p8130_hw4_yl5508

Yifei LIU (yl5508)

2023/11/15

```
library(tidyverse)
library(readxl)
library(BSDA)
```

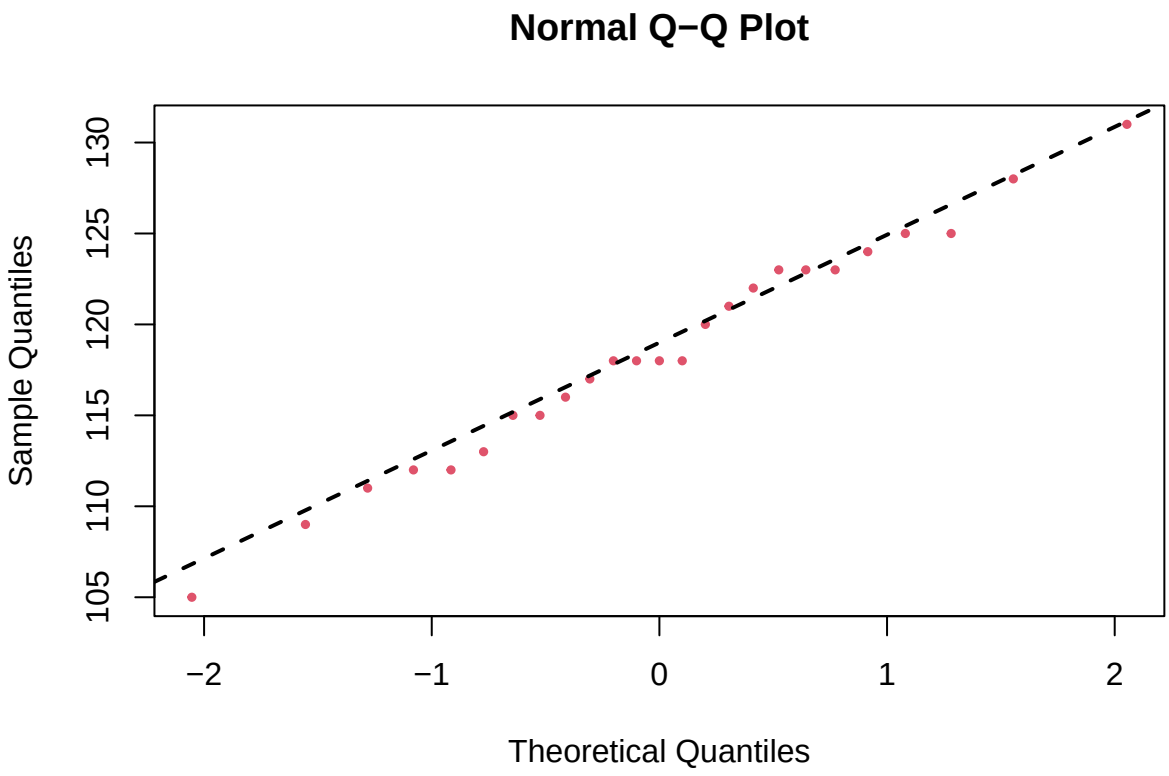
```
## Warning:   'BSDA' R 4.3.2
```

Problem 1

(a)

```
bsl = c(125, 123, 117, 123, 115, 112, 128, 118, 124, 111, 116, 109, 125, 120, 113, 123, 112, 118, 121, 
#bsl_data = tibble(bsl_value = bsl)

#check normality
#hist
bsl |>
  hist(xlab = "Blood Sugar Level", freq = T, col = 2)
```


```
#bsl_data |>
  #ggplot(aes(x = bsl_value)) +
  #geom_histogram(bins = 8, fill = "lightblue", color = "Black")

#Q-Q plot
qqnorm(bsl, col = 2, pch = 19, cex = 0.5)
qqline(bsl, col = 1, lwd = 2, lty = 2)
```




```
#directly added to existed plot

#Shapiro-Wilk test
res = shapiro.test(bsl)
norm_test = tibble(
  p_value = res$p.value,
  statistic = res$statistic
)
norm_test |>
  knitr::kable(digits = 5)
```

p_value	statistic
0.99294	0.98917

```
#sign test
res = SIGN.test(bsl, md = 120, alternative = "less", conf.level = 0.95)
mbs_sign_tidy = tibble(
  p_value = res$p.value,
  statistic = res$statistic
)
mbs_sign_tidy |>
  knitr::kable(digits = 5)
```

p_value	statistic
0.27063	10

Interpretation: Since p-value = 0.27 > 0.05 (sign test), we would fail to reject the null hypothesis at the =0.05 level, meaning we have no evidence that median blood sugar readings was less than 120 in the population.

(b)

Normal-Approximation: n* = 25-1 16

$H_0: \ median(bsl) - 120 = 0$ vs $H_1: \ median(bsl) - 120 < 0$

```
#Wilcoxon signed-rank test
wil_test = tibble(
  diff_abs = abs(bsl - 120),
  diff = bsl - 120
) |>
  mutate(
    pos_d = ifelse(diff>0, 1, 0),
    neg_d = ifelse(diff<0, 1, 0)
  ) |>
  arrange(- diff_abs) |>
  select(- diff) |>
  mutate(rank = ifelse(diff_abs > 0, rank(diff_abs[diff_abs > 0]), 0))

head(wil_test, 5)
```

```
## # A tibble: 5 x 4
##   diff_abs pos_d neg_d  rank
##      <dbl> <dbl> <dbl> <dbl>
## 1       15     0     1  24
## 2       11     0     1  22.5
## 3       11     1     0  22.5
## 4        9     0     1  21
## 5        8     0     1  19
```

```
#T+
T_sum = wil_test |>
  group_by(pos_d) |>
  summarise(sum_rank = sum(rank))

#T stat
T_pos = T_sum |> filter(pos_d == 1) |> pull(sum_rank)
T_stat = (abs(112.5 - 24*(24+1)/4)-1/2)/(sqrt(24*(24+1)*(24*2+1)/24-((2^3-2)*2+(4^3-4)*2)/48))
T_stat
```

```
## [1] 1.058331
```

```
#test statistic
z_5 = qnorm(0.05)
z_5
```

```
## [1] -1.644854
```

```
#p_value
1 - pnorm(T_stat)
```

```
## [1] 0.1449522
```

