1a

```r
lead <-  read.csv('D:/3rd_Semester/6611_biostatisticalmethod/hw10/lead2.csv')
```

```r
model1 <- glm(expose ~ iq, data=lead)
summary(model1)
summary(model1)$coefficients
```

```
Call:
glm(formula = expose ~ iq, data = lead)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.7199  -0.3769  -0.2732   0.5624   0.7804

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.084333   0.269727   4.020 0.000101 ***
iq          -0.007146   0.002668  -2.678 0.008421 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 0.2240058)

    Null deviance: 28.935  on 123  degrees of freedom
Residual deviance: 27.329  on 122  degrees of freedom
AIC: 170.37

Number of Fisher Scoring iterations: 2
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.084333427	0.269727112	4.020113	0.0001011308
iq	-0.007146336	0.002668307	-2.678228	0.0084214222

iq is linearly associated with lead exposure with p value

1b

```
gender = glm(iq ~ expose + sex, data = lead )
summary(gender)
```

```
Call:
glm(formula = iq ~ expose + sex, data = lead)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-52.664   -9.935    0.921    9.839   45.336

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  106.002      4.439  23.878   <2e-16 ***
expose        -7.916      2.911  -2.719   0.0075 **
sex           -2.338      2.887  -0.810   0.4196
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 244.2547)

    Null deviance: 31462  on 123  degrees of freedom
Residual deviance: 29555  on 121  degrees of freedom
AIC: 1038.6

Number of Fisher Scoring iterations: 2
```

1c

```
crude = model1$coefficients[2]
crude
b_adj = gender$coefficients[2]
b_adj
CR = (crude - b_adj)/crude
CR

```

iq: -0.00714633583677236

expose: -7.91631162507608

iq: -1106.74413712013

here, CR < 0.2 linearly associated with lead exposure with p value when adjusting with sex, not a cofounder with an operational criterion < 20%

2a

```r
crud_model <- glm(iq ~ miles , data=lead)
summary(crud_model)$coefficients
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	95.682821	3.444881	27.775364	1.384344e-54
miles	2.403658	1.819210	1.321264	1.888858e-01

```r
BCrude <- 2.403658
```

```r
Adj_model <- glm(iq ~ miles + first2y , data=lead)
summary(Adj_model)$coefficients
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	93.414592	4.253769	21.9604270	4.958560e-44
miles	3.194467	2.017205	1.5836105	1.158926e-01
first2y	3.210676	3.527586	0.9101625	3.645463e-01

```r
BAdj <- 3.194467
```

```r
BM <- 3.210676
```

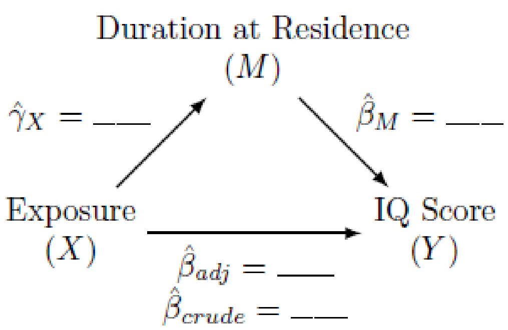
```r
BM_stdE <- 3.527586
```

```r
covariate <- glm(first2y ~  miles  , data=lead)
summary(covariate)$coefficients
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.7064646	0.08847536	7.984873	8.825372e-13
miles	-0.2463060	0.04672304	-5.271617	5.928597e-07

```r
GM <- -0.2463060
```

```r
GM_stdE <- 0.04672304
```



Beta M = 3.2106

Gamma X = -0.24630

Beta adj = 3.194467

Beta crude = 2.403658

2b

```r
prp<-(BCrude - BAdj)/BCrude
prp
```

-0.329002295667686

2c

```r
std_e <- sqrt((GM^2)*(BM_stdE^2)+(BM^2)*(GM_stdE^2))
std_e
```

0.881720584601002

```r
z <- prp/std_e
z
```

-0.373136684584229

```r
P <- 2*pnorm(z)
P
```

0.709046716770284

```r
CI_upper <- (BCrude - BAdj) + 1.96*std_e
CI_upper
CI_lower <- (BCrude - BAdj) - 1.96*std_e
CI_lower
```

