom
## Residual deviance: 180096592  on 581  degrees of freedom
```

```
##   (172 observations deleted due to missingness)
## AIC: 687702879
##
## Number of Fisher Scoring iterations: 4
```

```
SESassmklm <- lm(full$lnar ~ full$smkrate + full$Ascounty + full$MedIncome, weights=as.numeric(full$Popu
summary(SESassmklm)
```

```
##
## Call:
## lm(formula = full$lnar ~ full$smkrate + full$Ascounty + full$MedIncome,
##     weights = as.numeric(full$Population))
##
## Weighted Residuals:
##     Min      1Q  Median      3Q     Max
## -442.09  -28.41    7.35   42.54  234.40
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     3.292e+00  6.929e-02  47.506  < 2e-16 ***
## full$smkrate    2.173e+00  1.129e-01  19.247  < 2e-16 ***
## full$Ascounty   3.425e-03  1.231e-03   2.782 0.005583 **
## full$MedIncome -2.396e-06  7.225e-07  -3.317 0.000968 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 69.48 on 581 degrees of freedom
##   (172 observations deleted due to missingness)
## Multiple R-squared:  0.4433, Adjusted R-squared:  0.4404
## F-statistic: 154.2 on 3 and 581 DF,  p-value: < 2.2e-16
```

You can also embed plots, for example:

```
## Estimate the 25, 50, amd 75% quartile points for each variable for the quartiles interaction models
AsCut <- NA
AsCut[1] <- as.numeric(summary(full$lnAs)[2])
AsCut[2] <- as.numeric(summary(full$lnAs)[3])
AsCut[3] <- as.numeric(summary(full$lnAs)[5])

SmkCut <- NA
SmkCut[1] <- as.numeric(summary(full$lnsmk)[2])
SmkCut[2] <- as.numeric(summary(full$lnsmk)[3])
SmkCut[3] <- as.numeric(summary(full$lnsmk)[5])

SESCut <- NA
SESCut[1] <- as.numeric(summary(full$lnInc)[2])
SESCut[2] <- as.numeric(summary(full$lnInc)[3])
SESCut[3] <- as.numeric(summary(full$lnInc)[5])
```

```
## Continuous Interaction Models
## Arsenic and Smoking
AsSmk <- full$lnAs * full$lnsmk
intAsSmk <- aov(full$AdjRate ~ full$lnsmk + full$lnAs + full$lnInc + AsSmk, weights=as.numeric(full$Popu
summary(intAsSmk)
```

```
##               Df    Sum Sq   Mean Sq F value   Pr(>F)
## full$lnsmk     1 7.766e+09 7.766e+09  378.03  < 2e-16 ***
## full$lnAs      1 2.112e+08 2.112e+08   10.28  0.00142 **
## full$lnInc     1 6.238e+08 6.238e+08   30.36 5.39e-08 ***
## AsSmk          1 1.374e+08 1.374e+08    6.69  0.00994 **
## Residuals    580 1.192e+10 2.054e+07
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 172 observations deleted due to missingness
```

```
## Arsenic and SES
AsSES <- full$lnAs * full$lnInc
intAsSES <- aov(full$AdjRate ~ full$lnsmk + full$lnAs + full$lnInc + AsSES, weights=as.numeric(full$Popu
summary(intAsSES)
```

```
##               Df    Sum Sq   Mean Sq F value   Pr(>F)
## full$lnsmk     1 7.766e+09 7.766e+09 375.784  < 2e-16 ***
## full$lnAs      1 2.112e+08 2.112e+08  10.219  0.00147 **
## full$lnInc     1 6.238e+08 6.238e+08  30.183 5.89e-08 ***
## AsSES          1 6.627e+07 6.627e+07   3.206  0.07387 .
## Residuals    580 1.199e+10 2.067e+07
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 172 observations deleted due to missingness
```

```
## Calculate Strat Groups ##
## Smoking Quartiles
smkgrp <- ifelse(is.na(full$lnsmk), NA, ifelse(full$lnsmk < SmkCut[1], 1, ifelse(full$lnsmk >= SmkCut[1
```

```
## SES Low-Income Cutoffs
SESgrp <- ifelse(is.na(full$MedIncome), NA, ifelse(full$MedIncome < 24000 & !is.na(full$MedIncome), 1, i
```

```
## SES Quartiles
#SESgrp <- ifelse(full$lnInc< -0.158, 1, ifelse(full$lnInc>=-0.158 & full$lnInc< -0.00391, 2, ifelse(fu
## Arsenic Quartiles
AsQ <- ifelse(is.na(full$lnAs), NA, ifelse(full$lnAs < AsCut[1], 1, ifelse(full$lnAs >= AsCut[1] & full$
```

```
## Quartile-Based Interaction Models
## Convert quartiles to factors
AsQf <- as.factor(AsQ)
smkgrpf <- as.factor(smkgrp)
smkgrpfbak <- smkgrpf
SESgrpf <- as.factor(SESgrp)
```

```
##############
## ARSENIC ## figure 2
##############
## Arsenic and Smoking table 3 first line
smkgrpf <- smkgrpfbak
smkgrpf <- ifelse(is.na(smkgrpf), NA, ifelse(smkgrpf==1 | smkgrpf==2, 1, 2))
intAsSmk <- aov(full$AdjRate ~ SESgrpf + AsQf*smkgrpf, weights=as.numeric(full$Population))
summary(intAsSmk)
```

```
##                Df    Sum Sq   Mean Sq F value   Pr(>F)
## SESgrpf         3 2.695e+09 8.984e+08  37.943  < 2e-16 ***
## AsQf            3 1.567e+09 5.222e+08  22.057 1.57e-13 ***
## smkgrpf         1 2.614e+09 2.614e+09 110.390  < 2e-16 ***
## AsQf:smkgrpf    3 1.889e+08 6.297e+07   2.659   0.0475 *
## Residuals     574 1.359e+10 2.368e+07
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 172 observations deleted due to missingness
```

