EPID674 Epidemiologic Data Analysis using R

Regression in R

Kelly Bakulski, Lauren Middleton

Last compiled on December 07, 2022

# Install new packages

# Load packages

# Load data, recheck variables

# Check the file path  
here("nhanes\_class\_dataset.rda")

[1] "/cloud/project/nhanes\_class\_dataset.rda"

# Load the saved R data  
load(here("nhanes\_class\_dataset.rda"), verbose=TRUE)

Loading objects:  
 nhanes

# Check the variables  
str(nhanes)

'data.frame': 9254 obs. of 27 variables:  
 $ SEQN : int 93703 93704 93705 93706 93707 93708 93709 93710 93711 93712 ...  
 $ RIASEX : 'labelled' int 2 1 2 1 1 2 2 2 1 1 ...  
 ..- attr(\*, "label")= chr "Gender"  
 $ sex : Factor w/ 2 levels "Male","Female": 2 1 2 1 1 2 2 2 1 1 ...  
 $ RIDAGEYR : 'labelled' int 2 2 66 18 13 66 75 0 56 18 ...  
 ..- attr(\*, "label")= chr "Age in years at screening"  
 $ age\_groups : Factor w/ 5 levels "[0,16]","(16,32]",..: 1 1 5 2 1 5 5 1 4 2 ...  
 $ RIDRETH1 : 'labelled' int 5 3 4 5 5 5 4 3 5 1 ...  
 ..- attr(\*, "label")= chr "Race/Hispanic origin"  
 $ race\_eth : Factor w/ 5 levels "Non-Hispanic White",..: 5 1 3 5 5 5 3 1 5 2 ...  
 $ INDFMPIR : 'labelled' num 5 5 0.82 NA 1.88 1.63 0.41 4.9 5 0.76 ...  
 ..- attr(\*, "label")= chr "Ratio of family income to poverty"  
 $ DMDEDUC3 : 'labelled' int NA NA NA 15 6 NA NA NA NA 12 ...  
 ..- attr(\*, "label")= chr "Education level - Children/Youth 6-19"  
 $ DMDEDUC2 : 'labelled' int NA NA 2 NA NA 1 4 NA 5 NA ...  
 ..- attr(\*, "label")= chr "Education level - Adults 20+"  
 $ education\_youth: chr NA NA NA "More than high school" ...  
 $ education\_adult: chr NA NA "Less than high school" NA ...  
 $ education : Factor w/ 4 levels "Less than high school",..: NA NA 1 4 1 1 4 NA 4 1 ...  
 $ SDMVSTRA : 'labelled' int 145 143 145 134 138 138 136 134 134 147 ...  
 ..- attr(\*, "label")= chr "Masked variance pseudo-stratum"  
 $ SDMVPSU : 'labelled' int 2 1 2 2 1 2 1 1 2 2 ...  
 ..- attr(\*, "label")= chr "Masked variance pseudo-PSU"  
 $ LBXRBCSI : 'labelled' num NA 4.25 5.48 5.24 5.02 4.59 5.13 NA 4.65 5.12 ...  
 ..- attr(\*, "label")= chr "Red blood cell count (million cells/uL)"  
 $ LBXWBCSI : 'labelled' num NA 7.4 8.6 6.1 11.2 6 7.2 NA 5 7.1 ...  
 ..- attr(\*, "label")= chr "White blood cell count (1000 cells/uL)"  
 $ LBDLYMNO : 'labelled' num NA 3.5 3.4 1.5 4.2 1.9 1.9 NA 1.8 2.2 ...  
 ..- attr(\*, "label")= chr "Lymphocyte number (1000 cells/uL)"  
 $ LBDNENO : 'labelled' num NA 3.2 4.2 3.7 6.1 3.6 4.8 NA 2.7 4.2 ...  
 ..- attr(\*, "label")= chr "Segmented neutrophils num (1000 cell/uL)"  
 $ nlr : 'labelled' num NA 0.914 1.235 2.467 1.452 ...  
 ..- attr(\*, "label")= chr "Segmented neutrophils num (1000 cell/uL)"  
 $ LBXIRN : 'labelled' int NA NA 92 164 91 90 63 NA 56 225 ...  
 ..- attr(\*, "label")= chr "Iron frozen, Serum (ug/dL)"  
 $ iron\_status : Factor w/ 3 levels "Deficient","Excessive",..: NA NA 3 3 3 3 3 NA 1 2 ...  
 $ URXUAS : 'labelled' num NA NA NA NA 5.09 ...  
 ..- attr(\*, "label")= chr "Arsenic, Total - Urine (ug/L)"  
 $ LBXBCD : 'labelled' num NA 0.07 0.24 0.21 0.14 0.73 1.08 NA 0.38 0.26 ...  
 ..- attr(\*, "label")= chr "Blood cadmium (ug/L)"  
 $ LBXBPB : 'labelled' num NA NA 2.98 0.74 0.39 1.53 1.31 NA 2.15 0.27 ...  
 ..- attr(\*, "label")= chr "Blood lead (ug/dL)"  
 $ LBXCOT : 'labelled' num NA NA 0.028 0.138 0.555 0.011 54.3 NA 0.057 13.4 ...  
 ..- attr(\*, "label")= chr "Cotinine, Serum (ng/mL)"  
 $ cut\_groups : Factor w/ 5 levels "[0,8]","(8,20]",..: 1 1 5 2 2 5 5 1 4 2 ...

