HW 2

Q1

1.1

Taking derivative of $argmin_\beta E((Y_i - X_i^T\beta)^2)$ we obtain $E(X_i(Y_i - X_i\beta)) = 0$. Then we have

$$\begin{aligned} E(X_i(Y_i - X_i\beta)) &= 0 \\ E(X_iY_i) - E(X_iX_i^T\beta) &= 0 \\ E(X_iY_i) &= E(X_iX_i^T)\beta \\ \beta_{ols} &= (E(X_iX_i^T))^{-1}E(X_iY_i) \end{aligned}$$

1.2

$$\begin{aligned} Y_i &= X_i^T\beta^* + \epsilon_i \\ X_iY_i &= X_iX_i^T\beta^* + X_i\epsilon_i \\ E(X_iY_i) &= E(X_iX_i^T)\beta^* \\ \beta^* &= E(X_iX_i^T)^{-1}E(X_iY_i) \\ \beta^* &= \beta_{ols} \end{aligned}$$

1.3

Regarding $E(Y_i|X_i)$ as Y_i,

$$\begin{aligned} \beta_1 &= (E(X_iX_i^T))^{-1}E(X_iE(Y_i|X_i)) \\ &= (E(X_iX_i^T))^{-1}E(E(X_iY_i|X_i)) \\ &= (E(X_iX_i^T))^{-1}E(X_iY_i) \\ &= \beta_{ols} \end{aligned}$$

1.4

Let

$\beta_0 = E(Y_i|X_{1i} = 0, X_{2i} = 0)$,

$\beta_1 = E(Y_i|X_{1i} = 1, X_{2i} = 0) - \beta_0$,

$\beta_2 = E(Y_i|X_{1i} = 0, X_{2i} = 1) - \beta_0$,

$\beta_3 = E(Y_i|X_{1i} = 1, X_{2i} = 1) - \beta_0 - \beta_1 * X_{1i} - \beta_2 * X_{2i}$

Therefore,

$$\begin{aligned}
E(Y_i|X_{1i}, X_{2i}) &= E(Y_i|X_{1i} = 0, X_{2i} = 0) \\
&+ [E(Y_i|X_{1i} = 1, X_{2i} = 0) - E(Y_i|X_{1i} = 0, X_{2i} = 0)] * X_{1i} \\
&+ [E(Y_i|X_{1i} = 0, X_{2i} = 1) - E(Y_i|X_{1i} = 0, X_{2i} = 0)] * X_{2i} \\
&+ (E(Y_i|X_{1i} = 1, X_{2i} = 1) - \{E(Y_i|X_{1i} = 0, X_{2i} = 0) \\
&+ [E(Y_i|X_{1i} = 1, X_{2i} = 0) - E(Y_i|X_{1i} = 0, X_{2i} = 0)] * X_{1i} \\
&+ [E(Y_i|X_{1i} = 0, X_{2i} = 1) - E(Y_i|X_{1i} = 0, X_{2i} = 0)] * X_{2i}\}) * X_{1i} * X_{2i}
\end{aligned}$$

Therefore, the CEF $E(Y_i|X_{1i}, X_{2i})$ can be written as a linear function of $X = [1, X_{1i}, X_{2i}, X_{1i} * X_{2i}]$.

1.5

We can set $Z_i = X_{1i}^2$ and run the linear regression $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 Z_i$ and this is just a multivariate linear regression that we can use to estimate the CEF.

1.6

Let $\beta_1 = E(X_i X_i^T)^{-1} E(X_i Y_i)$, $\beta_2 = E(X_i X_i^T)^{-1} E(X_i D_i)$

$$\begin{aligned}
Cov(Y_i, D_i^{\perp X_i}) &= Cov(X_i\beta^T + e_i, D_i^{\perp X_i}) \\
&= Cov(X_i\beta^T, D_i^{\perp X_i}) + Cov(e_i, D_i^{\perp X_i}) \\
&= Cov(X_i\beta_r + D_i\tau_r, D_i^{\perp X_i}) \\
&= Cov(X_i\beta_r, D_i^{\perp X_i}) + Cov(D_i\tau_r, D_i^{\perp X_i}) \\
&= Cov(D_i\tau_r, D_i^{\perp X_i}) \\
&= \tau_r Cov(D_i^{\perp X_i}, D_i^{\perp X_i}) \\
&= \tau_r Var(D_i^{\perp X_i})
\end{aligned}$$

By properties of residuals, $Cov(Y_i^{\perp X_i}, D_i^{\perp X_i}) = Cov(Y_i, D_i^{\perp X_i})$. So, $\tau_r = \frac{Cov(Y_i^{\perp X_i}, D_i^{\perp X_i})}{Var(D_i^{\perp X_i})}$.

