Chapter 11 - Linear Regression & Modelling

Exercises

Brian Fogarty

15 May 2018

Contents

EXERCISE I

Using the 2012 Smoking and Drug Use Amongst English Pupils Dataset (`2012smokedrugs.dta`), analyse cigarette consumption using a multiple linear regression model.

1. Run a regression model with the outcome variable `cigs7` and the predictor variables `free`, `schyear`, and `sex`. Evaluate the overall model and identify any statistically significant relationships. Additionally, provide an interpretation of any statistically significant coefficients and discuss any significant relationships using plain language.
2. Using `ggplot()` create a histogram for `cigs7` and each statistically significant predictor. What do(es) the histogram(s) show you?
3. Run another multiple regression with the same predictors as the first regression model where the outcome variable `cigs7` only has positive values; that is, the 0s have been removed. Evaluate the overall model and identify any statistically significant relationships. How are these results different from the first model that you ran?
4. Using `ggplot()` create a histogram for the new `cigs7` variable and each statistically significant predictor. What do(es) the histogram(s) show you?

ANSWERS FOR EXERCISE I

Question 1.1

Read-in 2012 Smoking and Drug Use Amongst English Pupils.

```
setwd("C:/QSSD/Exercises/Chapter 11 - Exercises")
getwd()
```

```
[1] "C:/QSSD/Exercises/Chapter 11 - Exercises"
```

```
library(foreign)
drugs <- read.dta("2012smokedrugs.dta")

names(drugs)
```

```
 [1] "aage"           "aalcobottles7"  "aalcocans7"     "aalcunits7"
 [5] "abeerbottles7"  "abeerunits7"    "abottlesother7" "age1115"
 [9] "age1215"        "age1315"        "aglassliq7"     "aglassother7"
[13] "aglasswine7"    "ahalfother7"    "ahalfpints7"    "alargebeer7"
[17] "alargeother7"   "apeodrink"      "apints7"        "apintsother7"
[21] "apopunits7"     "asmallcans7"    "asmallother7"   "books"
[25] "cigs7"          "damph"          "dcoke"          "dcrack"
[29] "decstasy"       "dglue"          "dheroin"        "dketamine"
[33] "dlsd"           "dmagic"         "dmephed"        "dmethadone"
[37] "dother"         "dpop"           "drugocc"        "dtranqs"
[41] "dusefreq"       "dweed"          "famdrink"       "famsmoke"
[45] "free"           "lifediff"       "lifegood"       "liferight"
[49] "lifewant"       "lifewell"       "region"         "schyear"
[53] "sex"            "truant"         "truant12"
```

Model 1 with the number of cigarettes smoked in last week (`cigs7`, which goes from 0 to 140 cigarettes a week), whether the pupil receives free lunch (`free`, where 1 = "free lunch"), gender (`gender`, 1 = "female"), and the pupil's year in school (`schyear`, year 1 through 5).

```
model.1 <- lm(cigs7 ~ free + schyear + sex, data=drugs)
summary(model.1)
```

```

Call:
lm(formula = cigs7 ~ free + schyear + sex, data = drugs)

Residuals:
    Min      1Q  Median      3Q     Max
 -4.808  -2.347  -1.304  -0.096 138.696

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.82458    0.27663  -6.596 4.53e-11 ***
free         1.41863    0.29006   4.891 1.03e-06 ***
schyear      1.04281    0.07617  13.691  < 2e-16 ***
sex         -0.16525    0.21199  -0.780    0.436
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.961 on 7148 degrees of freedom
  (437 observations deleted due to missingness)
Multiple R-squared:  0.02821,   Adjusted R-squared:  0.02781
F-statistic: 69.18 on 3 and 7148 DF,  p-value: < 2.2e-16
```

```
confint(model.1, level=0.95)
```

```
                2.5 %     97.5 %
(Intercept) -2.3668488 -1.2823150
free         0.8500292  1.9872319
schyear      0.8935009  1.1921281
sex         -0.5808119  0.2503103
```

According to the F-statistic, we see that the overall model is statistically significant since $p \leq .05$. The R^2 (.028) and the adjusted R^2 (.028) are very close indicating that all of our predictors are contributing to explaining variance in the outcome variable. We can interpret the R^2 value as *our model explains 2.8% of the variance in the number of cigarettes smoked last week by pupils.*

