Project 2

Rosa Juan (rij87)

5/1/2020

Introduction

```
library(tidyverse)
library(sandwich)
library(lmtest)
library(plotROC)
library(glmnet)

class_diag <- function(probs, truth) {

    tab <- table(factor(probs > 0.5, levels = c("FALSE", "TRUE")),
        truth)
    acc = sum(diag(tab))/sum(tab)
    sens = tab[2, 2]/colSums(tab)[2]
    spec = tab[1, 1]/colSums(tab)[1]
    ppv = tab[2, 2]/rowSums(tab)[2]

    if (is.numeric(truth) == FALSE & is.logical(truth) == FALSE)
        truth <- as.numeric(truth) - 1

    ord <- order(probs, decreasing = TRUE)
    probs <- probs[ord]
    truth <- truth[ord]

    TPR = cumsum(truth)/max(1, sum(truth))
    FPR = cumsum(!truth)/max(1, sum(!truth))

    dup <- c(probs[-1] >= probs[-length(probs)], FALSE)
    TPR <- c(0, TPR[!dup], 1)
    FPR <- c(0, FPR[!dup], 1)

    n <- length(TPR)
    auc <- sum(((TPR[-1] + TPR[-n])/2) * (FPR[-1] - FPR[-n]))

    data.frame(acc, sens, spec, ppv, auc)
}

getwd()
```

```
## [1] "/Users/Rosa/Downloads/SDS 348"
```

```
data <- read.csv(file = "ModifiedData.csv")
glimpse(data)
```

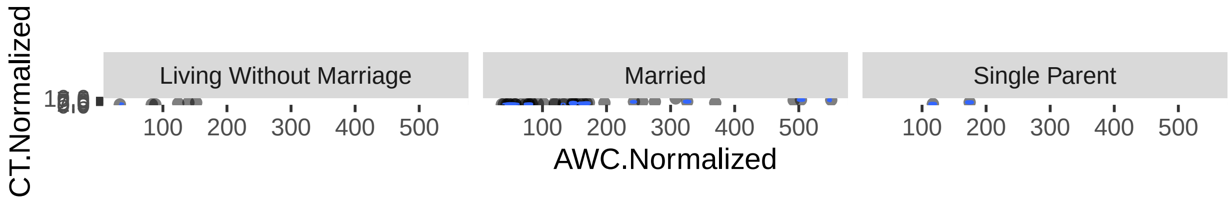
```
## Observations: 86
## Variables: 19
## $ Participant              <fct> P01, P02, P03, P04, P05, P06, P07, P08, P0...
## $ Duration_Secs.sum        <int> NA, NA, NA, NA, NA, 115640, 12860, 87128, ...
## $ Total.Recording.Hours    <dbl> NA, NA, NA, NA, NA, 32.122222, 3.572222, 2...
## $ AWC_COUNT.sum            <int> NA, NA, NA, NA, NA, 33103, 291, 12287, NA,...
## $ AWC.Normalized           <dbl> NA, NA, NA, NA, NA, 769.95870, 547.30233, ...
## $ CT_COUNT.sum             <int> NA, NA, NA, NA, NA, 273, 1, 181, NA, NA, 2...
## $ CT.Normalized            <dbl> NA, NA, NA, NA, NA, 6.349839, 1.880764, 7....
## $ Age.months               <dbl> 6.57, 8.27, 5.73, 6.30, 2.80, 1.70, 2.27, ...
## $ Family.Status            <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 5, 1, 1, ...
## $ Family.Annual.Income     <int> 6, 6, 3, 3, 5, 5, 4, 6, 3, 6, 4, 3, 6, 5, ...
## $ Education                <int> 5, 5, 4, 7, 5, 7, 7, 7, 5, 5, 5, 5, 7, 7, ...
## $ Occupation               <int> 3, 1, 3, 3, 3, 3, 3, 3, 1, 3, 2, 1, 3, 3, ...
## $ Caregiver.Race.Ethnicity <int> 4, 4, 6, 4, 4, 4, 4, 4, 4, 4, 4, 8, 8, 4, ...
## $ Language                 <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 1, 1, ...
## $ cryfreq                  <dbl> NA, NA, NA, 78.28520, NA, NA, NA, 70.00792...
## $ crydur                   <dbl> NA, NA, NA, 1.895284, NA, NA, NA, 1.972468...
## $ Baby.Gender              <int> 1, 2, 2, 1, 2, 1, 1, 1, 2, 1, 2, 1, 1, 1, ...
## $ Caregiver.Age            <int> 31, 33, 35, 33, 35, 31, 41, 42, 34, 35, 29...
## $ Infant.Race.Ethnicity    <int> 4, 4, 8, 4, 4, 4, 4, 8, 8, 4, 4, 8, 8, 4, ...
```

Education changes the potential knowledge a parent can have in raising a child. This can influence infant distress and development (Fouts et al, 2012). This data was obtained from the Daily Activity Lab (DAL) research project at UT Austin. Demographic and LENA outputs from 86 participants were gathered to study mother-infant interactions. The LENA recording device picks up language and distress variables which can be used to compare against demographic information. The demographic information collected includes: Age.months of the baby, Family.Status (1 = married; 2 = separated; 3 = divorced; 4 = single parent; 5 = living with a partner without marriage), Family.Annual.Income (1 = under $25K; 2 = $25,000 - $49,999; 3 = $50,000 - $74,999; 4 = $75,000 - $99,999; 5 = $100K - $124,999; 6 = $125K and above), Education (1 = less than 8th grade; 2 = some high school; 3 = high school diploma/GED; 4 = some college; 5 = college degree; 6 = some graduate school; 7 = graduate school degree; 8 = other), language (1 = English, 2 = Spanish, 8 = Other), and Baby.Gender (1 = Female; 2 = Male). The LENA outputs collected included: cryfreq (number of instances in a day); crydur (hours per day); AWC.Normlalized (Adult Word Count (AWC): the estimated amount of adult words spoken per day); CT.Normalized (Conversational turns (CT): the estimated adult-child interactions per day). This data is just one parameter that the DAL research center is trying to incorporate into a fitbit program or other related device that will tell mothers, relatively to other babies, how their baby is developing.

MANOVA

```
data1 <- data %>% filter(Family.Status != 3) %>% mutate(Family.Status = ifelse(as.character(Family.Statu
    "1", "Married", ifelse(as.character(Family.Status) == "4",
    "Single Parent", "Living Without Marriage")))

data1 %>% na.omit %>% ggplot(aes(x = AWC.Normalized, y = CT.Normalized)) +
    geom_point(alpha = 0.5) + geom_density_2d(h = 2) + coord_fixed() +
    facet_wrap(~Family.Status)
```



```
FamMan <- manova(cbind(AWC.Normalized, CT.Normalized, cryfreq,
    crydur) ~ Family.Status, data = data1)
summary(FamMan)
```

```
##               Df  Pillai approx F num Df den Df Pr(>F)
## Family.Status  2 0.20836   1.2502      8     86 0.2804
## Residuals     45
```

```
# not significant, so this is just for practice
summary.aov(FamMan)
```

```
##  Response AWC.Normalized :
##               Df Sum Sq Mean Sq F value Pr(>F)
## Family.Status  2  22986   11493  0.7509 0.4778
## Residuals     45 688799   15307
##
##  Response CT.Normalized :
##               Df  Sum Sq Mean Sq F value Pr(>F)
## Family.Status  2   4.954  2.4772   0.556 0.5774
## Residuals     45 200.503  4.4556
##
##  Response cryfreq :
##               Df Sum Sq Mean Sq F value Pr(>F)
## Family.Status  2  632.9  316.43  2.1663 0.1264
## Residuals     45 6573.2  146.07
##
##  Response crydur :
##               Df Sum Sq Mean Sq F value  Pr(>F)
## Family.Status  2  7.129  3.5644  3.2278 0.04898 *
## Residuals     45 49.693  1.1043
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## 32 observations deleted due to missingness
```

```r
data1 %>% group_by(Family.Status) %>% summarize(mean(AWC.Normalized,
    na.rm = T), mean(CT.Normalized, na.rm = T), mean(cryfreq,
    na.rm = T), mean(crydur, na.rm = T)) %>% na.omit
```

```
## # A tibble: 3 x 5
##   Family.Status `mean(AWC.Norma~ `mean(CT.Normal~ `mean(cryfreq, ~
##   <chr>                    <dbl>            <dbl>            <dbl>
## 1 Living Witho~            6118.            30.7              73.0
## 2 Married                   350.             5.46             81.4
## 3 Single Parent            6637.           112.               87.6
## # ... with 1 more variable: `mean(crydur, na.rm = T)` <dbl>
```

```r
pairwise.t.test(data1$crydur, data1$Family.Status, p.adj = "none")
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  data1$crydur and data1$Family.Status
##
##               Living Without Marriage Married
## Married       0.024                   -
## Single Parent 0.145                   0.888
##
## P value adjustment method: none
```

