EPID674 Epidemiologic Data Analysis using R

Regression in R

Kelly Bakulski

Last compiled on October 05, 2021

# Set up options

# Install new packages

# Load packages

# Load data, remake useful variables

# Check the file path  
here("nhanes3.rda")

## [1] "/cloud/project/nhanes3.rda"

# Load the saved R data  
load(here("nhanes3.rda"))  
  
  
# Remake a few variables from last class if they are no longer in your environment  
sex1 <- factor(nhanes$sex, levels = c(1, 2), labels = c("male", "female"))  
AGE5b <- cut(nhanes$age, quantile(nhanes$age, c(0, .2, .4, .6, .8, 1)), include.lowest = T) # quintiles  
AGE5c <- cut(nhanes$age, breaks = c(19, 40, 50, 60, 70, 90))  
age5c <- unclass(AGE5c)  
nhanes <- cbind(nhanes, sex1, AGE5b, AGE5c, age5c)  
  
  
# Clean up the dataset. Make sure NaN are NA for key covariates  
table(nhanes$educ, useNA = "always")

##   
## 1 2 3 NaN <NA>   
## 2032 2349 660 33 0

nhanes$educ[is.nan(nhanes$educ)] <- NA  
table(nhanes$educ, useNA = "always")

##   
## 1 2 3 <NA>   
## 2032 2349 660 33

table(nhanes$alc, useNA = "always")

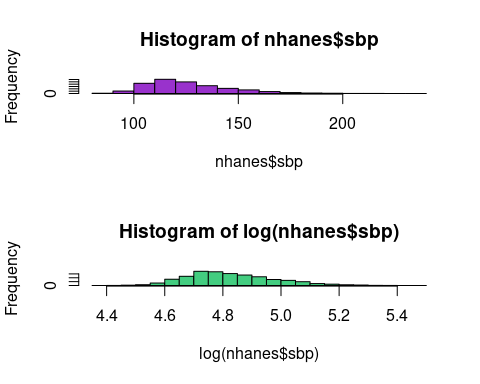
##   
## 0 1 NaN <NA>   
## 2753 2189 132 0

nhanes$alc[is.nan(nhanes$alc)] <- NA  
table(nhanes$alc, useNA = "always")

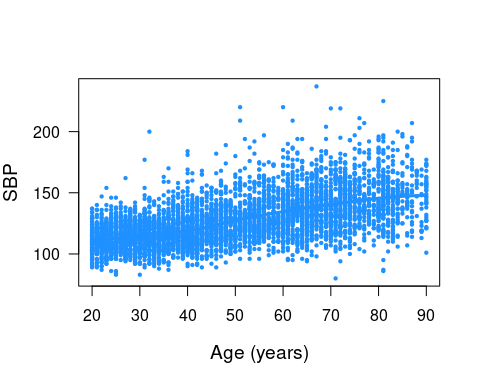
##   
## 0 1 <NA>   
## 2753 2189 132

### 6.1. Linear Models: Association between systolic blood pressure (SBP) and blood lead (bpb)

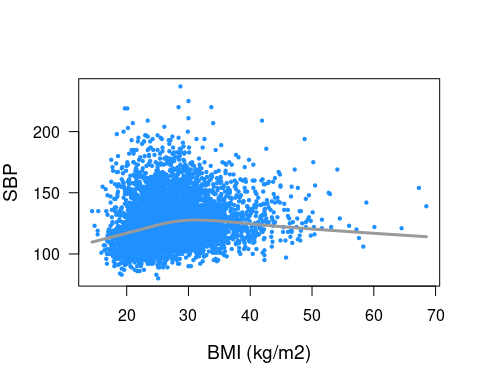
## Does the distribution of log(sbp) look closer to the normal distribution?  
par(mfrow = c(2, 1))  
hist(nhanes$sbp, nclass = 20, col = "darkorchid")  
hist(log(nhanes$sbp), nclass = 20, col = "seagreen3")



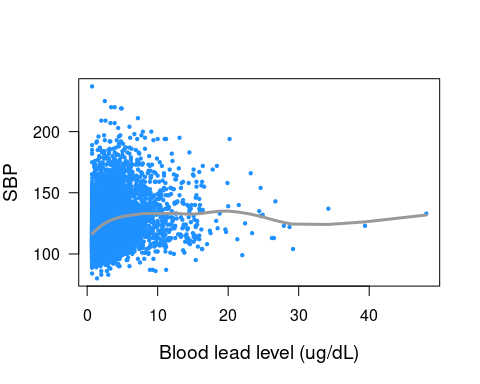
## Let's start with non-log transformed SBP first. Look at bivariate association between sbp and continuous covariates  
par(mfrow = c(1, 1))  
plot(nhanes$age, nhanes$sbp, pch = 20, cex = 0.7, col = "dodgerblue", cex.lab = 1.2, las = 1, ylab = "SBP", xlab = "Age (years)")  
lines(smooth.spline(nhanes$age, nhanes$sbp, df = 10), col = "dodgerblue", lwd = 3)



plot(nhanes$bmi, nhanes$sbp, pch = 20, cex = 0.7, col = "dodgerblue", cex.lab = 1.2, las = 1, ylab = "SBP", xlab = "BMI (kg/m2)")  
lines(smooth.spline(na.omit(nhanes$bmi), nhanes$sbp[na.omit(nhanes$bmi)], df = 3, ), col = "grey60", lwd = 3)



plot(nhanes$bpb, nhanes$sbp, pch = 20, cex = 0.7, col = "dodgerblue", cex.lab = 1.2, las = 1, ylab = "SBP", xlab = "Blood lead level (ug/dL)")  
lines(smooth.spline(nhanes$bpb, nhanes$sbp, df = 10), col = "grey60", lwd = 3)



# Simple linear regression

## Start creating a simple regression model for systolic blood pressure,  
## including only blood lead (bpb) (crude model).  
sbp.model <- lm(sbp ~ bpb, na.action = na.omit, data = nhanes)  
sbp.model

##   
## Call:  
## lm(formula = sbp ~ bpb, data = nhanes, na.action = na.omit)  
##   
## Coefficients:  
## (Intercept) bpb   
## 121.666 1.162

summary(sbp.model) # Is blood lead associated with SBP? Is this the direction you would expect?

##   
## Call:  
## lm(formula = sbp ~ bpb, data = nhanes, na.action = na.omit)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -51.586 -13.964 -3.757 10.612 114.521   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 121.66581 0.43552 279.36 <2e-16 \*\*\*  
## bpb 1.16166 0.08528 13.62 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 19.6 on 5072 degrees of freedom  
## Multiple R-squared: 0.03529, Adjusted R-squared: 0.0351   
## F-statistic: 185.6 on 1 and 5072 DF, p-value: < 2.2e-16

summary.aov(sbp.model)

## Df Sum Sq Mean Sq F value Pr(>F)   
## bpb 1 71318 71318 185.6 <2e-16 \*\*\*  
## Residuals 5072 1949387 384   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

anova(sbp.model)

## Analysis of Variance Table  
##   
## Response: sbp  
## Df Sum Sq Mean Sq F value Pr(>F)   
## bpb 1 71318 71318 185.56 < 2.2e-16 \*\*\*  
## Residuals 5072 1949387 384   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

## Add age in the model  
sbp.model1 <- lm(sbp ~ bpb + age, na.action = na.omit, data = nhanes)  
summary(sbp.model1) # Does anything change with age in the model?

##   
## Call:  
## lm(formula = sbp ~ bpb + age, data = nhanes, na.action = na.omit)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -62.405 -10.431 -1.427 8.470 101.240   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 95.00660 0.63062 150.655 < 2e-16 \*\*\*  
## bpb 0.46641 0.07063 6.604 4.41e-11 \*\*\*  
## age 0.60338 0.01181 51.077 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 15.93 on 5071 degrees of freedom  
## Multiple R-squared: 0.363, Adjusted R-squared: 0.3628   
## F-statistic: 1445 on 2 and 5071 DF, p-value: < 2.2e-16

# Add categorical variables linear regression

## Add race, which should be a categorical variable. We can use factor() or as.factor() to create indicator variables for each value of race  
table(nhanes$race)

##   
## 1 2   
## 3634 1440

sbp.model2 <- lm(sbp ~ bpb + age + factor(race), na.action = na.omit, data = nhanes)  
summary(sbp.model2)

##   
## Call:  
## lm(formula = sbp ~ bpb + age + factor(race), data = nhanes, na.action = na.omit)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -61.773 -10.420 -1.257 8.525 101.680   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 93.88241 0.66373 141.446 < 2e-16 \*\*\*  
## bpb 0.42848 0.07080 6.052 1.53e-09 \*\*\*  
## age 0.61400 0.01195 51.378 < 2e-16 \*\*\*  
## factor(race)2 2.66690 0.50308 5.301 1.20e-07 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 15.89 on 5070 degrees of freedom  
## Multiple R-squared: 0.3665, Adjusted R-squared: 0.3661   
## F-statistic: 977.8 on 3 and 5070 DF, p-value: < 2.2e-16

#### Change the reference level for a factor variable  
# Method 1: Make a new (or alter the old) variable  
table(nhanes$race)

##   
## 1 2   
## 3634 1440

race2 <- relevel(factor(nhanes$race), ref = 2)  
table(race2)

## race2  
## 2 1   
## 1440 3634

sbp.model2 <- lm(sbp ~ bpb + age + factor(race2), na.action = na.omit, data = nhanes)  
summary(sbp.model2)

##   
## Call:  
## lm(formula = sbp ~ bpb + age + factor(race2), data = nhanes,   
## na.action = na.omit)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -61.773 -10.420 -1.257 8.525 101.680   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 96.54931 0.69301 139.319 < 2e-16 \*\*\*  
## bpb 0.42848 0.07080 6.052 1.53e-09 \*\*\*  
## age 0.61400 0.01195 51.378 < 2e-16 \*\*\*  
## factor(race2)1 -2.66690 0.50308 -5.301 1.20e-07 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 15.89 on 5070 degrees of freedom  
## Multiple R-squared: 0.3665, Adjusted R-squared: 0.3661   
## F-statistic: 977.8 on 3 and 5070 DF, p-value: < 2.2e-16

# Method 2: Change the contrasts in the lm() statement  
sbp.model2 <- lm(sbp ~ bpb + age + C(factor(race), contr.treatment(2, base = 2)), na.action = na.omit, data = nhanes)  
summary(sbp.model2)

