p8130_hw_4

Chunxiao Zhai cz2544

11/8/2018

Problem 1 (10p):

Consider the simple linear regression model:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \epsilon_i \sim N(0, \sigma^2)$$

The 'estimated errors' of the model are called residuals and denoted by $e_i = Y_i - \hat{Y}_i$.

a) Write (not derive) the Least Squares estimators of β_0, β_1 are unbiased estimators of the true model parameters ' do not use matrix notation! (5p)

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2} \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

They are unbiased estimators of the true model parameters in that:

$$\sum_{i=1}^n (X_i - \bar{X}) = 0 \Rightarrow \bar{Y} \sum_{i=1}^n (X_i - \bar{X}) = 0 \Rightarrow \hat{\beta}_1 = \frac{\sum_{i=1}^n (X_i - \bar{X}) Y_i}{\sum_{i=1}^n (X_i - \bar{X})^2}$$

For all X_is have been observed, let $\frac{(X_i - \bar{X})}{\sum_{i=1}^n (X_i - \bar{X})^2} = c_i$, then:

$$\hat{\beta}_1 = \sum_{i=1}^n c_i Y_i E[\hat{\beta}_1] = \sum_{i=1}^n E[c_i Y_i] = \sum_{i=1}^n E[c_i(\beta_0 + \beta_1 X_i + \epsilon_i)] = \beta_0 \sum_{i=1}^n c_i + \beta_1 \sum_{i=1}^n E[c_i X_i] + \sum_{i=1}^n E[c_i \epsilon_i]$$

With $\sum_{i=1}^n c_i = 0$, each c_i and X_i known, $E[\epsilon_i] = 0$,

$$E[\hat{\beta}_1] = \beta_1 \sum_{i=1}^n c_i X_i + \sum_{i=1}^n c_i E[\epsilon_i] = \beta_1 \sum_{i=1}^n c_i X_i = \beta_1 \frac{\sum_{i=1}^n (X_i - \bar{X}) X_i}{\sum_{i=1}^n (X_i - \bar{X})^2} = \beta_1 \frac{\sum_{i=1}^n (X_i - \bar{X})(X_i - \bar{X})}{\sum_{i=1}^n (X_i - \bar{X})^2} = \beta_1$$

$\hat{\beta}_1 = \bar{Y} - \hat{\beta}_1 \bar{X}$, with$\bar{Y} = \beta_0 + \beta 1 \bar{X}$,

$$E[\hat{\beta}_0] = E[\bar{Y} - \hat{\beta}_1 \bar{X}] = \bar{Y} - E[\hat{\beta}_1 \bar{X}] = \beta_0 + \beta_1 \bar{X} - E[\hat{\beta}_1] \bar{X} = \beta_0$$

b) Write the Least Squares line equation and show that it always goes through the point $(\bar{X}, \bar{Y})$.(2p) Least Squares line equation is: $\hat{Y} = \hat{\beta}_1 X + \hat{\beta}_0$. When deriving β_0, β_1, the partial derivatives were set to 0, in which:

$$Q = \sum_{i=1}^n (Y_i - \beta_0 - \beta_1 X_i)^2 \frac{\partial Q}{\partial \beta_0} = -2 \sum_{i=1}^n (Y_i - \beta_0 - \beta_1 X_i) = 0 \Rightarrow \sum_{i=1}^n Y_i - n\beta_0 - \beta_1 \sum_{i=1}^n X_i = 0$$

For β_0, β_1 solved to minimize Q as $\hat{\beta}_0, \hat{\beta}_1$, $\sum_{i=1}^n Y_i = \hat{\beta}_1 \sum_{i=1}^n X_i + n\hat{\beta}_0$, thus $\bar{Y} = \hat{\beta}_1 \bar{X} + \hat{\beta}_0$, for $X = \bar{X}$, $\hat{Y} = \bar{Y}$, point $(\bar{X}, \bar{Y})$ is on the Least Squares line.