We see that whether pupils receive free lunch and the pupil's year in school have a positive, statistically significant effect on the expected number of cigarettes smoked per week. We can interpret the `free`'s coefficient as *pupils who receive free lunch are expected to smoke 1.42 more cigarettes per week than students who do not receive free lunch, controlling for other predictors.* A plain language discussion would be along the lines of *pupils who receive free lunch smoke more than pupils who do not receive free lunch, but the difference is small at in-between 1 to 2 more cigarettes a week. Free lunch is typically provided based on the economic status of pupils' families and poorer individuals tend to smoke more than wealthier individuals. As a proxy for family income, the free lunch effect may be demonstrating the link between wealth and smoking rates.*

For `schyear`, our interpretation is *for a one-year increase in a pupil's year in school, they are expected to smoke 1.04 more cigarettes a week.* A plain language discussion would possibly include *as we might expect, older pupils smoke more cigarettes than younger pupils. But the difference is not large as the increase is only expected to be 1 cigarette more for each year in school.*

Question 1.2

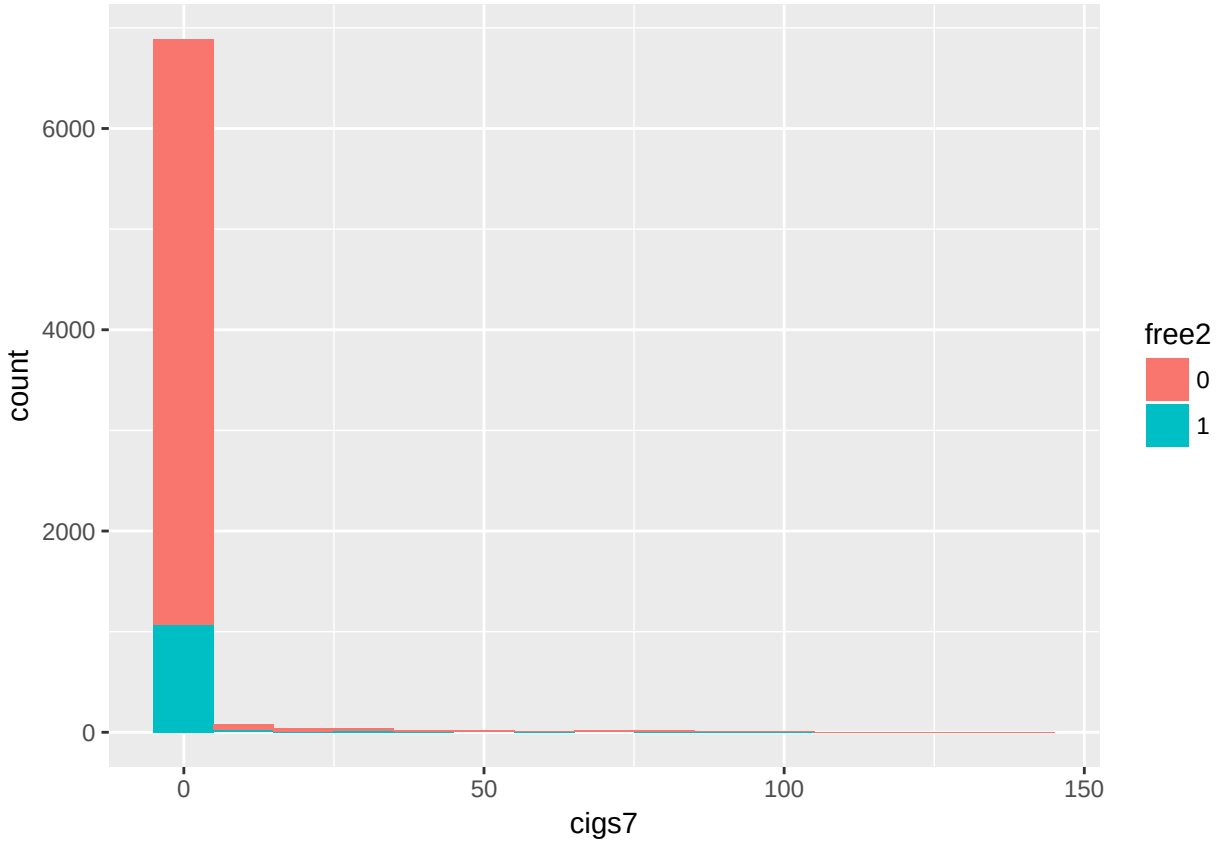
```
library(ggplot2)
class(drugs$free)
```

```
[1] "numeric"
```

```
drugs$free2 <- as.factor(drugs$free)

drugs2 <- subset(drugs, !is.na(cigs7) & !is.na(free2))

x11()
ggplot(drugs2, mapping=aes(cigs7)) +
  geom_histogram(mapping=aes(fill=free2), binwidth=10)
```



It is difficult to discern any relevant information about the relationship between `cigs7` and `free` from the histogram, except that most pupils have smoked 0 cigarettes in the past week and there are few pupils that receive free lunch.

