hw_10

lisa liubovich

2024-04-10

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.4     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.0
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become error
```

```
library(broom)
library(GGally)
```

```
## Warning: package ’GGally’ was built under R version 4.3.2
```

```
## Registered S3 method overwritten by ’GGally’:
##   method from
##   +.gg   ggplot2
```

Steroid

1.

```
steroid <- read_csv("https://dcgerard.github.io/stat_415_615/data/steroid.csv")
```

```
## Rows: 27 Columns: 2
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## dbl (2): steroid, age
##
## i Use ‘spec()‘ to retrieve the full column specification for this data.
## i Specify the column types or set ‘show_col_types = FALSE‘ to quiet this message.
```

```r
glimpse(steroid)
```

```
## Rows: 27
## Columns: 2
## $ steroid <dbl> 27.1, 22.1, 21.9, 10.7, 1.4, 18.8, 14.7, 5.7, 18.6, 20.4, 9.2,~
## $ age     <dbl> 23, 19, 25, 12, 8, 12, 11, 8, 17, 18, 9, 21, 10, 25, 9, 17, 9,~
```

```r
lm_steroid_quad <- lm(steroid ~ age + I(age^2), data = steroid)
tidy(lm_steroid_quad)
```

```
## # A tibble: 3 x 5
##   term        estimate std.error statistic    p.value
##   <chr>          <dbl>     <dbl>     <dbl>      <dbl>
## 1 (Intercept)  -26.3      5.88       -4.48 0.000157
## 2 age            4.87     0.775       6.29 0.00000169
## 3 I(age^2)      -0.118    0.0235     -5.05 0.0000371
```

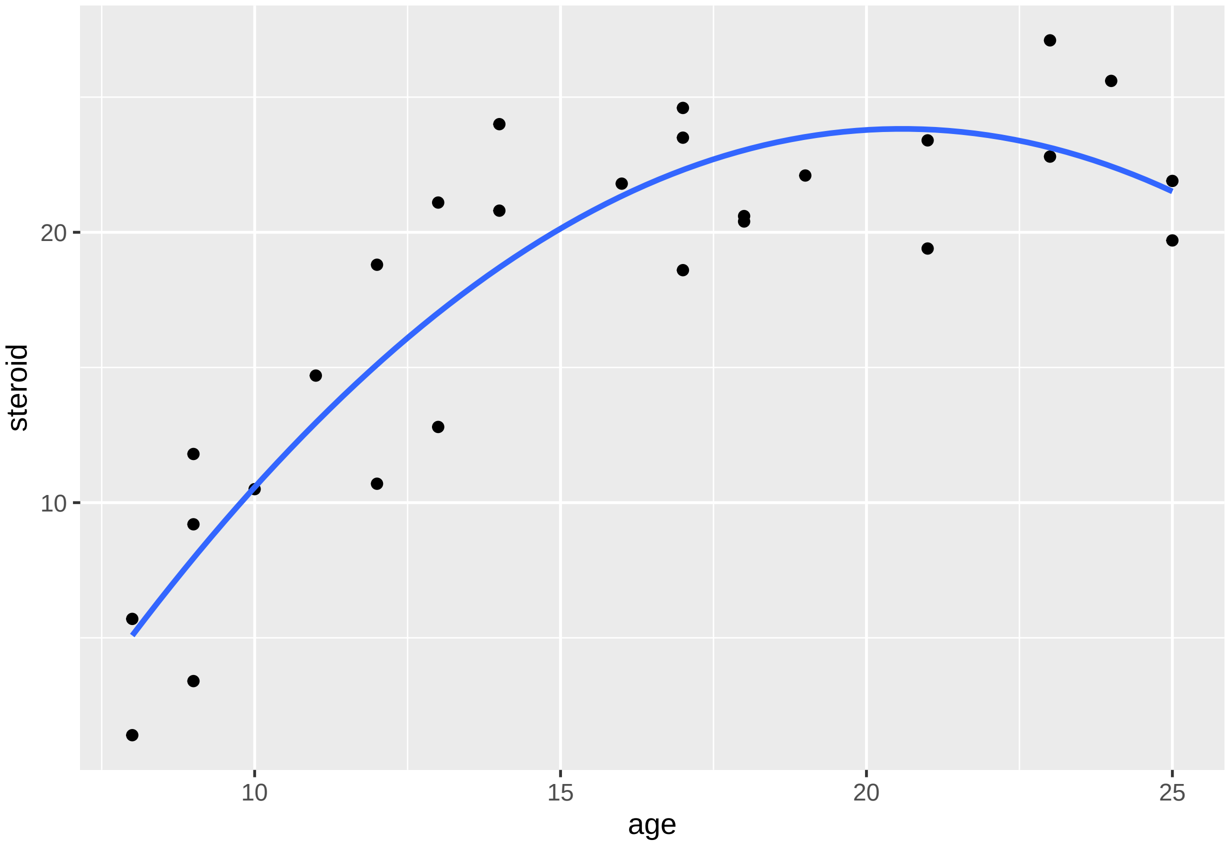
Model being fitted:

$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 \; X_{i2}^2 + \varepsilon_i$

Where Y_i is the predicted level of steroid for all observations i, X_{i1} is age in years for observations i, X_{i2} is age in years squared for observations i.