1.7

$$\begin{aligned}
\tau_r &= \frac{Cov(Y_i^{\perp X_i}, D_i^{\perp X_i})}{Var(D_i^{\perp X_i})} \\
&= \frac{Cov(\tau D_i^{\perp X_i} + \gamma Z_i^{\perp X_i} + e_i, D_i^{\perp X_i})}{Var(D_i^{\perp X_i})} \\
&= \frac{Cov(\tau D_i^{\perp X_i}, D_i^{\perp X_i})}{Var(D_i^{\perp X_i})} + \frac{Cov(\gamma Z_i^{\perp X_i}, D_i^{\perp X_i})}{Var(D_i^{\perp X_i})} \\
&= \tau + \gamma \frac{Cov(Z_i^{\perp X_i}, D_i^{\perp X_i})}{Var(D_i^{\perp X_i})} \\
&= \tau + \gamma\delta
\end{aligned}$$

where $\delta = \frac{Cov(Z_i^{\perp X_i}, D_i^{\perp X_i})}{Var(D_i^{\perp X_i})}$.

1.8

a With $V = X_1 + X_2$,

$$\begin{aligned} OLS(V|Z) &= Z^T \frac{Cov(V,Z)}{Var(Z)} \\ &= Z^T \frac{Cov(X_1,Z)}{Var(Z)} + Z^T \frac{Cov(X_2,Z)}{Var(Z)} \\ &= OLS(X_1|Z) + OLS(X_2|Z) \end{aligned}$$

Thus, $V^{\perp Z} = V - OLS(V|Z) = X_1 + X_2 - OLS(X_1|Z) - OLS(X_2|Z) = X_1^{\perp Z} + X_2^{\perp Z}$.

b

$$\begin{aligned} e^{\perp Z} &= e - OLS(e|Z) \\ &= e - Z^T \frac{Cov(e,Z)}{Var(Z)} \\ &= e \end{aligned}$$

c

$$\begin{aligned} Z^{\perp Z} &= Z - OLS(Z|Z) \\ &= Z - Z^T \frac{Cov(Z,Z)}{Var(Z)} \\ &= Z - Z \\ &= 0 \end{aligned}$$