```
## Without SES
intAsSmk2 <- aov(full$AdjRate ~ AsQf*smkgrpf, weights=as.numeric(full$Population))
summary(intAsSmk2)
```

```
##                Df    Sum Sq   Mean Sq F value   Pr(>F)
## AsQf            3 1.891e+09 6.302e+08  24.392 7.15e-15 ***
## smkgrpf         1 3.665e+09 3.665e+09 141.857  < 2e-16 ***
## AsQf:smkgrpf    3 1.913e+08 6.378e+07   2.469   0.0611 .
## Residuals     577 1.491e+10 2.584e+07
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 172 observations deleted due to missingness
```

```
## Arsenic and SES table 3 line 2
intAsSES <- aov(full$AdjRate ~ smkgrpf + AsQf*SESgrpf, weights=as.numeric(full$Population))
summary(intAsSES)
```

```
##                Df    Sum Sq   Mean Sq F value   Pr(>F)
## smkgrpf         1 4.252e+09 4.252e+09  180.22  < 2e-16 ***
## AsQf            3 1.304e+09 4.345e+08   18.42 2.05e-11 ***
## SESgrpf         3 1.320e+09 4.399e+08   18.64 1.51e-11 ***
## AsQf:SESgrpf    9 3.779e+08 4.199e+07    1.78   0.0691 .
## Residuals     568 1.340e+10 2.359e+07
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 172 observations deleted due to missingness
```

```
## Plot the Interaction between Arsenic and Smoking WITHOUT SES
smkgrpf <- smkgrpfbak
smkgrpf <- ifelse(is.na(smkgrpf), NA, ifelse(smkgrpf==1 | smkgrpf==2, 1, 2))
r1 <- glm(full[smkgrpf==1,]$AdjRate ~ full[smkgrpf==1,]$lnAs, family=poisson, weights=as.numeric(full[sm
summary(r1)
```

```
##
## Call:
## glm(formula = full[smkgrpf == 1, ]$AdjRate ~ full[smkgrpf ==
##     1, ]$lnAs, family = poisson, weights = as.numeric(full[smkgrpf ==
##     1, ]$Population))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -4555.0   -105.5    172.0    525.9   2252.1
```

```
##
## Coefficients:
##                             Estimate Std. Error  z value Pr(>|z|)
## (Intercept)                4.144e+00  1.604e-05 258419.4   <2e-16 ***
## full[smkgrpf == 1, ]$lnAs 1.931e-02  3.293e-05    586.4   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 175916938  on 289  degrees of freedom
## Residual deviance: 175572576  on 288  degrees of freedom
##   (167 observations deleted due to missingness)
## AIC: 543281859
##
## Number of Fisher Scoring iterations: 4
```

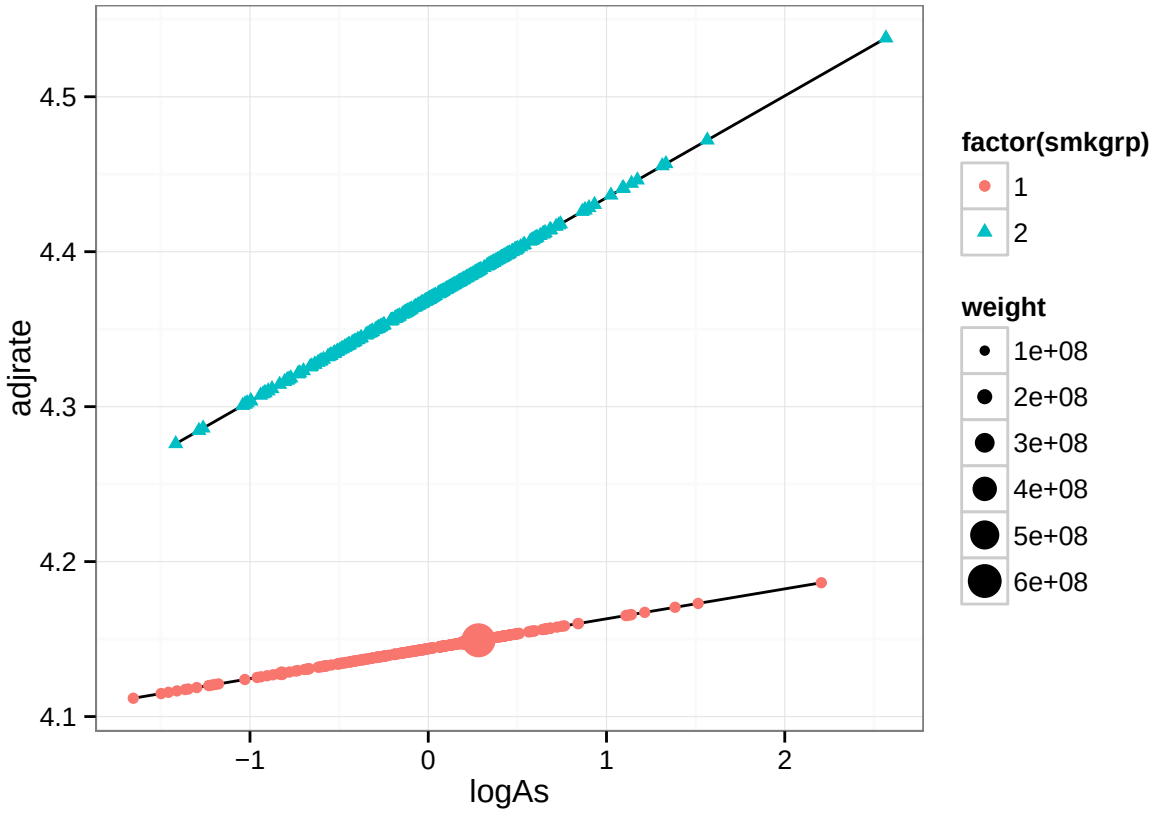
```
r2 <- glm(full[smkgrpf==2,]$AdjRate ~ full[smkgrpf==2,]$lnAs, family=poisson, weights=as.numeric(full[sm
summary(r2)
```