2.

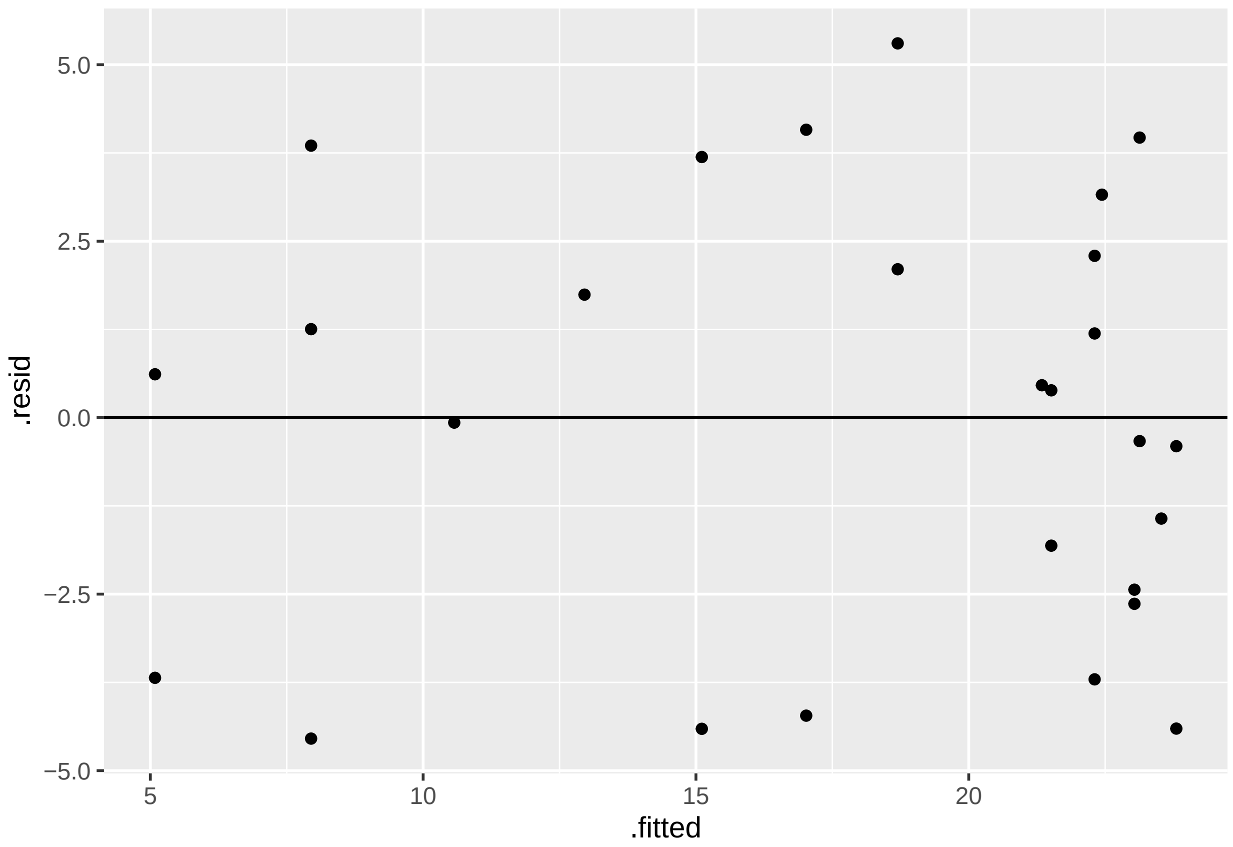
```r
ggplot(steroid, aes(x = age, y = steroid)) +
  geom_point() +
  geom_smooth(method = "lm",
              se = FALSE,
              formula = y ~ x + I(x^2))
```



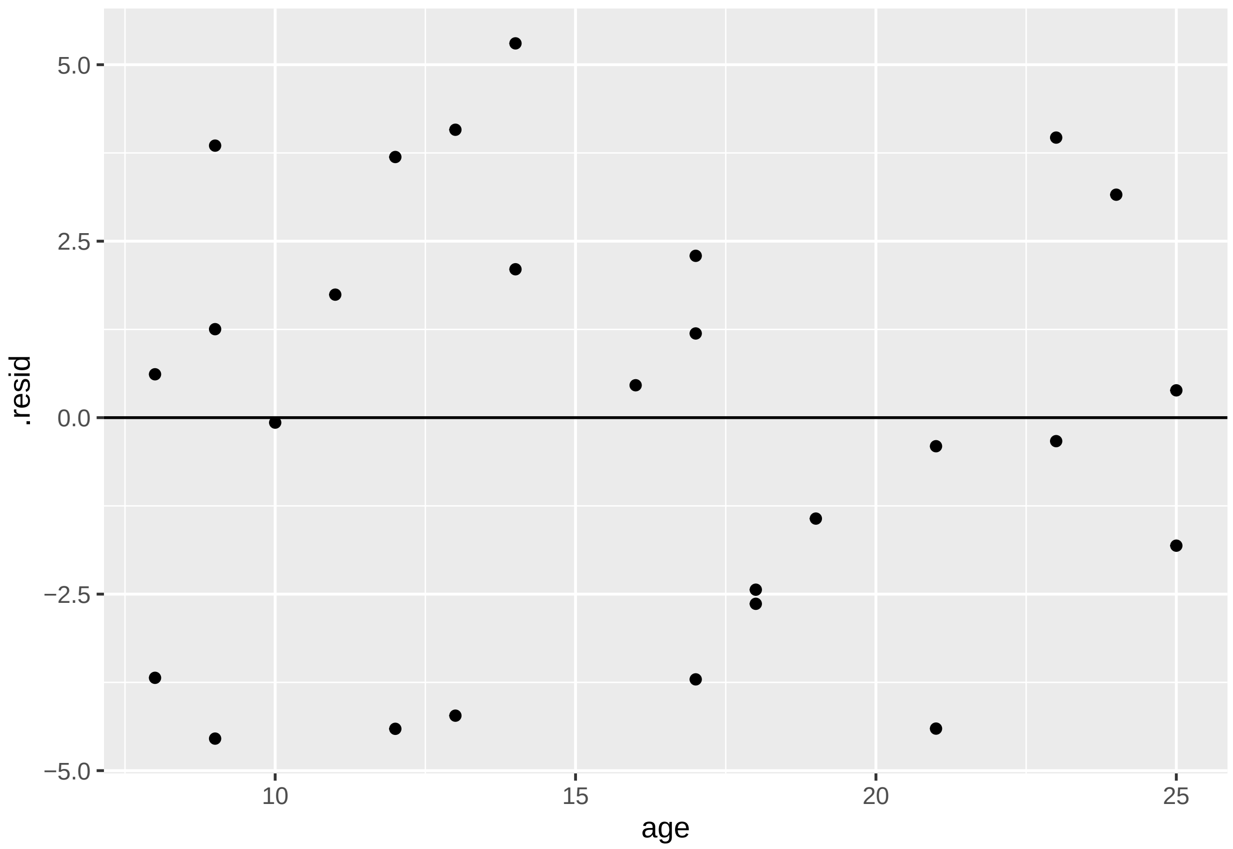
Yes, the quadratic regression function appears to be a good fit for the data here.

3.

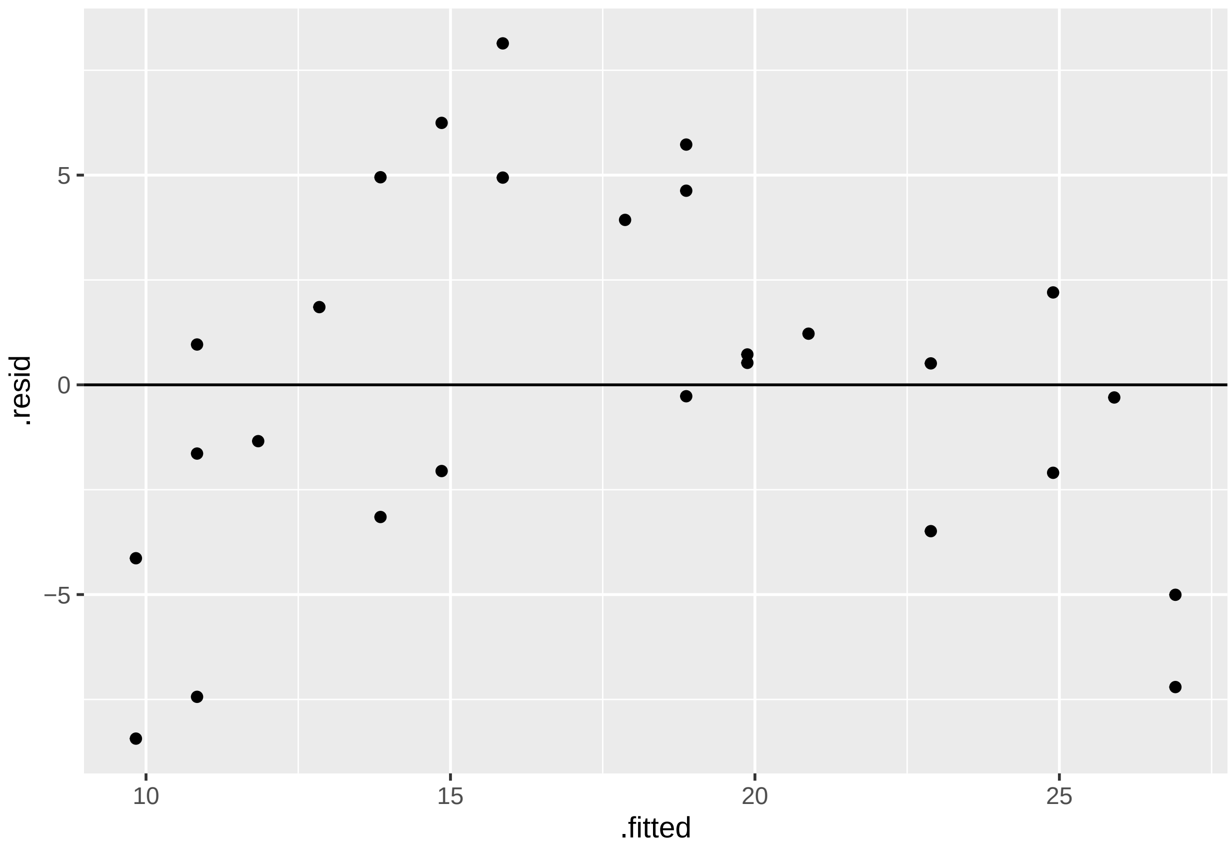
```
a_tlmsq <- augment(lm_steroid_quad)
ggplot(a_tlmsq, aes(x = .fitted, y = .resid)) +
  geom_point() +
  geom_hline(yintercept = 0)
```



