BIOS 545: Statistical Analysis

Dane Van Domelen

Department of Biostatistics and Bioinformatics
Rollins School of Public Health
Emory University
Atlanta, GA

February 22, 2017

Statistical analysis in R

- Analyzing data is what (most) statisticians do.
- General procedure:
    1. Load data into R.
    2. Clean data (boring am I right??)
    3. **Try to answer a research question.**

Format of this lecture

- **Basic idea:** Go through various research questions on a particular dataset to illustrate data analysis tools in R.
- **Dataset:** NHANES physical activity.
- **General procedure for each research question:**

  (1) Visualize data (review!)

  (2) Estimate parameter of interest.

  (3) Perform hypothesis test.

NHANES dataset

- National Health And Nutrition Examination Survey
  - Cross-sectional study in the US.
  - $n \approx 10,000$ in each 2-year cycle.
  - Demographics, questionnaires, lab tests, etc.
  - Publicly available!
    `https://www.cdc.gov/nchs/nhanes/`

Putting dataset together (FYI)

```r
# Load nhanesaccel and nhanesdata packages
install.packages("accelerometry")
install.packages("nhanesaccel", repos = "http://R-Forge.R-project.org")
install.packages("nhanesdata", repos = "http://R-Forge.R-project.org")
library("accelerometry")
library("nhanesaccel")
library("nhanesdata")

# Process NHANES 2003-2006 data
nhanes.pa <- nhanes.accel.process(waves = 1, valid.week.days = 5, valid.weekend.days = 2,
                                  weekday.weekend = TRUE, brevity = 2)

# Merge in demographics and body measurements datasets
data(demo_c)
names(demo_c) <- tolower(names(demo_c))

data(bmx_c)
names(bmx_c) <- tolower(names(bmx_c))

nhanes <- merge(x = demo_c, y = bmx_c)
nhanes <- merge(x = nhanes, y = nhanes.pa)

# Keep only variables of interest
nhanes <- nhanes[, c("seqn", "riagendr", "ridageyr", "ridreth2", "indfmpir",
                     "bmxbmi", "bmxwaist", "cpm", "wk_cpm", "we_cpm", "guideline_min")]

# Save dataset
save(nhanes, file = "nhanes.rda")
```

Initial look at dataset

```
# Download dataset from website
load("nhanes.rda")

# Look at data structure
class(nhanes)

## [1] "data.frame"

dim(nhanes)

## [1] 7176   11
```

Initial look at dataset

```
head(nhanes, n = 4)
```

```
##    seqn riagendr ridageyr ridreth2 indfmpir bmxbmi bmxwaist      cpm   wk_cp
## 1 21005        1       19        2     2.44  50.85    135.9 609.5422 649.294
## 2 21006        2       16        2     2.47  20.78     73.6 145.4638 171.183
## 3 21007        2       14        1     1.60  18.43     69.5 412.5396 372.763
## 4 21008        1       17        2     2.75  20.65     74.7 273.8358 287.469
##     we_cpm guideline_min
## 1 530.0384      19.33333
## 2 119.7441       2.75000
## 3 492.0927       0.00000
## 4 246.5692       0.00000
```

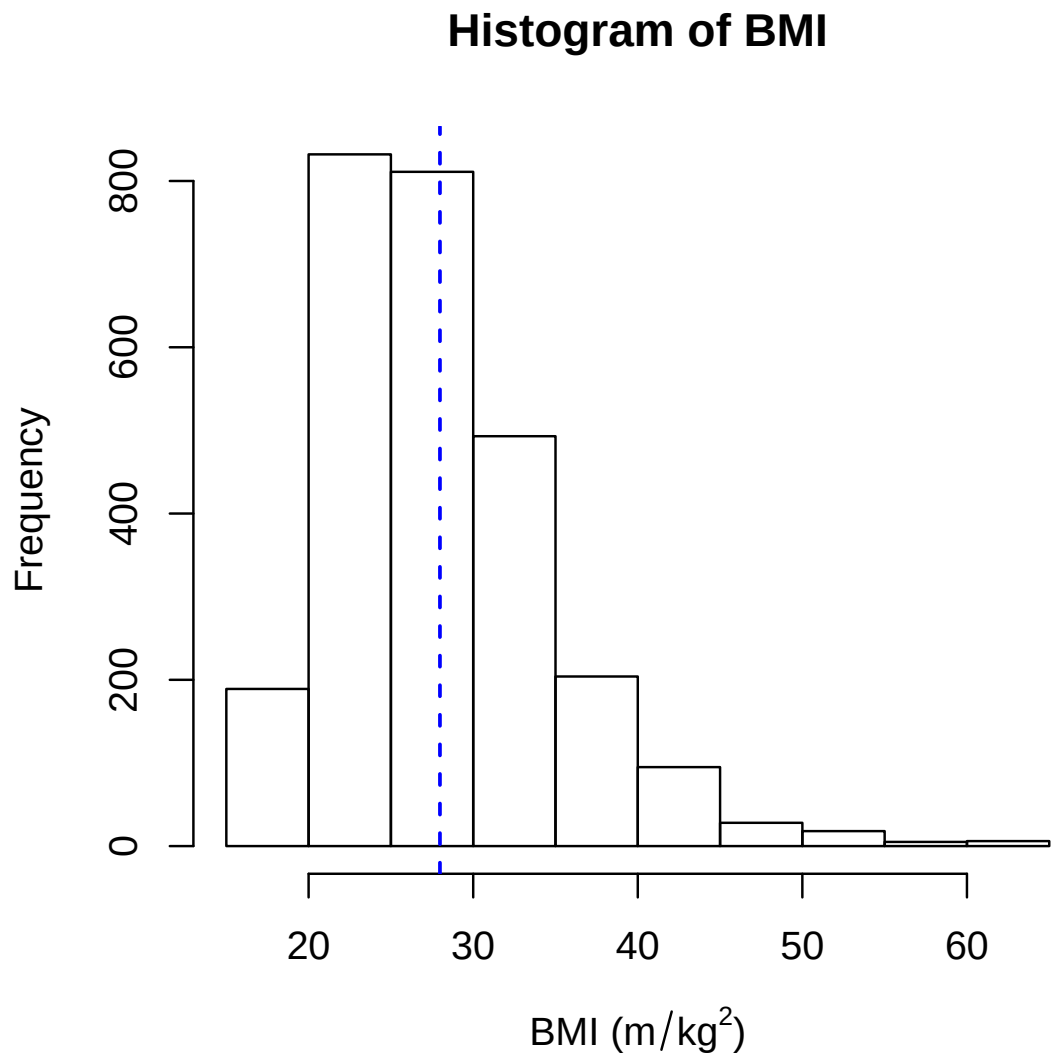
Research question: What is the mean BMI of American adults age 18-50?

Visualization

```r
# Get subset of participants age 18-50
nhanes.adults <- subset(nhanes, ridageyr >= 18 & ridageyr <= 50)

# Create histogram
hist(nhanes.adults$bmxbmi, main = "Histogram of BMI",
     xlab = expression(paste("BMI (", m/kg^2, ")")))
bmi.mean <- mean(nhanes.adults$bmxbmi, na.rm = T)
abline(v = bmi.mean, col = "blue", lty = 2, lwd = 1.5)
```

Visualization



Parameter estimation

- Statistical setup:
  - Let X_i = BMI for i^{th} participant, i = 1, ..., n
  - Assume $X_i \overset{iid}{\sim} (\mu,\ \sigma^2)$
- Estimators:

  $\hat{\mu} = \bar{X}$

  95% CI for μ : $\bar{X} \pm \frac{t_{1-\alpha/2,n-1}s}{\sqrt{n}}$

Parameter estimation

```r
# Calculate sample mean
(x.bar <- mean(nhanes.adults$bmxbmi, na.rm = T))

## [1] 27.97934
```

```r
# Calculate 95% CI manually
s <- sd(nhanes.adults$bmxbmi, na.rm = T)
n <- sum(!is.na(nhanes.adults$bmxbmi))
t <- qt(p = 0.975, df = n - 1)
c(x.bar - t * s / sqrt(n), x.bar + t * s / sqrt(n))

## [1] 27.72316 28.23551
```