0.937363345817964

-2.51898134581796

```r
prop_upper <- CI_upper/BCrude
prop_lower <- CI_lower/BCrude
```

```r
print(c(prop_lower, prop_upper))
```

```
[1] -1.0479783  0.3899737
```

2d

Std_e = 0.88, prp = -0.329

95%CI = -0.329 ± 1.96*0.88 = [-2.053, 1.39]

Proportioned mediation-

[-2.053/4.527, 1.39/4.527] = [-0.45,0.30]

3a

```r
mod2 <- glm(iq ~ miles + first2y + miles*first2y, data=lead)
summary(mod2)
```

```
Call:
glm(formula = iq ~ miles + first2y + miles * first2y, data = lead)

Deviance 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

(Dispersion parameter for gaussian family taken to be 236.2981)

    Null deviance: 31462  on 123  degrees of freedom
Residual deviance: 28356  on 120  degrees of freedom
AIC: 1035.5

Number of Fisher Scoring iterations: 2
```

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \hat{\beta}_3 X_1 X_2$$

= 98.45 + 0.59miles -19.61first2y + 17.65(miles * first2y)

$\beta 0$: The expected IQ for children living 0 miles from the smelter who did not live in the residence during the first two years of life is 98.451 points

$\beta 1$: estimated children who did not live in the current residence during the first two years of life, IQ increases, on average, by 0.591 points for every mile of distance the child currently lives from the smelter.

$\beta 2$: For children who live 0 miles from the smelter, IQ scores, on average, are 19.611 points lower for children who lived in the current residence during the first two years of life.

$\beta 3$: This is the difference between the effect of miles for those exposed during the first 2 years of life compared to those not exposed during the first 2 years of life. For children exposed in the first two years of life, a one mile increase in distance from the smelter results in an IQ score that is 17.655 points higher, on average

```r
vcov(mod2)
```

	(Intercept)	miles	first2y	miles:first2y
(Intercept)	19.238636	-8.570863	-19.238636	8.570863
miles	-8.570863	4.429659	8.570863	-4.429659
first2y	-19.238636	8.570863	61.752661	-38.835956
miles:first2y	8.570863	-4.429659	-38.835956	30.042297

3b

For this problem we need to interpret the interaction beta coefficient (β^miles∗first2y).

The relationship between IQ and miles is significantly different for children who lived in the residence during the first two years of life compared to children who did not live in the residence during the first two years of life ($p=0.0016$).

3c

who didn't live in the residence during the first two years of life
so First2y = 0

The regression equation for non-residence of 1st 2years:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \hat{\beta}_3 X_1 X_2$$

= 98.45 + 0.59miles -19.61first2y + 17.65(miles * first2y)

= 98.45 + 0.59miles -19.61*0 + 17.65(miles * 0)
= 98.45 + 0.59miles

3d

who lived in the current residence during the first two years
of life First2y = 1

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \hat{\beta}_3 X_1 X_2$$

= 98.45 + 0.59miles -19.61first2y + 17.65(miles * first2y)