##   
## Call:  
## lm(formula = sbp ~ bpb + age + C(factor(race), contr.treatment(2,   
## base = 2)), data = nhanes, na.action = na.omit)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -61.773 -10.420 -1.257 8.525 101.680   
##   
## Coefficients:  
## Estimate Std. Error t value  
## (Intercept) 96.54931 0.69301 139.319  
## bpb 0.42848 0.07080 6.052  
## age 0.61400 0.01195 51.378  
## C(factor(race), contr.treatment(2, base = 2))1 -2.66690 0.50308 -5.301  
## Pr(>|t|)   
## (Intercept) < 2e-16 \*\*\*  
## bpb 1.53e-09 \*\*\*  
## age < 2e-16 \*\*\*  
## C(factor(race), contr.treatment(2, base = 2))1 1.20e-07 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 15.89 on 5070 degrees of freedom  
## Multiple R-squared: 0.3665, Adjusted R-squared: 0.3661   
## F-statistic: 977.8 on 3 and 5070 DF, p-value: < 2.2e-16

# R provides Type I sequential SS, not the default Type III marginal SS reported by SAS and SPSS.  
# We can use the drop1() function to produce the familiar Type III results.  
# It will compare each term with the full model.  
anova(sbp.model2)

## Analysis of Variance Table  
##   
## Response: sbp  
## Df Sum Sq Mean Sq F value  
## bpb 1 71318 71318 282.470  
## age 1 662217 662217 2622.848  
## C(factor(race), contr.treatment(2, base = 2)) 1 7095 7095 28.102  
## Residuals 5070 1280075 252   
## Pr(>F)   
## bpb < 2.2e-16 \*\*\*  
## age < 2.2e-16 \*\*\*  
## C(factor(race), contr.treatment(2, base = 2)) 1.199e-07 \*\*\*  
## Residuals   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

drop1(sbp.model2, test = "F")

## Single term deletions  
##   
## Model:  
## sbp ~ bpb + age + C(factor(race), contr.treatment(2, base = 2))  
## Df Sum of Sq RSS AIC  
## <none> 1280075 28070  
## bpb 1 9247 1289322 28104  
## age 1 666469 1946544 30195  
## C(factor(race), contr.treatment(2, base = 2)) 1 7095 1287170 28096  
## F value Pr(>F)   
## <none>   
## bpb 36.625 1.535e-09 \*\*\*  
## age 2639.688 < 2.2e-16 \*\*\*  
## C(factor(race), contr.treatment(2, base = 2)) 28.102 1.199e-07 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

# Compare regression models

## Add other covariates to the model: which variables are biologically important?  
## Add sex, BMI, educ and smk.  
sbp.model3 <- lm(sbp ~ bpb + age + factor(race) + factor(sex) + bmi + factor(educ) + factor(smk), na.action = na.omit, data = nhanes)  
summary(sbp.model3) # What covariates are associated with SBP? Do they make sense biologically?

##   
## Call:  
## lm(formula = sbp ~ bpb + age + factor(race) + factor(sex) + bmi +   
## factor(educ) + factor(smk), data = nhanes, na.action = na.omit)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -59.202 -9.583 -1.317 7.926 101.116   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 83.160324 1.346404 61.765 < 2e-16 \*\*\*  
## bpb 0.277763 0.075759 3.666 0.000249 \*\*\*  
## age 0.615200 0.012258 50.189 < 2e-16 \*\*\*  
## factor(race)2 2.024195 0.498277 4.062 4.93e-05 \*\*\*  
## factor(sex)2 -3.912620 0.476734 -8.207 2.85e-16 \*\*\*  
## bmi 0.530437 0.038131 13.911 < 2e-16 \*\*\*  
## factor(educ)2 -0.003008 0.484418 -0.006 0.995046   
## factor(educ)3 -2.675500 0.711182 -3.762 0.000170 \*\*\*  
## factor(smk)2 -1.991441 0.562961 -3.537 0.000408 \*\*\*  
## factor(smk)3 -0.270575 0.559123 -0.484 0.628459   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 15.45 on 5021 degrees of freedom  
## (43 observations deleted due to missingness)  
## Multiple R-squared: 0.3977, Adjusted R-squared: 0.3966   
## F-statistic: 368.4 on 9 and 5021 DF, p-value: < 2.2e-16

# See output from models 1,2,and 3 side by side  
stargazer(sbp.model1, sbp.model2, sbp.model3, type = "text", dep.var.labels = "Systolic Blood Pressure (mmHg)")

##   
## ==============================================================================================================================  
## Dependent variable:   
## -------------------------------------------------------------------------------  
## Systolic Blood Pressure (mmHg)   
## (1) (2) (3)   
## ------------------------------------------------------------------------------------------------------------------------------  
## bpb 0.466\*\*\* 0.428\*\*\* 0.278\*\*\*   
## (0.071) (0.071) (0.076)   
##   
## age 0.603\*\*\* 0.614\*\*\* 0.615\*\*\*   
## (0.012) (0.012) (0.012)   
##   
## C(factor(race), contr.treatment(2, base = 2))1 -2.667\*\*\*   
## (0.503)   
##   
## factor(race)2 2.024\*\*\*   
## (0.498)   
##   
## factor(sex)2 -3.913\*\*\*   
## (0.477)   
##   
## bmi 0.530\*\*\*   
## (0.038)   
##   
## factor(educ)2 -0.003   
## (0.484)   
##   
## factor(educ)3 -2.675\*\*\*   
## (0.711)   
##   
## factor(smk)2 -1.991\*\*\*   
## (0.563)   
##   
## factor(smk)3 -0.271   
## (0.559)   
##   
## Constant 95.007\*\*\* 96.549\*\*\* 83.160\*\*\*   
## (0.631) (0.693) (1.346)   
##   
## ------------------------------------------------------------------------------------------------------------------------------  
## Observations 5,074 5,074 5,031   
## R2 0.363 0.367 0.398   
## Adjusted R2 0.363 0.366 0.397   
## Residual Std. Error 15.932 (df = 5071) 15.890 (df = 5070) 15.445 (df = 5021)   
## F Statistic 1,444.936\*\*\* (df = 2; 5071) 977.807\*\*\* (df = 3; 5070) 368.374\*\*\* (df = 9; 5021)  
## ==============================================================================================================================  
## Note: \*p<0.1; \*\*p<0.05; \*\*\*p<0.01

### Check if alcohol consumption is a confounder  
sbp.model4 <- lm(sbp ~ bpb + age + factor(race) + factor(sex) + bmi + factor(educ) + factor(smk) + factor(alc), na.action = na.omit, data = nhanes)  
summary(sbp.model4) # Is alcohol associated with sbp?

##   
## Call:  
## lm(formula = sbp ~ bpb + age + factor(race) + factor(sex) + bmi +   
## factor(educ) + factor(smk) + factor(alc), data = nhanes,   
## na.action = na.omit)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -58.495 -9.580 -1.311 7.928 101.235   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 82.33776 1.41399 58.231 < 2e-16 \*\*\*  
## bpb 0.24874 0.07705 3.228 0.001253 \*\*   
## age 0.61361 0.01276 48.106 < 2e-16 \*\*\*  
## factor(race)2 2.00975 0.50657 3.967 7.37e-05 \*\*\*  
## factor(sex)2 -3.87208 0.49390 -7.840 5.50e-15 \*\*\*  
## bmi 0.55795 0.03860 14.454 < 2e-16 \*\*\*  
## factor(educ)2 0.02761 0.49227 0.056 0.955279   
## factor(educ)3 -2.71562 0.72412 -3.750 0.000179 \*\*\*  
## factor(smk)2 -2.07944 0.56937 -3.652 0.000263 \*\*\*  
## factor(smk)3 -0.20883 0.57141 -0.365 0.714779   
## factor(alc)1 0.41798 0.49359 0.847 0.397144   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 15.38 on 4891 degrees of freedom  
## (172 observations deleted due to missingness)  
## Multiple R-squared: 0.3967, Adjusted R-squared: 0.3954   
## F-statistic: 321.6 on 10 and 4891 DF, p-value: < 2.2e-16

summary(sbp.model4)$coef # Pull out relevant information from the output: Coefficient table

## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 82.3377552 1.41399277 58.23067609 0.000000e+00  
## bpb 0.2487418 0.07704799 3.22840133 1.253061e-03  
## age 0.6136075 0.01275532 48.10601736 0.000000e+00  
## factor(race)2 2.0097527 0.50656503 3.96741297 7.369686e-05  
## factor(sex)2 -3.8720755 0.49390486 -7.83971931 5.502517e-15  
## bmi 0.5579520 0.03860066 14.45446563 2.100456e-46  
## factor(educ)2 0.0276076 0.49227488 0.05608169 9.552790e-01  
## factor(educ)3 -2.7156205 0.72411551 -3.75025867 1.786946e-04  
## factor(smk)2 -2.0794440 0.56937030 -3.65218209 2.627418e-04  
## factor(smk)3 -0.2088304 0.57140734 -0.36546680 7.147788e-01  
## factor(alc)1 0.4179774 0.49359253 0.84680655 3.971444e-01

summary(sbp.model4)$coef[2, 1] # Pull out just the beta coefficient blood Pb

## [1] 0.2487418

# 10% guideline for confounders: Does the addition of the new variable change the beta coefficient of interest by >10%?  
(summary(sbp.model4)$coef[2, 1] - summary(sbp.model3)$coef[2, 1]) / summary(sbp.model3)$coef[2, 1] # Calculate the percent change in the blood Pb coefficient before and after alcohol in the model

## [1] -0.1044827

# Does alcohol meet the guideline for a confounder?

# Reporting regression results for untransformed linear data

## Compute difference in SBP and 95% CI per one unit increase and IQR increase in bpb  
# For one unit increase in exposure  
summary(sbp.model4)$coef[2, 1] # Beta coefficient

## [1] 0.2487418

summary(sbp.model4)$coef[2, 1] - 1.96 \* summary(sbp.model4)$coef[2, 2] # Lower confidence interval

## [1] 0.09772778

summary(sbp.model4)$coef[2, 1] + 1.96 \* summary(sbp.model4)$coef[2, 2] # Upper confidence interval

## [1] 0.3997559

regress.display(sbp.model4) # Alternative: Convenience function to show all of the CI with the beta coefficients

