SOCI 30005_PS1_Hinojosa

Cintia Hinojosa

April 26, 2019

```
nals <- read_csv("C://Users/cinti/Box Sync/Booth 2017-2018/Spring 2019/Statistical Methods of Research 2/TA Sessions/nals_synthetic.csv", col_names = TRUE)
```

```
## Parsed with column specification:
## cols(
##   id = col_double(),
##   annearn = col_double(),
##   occprest = col_double(),
##   sei = col_double(),
##   age = col_double(),
##   yearsed = col_double(),
##   gender = col_double(),
##   ln_earn = col_double(),
##   literacy = col_double(),
##   unemp = col_double(),
##   parented = col_double(),
##   Ethnicity = col_double(),
##   Language = col_double(),
##   Education = col_double()
## )
```

```
ls()
```

```
## [1] "nals"
```

```
names(nals)
```

```
##  [1] "id"        "annearn"   "occprest"  "sei"       "age"
##  [6] "yearsed"   "gender"    "ln_earn"   "literacy"  "unemp"
## [11] "parented"  "Ethnicity" "Language"  "Education"
```

```
summary(nals)
```

```
##        id               annearn          occprest          sei
##  Min.   :1.010e+10   Min.   :    15   Min.   :17.00   Min.   :17.00
##  1st Qu.:3.110e+10   1st Qu.: 12954   1st Qu.:33.00   1st Qu.:32.00
##  Median :5.080e+10   Median : 20800   Median :43.00   Median :42.00
##  Mean   :4.791e+10   Mean   : 25903   Mean   :43.79   Mean   :48.52
##  3rd Qu.:6.670e+10   3rd Qu.: 33280   3rd Qu.:51.00   3rd Qu.:64.00
##  Max.   :8.570e+10   Max.   :416000   Max.   :86.00   Max.   :97.00
##                      NA's   :1316
##       age           yearsed          gender          ln_earn
##  Min.   :25.00   Min.   : 0.00   Min.   :0.0000   Min.   :1.099
##  1st Qu.:31.00   1st Qu.:12.00   1st Qu.:0.0000   1st Qu.:5.580
##  Median :38.00   Median :13.00   Median :0.0000   Median :6.043
##  Mean   :38.89   Mean   :13.31   Mean   :0.4902   Mean   :6.018
##  3rd Qu.:46.00   3rd Qu.:16.00   3rd Qu.:1.0000   3rd Qu.:6.479
##  Max.   :59.00   Max.   :18.00   Max.   :1.0000   Max.   :9.508
##                                                   NA's   :344
##     literacy          unemp           parented       Ethnicity
##  Min.   : 49.25   Min.   :0.0000   Min.   : 0.00   Min.   :1.000
##  1st Qu.:253.73   1st Qu.:0.0000   1st Qu.: 9.00   1st Qu.:2.000
##  Median :292.81   Median :0.0000   Median :12.00   Median :5.000
##  Mean   :286.18   Mean   :0.0879   Mean   :10.93   Mean   :3.982
##  3rd Qu.:329.13   3rd Qu.:0.0000   3rd Qu.:13.00   3rd Qu.:5.000
##  Max.   :441.40   Max.   :1.0000   Max.   :18.00   Max.   :5.000
##
##     Language       Education
##  Min.   :1.000   Min.   :1.000
##  1st Qu.:5.000   1st Qu.:3.000
##  Median :5.000   Median :3.000
##  Mean   :4.477   Mean   :3.487
##  3rd Qu.:5.000   3rd Qu.:5.000
##  Max.   :5.000   Max.   :6.000
##
```

Probability

> Consider the population of US adults in the labor force in 1992. We are interested in the relationship between educational attainment and the risk of unemployment. Educational attainment has six possible values: no degree, GED, high school degree, associates degree, bachelors degree, and masters degree or higher.