Comment: Using a = 0.05 significance level, T_stat = 1.06 Z_0.05 = -1.64 (or: p-value = 0.14 > 0.05), we would fail to reject H0 and conclude that there is no evidence that median blood sugar readings was less than 120 in the population.

```
#Wilcoxon signed-rank test using R code
res = wilcox.test(bsl, mu = 120, alternative = "less", conf.level = 0.95)
```

```
## Warning in wilcox.test.default(bsl, mu = 120, alternative = "less", conf.level
## = 0.95):      p
```

```
## Warning in wilcox.test.default(bsl, mu = 120, alternative = "less", conf.level
## = 0.95):  0    p
```

```
mbs_wil_tidy = tibble(
  p_value = res$p.value,
  statistic = res$statistic
)
mbs_wil_tidy |>
  knitr::kable(digits = 5)
```

p_value	statistic
0.14466	112.5

Interpretation: Since p-value = 0.14 > 0.05 (Wilcoxon signed-rank test), we would fail to reject the null hypothesis at the =0.05 level, meaning we have no evidence that median blood sugar readings was less than 120 in the population.

Problem 2

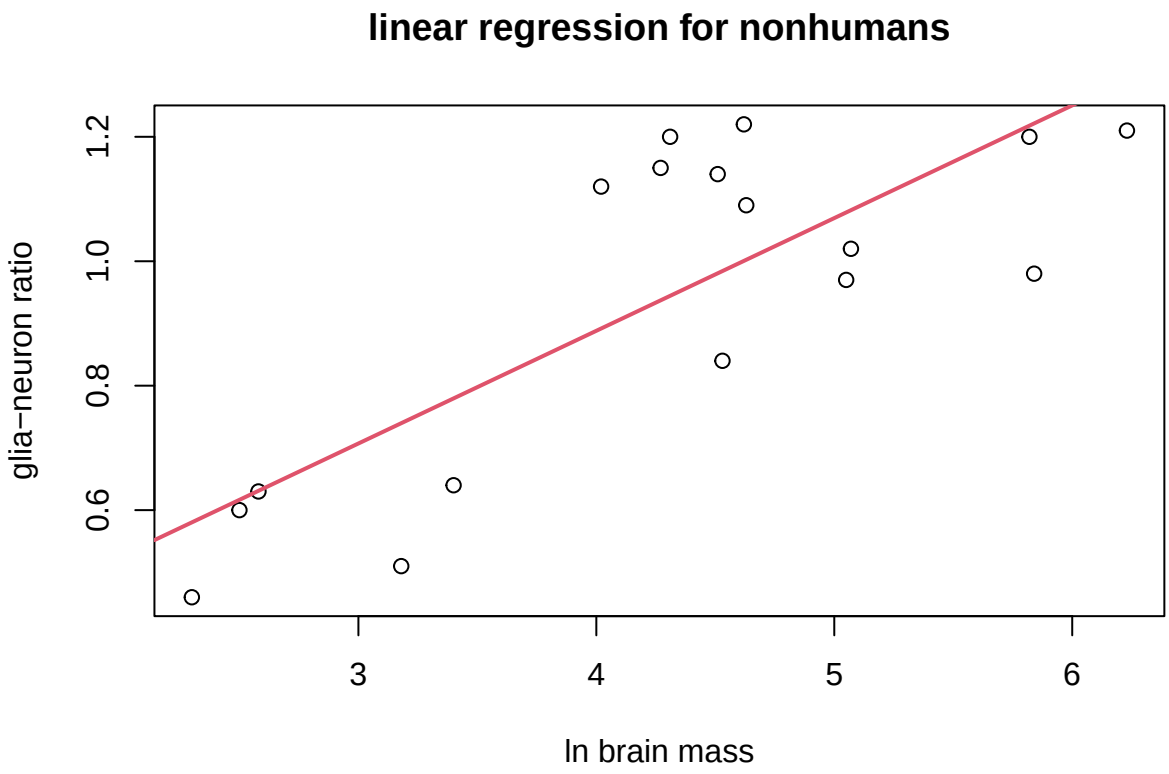
(a)

```
#loading nonhuman data
brain_data =
  read_excel("~/Biostat methods/p8130_Biostat Methods_hw/data/Brain.xlsx") |>
  janitor::clean_names()

brain_nonhuman =
  brain_data |>
  filter(species != "Homo sapiens") |>
  mutate(brain_mass_g = as.numeric(brain_mass_g))
```

```r
#generating a regression model
model_nh <- lm(data = brain_nonhuman, glia_neuron_ratio ~ ln_brain_mass)
summary(model_nh)
```

```
##
## Call:
## lm(formula = glia_neuron_ratio ~ ln_brain_mass, data = brain_nonhuman)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.24150 -0.12030 -0.01787  0.15940  0.25563
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    0.16370    0.15987   1.024 0.322093
## ln_brain_mass  0.18113    0.03604   5.026 0.000151 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1699 on 15 degrees of freedom
## Multiple R-squared:  0.6274, Adjusted R-squared:  0.6025
## F-statistic: 25.26 on 1 and 15 DF,  p-value: 0.0001507
```

```r
#plot with regression model
plot(brain_nonhuman$ln_brain_mass,
     brain_nonhuman$glia_neuron_ratio,
     main = "linear regression for nonhumans",
     xlab = "ln brain mass", ylab = "glia-neuron ratio")
abline(model_nh, lwd = 2, col = 2)
```



(b)

```
predict_human =
  brain_data |>
  slice(1) |>
  select(ln_brain_mass)

predict_human |>
  mutate(predict_ratio = predict.lm(model_nh, predict_human)) |>
  knitr::kable(digits = 5)
```

ln_brain_mass	predict_ratio
7.22	1.47146

Comment: Given humans brain mass, the predicted glia-neuron ratio for humans is 1.47 according to the generated linear regression.

(c)

The first interval is confidence interval. It would be suitable for estimating the mean response for the overall population.
The second interval is prediction interval. It would be suitable when we predict the result of one specific individual.
So, for this case, confidence interval would be reasonable.

(d)

```
#confidence interval
predict_human |>
  bind_cols(predict.lm(model_nh, predict_human, interval = "predict", conf.level = 0.95)) |>
  rename(predict_ratio = fit, lower_bound = lwr, upper_bound = upr) |>
  knitr::kable(digits = 5)
```

ln_brain_mass	predict_ratio	lower_bound	upper_bound
7.22	1.47146	1.03605	1.90687

Comment: 95% confidence interval of human glia-neuron is (1.03, 1.91). The predicted value is 1.47. Humans have a higher glia-neuron ratio that nonhumans, so it would be deemed as an outlier from the regression model (prediction interval doesn't contain value of nonhumans).