Q2

2.1

```
rm(list = ls())
library(tidyverse)
```

a

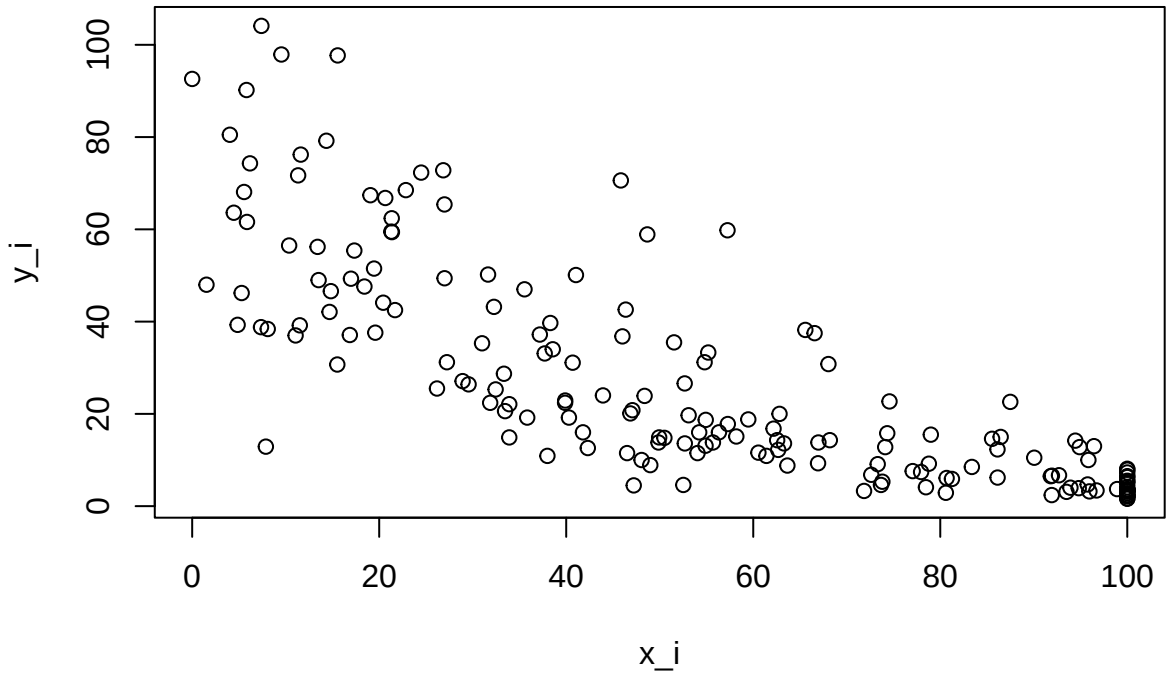
```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.5     v dplyr   1.0.7
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.0.2     v forcats 0.5.1
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
df = read.csv('qog_jan16.csv')

summary(df)
```

```
##     cname           wdi_mortinftot      epi_watsup     wdi_accelectr
##  Length:184         Min.   :  1.600   Min.   :  0.00   Min.   :  6.40
##  Class :character   1st Qu.:  6.775   1st Qu.: 27.18   1st Qu.: 56.33
```

```
##  Mode  :character   Median : 16.000   Median : 53.56   Median : 97.70
##                     Mean   : 26.337   Mean   : 55.29   Mean   : 77.63
##                     3rd Qu.: 39.225   3rd Qu.: 86.21   3rd Qu.:100.00
##                     Max.   :104.100   Max.   :100.00   Max.   :100.00
```

```
df$y_i <- df$wdi_mortinftot
df$x_i <- df$epi_watsup
df$z_i <- df$wdi_accelectr

attach(df)
#hist(x_i)
#hist(y_i)
plot(x_i, y_i)
```




```
detach(df)
```

There appears a negative correlation between infant mortality and access to clean water. Better the access to clean water, lower rates of infant mortality.

```
attach(df)
linear_model = lm(y_i ~ x_i, data = df)
summary(linear_model)
```