MANOVA tests were run of various demographic variables against the LENA outputs, but there was no significant p-value which means that for each response variable, the means of the groups are equal. Unfortunately, the assumptions are not likely to have been met as we do have independent observations but not random samples since the data was gathered mainly from one area. Multivariate normality of DVs is also not met as not even each group has 25 or more observations. Homogeneity, linear relationships, and multicolinearity assumptions are also most likely not met. Even though there was no significant difference in means, an ANOVA and post-hoc tests were performed for practice. However, since 6 total tests were done, the new alpha value was 0.0083 and there were no significant differences in means for cry duration based on family status.

Randomization Test

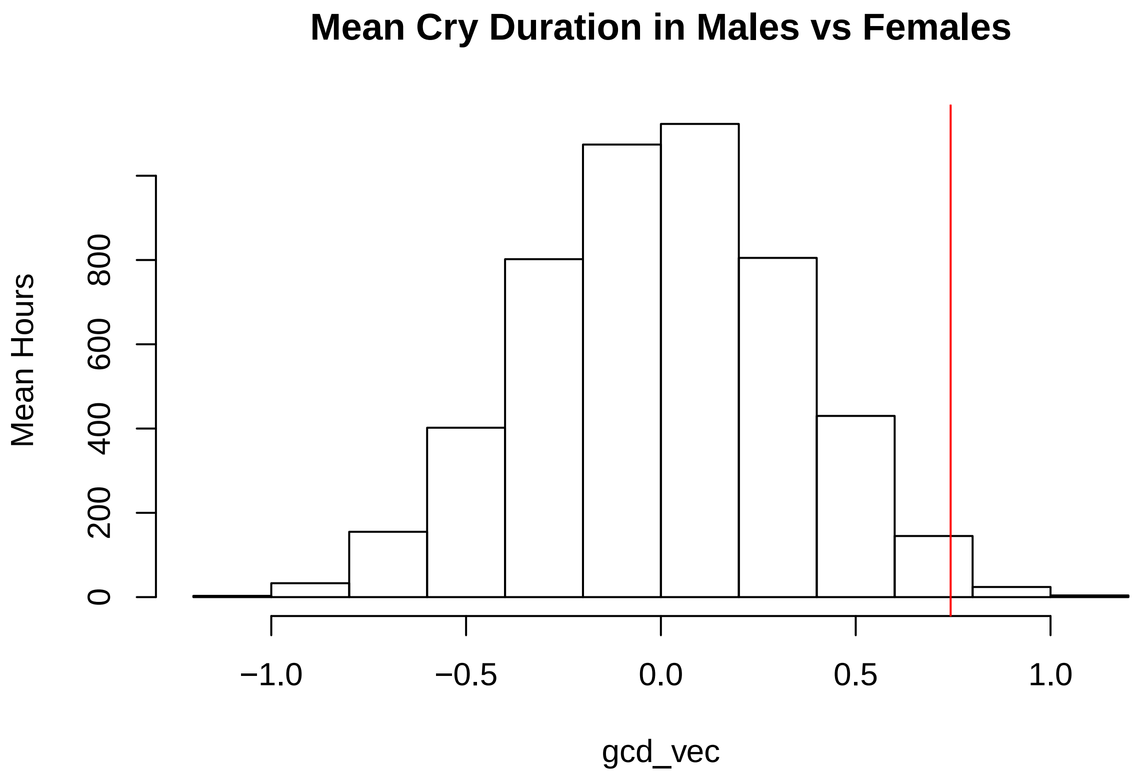
```r
data2 <- data %>% mutate(Baby.Gender = ifelse(as.character(Baby.Gender) ==
    "1", "Female", "Male")) %>% na.omit
data2 %>% group_by(Baby.Gender) %>% summarize(m = mean(crydur)) %>%
    summarize(diff(m))
```

```
## # A tibble: 1 x 1
##   `diff(m)`
##       <dbl>
## 1     0.744
```

```
gcd_vec <- vector()
for (i in 1:5000) {
    new <- data.frame(CryDur = sample(data2$crydur), Gender = data2$Baby.Gender)
    gcd_vec[i] <- mean(new[new$Gender == "Male", ]$CryDur) -
        mean(new[new$Gender == "Female", ]$CryDur)
}
mean(gcd_vec > 0.7436 | gcd_vec < -0.7436)
```

```
## [1] 0.023
```

```
{
    hist(gcd_vec, main = "Mean Cry Duration in Males vs Females",
        ylab = "Mean Hours")
    abline(v = 0.7436, col = "red")
}
```



The null hypothesis is that the mean cry duration is the same for males and females in this sample.The alternative hypothesis is that the mean cry duration is different for males versus females. After performing the test, we compute the two-tailed p-value which is 0.0248. Since the p-value is less than 0.05, we reject *the null hypothesis and there is a difference for males versus females in the mean cry duration. This is also visualized in the histogram. The mean cry duration for males is higher than the mean cry duration for females.*

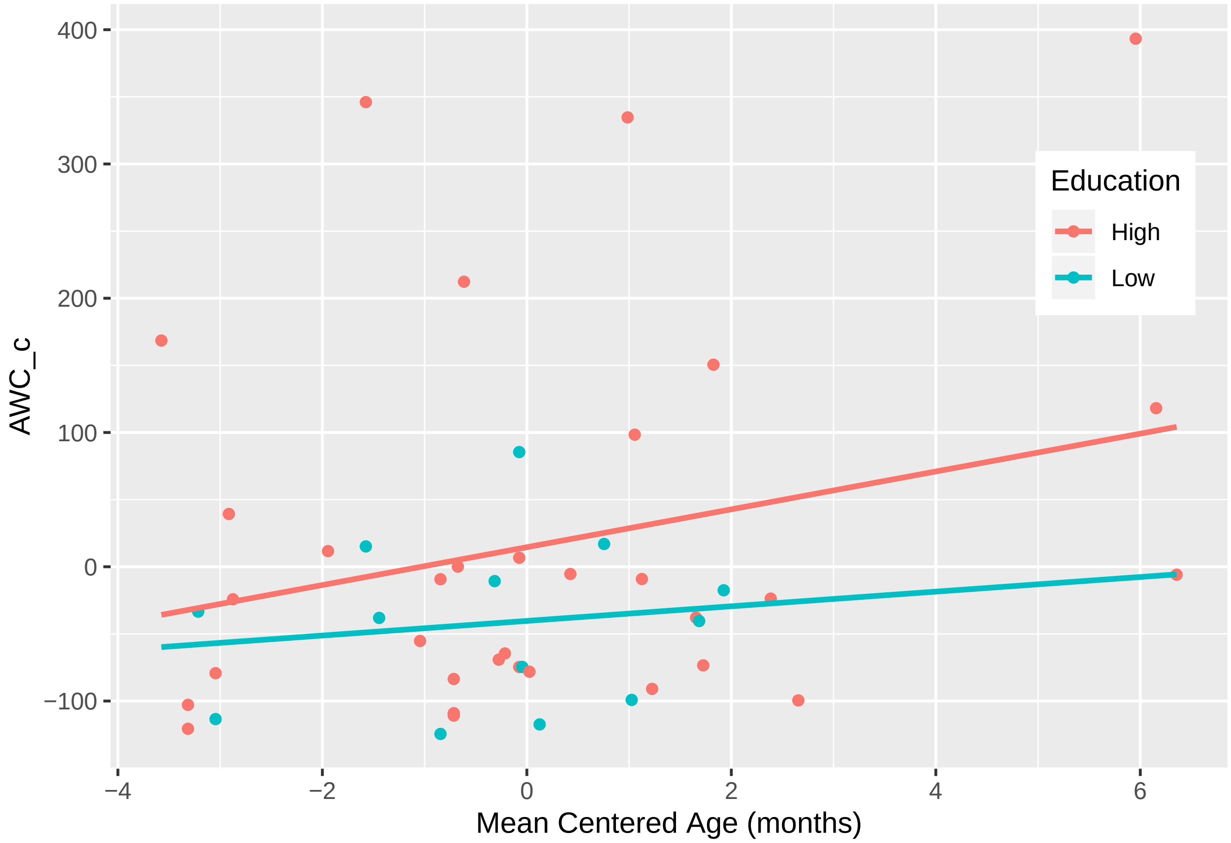
Linear Regression Model

```
data3 <- data2 %>% mutate(Age.M_c = data2$Age.months - mean(data2$Age.months)) %>%
    mutate(AWC_c = data2$AWC.Normalized - mean(data2$AWC.Normalized)) %>%
    mutate(Education = ifelse(as.character(Education) == "1" |
        as.character(Education) == "2" | as.character(Education) ==
        "3" | as.character(Education) == "4", "Low", "High"))
fit <- lm(AWC_c ~ Age.M_c * Baby.Gender * Education, data = data3)
summary(fit)
```