```
##
## Call:
## glm(formula = full[smkgrpf == 2, ]$AdjRate ~ full[smkgrpf ==
##     2, ]$lnAs, family = poisson, weights = as.numeric(full[smkgrpf ==
##     2, ]$Population))
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1773.74   -228.76     53.34    326.82   1574.57
##
## Coefficients:
##                             Estimate Std. Error z value Pr(>|z|)
## (Intercept)                4.369e+00  2.369e-05  184430   <2e-16 ***
## full[smkgrpf == 2, ]$lnAs 6.571e-02  4.283e-05    1534   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 68053265  on 294  degrees of freedom
## Residual deviance: 65713881  on 293  degrees of freedom
##   (162 observations deleted due to missingness)
## AIC: 205610885
##
## Number of Fisher Scoring iterations: 4
```

```
data1 <- cbind( c(t(r1$model[2]),t(r2$model[2])),
                c(log(r1$fitted.values),log(r2$fitted.values)), # figure 2 replace fitted
                c(r1$weights, r2$weights),
                c(rep("1",dim(r1$model[2])[1]), rep("2",dim(r2$model[2])[1])))
data1 <- as.data.frame(data1, stringsAsFactors=FALSE)
names(data1) <- c("logAs","logRate", "weight", "smkgrp")
data1$logAs  <- as.numeric(as.character(data1$logAs))
```

```
data1$logRate <- as.numeric(as.character(data1$logRate))
data1$weight  <- as.numeric(as.character(data1$weight))
data1$smkgrp  <- as.numeric(as.character(data1$smkgrp))
data1$adjinc  <- c(as.numeric(coef(r1)[2])*r1$model[,2]*0, as.numeric(coef(r2)[2])*r2$model[,2]*0)
data1$adjrate <- data1$adjinc+data1$logRate

assmkp <- ggplot(data1, aes(x=logAs, y=adjrate, shape=factor(smkgrp), color=factor(smkgrp)))
assmkp + stat_smooth(method = "glm", level=0.95, alpha=1, fill="grey80", color="black") +
         #scale_color_manual(values=c("grey50","grey70")) +
         geom_point(aes(size=weight)) +
         geom_point() +
         theme(legend.position = "right") +
         theme_bw()
```




```
## GLMS for smoking levels WITH SES
smkgrpf <- smkgrpfbak
## Bottom 50% vs Top 50%
smkgrpf <- ifelse(is.na(smkgrpf), NA, ifelse(smkgrpf==1 | smkgrpf==2, 1, 2))
r1 <- glm(full[smkgrpf==1,]$AdjRate ~ full[smkgrpf==1,]$lnAs + full[smkgrpf==1,]$MedIncome, family=pois
summary(r1)
```

```
##
## Call:
## glm(formula = full[smkgrpf == 1, ]$AdjRate ~ full[smkgrpf ==
##     1, ]$lnAs + full[smkgrpf == 1, ]$MedIncome, family = poisson,
##     weights = as.numeric(full[smkgrpf == 1, ]$Population))
##
```

```
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -4941.6   -151.1    150.1    425.7   2249.4
##
## Coefficients:
##                                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)                    4.326e+00  7.146e-05 60531.0   <2e-16 ***
## full[smkgrpf == 1, ]$lnAs      7.596e-03  3.328e-05   228.2   <2e-16 ***
## full[smkgrpf == 1, ]$MedIncome -3.862e-06  1.486e-09 -2599.0   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 175916938  on 289  degrees of freedom
## Residual deviance: 168750465  on 287  degrees of freedom
##   (167 observations deleted due to missingness)
## AIC: 536459750
##
## Number of Fisher Scoring iterations: 4
```

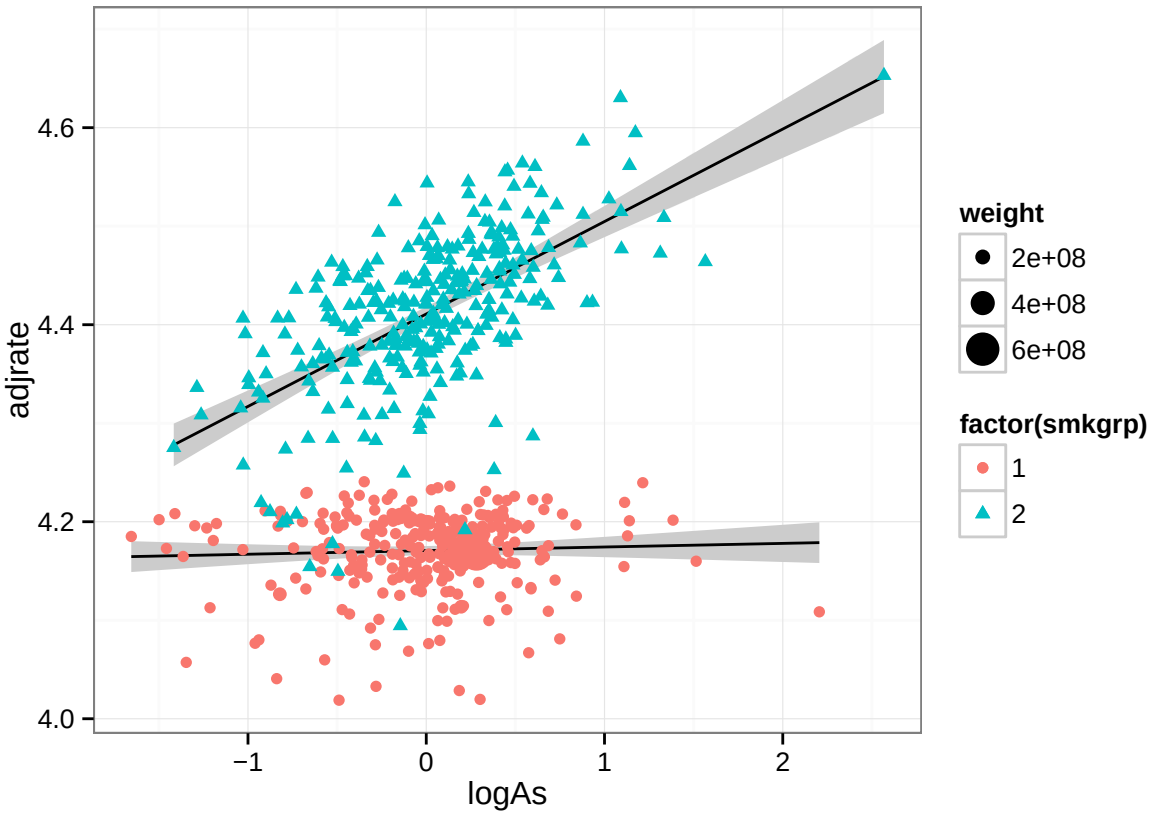
```
r2 <- glm(full[smkgrpf==2,]$AdjRate ~ full[smkgrpf==2,]$lnAs + full[smkgrpf==2,]$MedIncome, family=pois
summary(r2)
```