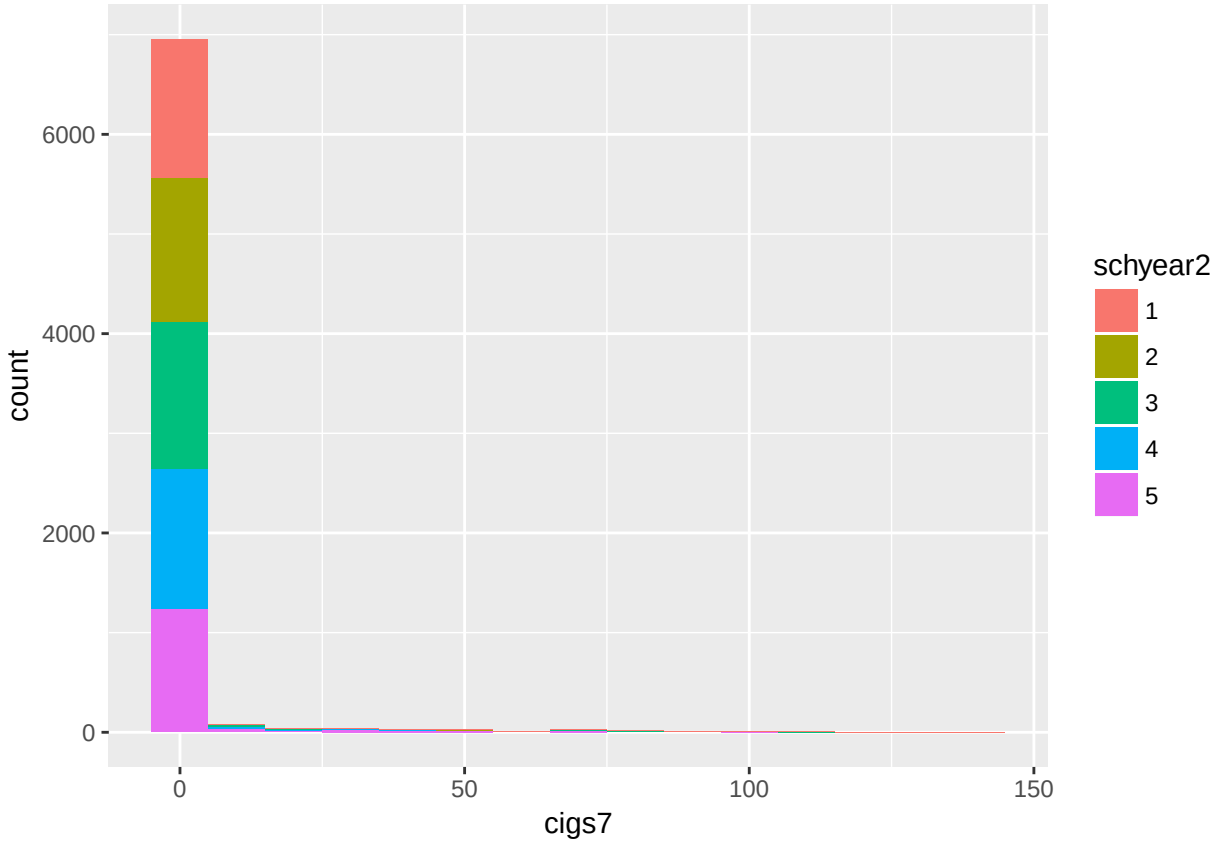
```
class(drugs$schyear)
```

```
[1] "numeric"
```

```
drugs$schyear2 <- as.factor(drugs$schyear)

drugs2 <- subset(drugs, !is.na(cigs7) & !is.na(schyear2))

x11()
ggplot(drugs2, mapping=aes(cigs7)) +
  geom_histogram(mapping=aes(fill=schyear2), binwidth=10)
```



As in the first histogram, it is difficult here to discern any relevant information about the relationship between `cigs7` and `schyear` from the histogram, except that most pupils, across all school years, have smoked 0 cigarettes in the past week and that most pupils who have smoked in the past week tend to be older.