##   
## Coeff lower095ci upper095ci Pr>|t|  
## (Intercept) 82.3377552 79.56569429 85.1098161 0.000000e+00  
## bpb 0.2487418 0.09769317 0.3997905 1.253061e-03  
## age 0.6136075 0.58860131 0.6386136 0.000000e+00  
## factor(race)2 2.0097527 1.01665770 3.0028476 7.369686e-05  
## factor(sex)2 -3.8720755 -4.84035080 -2.9038001 5.502517e-15  
## bmi 0.5579520 0.48227733 0.6336266 2.100456e-46  
## factor(educ)2 0.0276076 -0.93747225 0.9926875 9.552790e-01  
## factor(educ)3 -2.7156205 -4.13521210 -1.2960289 1.786946e-04  
## factor(smk)2 -2.0794440 -3.19566554 -0.9632225 2.627418e-04  
## factor(smk)3 -0.2088304 -1.32904543 0.9113846 7.147788e-01  
## factor(alc)1 0.4179774 -0.54968566 1.3856404 3.971444e-01

# To improve communication of findings, report association for an IQR increase in exposure   
IQR(nhanes$bpb)

## [1] 3.1

IQR(nhanes$bpb) \* summary(sbp.model4)$coef[2, 1] # Per IQR increase in exposure, beta coefficient

## [1] 0.7710997

l95ci.iqr <- IQR(nhanes$bpb) \* (summary(sbp.model4)$coef[2, 1] - 1.96 \* summary(sbp.model4)$coef[2, 2]) # Per IQR increase in exposure, lower confidence interval  
u95ci.iqr <- IQR(nhanes$bpb) \* (summary(sbp.model4)$coef[2, 1] + 1.96 \* summary(sbp.model4)$coef[2, 2]) # Per IQR increase in exposure, upper confidence interval  
regress.display(sbp.model4)$table[2, ] \* IQR(nhanes$bpb) # Multiply the convenience table by the IQR

## Coeff lower095ci upper095ci Pr>|t|   
## 0.77109973 0.30284884 1.23935061 0.00388449

# Compare regression results

## Does the variable packyrs give a better estimate than smk?  
sbp.model5 <- lm(sbp ~ bpb + age + factor(race) + factor(sex) + bmi + factor(educ) + packyrs + factor(alc), na.action = na.omit, data = nhanes)  
summary(sbp.model5)

##   
## Call:  
## lm(formula = sbp ~ bpb + age + factor(race) + factor(sex) + bmi +   
## factor(educ) + packyrs + factor(alc), data = nhanes, na.action = na.omit)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -59.679 -9.666 -1.233 8.026 101.502   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 82.04513 1.42080 57.746 < 2e-16 \*\*\*  
## bpb 0.24963 0.07906 3.157 0.001602 \*\*   
## age 0.61464 0.01317 46.658 < 2e-16 \*\*\*  
## factor(race)2 2.22727 0.51892 4.292 1.81e-05 \*\*\*  
## factor(sex)2 -3.70826 0.50585 -7.331 2.68e-13 \*\*\*  
## bmi 0.55073 0.03930 14.012 < 2e-16 \*\*\*  
## factor(educ)2 0.04590 0.50476 0.091 0.927547   
## factor(educ)3 -2.60357 0.73817 -3.527 0.000424 \*\*\*  
## packyrs -0.02462 0.01097 -2.244 0.024860 \*   
## factor(alc)1 0.32565 0.49934 0.652 0.514333   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 15.42 on 4677 degrees of freedom  
## (387 observations deleted due to missingness)  
## Multiple R-squared: 0.3952, Adjusted R-squared: 0.394   
## F-statistic: 339.6 on 9 and 4677 DF, p-value: < 2.2e-16

AIC(sbp.model4)

## [1] 40722.09

AIC(sbp.model5)

## [1] 38957.68

## Which model has the lower AIC? The one with smk (model 4) or the one with packyrs (model 5)?   
## Model 5 also dropped 215 more people. This is an analyst judgment call (no perfect answer). I would probably stick with model 4 since the AIC are close, to keep the people in the study.  
  
## Run two models: One with age and one adding age as a quadratic function.  
## Does the quadratic term improve the fit of the model?  
sbp.model6 <- lm(sbp ~ bpb + age + factor(race) + factor(sex) + bmi + factor(educ) + factor(smk) + factor(alc) + I(age^2), na.action = na.omit, data = nhanes)  
summary(sbp.model6) # Is the age squared term significant?

##   
## Call:  
## lm(formula = sbp ~ bpb + age + factor(race) + factor(sex) + bmi +   
## factor(educ) + factor(smk) + factor(alc) + I(age^2), data = nhanes,   
## na.action = na.omit)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -59.724 -9.518 -1.413 7.689 101.809   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 87.0142129 1.8737032 46.440 < 2e-16 \*\*\*  
## bpb 0.2582957 0.0769837 3.355 0.000799 \*\*\*  
## age 0.3452699 0.0718091 4.808 1.57e-06 \*\*\*  
## factor(race)2 2.0928803 0.5063451 4.133 3.64e-05 \*\*\*  
## factor(sex)2 -3.7960138 0.4936354 -7.690 1.77e-14 \*\*\*  
## bmi 0.5913718 0.0395399 14.956 < 2e-16 \*\*\*  
## factor(educ)2 0.0989887 0.4919604 0.201 0.840541   
## factor(educ)3 -2.4208295 0.7272801 -3.329 0.000879 \*\*\*  
## factor(smk)2 -1.8118191 0.5729428 -3.162 0.001575 \*\*   
## factor(smk)3 0.0969062 0.5762782 0.168 0.866465   
## factor(alc)1 0.4773454 0.4931648 0.968 0.333131   
## I(age^2) 0.0026033 0.0006856 3.797 0.000148 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 15.36 on 4890 degrees of freedom  
## (172 observations deleted due to missingness)  
## Multiple R-squared: 0.3984, Adjusted R-squared: 0.3971   
## F-statistic: 294.4 on 11 and 4890 DF, p-value: < 2.2e-16

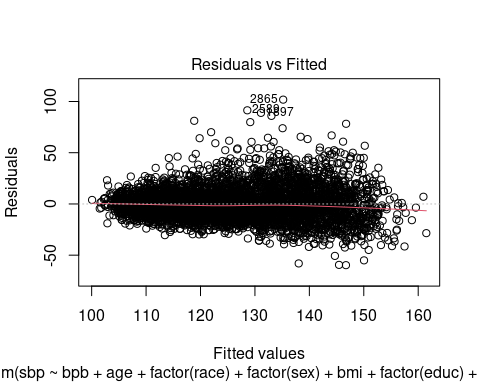
## Use the anova function to compare the 2 models and  
## see if the quadratic term improves the model.  
anova(sbp.model4, sbp.model6, test = "F")

## Analysis of Variance Table  
##   
## Model 1: sbp ~ bpb + age + factor(race) + factor(sex) + bmi + factor(educ) +   
## factor(smk) + factor(alc)  
## Model 2: sbp ~ bpb + age + factor(race) + factor(sex) + bmi + factor(educ) +   
## factor(smk) + factor(alc) + I(age^2)  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 4891 1157606   
## 2 4890 1154203 1 3403 14.418 0.0001482 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

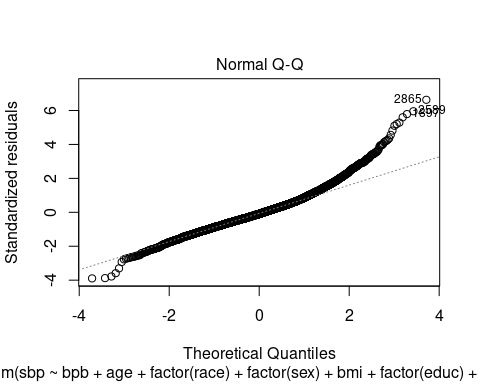
## We'll use sbp.model6 as our final model  
## What is the relationship between bpb and sbp?

# Regression diagnostics

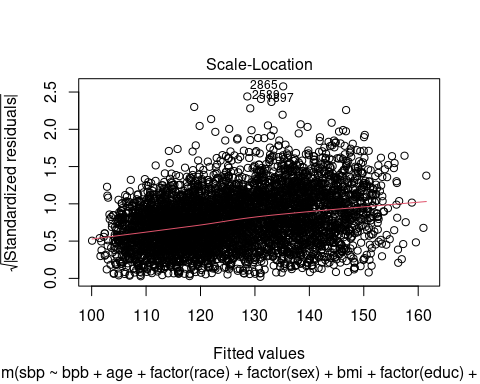
### Regression Diagnostics  
## In the case of linear model, the plot of the model gives diagnostic plots  
plot(sbp.model6, which = 1)



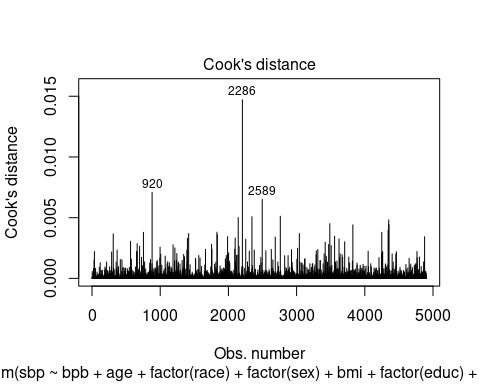
plot(sbp.model6, which = 2)



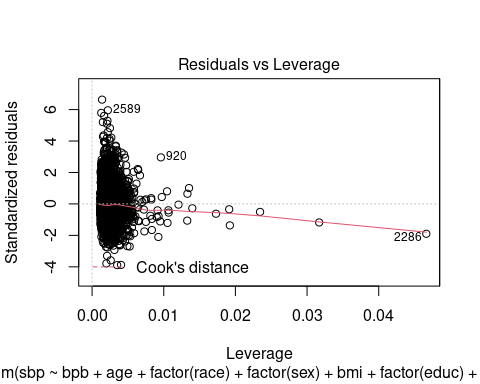
plot(sbp.model6, which = 3)



plot(sbp.model6, which = 4)