$\bar{Y} = \hat{\beta}_1 \bar{X} + \hat{\beta}_0$ is the formular used to solve the $\hat{\beta}_0$, so it is always true for any pair of solved $\hat{\beta}_0, \hat{\beta}_1$, which means the line always goes through point $(\bar{X}, \bar{Y})$.

c) Use maximum likelihood method to derive an estimator of σ^2. Find its expected value and comment on the unbiasness property. (3p) for $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \epsilon_i \sim N(0, \sigma^2)$, $Y_i \sim N(\beta_0 + \beta_1 X_i, \sigma^2)$, the probability density function for the ith observation (X_i, Y_i) is:

$$f_i = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{Y_i - \beta_0 - \beta_1 X_i}{\sigma})^2}$$

The likelihood function:

$$L(\beta_0, \beta 1, \sigma) = \prod_{i=1}^{n} f_i = (\sigma\sqrt{2\pi})^{-n} \prod_{i=1}^{n} e^{-\frac{1}{2}(\frac{Y_i - \beta_0 - \beta_1 X_i}{\sigma})^2}$$

The log-likelihood function:

$$l = ln(L) = -n * ln(\sigma\sqrt{2\pi}) + \sum_{i=1}^{n} -\frac{1}{2}(\frac{Y_i - \beta_0 - \beta_1 X_i}{\sigma})^2$$

to maximize likelihood all partial derivatives are set to 0, in which:

$$\frac{\partial l}{\partial \sigma} = -\frac{n}{\sigma} + \sum_{i=1}^{n} \frac{(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2}{\sigma^{-3}} = 0$$

The estimators of β_0, β_1 are the same in Maximum Likelihood Estimation method and the Least Squares Estimation method, so:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2}{n} E[\hat{\sigma}^2] = E[\frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n}] = \frac{E[\sum_{i=1}^{n} e_i{}^2]}{n} = \frac{E[SSE]}{n}$$

For $MSE = SSE/df = SSE/(n-2)$, $E[MSE] = \sigma^2$ the expected value of LSE estimator of variance is:

$$E[\hat{\sigma}^2] = \frac{(n-2)E[MSE]}{n} = \frac{(n-2)}{n}\sigma^2$$

It is unbiased only when n is a large number.

For all problems below, assume a significance level of 0.05 unless stated otherwise. You can use R to perform the analyses, but you need to write the hypotheses where specified.

Problem 2 (25p)

For this problem, you will be using data 'HeartDisease.csv'. The investigator is mainly interested if there is an association between 'total cost' (in dollars) of patients diagnosed with heart disease and the 'number of emergency room (ER) visits'. Further, the model will need to be adjusted for other factors, including 'age', 'gender', 'number of complications' that arose during treatment, and 'duration of treatment condition'.

a) Provide a short description of the data set: what is the main outcome, main predictor and other important covariates. Also, generate appropriate descriptive statistics for all variables of interest (continuous and categorical) ' no test required. (5p)

```
heart_data = read.csv(file = "./HeartDisease.csv")
```

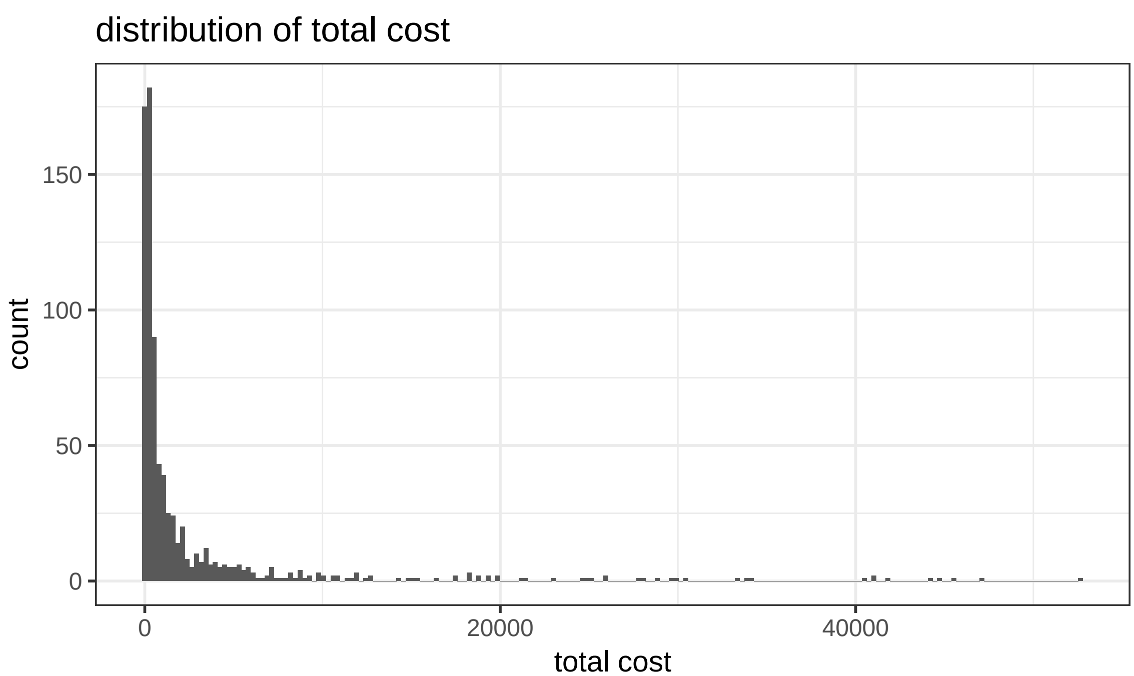
This data set contains 788 observations of 10 variables: id, totalcost, age, gender, interventions, drugs, ERvisits, complications, comorbidities, duration. Among these, the main outcome is "total cost" with mean of 2799.9559645, and standard error of6690.2604647, the main predictor is "ERvisits", with mean of 3.4251269 and standard error of2.6374737,. In all other variables, gender is the only categorical variable with 180 males and 608 females. The descriptive statistics for all variables of interest are listed in the table below:

```r
heart_data %>%
  select(-gender, -id) %>%
  skimr::skim() %>%
  filter(stat %in% c("n", "mean", "sd", "p25", "p50", "p75")) %>%
  group_by(variable, type) %>%
  nest(stat, formatted) %>% unnest() %>% spread(stat, formatted) %>%
  select(variable, type, n, mean, sd, everything()) %>%
  knitr::kable(digits = 1)
```