```
##
## Call:
## glm(formula = full[smkgrpf == 2, ]$AdjRate ~ full[smkgrpf ==
##     2, ]$lnAs + full[smkgrpf == 2, ]$MedIncome, family = poisson,
##     weights = as.numeric(full[smkgrpf == 2, ]$Population))
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1769.12   -244.11     -9.96    270.72   1591.53
##
## Coefficients:
##                                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)                    4.689e+00  1.135e-04   41314   <2e-16 ***
## full[smkgrpf == 2, ]$lnAs      5.935e-02  4.312e-05    1377   <2e-16 ***
## full[smkgrpf == 2, ]$MedIncome -7.985e-06  2.792e-09   -2860   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 68053265  on 294  degrees of freedom
## Residual deviance: 57437520  on 292  degrees of freedom
##   (162 observations deleted due to missingness)
## AIC: 197334525
##
## Number of Fisher Scoring iterations: 4
```

```
## Plot the Interaction between Arsenic and Smoking with SES
data1 <- cbind( c(t(r1$model[2]),t(r2$model[2])),
```

```
                c(log(r1$fitted.values),log(r2$fitted.values)),
                c(r1$weights, r2$weights),
                c(rep("1",dim(r1$model[2])[1]), rep("2",dim(r2$model[2])[1])))
data1 <- as.data.frame(data1, stringsAsFactors=FALSE)
names(data1) <- c("logAs","logRate", "weight", "smkgrp")
data1$logAs  <- as.numeric(as.character(data1$logAs))
data1$logRate <- as.numeric(as.character(data1$logRate))
data1$weight  <- as.numeric(as.character(data1$weight))
data1$smkgrp  <- as.numeric(as.character(data1$smkgrp))
data1$adjinc  <- c(as.numeric(coef(r1)[2])*r1$model[,2], as.numeric(coef(r2)[2])*r2$model[,2])
data1$adjrate <- data1$adjinc+data1$logRate

assmkp <- ggplot(data1, aes(x=logAs, y=adjrate, shape=factor(smkgrp), color=factor(smkgrp)))
assmkp + stat_smooth(method = "glm", level=0.95, alpha=1, fill="grey80", color="black") +
         #scale_color_manual(values=c("grey50","grey70")) +
         geom_point(aes(size=weight)) +
     geom_point() +
         theme(legend.position = "right") +
        theme_bw()
```




```
## Determine the concentration of each heavy metal in KY|WV and !KY&!WV
## Average County averages for both states

## Arsenic
KYWVas    <- mean(na.omit(full[full$SFIPS==21 | full$SFIPS==54,]$Ascounty))
KYWVas
```

```
## [1] 11.50363
```

```
notKYWVas <- mean(na.omit(full[full$SFIPS!=21 & full$SFIPS!=54,]$Ascounty))
notKYWVas
```

```
## [1] 9.670494
```

```
t.test(na.omit(full[full$SFIPS==21 | full$SFIPS==54,]$Ascounty),
       na.omit(full[full$SFIPS!=21 & full$SFIPS!=54,]$Ascounty))
```

```
##
##  Welch Two Sample t-test
##
## data:  na.omit(full[full$SFIPS == 21 | full$SFIPS == 54, ]$Ascounty) and na.omit(full[full$SFIPS != 2
## t = 2.4517, df = 217.87, p-value = 0.015
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.359501 3.306767
## sample estimates:
## mean of x mean of y
## 11.503628  9.670494
```

```
## Smoking Prevalence
KYWVsmk       <- mean(na.omit(full[full$SFIPS==21 | full$SFIPS==54,]$smkrate))
KYWVsmk
```

```
## [1] 0.5220805
```

```
notKYWVsmk <- mean(na.omit(full[full$SFIPS!=21 & full$SFIPS!=54,]$smkrate))
notKYWVsmk
```

```
## [1] 0.4714118
```

```
t.test(na.omit(full[full$SFIPS==21 | full$SFIPS==54,]$smkrate),
       na.omit(full[full$SFIPS!=21 & full$SFIPS!=54,]$smkrate))
```

```
##
##  Welch Two Sample t-test
##
## data:  na.omit(full[full$SFIPS == 21 | full$SFIPS == 54, ]$smkrate) and na.omit(full[full$SFIPS != 2
## t = 6.3875, df = 292.6, p-value = 6.617e-10
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.03505673 0.06628065
## sample estimates:
## mean of x mean of y
## 0.5220805 0.4714118
```