b

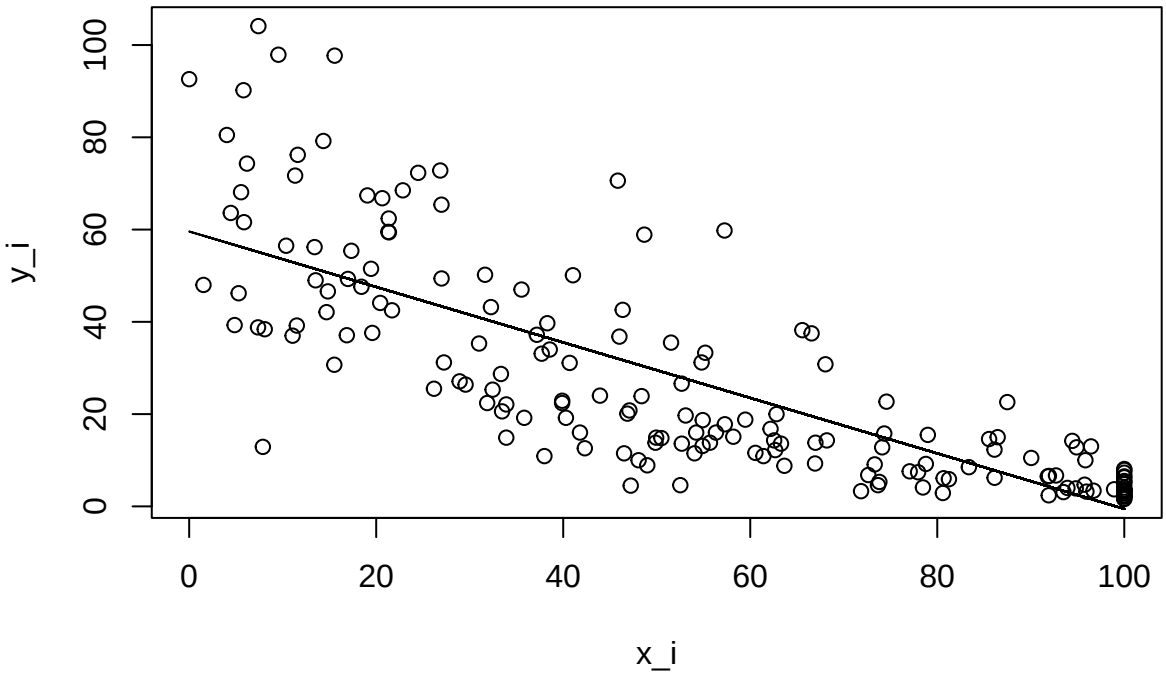
```
##
## Call:
## lm(formula = y_i ~ x_i, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -41.945 -10.277   0.429   6.294  48.966
##
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 59.58323    2.13802   27.87   <2e-16 ***
## x_i         -0.60126    0.03351  -17.94   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.47 on 182 degrees of freedom
## Multiple R-squared:  0.6388, Adjusted R-squared:  0.6369
## F-statistic: 321.9 on 1 and 182 DF,  p-value: < 2.2e-16
```

```
coef(linear_model)
```

```
## (Intercept)         x_i
##   59.583229   -0.601256
```

```
predict_linear1 <- predict(linear_model, newdata = data.frame(x_i =  x_i))

plot(x_i, y_i)
lines(x_i, predict_linear1)
```




```
detach(df)
```

The estimate for the slope is -0.60126, for the intercept is 59.58323. For countries with no access to clean water, the infant mortality rate is expected to be 59.58323%. Also, one percentage point increase in access to clean water is associated with 0.60126 percentage point decrease in infant mortality rate.

```
set.seed(42)
B = 10000
attach(df)
intercept_linear = rep(NA,B)
slope_linear = rep(NA,B)
for (i in 1:B){
```

```
  index = sample(c(1:184), size = 184, replace = T)
  x.boot = x_i[index]
  y.boot = y_i[index]
  model = lm(y.boot ~ x.boot)
  intercept_linear[i] =coef(model)[1]
  slope_linear[i] =coef(model)[2]
}
detach(df)

#summary(slope_linear)
quantile(intercept_linear, c(0.025,0.975))
```

c

```
##     2.5%    97.5%
## 53.98365 65.03107
```

```
quantile(slope_linear, c(0.025,0.975))
```

```
##       2.5%      97.5%
## -0.6703358 -0.5319351
```

The 95% confidence interval for intercept is [53.98365, 65.03107]; for slope is [-0.6703358, -0.5319351].

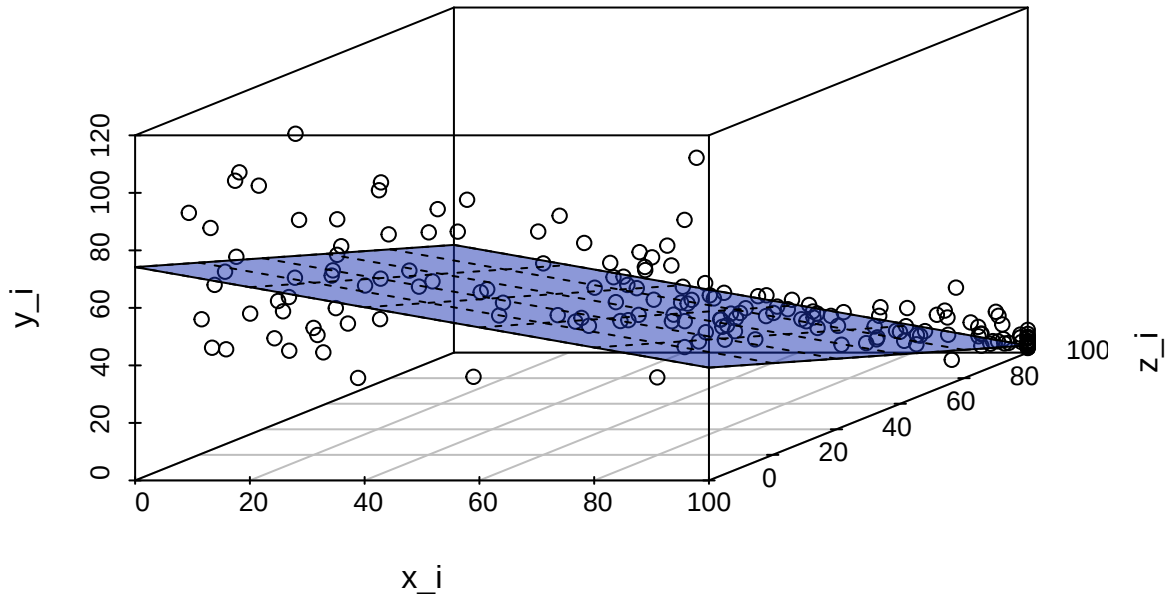
```
attach(df)

linear_model2 <- lm(y_i ~ x_i + z_i)

summary(linear_model2)
```