Q1. Construct a theoretical contingency table with two rows (values of unemployment) and six columns (values of educational attainment) in which the entries are the joint and marginal probabilities. Use Greek letters to represent these (for example use it).

```r
# Label values for easy interpretation
nals$unemp <- factor(nals$unemp,
levels = c(0,1),
labels = c("employed", "unemployed"))

nals$Education <- factor(nals$Education,
levels = c(1,2,3,4,5,6),
labels = c("none", "ged", "hs", "aa", "ba", "grad"))



# Create a new contingency table with names for x and y dimensions (rows and columns)
nals.tab <- prop.table(table(nals$Education,nals$unemp,
                                    dnn = c("Education Attainment", "Unemployment Status")))

# Display contingency tables
nals.tab
```

```
##                     Unemployment Status
## Education Attainment    employed  unemployed
##                 none 0.106548191 0.019132245
##                 ged  0.033301313 0.005203330
##                 hs   0.392251041 0.040185719
##                 aa   0.117114954 0.010246558
##                 ba   0.169708614 0.008885687
##                 grad 0.093179635 0.004242715
```

```r
# Convert the contingency table into data frame for getting the marginals
nals.tab <- as.data.frame.matrix(nals.tab)

# add the joint probabilities across each row to get the marginals for various levels of educati
on attainment (rows)
nals.tab$marginal.education <- rowSums(nals.tab)

# add the joint probabilities across each row to get the marginals for the two levels of unemplo
yment (rows)
nals.tab["marginal.employment",] <- colSums(nals.tab)

# looking at the new table
nals.tab
```

```
##                       employed  unemployed marginal.education
## none                0.10654819 0.019132245         0.12568044
## ged                 0.03330131 0.005203330         0.03850464
## hs                  0.39225104 0.040185719         0.43243676
## aa                  0.11711495 0.010246558         0.12736151
## ba                  0.16970861 0.008885687         0.17859430
## grad                0.09317963 0.004242715         0.09742235
## marginal.employment 0.91210375 0.087896254         1.00000000
```

```r
# Crosstabulation of unemployment status and educational attainment
# N/T = joint probability
# Col and row percentages = marginal probaility
crosstab <- CrossTable(nals$unemp, nals$Education,
           expected = FALSE,
           prop.r = TRUE,
           prop.c = TRUE,
           prop.t = TRUE,
           prop.chisq = FALSE)
```

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  12492
##
##
##              | nals$Education
##   nals$unemp |      none |       ged |        hs |        aa |        ba |      grad | Row To
tal |
## -------------|-----------|-----------|-----------|-----------|-----------|-----------|-------
----|
##     employed |      1331 |       416 |      4900 |      1463 |      2120 |      1164 |     11
394 |
##              |     0.117 |     0.037 |     0.430 |     0.128 |     0.186 |     0.102 |     0.
912 |
##              |     0.848 |     0.865 |     0.907 |     0.920 |     0.950 |     0.956 |
|
##              |     0.107 |     0.033 |     0.392 |     0.117 |     0.170 |     0.093 |
|
## -------------|-----------|-----------|-----------|-----------|-----------|-----------|-------
----|
##   unemployed |       239 |        65 |       502 |       128 |       111 |        53 |      1
098 |
##              |     0.218 |     0.059 |     0.457 |     0.117 |     0.101 |     0.048 |     0.
088 |
##              |     0.152 |     0.135 |     0.093 |     0.080 |     0.050 |     0.044 |
|
##              |     0.019 |     0.005 |     0.040 |     0.010 |     0.009 |     0.004 |
|
## -------------|-----------|-----------|-----------|-----------|-----------|-----------|-------
----|
## Column Total |      1570 |       481 |      5402 |      1591 |      2231 |      1217 |     12
492 |
##              |     0.126 |     0.039 |     0.432 |     0.127 |     0.179 |     0.097 |
|

