Quiz2

Matthew Stoebe

2025-04-20

Question 1

a

```r
fiz <- read.table("Data/fizzergy.txt",
                  header = FALSE,
                  col.names = c("consume","age","chol"))

mod0 <- lm(chol ~ consume, data=fiz)
summary(mod0)
```

```
##
## Call:
## lm(formula = chol ~ consume, data = fiz)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -48.037 -11.688   1.312  11.312  51.312
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  205.688      1.956 105.161  < 2e-16 ***
## consume      -12.651      4.123  -3.068  0.00267 **
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 18.86 on 118 degrees of freedom
## Multiple R-squared:  0.07388,    Adjusted R-squared:  0.06603
## F-statistic: 9.413 on 1 and 118 DF,  p-value: 0.002673
```

There is a signifficant relationship between consumption and cholesteral levels before controlling for other covariates.

b

```r
mod1 <- lm(chol ~ consume + age, data=fiz)
summary(mod1)
```

```
## 
## Call:
## lm(formula = chol ~ consume + age, data = fiz)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -31.797  -7.318  -0.111   8.459  33.161 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 143.1364     5.5665  25.714   <2e-16 ***
## consume       1.3794     3.0758   0.448    0.655    
## age           1.3450     0.1162  11.578   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
## 
## Residual standard error: 12.93 on 117 degrees of freedom
## Multiple R-squared:  0.5684, Adjusted R-squared:  0.561 
## F-statistic: 77.04 on 2 and 117 DF,  p-value: < 2.2e-16
```

After we control for age, we see that consumption no longer has a significant relationship with cholesteral levels

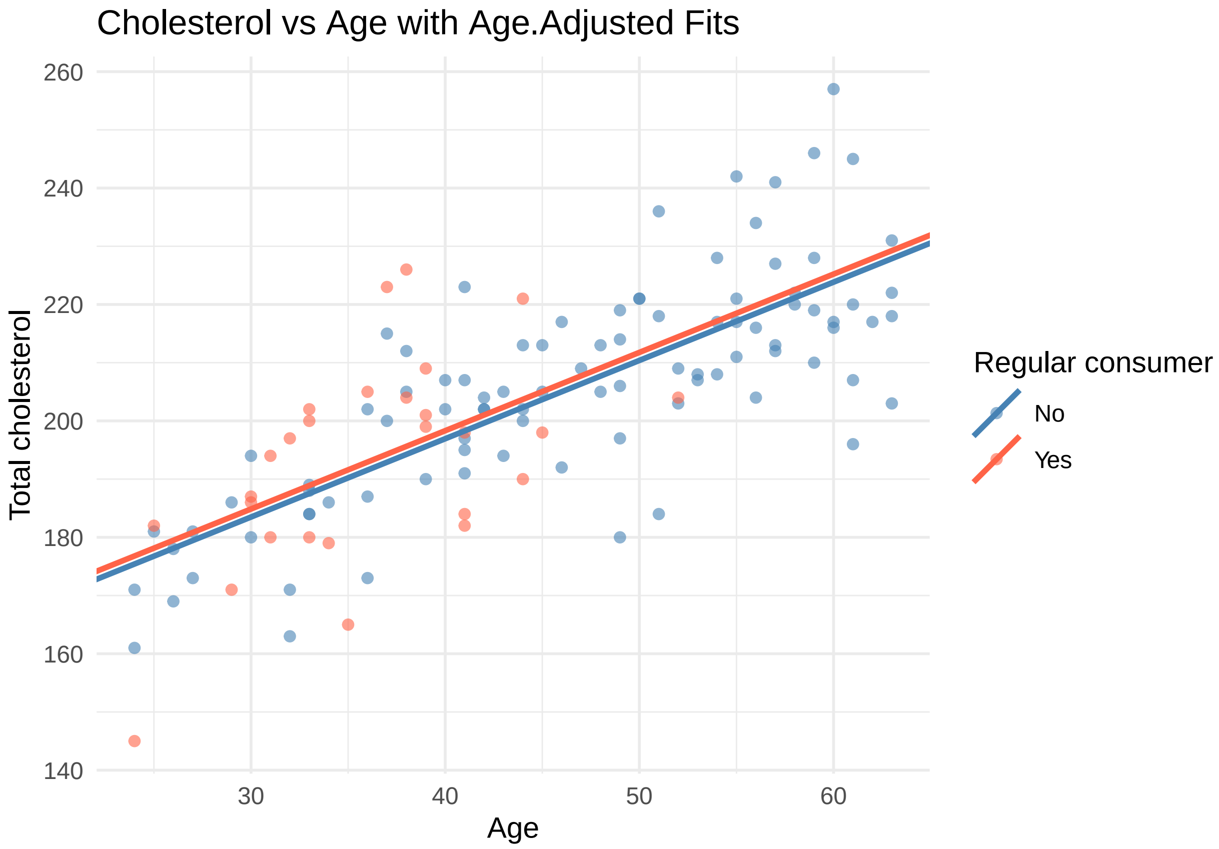
C

```r
library(ggplot2)

mod1  <- lm(chol ~ consume + age, data = fiz)
coefs <- coef(mod1)
slope <- coefs["age"]
int0  <- coefs["(Intercept)"]
int1  <- int0 + coefs["consume"]

# Create a little data-frame for the two lines
line_df <- data.frame(
  consume   = factor(c(0, 1)),
  intercept = c(int0, int1),
  slope     = slope
)
```

```
## Warning in data.frame(consume = factor(c(0, 1)), intercept = c(int0, int1), :
## row names were found from a short variable and have been discarded
```

```r
ggplot(fiz, aes(x = age, y = chol, color = factor(consume))) +
  geom_point(alpha = 0.6) +
  geom_abline(aes(intercept = intercept, slope = slope, color = consume),
              data = line_df,
              size = 1) +
  scale_color_manual(
    name   = "Regular consumer",
    values = c("0" = "steelblue", "1" = "tomato"),
    labels = c("No", "Yes")
  ) +
  labs(
    x     = "Age",
    y     = "Total cholesterol",
    title = "Cholesterol vs Age with Age-Adjusted Fits"
  ) +
  theme_minimal()
```

```
## Warning: Using ‘size‘ aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use ‘linewidth‘ instead.
## This warning is displayed once every 8 hours.
## Call ‘lifecycle::last_lifecycle_warnings()‘ to see where this warning was
## generated.
```