d

```
##
## Call:
## lm(formula = y_i ~ x_i + z_i)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -31.155  -5.976  -0.545   3.977  46.824
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 74.18601    2.46742  30.066  < 2e-16 ***
## x_i         -0.34962    0.04051  -8.630 3.08e-15 ***
## z_i         -0.36735    0.04239  -8.666 2.47e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.2 on 181 degrees of freedom
## Multiple R-squared:  0.7448, Adjusted R-squared:  0.7419
## F-statistic: 264.1 on 2 and 181 DF,  p-value: < 2.2e-16
```

```
library("scatterplot3d")
s3d <- scatterplot3d(x= x_i, y=z_i, z=y_i)
s3d$plane3d(linear_model2, draw_polygon = TRUE, draw_lines = TRUE,
```

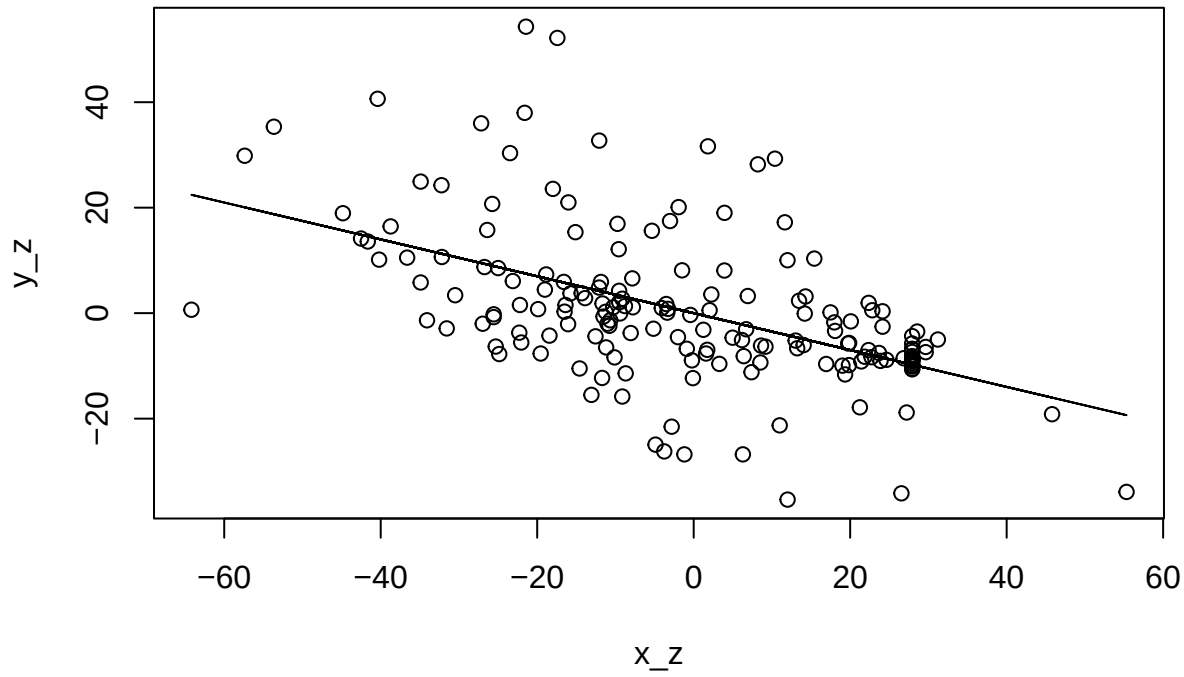
```
            polygon_args = list(col = rgb(.1, .2, .7, .5)))
```




```
detach(df)
```