```
## Median Income
KYWVmed    <- mean(na.omit(full[full$SFIPS==21 | full$SFIPS==54,]$MedIncome))
KYWVmed
```

```
## [1] 31011.92
```

```
notKYWVmed <- mean(na.omit(full[full$SFIPS!=21 & full$SFIPS!=54,]$MedIncome))
notKYWVmed
```

```
## [1] 37771.8
```

```
t.test(na.omit(full[full$SFIPS==21 | full$SFIPS==54,]$MedIncome),
       na.omit(full[full$SFIPS!=21 & full$SFIPS!=54,]$MedIncome))
```

```
##
##  Welch Two Sample t-test
##
## data:  na.omit(full[full$SFIPS == 21 | full$SFIPS == 54, ]$MedIncome) and na.omit(full[full$SFIPS !=
## t = -9.7384, df = 359.1, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -8124.976 -5394.775
## sample estimates:
## mean of x mean of y
##  31011.92  37771.80
```

```
## Lung cancer incidence
KYWVrate    <- mean(na.omit(full[full$SFIPS==21 | full$SFIPS==54,]$AdjRate))
KYWVrate
```

```
## [1] 100.2743
```

```
notKYWVrate <- mean(na.omit(full[full$SFIPS!=21 & full$SFIPS!=54,]$AdjRate))
notKYWVrate
```

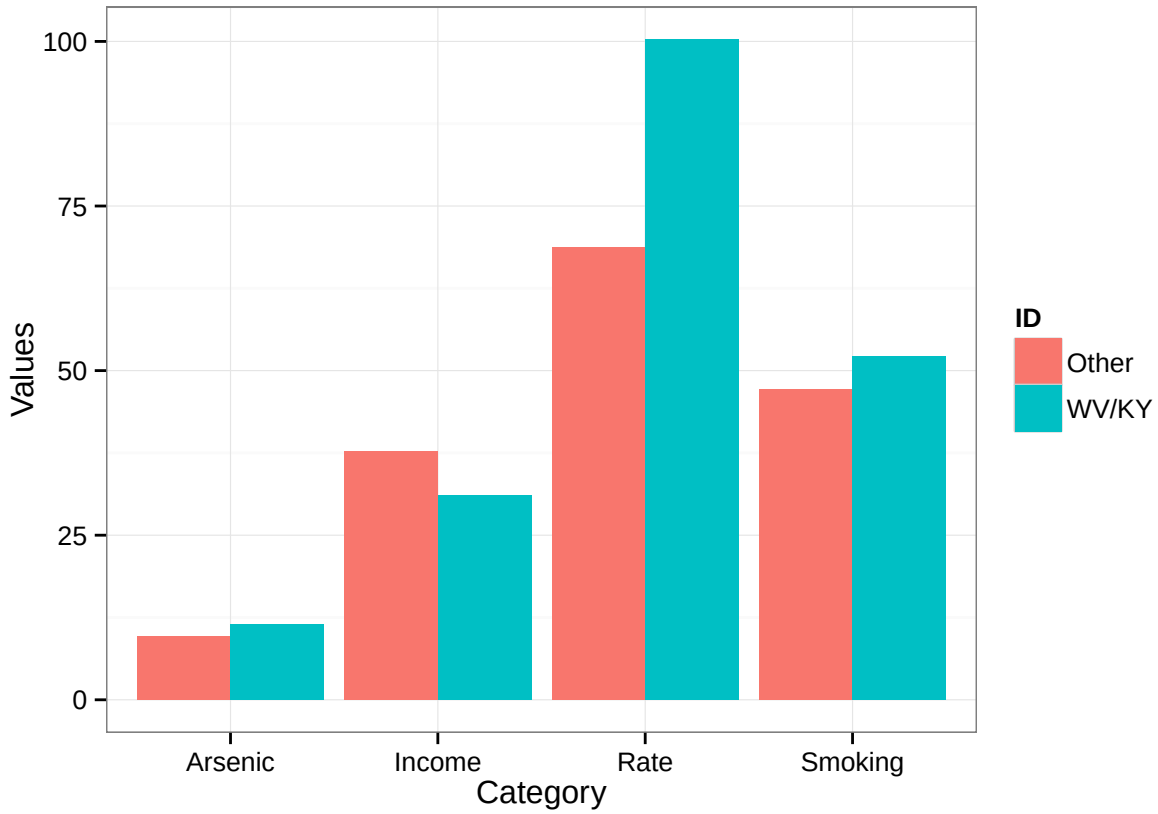
```
## [1] 68.69931
```

```
t.test(na.omit(full[full$SFIPS==21 | full$SFIPS==54,]$AdjRate),
       na.omit(full[full$SFIPS!=21 & full$SFIPS!=54,]$AdjRate))
```

```
##
##  Welch Two Sample t-test
##
## data:  na.omit(full[full$SFIPS == 21 | full$SFIPS == 54, ]$AdjRate) and na.omit(full[full$SFIPS != 2
## t = 18.977, df = 247.42, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  28.29777 34.85218
## sample estimates:
## mean of x mean of y
## 100.27429  68.69931
```