Parameter estimation

```
# Calculate 95% CI using built-in R function
t.test(nhanes.adults$bmxbmi)


##
##  One Sample t-test
##
## data:  nhanes.adults$bmxbmi
## t = 214.16, df = 2680, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  27.72316 28.23551
## sample estimates:
## mean of x
##  27.97934
```

Hypothesis testing

Suppose we want to test:

$$H_0 : \mu = 25$$

$$H_A : \mu \neq 25$$

Hypothesis testing

```
(ttest.fit <- t.test(nhanes.adults$bmxbmi, mu = 25))
```

```
##
##  One Sample t-test
##
## data:  nhanes.adults$bmxbmi
## t = 22.805, df = 2680, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 25
## 95 percent confidence interval:
##  27.72316 28.23551
## sample estimates:
## mean of x
##  27.97934
```

Hypothesis testing

```
names(ttest.fit)

## [1] "statistic"   "parameter"   "p.value"     "conf.int"    "estimate"
## [6] "null.value"  "alternative" "method"      "data.name"

ttest.fit$estimate

## mean of x
##  27.97934

ttest.fit$conf.int

## [1] 27.72316 28.23551
## attr(,"conf.level")
## [1] 0.95

ttest.fit$p.value

## [1] 2.344613e-105
```

Research question: In American adults age 18-50, is the population mean BMI for males the same as for females?

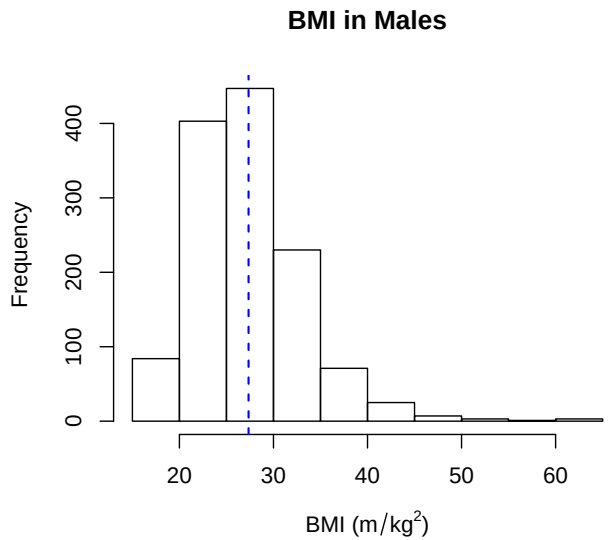
Visualization

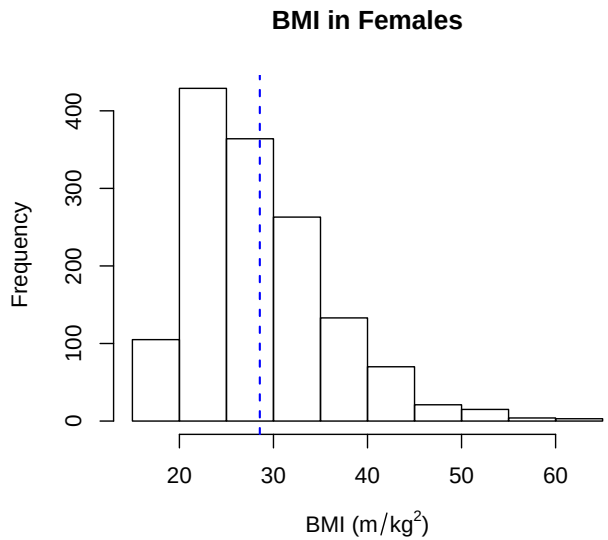
```r
# Create histogram of BMI by sex
par(mfrow = c(2, 1))

locs.m <- which(nhanes.adults$riagendr == 1)
hist(nhanes.adults$bmxbmi[locs.m], main = "BMI in Males",
     xlab = expression(paste("BMI (", m/kg^2, ")")))
bmi.mean.m <- mean(nhanes.adults$bmxbmi[locs.m], na.rm = T)
abline(v = bmi.mean.m, col = "blue", lty = 2, lwd = 1.5)

locs.f <- which(nhanes.adults$riagendr == 2)
hist(nhanes.adults$bmxbmi[locs.f], main = "BMI in Females",
     xlab = expression(paste("BMI (", m/kg^2, ")")))
bmi.mean.f <- mean(nhanes.adults$bmxbmi[locs.f], na.rm = T)
abline(v = bmi.mean.f, col = "blue", lty = 2, lwd = 1.5)
```

Visualization





Parameter estimation

- Statistical setup:
  - Let $X_{m,i}$ = BMI for i^{th} male, $i = 1, ..., n_m$
  - Let $X_{f,j}$ = BMI for j^{th} female, $j = 1, ..., n_f$
  - Assume $X_{m,i} \overset{iid}{\sim} (\mu_m,\ \sigma_m^2)$ and $X_{f,j} \overset{iid}{\sim} (\mu_f,\ \sigma_f^2)$
- Parameters/Estimators:

  $\mu_\Delta = \mu_m - \mu_f$

  $\widehat{\mu_\Delta} = \bar{X}_m - \bar{X}_f$

  95% CI based on t-distribution $\Rightarrow$ 2 versions

Parameter estimation

```
# Fit two-sample t-test by giving t.test two vectors
(ttest.fit <- t.test(nhanes.adults$bmxbmi[locs.m],
                     nhanes.adults$bmxbmi[locs.f]))


##
##  Welch Two Sample t-test
##
## data:  nhanes.adults$bmxbmi[locs.m] and nhanes.adults$bmxbmi[locs.f]
## t = -4.6454, df = 2625.7, p-value = 3.561e-06
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.7015268 -0.6914384
## sample estimates:
## mean of x mean of y
##  27.35142  28.54790
```

Parameter estimation

```r
# Fit two-sample t-test using formula notation (easier!)
(ttest.fit <- t.test(bmxbmi ~ riagendr, data = nhanes.adults))


##
##  Welch Two Sample t-test
##
## data:  bmxbmi by riagendr
## t = -4.6454, df = 2625.7, p-value = 3.561e-06
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.7015268 -0.6914384
## sample estimates:
## mean in group 1 mean in group 2
##        27.35142        28.54790
```

Hypothesis testing

Already saw results for two-sample t-test:

$$H_0 : \mu_\Delta = 0$$

$$H_A : \mu_\Delta \neq 0$$

$\Rightarrow$ Can also test whether μ_Δ equals some non-zero value, but this is less common.

In-class activity

(1) Find out whether we assumed equal variance.

(2) Decide whether we *should* assume equal variance.

(3) Test $H_0 : \mu_\Delta = -1$, using appropriate test.