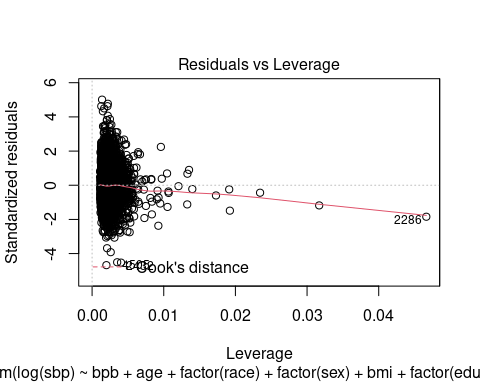
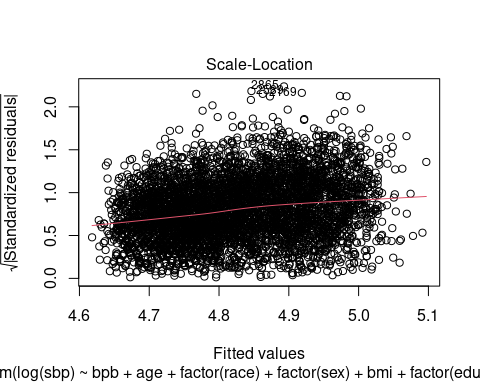
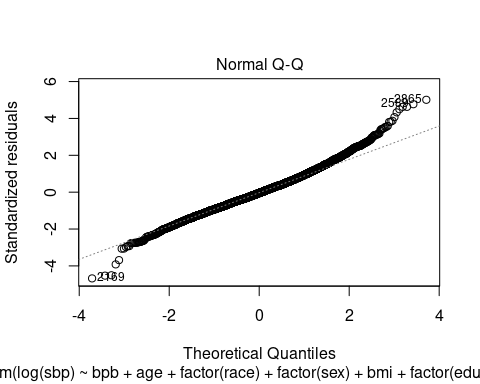
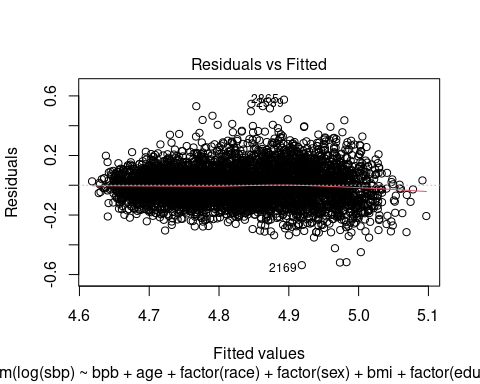
plot(sbp.model6, which = 5)



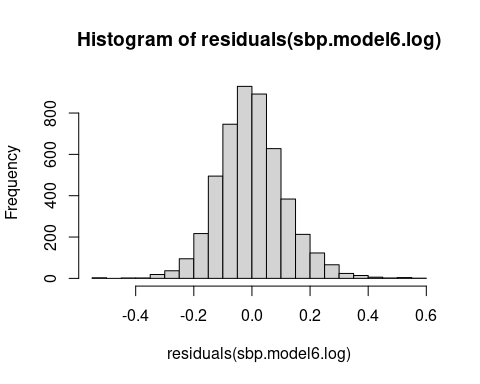
## How about log(sbp)?  
## Check how diagnostic plots using log-transformed sbp look like  
  
sbp.model6.log <- lm(log(sbp) ~ bpb + age + factor(race) + factor(sex) + bmi + factor(educ) + factor(smk) + factor(alc) + I(age^2), na.action = na.omit, data = nhanes)  
summary(sbp.model6.log)

##   
## Call:  
## lm(formula = log(sbp) ~ bpb + age + factor(race) + factor(sex) +   
## bmi + factor(educ) + factor(smk) + factor(alc) + I(age^2),   
## data = nhanes, na.action = na.omit)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.53660 -0.07304 -0.00487 0.06676 0.57499   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 4.505e+00 1.401e-02 321.638 < 2e-16 \*\*\*  
## bpb 1.974e-03 5.755e-04 3.430 0.000609 \*\*\*  
## age 3.281e-03 5.368e-04 6.113 1.06e-09 \*\*\*  
## factor(race)2 1.535e-02 3.785e-03 4.054 5.11e-05 \*\*\*  
## factor(sex)2 -3.504e-02 3.690e-03 -9.496 < 2e-16 \*\*\*  
## bmi 4.864e-03 2.956e-04 16.455 < 2e-16 \*\*\*  
## factor(educ)2 1.928e-03 3.678e-03 0.524 0.600129   
## factor(educ)3 -1.837e-02 5.437e-03 -3.379 0.000734 \*\*\*  
## factor(smk)2 -1.415e-02 4.283e-03 -3.303 0.000964 \*\*\*  
## factor(smk)3 5.033e-04 4.308e-03 0.117 0.906992   
## factor(alc)1 5.098e-03 3.687e-03 1.383 0.166742   
## I(age^2) 1.389e-05 5.125e-06 2.710 0.006758 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.1148 on 4890 degrees of freedom  
## (172 observations deleted due to missingness)  
## Multiple R-squared: 0.4159, Adjusted R-squared: 0.4146   
## F-statistic: 316.5 on 11 and 4890 DF, p-value: < 2.2e-16

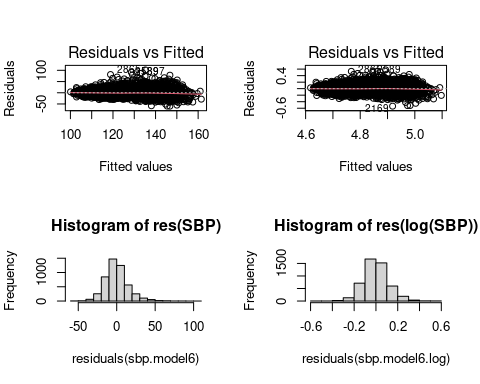
plot(sbp.model6.log)



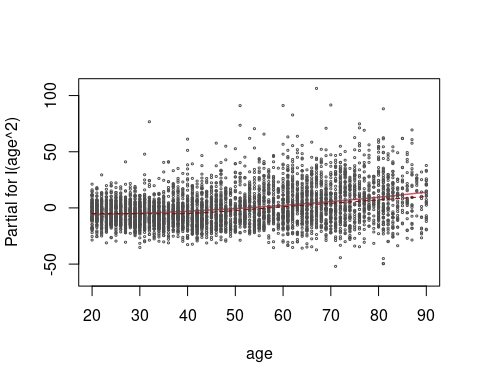
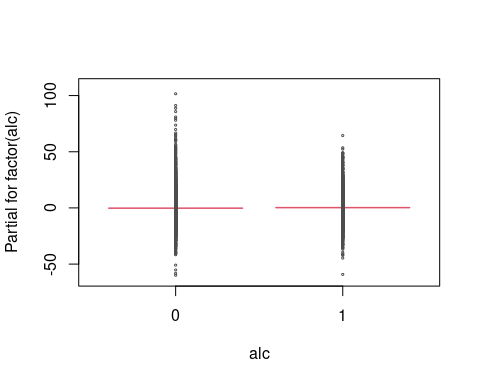
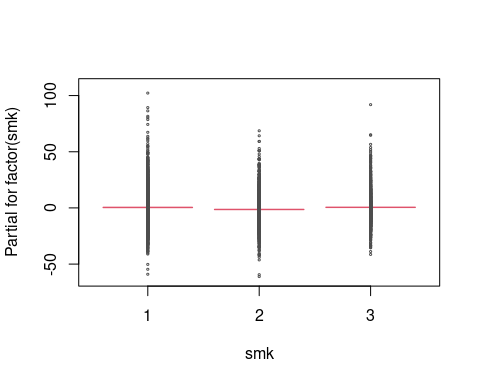
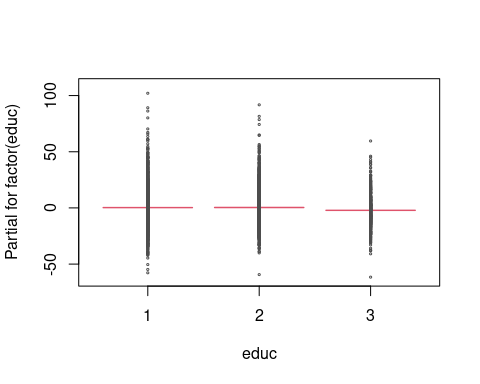
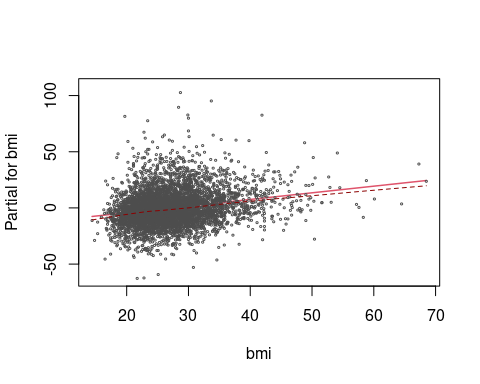
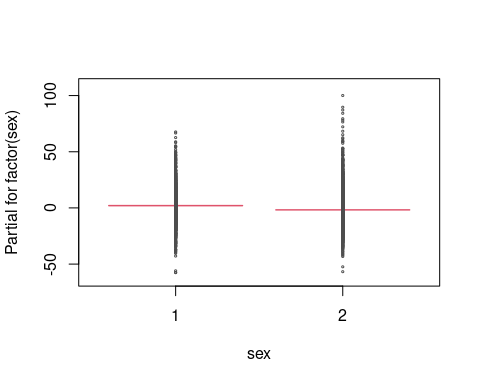
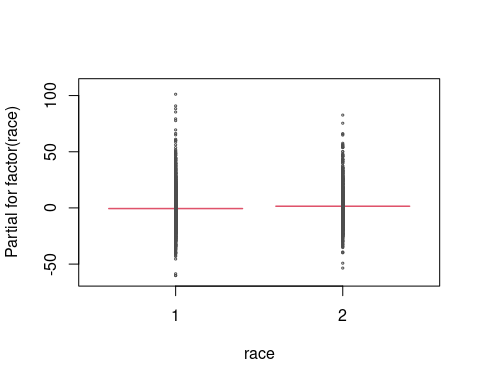
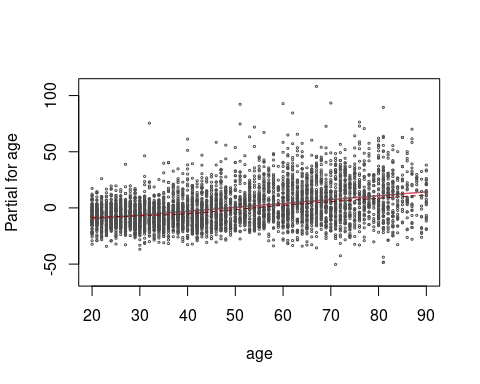
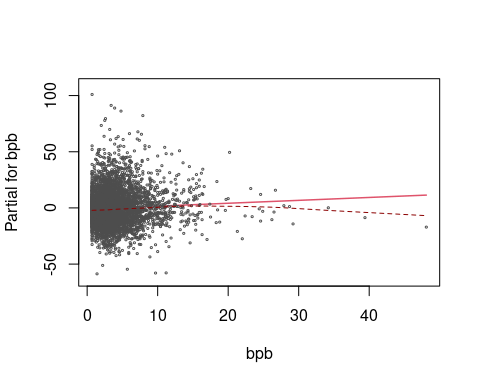
hist(residuals(sbp.model6.log), nclass = 20)



par(mfrow = c(2, 2))  
plot(sbp.model6, which = 1)  
plot(sbp.model6.log, which = 1)  
hist(residuals(sbp.model6), main = "Histogram of res(SBP)")  
hist(residuals(sbp.model6.log), main = "Histogram of res(log(SBP))")



par(mfrow = c(1, 1))  
  