(e)

Comment: The data point of humans is actually a outlier for those of nonhumans. It would interfere the generation of the correct regression model for nonhumans.

Problem 3

(a)

```
#loading heart disease data
hd_data =
  read_csv("~/Biostat methods/p8130_Biostat Methods_hw/data/HeartDisease.csv") |>
  janitor::clean_names() |>
  select(totalcost, e_rvisits, age, gender, complications, duration)
```

```
## Rows: 788 Columns: 10
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## dbl (10): id, totalcost, age, gender, interventions, drugs, ERvisits, compli...
##
## 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.
```

```
hd_data |>
  summary() |>
  knitr::kable(digits = 1)
```

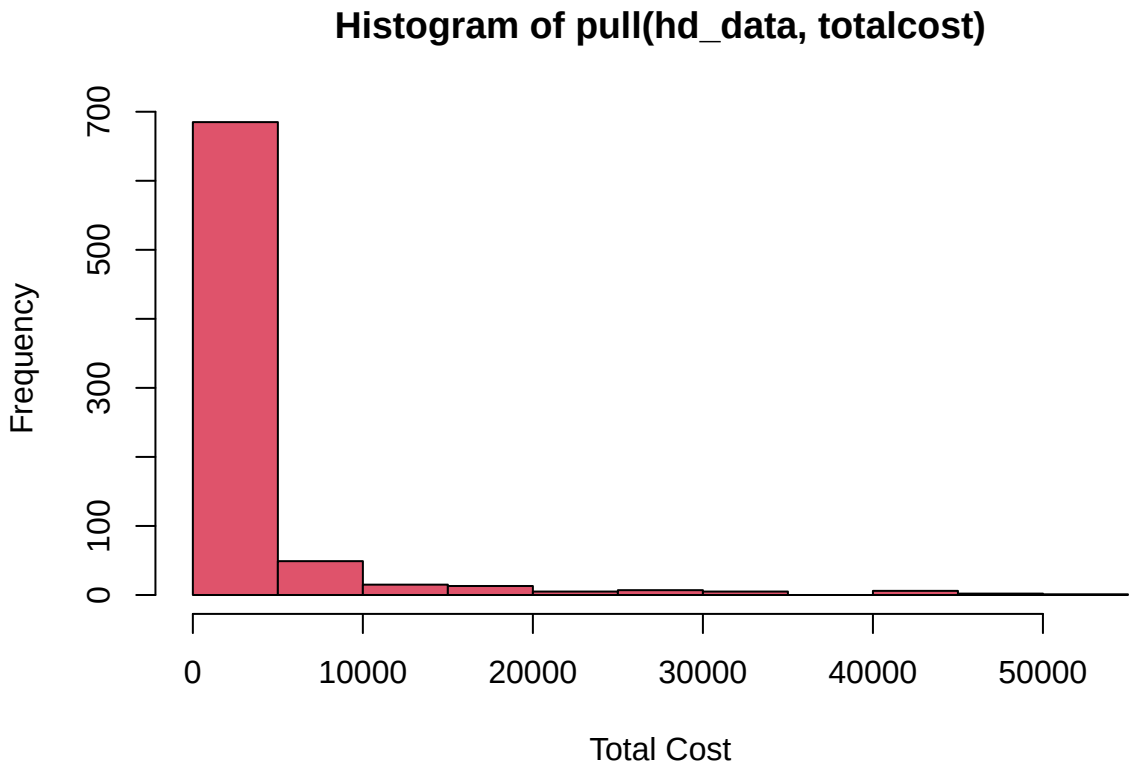
	totalcost	e_rvisits	age	gender	complications	duration
	Min. : 0.0	Min. : 0.000	Min. :24.00	Min. :0.0000	Min. :0.00000	Min. : 0.00
	1st Qu.: 161.1	1st Qu.: 2.000	1st Qu.:55.00	1st Qu.:0.0000	1st Qu.:0.00000	1st Qu.: 41.75
	Median : 507.2	Median : 3.000	Median :60.00	Median :0.0000	Median :0.00000	Median :165.50

	totalcost	e_rvisits	age	gender	complications	duration
	Mean : 2800.0	Mean : 3.425	Mean :58.72	Mean :0.2284	Mean :0.05711	Mean :164.03
	3rd Qu.: 1905.5	3rd Qu.: 5.000	3rd Qu.:64.00	3rd Qu.:0.0000	3rd Qu.:0.00000	3rd Qu.:281.00
	Max. :52664.9	Max. :20.000	Max. :70.00	Max. :1.0000	Max. :3.00000	Max. :372.00

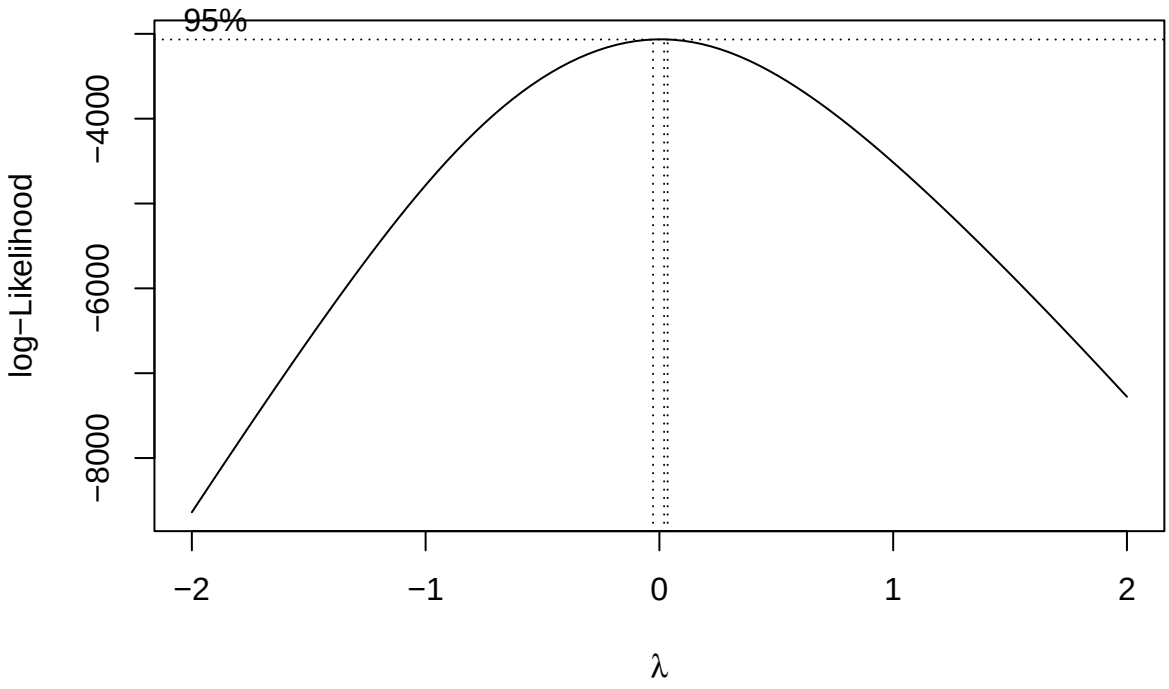
Comment:
(a) The main outcome of the data set is `totalcost` (continuous variable).
(b) The main predictor is `e_rvisits` (continuous variable).
(c) Important covariates are `age` (categorical variable), `gender` (categorical variable), `complications` (categorical variable), `duration` (continuous variable).