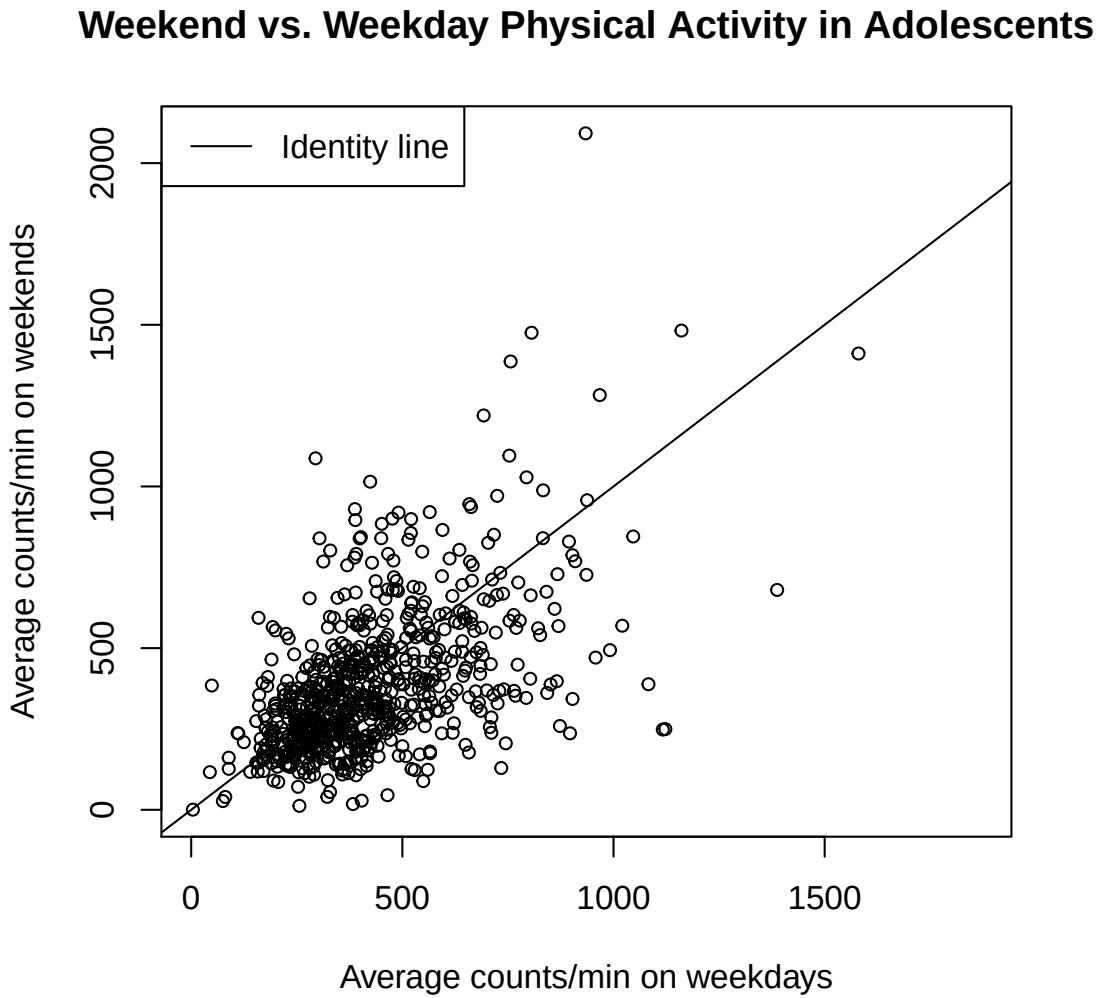
Research question: Are American adolescents age 13-17 more physically active on weekdays, or on weekend days?

Visualization #1: Scatterplot

```r
# Get subset of data for adolescents age 13-17
nhanes.adol <- subset(nhanes, ridageyr >= 13 & ridageyr <= 17)

# Plot weekend physical activity vs. weekday physical activity
plot(nhanes.adol$wk_cpm, nhanes.adol$we_cpm,
     main = "Weekend vs. Weekday Physical Activity in Adolescents")
abline(a = 0, b = 1)
legend("topleft", lty = 1, col = "black", legend = "Identity line")
```

Visualization #1: Scatterplot

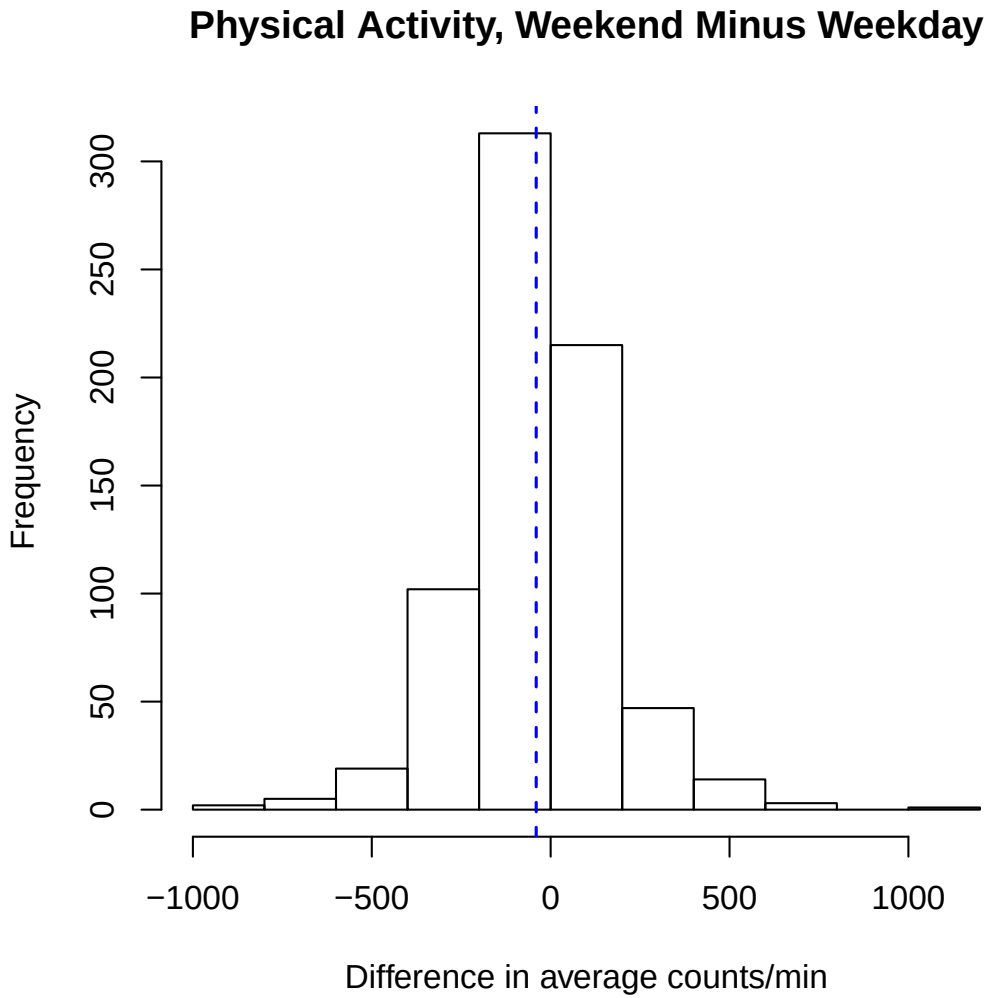


Visualization #2: Histogram

```r
# Calculate difference between weekend and weekday PA for each participant
nhanes.adol$cpm_diff <- nhanes.adol$we_cpm - nhanes.adol$wk_cpm

# Create histogram of differences
hist(nhanes.adol$cpm_diff,
     main = "Physical Activity, Weekend Minus Weekday",
     xlab = "Difference in average counts/min")
cpm.diff.mean <- mean(nhanes.adol$cpm_diff, na.rm = T)
abline(v = cpm.diff.mean, col = "blue", lty = 2, lwd = 1.5)
```

Visualization #2: Histogram



Parameter estimation

- Statistical setup:
  - Let $X_{d,i}$ = Difference between average weekend physical activity and average weekday physical activity for i^{th} participant, $i = 1, ..., n$.
  - Assume $X_{d,i} \overset{iid}{\sim} (\mu_d,\ \sigma_d^2)$
- Estimators:

  $\widehat{\mu_d} = \bar{X}_d$

  95% CI for μ_d : $\bar{X}_d \pm \frac{t_{1-\alpha/2,n-1} s_d}{\sqrt{n}}$

Parameter estimation

```r
# Fit paired t-test by giving t.test two vectors
t.test(nhanes.adol$we_cpm, nhanes.adol$wk_cpm, paired = T)


##
##  Paired t-test
##
## data:  nhanes.adol$we_cpm and nhanes.adol$wk_cpm
## t = -5.3711, df = 720, p-value = 1.057e-07
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -55.16344 -25.63108
## sample estimates:
## mean of the differences
##               -40.39726
```

Parameter estimation

```r
# Fit paired t-test by giving t.test single vector of differences
t.test(nhanes.adol$cpm_diff)
```

```
##
##  One Sample t-test
##
## data:  nhanes.adol$cpm_diff
## t = -5.3711, df = 720, p-value = 1.057e-07
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -55.16344 -25.63108
## sample estimates:
## mean of x
## -40.39726
```