### Partial Residual Plots  
## Plot sbp vs. bpb given that other variables are in the model (adjusted)  
  
## This can be done by 'termplot'  
termplot(sbp.model6, partial.resid = TRUE, col.res = "gray30", cex = 0.3, smooth = panel.smooth)



## Suppose you decide to go with log transformed SBP. What do you report if y is log-transformed?

## Percent increase (difference) vs. absolute increase (difference)

## Compute percent increase in SBP and 95% CI per one unit increase and IQR increase in bpb

summary(sbp.model6.log)$coef

## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 4.504957e+00 1.400631e-02 321.6375700 0.000000e+00  
## bpb 1.973727e-03 5.754686e-04 3.4297737 6.090800e-04  
## age 3.281259e-03 5.367875e-04 6.1127716 1.055187e-09  
## factor(race)2 1.534593e-02 3.785033e-03 4.0543714 5.105082e-05  
## factor(sex)2 -3.504012e-02 3.690025e-03 -9.4959008 3.321831e-21  
## bmi 4.863685e-03 2.955687e-04 16.4553462 2.932179e-59  
## factor(educ)2 1.927926e-03 3.677504e-03 0.5242485 6.001295e-01  
## factor(educ)3 -1.836873e-02 5.436567e-03 -3.3787365 7.338916e-04  
## factor(smk)2 -1.414501e-02 4.282864e-03 -3.3026976 9.644763e-04  
## factor(smk)3 5.033236e-04 4.307797e-03 0.1168401 9.069915e-01  
## factor(alc)1 5.098266e-03 3.686508e-03 1.3829529 1.667424e-01  
## I(age^2) 1.388725e-05 5.125029e-06 2.7096926 6.758042e-03

summary(sbp.model6.log)$coef[2, 1]

## [1] 0.001973727

summary(sbp.model6.log)$coef[2, 2]

## [1] 0.0005754686

exp(summary(sbp.model6.log)$coef[2, 1])

## [1] 1.001976

# exp(sbp.model6.log$coef[2])  
100 \* (exp(summary(sbp.model6.log)$coef[2, 1]) - 1) # Per one unit increase in blood Pb, percent change in systolic blood pressure

## [1] 0.1975676

100 \* (exp(summary(sbp.model6.log)$coef[2, 1] - 1.96 \* summary(sbp.model6.log)$coef[2, 2]) - 1) # Lower confidence interval

## [1] 0.08461663

100 \* (exp(summary(sbp.model6.log)$coef[2, 1] + 1.96 \* summary(sbp.model6.log)$coef[2, 2]) - 1) # Upper confidence interval

## [1] 0.310646

IQR(nhanes$bpb)

## [1] 3.1

100 \* (exp(IQR(nhanes$bpb) \* summary(sbp.model6.log)$coef[2, 1]) - 1) # Per IQR unit increase in blood Pb, percent change in systolic blood pressure

## [1] 0.613731

100 \* (exp(IQR(nhanes$bpb) \* (summary(sbp.model6.log)$coef[2, 1] - 1.96 \* summary(sbp.model6.log)$coef[2, 2])) - 1) # Lower confidence interval

## [1] 0.2625447

100 \* (exp(IQR(nhanes$bpb) \* (summary(sbp.model6.log)$coef[2, 1] + 1.96 \* summary(sbp.model6.log)$coef[2, 2])) - 1) # Upper confidence interval

## [1] 0.9661474

## linearity assumption

# Non-linear associations

# What if the relationship between blood lead and sbp is non-linear?   
library(mgcv)

## Loading required package: nlme

## This is mgcv 1.8-35. For overview type 'help("mgcv-package")'.

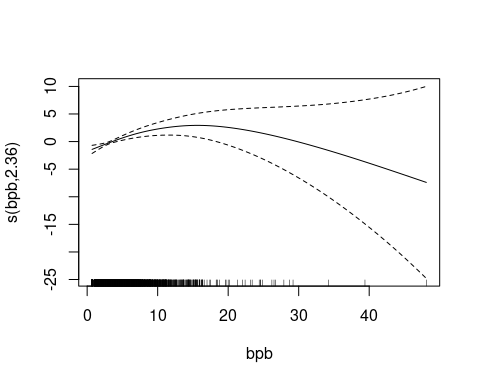
##   
## Attaching package: 'mgcv'

## The following object is masked from 'package:nnet':  
##   
## multinom

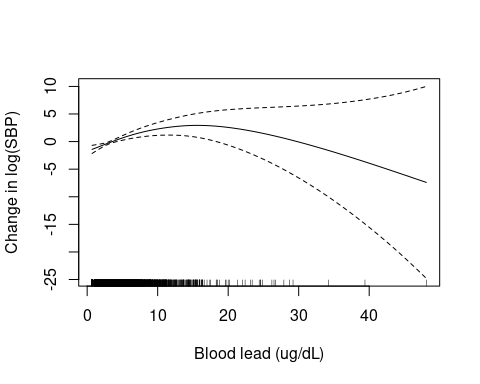
sbp.model6.gam <- gam(sbp ~ s(bpb) + age + factor(race) + factor(sex) + bmi + factor(educ) + factor(smk) + factor(alc) + I(age^2), na.action = na.omit, data = nhanes)  
summary(sbp.model6.gam)

##   
## Family: gaussian   
## Link function: identity   
##   
## Formula:  
## sbp ~ s(bpb) + age + factor(race) + factor(sex) + bmi + factor(educ) +   
## factor(smk) + factor(alc) + I(age^2)  
##   
## Parametric coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 88.2143968 1.8617921 47.381 < 2e-16 \*\*\*  
## age 0.3363437 0.0718470 4.681 2.93e-06 \*\*\*  
## factor(race)2 2.0313316 0.5062947 4.012 6.11e-05 \*\*\*  
## factor(sex)2 -3.5783017 0.5001717 -7.154 9.67e-13 \*\*\*  
## bmi 0.5931173 0.0395125 15.011 < 2e-16 \*\*\*  
## factor(educ)2 0.1625592 0.4919648 0.330 0.741090   
## factor(educ)3 -2.2564780 0.7288554 -3.096 0.001973 \*\*   
## factor(smk)2 -1.8333930 0.5724650 -3.203 0.001371 \*\*   
## factor(smk)3 -0.0462355 0.5780681 -0.080 0.936254   
## factor(alc)1 0.4435799 0.4929664 0.900 0.368262   
## I(age^2) 0.0026407 0.0006852 3.854 0.000118 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Approximate significance of smooth terms:  
## edf Ref.df F p-value   
## s(bpb) 2.36 3.014 6.61 0.000184 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## R-sq.(adj) = 0.398 Deviance explained = 40%  
## GCV = 236.21 Scale est. = 235.57 n = 4902

plot(sbp.model6.gam)



plot(sbp.model6.gam, xlab = "Blood lead (ug/dL)", ylab = "Change in log(SBP)") # Does this relationship look linear?



# Option to log transform the exposure  
sbp.model7 <- lm(sbp ~ log(bpb) + age + factor(race) + factor(sex) + bmi + factor(educ) + factor(smk) + factor(alc) + I(age^2), na.action = na.omit, data = nhanes)  
summary(sbp.model7)

##   
## Call:  
## lm(formula = sbp ~ log(bpb) + age + factor(race) + factor(sex) +   
## bmi + factor(educ) + factor(smk) + factor(alc) + I(age^2),   
## data = nhanes, na.action = na.omit)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -59.464 -9.543 -1.339 7.746 103.034   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 86.8708956 1.8722723 46.399 < 2e-16 \*\*\*  
## log(bpb) 1.4222563 0.3526023 4.034 5.58e-05 \*\*\*  
## age 0.3316940 0.0719935 4.607 4.18e-06 \*\*\*  
## factor(race)2 2.0361714 0.5066138 4.019 5.93e-05 \*\*\*  
## factor(sex)2 -3.5467680 0.5053567 -7.018 2.55e-12 \*\*\*  
## bmi 0.5902534 0.0394943 14.945 < 2e-16 \*\*\*  
## factor(educ)2 0.1417779 0.4914805 0.288 0.77300   
## factor(educ)3 -2.2839018 0.7293924 -3.131 0.00175 \*\*   
## factor(smk)2 -1.8271820 0.5726692 -3.191 0.00143 \*\*   
## factor(smk)3 -0.0474241 0.5795115 -0.082 0.93478   
## factor(alc)1 0.4312724 0.4933101 0.874 0.38203   
## I(age^2) 0.0026700 0.0006858 3.893 0.00010 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 15.36 on 4890 degrees of freedom  
## (172 observations deleted due to missingness)  
## Multiple R-squared: 0.399, Adjusted R-squared: 0.3977   
## F-statistic: 295.2 on 11 and 4890 DF, p-value: < 2.2e-16

summary(sbp.model6)

##   
## Call:  
## lm(formula = sbp ~ bpb + age + factor(race) + factor(sex) + bmi +   
## factor(educ) + factor(smk) + factor(alc) + I(age^2), data = nhanes,   
## na.action = na.omit)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -59.724 -9.518 -1.413 7.689 101.809   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 87.0142129 1.8737032 46.440 < 2e-16 \*\*\*  
## bpb 0.2582957 0.0769837 3.355 0.000799 \*\*\*  
## age 0.3452699 0.0718091 4.808 1.57e-06 \*\*\*  
## factor(race)2 2.0928803 0.5063451 4.133 3.64e-05 \*\*\*  
## factor(sex)2 -3.7960138 0.4936354 -7.690 1.77e-14 \*\*\*  
## bmi 0.5913718 0.0395399 14.956 < 2e-16 \*\*\*  
## factor(educ)2 0.0989887 0.4919604 0.201 0.840541   
## factor(educ)3 -2.4208295 0.7272801 -3.329 0.000879 \*\*\*  
## factor(smk)2 -1.8118191 0.5729428 -3.162 0.001575 \*\*   
## factor(smk)3 0.0969062 0.5762782 0.168 0.866465   
## factor(alc)1 0.4773454 0.4931648 0.968 0.333131   
## I(age^2) 0.0026033 0.0006856 3.797 0.000148 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 15.36 on 4890 degrees of freedom  
## (172 observations deleted due to missingness)  
## Multiple R-squared: 0.3984, Adjusted R-squared: 0.3971   
## F-statistic: 294.4 on 11 and 4890 DF, p-value: < 2.2e-16

## compare model6 and model7  
AIC(sbp.model6, sbp.model7) # Which model has the lower AIC? The spline term or the log term for blood Pb?