variable	type	n	mean	sd	p25	p50	p75
ERvisits	integer	788	3.43	2.64	2	3	5
age	integer	788	58.72	6.75	55	60	64
comorbidities	integer	788	3.77	5.95	0	1	5
complications	integer	788	0.057	0.25	0	0	0
drugs	integer	788	0.45	1.06	0	0	0
duration	integer	788	164.03	120.92	41.75	165.5	281
interventions	integer	788	4.71	5.59	1	3	6
totalcost	numeric	788	2799.96	6690.26	161.12	507.2	1905.45

b) Investigate the shape of the distribution for variable ‘total cost’ and try different transformations, if needed. (2p)

```r
heart_data %>%
  ggplot(aes(x = totalcost)) +
  geom_histogram(bins = 200) +
  labs(title = "distribution of total cost",
       x = "total cost")
```

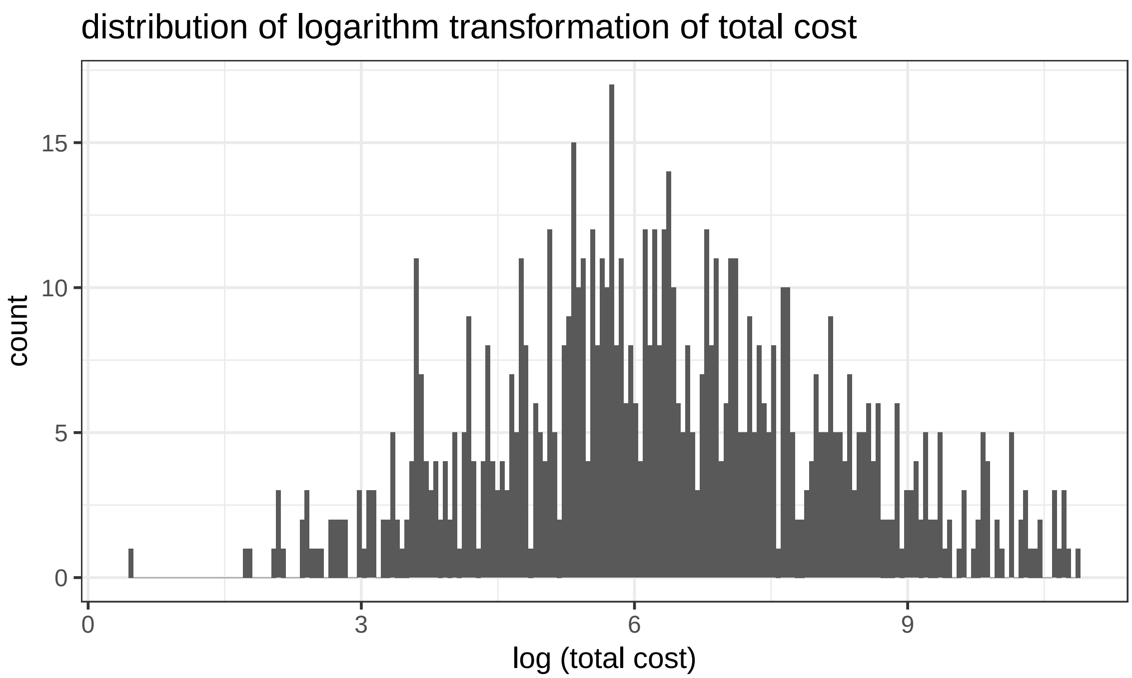


The histogram shows that the distribution of ‘total cost’ is very right skewed, with majority of observations less than 508.

Try logarithm transformation :

```r
log = heart_data %>%
  mutate(log_tolcost = log(totalcost)) %>%
  ggplot(aes(x = log_tolcost)) +
  geom_histogram(bins = 200) +
  labs(title = "distribution of logarithm transformation of total cost",
       x = "log (total cost)")
log
```

```
## Warning: Removed 3 rows containing non-finite values (stat_bin).
```

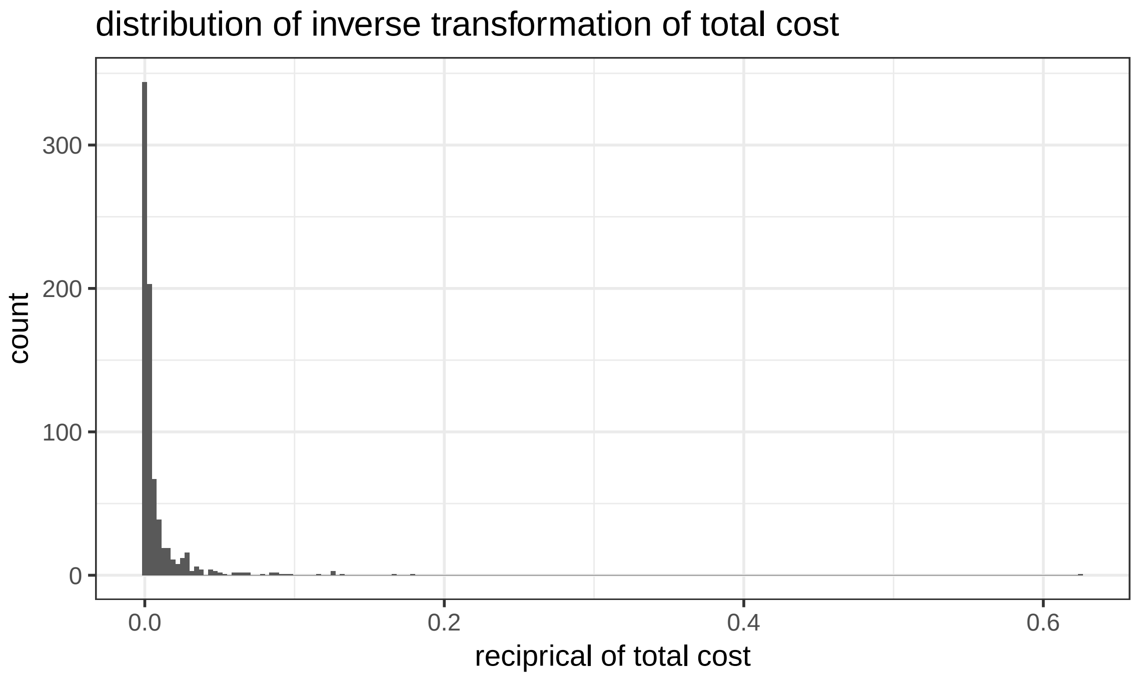


The transformed variavle have a good bell curve distribution, although still has some outliers.