(b)

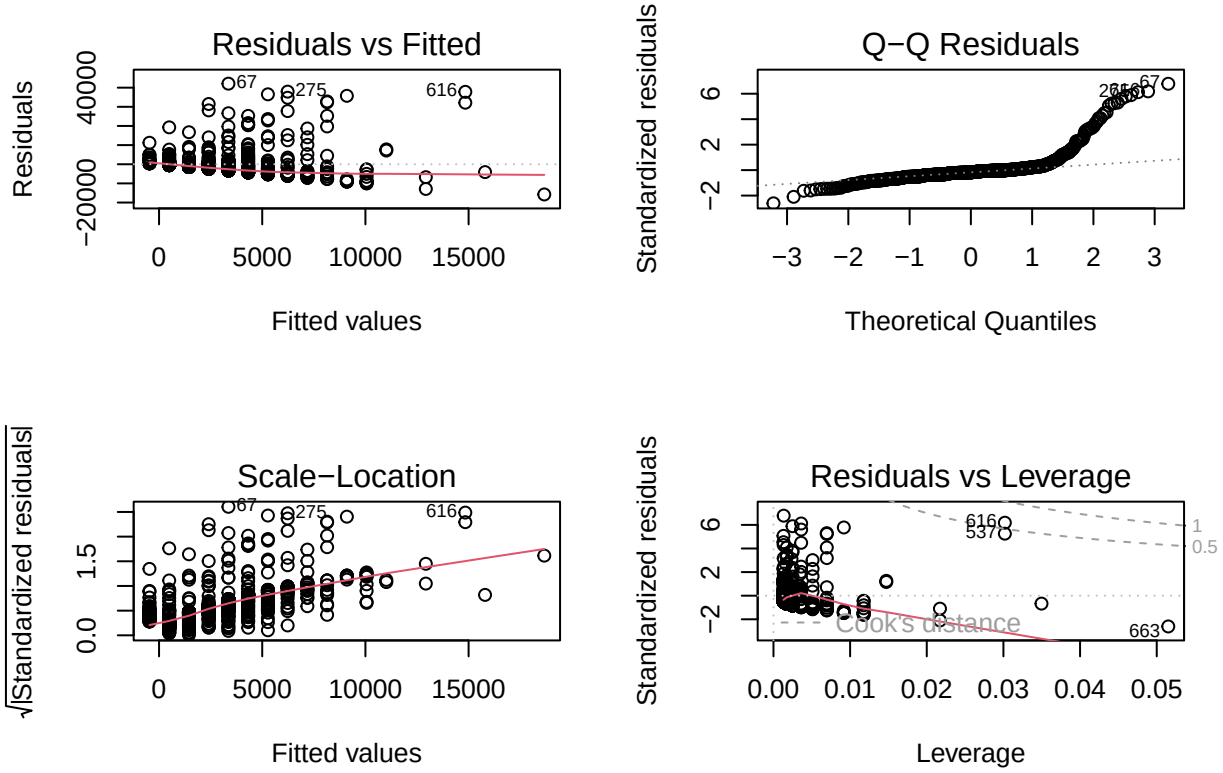
```
#check normality
#hist
hd_data |>
  pull(totalcost) |>
  hist(xlab = "Total Cost", freq = T, col = 2)
```



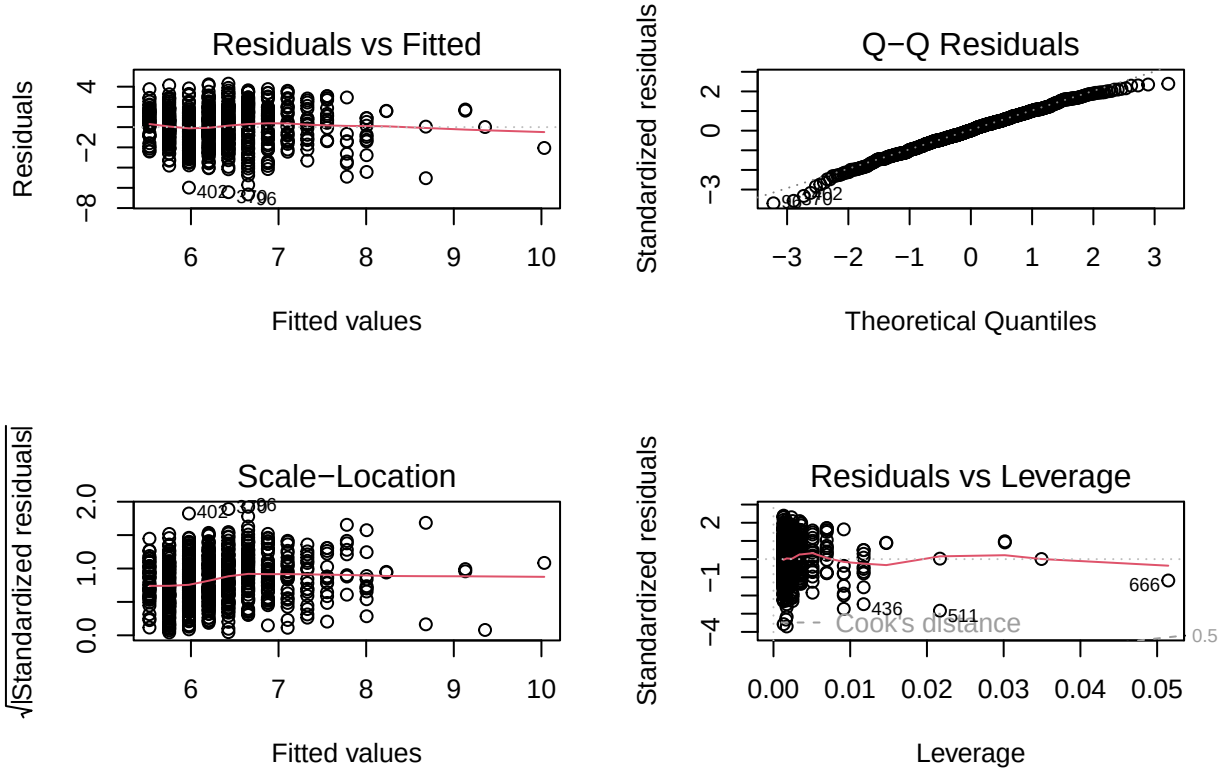

```
#try box-cox transformation
model_hd = lm(totalcost ~ e_rvisits, data = hd_data |> filter(totalcost != 0))
MASS::boxcox(model_hd)
```