## -------------|-----------|-----------|-----------|-----------|-----------|-----------|-------
----|
##
##
```

$$Pr(Y = y) = \phi^y(1-\phi)^{1-y}, \text{for } y\,\epsilon\{0,1\}$$
$$y : \{1 = \text{unemployed}, 0 = \text{employed}\}$$
$$Pr(Y = 1) = \phi$$
$$Pr(Y = 0) = 1 - \phi$$

Q2. Define the marginal probability of unemployment and decompose it into the sum of the relevant joint probabilities.

Marginal Probabilities for Unemployment Status:

$$Pr(Y = y) = \sum_x Pr(Y = y, X = x)$$

$$\begin{aligned}
\text{Marginal probability of Unemployment:} \quad & Pr(Y = 0) = \sum_x Pr(Y = 1, X = x) && = 0.088 \\
\text{Marginal probability of Employment:} \quad & Pr(Y = 1) = \sum_x Pr(Y = 0, X = x) && = 0.912
\end{aligned}$$

The marginal probabilities tell us the probability of an event independent of other variables. Here, there is a 91.2% probability that a randomly selected individual will be employed and 8.8% probability they would be unemployed.

Joint Probabilities for Unemployment Status and Educational Attainment:

$$Pr(Y = 1, X = x) = x\epsilon\{1,2,4,5,6\}, y\epsilon\{0,1\}$$

$$\begin{aligned}
\text{Unemployed and no degree:} \quad & Pr(Y = 1) = Pr(Y = 1, X = 1) && = 0.019 \\
\text{Unemployed and GED:} \quad & Pr(Y = 1) = Pr(Y = 1, X = 2) && = 0.005 \\
\text{Unemployed and HS:} \quad & Pr(Y = 1) = Pr(Y = 1, X = 3) && = 0.040 \\
\text{Unemployed and AA:} \quad & Pr(Y = 1) = Pr(Y = 1, X = 4) && = 0.010 \\
\text{Unemployed and BA:} \quad & Pr(Y = 1) = Pr(Y = 1, X = 5) && = 0.009 \\
\text{Unemployed and grad:} \quad & Pr(Y = 1) = Pr(Y = 1, X = 6) && = 0.004
\end{aligned}$$

The joint probabilities tell us the probability that an individual has two attributes of certain levels. For example, from observing the joint probabilities in the data, there is a 1.9% probability of an individual being unemployed and having no educational degree, and for the most part, this probability shrinks as the level of educational attainment grows. Relative to the other probability estimates, unemployed individuals with a high school degree break away from this pattern, with the highest observed joint probability with unemployment at 4%.