Question 1.3

One issue to consider is that `cigs7` has a substantial number of 0s - over 94%. For this question, we want to recode `cigs7` to exclude the 0s and then re-run the regression. To do so, we will use the `car` package and simply recode all the 0s as "NA" and create a new variable called `cigs7a`.

```
library(car)
```

```
Warning: package 'car' was built under R version 3.4.3
```

```
drugs$cigs7a <- recode(drugs$cigs7, "0=NA")
table(drugs$cigs7a)
```

```
 1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18
54 28 22 28 12 16 16  4  7  8  7  6  5  8  1  5  6  5
19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36
 3  2  3  3  2  2  4  4  2  5  4  2  6  1  1  3  9  3
37 38 39 40 41 42 43 45 46 47 49 50 51 52 53 54 55 56
 3  2  1  4  2  6  2  3  4  1  3  3  2  4  4  1  1  2
60 62 63 66 67 69 70 71 73 74 75 76 77 79 80 83 84 85
 7  2  1  2  2  3 10  1  1  1  3  1  1  1  8  1  1  1
86 89 90 92 95 100 104 105 110 140
```

```
  2   1   4   1   1   2   1   3   3   2
```

Now, let's re-run the regression

```
model.2 <- lm(cigs7a ~ free + schyear + sex, data=drugs)
summary(model.2)
```

```

Call:
lm(formula = cigs7a ~ free + schyear + sex, data = drugs)

Residuals:
   Min     1Q Median     3Q    Max
-33.31 -20.80 -12.47  15.99 122.04

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    6.549      7.452   0.879   0.3800
free           5.940      3.365   1.765   0.0783 .
schyear        3.803      1.625   2.340   0.0197 *
sex            2.811      2.855   0.984   0.3255
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 28.7 on 404 degrees of freedom
  (7181 observations deleted due to missingness)
Multiple R-squared:  0.01964,   Adjusted R-squared:  0.01236
F-statistic: 2.698 on 3 and 404 DF,  p-value: 0.04555
```

```
confint(model.2, level=0.95)
```

```
                2.5 %    97.5 %
(Intercept) -8.1007228 21.199072
free        -0.6754016 12.556060
schyear      0.6087239  6.997143
sex         -2.8026697  8.424100
```