Try inverse transformation :

```r
inv = heart_data %>%
  mutate(inverse_tolcost = 1/totalcost) %>%
  ggplot(aes(x = inverse_tolcost)) +
  geom_histogram(bins = 200) +
  labs(title = "distribution of inverse transformation of total cost",
       x = "reciprical of total cost")
inv
```

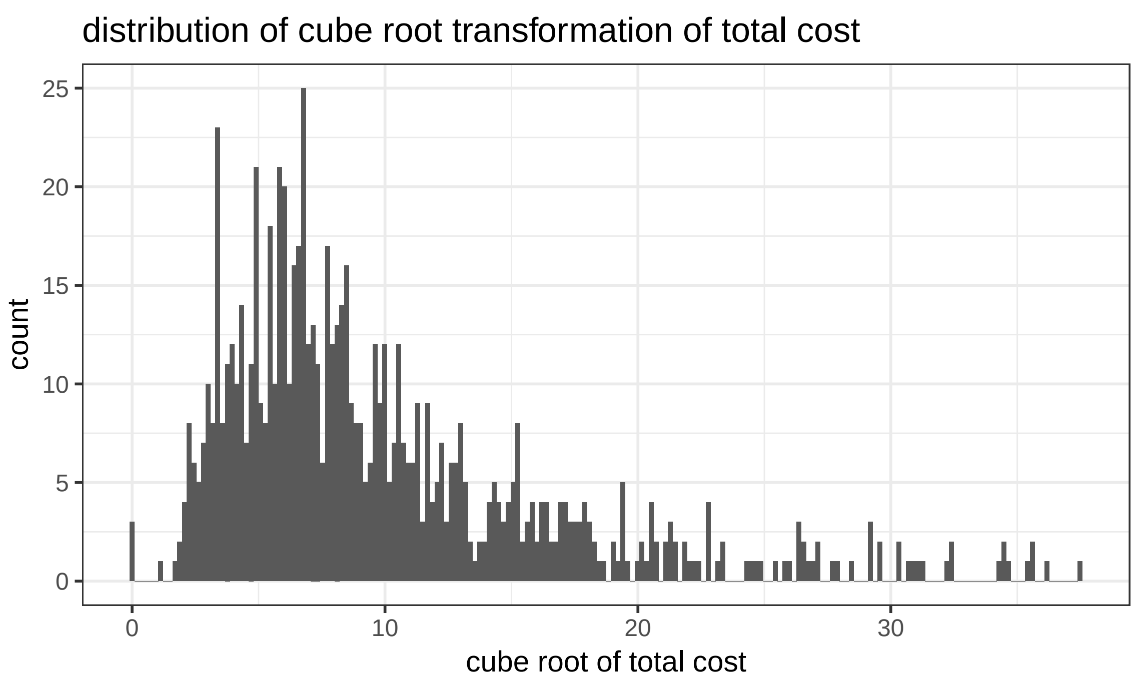
```
## Warning: Removed 3 rows containing non-finite values (stat_bin).
```



The distribution of recipricals of totalcost is still extremely right skewed, this transformation is not effective.

Try cube root transformation :

```
cbrt = heart_data %>%
  mutate(cbrt_tolcost = totalcost^(1/3)) %>%
  ggplot(aes(x = cbrt_tolcost)) +
  geom_histogram(bins = 200) +
  labs(title = "distribution of cube root transformation of total cost",
       x = "cube root of total cost")
cbrt
```



This transformation is better than inverse transformation, but the transformed distribution is still right skewed. The square root transformation will be weaker than cube root. The logarithm transformation is the best way to approach narmality in this case.

c) Create a new variable called 'comp_bin' by dichotomizing 'complications': 0 if no complications, and 1 otherwise. (1p)

```
heart_data = heart_data %>%
  mutate(comp_bin = if_else(complications == 0, 0, 1))
```