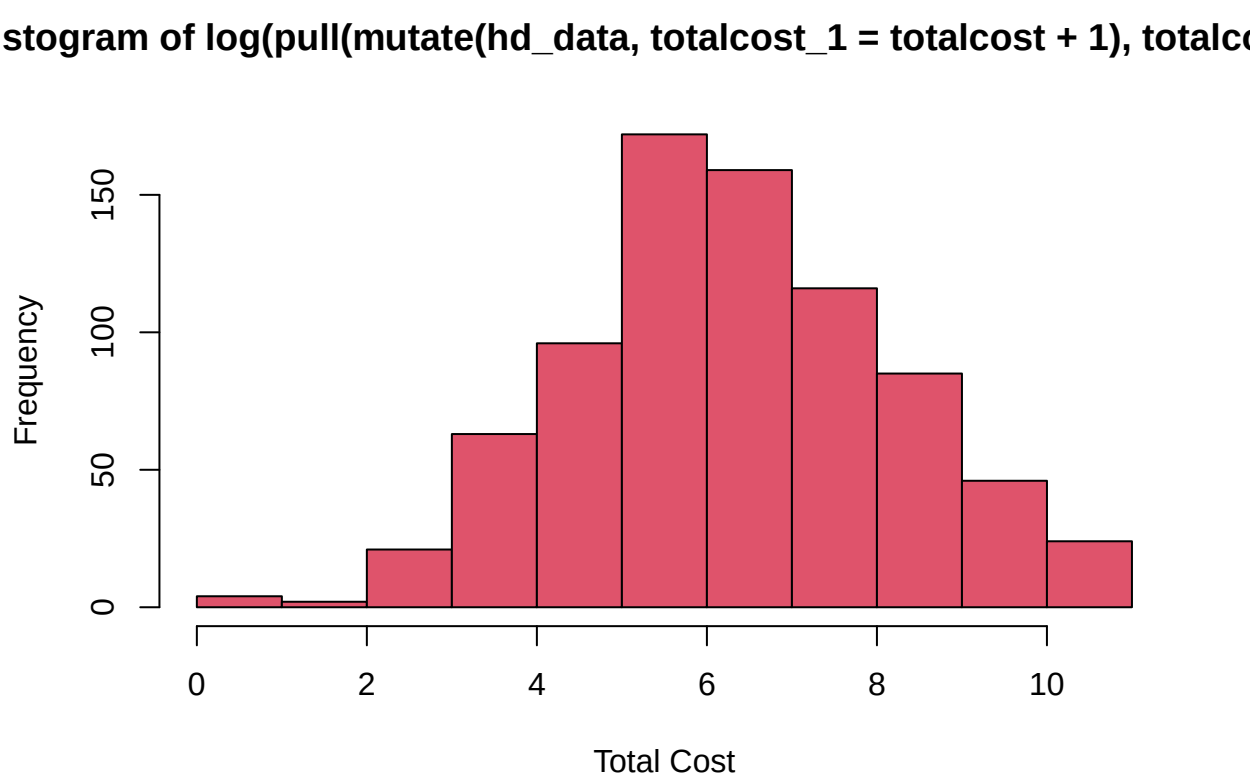

```
par(mfrow = c(2, 2))
plot(model_hd)
```



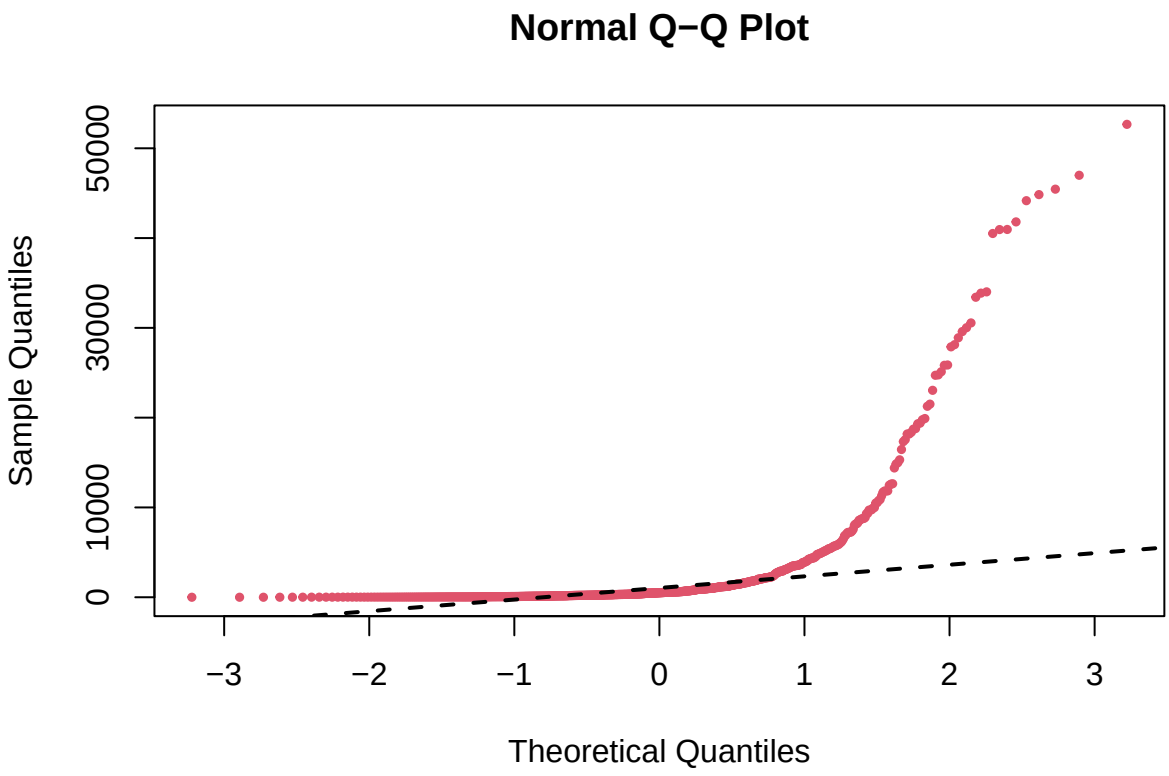

```r
#try log transformation
model_hd_log = lm(log(totalcost + 1) ~ e_rvisits, data = hd_data)

par(mfrow = c(2, 2))
plot(model_hd_log)
```