Q3.For each possible level of education, define the conditional probability of unemployment.

Conditional Probabilities for Unemployment:

$$\text{Conditional Probabilities for Unempolyment} = \frac{\text{Joint Probability}}{\text{Marginal Probability}}$$

$$Pr(Y = 1|X = x) = \frac{Pr(Y = 1, X = x)}{Pr(X = x)}$$

$$\begin{aligned}
\text{Unemployed and no degree:} \quad & Pr(Y=1|X=1) = \frac{Pr(Y=1, X=1)}{Pr(X=1)} && = 0.1507937 \\
\text{Unemployed and GED:} \quad & Pr(Y=1|X=2) = \frac{Pr(Y=1, X=2)}{Pr(X=2)} && = 0.1282051 \\
\text{Unemployed and HS:} \quad & Pr(Y=1|X=3) = \frac{Pr(Y=1, X=3)}{Pr(X=3)} && = 0.0925926 \\
\text{Unemployed and AA:} \quad & Pr(Y=1|X=4) = \frac{Pr(Y=1, X=4)}{Pr(X=4)} && = 0.0787402 \\
\text{Unemployed and BA:} \quad & Pr(Y=1|X=5) = \frac{Pr(Y=1, X=5)}{Pr(X=5)} && = 0.0502793 \\
\text{Unemployed and grad:} \quad & Pr(Y=1|X=6) = \frac{Pr(Y=1, X=6)}{Pr(X=6)} && = 0.0412371
\end{aligned}$$

Conditional probability shows us the probability that an individual will experience an event, given another event. In this case, we are examining the probability that an individual is unemployed *given* the level of education they have attained. The data reveals a pattern in which the conditional probability of unemployment decreases as level of education attained increases, with the joint probability of being unemployed and having no degree at 15.1% while being unemployed and having a master's degree or higher is at 4.1%.

Q4. Decompose the joint probability of having no degree and being unemployed into the relevant marginal and conditional probabilities.

$$\text{Joint Probability} = \text{Conditional Probability} * \text{Marginal Probability}$$

$$\begin{aligned}
