Question 3

Joseph Froelicher

December 3, 2020

Part A

```r
fit1 = lm(iq ~ miles + first2y + miles * first2y, data = lead)
fit1$coefficients
```

```
##   (Intercept)         miles       first2y miles:first2y 
##    98.4513331     0.5913399   -19.6116047    17.6546119
```

$$Y_{iq} = \beta_{intercept} + \beta_{miles}X_1 + \beta_{first2y}X_2 + \beta_{interaction}X_1X_2$$

$$\text{Y} = 98.4513331 + 0.5913399\text{X1} + \text{-}19.6116047\text{X2} + 17.6546119\text{X1X2}$$

The average IQ score for those who live 0 miles from the nearest smelter, and were not exposed during their first two years of life is 98.4513331. The difference in IQ score at 0 miles from the nearest smelter between those who were not exposed during their first two years of life, and those that were exposed during their first two years of life is 0.5913399. The slope for those who were not exposed during their first two years of life is -19.6116047. And the difference in slope between those exposed in their first two years, and those not exposed in their first two years is 17.6546119.

Part B

```r
summary(fit1)
```

```
## 
## Call:
## lm(formula = iq ~ miles + first2y + miles * first2y, data = lead)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -48.549  -9.405   0.154   9.926  45.861 
## 
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)    
## (Intercept)    98.4513     4.3862  22.446  < 2e-16 ***
## miles           0.5913     2.1047   0.281  0.77922    
## first2y       -19.6116     7.8583  -2.496  0.01393 *  
## miles:first2y  17.6546     5.4811   3.221  0.00164 ** 
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 15.37 on 120 degrees of freedom
```

```
## Multiple R-squared:  0.09873,    Adjusted R-squared:  0.0762
## F-statistic: 4.382 on 3 and 120 DF,  p-value: 0.005783
```

$$H_0 : \beta_{interaction} = 0$$

$$H_A : \beta_{interaction} \neq 0$$

There is evidence to suggest that the difference of of slopes between those who were exposed in their first two years, and those who were not exposed in their first two years is not 0 (p = 0.0016442, [6.8024443, 28.5067795]).

Part C

$$Y_{iq} = \beta_{intercept} + \beta_{miles} X_1$$

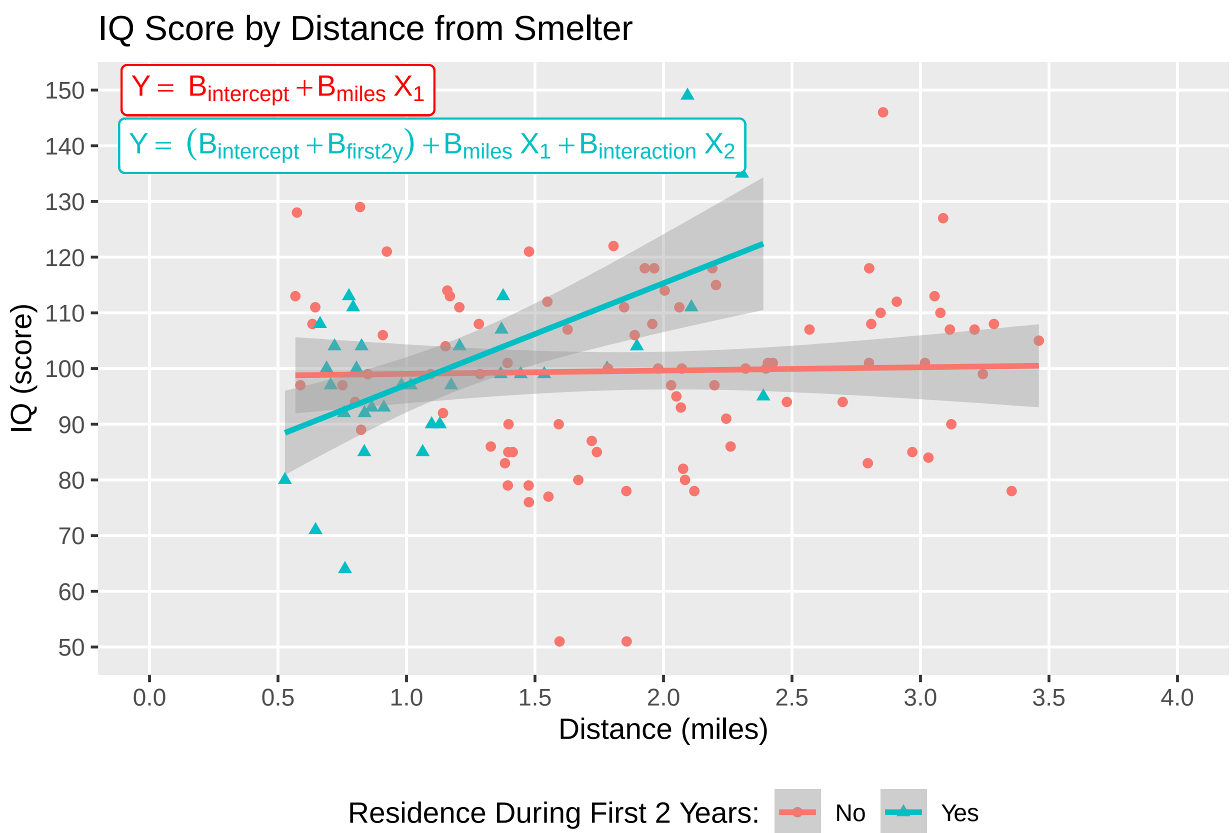
$$\text{Y} = 98.4513331 + 0.5913399\text{X1}$$