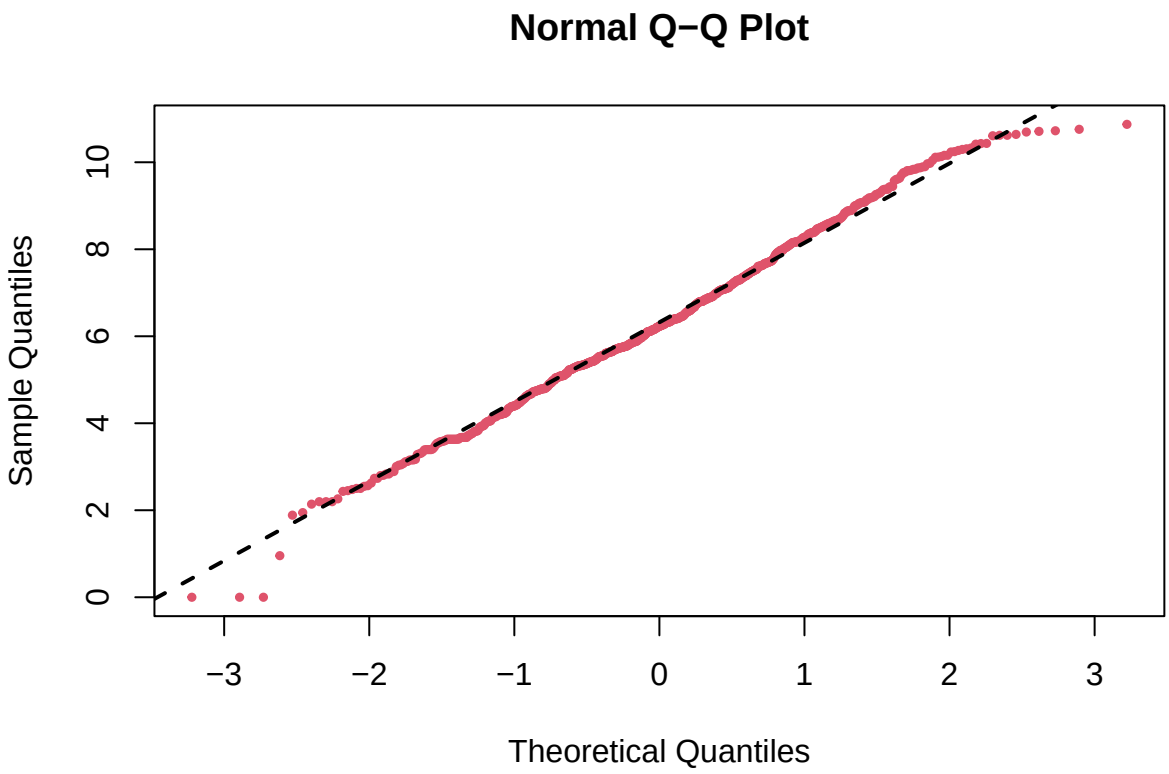

```r
#transformed hist
hd_data |>
  mutate(totalcost_1 = totalcost + 1) |>
  pull(totalcost_1) |>
  log() |>
  hist(xlab = "Total Cost", freq = T, col = 2)
```




```
#Q-Q plot
#untransformed
qqnorm(hd_data |> pull(totalcost), col = 2, pch = 19, cex = 0.5)
qqline(hd_data |> pull(totalcost), col = 1, lwd = 2, lty = 2)
```




```
#log transformed
qqnorm(hd_data |> mutate(totalcost_1 = totalcost + 1) |> pull(totalcost_1) |> log(), col = 2, pch = 19,
qqline(hd_data |> mutate(totalcost_1 = totalcost + 1) |> pull(totalcost_1) |> log(), col = 1, lwd = 2, l
```



Comment: Use log(x+1) function to transform the variable `totalcost` and turn it into a nice bell-shape distribution.

(c)