```
## Draw a bar chart of the values used in the T-tests
bardata <- c(notKYWVas, KYWVas,
             notKYWVsmk*100, KYWVsmk*100, notKYWVmed/1000, KYWVmed/1000, notKYWVrate, KYWVrate)
barcats <- c("Arsenic", "Arsenic",
             "Smoking", "Smoking", "Income", "Income", "Rate", "Rate")
barid <- c("Other", "WV/KY",
           "Other", "WV/KY", "Other", "WV/KY", "Other", "WV/KY")
bardata1 <- data.frame(barid, barcats, bardata)
names(bardata1) <- c("ID", "Category", "Values")

bar1 <- ggplot(bardata1, aes(Category, fill=ID, y=Values))
bar1 +  geom_bar(position="dodge", stat="identity") + theme_bw()
```



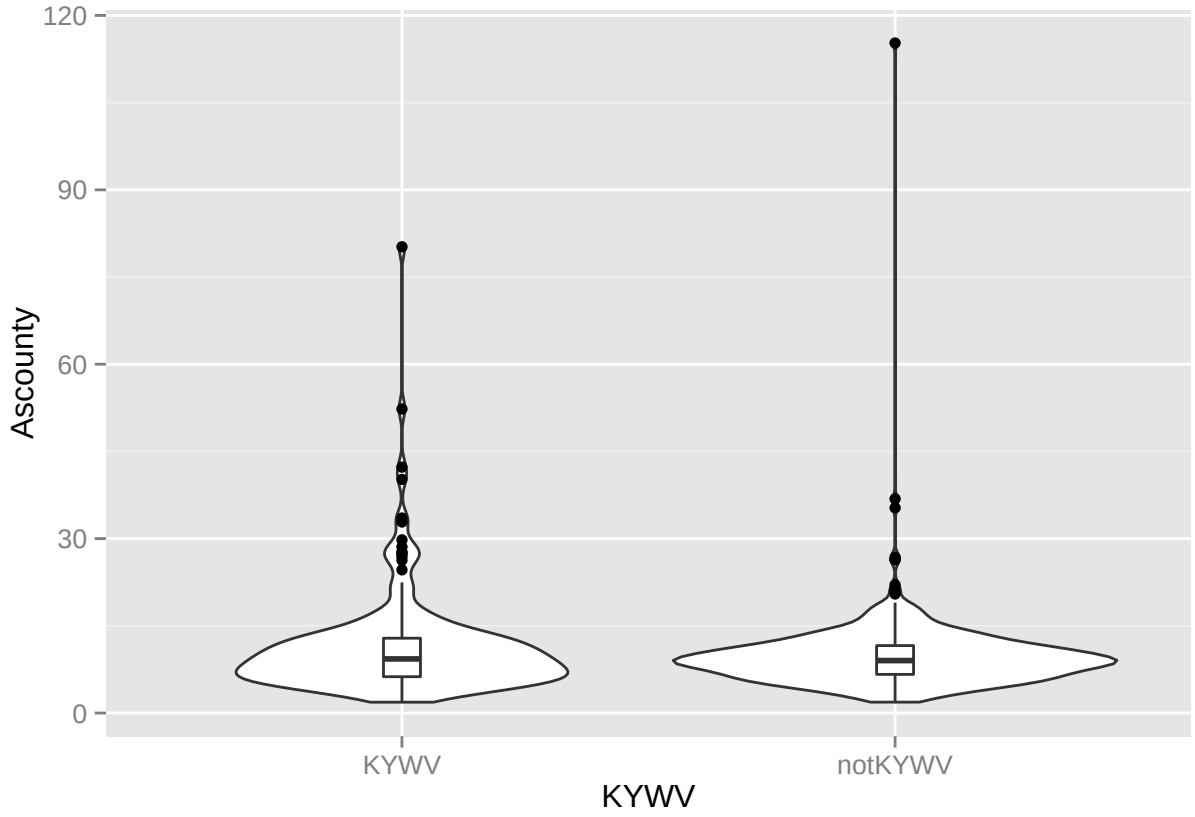

```
# Violin Plots

## Arsenic
KYWVas    <- na.omit(full[full$SFIPS==21 | full$SFIPS==54,]$Ascounty)
notKYWVas <- na.omit(full[full$SFIPS!=21 & full$SFIPS!=54,]$Ascounty)

length(KYWVas) = length(notKYWVas)
KYWVdf    <- KYWVas    %>% as.data.frame() %>% mutate(KYWV="KYWV") %>% select("Ascounty"=1,"KYWV"=2)
notKYWVdf <- notKYWVas %>% as.data.frame() %>% mutate(KYWV="notKYWV") %>% select("Ascounty"=1,"KYWV"=2)
KYWVcombined = merge(KYWVdf,notKYWVdf,all=TRUE)
KYWVcombined %>% ggplot(aes(x=KYWV,y=Ascounty)) + geom_violin() +
                          geom_boxplot(width=0.1)
```

```
## Warning: Removed 398 rows containing non-finite values (stat_ydensity).
```