Hypothesis testing

Already saw results for paired t-test:

$$H_0 : \mu_d = 0$$

$$H_A : \mu_d \neq 0$$

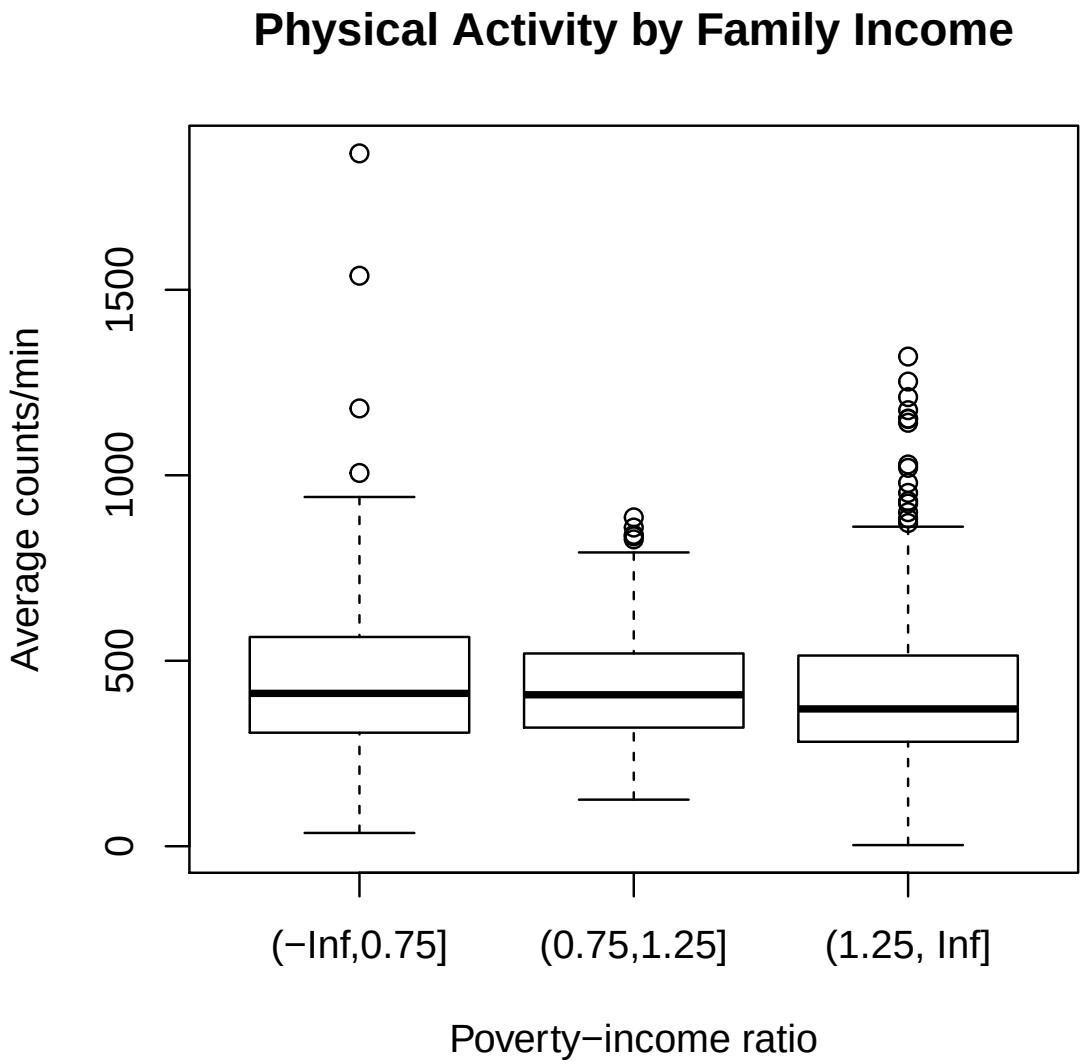
Research question: Does physical activity differ by family income level (low, medium, high) in American adolescents?

Visualization

```r
# Create three categories of poverty income ratio variable
nhanes.adol$pir.f <- cut(nhanes.adol$indfmpir,
                         breaks = c(-Inf, 0.75, 1.25, Inf))

# Create boxplot of physical activity by PIR
boxplot(cpm ~ pir.f, data = nhanes.adol,
        main = "Physical Activity by Family Income",
        ylab = "Average counts/min", xlab = "Poverty-income ratio")
```

Visualization



Parameter estimation

- Statistical setup:
  - Let $X_{i,j}$ = Average physical activity for j^{th} participant in i^{th} PIR group, $i = 1, 2, 3$; $j = 1, ..., n_i$
  - Assume $X_{i,j} \overset{ind}{\sim} (\mu_i,\ \sigma^2)$, $i = 1, 2, 3$
- Estimators:

  $\hat{\mu}_i = \bar{X}_i$

  95% CI for each μ_i : $\bar{X}_i \pm \frac{t_{1-\alpha/2, n_i-1} s_i}{\sqrt{n_i}}$

Parameter estimation

```
# Point estimates for mu's
tapply(nhanes.adol$cpm, nhanes.adol$pir.f,
       function(x) mean(x, na.rm = T))


## (-Inf,0.75] (0.75,1.25] (1.25, Inf]
##    453.5996    424.4943    414.3118
```

Parameter estimation

```r
# Interval estimates for mu's
tapply(nhanes.adol$cpm, nhanes.adol$pir.f,
       function(x) t.test(x)$conf.int)
```

```
## $`(-Inf,0.75]`
## [1] 424.2567 482.9426
## attr(,"conf.level")
## [1] 0.95
##
## $`(0.75,1.25]`
## [1] 398.8581 450.1305
## attr(,"conf.level")
## [1] 0.95
##
## $`(1.25, Inf]`
## [1] 399.3683 429.2554
## attr(,"conf.level")
## [1] 0.95
```

Hypothesis testing

Natural thing to test:

$$H_0 : \mu_1 = \mu_2 = \mu_3$$

$$H_A : \text{Not all } \mu's \text{ equal}$$

$\Rightarrow$ One-way ANOVA

Hypothesis testing

```
# Fit ANOVA
anova.fit <- aov(cpm ~ pir.f, data = nhanes.adol)
summary(anova.fit)
```

```
##               Df   Sum Sq Mean Sq F value Pr(>F)
## pir.f          2   259455  129727   3.329 0.0362 *
## Residuals   1042 40611061   38974
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 246 observations deleted due to missingness
```

Hypothesis testing

```
# Multiple comparisons
TukeyHSD(anova.fit)


##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = cpm ~ pir.f, data = nhanes.adol)
##
## $pir.f
##                                diff       lwr       upr     p adj
## (0.75,1.25]-(-Inf,0.75] -29.10530 -76.71982 18.509224 0.3234266
## (1.25, Inf]-(-Inf,0.75] -39.28779 -75.02666 -3.548924 0.0270331
## (1.25, Inf]-(0.75,1.25] -10.18249 -50.73418 30.369186 0.8258712
```

Research question: In older American males, Is there an association between race/ethnicity (4 levels) and obesity status (3 levels)?