d

Age is related to both energy drink consumption habits and cholesterol. As such, omitting it biases the comparison to show non-existant relationships between teh two variables.

Question 2

a

```r
df <- read.table("Data/tea.txt", header=TRUE)

df$teaF <- factor(df$tea, levels=0:2, labels=c("rarely","sometimes","frequently"))
anova_res <- aov(noreph ~ teaF, data=df)
summary(anova_res)
```

```
##              Df  Sum Sq Mean Sq F value Pr(>F)
## teaF          2  311958  155979   3.644 0.0285 *
## Residuals   147 6291731   42801
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

There is a signifficant relationship between drinking tea and norepinephrine levels which I assume are related to migrane levels..

b

We cannot make a causal conclusion such are recommending tea drinking based on this data because it is an observational study AND does not apply any more advanced causal techniques to attempt to measure causality. We can merely say that these two things appear to be related. We cannot drive recomendations.

Question 3

##a

```r
schools <- read.table("Data/schoolscore.txt", header=TRUE)

res <- t.test(schools$sc24, schools$sc23,
              paired      = TRUE,
              alternative = "greater")
print(res)
```

```
##
##  Paired t-test
##
## data:  schools$sc24 and schools$sc23
## t = 5.5594, df = 24, p-value = 5.081e-06
## alternative hypothesis: true mean difference is greater than 0
## 95 percent confidence interval:
##  3.35327     Inf
## sample estimates:
## mean difference
##           4.844
```

The pass rates ARE significantly higher in 2024 than in 2023 with a p value of 5.081e-06.

b

No. this is not an RCT and does not apply more advanced causal techniques. Additionall by sampling the lowest performing schools and remeasuring without controll group, regression to the mean is likely to have occured and may explain the observed gain. Without randomized assigment or comparable controll, we cannot attribute the improvement to smaller class sizes.

Question 4

```
hosp <- read.table("Data/hosps.txt", header = TRUE)

hosp$publicF <- factor(hosp$public,
                       levels = c(0, 1),
                       labels = c("private", "public"))

aggregate(percinf ~ publicF, data = hosp,
          FUN = function(x) c(n=length(x),
                              mean=round(mean(x),1),
                              sd=round(sd(x),1)))
```

```
##   publicF percinf.n percinf.mean percinf.sd
## 1 private     211.0          3.0        2.8
## 2  public     209.0          3.7        2.9
```

```
tt <- t.test(percinf ~ public,
             data        = hosp,
             var.equal   = FALSE)
tt
```

```
## 
##  Welch Two Sample t-test
## 
## data:  percinf by public
## t = -2.6415, df = 416.96, p-value = 0.008564
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -1.2991408 -0.1905755
## sample estimates:
## mean in group 0 mean in group 1 
##        2.987678        3.732536
```

```
lm_mod <- lm(percinf ~ public + pmedicaid + npat, data = hosp)
summary(lm_mod)
```

```
## 
## Call:
## lm(formula = percinf ~ public + pmedicaid + npat, data = hosp)
## 
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max 
## -4.8772 -1.8935 -0.7217  1.0979 12.2125 
## 
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  2.468e+00  2.956e-01   8.348 1.04e-15 ***
## public      -1.924e-01  3.147e-01  -0.611    0.541    
## pmedicaid   -1.394e-02  2.332e-02  -0.598    0.550    
## npat         1.932e-06  3.250e-07   5.946 5.83e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 2.78 on 416 degrees of freedom
## Multiple R-squared:  0.09348,    Adjusted R-squared:  0.08694 
## F-statistic:  14.3 on 3 and 416 DF,  p-value: 6.948e-09
```

A non - adjusted test comparing public to private hospitals does show significantly higher infection rates at public hospitals (p <.05). However, accounting for hospital size and medicad mix removes this significance. After controlling for these factors, we see that the number of patients has the significant relationship with the infection rate, but whether the hospital is public or private does not. all signifficance drawn at .05 level.