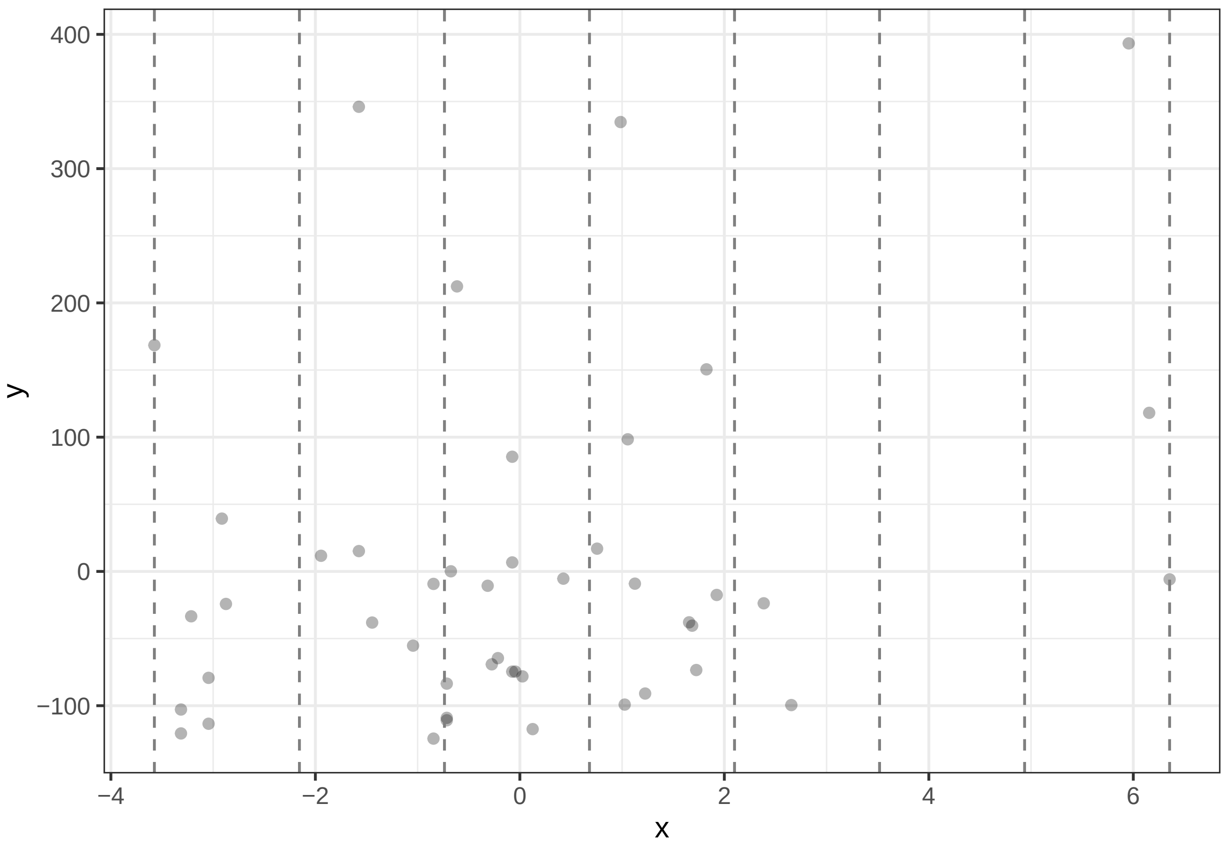
```
##
## Call:
## lm(formula = AWC_c ~ Age.M_c * Baby.Gender * Education, data = data3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -175.54  -67.76  -29.44   20.21  379.61
##
## Coefficients:
##                                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)                             7.233     31.312   0.231   0.8186
## Age.M_c                                25.895     12.015   2.155   0.0375 *
## Baby.GenderMale                         9.129     43.400   0.210   0.8345
## EducationLow                          -63.032     72.194  -0.873   0.3881
## Age.M_c:Baby.GenderMale               -23.754     17.119  -1.388   0.1733
## Age.M_c:EducationLow                  -14.988     42.201  -0.355   0.7244
## Baby.GenderMale:EducationLow           13.427     88.891   0.151   0.8807
## Age.M_c:Baby.GenderMale:EducationLow   15.447     51.321   0.301   0.7651
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 123.9 on 38 degrees of freedom
## Multiple R-squared:  0.1559, Adjusted R-squared:  0.0004254
## F-statistic: 1.003 on 7 and 38 DF,  p-value: 0.4446
```

For a female infant with the mean age whose caregiver has a high level of education, the predicted AWC is 7.233 words. Controlling for education and gender, for every one unit of increase in infant age, the AWC goes up by 25.895 words. For infants with average age and controlling for education, males have a predicted AWC that is 9.129 higher than females. For female infants with average age, caregivers with low education levels have a predicted AWC that is 63.032 words lower than caregivers with high education levels. Controlling for education, the slope for age on AWC is 23.754 lower for males compared to females. Controlling for gender, the slope for age on AWC is 14.988 lower for low compared to high education levels. Controlling for age, 13.427 AWC is accounted by the interaction between male infants of caregivers with low education levels. About 15.447 AWC is accounted for by the interaction between male infants of average age with caregivers of low education status.