# Add a label for the race\_eth variable so we don't need add it in every model  
set\_label(nhanes$race\_eth) <- "Race/Ethnicity"  
set\_label(nhanes$sex) <- "Sex"  
# Check the labeled variable  
str(nhanes$race\_eth)

Factor w/ 5 levels "Non-Hispanic White",..: 5 1 3 5 5 5 3 1 5 2 ...  
 - attr(\*, "label")= chr "Race/Ethnicity"

str(nhanes$sex)

Factor w/ 2 levels "Male","Female": 2 1 2 1 1 2 2 2 1 1 ...  
 - attr(\*, "label")= chr "Sex"

# Linear Models: Association between white blood cells (LBXWBCSI) and blood iron (LBXIRN)

## Outline

* Check correlations between variables
* Run and interpret linear regression
* Add covariates
* Check regression diagnostics
* Update linear regression model

## Look at bivariate association between white blood cells and continuous covariates

# Set up the plots  
iron <- ggplot(nhanes,  
 aes(x = LBXIRN,  
 y = LBXWBCSI)) +  
 geom\_point(alpha = 0.1) +  
 ylim(0,20) +  
 geom\_smooth(method = 'lm')  
  
  
# Set up the plots  
age <- ggplot(nhanes,  
 aes(x = RIDAGEYR,  
 y = LBXWBCSI)) +  
 geom\_point(alpha = 0.1) +  
 ylim(0,20) +  
 geom\_smooth(method = 'lm')  
  
pir <- ggplot(nhanes,  
 aes(x = INDFMPIR,  
 y = LBXWBCSI)) +  
 geom\_point(alpha = 0.1) +  
 ylim(0,20) +  
 geom\_smooth(method = 'lm')  
  
# Plot the distributions  
ggarrange(iron, age, pir)

Don't know how to automatically pick scale for object of type  
<labelled/integer>. Defaulting to continuous.  
`geom\_smooth()` using formula = 'y ~ x'

Warning: Removed 3341 rows containing non-finite values (`stat\_smooth()`).

Warning: Removed 3341 rows containing missing values (`geom\_point()`).

Don't know how to automatically pick scale for object of type  
<labelled/integer>. Defaulting to continuous.  
`geom\_smooth()` using formula = 'y ~ x'

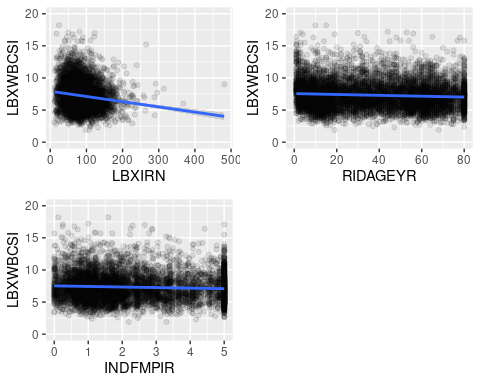
Warning: Removed 1733 rows containing non-finite values (`stat\_smooth()`).

Warning: Removed 1733 rows containing missing values (`geom\_point()`).

`geom\_smooth()` using formula = 'y ~ x'

Warning: Removed 2650 rows containing non-finite values (`stat\_smooth()`).

Warning: Removed 2650 rows containing missing values (`geom\_point()`).