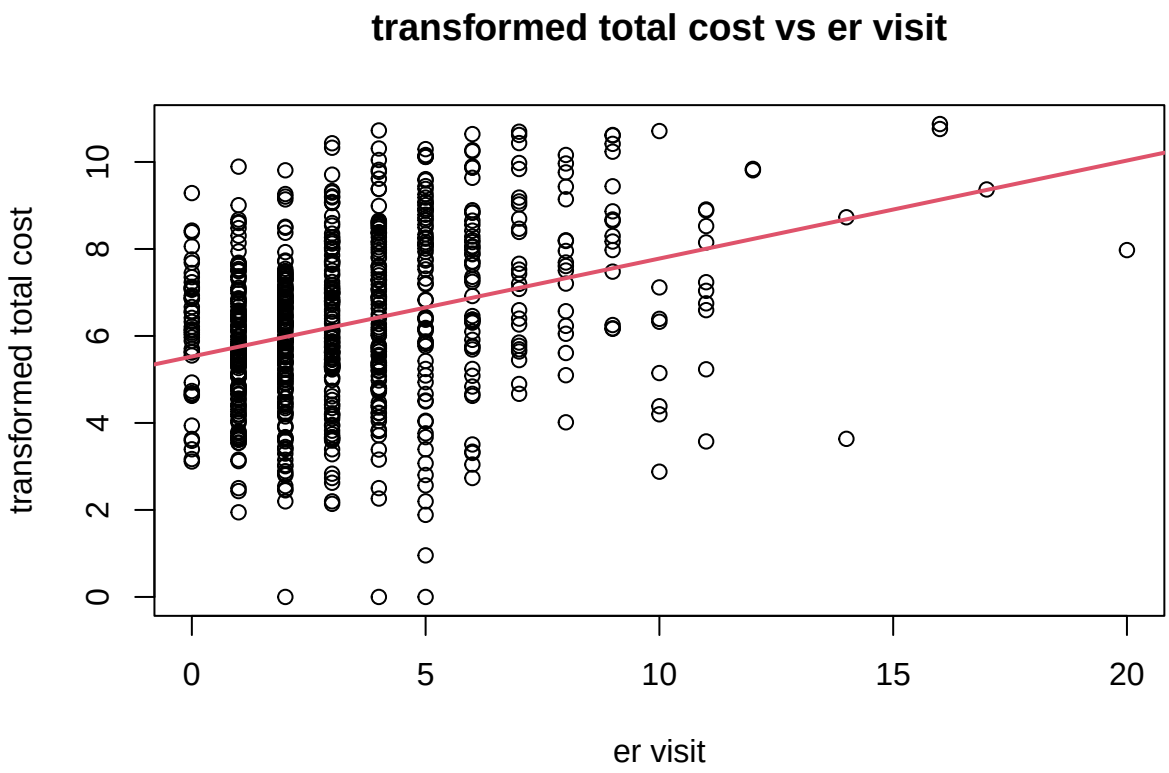
```
hd_data_tidy =
  hd_data |>
  mutate(comp_bin = ifelse(complications == 0, 0, 1))
hd_data_tidy |> head(5)
```

```
## # A tibble: 5 x 7
##   totalcost e_rvisits   age gender complications duration comp_bin
##       <dbl>     <dbl> <dbl>  <dbl>         <dbl>    <dbl>    <dbl>
## 1      179.         4    63      0             0      300        0
## 2      319          6    59      0             0      120        0
## 3     9311.         2    62      0             0      353        0
## 4      281.         7    60      1             0      332        0
## 5    18727.         7    55      0             0       18        0
```

(d)

```
model_hd_log = lm(log(totalcost + 1) ~ e_rvisits, data = hd_data_tidy)

plot(log(totalcost + 1) ~ e_rvisits, data = hd_data,
     main = "transformed total cost vs er visit",
```

```
     xlab = "er visit", ylab = "transformed total cost")

abline(model_hd_log, lwd = 2, col = 2)
```




```
summary(model_hd_log)
```

```
##
## Call:
## lm(formula = log(totalcost + 1) ~ e_rvisits, data = hd_data_tidy)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.6532 -1.1230  0.0309  1.2797  4.2964
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.52674    0.10510  52.584   <2e-16 ***
## e_rvisits    0.22529    0.02432   9.264   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.799 on 786 degrees of freedom
## Multiple R-squared:  0.09844,    Adjusted R-squared:  0.09729
## F-statistic: 85.82 on 1 and 786 DF,  p-value: < 2.2e-16
```

Comment: For adding 1 unit to emergency room visit, the log of (total cost plus one) would increase by 0.23 unit. The p_value is less than 0.05, so er visit has a significant effect on the log of (total cost plus one).

(e)

(i)

```
regmulti_hd = lm(log(totalcost + 1) ~ e_rvisits + comp_bin, data = hd_data_tidy)
summary(regmulti_hd)
```

```
##
## Call:
## lm(formula = log(totalcost + 1) ~ e_rvisits + comp_bin, data = hd_data_tidy)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.5249 -1.0769 -0.0074  1.1847  4.4024
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.51020    0.10279  53.606  < 2e-16 ***
## e_rvisits    0.20295    0.02405   8.437  < 2e-16 ***
## comp_bin     1.70573    0.27915   6.111 1.56e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.759 on 785 degrees of freedom
## Multiple R-squared:  0.1394, Adjusted R-squared:  0.1372
## F-statistic: 63.57 on 2 and 785 DF,  p-value: < 2.2e-16
```

Comment: There are significant effects of variables `er_visit` and `comp_bin` on `log(totalcost + 1)`, since p_value is less than 0.05.