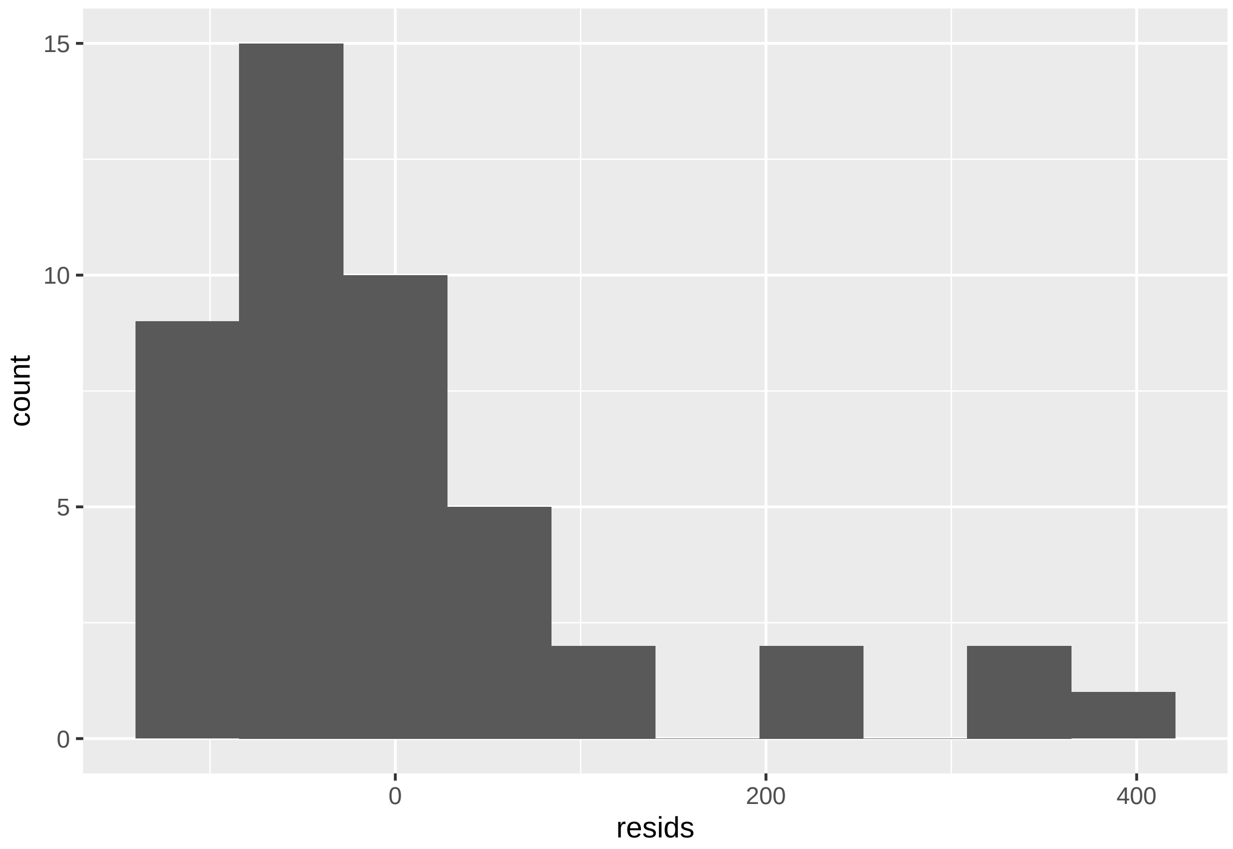
```
# regression plot
ggplot(data3, aes(x = Age.M_c, y = AWC_c, group = Education)) +
    geom_point(aes(color = Education)) + geom_smooth(method = "lm",
    se = F, fullrange = T, aes(color = Education)) + theme(legend.position = c(0.9,
    0.7)) + xlab("Mean Centered Age (months)")
```



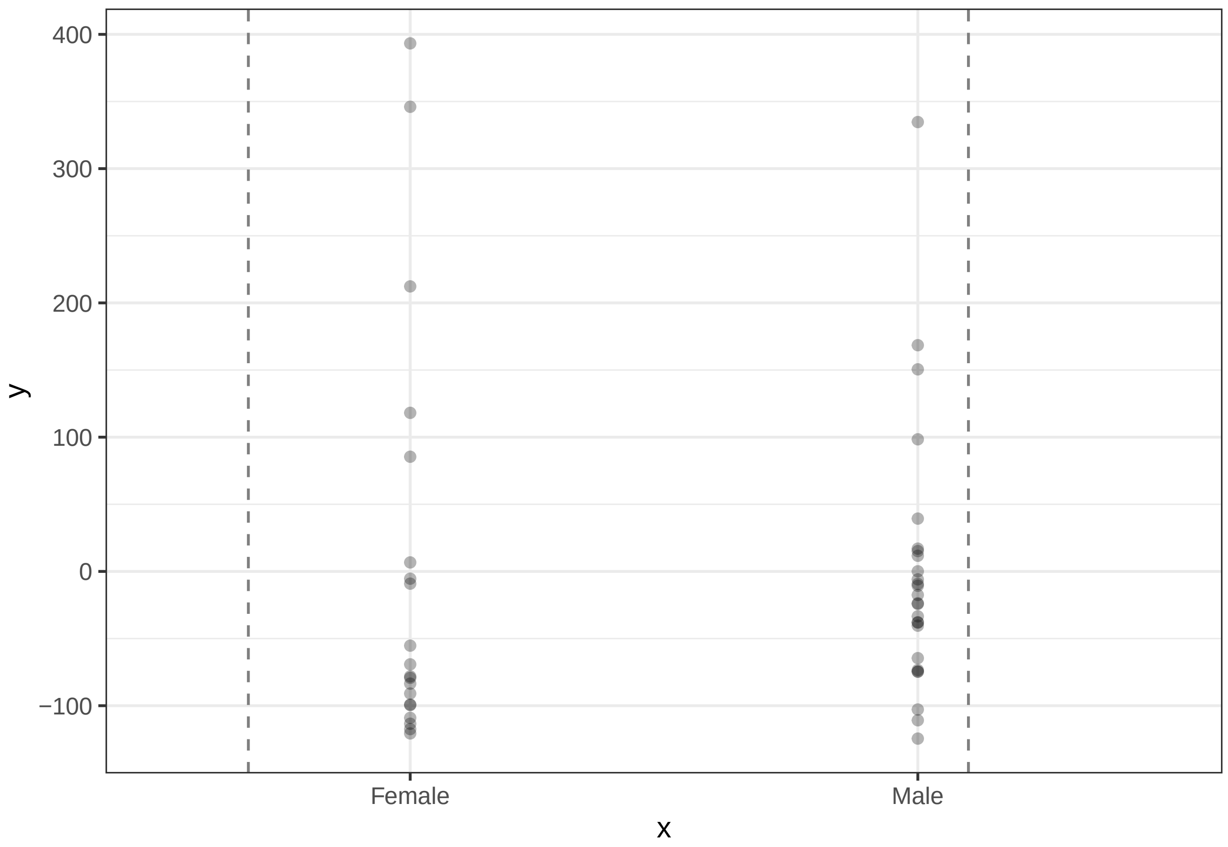
```
# linearity and normality for Age vs AWC
dat <- data.frame(x = data3$Age.M_c, y = data3$AWC_c)
breaks <- seq(min(dat$x), max(dat$x), len = 8)
ggplot(dat, aes(x, y)) + geom_point(alpha = 0.3) + theme_bw() +
    geom_vline(xintercept = breaks, lty = 2, color = "gray50")
```