```
## Warning: Removed 398 rows containing non-finite values (stat_boxplot).
```

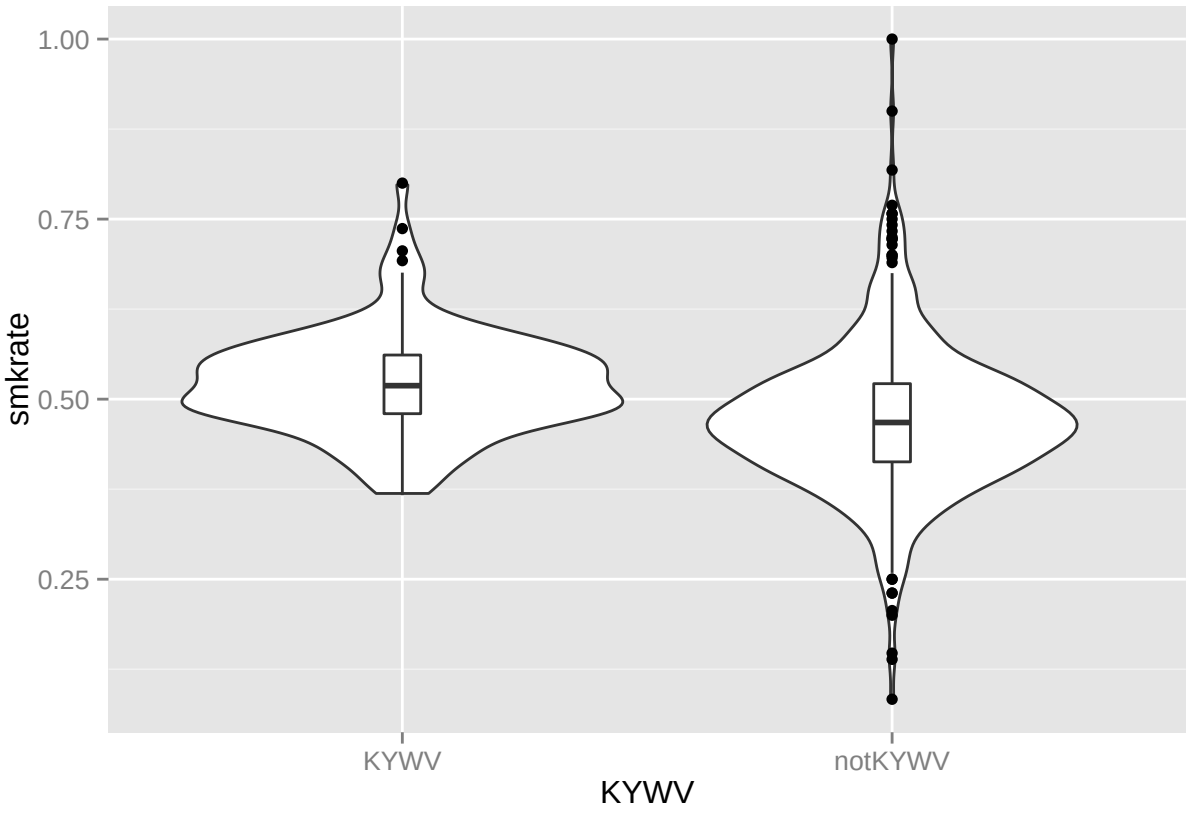



```
## Smoking Prevalence
KYWVsmk      <- na.omit(full[full$SFIPS==21 | full$SFIPS==54,]$smkrate)
notKYWVsmk <- na.omit(full[full$SFIPS!=21 & full$SFIPS!=54,]$smkrate)

length(KYWVsmk) = length(notKYWVsmk)
KYWVdf    <- KYWVsmk    %>% as.data.frame() %>% mutate(KYWV="KYWV") %>% select("smkrate"=1,"KYWV"=2)
notKYWVdf <- notKYWVsmk %>% as.data.frame() %>% mutate(KYWV="notKYWV") %>% select("smkrate"=1,"KYWV"=2)
KYWVcombined = merge(KYWVdf,notKYWVdf,all=TRUE)
KYWVcombined %>% ggplot(aes(x=KYWV,y=smkrate)) + geom_violin() +
                          geom_boxplot(width=0.1)
```

```
## Warning: Removed 340 rows containing non-finite values (stat_ydensity).

## Warning: Removed 340 rows containing non-finite values (stat_boxplot).
```

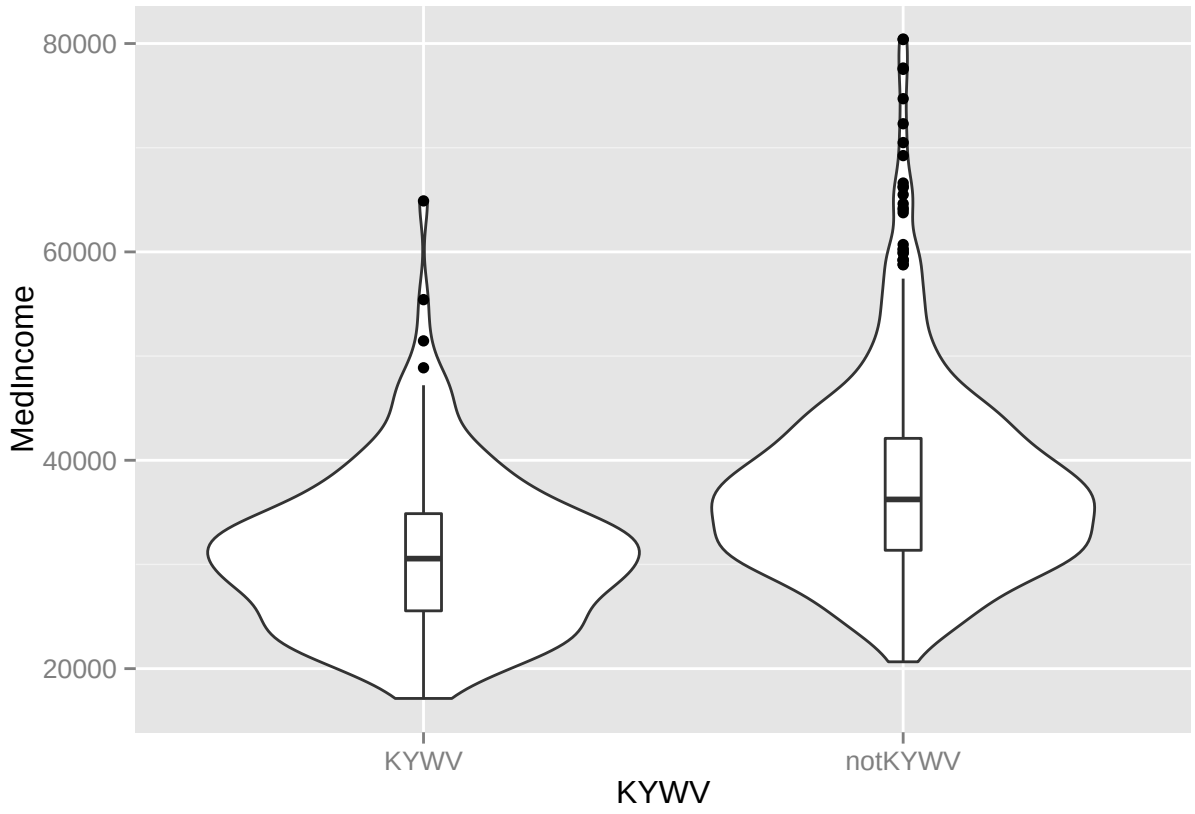



```
## Median Income
KYWVmed    <- na.omit(full[full$SFIPS==21 | full$SFIPS==54,]$MedIncome)
notKYWVmed <- na.omit(full[full$SFIPS!=21 & full$SFIPS!=54,]$MedIncome)