```
ggplot(a_tlmsq, aes(x = age, y = .resid)) +
  geom_point() +
  geom_hline(yintercept = 0)
```



```
lm_steroid <- lm(steroid ~ age, steroid)
a_lmsteroid <- augment(lm_steroid)
ggplot(a_lmsteroid, aes(x = .fitted, y = .resid)) +
  geom_point() +
  geom_hline(yintercept = 0)
```



Without the quadratic term, the curve is a lot more pronounced. The vertical spread of the residual plot with the quadratic term isn't the best, but it isn't crazy, so we can assume constant variance.

4.

H_0: $\beta_1 = \beta_2 = 0 \rightarrow Y_i = \beta_0 + \varepsilon_i$

H_A: at least one β_1 or β_2 is non-zero $\rightarrow Y_i = \beta_0 + \beta_1\ X_{i1} + \beta_2\ X_{i2}^2 + \varepsilon_i$

```
glance(lm_steroid_quad)
```

```
## # A tibble: 1 x 12
##   r.squared adj.r.squared sigma statistic       p.value    df logLik   AIC   BIC
##       <dbl>         <dbl> <dbl>     <dbl>         <dbl> <dbl>  <dbl> <dbl> <dbl>
## 1     0.814         0.799  3.15      52.6 0.00000000168     2  -67.7  143.  149.
## # i 3 more variables: deviance <dbl>, df.residual <int>, nobs <int>
```

We have very strong evidence of an association between age and steroid (p-value = 1.67764e^{-09}).

5.

```
summary(lm_steroid_quad)
```

```
##
## Call:
## lm(formula = steroid ~ age + I(age^2), data = steroid)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.5463 -2.5369  0.3868  2.1973  5.3020
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -26.32541    5.88154  -4.476 0.000157 ***
## age           4.87357    0.77515   6.287 1.69e-06 ***
## I(age^2)     -0.11840    0.02347  -5.045 3.71e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.153 on 24 degrees of freedom
## Multiple R-squared:  0.8143, Adjusted R-squared:  0.7989
## F-statistic: 52.63 on 2 and 24 DF,  p-value: 1.678e-09
```

The multiple R^2 is 0.8143, meaning that 81.43% is the proportionate reduction of total variation in Y associated with the use of the set of the variables X_{i1} and X_{i2}^2 . This tells us that this model is a relatively good fit for the data.

6.

```r
ages <- data.frame(age = c(10, 15, 20))
predict(lm_steroid_quad, newdata = ages, interval = "confidence", level = 0.95)
```

```
##        fit       lwr      upr
## 1 10.57021  8.663491 12.47692
## 2 20.13792 18.295355 21.98049
## 3 23.78558 22.015714 25.55544
```

Mean steroid level @ age 10: 10.57 (will fall between 8.66 and 12.48 in 95% of repeated samples)

Mean steroid level @ age 15: 20.14 (will fall between 18.30 and 21.98 in 95% of repeated samples)

Mean steroid level @ age 20: 23.79 (will fall between 22.02 and 25.56 in 95% of repeated samples)

7.

first, I would check if an age of 4 is within the range of data.

```r
range(steroid$age)
```

```
## [1]  8 25
```

Since 4 is out of range of our data, I would refuse to make a prediction for a 4 year old female since it would be an inappropriate extrapolation.

8.

```
anova(lm_steroid, lm_steroid_quad)
```

```
## Analysis of Variance Table
##
## Model 1: steroid ~ age
## Model 2: steroid ~ age + I(age^2)
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     25 491.53
## 2     24 238.54  1    252.99 25.453 3.708e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

H_0: the quadratic term does not contribute significantly to the model

H_A: the quadratic term does contribute significantly to the model

We have strong evidence to reject the null hypothesis and conclude that the quadratic term does contribute significantly to the model (p-value of 3.708e^{-05}).