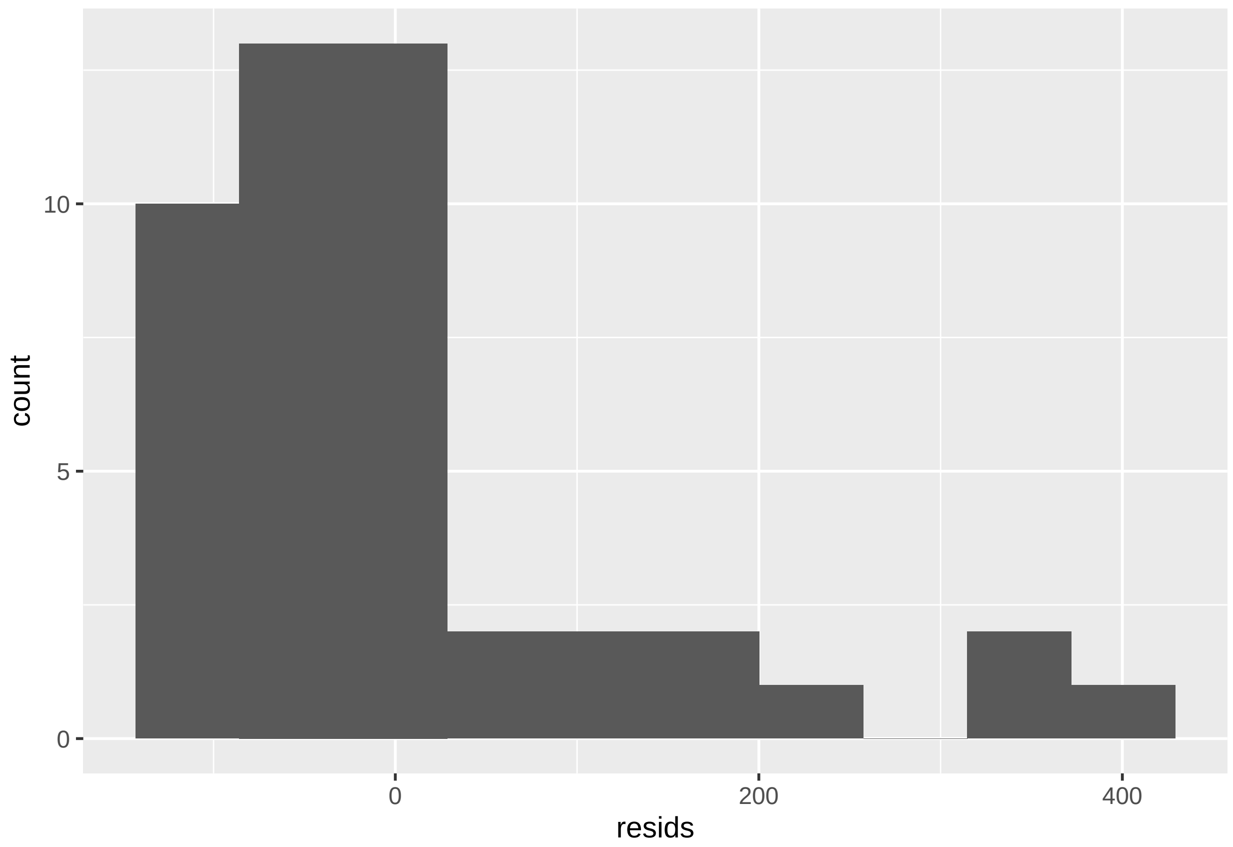
```
resids <- lm(y ~ x, data = dat)$residuals
ggplot() + geom_histogram(aes(resids), bins = 10)
```



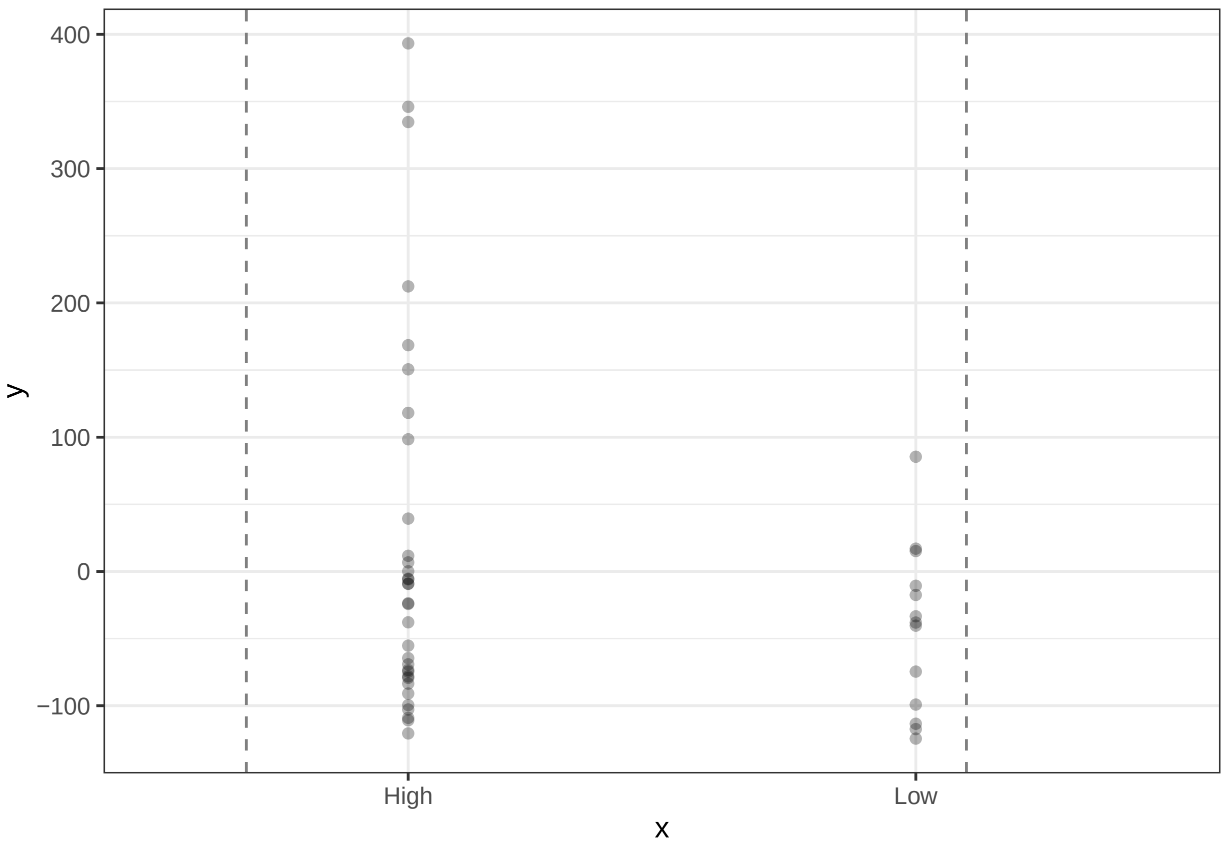
```
# linearity and normality for Gender vs AWC
dat <- data.frame(x = data3$Baby.Gender, y = data3$AWC_c)
ggplot(dat, aes(x, y)) + geom_point(alpha = 0.3) + theme_bw() +
    geom_vline(xintercept = breaks, lty = 2, color = "gray50")
```



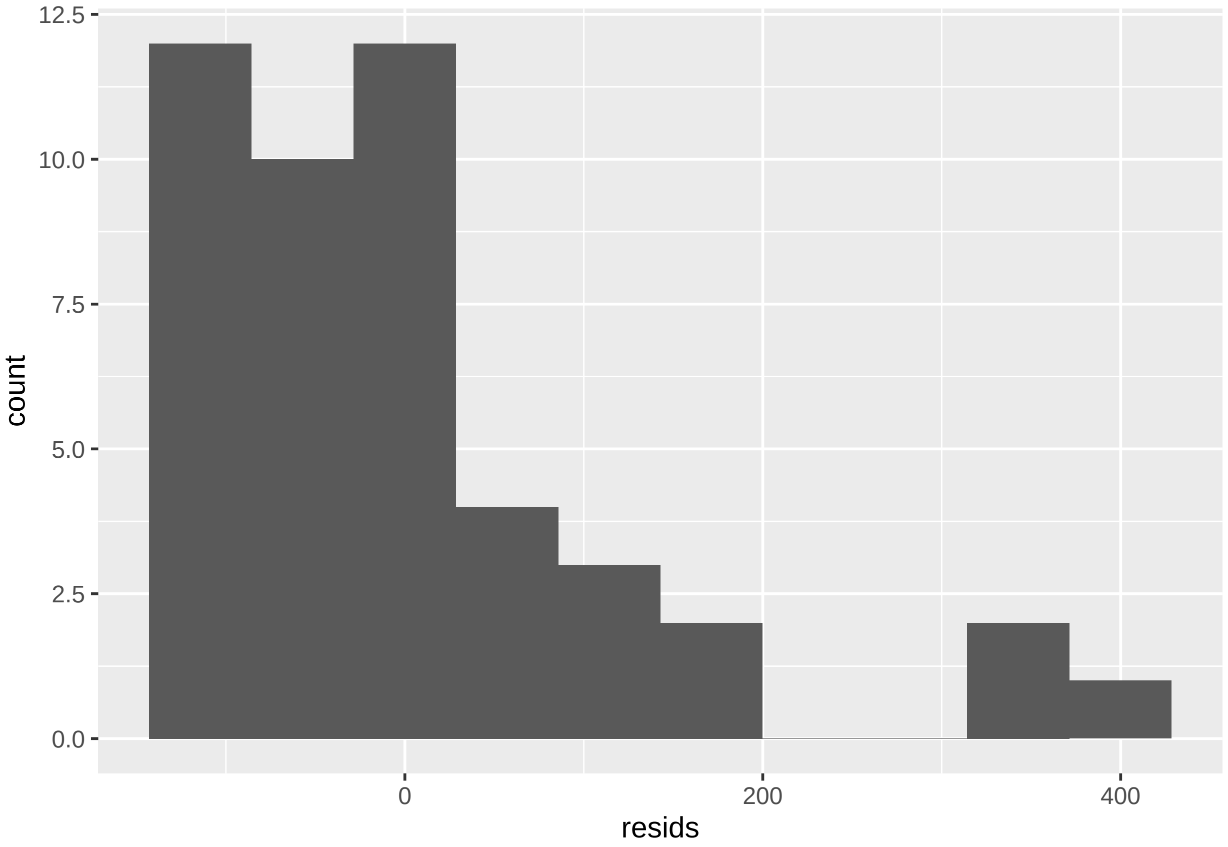
```
resids <- lm(y ~ x, data = dat)$residuals
ggplot() + geom_histogram(aes(resids), bins = 10)
```