Generate variables

```
# Get subset of participants age 60+
nhanes.olderm <- subset(nhanes, riagendr == 1 & ridageyr >= 60)
nhanes.olderm <- subset(nhanes, riagendr == 1 & ridageyr >= 70)

# Create 4-level factor version of race/ethnicity
nhanes.olderm$ridreth2[nhanes.olderm$ridreth2 == 5] <- 4
nhanes.olderm$race.f <- factor(nhanes.olderm$ridreth2,
                               levels = 1:4,
                               labels = c("Non-Hisp. White", "Non-Hisp. Black",
                                          "Mex. Amer.", "Other"))

# Create obesity variable
nhanes.olderm$obesity.f <- cut(nhanes.olderm$bmxbmi,
                               breaks = c(-Inf, 25, 30, Inf), right = F,
                               labels = c("Normal", "Overweight", "Obese"))
```

Visualization

```
# Contingency table with frequencies
(table.freq <- table(nhanes.olderm$race.f, nhanes.olderm$obesity.f))
```

```
##
##                    Normal Overweight Obese
##   Non-Hisp. White      94         129    75
##   Non-Hisp. Black      17          21    11
##   Mex. Amer.           23          47     9
##   Other                 8           4     2
```

Visualization

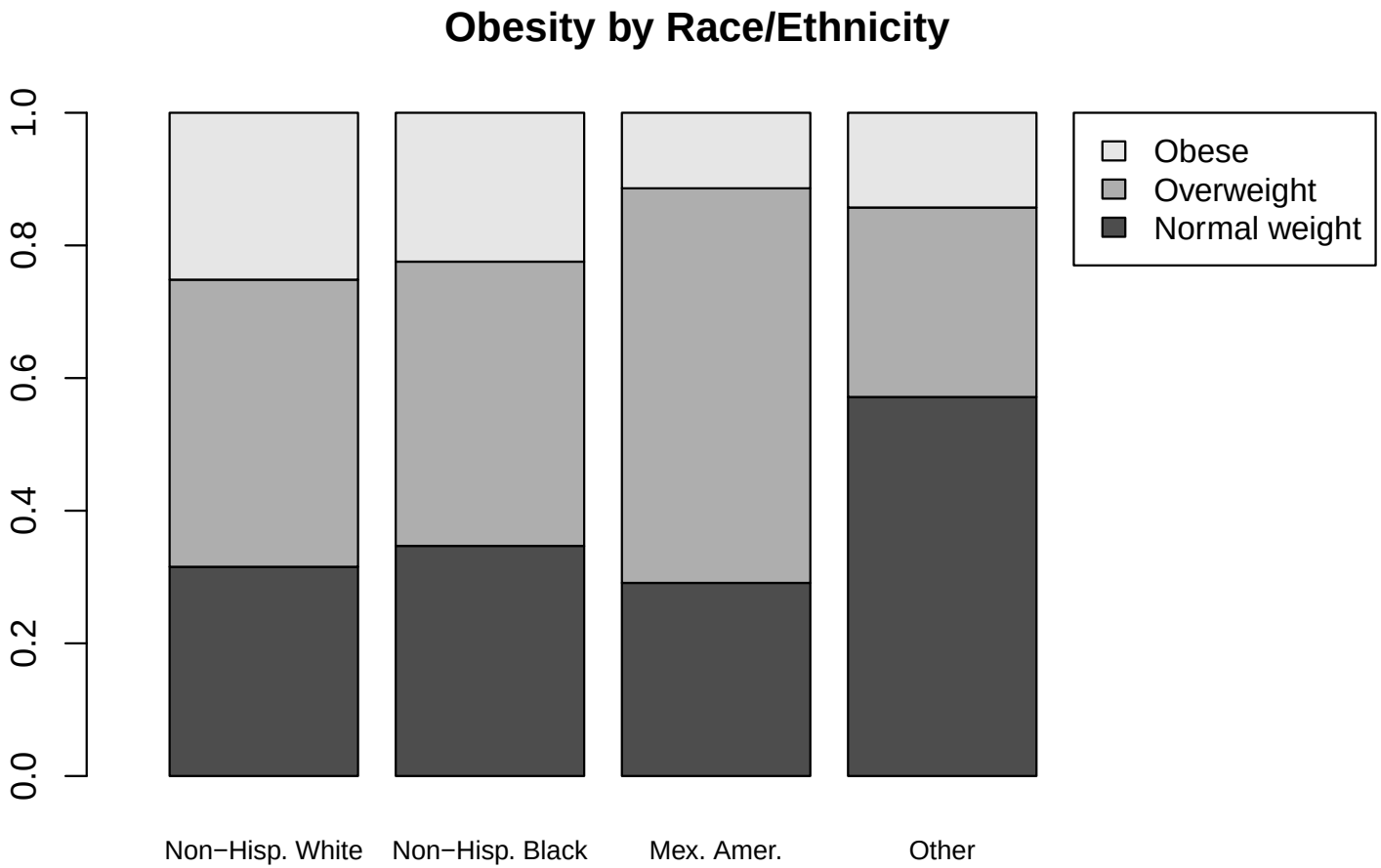
```
# Contingency table with row proportions
(table.rowprops <- prop.table(table.freq, margin = 1))
```

```
##
##                       Normal Overweight     Obese
##   Non-Hisp. White 0.3154362  0.4328859 0.2516779
##   Non-Hisp. Black 0.3469388  0.4285714 0.2244898
##   Mex. Amer.      0.2911392  0.5949367 0.1139241
##   Other           0.5714286  0.2857143 0.1428571
```

Visualization

```r
# Create bar plot
par(xpd = TRUE)
barplot(t(table.rowprops),
        main = "Obesity by Race/Ethnicity", xlim = c(0, 6),
        cex.names = 0.75,
        args.legend = list(bg = "white", x = 6.6, y = 1, cex = 0.9),
        legend.text = c("Normal weight", "Overweight", "Obese"))
```

Visualization



Hypothesis testing

Typical test for two categorical variables:

H_0 : Race and obesity are not associated.

H_A : Race and obesity are associated.

$\Rightarrow$ Chi-square test of association

Hypothesis testing

```
# Chi-square test of association
chisq.test(nhanes.olderm$race.f, nhanes.olderm$obesity.f)

## Warning in chisq.test(nhanes.olderm$race.f, nhanes.olderm$obesity.f):
Chi-squared approximation may be incorrect

##
##  Pearson's Chi-squared test
##
## data:  nhanes.olderm$race.f and nhanes.olderm$obesity.f
## X-squared = 13.492, df = 6, p-value = 0.03586
```

Hypothesis testing

```
# Chi-square test of association
chisq.test(nhanes.olderm$race.f, nhanes.olderm$obesity.f)

## Warning in chisq.test(nhanes.olderm$race.f, nhanes.olderm$obesity.f):
Chi-squared approximation may be incorrect

##
##  Pearson's Chi-squared test
##
## data:  nhanes.olderm$race.f and nhanes.olderm$obesity.f
## X-squared = 13.492, df = 6, p-value = 0.03586
```

$\Rightarrow$ P-value suggests race is significantly assoc. with obesity.

$\Rightarrow$ But Chi-square test may not be valid due to small sample.

Hypothesis testing

```r
# Look at expected cell counts for each cell
(expected.counts <- matrix(rowSums(table.freq), ncol = 1) %*%
   colSums(table.freq) / sum(table.freq))
```

```
##           [,1]       [,2]      [,3]
## [1,] 96.172727 136.131818 65.695455
## [2,] 15.813636  22.384091 10.802273
## [3,] 25.495455  36.088636 17.415909
## [4,]  4.518182   6.395455  3.086364
```

In-class activity

(1) Figure out what small-sample test we could use here.

(2) Figure out what R function does it.