Pr(X=1, Y=1) &= Pr(Y=1|X=1) * Pr(X=1) \\
Pr(Y=1, X=1) &= Pr(X=1|Y=1) * Pr(Y=1)
\end{aligned}$$

Unemployed and no degree:

$$\begin{aligned}
Pr(Y=1|X=1) &= Pr(Y=1|X=1) * Pr(X=1) \\
&= (.019/.126) * (.126) \\
&= (.151) * (.126) \\
&= .019
\end{aligned}$$

We can break down the joint probability equation to see observe the conditional and marginal probabilities of unemployment and educational attainment. The joint probability for an individual to be unemployed and have no educational degree is 1.9%, the conditional probability that they will be unemployed given that they have no degree is 15.1%, and the marginal probability of having no educational degree is 12.6%.

Q5. Now decompose the marginal probability of being unemployed into a function of the relevant marginal and conditional probabilities.

Marginal probability derived from conditional probability and marginal probability:

$$Pr(Y = y) = \sum_x Pr(Y = y|X = x) * Pr(X = x)$$

$ $

$$\begin{aligned} Pr(Y = 1) &= Pr(Y = 1|X = 1) * Pr(X = 1) \\ &= (.019) * (.126) \\ &= (.002) \end{aligned}$$

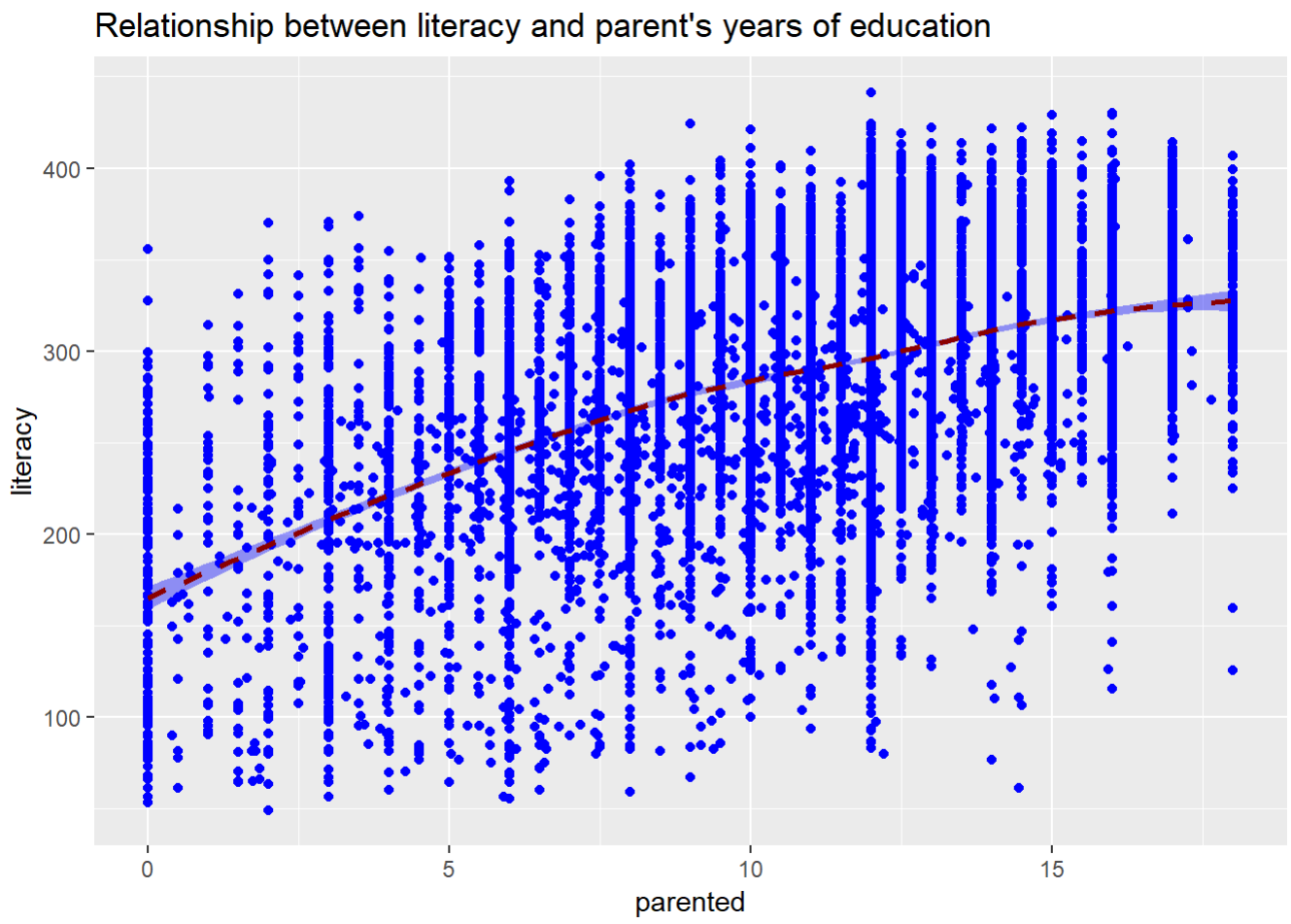
Q6. Using the NALS data, estimate the conditional probabilities of unemployment for each level of education. What does this seem to say about the association between education and unemployment?

Q7. Again using NALS, assume that unemployment and education were independent. What would then be the estimated conditional probability of unemployment given no degree? Under this scenario, how many of those with degree would we expect to be unemployed? Compare this to the number of those with no degree who were in fact unemployed and comment on how education is associated with joblessness for this group.

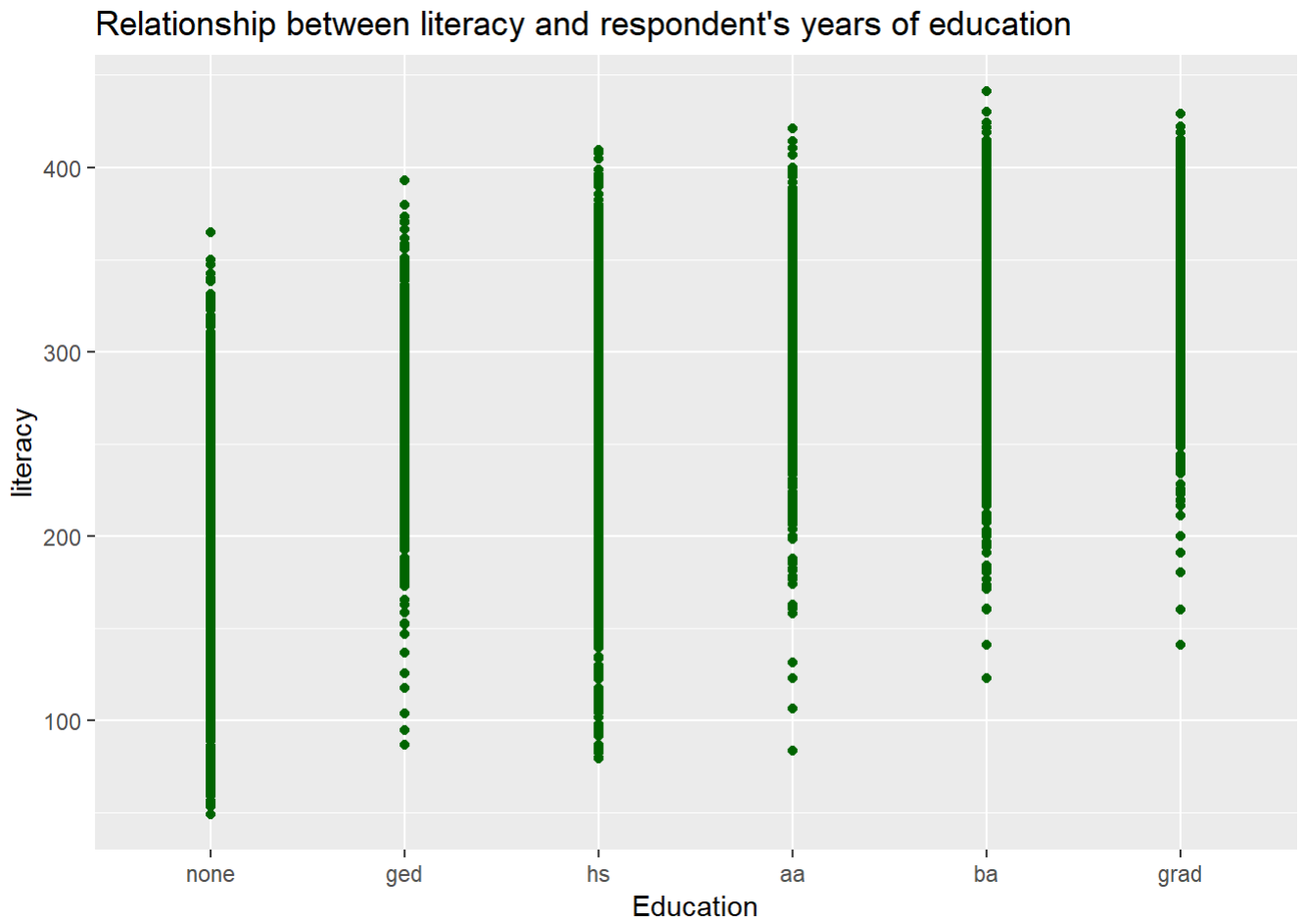
Expectation

> We are again going to work with NALS, now using three variables: parent years of education (X), respondent years of education (Z), and adult literacy (Y).