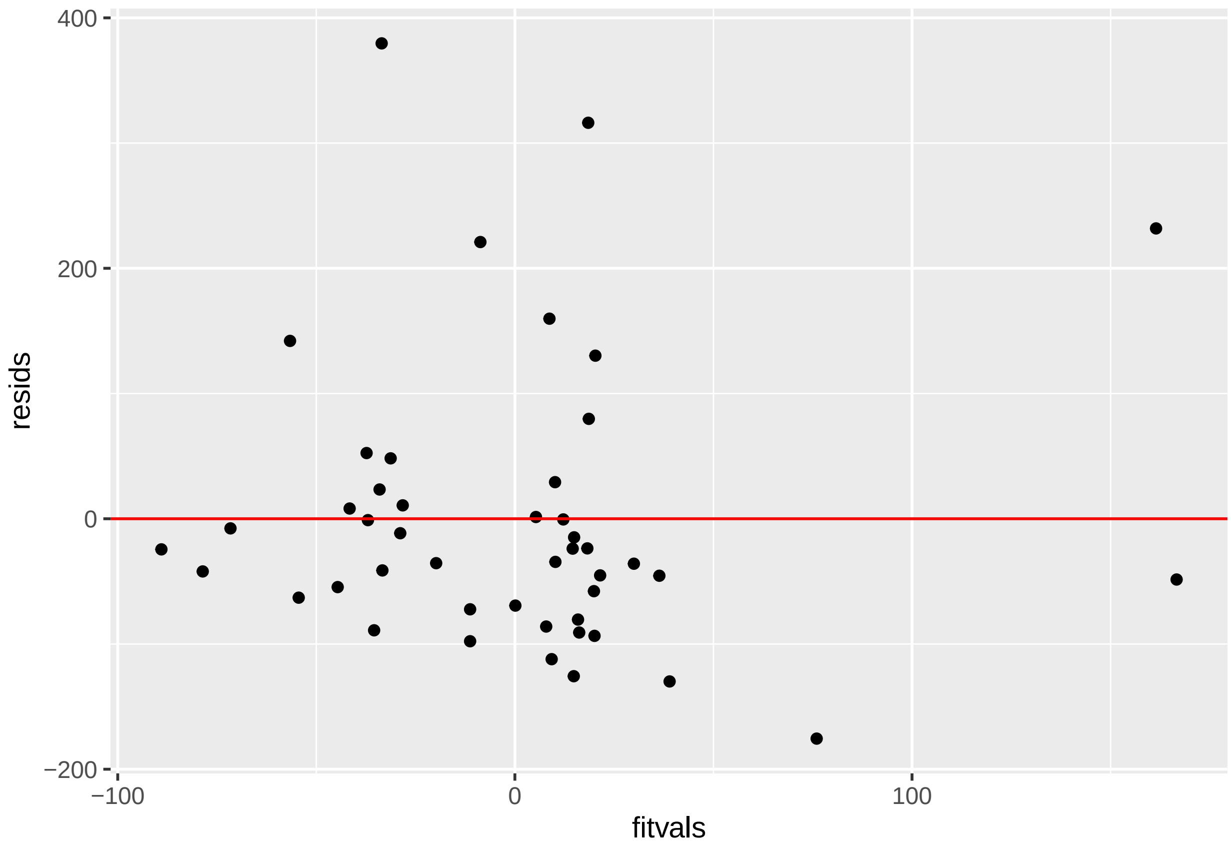
```
# linearity and normality for Education vs AWC
dat <- data.frame(x = data3$Education, y = data3$AWC_c)
ggplot(dat, aes(x, y)) + geom_point(alpha = 0.3) + theme_bw() +
    geom_vline(xintercept = breaks, lty = 2, color = "gray50")
```



```r
resids <- lm(y ~ x, data = dat)$residuals
ggplot() + geom_histogram(aes(resids), bins = 10)
```



```
# homoskedasticity
resids <- fit$residuals
fitvals <- fit$fitted.values
ggplot() + geom_point(aes(fitvals, resids)) + geom_hline(yintercept = 0,
    color = "red")
```



```
summary(fit)$coef
```

```
##                                        Estimate Std. Error    t value
## (Intercept)                            7.232556   31.31237  0.2309808
## Age.M_c                               25.895327   12.01527  2.1552018
## Baby.GenderMale                        9.128573   43.40001  0.2103357
## EducationLow                         -63.032390   72.19392 -0.8730983
## Age.M_c:Baby.GenderMale              -23.754138   17.11900 -1.3875888
## Age.M_c:EducationLow                 -14.988492   42.20146 -0.3551652
## Baby.GenderMale:EducationLow          13.426596   88.89097  0.1510457
## Age.M_c:Baby.GenderMale:EducationLow  15.447302   51.32128  0.3009922
##                                        Pr(>|t|)
## (Intercept)                          0.81856887
## Age.M_c                              0.03754559
## Baby.GenderMale                      0.83452943
## EducationLow                         0.38809354
## Age.M_c:Baby.GenderMale              0.17334617
## Age.M_c:EducationLow                 0.72442966
## Baby.GenderMale:EducationLow         0.88073870
## Age.M_c:Baby.GenderMale:EducationLow 0.76506132
```

```
coeftest(fit, vcov = vcovHC(fit))
```

```
##
```

```
## t test of coefficients:
##
##                                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)                            7.2326    41.2328  0.1754   0.8617
## Age.M_c                               25.8953    21.4975  1.2046   0.2358
## Baby.GenderMale                        9.1286    50.9736  0.1791   0.8588
## EducationLow                         -63.0324    95.1615 -0.6624   0.5117
## Age.M_c:Baby.GenderMale              -23.7541    23.8534 -0.9958   0.3256
## Age.M_c:EducationLow                 -14.9885   124.7649 -0.1201   0.9050
## Baby.GenderMale:EducationLow          13.4266   101.0130  0.1329   0.8950
## Age.M_c:Baby.GenderMale:EducationLow  15.4473   125.3412  0.1232   0.9026
```

The assumptions of linearity, normality, and homoskedasticity for this data are not met as shown by the plots. From the linear regression model, the only significant coefficient is the one for the average age. However, after the standard errors are corrected, no coefficients are significant. Additionally, the uncorrected SE for all variables and interactions are lower than the corrected SE. R^2 says that 15.6% of variability in AWC is explained by the model. However, the adjusted R^2 says that 0% of variation in AWC is explained by the model which means that the variation previously seen was due to chance.

Bootstrapped SEs

```r
samp_boot <- replicate(5000, {
    boot_dat <- sample_frac(data3, replace = T)
    fit <- lm(AWC_c ~ Age.M_c * Baby.Gender * Education, data = boot_dat)
    coef(fit)
})
samp_boot %>% t %>% as.data.frame %>% na.omit %>% summarize_all(sd)
```

```
##   (Intercept) Age.M_c Baby.GenderMale EducationLow Age.M_c:Baby.GenderMale
## 1    38.62479 21.8671        49.31137      75.4464                25.05704
##   Age.M_c:EducationLow Baby.GenderMale:EducationLow
## 1             262.3987                     84.81289
##   Age.M_c:Baby.GenderMale:EducationLow
## 1                             263.1749
```

The bootstrapped SEs are higher than the original SEs and the robust SEs for all variables and interactions. Since there is greater error, this most likely indicates that the data violates multiple assumptions and the p-values are not significant.