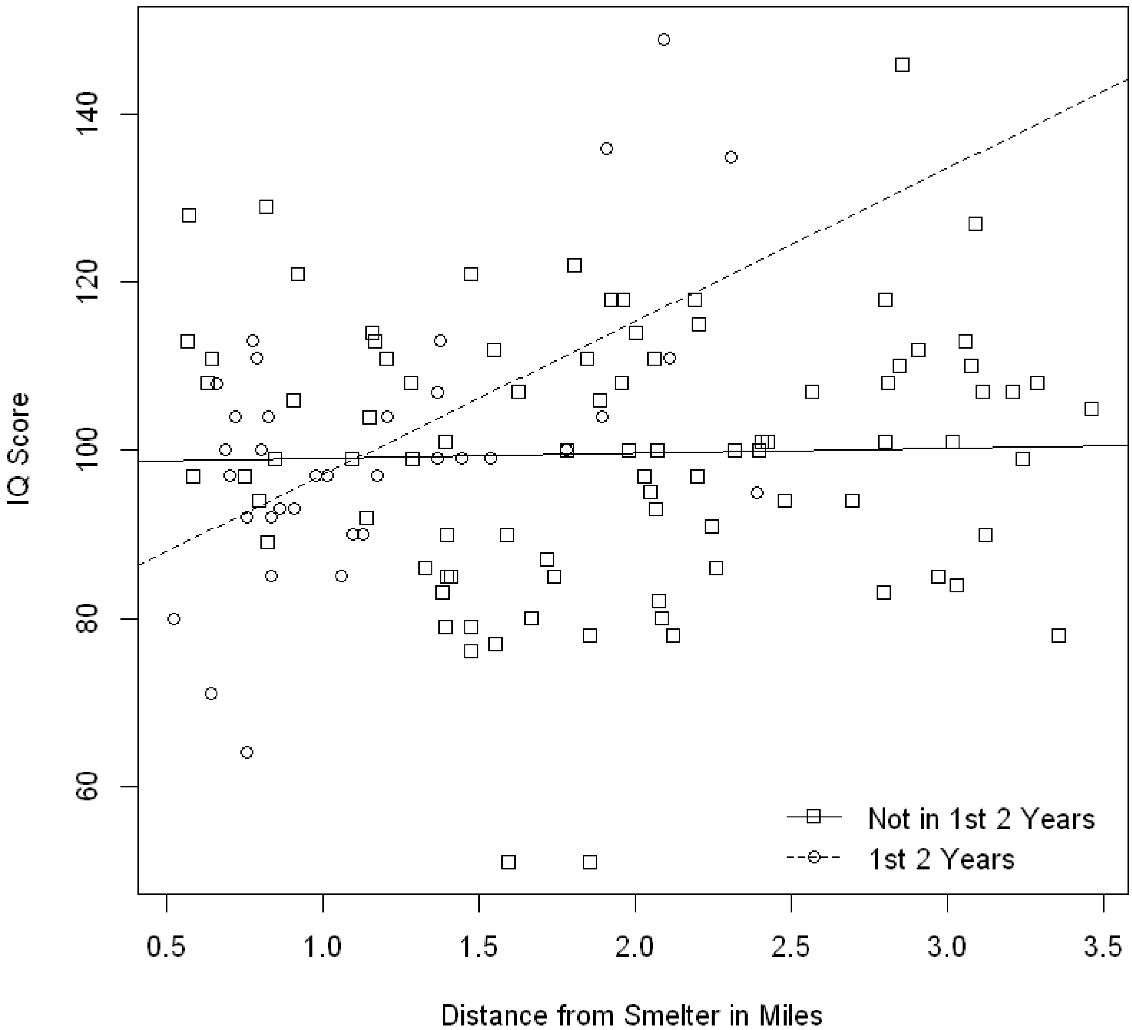
= 98.45 + 0.59miles -19.61*1 + 17.65(miles * 1)

= 98.45 + 0.59miles + 17.65miles – 19.61
= (98.45 – 19.61) + (0.59 + 17.65)*miles

= 78.84 + 18.24*miles

3e

```r
plot(x=lead$miles, y=lead$iq, pch=lead$first2y, xlab='Distance from Smelter in Miles', ylab='IQ Score')
abline(a=98.451, b=0.591, lty=1) # didn't live within first 2 years
abline(a=78.840, b=18.246, lty=2) # did live within first 2 years
legend('bottomright', bty='n', pch=c(0,1), lty=c(1,2), legend=c('Not in 1st 2 Years','1st 2 Years'))
```



There is definitely an interaction as at 1.0 distance from smelter in miles where the IQ score is exactly 100.

3f.

t-test result from 3A, for miles p = 0.7792
we fail to reject the null hypothesis that the slope is significantly different from 0 for those who did not live in the residence during the 1^{st} 2 years of life.