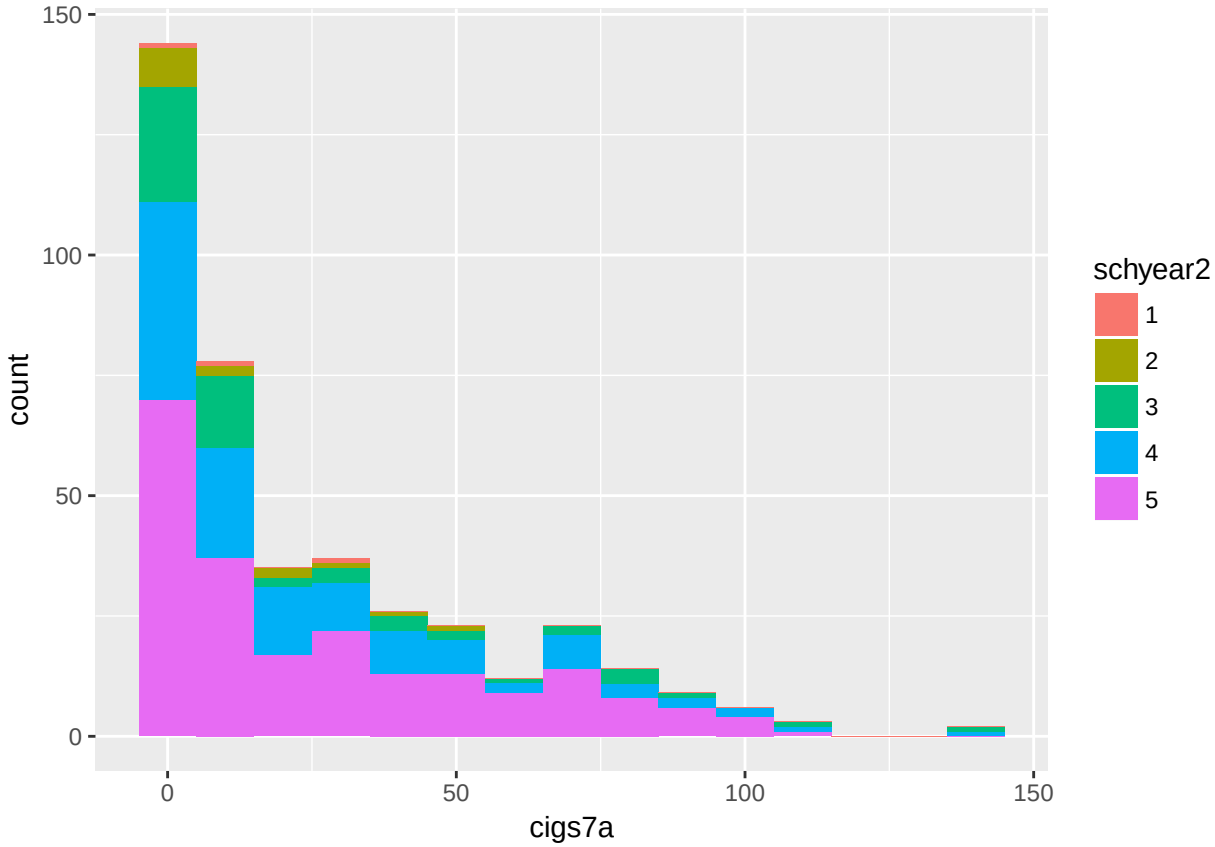
According to the F-statistic, we see that the overall model is statistically significant since $p \leq .05$. The R^2 (.020) and the adjusted R^2 (.012) are not very close indicating that one (or more) of our predictors are not contributing to explaining variance in the outcome variable. We can interpret the R^2 value as *our model explains 2.0% of the variance in the number of cigarettes smoked last week by pupils.*

In this new regression, we see that only a pupil's school year is statistically significant. Also notice that there are a substantial number of missing values - roughly the 94% of pupils who never smoked. Therefore, we are really looking at a subset of our sample. Given that, we might want to qualify our interpretation with a statement that we are only looking at pupils who have smoked in the past week. With that information in mind, we interpret `schyear`'s coefficient as *among pupils who reported smoking in the past week, for a one-year increase in school year, pupils are expected to smoke 3.8 more cigarettes, while controlling for other predictors.* This is a larger effect size than when we included all pupils. Our plain language discussion might be *among pupils who do smoke, older pupils smoke more than younger pupils. In fact, amongst these pupils, those in their 5th year are expected to smoke almost a pack more than those in their 1st year*; pupils in their first year are expected to smoke 10.35 cigarettes a week $(6.55 + 3.8(1))$ and pupils in their fifth year are expected to smoke 25.55 cigarettes a week $(6.55 + 3.8(5))$.

Question 1.4

```
drugs2 <- subset(drugs, !is.na(cigs7a) & !is.na(schyear2))

x11()
ggplot(drugs2, mapping=aes(cigs7a)) +
  geom_histogram(mapping=aes(fill=schyear2), binwidth=10)
```



The histogram shows that even among pupils who smoke most of them do not smoke alot of cigarettes per week. Young pupils smoke less cigarettes per week than older pupils with 5th year pupils making up most of the heavier smokers.

EXERCISE II

Using the 2011 England Health Survey dataset (`2011 England Health.dta`):

1. Run a regression model with the outcome variable `bmival` and the predictor variables `employed`, `cigs`, `alcohol`, and `fruitveg`. Evaluate the overall model and identify any statistically significant relationships. Additionally, provide an interpretation of any statistically significant coefficients and discuss any significant relationships using plain language.
2. Using `ggplot()` create a scatterplot for `bmival` with `fruitveg` on the x-axis and two regression lines representing employed and not employed (from the variable `employed`). What does it show you?