9.

H_0: $Y_i = \beta_0 + \beta_1\ X_{i1} + \varepsilon_i$

H_A: $Y_i = \beta_0 + \beta_1\ X_{i1} + \beta_{11}\ X_{i1}^2 + \varepsilon_i$

implicit assumption: the reduced model is a subset of the full model

```
anova(lm_steroid, lm_steroid_quad)
```

```
## Analysis of Variance Table
##
## Model 1: steroid ~ age
## Model 2: steroid ~ age + I(age^2)
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     25 491.53
## 2     24 238.54  1    252.99 25.453 3.708e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We have strong evidence to suggest a lack of fit of the reduced model; that is, we have strong evidence to reject the null hypothesis that the reduced model is true (p-value of 3.708e^{-05})

County Demographic Info

```
cdi <- read_csv("https://dcgerard.github.io/stat_415_615/data/cdi.csv")
```

```
## Rows: 440 Columns: 17
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (3): county, state, region
## dbl (14): id, area, pop, percent_18_34, percent_65, physicians, beds, crimes...
##
## i Use ‘spec()‘ to retrieve the full column specification for this data.
## i Specify the column types or set ‘show_col_types = FALSE‘ to quiet this message.
```

1.

```
cdi <- mutate(cdi,
              region = factor(region))
lm_cdi <- lm(physicians ~ region + pop + total_income, cdi)
tidy(lm_cdi)
```

```
## # A tibble: 6 x 5
##   term            estimate std.error statistic  p.value
##   <chr>              <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)   -62.0      57.0        -1.09   2.78e- 1
## 2 regionNE        3.49     78.8         0.0443 9.65e- 1
## 3 regionS        45.7      71.4         0.640  5.23e- 1
## 4 regionW      -146.       85.2        -1.71   8.82e- 2
## 5 pop             0.000551  0.000284    1.94   5.24e- 2
## 6 total_income    0.107     0.0133      8.07   6.80e-15
```

Model fit:

$Y_i = \beta_0 + \beta_1 \ X_{i1} + \beta_2 \ X_{i2} + \beta_3 \ X_{i3} + \beta_4 \ X_{i4} + \beta_5 \ X_{i5} + \varepsilon_i$

where:

Y_i is the number of professionally active non-federal physicians during 1990 for observations i

β_0 is the y - intercept

Region has 4 levels, so there are 3 classes which can be represented with three indicator variables:

X_{i1} : 1 if NE, 0 otherwise

X_{i2} : 1 if S, 0 otherwise

X_{i3} : 1 if W, 0 otherwise

NC is the base class

X_{i4} is the total population

X_{i5} is the total personal income

2.

```
lm_reduced <- lm(physicians ~ 1, cdi)
anova(lm_reduced, lm_cdi)
```

```
## Analysis of Variance Table
##
## Model 1: physicians ~ 1
## Model 2: physicians ~ region + pop + total_income
##   Res.Df        RSS Df  Sum of Sq      F    Pr(>F)
## 1    439 1406206299
## 2    434  139093455  5 1267112844 790.73 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We have strong evidence to reject the null hypothesis that the reduced model is true; that is, we have evidence of lack-of-fit for the reduced model and should include the other variables in order to have a good fit (p-value of 2.2e^{-16}).

3.