# Simple linear regression

# Create a simple regression model for white blood cell count and blood iron (crude model)  
model\_crude <- lm(LBXWBCSI ~ LBXIRN,  
 na.action = na.omit,  
 data = nhanes)  
  
# print regression results  
summary(model\_crude)

Call:  
lm(formula = LBXWBCSI ~ LBXIRN, data = nhanes, na.action = na.omit)  
  
Residuals:  
 Min 1Q Median 3Q Max   
 -5.31 -1.59 -0.32 1.12 392.38   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 8.133376 0.186516 43.61 < 2e-16 \*\*\*  
LBXIRN -0.009411 0.001990 -4.73 2.3e-06 \*\*\*  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 5.621 on 5917 degrees of freedom  
 (3335 observations deleted due to missingness)  
Multiple R-squared: 0.003767, Adjusted R-squared: 0.003598   
F-statistic: 22.37 on 1 and 5917 DF, p-value: 2.298e-06

# Is blood iron associated with white blood cell count?  
# Do these findings match your expectations?

# Make a table of regression results

# More detailed information about regression results  
summary.aov(model\_crude)

Df Sum Sq Mean Sq F value Pr(>F)   
LBXIRN 1 707 706.9 22.37 2.3e-06 \*\*\*  
Residuals 5917 186951 31.6   
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
3335 observations deleted due to missingness

glance(model\_crude)

# A tibble: 1 × 12  
 r.squared adj.r.sq…¹ sigma stati…² p.value df logLik AIC BIC devia…³  
 <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
1 0.00377 0.00360 5.62 22.4 2.30e-6 1 -18617. 37240. 37260. 186951.  
# … with 2 more variables: df.residual <int>, nobs <int>, and abbreviated  
# variable names ¹​adj.r.squared, ²​statistic, ³​deviance

# Make an output table of the findings, default settings  
tbl\_regression(model\_crude)

Table printed with `knitr::kable()`, not {gt}. Learn why at  
https://www.danieldsjoberg.com/gtsummary/articles/rmarkdown.html  
To suppress this message, include `message = FALSE` in the code chunk header.

| **Characteristic** | **Beta** | **95% CI** | **p-value** |
| --- | --- | --- | --- |
| Iron frozen, Serum (ug/dL) | -0.01 | -0.01, -0.01 | <0.001 |

# Make an output table of the findings, update settings  
tbl\_regression(model\_crude, estimate\_fun = partial(style\_sigfig, digits = 3)) %>%  
 bold\_labels() %>%  
 add\_glance\_table(include = c(adj.r.squared, AIC, nobs))

Table printed with `knitr::kable()`, not {gt}. Learn why at  
https://www.danieldsjoberg.com/gtsummary/articles/rmarkdown.html  
To suppress this message, include `message = FALSE` in the code chunk header.

| **Characteristic** | **Beta** | **95% CI** | **p-value** |
| --- | --- | --- | --- |
| **Iron frozen, Serum (ug/dL)** | -0.009 | -0.013, -0.006 | <0.001 |
| Adjusted R² | 0.004 |  |  |
| AIC | 37,240 |  |  |
| No. Obs. | 5,919 |  |  |

# Check your understanding!

Perform linear regression with blood cadmium (variable name: LBXBCD) as a predictor and neutrophil lymphocyte ratio (variable name: nlr) as the outcome. Are cadmium levels associated with the neutrophil lymphocyte ratio?

# Multivariable linear regression: Adding a numeric covariate

## Add age into the model  
model\_age <- lm(LBXWBCSI ~ LBXIRN +  
 RIDAGEYR,  
 na.action = na.omit,  
 data = nhanes)  
summary(model\_age)

Call:  
lm(formula = LBXWBCSI ~ LBXIRN + RIDAGEYR, data = nhanes, na.action = na.omit)  
  
Residuals:  
 Min 1Q Median 3Q Max   
 -5.36 -1.59 -0.32 1.12 392.33   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 8.035146 0.246599 32.584 < 2e-16 \*\*\*  
LBXIRN -0.009389 0.001990 -4.718 2.44e-06 \*\*\*  
RIDAGEYR 0.002118 0.003479 0.609 0.543   
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 5.621 on 5916 degrees of freedom  
 (3335 observations deleted due to missingness)  
Multiple R-squared: 0.003829, Adjusted R-squared: 0.003493   
F-statistic: 11.37 on 2 and 5916 DF, p-value: 1.178e-05

# Does anything change with age in the model?

# Multivariable linear regression: Adding categorical covariates

# Check that variables are factors, check the reference level  
str(nhanes$race\_eth) #5 levels for race with Non-Hispanic White as the reference

Factor w/ 5 levels "Non-Hispanic White",..: 5 1 3 5 5 5 3 1 5 2 ...  
 - attr(\*, "label")= chr "Race/Ethnicity"

str(nhanes$sex) #2 levels for sex with Male as the reference

Factor w/ 2 levels "Male","Female": 2 1 2 1 1 2 2 2 1 1 ...  
 - attr(\*, "label")= chr "Sex"

# Construct adjusted model  
model\_adj <- lm(LBXWBCSI ~ LBXIRN +  
 RIDAGEYR +  
 race\_eth +  
 sex +  
 INDFMPIR,  
 na.action = na.omit,  
 data = nhanes)  
summary(model\_adj)