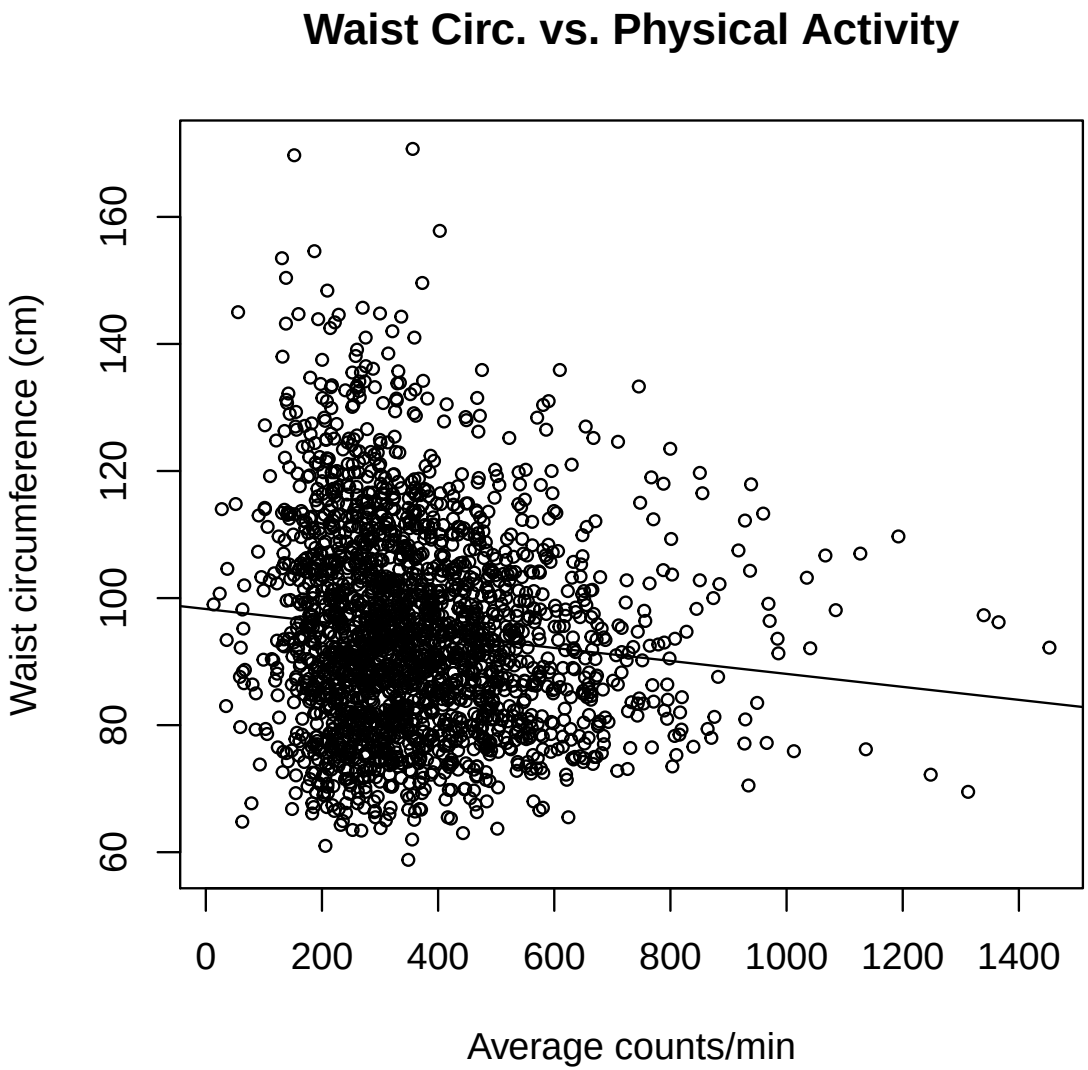
(3) Perform test and see if conclusion is same as Chi-square.

Research question: In American adults age 18-50, is there a correlation between physical activity and waist circumference? If so, how strong is it?

Visualization

```r
# Scatterplot of waist circumference vs. physical activity
plot(nhanes.adults$cpm, nhanes.adults$bmxwaist, cex = 0.7,
     main = "Waist Circ. vs. Physical Activity",
     ylab = "Waist circumference (cm)", xlab = "Average counts/min")
linear.fit <- lm(bmxwaist ~ cpm, data = nhanes.adults)
abline(linear.fit)
```

Visualization



Correlation analysis

- Statistical setup:
  - Let X_i = physical activity and Y_i = waist circumference for i^{th} participant.
  - Observe (X_i, Y_i), $i = 1, ..., n$.
- Parameter of interest:

$$\rho_{xy} = \text{Cor}(X, Y) = \frac{\text{Cov}(X,Y)}{\sqrt{V(X)V(Y)}} = \frac{E(XY)-E(X)E(Y)}{\sqrt{V(X)V(Y)}}$$

- Hypothesis test:

$$H_0 : \rho = 0$$

$$H_A : \rho \neq 0$$

Correlation analysis

```r
# Calculate Pearson correlation coefficient
cor.test(nhanes.adults$cpm, nhanes.adults$bmxwaist)
```

```
##
##  Pearson's product-moment correlation
##
## data:  nhanes.adults$cpm and nhanes.adults$bmxwaist
## t = -5.1771, df = 2253, p-value = 2.454e-07
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.14903739 -0.06745061
## sample estimates:
##        cor
## -0.1084266
```

Correlation analysis

```
# Calculate Spearman correlation coefficient
cor.test(nhanes.adults$cpm, nhanes.adults$bmxwaist, method = "spearman")

## Warning in cor.test.default(nhanes.adults$cpm,
nhanes.adults$bmxwaist, method = "spearman"):  Cannot compute exact
p-value with ties

##
##  Spearman's rank correlation rho
##
## data:  nhanes.adults$cpm and nhanes.adults$bmxwaist
## S = 2111400000, p-value = 6.092e-07
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##        rho
## -0.1048143
```

Regression analysis

- Statistical setup:
  - Assume $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i,\ \epsilon_i \overset{iid}{\sim} (0,\ \sigma^2)$
- Parameters:

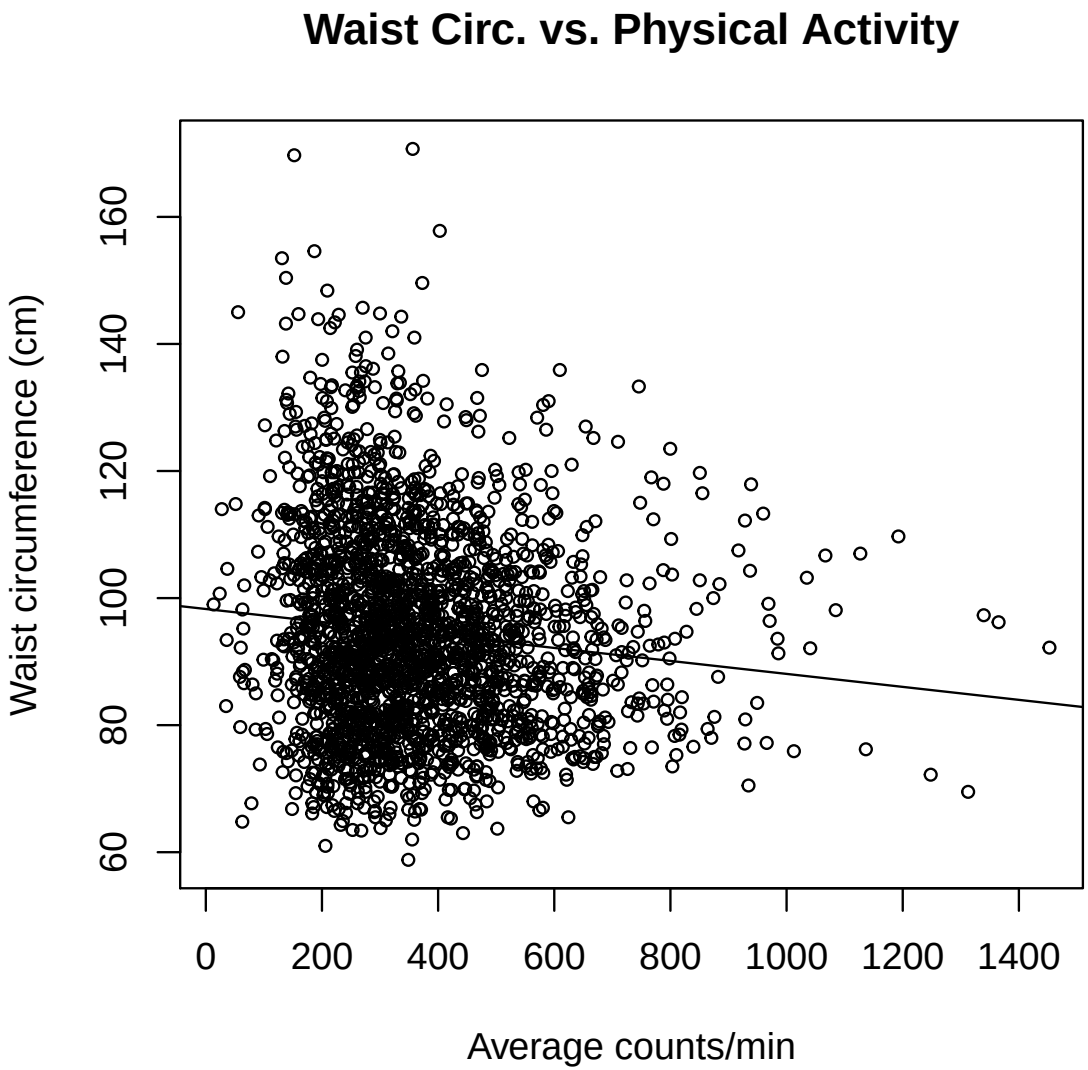
  $\beta_0 =$ Intercept (typically not of interest)