```
# Relationship between literacy and parental years of education
ggplot(nals, aes(parented,literacy)) + geom_point(color="blue") +
  labs(title="Relationship between literacy and parent's years of education") +
  geom_smooth(method="loess", linetype="dashed",color="darkred", fill="blue")
```




```
# Relationship between literacy and respondent's education
ggplot(nals, aes(Education,literacy)) + geom_point(color="darkgreen") +
  labs(title="Relationship between literacy and respondent's years of education") +
  geom_smooth(method="loess", linetype="dashed",color="darkred", fill="green")
```



There is a general positive linear relationships between the outcome variable, respondent's literacy level, with both of the independent variables, parent's years of education and respondent's educational attainment. There also doesn't seem to be any extreme outliners that stand out, so we can continue examining for statistically significant linear relationships through regression models.

Q1. Write down a theoretical linear regression model in which Y is a function of X and Z. Define the terms in the model. We will assume this is the true model" of the relationship between X,Z, and Y.

$$Y_i = \alpha + \beta_1 x + \gamma_2 z + \varepsilon_i,$$
$$\varepsilon_i \; N(0, \sigma_{y|x,z})$$

In the above model, Y_i is the literacy level of a particular individual in the population, α is the y-intercept for the model's regression line, meaning that it is the average literacy level for an individual who has no educational degree ($\gamma_2 z$) and parents with 0 years of education ($\beta_1 x$). ε_i is the difference between an individual's actual literacy level and the average literacy level for all individuals in the population who share the same values in regards to their educational attainment, Z, and their parent's years of education, X. $\sigma_{y|x,z}$ is the standard deviation for a particular subset of individuals in the populations with the same values for Z and X.

Q2. Write down two other linear regression models: a) Z is a function of X; and b) X is a function of Z;

$$Y_i = \alpha + \beta_1 x + \gamma_2 z + \varepsilon_i,$$
$$\varepsilon_i \ N(0, \sigma_{y|x,z})$$

Q3. Now suppose someone estimated a model using only Z as a predictor. That means the person would be studying the expected value of Y given Z alone.

a. Using (1), find E(Y|Z). b. Using (1) and (2b), define the bias involved. Show that it has two parts and define them.

$$\mathbb{E}\, Y = \mu(z)$$

$$Y_i = \beta_z - \gamma(z_i) + \varepsilon_i$$

In the above model, $Y_i = \beta_z$ is the naive estimate of the coefficient for the respondent's educational attainment as a sole predictor of their literacy level. The potential bias in the model is captured by $\gamma(z_i) + \varepsilon_i$, which is likely. The two conditions for bias to equal 0 is for $\gamma = 0$ or if there was no difference in the predicted average literacy level between individuals with varying years of education.