Call:  
lm(formula = LBXWBCSI ~ LBXIRN + RIDAGEYR + race\_eth + sex +   
 INDFMPIR, data = nhanes, na.action = na.omit)  
  
Residuals:  
 Min 1Q Median 3Q Max   
 -5.31 -1.57 -0.38 1.10 392.67   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 8.405367 0.353643 23.768 < 2e-16 \*\*\*  
LBXIRN -0.009690 0.002307 -4.201 2.7e-05 \*\*\*  
RIDAGEYR 0.003400 0.004007 0.849 0.3962   
race\_ethMexican American 0.050520 0.263667 0.192 0.8481   
race\_ethNon-Hispanic Black -0.754378 0.227093 -3.322 0.0009 \*\*\*  
race\_ethOther Hispanic -0.278759 0.317048 -0.879 0.3793   
race\_ethOther Race -0.243246 0.234123 -1.039 0.2989   
sexFemale 0.133446 0.168931 0.790 0.4296   
INDFMPIR -0.090734 0.053015 -1.711 0.0871 .   
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 5.955 on 5164 degrees of freedom  
 (4081 observations deleted due to missingness)  
Multiple R-squared: 0.006668, Adjusted R-squared: 0.005129   
F-statistic: 4.333 on 8 and 5164 DF, p-value: 3.204e-05

# Why are there only 4 levels for race, not 5?

# Make a nice table of regression statistics

# For more information: *https://www.danieldsjoberg.com/gtsummary/articles/tbl\_regression.html*

# Model summaries as an object  
glance\_model\_adj <- as.data.frame(glance(model\_adj))  
glance\_model\_adj

r.squared adj.r.squared sigma statistic p.value df logLik  
1 0.006667901 0.005129044 5.955087 4.333022 3.203781e-05 8 -16565.57  
 AIC BIC deviance df.residual nobs  
1 33151.14 33216.65 183131.3 5164 5173

# Model statistics as an object  
model\_stats <- as.data.frame(tidy(model\_adj))  
model\_stats

term estimate std.error statistic p.value  
1 (Intercept) 8.405366887 0.353643447 23.7679136 1.366252e-118  
2 LBXIRN -0.009689794 0.002306576 -4.2009424 2.702869e-05  
3 RIDAGEYR 0.003400143 0.004007022 0.8485462 3.961732e-01  
4 race\_ethMexican American 0.050520073 0.263666760 0.1916058 8.480586e-01  
5 race\_ethNon-Hispanic Black -0.754377501 0.227093158 -3.3218856 9.003281e-04  
6 race\_ethOther Hispanic -0.278759350 0.317048367 -0.8792329 3.793159e-01  
7 race\_ethOther Race -0.243245611 0.234122927 -1.0389654 2.988695e-01  
8 sexFemale 0.133445900 0.168930866 0.7899439 4.295968e-01  
9 INDFMPIR -0.090734449 0.053014890 -1.7114899 8.705076e-02

# Produce a formatted model summary table  
tbl\_regression(model\_adj, estimate\_fun = partial(style\_sigfig, digits = 3)) %>%  
 bold\_labels()

Table printed with `knitr::kable()`, not {gt}. Learn why at  
https://www.danieldsjoberg.com/gtsummary/articles/rmarkdown.html  
To suppress this message, include `message = FALSE` in the code chunk header.

| **Characteristic** | **Beta** | **95% CI** | **p-value** |
| --- | --- | --- | --- |
| **Iron frozen, Serum (ug/dL)** | -0.010 | -0.014, -0.005 | <0.001 |
| **Age in years at screening** | 0.003 | -0.004, 0.011 | 0.4 |
| **Race/Ethnicity** |  |  |  |
| Non-Hispanic White | — | — |  |
| Mexican American | 0.051 | -0.466, 0.567 | 0.8 |
| Non-Hispanic Black | -0.754 | -1.20, -0.309 | <0.001 |
| Other Hispanic | -0.279 | -0.900, 0.343 | 0.4 |
| Other Race | -0.243 | -0.702, 0.216 | 0.3 |
| **Sex** |  |  |  |
| Male | — | — |  |
| Female | 0.133 | -0.198, 0.465 | 0.4 |
| **Ratio of family income to poverty** | -0.091 | -0.195, 0.013 | 0.087 |

# Add features to the formatted model summary table  
tbl\_regression(model\_adj, estimate\_fun = partial(style\_sigfig, digits = 3)) %>% #change significant figures on the estimates  
 bold\_labels() %>%  
 add\_glance\_table(include = c(adj.r.squared, AIC, nobs)) %>% # Adds rows for model summary statistics  
 add\_significance\_stars() # Adds asterisks instead of p-values, standard error instead of confidence interval