  $\beta_1 =$ Slope
- Hypothesis test of primary interest:

  $H_0 : \beta_1 = 0$

  $H_A : \beta_1 \neq 0$

Intercept and slope



Regression analysis

```r
# Fit simple linear regression of waist circumference vs. physical activity
linear.fit <- lm(bmxwaist ~ cpm, data = nhanes.adults)
summary(linear.fit)
```

Regression analysis

```
##
## Call:
## lm(formula = bmxwaist ~ cpm, data = nhanes.adults)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -35.920 -11.736  -1.670   9.746  76.057
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 98.280842   0.793712 123.824  < 2e-16 ***
## cpm         -0.010214   0.001973  -5.177 2.45e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.88 on 2253 degrees of freedom
##   (458 observations deleted due to missingness)
## Multiple R-squared:  0.01176, Adjusted R-squared:  0.01132
## F-statistic:  26.8 on 1 and 2253 DF,  p-value: 2.454e-07
```

Regression analysis

```r
# Divide CPM by 100 to make slope easier to interpret
nhanes.adults$cpm_100 <- nhanes.adults$cpm / 100
linear.fit <- lm(bmxwaist ~ cpm_100, data = nhanes.adults)
summary(linear.fit)
```

Regression analysis

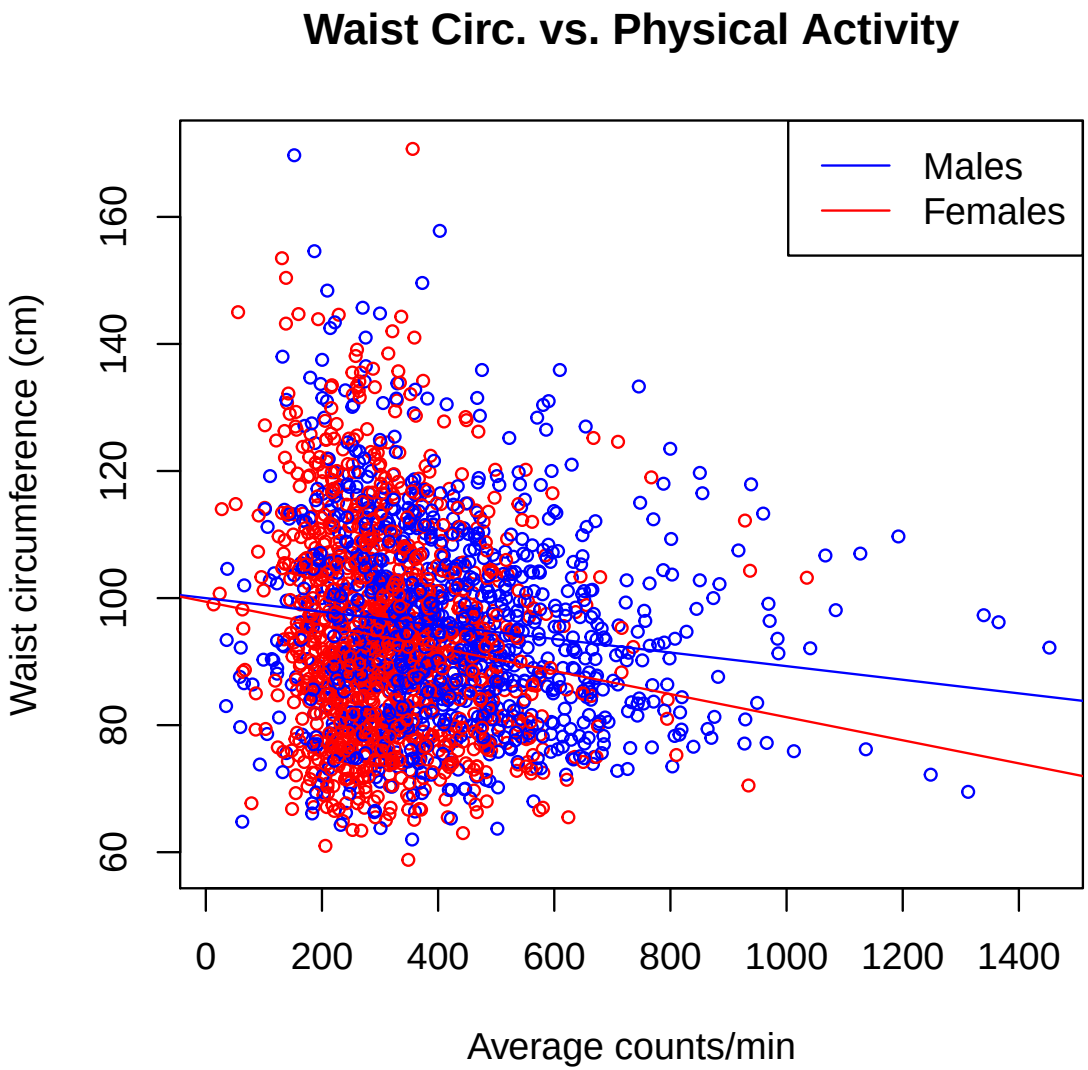
```
##
## Call:
## lm(formula = bmxwaist ~ cpm_100, data = nhanes.adults)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -35.920 -11.736  -1.670   9.746  76.057
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  98.2808     0.7937 123.824  < 2e-16 ***
## cpm_100      -1.0214     0.1973  -5.177 2.45e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.88 on 2253 degrees of freedom
##   (458 observations deleted due to missingness)
## Multiple R-squared:  0.01176, Adjusted R-squared:  0.01132
## F-statistic:  26.8 on 1 and 2253 DF,  p-value: 2.454e-07
```

Research question: Does the relationship between physical activity and waist circumference differ by sex?

Visualization

```
# Scatterplot of waist circumference vs. physical activity by sex
plot(nhanes.adults$cpm, nhanes.adults$bmxwaist, cex = 0.7,
     main = "Waist Circ. vs. Physical Activity",
     ylab = "Waist circumference (cm)", xlab = "Average counts/min",
     col = ifelse(nhanes.adults$riagendr == 1, "blue", "red"))
legend("topright", col = c("blue", "red"), legend = c("Males", "Females"), lty
fit.m <- lm(bmxwaist ~ cpm, data = subset(nhanes.adults, riagendr == 1))
fit.f <- lm(bmxwaist ~ cpm, data = subset(nhanes.adults, riagendr == 2))
abline(fit.m, col = "blue")
abline(fit.f, col = "red")
```

Visualization



Regression analysis

- Statistical setup:
  - Let X_i = physical activity, Y_i = waist circumference, and $M_i = 1$ if male, 0 if female.
  - $Y_i = \beta_0 + \beta_1 X_i + \beta_2 M_i + \beta_3 X_i M_i + \epsilon_i,\ \epsilon_i \overset{iid}{\sim} (0,\ \sigma^2)$
- Parameters:
  - Interaction term, β_3, is of primary interest.
  - From model, what is the slope for males? Females?
  - What does it mean if $\beta_3 = 0$?

Regression analysis

```r
# Fit model with interaction term
nhanes.adults$male <- ifelse(nhanes.adults$riagendr == 1, 1, 0)
linear.fit <- lm(bmxwaist ~ cpm + male + cpm * male, data = nhanes.adults)
summary(linear.fit)
```

Regression analysis

```
##
## Call:
## lm(formula = bmxwaist ~ cpm + male + cpm * male, data = nhanes.adults)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -34.683 -11.632  -1.708   9.483  77.748
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 99.425633   1.235776  80.456  < 2e-16 ***
## cpm         -0.018174   0.003643  -4.988 6.55e-07 ***
## male         0.568081   1.693916   0.335   0.7374
## cpm:male     0.007465   0.004422   1.688   0.0915 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.81 on 2251 degrees of freedom
##   (458 observations deleted due to missingness)
## Multiple R-squared:  0.02188, Adjusted R-squared:  0.02058
## F-statistic: 16.79 on 3 and 2251 DF,  p-value: 8.726e-11
```

Regression analysis

Question: What if we drop the interaction term? It was not significant after all.