(ii)

```
anova(regmulti_hd)
```

```
## Analysis of Variance Table
##
## Response: log(totalcost + 1)
##            Df  Sum Sq Mean Sq F value    Pr(>F)
## e_rvisits   1  277.87 277.870  89.792 < 2.2e-16 ***
## comp_bin    1  115.55 115.549  37.339 1.563e-09 ***
## Residuals 785 2429.27   3.095
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

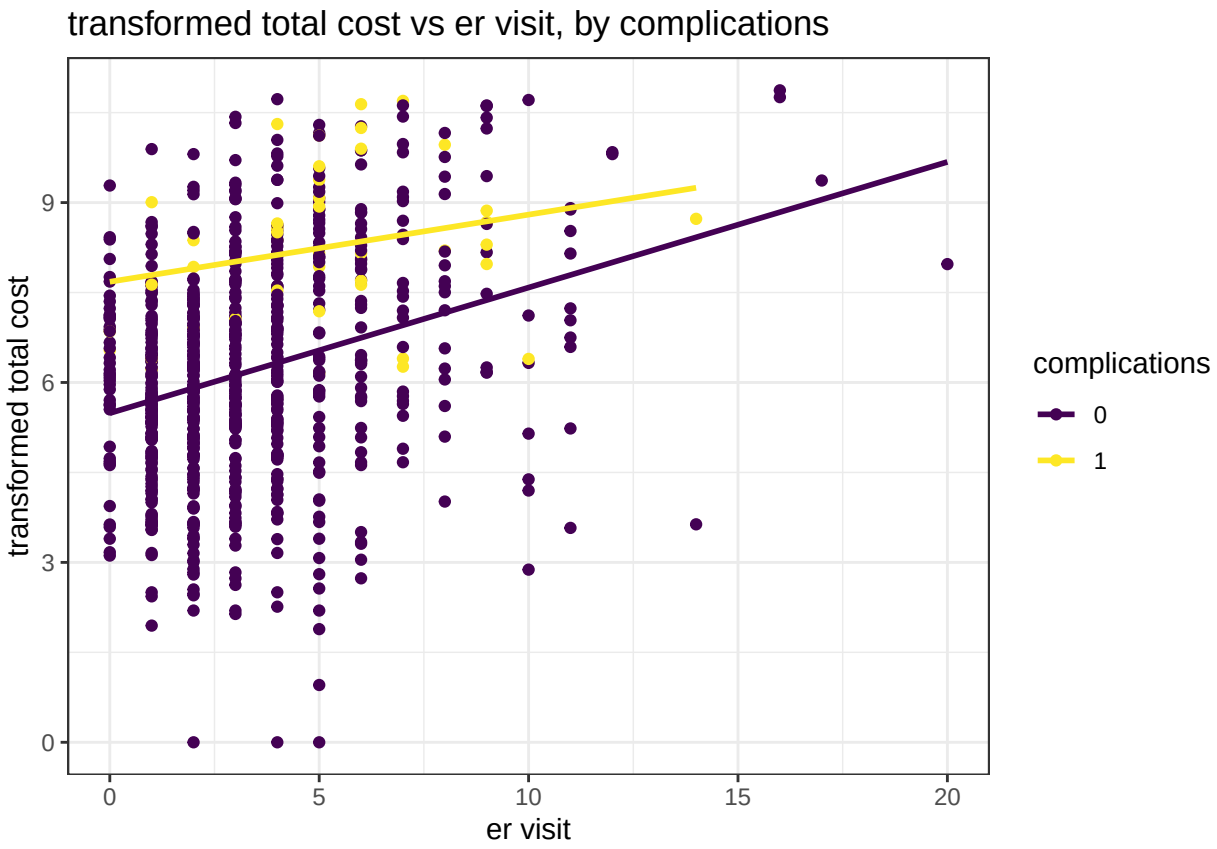
```
regmulti_hd_interact = lm(log(totalcost + 1) ~ e_rvisits * comp_bin, data = hd_data_tidy)
summary(regmulti_hd_interact)
```

```
##
```

```
## Call:
## lm(formula = log(totalcost + 1) ~ e_rvisits * comp_bin, data = hd_data_tidy)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -6.536 -1.083  0.004  1.200  4.398
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)          5.48849    0.10500  52.271  < 2e-16 ***
## e_rvisits            0.20947    0.02490   8.412  < 2e-16 ***
## comp_bin             2.19096    0.55447   3.951 8.47e-05 ***
## e_rvisits:comp_bin -0.09753    0.09630  -1.013    0.311
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.759 on 784 degrees of freedom
## Multiple R-squared:  0.1405, Adjusted R-squared:  0.1372
## F-statistic: 42.72 on 3 and 784 DF,  p-value: < 2.2e-16
```

```
hd_data_tidy |>
  ggplot(aes(x = e_rvisits, y = log(totalcost + 1), color = factor(comp_bin))) +
  geom_point() +
  geom_smooth(method="lm", se=F, aes(group = comp_bin, color = factor(comp_bin))) +
  labs(title = "transformed total cost vs er visit, by complications",
    x = "er visit",
    y = "transformed total cost") +
  viridis::scale_color_viridis(name = "complications", discrete = TRUE, option = "viridis") +
  theme_bw()
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



Comment:
(a) We can start an anova test on the model, and we notice that both variables have a significant effects on total cost.
(b) Then, I use formula of `log(totalcost + 1) ~ e_rvisits * comp_bin` to take interaction effect into account. And I find that there is no evidence showing that interaction effect would exist in their relationship.
(c) Lastly, I draw two lines colored by complication situation. Since two lines are not parallel, I can reach the conclusion that complications would not be a confounder in this case.

(iii)

Comment: Since `comp_bin` is not a confounder (nor has a interaction effect with `er_visit`) in the model and it shows a significant effect on `total_cost`, it should be contained in the model and it will not affect the effect that `er_visit` has over `total_cost`.

(f)

(i)

```
regmulti_hd_5 = lm(log(totalcost + 1) ~ e_rvisits + comp_bin + age + gender + duration, data = hd_data_
summary(regmulti_hd_5)
```

```
##
## Call:
## lm(formula = log(totalcost + 1) ~ e_rvisits + comp_bin + age +
##     gender + duration, data = hd_data_tidy)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.4711 -1.0340 -0.1158  0.9493  4.3372
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.9404610  0.5104064  11.639  < 2e-16 ***
## e_rvisits    0.1745975  0.0225736   7.735 3.20e-14 ***
## comp_bin     1.5044946  0.2584882   5.820 8.57e-09 ***
## age         -0.0206475  0.0086746  -2.380   0.0175 *
## gender      -0.2067662  0.1387002  -1.491   0.1364
## duration     0.0057150  0.0004888  11.691  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.624 on 782 degrees of freedom
## Multiple R-squared:  0.2694, Adjusted R-squared:  0.2647
## F-statistic: 57.68 on 5 and 782 DF,  p-value: < 2.2e-16
```

Comment: `age` is statistically significant at = 0.05 significant level (p_value < 0.05) and `duration` is statistically significant at any regular significant level (p_value < 0.001). We shall include `duration` when generate a linear model.
Given other conditions unchanged, a unit increase in `duration` would lead to a 0.0057 unit increase in log(total cost + 1).

(ii)

```
regmulti_hd_test = lm(log(totalcost + 1) ~ e_rvisits + comp_bin + duration, data = hd_data_tidy)
summary(regmulti_hd_test)
```

```
##
## Call:
## lm(formula = log(totalcost + 1) ~ e_rvisits + comp_bin + duration,
##     data = hd_data_tidy)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.5679 -1.0946 -0.1217  0.9612  4.6414
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 4.7266036  0.1173587  40.275  < 2e-16 ***
## e_rvisits   0.1682654  0.0224921   7.481 1.98e-13 ***
## comp_bin    1.5389634  0.2590331   5.941 4.25e-09 ***
## duration    0.0055568  0.0004864  11.425  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.63 on 784 degrees of freedom
## Multiple R-squared:  0.2622, Adjusted R-squared:  0.2594
## F-statistic: 92.88 on 3 and 784 DF,  p-value: < 2.2e-16
```

Comment: Given the adjusted r squared shown in both model analysis, MLR model (0.2594) would be more appropriate than SLR model (0.0973).