Previously, the coefficient for x_i is -0.60126, but now it is -0.34962. It is not the same as before and the magnitude is lower. Its interpretation is that, if z_i stays the same, one percentage point increase in access to clean water is associated with 0.34962 percentage point decrease in infant mortality rate.

```
attach(df)

linear_model3 <- lm(x_i ~ z_i)
x_z <- resid(linear_model3)
#summary(linear_model3)
#summary(x_z)

linear_model4 <- lm(y_i ~ z_i)
y_z <- resid(linear_model4)

plot(x_z, y_z)
linear_model5 <- lm(y_z ~ x_z)
predict_linear5 <- predict(linear_model5, newdata = data.frame(x_z =  x_z))
lines(x_z,predict_linear5)
```



e

```
# check
coef(linear_model5)[2]
```

```
##        x_z
## -0.3496213
```

```
detach(df)
```

Yes, the coefficient is identical to the previous regression coefficient of X_i.

```
attach(df)
linear_model6 <- lm(z_i ~ x_i)
coef(linear_model6)[2]
```

f

```
##       x_i
## 0.6850051
```

```
c(coef(linear_model)[2] - coef(linear_model6)[2] *coef(linear_model2)[3], coef(linear_model2)[2])
```

```
##        x_i        x_i
## -0.3496213 -0.3496213
```

```
detach(df)
```

The estimated value is 0.6850051. This means that one percentage point increase in access to clean water is associated with a 0.6850051 percentage point increase in access to electricity.

Both ways, the result is -0.3496213.

2.2

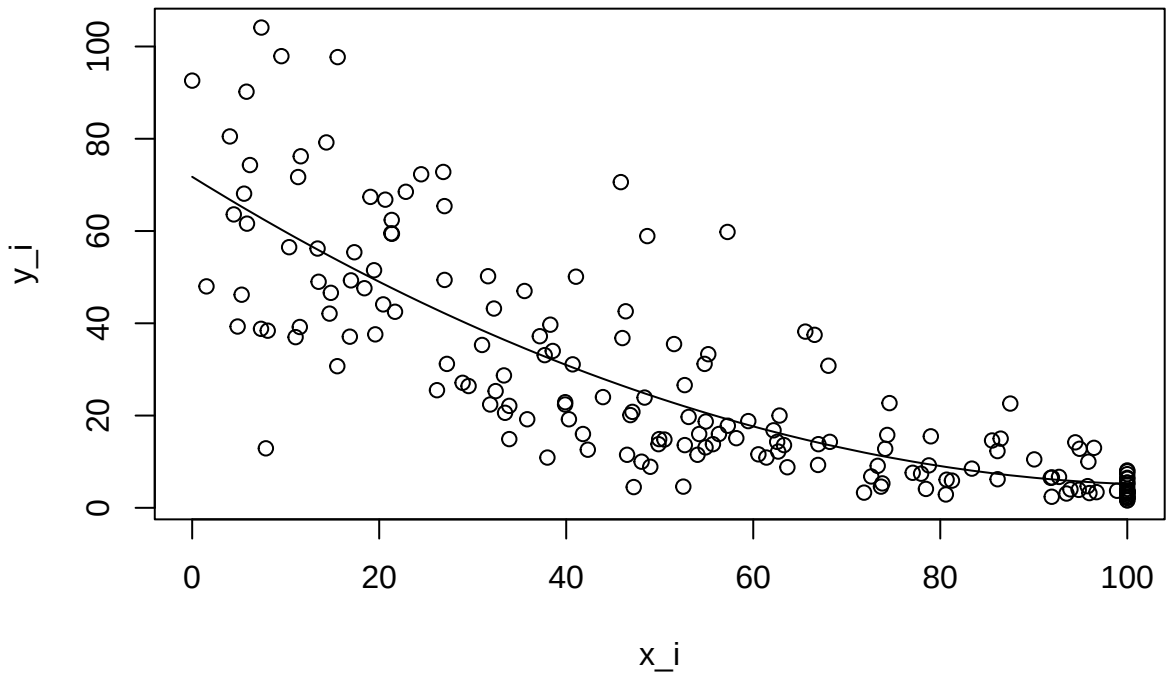
```
attach(df)
quadratic_model_a <- lm(y_i ~ x_i + I(x_i^2))
```

```
summary(quadratic_model_a)
```