ANSWERS FOR EXERCISE II

Question 2.1

Read-in the 2011 England Health Survey data.

```
health <- read.dta("2011 England Health.dta", convert.factors=FALSE)

names(health)
```

```
 [1] "hserial"  "pserial"  "HHSize"   "tenureb"  "Sex"      "Age"
 [7] "MonthAge" "WeekAge"  "PersNo"   "topqual3" "HRPID"    "econact"
[13] "nssec8"   "Origin"   "totinc"   "eqvinc"   "NurOutc"  "relto01"
[19] "relto02"  "relto03"  "relto04"  "relto05"  "relto06"  "relto07"
[25] "relto08"  "relto09"  "Relto10"  "Relto11"  "Relto12"  "ReltoHRP"
[31] "marstatc" "SHA"      "gor1"     "wt_int"   "wt_nurse" "SayWgt"
[37] "SayDiet"  "htval"    "wtval"    "bmival"   "whval"    "omdiaval"
[43] "omsysval" "dnnow"    "totalwu"  "porfv"    "acutill"  "IllsM1"
[49] "IllsM2"   "IllsM3"   "IllsM4"   "IllsM5"   "IllsM6"   "limitill"
[55] "medcnj"   "genhelf2" "cigst1"   "cigst2"   "cigs"     "health"
[61] "ill"      "marital"  "gender"   "employed" "alcohol"  "fruitveg"
[67] "age"
```

Model 3 with respondents' BMI values (`bmival`, which goes from 8.34 to 65.28), whether the respondent is employed (`employed`, where 1 = "employed"), cigarette consumption (`cigs`, a 4-value ordinal variable with higher numbers indicating more smoking), alcohol consumption by unit per week (`alcohol`, which goes from 0 to 461.5), and the number of fruits and vegetable consumed per day (`fruitveg`, which goes from 0 to 30).

```
summary(model.3 <- lm(bmival ~ employed + cigs + alcohol + fruitveg, data=health))
```

```
Call:
lm(formula = bmival ~ employed + cigs + alcohol + fruitveg, data = health)

Residuals:
    Min      1Q  Median      3Q     Max
-13.611  -3.635  -0.771   2.699  37.286

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 28.653393   0.185322 154.614  < 2e-16 ***
employed    -0.492066   0.129601  -3.797 0.000148 ***
cigs        -0.490416   0.080824  -6.068 1.37e-09 ***
```

```
alcohol     -0.001092   0.003174  -0.344 0.730945
fruitveg    -0.057178   0.025014  -2.286 0.022292 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.32 on 6865 degrees of freedom
  (3747 observations deleted due to missingness)
Multiple R-squared:  0.007778,  Adjusted R-squared:  0.007199
F-statistic: 13.45 on 4 and 6865 DF,  p-value: 6.353e-11
```

```
confint(model.3, level=0.95)
```

```
                  2.5 %       97.5 %
(Intercept) 28.290105158 29.016681753
employed    -0.746123339 -0.238008304
cigs        -0.648856897 -0.331975336
alcohol     -0.007314339  0.005131119
fruitveg    -0.106213102 -0.008143738
```

According to the F-statistic, we see that the overall model is statistically significant since $p \leq .05$. The R^2 (.008) and the adjusted R^2 (.007) are very close indicating that all of our predictors are contributing to explaining variance in the outcome variable. Even though R^2 values are relative, this R^2 is very small, suggesting we have a poor model in terms of explanatory power. We can interpret the R^2 value as *our model explains .8% of the variance in the number of cigarettes smoked last week by pupils.*

We see that being employed, smoking, and fruit and vegetable consumption all have a negative, statistically significant effect on respondents' expected BMI values. We can interpret `employed` as *employed respondents are expected to have a BMI value .49 points lower than unemployed respondents, while controlling for other predictors.* A plain language discussion may include *people who are employed are expected to have slightly lower BMI values than unemployed people. People who are employed are likely more physically active than people who are not employed, which may result in a lower BMI.* For `cigs`, *for a one-unit increase in cigarette consumption, respondents' BMI values are expected to decrease by .49 points.* Notice that `cigs` is a four-category ordinal variable and thus we need to use the generic *for a one-unit increase* in our interpretation. A plain language discussion might include *as people smoke more, they are expected to have lower BMI values. This may be due to the effect of nicotine on people's appetite and metabolism. Although smokers are expected to have lower BMI values, this says nothing about life expectancy since, although BMI is related to longevity, smoking has a negative impact on life expectancy.* For `fruitveg`, *for each additional fruit or vegetable consumed per day, respondents' BMI values are expected to decrease by .06 points.* A plain language discussion might include *as people eat more fruits and vegetables per day, their BMI values are expected to slightly decrease. If people are eating more fruits and vegetables, they are likely eating less junk food or high calorie foods.*