Part D

$$Y_{iq} = (\beta_{intercept} + \beta_{first2y}) + \beta_{miles} X_1 + \beta_{interaction} X_1$$

$$\text{Y} = 78.8397284 + 0.5913399\text{X1} + 17.6546119\text{X1}$$

Part E



Part F

$$H_0 : \beta_{miles} = 0$$
$$H_A : \beta_{miles} \neq 0$$

There is not enough evidence to suggest that the slope of those who were not exposed in their first two years is not 0 (p = 0.7792212, ci = [-3.5757711, 4.758451]).

Part G

```
vcov = vcov(fit1)
se = sqrt(vcov[2, 2] + vcov[4, 4] + 2 * vcov[2, 4])
alpha = 0.05

ci = c(
  fit1$coefficients[2] + fit1$coefficients[2] - qnorm(1 - (alpha / 2)) * se,
  fit1$coefficients[2] + fit1$coefficients[2] + qnorm(1 - (alpha / 2)) * se
)

t = (fit1$coefficients[2] + fit1$coefficients[2]) / se
p = 2 * pt(t, dim(lead)[1] - 1, lower.tail = FALSE)
```

$$H_0 : \beta_{miles} + \beta_{interaction} = 0$$
$$H_A : \beta_{miles} + \beta_{interaction} \neq 0$$

There is not enough evidence to suggest that the slope of those who were exposed in their first two years is not 0 (p = 0.8156142, ci = [-8.7364881, 11.1018479]).

Part H

Those who were not exposed during their first two years of life are accounting for a 0.5913399 increase in IQ score for each unit increase in miles from the smelter (95% Confidence Interval: [-3.5757711, 4.758451]). Those who were exposed in their first two years of life are accounting for a 1.1826799 increase in IQ score (95% Confidence Interval: [-8.7364881, 11.1018479]). Neither of which is a significant result.

Part I

```
fit_empty = lm(iq ~ miles, lead)
full = anova(fit1)
restricted = anova(fit_empty)

rss_restricted = restricted$`Sum Sq`[2]
rss_full = full$`Sum Sq`[4]
k_full = ( sum(full$Df)-full$Df[4] )
k_restricted = ( sum(restricted$Df) - restricted$Df[2] )
n = nrow(lead)

f = ( (rss_restricted - rss_full) / (k_full - k_restricted) ) / ( rss_full / (n - k_full) )
```

```r
p = pf(f, k_full - k_restricted, n - k_full, lower.tail = FALSE)
q = qf(1 - alpha, k_full - k_restricted, n - k_full, lower.tail = TRUE)
```

Null Hypothesis: The full model with miles and first two years and the interaction between, does not account for the variability in IQ score

Alternative Hypothesis: The full model with miles and first two years and the interaction between, accounts for a significantly more amount of variability in IQ score.

$$F = \frac{(\frac{RSS_{restricted} - RSS_{full}}{k_{full} - k_{restricted}})}{(\frac{RSS_{full}}{n - k_{full}})}$$

F-statistic: 5.6806786
Critical value: 3.0711405
p-value: 0.0043849

We reject the null hypothesis that the full model is not better ($F(2, 121) = 5.6806786$, $p(F > 3.0711405) = 0.0043849$), in favor of using the full model. There is evidence to suggest that the full model is accounting for significantly more variability than the restricted model.