Logistic Regression

```r
data5 <- data3 %>% mutate(Caregiver.Race.Ethnicity = ifelse(as.character(Caregiver.Race.Ethnicity) ==
    "4", "White", "Non-White")) %>% mutate(Education = ifelse(Education ==
    "Low", 1, 0))
data5$C.Age_c <- data5$Caregiver.Age - mean(data5$Caregiver.Age)
fit2 <- glm(Education ~ Family.Annual.Income + Caregiver.Race.Ethnicity +
    C.Age_c, data = data5, family = "binomial")
coeftest(fit2)
```

```
##
## z test of coefficients:
```

```
##
##                                Estimate Std. Error z value Pr(>|z|)
## (Intercept)                     2.18723    1.19621  1.8285  0.06748 .
## Family.Annual.Income           -0.87073    0.35272 -2.4686  0.01357 *
## Caregiver.Race.EthnicityWhite -0.67029    0.91711 -0.7309  0.46486
## C.Age_c                        -0.19882    0.10325 -1.9257  0.05414 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

For a non-White caregiver of average age with zero family annual income, the log-odds is 2.187. For non-White caregivers and holding caregiver age constant, going up $1 in family annual income decreases log-odds by -0.871. Holding family annual income and caregiver age constant, being White decreases the log-odds by 0.670 as compared to being non-White. For non-White caregivers and holding family annual income constant, increasing 1 year of caregiver age decreases log-odds by 0.199. The only significant p-value is for family annual income which tells us that education significantly differs based on income.

```r
# confusion matrix
prob <- predict(fit2, type = "response")
pred <- ifelse(prob > 0.5, 1, 0)
table(prediction = pred, truth = data5$Education) %>% addmargins
```

```
##            truth
## prediction  0  1 Sum
##        0   30  4  34
##        1    3  9  12
##        Sum 33 13  46
```

```r
(30 + 9)/46  #Accuracy
```

```
## [1] 0.8478261
```

```r
9/13  #TPR
```

```
## [1] 0.6923077
```

```r
30/33  #TNR
```

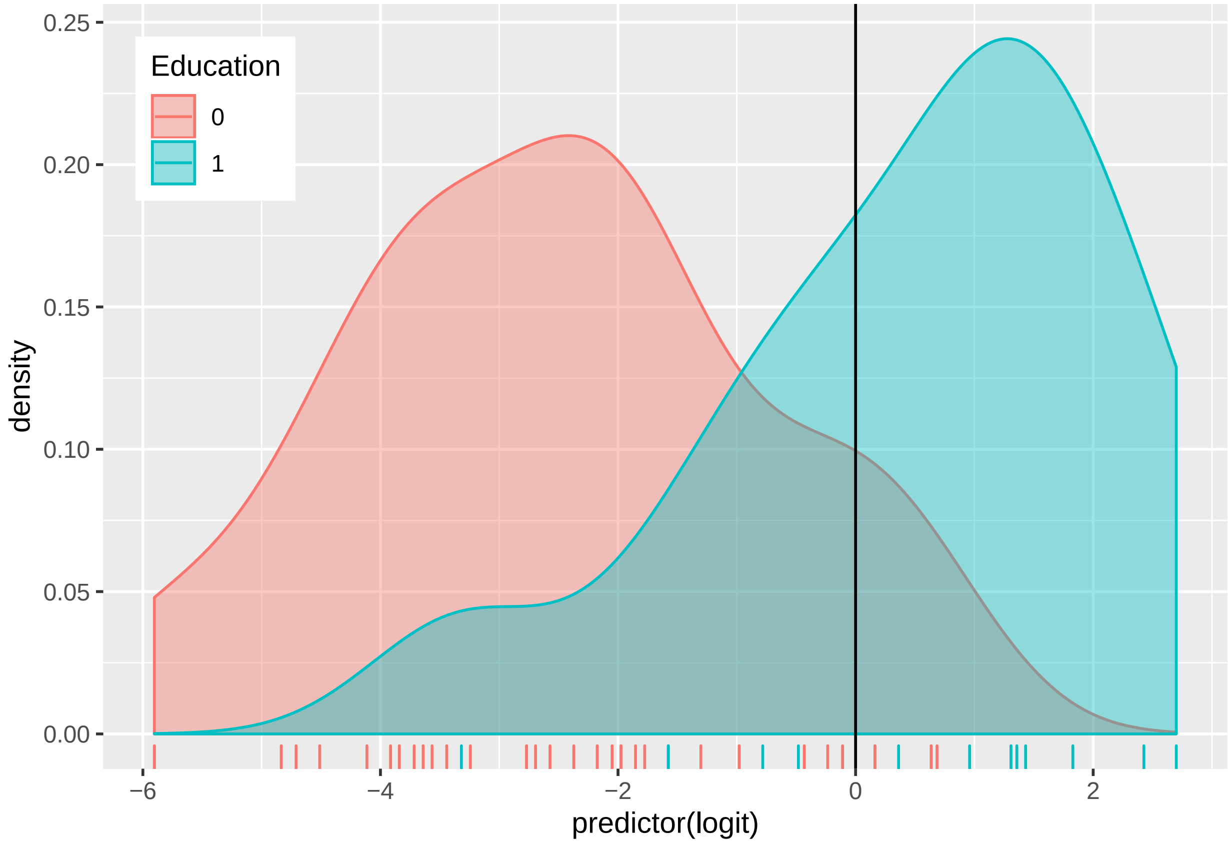
```
## [1] 0.9090909
```

```r
9/12  #PPV
```

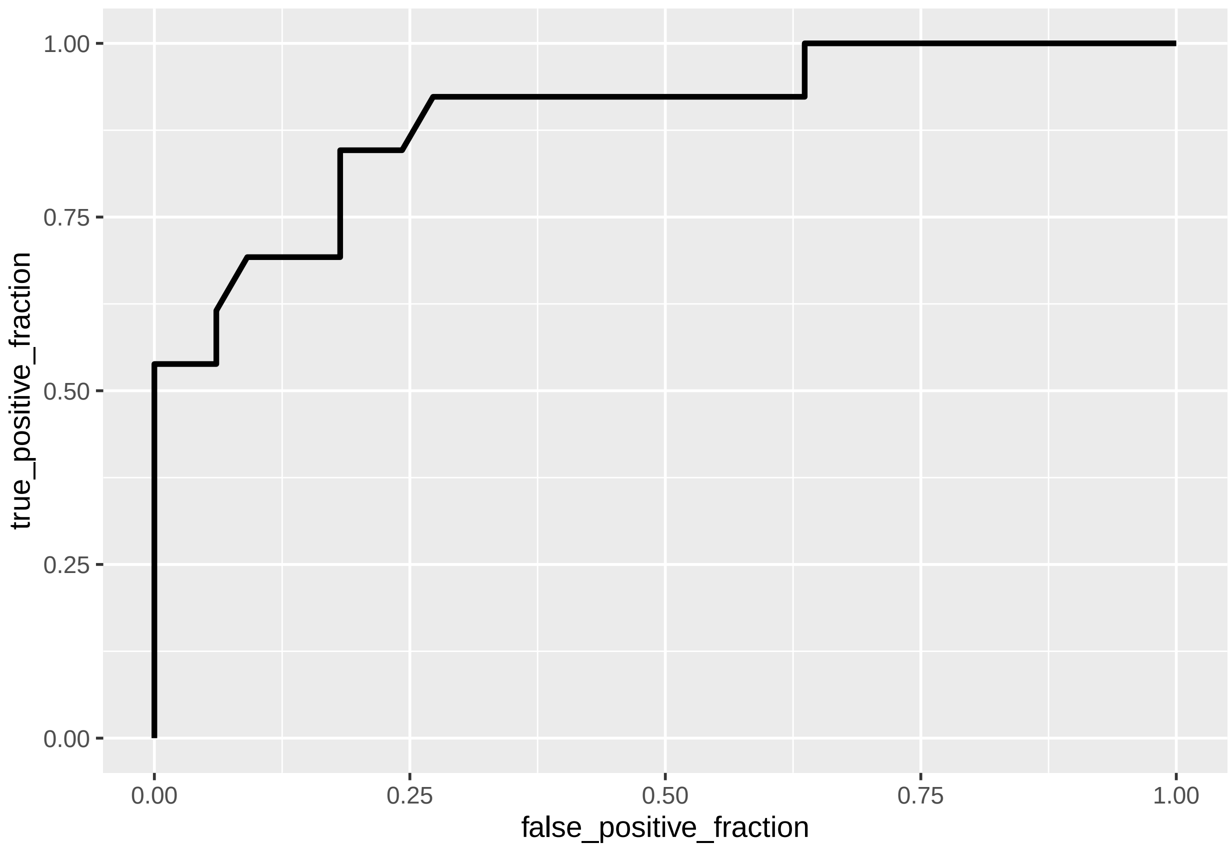
```
## [1] 0.75
```

This regression model has 84.8% accuracy of predicting low versus high education of a caregiver based on family annual income, caregiver age, and caregiver race. The probability of predicting a caregiver has a low education level, the sensitivity, is 69.2%. The probability of predicting a high level education, the specificity, is 90.9%. The percentage of caregivers who are classified as having a low education level and do have a low education level is 75.0%.