## df AIC  
## sbp.model6 13 40709.65  
## sbp.model7 13 40704.64

# Effect modification by sex

# What if the relationship between blood Pb and sbp varies by sex?  
  
### Try interaction model  
sbp.model6.int <- lm(sbp ~ bpb + factor(sex) + bpb \* factor(sex) + age + factor(race) + bmi + factor(educ)  
 + factor(smk) + factor(alc) + I(age^2), data = nhanes)  
# below is the same  
sbp.model6.int <- lm(sbp ~ bpb \* factor(sex) + age + factor(race) + bmi + factor(educ)  
 + factor(smk) + factor(alc) + I(age^2), data = nhanes)  
summary(sbp.model6.int) # Is the interaction term significant?

##   
## Call:  
## lm(formula = sbp ~ bpb \* factor(sex) + age + factor(race) + bmi +   
## factor(educ) + factor(smk) + factor(alc) + I(age^2), data = nhanes)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -58.894 -9.530 -1.369 7.642 102.461   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 87.8854794 1.9054676 46.123 < 2e-16 \*\*\*  
## bpb 0.1317420 0.0923617 1.426 0.153825   
## factor(sex)2 -5.1875500 0.7476820 -6.938 4.49e-12 \*\*\*  
## age 0.3385020 0.0718234 4.713 2.51e-06 \*\*\*  
## factor(race)2 2.0885416 0.5060825 4.127 3.74e-05 \*\*\*  
## bmi 0.5905818 0.0395204 14.944 < 2e-16 \*\*\*  
## factor(educ)2 0.1078127 0.4917151 0.219 0.826458   
## factor(educ)3 -2.4434079 0.7269556 -3.361 0.000782 \*\*\*  
## factor(smk)2 -1.7832085 0.5727587 -3.113 0.001860 \*\*   
## factor(smk)3 0.1024602 0.5759802 0.178 0.858818   
## factor(alc)1 0.4787710 0.4929064 0.971 0.331436   
## I(age^2) 0.0026405 0.0006854 3.853 0.000118 \*\*\*  
## bpb:factor(sex)2 0.3796814 0.1532848 2.477 0.013284 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 15.36 on 4889 degrees of freedom  
## (172 observations deleted due to missingness)  
## Multiple R-squared: 0.3992, Adjusted R-squared: 0.3977   
## F-statistic: 270.7 on 12 and 4889 DF, p-value: < 2.2e-16

### Stratified by sex  
table(nhanes$sex)

##   
## 1 2   
## 2335 2739

## Male (sex==1)  
sbp.model6.male <- lm(sbp ~ bpb + age + factor(race) + bmi + factor(educ)  
 + factor(smk) + factor(alc) + I(age^2), data = nhanes, subset = (sex == 1))  
summary(sbp.model6.male) # What is the beta coefficient for blood Pb in males?

##   
## Call:  
## lm(formula = sbp ~ bpb + age + factor(race) + bmi + factor(educ) +   
## factor(smk) + factor(alc) + I(age^2), data = nhanes, subset = (sex ==   
## 1))  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -55.951 -9.164 -1.214 7.452 67.763   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 94.0698502 2.7370216 34.369 < 2e-16 \*\*\*  
## bpb 0.2201841 0.0907801 2.425 0.015367 \*   
## age 0.0247105 0.1020797 0.242 0.808748   
## factor(race)2 2.6271948 0.7251217 3.623 0.000298 \*\*\*  
## bmi 0.7190556 0.0667827 10.767 < 2e-16 \*\*\*  
## factor(educ)2 -0.1580127 0.6931503 -0.228 0.819696   
## factor(educ)3 -1.9583250 0.9739884 -2.011 0.044485 \*   
## factor(smk)2 -0.2674489 0.7894817 -0.339 0.734818   
## factor(smk)3 1.5625824 0.8126909 1.923 0.054641 .   
## factor(alc)1 -0.1923135 0.6709470 -0.287 0.774422   
## I(age^2) 0.0043354 0.0009734 4.454 8.84e-06 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 14.6 on 2240 degrees of freedom  
## (84 observations deleted due to missingness)  
## Multiple R-squared: 0.3134, Adjusted R-squared: 0.3103   
## F-statistic: 102.3 on 10 and 2240 DF, p-value: < 2.2e-16

## Female (sex==2)  
sbp.model6.female <- lm(sbp ~ bpb + age + factor(race) + bmi + factor(educ)  
 + factor(smk) + factor(alc) + I(age^2), data = nhanes, subset = (sex == 2))  
summary(sbp.model6.female) # What is the beta coefficient for blood Pb in females? Is it similar to that in males?

##   
## Call:  
## lm(formula = sbp ~ bpb + age + factor(race) + bmi + factor(educ) +   
## factor(smk) + factor(alc) + I(age^2), data = nhanes, subset = (sex ==   
## 2))  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -57.975 -9.454 -1.348 7.659 99.869   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 77.0605375 2.4865167 30.991 < 2e-16 \*\*\*  
## bpb 0.2931960 0.1343263 2.183 0.02914 \*   
## age 0.5576526 0.0990458 5.630 1.99e-08 \*\*\*  
## factor(race)2 2.0982646 0.6949672 3.019 0.00256 \*\*   
## bmi 0.5305908 0.0492797 10.767 < 2e-16 \*\*\*  
## factor(educ)2 0.3947007 0.6828269 0.578 0.56329   
## factor(educ)3 -2.1735912 1.0641023 -2.043 0.04119 \*   
## factor(smk)2 -1.4335988 0.8458868 -1.695 0.09023 .   
## factor(smk)3 -0.5512632 0.8104050 -0.680 0.49642   
## factor(alc)1 0.6533917 0.7063162 0.925 0.35501   
## I(age^2) 0.0016206 0.0009452 1.715 0.08655 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 15.66 on 2640 degrees of freedom  
## (88 observations deleted due to missingness)  
## Multiple R-squared: 0.4611, Adjusted R-squared: 0.4591   
## F-statistic: 225.9 on 10 and 2640 DF, p-value: < 2.2e-16

Exercise 6A

# This is your last graded homework assignment for the semester.

### 6.3. Generalized Linear Models: Association between hypertension (htn) and blood lead (bpb)

## Logistic regression for hypertension  
## Look at hypertension status (htn)  
tab1(nhanes$htn, graph = F)

## nhanes$htn :   
## Frequency Percent Cum. percent  
## 0 3135 61.8 61.8  
## 1 1939 38.2 100.0  
## Total 5074 100.0 100.0

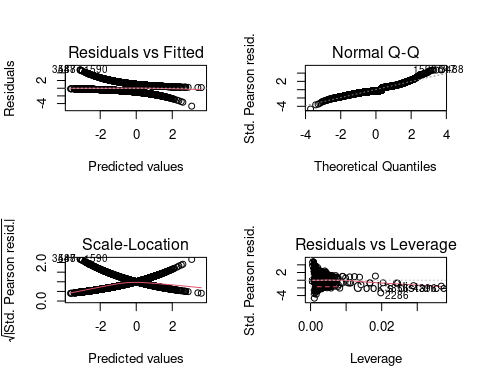
htn.model <- glm(htn ~ bpb + age + factor(sex) + factor(race) + bmi + factor(educ) + factor(smk) + factor(alc),  
 family = binomial,  
 na.action = na.omit, data = nhanes)  
summary(htn.model)

##   
## Call:  
## glm(formula = htn ~ bpb + age + factor(sex) + factor(race) +   
## bmi + factor(educ) + factor(smk) + factor(alc), family = binomial,   
## data = nhanes, na.action = na.omit)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.4998 -0.7949 -0.4272 0.8959 2.5147   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -6.635396 0.261535 -25.371 < 2e-16 \*\*\*  
## bpb 0.023959 0.011750 2.039 0.0414 \*   
## age 0.061698 0.002235 27.602 < 2e-16 \*\*\*  
## factor(sex)2 0.063134 0.077426 0.815 0.4148   
## factor(race)2 0.494599 0.080020 6.181 6.37e-10 \*\*\*  
## bmi 0.097071 0.006285 15.446 < 2e-16 \*\*\*  
## factor(educ)2 0.075786 0.076618 0.989 0.3226   
## factor(educ)3 -0.085220 0.114171 -0.746 0.4554   
## factor(smk)2 -0.037067 0.086530 -0.428 0.6684   
## factor(smk)3 0.080902 0.091358 0.886 0.3759   
## factor(alc)1 0.008554 0.077201 0.111 0.9118   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 6500.8 on 4901 degrees of freedom  
## Residual deviance: 5080.9 on 4891 degrees of freedom  
## (172 observations deleted due to missingness)  
## AIC: 5102.9  
##   
## Number of Fisher Scoring iterations: 4