Table printed with `knitr::kable()`, not {gt}. Learn why at  
https://www.danieldsjoberg.com/gtsummary/articles/rmarkdown.html  
To suppress this message, include `message = FALSE` in the code chunk header.

| **Characteristic** | **Beta** | **SE** |
| --- | --- | --- |
| **Iron frozen, Serum (ug/dL)** | -0.010\*\*\* | 0.002 |
| **Age in years at screening** | 0.003 | 0.004 |
| **Race/Ethnicity** |  |  |
| Non-Hispanic White | — | — |
| Mexican American | 0.051 | 0.264 |
| Non-Hispanic Black | -0.754\*\*\* | 0.227 |
| Other Hispanic | -0.279 | 0.317 |
| Other Race | -0.243 | 0.234 |
| **Sex** |  |  |
| Male | — | — |
| Female | 0.133 | 0.169 |
| **Ratio of family income to poverty** | -0.091 | 0.053 |
| Adjusted R² | 0.005 |  |
| AIC | 33,151 |  |
| No. Obs. | 5,173 |  |

# Make a table comparing regression models

# What covariates are associated with white blood cells? Do they make sense biologically?  
# Add other covariates to the model: which variables are biologically important?  
# Set up models that add variables sequentially  
# Run the models and convert them into tbl formatted objects  
model\_0 <- tbl\_regression(lm(LBXWBCSI ~ LBXIRN,  
 na.action = na.omit, data = nhanes),  
 estimate\_fun = partial(style\_sigfig, digits = 3))  
model\_1 <- tbl\_regression(lm(LBXWBCSI ~ LBXIRN + RIDAGEYR,  
 na.action = na.omit, data = nhanes),  
 estimate\_fun = partial(style\_sigfig, digits = 3))  
model\_2 <- tbl\_regression(lm(LBXWBCSI ~ LBXIRN + RIDAGEYR + race\_eth,  
 na.action = na.omit, data = nhanes),  
 estimate\_fun = partial(style\_sigfig, digits = 3))  
model\_3 <- tbl\_regression(lm(LBXWBCSI ~ LBXIRN + RIDAGEYR + race\_eth + sex,  
 na.action = na.omit, data = nhanes),  
 estimate\_fun = partial(style\_sigfig, digits = 3))  
model\_4 <- tbl\_regression(lm(LBXWBCSI ~ LBXIRN + RIDAGEYR + race\_eth + sex + INDFMPIR,  
 na.action = na.omit, data = nhanes),  
 estimate\_fun = partial(style\_sigfig, digits = 3))  
  
# Compare outputs from models side by side, save as html for viewing  
tbl\_merge(tbls = list(model\_0, model\_1, model\_2, model\_3, model\_4),  
 tab\_spanner = c("\*\*Crude\*\*", "\*\*Model 1\*\*", "\*\*Model 2\*\*", "\*\*Model 3\*\*", "\*\*Model 4\*\*"))

Table printed with `knitr::kable()`, not {gt}. Learn why at  
https://www.danieldsjoberg.com/gtsummary/articles/rmarkdown.html  
To suppress this message, include `message = FALSE` in the code chunk header.

| **Characteristic** | **Beta** | **95% CI** | **p-value** | **Beta** | **95% CI** | **p-value** | **Beta** | **95% CI** | **p-value** | **Beta** | **95% CI** | **p-value** | **Beta** | **95% CI** | **p-value** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Iron frozen, Serum (ug/dL) | -0.009 | -0.013, -0.006 | <0.001 | -0.009 | -0.013, -0.005 | <0.001 | -0.010 | -0.014, -0.006 | <0.001 | -0.010 | -0.014, -0.006 | <0.001 | -0.010 | -0.014, -0.005 | <0.001 |
| Age in years at screening |  |  |  | 0.002 | -0.005, 0.009 | 0.5 | 0.002 | -0.005, 0.009 | 0.6 | 0.002 | -0.005, 0.009 | 0.6 | 0.003 | -0.004, 0.011 | 0.4 |
| Race/Ethnicity |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Non-Hispanic White |  |  |  |  |  |  | — | — |  | — | — |  | — | — |  |
| Mexican American |  |  |  |  |  |  | 0.025 | -0.425, 0.476 | >0.9 | 0.022 | -0.428, 0.473 | >0.9 | 0.051 | -0.466, 0.567 | 0.8 |
| Non-Hispanic Black |  |  |  |  |  |  | -0.761 | -1.15, -0.370 | <0.001 | -0.761 | -1.15, -0.370 | <0.001 | -0.754 | -1.20, -0.309 | <0.001 |
| Other Hispanic |  |  |  |  |  |  | -0.281 | -0.813, 0.250 | 0.3 | -0.286 | -0.818, 0.246 | 0.3 | -0.279 | -0.900, 0.343 | 0.4 |
| Other Race |  |  |  |  |  |  | -0.292 | -0.699, 0.116 | 0.2 | -0.291 | -0.699, 0.117 | 0.2 | -0.243 | -0.702, 0.216 | 0.3 |
| Sex |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Male |  |  |  |  |  |  |  |  |  | — | — |  | — | — |  |
| Female |  |  |  |  |  |  |  |  |  | 0.126 | -0.166, 0.418 | 0.4 | 0.133 | -0.198, 0.465 | 0.4 |
| Ratio of family income to poverty |  |  |  |  |  |  |  |  |  |  |  |  | -0.091 | -0.195, 0.013 | 0.087 |