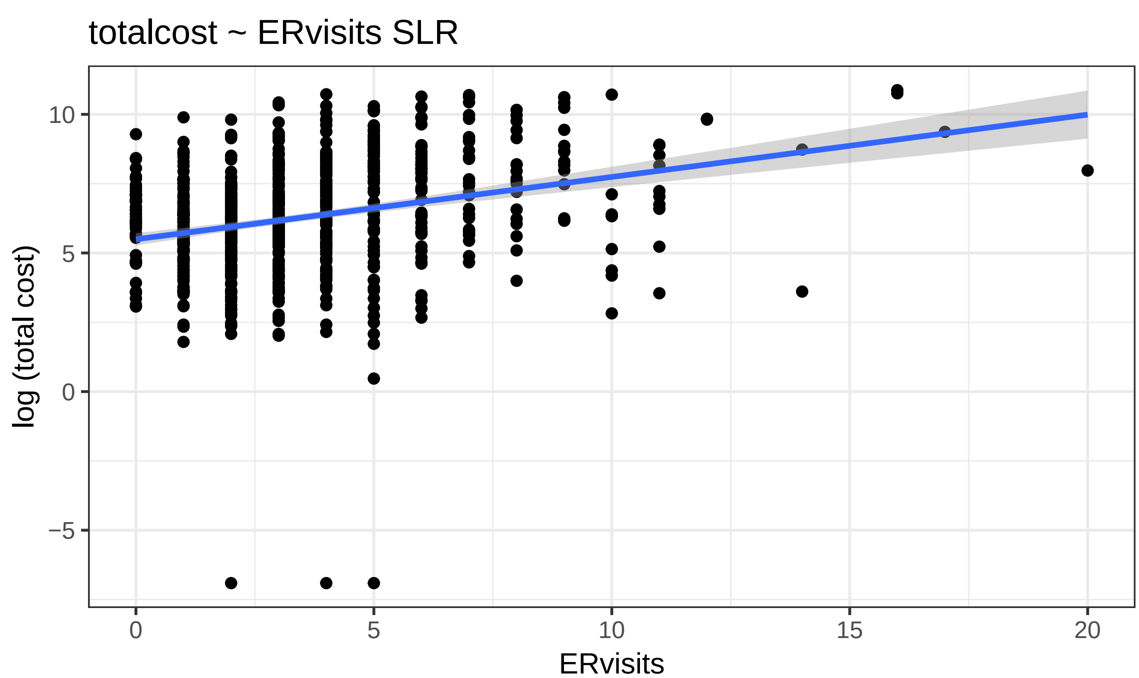
d) Based on our decision in part b), fit a simple linear regression (SLR) between the original or transformed 'total cost' and predictor 'ERvisits'. This includes a scatterplot and results of the regression, with appropriate comments on significance and interpretation of the slope. (5p)

```
heart_data_log = heart_data %>%
  mutate(totalcost = if_else(totalcost == 0, 0.001, totalcost),
         log_tolcost = log(totalcost))

fit_slr_trans = lm(log_tolcost ~ ERvisits, data = heart_data_log)

scatter_trans = ggplot(data = heart_data_log, aes(x = ERvisits, y = log_tolcost)) +
  geom_point() +
  geom_smooth(method = lm) +
  labs(title = "totalcost ~ ERvisits SLR",
       y = "log (total cost)")

scatter_trans
```




```
summary(fit_slr_trans)
```

```
##
## Call:
## lm(formula = log_tolcost ~ ERvisits, data = heart_data_log)
##
## Residuals:
```

```
##      Min       1Q   Median       3Q      Max
## -13.5255  -1.0922   0.0608   1.3147   4.3314
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.49384    0.11387  48.248   <2e-16 ***
## ERvisits     0.22477    0.02635   8.531   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.949 on 786 degrees of freedom
## Multiple R-squared:  0.08475,    Adjusted R-squared:  0.08359
## F-statistic: 72.78 on 1 and 786 DF,  p-value: < 2.2e-16
```

```r
beta_transback = broom::tidy(fit_slr_trans) %>% pull(estimate) %>% exp()
beta_transback
```

```
## [1] 243.190361   1.252036
```

comments : The value of slope(1.25) is the change in the ratio of the expected geometric means of 'total cost' as 'ERvisits' increase by 1. The intercept(243.2) is the geometric mean of 'total cost'. The Pr(>|t|) is the chance to observe this value of 1.25 when we assume the slope is 0, it is <2e-16. The model is significant with $\alpha = 0.05$

e) Fit a multiple linear regression (MLR) with 'comp_bin' and 'ERvisits' as predictors.

```r
fit_mlr_trans = lm(log_tolcost ~ ERvisits + comp_bin , data = heart_data_log)
summary(fit_mlr_trans)
```

```
##
## Call:
## lm(formula = log_tolcost ~ ERvisits + comp_bin, data = heart_data_log)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.3943  -1.0451   0.0252   1.2191   4.4397
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.47693    0.11165  49.054  < 2e-16 ***
## ERvisits     0.20193    0.02613   7.728 3.33e-14 ***
## comp_bin     1.74365    0.30321   5.751 1.27e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.911 on 785 degrees of freedom
## Multiple R-squared:  0.1218, Adjusted R-squared:  0.1195
## F-statistic: 54.41 on 2 and 785 DF,  p-value: < 2.2e-16
```