Model for south:

$Y_i = \beta_0 + \beta_1\ (0) + \beta_2\ (1) + \beta_3\ (0) + \beta_4\ X_{i4} + \beta_5\ X_{i5} + \varepsilon_i \rightarrow Y_i = \beta_0 + \beta_2 + \beta_4\ X_{i4} + \beta_5\ X_{i5} + \varepsilon_i$

Model for west:

$Y_i = \beta_0 + \beta_1\ (0) + \beta_2\ (0) + \beta_3\ (1) + \beta_4\ X_{i4} + \beta_5\ X_{i5} + \varepsilon_i \rightarrow Y_i = \beta_0 + \beta_3 + \beta_4\ X_{i4} + \beta_5\ X_{i5} + \varepsilon_i$

model for south - model for west = β_2 - β_3

```
tidy(lm_cdi)
```

```
## # A tibble: 6 x 5
##   term            estimate std.error statistic  p.value
##   <chr>              <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)   -62.0      57.0         -1.09   2.78e- 1
## 2 regionNE        3.49     78.8          0.0443 9.65e- 1
## 3 regionS        45.7      71.4          0.640  5.23e- 1
## 4 regionW      -146.       85.2         -1.71   8.82e- 2
## 5 pop             0.000551  0.000284     1.94   5.24e- 2
## 6 total_income    0.107     0.0133       8.07   6.80e-15
```

β_2 = 45.68986, β_3 = -145.5264

s(β_2) = 71.395417787, s(β_3) = 85.152925853

```
45.68986 - (-145.5264)
```

```
## [1] 191.2163
```

```
sqrt(71.395417787^2 + 85.152925853^2)
```

```
## [1] 111.123
```

```
191.2163 - 1.96*111.123
```

```
## [1] -26.58478
```

```
191.2163 + 1.96*111.123
```

```
## [1] 409.0174
```

The southern region has on average 191.2163 more physicians (95% CI: 26.58478 less physicians to 409.0174 more physicians) than the western region, adjusting for total population and annual income.

4.

H_0: $\beta_1 = \beta_2 = \beta_3 = 0$

H_A: at least one of β_1, β_2, or β_3 are non-zero

```
lm_cdi_reduced <- lm(physicians ~ pop + total_income, cdi)
anova(lm_cdi_reduced, lm_cdi)
```

```
## Analysis of Variance Table
##
## Model 1: physicians ~ pop + total_income
## Model 2: physicians ~ region + pop + total_income
##   Res.Df       RSS Df Sum of Sq      F Pr(>F)
## 1    437 140967081
## 2    434 139093455  3   1873626 1.9487  0.121
```

We do not have evidence to reject the null hypothesis that there is no geographic effect (p-value of 0.121).