# Consider log transformed exposure

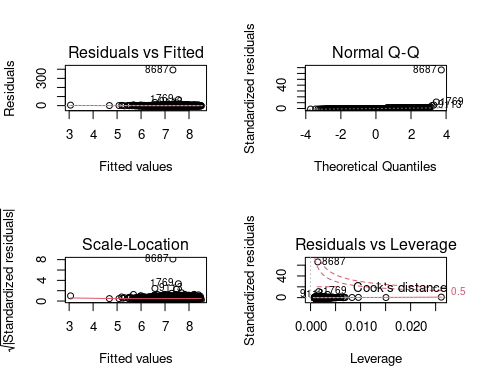
# Calculate multivariable adjusted model with a log transformed iron variable  
model\_adj\_log <- lm(LBXWBCSI ~ log2(LBXIRN) +  
 RIDAGEYR +  
 race\_eth +  
 sex +  
 INDFMPIR,  
 na.action = na.omit,  
 data = nhanes)  
  
# print regression results  
summary(model\_adj\_log)

Call:  
lm(formula = LBXWBCSI ~ log2(LBXIRN) + RIDAGEYR + race\_eth +   
 sex + INDFMPIR, data = nhanes, na.action = na.omit)  
  
Residuals:  
 Min 1Q Median 3Q Max   
 -6.22 -1.56 -0.39 1.11 392.67   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 11.303028 0.882410 12.809 < 2e-16 \*\*\*  
log2(LBXIRN) -0.596985 0.131167 -4.551 5.45e-06 \*\*\*  
RIDAGEYR 0.004149 0.004006 1.036 0.300424   
race\_ethMexican American 0.045388 0.263582 0.172 0.863291   
race\_ethNon-Hispanic Black -0.761116 0.227027 -3.353 0.000807 \*\*\*  
race\_ethOther Hispanic -0.279339 0.316954 -0.881 0.378184   
race\_ethOther Race -0.249981 0.234063 -1.068 0.285566   
sexFemale 0.103170 0.169733 0.608 0.543321   
INDFMPIR -0.086706 0.053037 -1.635 0.102147   
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 5.953 on 5164 degrees of freedom  
 (4081 observations deleted due to missingness)  
Multiple R-squared: 0.007255, Adjusted R-squared: 0.005718   
F-statistic: 4.718 on 8 and 5164 DF, p-value: 8.827e-06

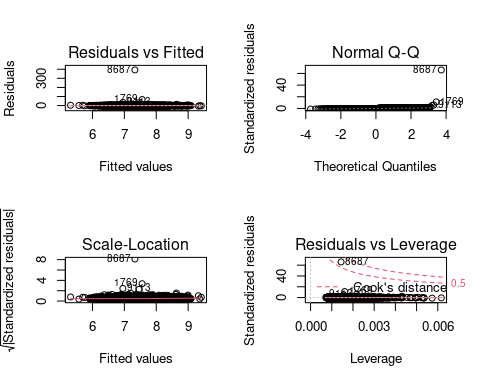
# Is log2 of blood iron associated with white blood cell count?

# Regression diagnostics: LBXIRN vs log2(LBXIRN)

### Regression Diagnostics  
  
## In the case of linear model, the plot of the model gives diagnostic plots  
# Residuals vs Fitted: linearity  
# Scale-location: homogeneity of variance  
# QQ: normality  
# Residuals vs Leverage: outliers and leverage points  
  
# non log-transformed model  
par(mfrow = c(2, 2))  
plot(model\_adj)



# log-transformed model  
par(mfrow = c(2, 2))  
plot(model\_adj\_log)



# Let's go with the log transformed exposure

# Test for effect modification by sex

# Does the relationship between blood iron and white blood cells vary by sex?  
  
model\_int <- lm(LBXWBCSI ~ log2(LBXIRN) +  
 RIDAGEYR +  
 race\_eth +  
 sex +  
 log2(LBXIRN) \* sex +  
 INDFMPIR,  
 na.action = na.omit,  
 data = nhanes)  
summary(model\_int)