3g

```
vcov(mod2)
```

	(Intercept)	miles	first2y	miles:first2y
(Intercept)	19.238636	-8.570863	-19.238636	8.570863
miles	-8.570863	4.429659	8.570863	-4.429659
first2y	-19.238636	8.570863	61.752661	-38.835956
miles:first2y	8.570863	-4.429659	-38.835956	30.042297

```
lead$nt_1st2y <- abs(lead$first2y -1)
```

```
m_reverse <- lm(iq ~ miles+nt_1st2y+ miles*nt_1st2y, data = lead )
round(cbind(summary(m_reverse)$coefficients, confint(m_reverse)), 4)
```

	Estimate	Std. Error	t value	Pr(>|t|)	2.5 %	97.5 %
(Intercept)	78.8397	6.5203	12.0915	0.0000	65.9300	91.7494
miles	18.2460	5.0609	3.6053	0.0005	8.2257	28.2662
nt_1st2y	19.6116	7.8583	2.4957	0.0139	4.0527	35.1705
miles:nt_1st2y	-17.6546	5.4811	-3.2210	0.0016	-28.5068	-6.8024

P value is 5 x 10^(-4)

3h

#The relationship between IQ and distance from smelter differs significantly for exposed in the first 2 years compared to unexposed children (p=0.0016). There is a significant interaction with exposure during the first 2 years of life.

On an average, IQ increases by 0.59 points for a mile increase in distance from the smelter (95% CI: -3.53, 4.72 points) for children unexposed in the first 2 years, but this is not significantly different from 0 (p=0.78).

On an average, IQ increases by 18.25 points for a mile increase in distance from the smelter (95% CI: 8.22, 28.26 points) for children exposed in the first 2 years, this is significantly different from 0 (p=0. 0005).

On an average, IQ increase an average of 17.65 points more per mile of distance in children exposed in the first 2 years compared to unexposed (95% CI: 6.91 to 28.4 points/mile).

3i

```
one.way <- aov(iq ~ first2y, data = lead)

summary(one.way)
```

```
             Df Sum Sq Mean Sq F value Pr(>F)
first2y       1     16   16.26   0.063  0.802
Residuals   122  31446  257.75
```

```
two.way <- aov(iq ~ first2y + miles, data = lead)

summary(two.way)
```

```
             Df Sum Sq Mean Sq F value Pr(>F)
first2y       1     16    16.3   0.064  0.801
miles         1    639   638.5   2.508  0.116
Residuals   121  30807   254.6
```

Adding miles to the model seems to have made the model better: it slightly reduced the residual variance (the residual sum of squares went from 257.75 to 254.6)

```
interaction <- aov(iq ~ first2y*miles, data = lead)

summary(interaction)
```

```
               Df Sum Sq Mean Sq F value  Pr(>F)
first2y         1     16    16.3   0.069 0.79353
miles           1    639   638.5   2.702 0.10283
first2y:miles   1   2452  2451.6  10.375 0.00164 **
Residuals     120  28356   236.3
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

I think after adding interaction between models, have made the model better: it reduced the residual variance (the residual sum of squares went from 254.6 to 236.3)

```
interaction_m <- aov(iq ~ miles+ first2y + first2y*miles, data = lead)

summary(interaction_m)
```

```
               Df Sum Sq Mean Sq F value  Pr(>F)
miles           1    444   443.9   1.878 0.17308
first2y         1    211   210.9   0.893 0.34668
miles:first2y   1   2452  2451.6  10.375 0.00164 **
Residuals     120  28356   236.3
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
interaction_1st <- aov(iq ~ miles + first2y*miles, data = lead)

summary(interaction_1st)
```

```
               Df Sum Sq Mean Sq F value  Pr(>F)
miles           1    444   443.9   1.878 0.17308
first2y         1    211   210.9   0.893 0.34668
miles:first2y   1   2452  2451.6  10.375 0.00164 **
Residuals     120  28356   236.3
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The addition of first2y and the interaction between miles and first2y has same impact so even if we don’t add still the residual sum of squares is 236.3. I don’t see any significant change.