5.

```
lm_cdi_int <- lm(physicians ~ region + (pop*total_income), cdi)
anova(lm_cdi, lm_cdi_int)
```

```
## Analysis of Variance Table
##
## Model 1: physicians ~ region + pop + total_income
## Model 2: physicians ~ region + (pop * total_income)
##   Res.Df       RSS Df Sum of Sq      F  Pr(>F)
## 1    434 139093455
## 2    433 137997531  1   1095924 3.4387 0.06436 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

There is weak evidence to suggest that there is an interaction effect between population and total income, thus there is weak evidence to suggest we should add such an interaction term to our model (p-value of 0.06436).

Conceptual Exercises

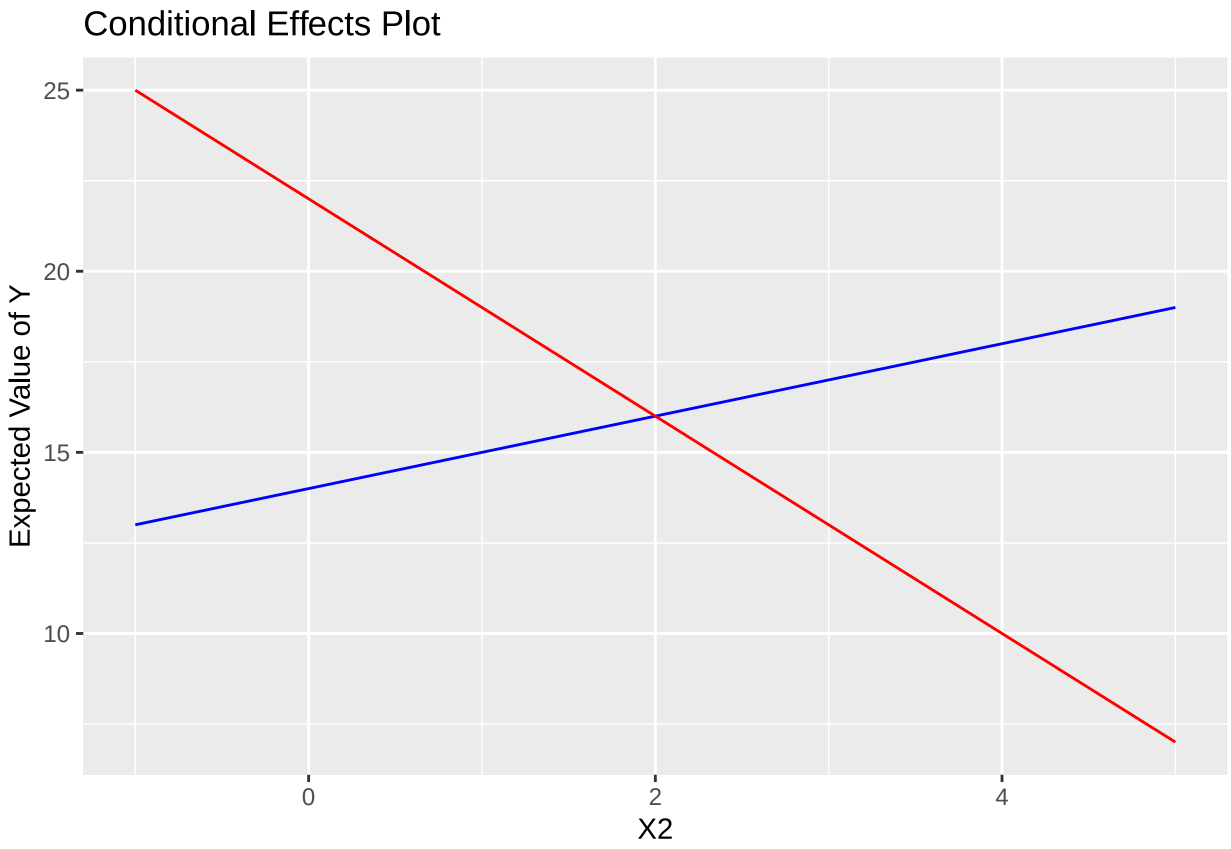
1.

with help from ChatGPT for syntax: https://chat.openai.com/c/dcef4e40-386b-49de-bd50-fea283b27a60

```
response_function <- function(X1, X2) 10 + 4*X1 + 3*X2 - 2*X1*X2
X2_range <- seq(-1, 5, by = 0.1)

data_plot <- data.frame(X2 = X2_range) %>%
  mutate(Y_X1_1 = response_function(X1 = 1, X2 = X2_range),
         Y_X1_3 = response_function(X1 = 3, X2 = X2_range))

ggplot(data_plot, aes(x = X2)) +
  geom_line(aes(y = Y_X1_1), color = "blue") +
  geom_line(aes(y = Y_X1_3), color = "red") +
  labs(x = "X2", y = "Expected Value of Y", title = "Conditional Effects Plot") +
  scale_color_manual(values = c("blue", "red"), name = "X1", labels = c("1", "3"))
```



The plot shows that the effects of X_1 and X_2 on Y are not additive by showing that the association between X_1 and Y are dependent on the levels of X_2 and vice versa; specifically, if the effects were additive, the lines would be parallel. The lines are clearly intersecting, indicating that the relationship between X_2 and Y is not the same at different levels of X_1.

2.

I would hesitate to say that this is an unfair coding scheme, as our domain knowledge indicates that there are societal factors that systematically disadvantage women in learning opportunities or environments related to the task being studied. The positive coefficient for gender could reflect underlying biases rather than inherent learning differences between men and women. If you recoded the variable to be 1 for females and 0 otherwise and you found a negative coefficient for the variable, that could tell you that the coding was fair.