Call:  
lm(formula = LBXWBCSI ~ log2(LBXIRN) + RIDAGEYR + race\_eth +   
 sex + log2(LBXIRN) \* sex + INDFMPIR, data = nhanes, na.action = na.omit)  
  
Residuals:  
 Min 1Q Median 3Q Max   
 -6.15 -1.56 -0.39 1.11 392.66   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 11.136365 1.370370 8.127 5.48e-16 \*\*\*  
log2(LBXIRN) -0.571384 0.207707 -2.751 0.005964 \*\*   
RIDAGEYR 0.004191 0.004015 1.044 0.296617   
race\_ethMexican American 0.044523 0.263663 0.169 0.865910   
race\_ethNon-Hispanic Black -0.761390 0.227055 -3.353 0.000804 \*\*\*  
race\_ethOther Hispanic -0.278877 0.316998 -0.880 0.379038   
race\_ethOther Race -0.250433 0.234102 -1.070 0.284778   
sexFemale 0.371978 1.699429 0.219 0.826749   
INDFMPIR -0.086706 0.053042 -1.635 0.102179   
log2(LBXIRN):sexFemale -0.042460 0.267096 -0.159 0.873698   
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 5.954 on 5163 degrees of freedom  
 (4081 observations deleted due to missingness)  
Multiple R-squared: 0.00726, Adjusted R-squared: 0.00553   
F-statistic: 4.195 on 9 and 5163 DF, p-value: 2.016e-05

model\_int <- as.data.frame(tidy(model\_int))  
model\_int

term estimate std.error statistic p.value  
1 (Intercept) 11.136365118 1.370370040 8.1265387 5.480650e-16  
2 log2(LBXIRN) -0.571383798 0.207707103 -2.7509112 5.963695e-03  
3 RIDAGEYR 0.004191296 0.004015342 1.0438205 2.966173e-01  
4 race\_ethMexican American 0.044523414 0.263662883 0.1688649 8.659095e-01  
5 race\_ethNon-Hispanic Black -0.761390063 0.227054900 -3.3533302 8.042025e-04  
6 race\_ethOther Hispanic -0.278876963 0.316997501 -0.8797450 3.790384e-01  
7 race\_ethOther Race -0.250432711 0.234101929 -1.0697593 2.847777e-01  
8 sexFemale 0.371978200 1.699429168 0.2188842 8.267489e-01  
9 INDFMPIR -0.086706208 0.053042010 -1.6346705 1.021791e-01  
10 log2(LBXIRN):sexFemale -0.042460474 0.267096365 -0.1589706 8.736982e-01

# Is the interaction term significant?

# Stratify by levels of sex and perform regression

### Stratify dataset by sex  
  
# Check distribution  
table(nhanes$sex)

Male Female   
 4557 4697

# Filter dataset to only include male, run linear regression, save results as dataframe  
model\_male <- nhanes %>%  
 filter(sex == "Male") %>%  
 lm(LBXWBCSI ~ log2(LBXIRN) +  
 RIDAGEYR +  
 race\_eth +  
 INDFMPIR,  
 na.action = na.omit,  
 data = .) %>%  
 tidy() %>%  
 as.data.frame()  
model\_male

term estimate std.error statistic p.value  
1 (Intercept) 11.419492043 0.612073685 18.65705440 7.333265e-73  
2 log2(LBXIRN) -0.604605274 0.091079569 -6.63820967 3.875811e-11  
3 RIDAGEYR 0.004132096 0.002474667 1.66975807 9.509168e-02  
4 race\_ethMexican American -0.008868574 0.165830686 -0.05347969 9.573540e-01  
5 race\_ethNon-Hispanic Black -1.267674137 0.142236898 -8.91241411 9.372931e-19  
6 race\_ethOther Hispanic -0.301455821 0.199909783 -1.50795932 1.316905e-01  
7 race\_ethOther Race -0.232150952 0.145190398 -1.59894149 1.099592e-01  
8 INDFMPIR -0.068857794 0.033238624 -2.07161985 3.840264e-02