Question 2.2

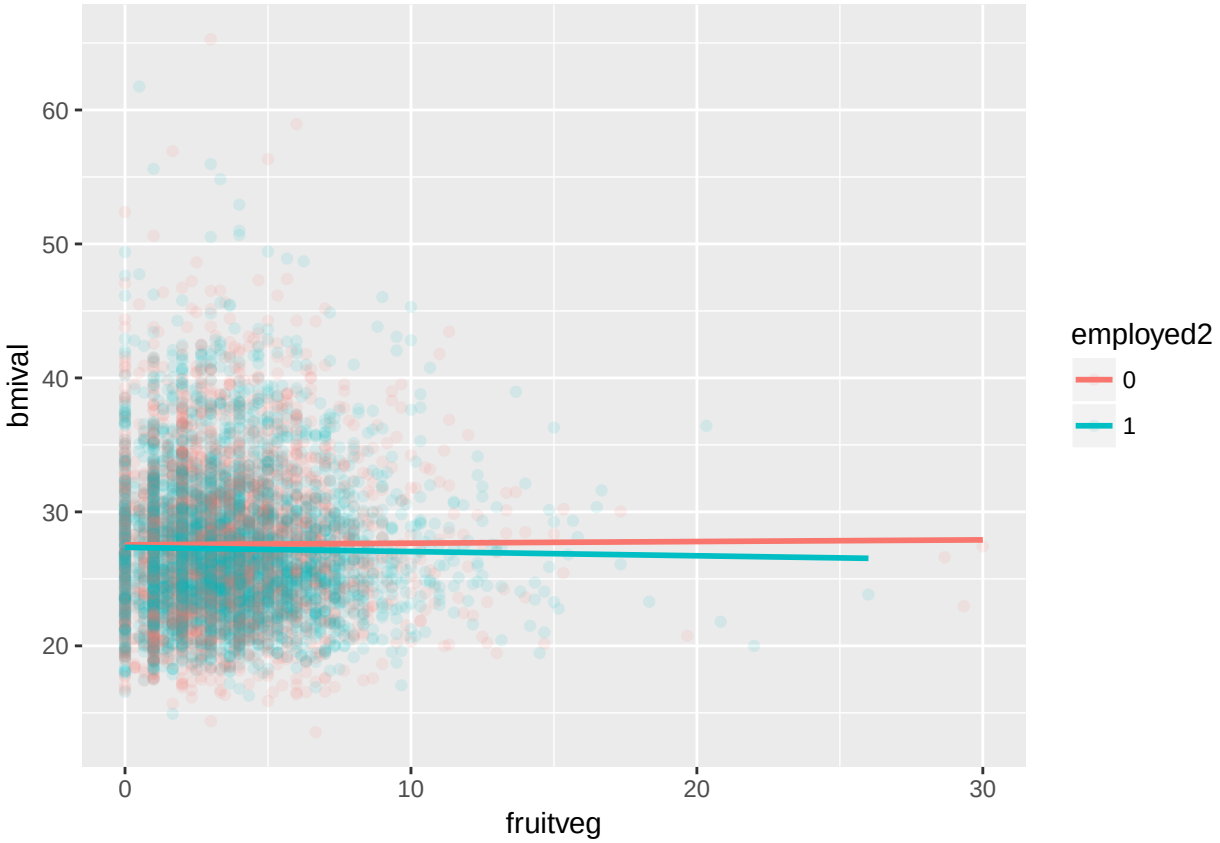
```
class(health$employed)
```

```
[1] "integer"
```

```
health2 <- subset(health, !is.na(employed) & !is.na(bmival) & !is.na(fruitveg))

health2$employed2 <- as.factor(health2$employed)

x11()
ggplot(health2, mapping=aes(x=fruitveg, y=bmival, colour=employed2)) +
      geom_point(alpha=1/10) +
      geom_smooth(method=lm, se=FALSE)
```



The scatterplot shows employed respondents' BMI values decrease as fruit and vegetable consumption increases, while unemployed respondents' BMI remain the same as fruit and vegetable consumption increases. We also see that most respondents consume less than 10 fruits and/or vegetables a day. Therefore, we might want to use a recoded version of `fruitveg` that only includes respondents that eat less than 10 to 15 fruits and/or vegetables a day.