```
# linear regression of respondent's educational attainment on adult literacy
lm_literacy_ed <- lm(nals$literacy ~ nals$Education)
print(lm_literacy_ed)
```

```
##
## Call:
## lm(formula = nals$literacy ~ nals$Education)
##
## Coefficients:
##        (Intercept)   nals$Educationged    nals$Educationhs
##             204.40               65.20               74.70
##   nals$Educationaa    nals$Educationba  nals$Educationgrad
##              98.98              120.80              131.24
```

In a linear regression of respondent educational on adult literacy, the coefficients vary by level of educational attainment. The model predicts an average literacy score increase of 65.2 for respondents with a ged, 74.7 for those with a high school degree, 99 for those with an associates degree, 120.8 for those with a bachelor's degree, and 131.2 for respondents with a master's degree or higher. Respondents with no educational degree having a predicted average literacy score of 204.4 under this model.

$$y_i = 204.4 + 65.2 z_{ged} + \epsilon$$
$$y_i = 204.4 + 74.7 z_{hs} + \epsilon$$
$$y_i = 204.4 + 98.98 z_{aa} + \epsilon$$
$$y_i = 204.4 + 120.8 z_{ba} + \epsilon$$
$$y_i = 204.4 + 131.24 z_{grad} + \epsilon$$

Q4. Now suppose someone estimated a model using only X as a predictor. That means this person would be studying the expected value of Y given X alone (2c).

a. Using (1), find E(Y|X). Define the “total effect” of X on Y. b. What is the direct effect of X on Y based on your theoretical model (1)? c. Find the indirect effect of X on Y as it operates through Z. d. Show that the total effect of X on Y is the sum of the direct and indirect effects you have defined.

$$\mathbb{E}\, Y = \mu(x)$$

$$Y_i = \beta_x - \gamma(x_i) + \varepsilon_i$$

```
# linear regression of parental education on adult literacy
lm_literacy_ped <- lm(nals$literacy ~ nals$parented)
print(lm_literacy_ped)
```

```
##
## Call:
## lm(formula = nals$literacy ~ nals$parented)
##
## Coefficients:
##   (Intercept)  nals$parented
##        192.47           8.57
```

In a linear regression of parental education on adult literacy, parental education has a coefficient of 8.57, meaning that the model estimates an average increase of 8.57 in a a respondent’s literacy level for every added 1 year of parental education, with respondents with 0 years of parental edcuation having a predicted average literacy score of 192.47.

$$y_i = 192.47 + 8.57x + \epsilon$$

Q5. Estimate these total, direct, and indirect effects using NALS, and comment on what you have learned about how parent education and respondent education are linked to adult literacy.