a

```
##
## Call:
## lm(formula = y_i ~ x_i + I(x_i^2))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -49.324  -6.074  -1.882   4.659  44.057
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 71.750805   3.163359  22.682  < 2e-16 ***
## x_i         -1.255507   0.135049  -9.297  < 2e-16 ***
## I(x_i^2)     0.005898   0.001184   4.982 1.46e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.61 on 181 degrees of freedom
## Multiple R-squared:  0.6824, Adjusted R-squared:  0.6789
## F-statistic: 194.5 on 2 and 181 DF,  p-value: < 2.2e-16
```

```
predict_quadratic1 <- predict(quadratic_model_a, newdata = data.frame(x_i = sort(x_i)))

plot(x_i, y_i)
lines(sort(x_i),predict_quadratic1)
```




```
#reorder <- order(x_i)
#lines(x_i[reorder], predict_quadratic1[reorder])

detach(df)
```

This seems to fit the data better. The intercept is 71.750805, estimator for x_i is -1.255507, for x-squared is 0.005898. In this case, we cannot simply interpret the coefficient for the quadratic term. It makes no practical sense when we interpret it as the "one unit change in the square of access to clean water is associated with a 0.005898 increase in infant mortality." The trend is obvious that infant mortality is negatively correlated with access to clean water. Moreover, the quadratic term of a percentage is not intuitive to understand.

b

$$\begin{aligned} APD_{yx} &= E(\frac{dE(Y_i|X_i)}{dX_i}) \\ &= E(\frac{d(\beta_{y1.xx^2} + \beta_{yx.1x^2}X_i + \beta_{yx^2.1x}X_i^2)}{dX_i}) \\ &= E(\beta_{yx.1x^2} + 2\beta_{yx^2.1x}X_i) \\ &= \beta_{yx.1x^2} + 2\beta_{yx^2.1x}E(X_i) \end{aligned}$$

```
set.seed(42)
B = 10000
attach(df)
intercept_quad = rep(NA,B)
slope_quad = rep(NA,B)
apd = rep(NA,B)
for (i in 1:B){
  index = sample(c(1:184), size = 184, replace = T)
  x.boot = x_i[index]
  y.boot = y_i[index]
  model <- lm(y.boot ~ x.boot + I(x.boot^2))
  intercept_quad[i] =coef(model)[2]
  slope_quad[i] =coef(model)[3]
  apd[i] = intercept_quad[i] + 2*slope_quad[i]*mean(x.boot)
}
detach(df)

quantile(apd, c(0.025,0.975))
```

c

```
##       2.5%      97.5%
## -0.6929937 -0.5219206
```

```
summary(apd)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -0.7804 -0.6312 -0.6017 -0.6032 -0.5732 -0.4442
```

The confidence interval is [-0.6929937, -0.5219206]. The bootstrap estimate is quite close to the coefficient of the simple linear model.

```
attach(df)
h = 0.0001
predict_quadratic2 <- predict(quadratic_model_a, newdata = data.frame(x_i = sort(x_i+h)))
predict_quadratic3 <- predict(quadratic_model_a, newdata = data.frame(x_i = sort(x_i-h)))

res = mean((predict_quadratic2-predict_quadratic3)/(2*h))
res
```

d

```
## [1] -0.6032968
```

```
detach(df)
```

This is also a pretty good estimate for -0.601256.

```
attach(df)
quadratic_model_e <- lm(y_i ~ x_i + I(x_i^2) + z_i + I(z_i^2) )
summary(quadratic_model_e)
```