The MLR model is $log(total\ cost) = 5.5 + 0.2 * ERvisits + 1.7 * comp\_bin$

i) Test if 'comp_bin' is an effect modifier of the relationship between 'total cost' and 'ERvisits'. Comment. (2p)

```r
fit_mlr_interact = lm(log_tolcost ~ ERvisits + comp_bin + ERvisits:comp_bin, data = heart_data_log)
summary(fit_mlr_interact)
```

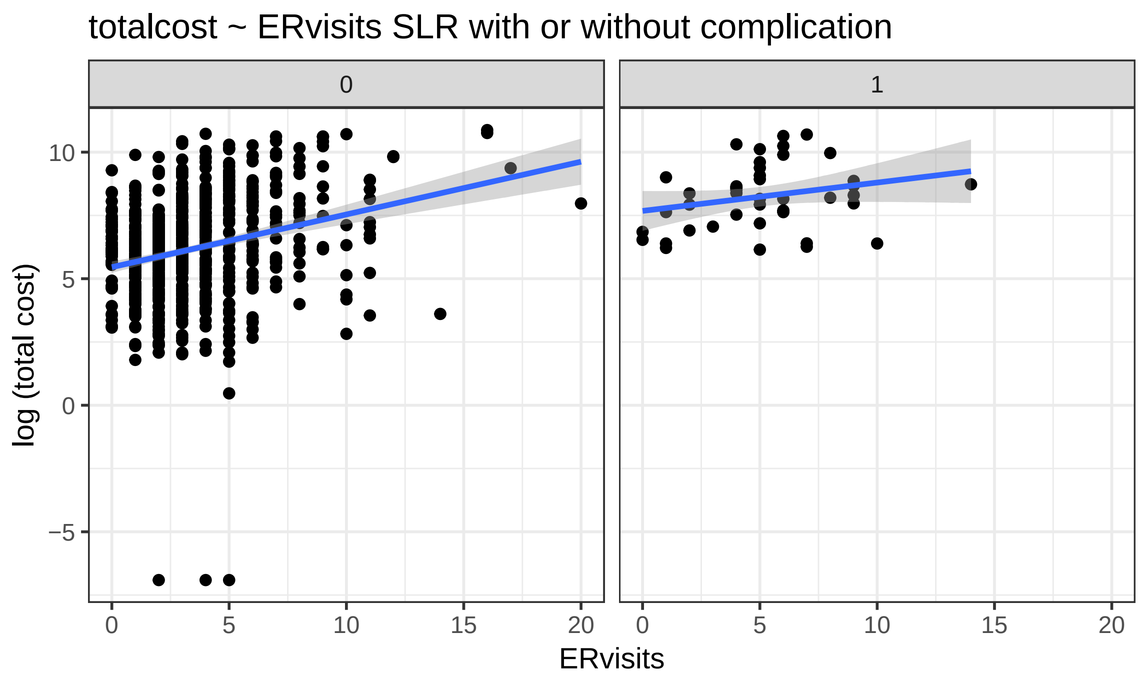
```
##
```

```
## Call:
## lm(formula = log_tolcost ~ ERvisits + comp_bin + ERvisits:comp_bin,
##     data = heart_data_log)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.4051  -1.0559   0.0325   1.2269   4.4353
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        5.45548    0.11406  47.828  < 2e-16 ***
## ERvisits           0.20837    0.02705   7.703 4.01e-14 ***
## comp_bin           2.22320    0.60233   3.691 0.000239 ***
## ERvisits:comp_bin -0.09639    0.10461  -0.921 0.357103
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.911 on 784 degrees of freedom
## Multiple R-squared:  0.1227, Adjusted R-squared:  0.1193
## F-statistic: 36.55 on 3 and 784 DF,  p-value: < 2.2e-16
```

'comp_bin' is not an effect modifier of the relationship between 'total cost' and 'ERvisits' at significance level of $\alpha = 0.05$.

```
p_strat =
  heart_data_log %>%
  group_by(comp_bin) %>%
  ggplot(aes(x = ERvisits, y = log_tolcost)) +
  geom_point() +
  geom_smooth(method = lm) +
  labs(title = "totalcost ~ ERvisits SLR with or without complication",
       y = "log (total cost)") +
  facet_grid(~ comp_bin)

p_strat
```




```r
strat_comp_0 = heart_data_log %>% filter(comp_bin==0)
strat_comp_1 = heart_data_log %>% filter(comp_bin==1)
fit_comp_0 = lm(log_tolcost ~ ERvisits, data = strat_comp_0) %>% broom::tidy()
fit_comp_1 = lm(log_tolcost ~ ERvisits, data = strat_comp_1) %>% broom::tidy()
fit_comp_0
```