length(KYWVmed) = length(notKYWVmed)
KYWVdf    <- KYWVmed    %>% as.data.frame() %>% mutate(KYWV="KYWV") %>% select("MedIncome"=1,"KYWV"=2)
notKYWVdf <- notKYWVmed %>% as.data.frame() %>% mutate(KYWV="notKYWV") %>% select("MedIncome"=1,"KYWV"=
KYWVcombined = merge(KYWVdf,notKYWVdf,all=TRUE)
KYWVcombined %>% ggplot(aes(x=KYWV,y=MedIncome)) + geom_violin() +
                         geom_boxplot(width=0.1)
```

```
## Warning: Removed 407 rows containing non-finite values (stat_ydensity).
```

```
## Warning: Removed 407 rows containing non-finite values (stat_boxplot).
```

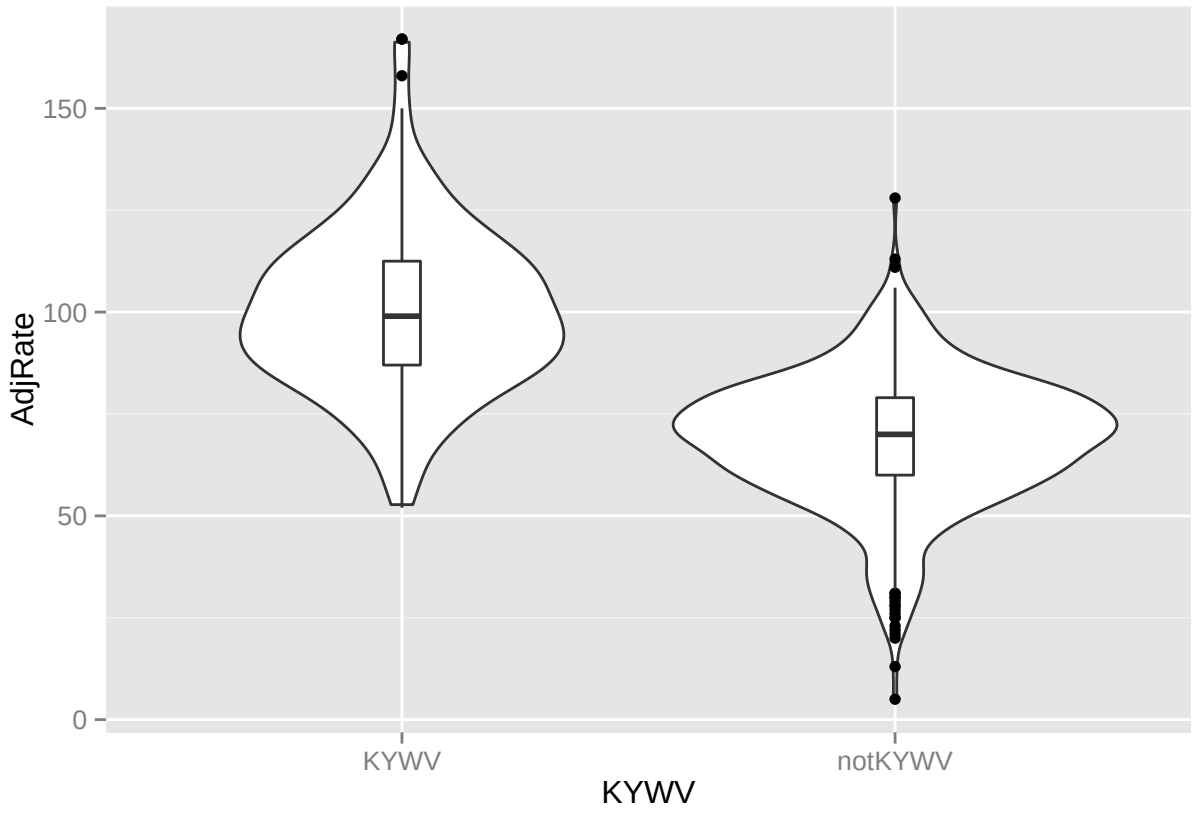



```
## Lung cancer incidence
KYWVrate      <- na.omit(full[full$SFIPS==21 | full$SFIPS==54,]$AdjRate)
notKYWVrate <- na.omit(full[full$SFIPS!=21 & full$SFIPS!=54,]$AdjRate)

length(KYWVrate) = length(notKYWVrate)
KYWVdf    <- KYWVrate    %>% as.data.frame() %>% mutate(KYWV="KYWV") %>% select("AdjRate"=1,"KYWV"=2)
notKYWVdf <- notKYWVrate %>% as.data.frame() %>% mutate(KYWV="notKYWV") %>% select("AdjRate"=1,"KYWV"=2
KYWVcombined = merge(KYWVdf,notKYWVdf,all=TRUE)
KYWVcombined %>% ggplot(aes(x=KYWV,y=AdjRate)) + geom_violin() +
                          geom_boxplot(width=0.1)
```

```
## Warning: Removed 407 rows containing non-finite values (stat_ydensity).
```

```
## Warning: Removed 407 rows containing non-finite values (stat_boxplot).
```


150
100
50
0
AdjRate
KYWV
notKYWV
KYWV