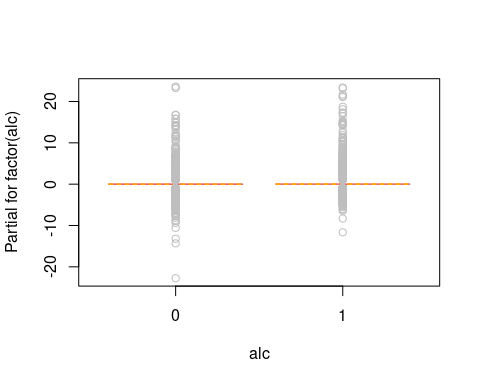
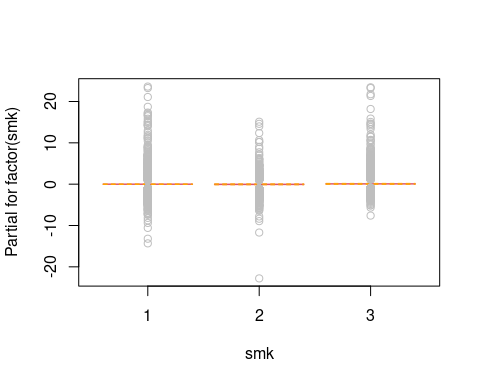
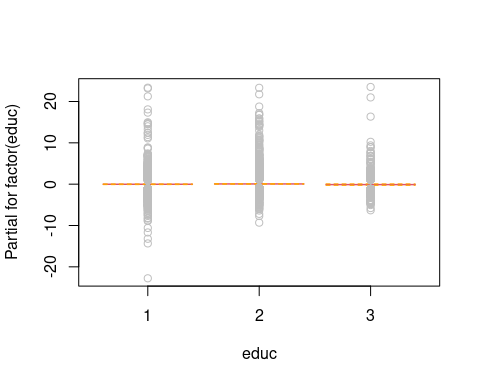
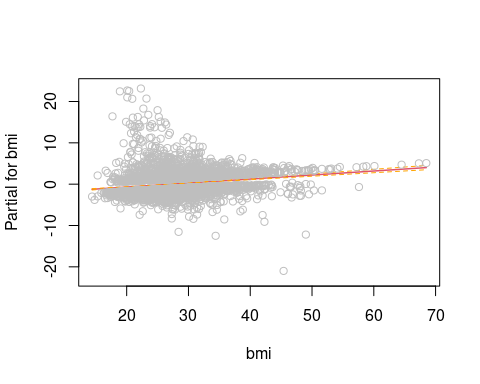
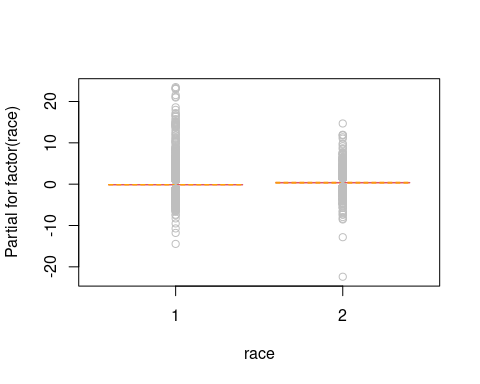
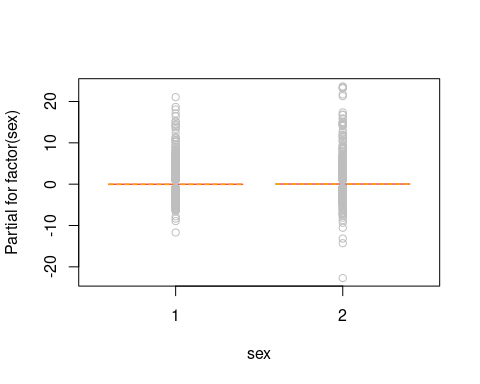
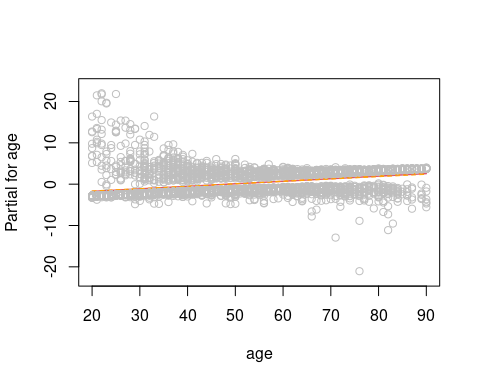
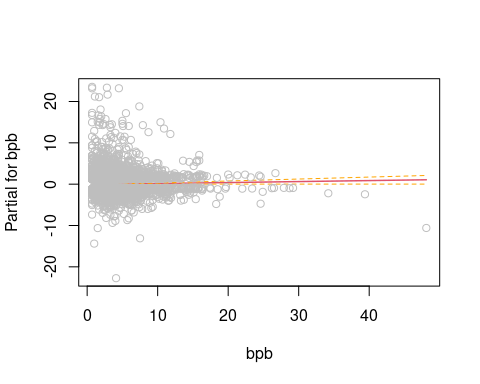
# Compute ORs  
logistic.display(htn.model)

##   
## OR lower95ci upper95ci Pr(>|Z|)  
## bpb 1.0242485 1.0009304 1.048110 4.143898e-02  
## age 1.0636409 1.0589913 1.068311 1.040118e-167  
## factor(sex)2 1.0651698 0.9151960 1.239720 4.148329e-01  
## factor(race)2 1.6398400 1.4018069 1.918292 6.372658e-10  
## bmi 1.1019384 1.0884485 1.115596 8.036933e-54  
## factor(educ)2 1.0787320 0.9283169 1.253519 3.225922e-01  
## factor(educ)3 0.9183104 0.7341876 1.148609 4.554123e-01  
## factor(smk)2 0.9636111 0.8132928 1.141712 6.683781e-01  
## factor(smk)3 1.0842650 0.9065078 1.296879 3.758574e-01  
## factor(alc)1 1.0085903 0.8669650 1.173351 9.117772e-01

# Regression diagnostics  
par(mfrow = c(2, 2))  
plot(htn.model)



par(mfrow = c(1, 1))  
termplot(htn.model, se = T, partial.resid = T)



# Stepwise variable selection: Machine learning

# Poisson Regression for count outcomes.

# Code is available, but we will skip going through this during class

# ## Poisson regression for respiratory death: Montana dataset from epiDisplay  
# data(Montana)  
# summ(Montana)  
# head(Montana, 10)  
# hist(Montana$respdeath)  
#   
# par(mfrow = c(2, 2))  
# tab1(Montana$agegr)  
# tab1(Montana$period)  
# tab1(Montana$start)  
# tab1(Montana$arsenic)  
#   
# ## label categorical variables  
# Montana$agegr <- factor(Montana$agegr, labels = c("40-49", "50-59", "60-69", "70-79"))  
# Montana$period <- factor(Montana$period, labels = c("1938-1949", "1950-1959", "1960-1969", "1970-1977"))  
# Montana$start <- factor(Montana$start, labels = c("pre-1925", "1925 & after"))  
# Montana$arsenic <- factor(Montana$arsenic, labels = c("<1 year", "1-4 years", "5-14 years", "15+ years"))  
#   
# tab1(Montana$agegr, missing = F)  
# tab1(Montana$period, missing = F)  
# tab1(Montana$start, missing = F)  
# tab1(Montana$arsenic, missing = F)  
# par(mfrow = c(1, 1))  
#   
# ## Compute incidence rate by age and period  
# table.pyears <- tapply(Montana$personyrs, list(Montana$period, Montana$agegr), sum)  
# table.deaths <- tapply(Montana$respdeath, list(Montana$period, Montana$agegr), sum)  
# table.inc10000 <- table.deaths / table.pyears \* 10000  
# table.inc10000  
#   
# ## create a time-series plot of the incidence  
# plot.ts(table.inc10000, plot.type = "single", xlab = "", ylab = "#/10,000 person-years", xaxt = "n", col = c("black", "blue", "red", "green"), lty = c(2, 1, 1, 2), las = 1)  
# points(rep(1:4, 4), table.inc10000, pch = 22, cex = table.pyears / sum(table.pyears) \* 20)  
# title(main = "Incidence by age and period")  
# axis(side = 1, at = 1:4, labels = levels(Montana$period))  
# legend("topleft", legend = levels(Montana$agegr)[4:1], col = c("green", "red", "blue", "black"), bg = "white", lty = c(2, 1, 1, 2))  
#   
# ## check arsenic  
# tab1(Montana$arsenic)  
# tapply(Montana$respdeath, Montana$arsenic, mean)  
# tapply(Montana$personyrs, Montana$arsenic, mean)  
#   
# ## Poisson model  
# resp.mode11 <- glm(respdeath ~ period, offset = log(personyrs), family = poisson, data = Montana)  
# summary(resp.mode11)  
#   
# resp.mode12 <- glm(respdeath ~ agegr, offset = log(personyrs), family = poisson, data = Montana)  
# summary(resp.mode12)  
#   
# resp.mode13 <- glm(respdeath ~ period + agegr, offset = log(personyrs), family = poisson, data = Montana)  
# summary(resp.mode13)  
#   
# AIC(resp.mode11, resp.mode12, resp.mode13)  
# ## model2 is better  
#   
# resp.mode14 <- glm(respdeath ~ agegr + arsenic, offset = log(personyrs), family = poisson, data = Montana)  
# summary(resp.mode14)  
#   
# ## is there a linear trend across arsenic exposure?  
# resp.mode14.lin <- glm(respdeath ~ agegr + as.numeric(arsenic), offset = log(personyrs), family = poisson, data = Montana)  
# summary(resp.mode14.lin)  
#   
# ## compute IRR  
# idr.display(resp.mode14)

Optional: Exercise 6B

# Not a graded assignment. Optional.

### 6.5. Matched Case-Control Study: VC1to1 from epiDisplay

# Matched case-control dataset available in the epiDisplay package (packages can contain practice data as well as functions)  
data(VC1to1)  
summ(VC1to1) # What is the shape/structure of this dataset?

##   
## No. of observations = 52  
##   
## Var. name obs. mean median s.d. min. max.   
## 1 matset 52 13.5 13.5 7.57 1 26   
## 2 case 52 0.5 0.5 0.5 0 1   
## 3 smoking 52 0.81 1 0.4 0 1   
## 4 rubber 52 0.33 0 0.47 0 1   
## 5 alcohol 52 0.52 1 0.5 0 1

head(VC1to1) # 1 case matched to 1 control

## matset case smoking rubber alcohol  
## 1 1 1 1 0 0  
## 2 1 0 1 0 0  
## 3 2 1 1 0 1  
## 4 2 0 1 1 0  
## 5 3 1 1 1 0  
## 6 3 0 1 1 0

# Reshape the data to facilitate data exploration  
# function 'reshape' converts wide to long or vice versa  
wide <- reshape(VC1to1, timevar = "case", v.names = c("smoking", "rubber", "alcohol"), idvar = "matset", direction = "wide")  
head(wide, 3)

## matset smoking.1 rubber.1 alcohol.1 smoking.0 rubber.0 alcohol.0  
## 1 1 1 0 0 1 0 0  
## 3 2 1 0 1 1 1 0  
## 5 3 1 1 0 1 1 0

table(wide$smoking.1, wide$smoking.0, dnn = c("smoking in case", "smoking in control"))

## smoking in control  
## smoking in case 0 1  
## 0 0 5  
## 1 5 16

# dnn: dimnames names  
  
# matchTab() is for the conditional OR (McNemar's OR)  
matchTab(VC1to1$case, VC1to1$smoking, strata = VC1to1$matset) # Is smoking exposure associated with cancer?