# What is the beta coefficient for blood iron in males?  
  
# Filter dataset to only include female, run linear regression, save results as dataframe  
model\_female <- nhanes %>%  
 filter(sex == "Female") %>%  
 lm(LBXWBCSI ~ log2(LBXIRN) +  
 RIDAGEYR +  
 race\_eth +  
 INDFMPIR,  
 na.action = na.omit,  
 data = .) %>%  
 tidy() %>%  
 as.data.frame()  
model\_female

term estimate std.error statistic p.value  
1 (Intercept) 11.215775946 1.445799984 7.7574879 1.227768e-14  
2 log2(LBXIRN) -0.582910933 0.225791720 -2.5816311 9.886867e-03  
3 RIDAGEYR 0.004382003 0.007577939 0.5782579 5.631393e-01  
4 race\_ethMexican American 0.111462823 0.487802252 0.2285000 8.192752e-01  
5 race\_ethNon-Hispanic Black -0.292143334 0.421946588 -0.6923704 4.887656e-01  
6 race\_ethOther Hispanic -0.245573443 0.584952541 -0.4198177 6.746528e-01  
7 race\_ethOther Race -0.266938051 0.439026395 -0.6080228 5.432246e-01  
8 INDFMPIR -0.094697353 0.098571340 -0.9606986 3.367916e-01

# What is the beta coefficient for blood iron in females? Is it similar to that in males?

# Optional: Automate linear regressions with multiple outcomes and exposures

# Goal: run individual linear regressions for 5 immune measure outcomes and 5 chemical exposures  
  
# First create a long dataset from nhanes  
 # Remove labels so columns can be vertically combined  
 # Keep only the columns we want  
 # Log transform the chemical measures  
 # Make both the chemicals and immune measures long  
 # Remove any rows with missing values  
  
# Remove all labels  
nhanes\_unlab <- nhanes %>%  
 remove\_all\_labels()  
  
# Create vectors of the variables to keep  
demog <- c("SEQN",  
 "sex",  
 "RIDAGEYR",  
 "race\_eth",  
 "INDFMPIR",  
 "education")  
cells <- c("LBXRBCSI",  
 "LBXWBCSI",  
 "LBDLYMNO",  
 "LBDNENO",  
 "nlr")  
chems <- c("LBXIRN",  
 "URXUAS",  
 "LBXCOT",  
 "LBXBCD",  
 "LBXBPB")  
  
# Create the long dataset  
long\_nhanes <- nhanes\_unlab %>%  
 select(all\_of(demog),  
 all\_of(cells),  
 all\_of(chems)) %>%  
 mutate(LBXIRN = log2(LBXIRN),  
 LBXCOT = log2(LBXCOT),  
 URXUAS = log2(URXUAS),  
 LBXBCD = log2(LBXBCD),  
 LBXBPB = log2(LBXBPB)) %>%  
 pivot\_longer(cols = all\_of(chems),  
 names\_to = "chemical\_codename",  
 values\_to = "chem\_measurement") %>%  
 pivot\_longer(cols = all\_of(cells),  
 names\_to = "celltype\_codename",  
 values\_to = "cell\_measurement") %>%  
 na.omit()  
  
# Run multiple linear regressions  
 df\_regressions\_i <- long\_nhanes %>%  
 group\_by(celltype\_codename,  
 chemical\_codename) %>%  
 do(lm(cell\_measurement ~ chem\_measurement +  
 RIDAGEYR +  
 race\_eth +  
 sex +  
 INDFMPIR,  
 na.action = na.omit,  
 data = .) %>%  
 tidy(.)) %>%  
 ungroup(.)  
 # How many regressions did we just run?  
 # What is the formula for the first regression?  
   
df\_regressions\_i

# A tibble: 225 × 7  
 celltype\_codename chemical\_codename term estimate std.e…¹ stati…² p.value  
 <chr> <chr> <chr> <dbl> <dbl> <dbl> <dbl>  
 1 LBDLYMNO LBXBCD (Inte… 2.54 6.77e-2 37.5 4.11e-276  
 2 LBDLYMNO LBXBCD chem\_… 0.00650 1.48e-2 0.438 6.61e- 1  
 3 LBDLYMNO LBXBCD RIDAG… -0.00691 8.22e-4 -8.41 5.04e- 17  
 4 LBDLYMNO LBXBCD race\_… 0.127 5.17e-2 2.46 1.39e- 2  
 5 LBDLYMNO LBXBCD race\_… -0.0273 4.46e-2 -0.612 5.41e- 1  
 6 LBDLYMNO LBXBCD race\_… -0.00171 6.31e-2 -0.0270 9.78e- 1  
 7 LBDLYMNO LBXBCD race\_… 0.0364 4.66e-2 0.780 4.35e- 1  
 8 LBDLYMNO LBXBCD sexFe… 0.0743 3.29e-2 2.26 2.42e- 2  
 9 LBDLYMNO LBXBCD INDFM… 0.00732 1.06e-2 0.692 4.89e- 1  
10 LBDLYMNO LBXBPB (Inte… 2.52 5.50e-2 45.8 0   
# … with 215 more rows, and abbreviated variable names ¹​std.error, ²​statistic