e

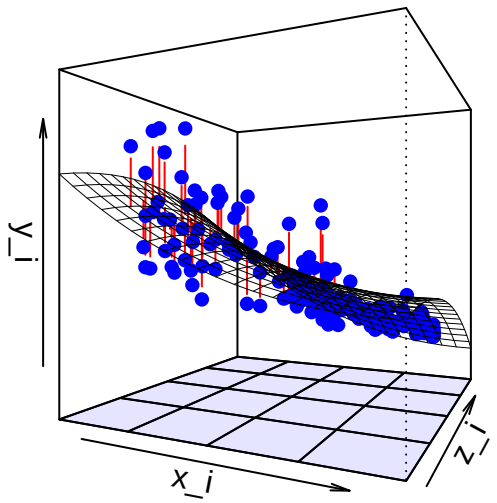
```
##
## Call:
## lm(formula = y_i ~ x_i + I(x_i^2) + z_i + I(z_i^2))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -28.249  -4.427  -1.044   4.303  48.355
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 68.113336   4.538473  15.008  < 2e-16 ***
## x_i         -0.706417   0.147836  -4.778 3.67e-06 ***
## I(x_i^2)     0.003103   0.001176   2.640  0.00903 **
## z_i          0.157026   0.186522   0.842  0.40099
## I(z_i^2)    -0.003961   0.001482  -2.673  0.00821 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.89 on 179 degrees of freedom
## Multiple R-squared:  0.7602, Adjusted R-squared:  0.7548
## F-statistic: 141.8 on 4 and 179 DF,  p-value: < 2.2e-16
```

```
predict_quadratic5 <- predict(quadratic_model_a, newdata = data.frame(x_i = sort(x_i), z_i = sort(z_i))

library(rockchalk)
```

```
##
## Attaching package: 'rockchalk'
```

```
## The following object is masked from 'package:dplyr':
##
##     summarize
```

```
plotPlane(quadratic_model_e, "x_i", "z_i", pch=16, col=rgb(0,0,1,0.1), drawArrows=TRUE, alength=0,
          acol="red", alty=1,alwd=1, theta=25, phi=0)
```




```
detach(df)
```

The coefficient for x_i is -0.706417, for x_i^2 it is 0.003103, for z_i it is 0.157026, for z_i^2 it is -0.003961.

```
attach(df)
h = 0.0001
predict_quadratic6 <- predict(quadratic_model_e, newdata = data.frame(x_i = sort(x_i+h), z_i = sort(z_i
predict_quadratic7 <- predict(quadratic_model_e, newdata = data.frame(x_i = sort(x_i-h), z_i = sort(z_i

res = mean((predict_quadratic6-predict_quadratic7)/(2*h))
res
```

f

```
## [1] -0.3632335
```

```
detach(df)
```

This is a pretty good estyimate of the previous regression coefficient -0.36735.

```
set.seed(42)
B = 10000
attach(df)
intercept_quad = rep(NA,B)
slope_quad = rep(NA,B)
apd1 = rep(NA,B)
for (i in 1:B){
  index = sample(c(1:184), size = 184, replace = T)
  x.boot = x_i[index]
  y.boot = y_i[index]
  z.boot = z_i[index]
  model <- lm(y.boot ~ x.boot + I(x.boot^2) + z.boot + I(z.boot^2))
  intercept_quad[i] =coef(model)[2]
  slope_quad[i] =coef(model)[3]
  apd1[i] = intercept_quad[i] + 2*slope_quad[i]*mean(x.boot)
}
detach(df)
```

```
quantile(apd1, c(0.025,0.975))
```

g

```
##       2.5%      97.5%
## -0.4616677 -0.2679650
```

```
summary(apd1)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -0.5543 -0.3970 -0.3628 -0.3634 -0.3299 -0.1844
```

A confidence interval is [-0.4616677, -0.2679650].

2.3

Although there is an obvious correlation and potenrially some causality, it is too hasty to make causal inference. As we have observed, the addition of the access to electricity variable has reduced the magnitude of access to water by half. It's possible that another variable causes both low access to clean water and high infant mortality rates simultaneously. For example, poor infrastructure or pollution can both cause the water to be unclean, at the same time affecting infant's health through other channels as well, for example, poor medical healthcare or worse food quality, etc.