##   
## Exposure status: $ = 1   
##   
## Exposure status: VC1to1 = 1   
##   
## Exposure status: smoking = 1   
##   
## Total number of match sets in the tabulation = 26   
##   
## Number of controls = 1   
## No. of controls exposed  
## No. of cases exposed 0 1  
## 0 0 5  
## 1 5 16  
##   
## Odds ratio by Mantel-Haenszel method = 1   
##   
## Odds ratio by maximum likelihood estimate (MLE) method = 1   
## 95%CI= 0.29 , 3.454   
##

matchTab(VC1to1$case, VC1to1$rubber, strata = VC1to1$matset) # Is working in the rubber industry associated with cancer?

##   
## Exposure status: $ = 1   
##   
## Exposure status: VC1to1 = 1   
##   
## Exposure status: rubber = 1   
##   
## Total number of match sets in the tabulation = 26   
##   
## Number of controls = 1   
## No. of controls exposed  
## No. of cases exposed 0 1  
## 0 13 5  
## 1 4 4  
##   
## Odds ratio by Mantel-Haenszel method = 0.8   
##   
## Odds ratio by maximum likelihood estimate (MLE) method = 0.8   
## 95%CI= 0.215 , 2.979   
##

matchTab(VC1to1$case, VC1to1$alcohol, strata = VC1to1$matset) # Is alcohol consumption associated iwth cancer?

##   
## Exposure status: $ = 1   
##   
## Exposure status: VC1to1 = 1   
##   
## Exposure status: alcohol = 1   
##   
## Total number of match sets in the tabulation = 26   
##   
## Number of controls = 1   
## No. of controls exposed  
## No. of cases exposed 0 1  
## 0 7 2  
## 1 9 8  
##   
## Odds ratio by Mantel-Haenszel method = 4.5   
##   
## Odds ratio by maximum likelihood estimate (MLE) method = 4.5   
## 95%CI= 0.972 , 20.827   
##

## look at the full dataset VC1to6  
data(VC1to6) # 1 case matched to up to 6 controls  
summ(VC1to6)

##   
## No. of observations = 119  
##   
## Var. name obs. mean median s.d. min. max.   
## 1 matset 119 15.75 17 6.96 1 26   
## 2 case 119 0.22 0 0.41 0 1   
## 3 smoking 119 0.7 1 0.46 0 1   
## 4 rubber 119 0.37 0 0.48 0 1   
## 5 alcohol 119 0.42 0 0.5 0 1

VC1to6[, ]

## matset case smoking rubber alcohol  
## 1 1 1 1 0 0  
## 2 1 0 1 0 0  
## 3 2 1 1 0 1  
## 4 2 0 1 1 0  
## 5 3 1 1 1 0  
## 6 3 0 1 1 0  
## 7 4 1 1 0 0  
## 8 4 0 1 1 1  
## 9 4 0 0 1 1  
## 10 5 1 0 0 1  
## 11 5 0 1 0 0  
## 12 5 0 1 0 0  
## 13 6 1 1 0 1  
## 14 6 0 0 0 0  
## 15 6 0 0 0 0  
## 16 7 1 1 0 1  
## 17 7 0 1 0 0  
## 18 7 0 1 0 1  
## 19 7 0 1 1 0  
## 20 8 1 1 0 0  
## 21 8 0 1 0 0  
## 22 8 0 1 0 1  
## 23 8 0 0 1 0  
## 24 9 1 1 1 1  
## 25 9 0 1 1 0  
## 26 9 0 1 1 0  
## 27 9 0 0 1 0  
## 28 10 1 0 0 0  
## 29 10 0 1 1 0  
## 30 10 0 0 0 0  
## 31 10 0 1 0 0  
## 32 11 1 1 0 1  
## 33 11 0 1 0 1  
## 34 11 0 0 0 0  
## 35 11 0 1 0 1  
## 36 12 1 1 0 0  
## 37 12 0 1 0 1  
## 38 12 0 1 0 0  
## 39 12 0 0 0 0  
## 40 13 1 1 1 0  
## 41 13 0 0 0 0  
## 42 13 0 0 0 0  
## 43 13 0 0 0 0  
## 44 14 1 1 0 1  
## 45 14 0 1 0 1  
## 46 14 0 0 0 0  
## 47 14 0 1 1 0  
## 48 14 0 1 0 1  
## 49 15 1 1 0 1  
## 50 15 0 1 0 0  
## 51 15 0 0 0 0  
## 52 15 0 1 0 0  
## 53 15 0 1 0 1  
## 54 16 1 1 0 1  
## 55 16 0 0 0 1  
## 56 16 0 1 0 1  
## 57 16 0 1 0 1  
## 58 16 0 1 1 0  
## 59 17 1 1 1 1  
## 60 17 0 1 0 1  
## 61 17 0 1 1 1  
## 62 17 0 1 0 0  
## 63 17 0 0 1 0  
## 64 18 1 1 0 1  
## 65 18 0 1 0 1  
## 66 18 0 0 1 0  
## 67 18 0 0 1 0  
## 68 18 0 1 0 0  
## 69 18 0 1 0 1  
## 70 19 1 0 0 0  
## 71 19 0 1 0 0  
## 72 19 0 0 0 0  
## 73 19 0 0 0 0  
## 74 19 0 0 1 0  
## 75 19 0 0 0 0  
## 76 20 1 1 1 1  
## 77 20 0 0 0 0  
## 78 20 0 0 0 0  
## 79 20 0 0 0 0  
## 80 20 0 1 0 0  
## 81 20 0 0 0 0  
## 82 21 1 1 1 1  
## 83 21 0 1 1 1  
## 84 21 0 1 1 1  
## 85 21 0 1 0 1  
## 86 21 0 1 1 1  
## 87 21 0 1 1 1  
## 88 21 0 1 1 1  
## 89 22 1 0 0 1  
## 90 22 0 1 1 1  
## 91 22 0 0 0 1  
## 92 22 0 1 1 0  
## 93 22 0 0 0 0  
## 94 22 0 0 1 0  
## 95 22 0 1 0 0  
## 96 23 1 1 1 1  
## 97 23 0 1 1 1  
## 98 23 0 1 1 1  
## 99 23 0 1 1 0  
## 100 23 0 1 1 1  
## 101 23 0 1 1 0  
## 102 23 0 1 1 0  
## 103 24 1 1 1 1  
## 104 24 0 1 0 0  
## 105 24 0 1 0 0  
## 106 24 0 0 1 1  
## 107 24 0 1 0 0  
## 108 24 0 1 0 1  
## 109 24 0 1 0 0  
## 110 25 1 0 0 0  
## 111 25 0 1 1 0  
## 112 25 0 1 1 1  
## 113 25 0 1 0 1  
## 114 25 0 1 0 0  
## 115 26 1 1 0 1  
## 116 26 0 0 0 0  
## 117 26 0 1 1 0  
## 118 26 0 0 0 0  
## 119 26 0 1 1 1

# what is the effect of smoking?  
matchTab(VC1to6$case, VC1to6$smoking, strata = VC1to6$matset) # Is smoking exposure associated with cancer?

##   
## Exposure status: $ = 1   
##   
## Exposure status: VC1to6 = 1   
##   
## Exposure status: smoking = 1   
##   
## Total number of match sets in the tabulation = 26   
##   
## Number of controls = 1   
## No. of controls exposed  
## No. of cases exposed 0 1  
## 0 0 0  
## 1 0 3  
##   
## Number of controls = 2   
## No. of controls exposed  
## No. of cases exposed 0 1 2  
## 0 0 0 1  
## 1 1 1 0  
##   
## Number of controls = 3   
## No. of controls exposed  
## No. of cases exposed 0 1 2 3  
## 0 0 0 1 0  
## 1 1 0 4 1  
##   
## Number of controls = 4   
## No. of controls exposed  
## No. of cases exposed 0 1 2 3 4  
## 0 0 0 0 0 1  
## 1 0 0 1 4 0  
##   
## Number of controls = 5   
## No. of controls exposed  
## No. of cases exposed 0 1 2 3 4 5  
## 0 0 1 0 0 0 0  
## 1 0 1 0 1 0 0  
##   
## Number of controls = 6   
## No. of controls exposed  
## No. of cases exposed 0 1 2 3 4 5 6  
## 0 0 0 0 1 0 0 0  
## 1 0 0 0 0 0 1 2  
##   
## Odds ratio by Mantel-Haenszel method = 1.988   
##   
## Odds ratio by maximum likelihood estimate (MLE) method = 2.066   
## 95%CI= 0.678 , 6.301   
##

matchTab(VC1to6$case, VC1to6$alcohol, strata = VC1to6$matset)

##   
## Exposure status: $ = 1   
##   
## Exposure status: VC1to6 = 1   
##   
## Exposure status: alcohol = 1   
##   
## Total number of match sets in the tabulation = 26   
##   
## Number of controls = 1   
## No. of controls exposed  
## No. of cases exposed 0 1  
## 0 2 0  
## 1 1 0  
##   
## Number of controls = 2   
## No. of controls exposed  
## No. of cases exposed 0 1 2  
## 0 0 0 1  
## 1 2 0 0  
##   
## Number of controls = 3   
## No. of controls exposed  
## No. of cases exposed 0 1 2 3  
## 0 2 2 0 0  
## 1 1 1 1 0  
##   
## Number of controls = 4   
## No. of controls exposed  
## No. of cases exposed 0 1 2 3 4  
## 0 0 0 1 0 0  
## 1 0 2 2 1 0  
##   
## Number of controls = 5   
## No. of controls exposed  
## No. of cases exposed 0 1 2 3 4 5  
## 0 1 0 0 0 0 0  
## 1 1 0 1 0 0 0  
##   
## Number of controls = 6   
## No. of controls exposed  
## No. of cases exposed 0 1 2 3 4 5 6  
## 0 0 0 0 0 0 0 0  
## 1 0 0 2 1 0 0 1  
##   
## Odds ratio by Mantel-Haenszel method = 5.386   
##   
## Odds ratio by maximum likelihood estimate (MLE) method = 5.655   
## 95%CI= 1.811 , 17.659   
##