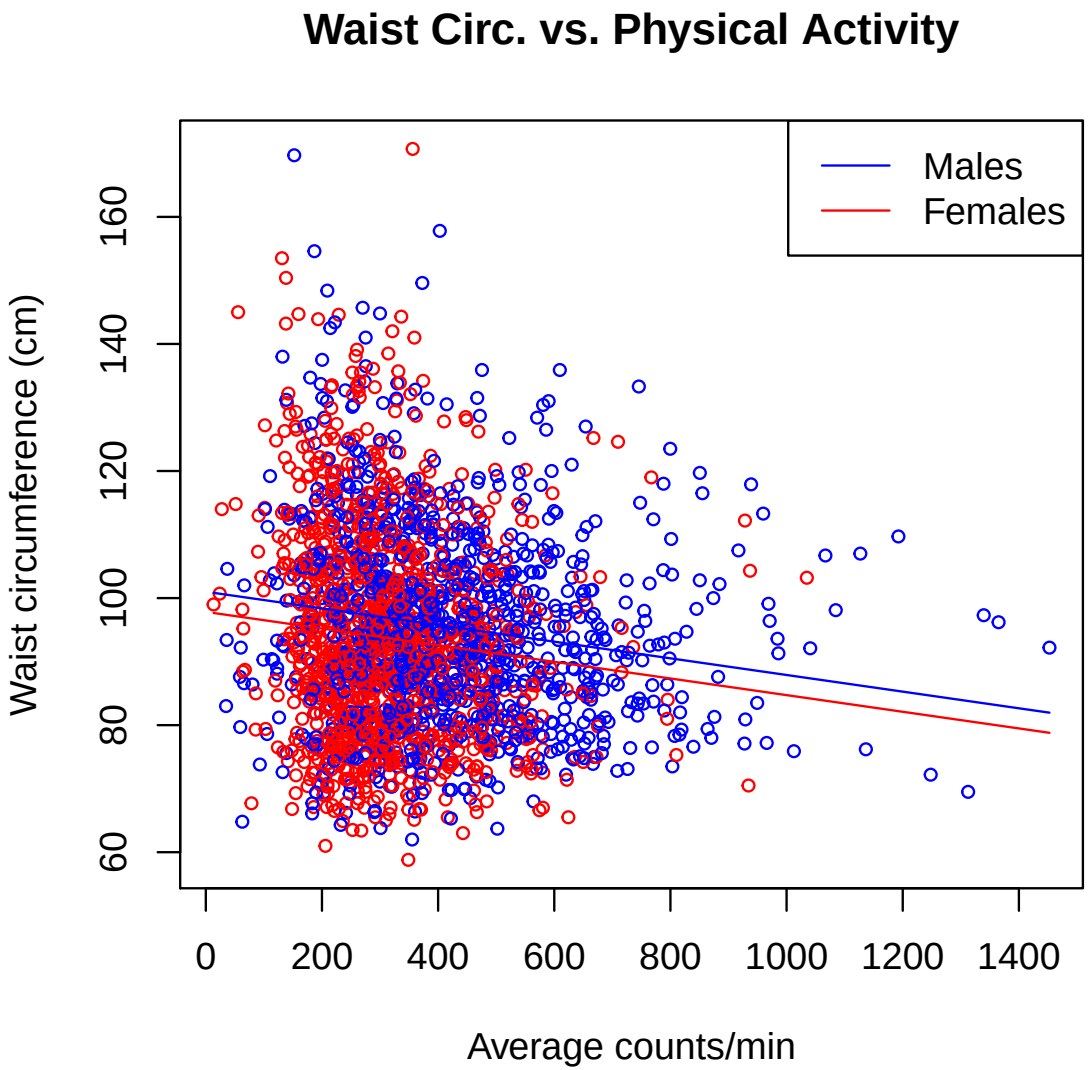
Regression analysis

- Previously, with interaction term:
  - $Y_i = \beta_0 + \beta_1 X_i + \beta_2 M_i + \beta_3 X_i M_i + \epsilon_i,\ \epsilon_i \overset{iid}{\sim} (0,\ \sigma^2)$
- Now:
  - $Y_i = \beta_0^* + \beta_1^* X_i + \beta_2^* M_i + \epsilon_i,\ \epsilon_i \overset{iid}{\sim} (0,\ \sigma^{*2})$
- Parameters:
  - Now, what is the slope for males? Females?
  - Are the regression lines the same?

Visualization

```
# Scatterplot with regression line that does not include interaction term
plot(nhanes.adults$cpm, nhanes.adults$bmxwaist, cex = 0.7,
     main = "Waist Circ. vs. Physical Activity",
     ylab = "Waist circumference (cm)", xlab = "Average counts/min",
     col = ifelse(nhanes.adults$riagendr == 1, "blue", "red"))
fit <- lm(bmxwaist ~ cpm + male, data = nhanes.adults)
xvals <- range(nhanes.adults$cpm, na.rm = T)
yvals.m <- fit$coef[1] + fit$coef[2] * xvals + fit$coef[3]
yvals.f <- fit$coef[1] + fit$coef[2] * xvals
points(xvals, yvals.m, type = "l", col = "blue")
points(xvals, yvals.f, type = "l", col = "red")
legend("topright", col = c("blue", "red"), legend = c("Males", "Females"), lty
```

Visualization



Regression analysis

```r
# Fit model with interaction term
linear.fit <- lm(bmxwaist ~ cpm + male, data = nhanes.adults)
summary(linear.fit)
```

Regression analysis

```
##
## Call:
## lm(formula = bmxwaist ~ cpm + male, data = nhanes.adults)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -35.378 -11.675  -1.711   9.478  77.538
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 97.830292   0.796567 122.815  < 2e-16 ***
## cpm         -0.013107   0.002066  -6.344 2.70e-10 ***
## male         3.171136   0.701374   4.521 6.46e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.81 on 2252 degrees of freedom
##   (458 observations deleted due to missingness)
## Multiple R-squared:  0.02065, Adjusted R-squared:  0.01978
## F-statistic: 23.74 on 2 and 2252 DF,  p-value: 6.28e-11
```

Final thoughts

- The internet exists. No need to memorize!
- Function help files are...helpful.
- In most cases, a graph is (at least) as good as a test.
- Make graph $\Rightarrow$ eyeball association $\Rightarrow$ confirm with p-value.

Lab

You may or may not be familiar with logistic regression. Basically, logistic regression is what you use to test whether one or more variables are associated with a **binary** outcome variable.

It might look a little funny, but for logistic regression with a binary outcome variable Y and two predictors X_1 and X_2, we assume the following model:

$$\log[\frac{P(Y_i=1)}{1-P(Y_i=1)}] = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i}$$

In other words, we assume that there is a linear relationship between each predictor and the **log-odds** of Y.

If $\beta_1 = 0$, then X_1 is not associated with Y. If β_1 is positive, then people with higher values for X_1 are **more likely** to experience the outcome than people with lower values of X_1; and vice versa if β_1 is negative.

Anyway, in this lab, you will learn how to fit a logistic regression model in R. You will have to use the **glm** function. It works like **lm**, but can handle various regression models, not just linear regression.

Please fit a logistic regression model to see whether sex and waist circumference are associated with odds of meeting the US physical activity guidelines in American adults. Here is an outline:

1. Create a variable called **met_guideline** that is 1 if **guideline_min** > 21.4, and 0 otherwise.
2. Create a variable called **male** that is 1 if **riagendr** $= 1$ and 0 otherwise.
3. Look at the help file for **glm** and see what input you have to specify to get R to do logistic regression.
4. Fit and interpret the logistic regression model.
5. Re-fit the model with a sex-by-waist circumference interaction term. (If it is significant, then the relationship between waist circ. and odds of meeting the physical activity guidelines differs in males vs. females)