```
# density of log-odds
data5$logit <- predict(fit2, type = "link")
data5$Education <- as.factor(data5$Education)
data5 %>% ggplot(aes(logit, color = Education, fill = Education)) +
    geom_density(alpha = 0.4) + theme(legend.position = c(0.1,
    0.85)) + geom_vline(xintercept = 0) + xlab("predictor(logit)") +
    geom_rug(aes(logit, color = Education))
```



```
# ROC & AUC
data5 <- data5 %>% mutate(Education = ifelse(Education == 0,
    "High", "Low"))
ROCplot <- ggplot(data5) + geom_roc(aes(d = Education, m = prob),
    n.cuts = 0)
ROCplot
```



```
calc_auc(ROCplot)
```

```
##   PANEL group       AUC
## 1     1    -1 0.8927739
```

The probability that a randomly selected caregiver with a low education level has a higher predicted probability than a randomly selected person with a high education level, the AUC, is 89.3%, which is good but not the best.

```
# repeated random sub-sampling CV
data5 <- data5 %>% mutate(Education = ifelse(as.factor(Education) ==
    "Low", 1, 0))
fraction <- 0.5
train_n <- floor(fraction * nrow(data5))
iter <- 500
diags <- NULL
for (i in 1:iter) {
    train_index <- sample(1:nrow(data5), size = train_n)
    train <- data5[train_index, ]
    test <- data5[-train_index, ]
    truth <- test$Education
    fit <- glm(Education ~ Family.Annual.Income + Caregiver.Race.Ethnicity +
        C.Age_c, data = train, family = "binomial")
    probs <- predict(fit, newdata = test, type = "response")
    diags <- rbind(diags, class_diag(probs, truth))
```

```r
}
diags <- diags %>% na.omit
summarize_all(diags, mean)
```

```
##         acc      sens      spec       ppv       auc
## 1 0.7778261 0.6020532 0.8570603 0.6485783 0.8204162
```

The average out-of-sample accuracy, sensitivity, and recall is 0.78, 0.60, and 0.65, respectively.

LASSO Regression

```r
y <- as.matrix(data5$crydur)
data6 <- data5 %>% select(-Participant, -Duration_Secs.sum, -AWC_COUNT.sum,
    -AWC_c, -CT_COUNT.sum, -Age.M_c, -C.Age_c, -logit)
x <- model.matrix(crydur ~ ., data = data6)
# head(x)
cv <- cv.glmnet(x, y)
lasso <- glmnet(x, y, lambda = cv$lambda.1se)
coef(lasso)
```

```
## 16 x 1 sparse Matrix of class "dgCMatrix"
##                                          s0
## (Intercept)                     -1.63772535
## (Intercept)                      .
## Total.Recording.Hours            .
## AWC.Normalized                   .
## CT.Normalized                    .
## Age.months                       .
## Family.Status                    .
## Family.Annual.Income             .
## Education                       -0.11163866
## Occupation                       .
## Caregiver.Race.EthnicityWhite    .
## Language                         .
## cryfreq                          0.04712618
## Baby.GenderMale                  .
## Caregiver.Age                    .
## Infant.Race.Ethnicity            .
```

The LASSO regression says that education and cry frequency are the most important predictors of cry duration.

```r
# 10-fold CV
k = 10
dat1 <- data6[sample(nrow(data6)), ]
folds <- cut(seq(1:nrow(data6)), breaks = k, labels = F)
diags <- NULL
for (i in 1:k) {
    train <- dat1[folds != i, ]
    test <- dat1[folds == i, ]
    truth <- test$crydur
    fit <- glm(crydur ~ Education + cryfreq, data = train)
    probs <- predict(fit, newdata = test, type = "response")
    diags <- rbind(diags, class_diag(probs, truth))
}
diags %>% summarize_all(mean)
```

```
##    acc sens spec   ppv auc
## 1 0.24    1  0.1 0.225   1
```

```
fit4 <- lm(crydur ~ Education + cryfreq, data = data6)
summary(fit4)
```

```
##
## Call:
## lm(formula = crydur ~ Education + cryfreq, data = data6)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.16215 -0.49704 -0.01783  0.36064  1.80689
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.776340   0.734904  -3.778 0.000482 ***
## Education   -0.538957   0.238627  -2.259 0.029038 *
## cryfreq      0.062700   0.008697   7.210 6.41e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6898 on 43 degrees of freedom
## Multiple R-squared:  0.6369, Adjusted R-squared:  0.6201
## F-statistic: 37.72 on 2 and 43 DF,  p-value: 3.463e-10
```

```
fit5 <- lm(crydur ~ ., data = data6)
summary(fit5)
```

```
##
## Call:
## lm(formula = crydur ~ ., data = data6)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.16718 -0.25116 -0.06854  0.23952  1.51151
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   -1.6445948  1.3287554  -1.238   0.2251
## Total.Recording.Hours         -0.0032217  0.0047831  -0.674   0.5056
## AWC.Normalized                -0.0009718  0.0017085  -0.569   0.5736
## CT.Normalized                  0.0462738  0.1011231   0.458   0.6504
## Age.months                    -0.0699397  0.0478916  -1.460   0.1542
## Family.Status                 -0.0740660  0.0907167  -0.816   0.4205
## Family.Annual.Income          -0.0825739  0.0907370  -0.910   0.3698
## Education                     -0.5763642  0.3256890  -1.770   0.0866 .
## Occupation                    -0.2297364  0.1209190  -1.900   0.0668 .
## Caregiver.Race.EthnicityWhite  0.0396918  0.2891598   0.137   0.8917
## Language                      -0.1799802  0.1305068  -1.379   0.1777
## cryfreq                        0.0636273  0.0105234   6.046 1.08e-06 ***
## Baby.GenderMale                0.2431132  0.2347887   1.035   0.3085
## Caregiver.Age                  0.0220908  0.0263931   0.837   0.4090
## Infant.Race.Ethnicity         -0.0631688  0.0751679  -0.840   0.4071
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6394 on 31 degrees of freedom
## Multiple R-squared:  0.7751, Adjusted R-squared:  0.6735
## F-statistic: 7.632 on 14 and 31 DF,  p-value: 1.392e-06
```

The residual standard error for the model using LASSO coefficients is 0.6898 which is slightly higher than the residual standard error for a model using all coefficients which is 0.6395. This means that the model using all coefficients has a slight better fit than the model using LASSO coefficients. This is an unusual answer, but the results from the 10-fold CV are also unusual. This is most likely due to the non-normal distribution of the data into the train and test sets. Furthermore, many of the coefficients did not have enough data for all categorical levels.