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic   p.value
##   <chr>          <dbl>     <dbl>     <dbl>     <dbl>
## 1 (Intercept)    5.46     0.116      47.1  2.89e-225
## 2 ERvisits       0.208    0.0275      7.59 9.95e- 14
```

```r
fit_comp_1
```

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)    7.68     0.389      19.8  1.46e-22
## 2 ERvisits       0.112    0.0664      1.69 9.94e- 2
```

In epidemiology, to decide if a variavle is a effect measurement modifier, we compare the measurement in each stratum to the crude. The stratum without complication has a slope of 1.2312132 which is less than the crude of 1.2523227, but the difference is less than 10%. While the stratum with complications has a slope of 1.1185129 which is also less than the crude of 1.2523227, but the difference is about 11%.

ii)Test if 'comp_bin' is a confounder of the relationship between 'total cost' and 'ERvisits'.Comment. (2p)

```r
anova(fit_slr_trans, fit_mlr_trans)
```

```
## Analysis of Variance Table
##
## Model 1: log_tolcost ~ ERvisits
## Model 2: log_tolcost ~ ERvisits + comp_bin
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    786 2986.9
```

```
## 2    785 2866.1  1    120.74 33.07 1.273e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
exp(coef(fit_slr_trans)[2])
```

```
## ERvisits
## 1.252036
```

```
exp(coef(fit_mlr_trans)[2])
```

```
## ERvisits
## 1.223761
```

```
1-exp(coef(fit_mlr_trans)[2])/exp(coef(fit_slr_trans)[2])
```

```
##   ERvisits
## 0.02258374
```

When add 'comp_bin' into the model, the slope(ratio of the expected geometric means of 'total cost' as 'ERvisits' increase by 1) will decrease from 1.25 to 1.22, by 2.2%. By 10% criterion, not considered as confounder between 'total cost' and 'ERvisits'.

iii) Decide if 'comp_bin' should be included along with 'ERvisits. Why or why not?(1p)

```
anova(fit_slr_trans, fit_mlr_trans)
```

```
## Analysis of Variance Table
##
## Model 1: log_tolcost ~ ERvisits
## Model 2: log_tolcost ~ ERvisits + comp_bin
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    786 2986.9
## 2    785 2866.1  1    120.74 33.07 1.273e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

'comp_bin' should be included given when add into the model, the adjusted R-squared increase from 0.08359 to 0.1195, the anova test prefer the larger model, and it is a effect measurement modifier in the stratum with complications.

f) Use your choice of model in part e) and add additional covariates (age, gender, and duration of treatment).

g) Fit a MLR, show the regression results and comment. (5p)

```
fit_mlr_add = lm(log_tolcost ~ ERvisits + comp_bin + age + gender + duration ,data = heart_data_log)
summary(fit_mlr_add)
```

```
##
## Call:
## lm(formula = log_tolcost ~ ERvisits + comp_bin + age + gender +
##     duration, data = heart_data_log)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.1885  -0.9962  -0.0838   1.0099   4.3499
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)  5.8016080  0.5559910  10.435  < 2e-16 ***
## ERvisits     0.1732359  0.0245897   7.045 4.07e-12 ***
## comp_bin     1.5335773  0.2815738   5.446 6.89e-08 ***
## age         -0.0193389  0.0094493  -2.047   0.0410 *
## gender      -0.3234418  0.1510875  -2.141   0.0326 *
## duration     0.0060629  0.0005325  11.386  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.769 on 782 degrees of freedom
## Multiple R-squared:  0.2502, Adjusted R-squared:  0.2454
## F-statistic: 52.18 on 5 and 782 DF,  p-value: < 2.2e-16
```

```
fit_mlr_add_satuated = lm(log_tolcost ~ ERvisits + comp_bin + age + gender + duration
                          + ERvisits*comp_bin + ERvisits*gender
                          + comp_bin*age + comp_bin*gender + comp_bin*duration
                          + age*gender
                          + gender*duration
                          ,data = heart_data_log)
summary(fit_mlr_add_satuated)
```

```
##
## Call:
## lm(formula = log_tolcost ~ ERvisits + comp_bin + age + gender +
##     duration + ERvisits * comp_bin + ERvisits * gender + comp_bin *
##     age + comp_bin * gender + comp_bin * duration + age * gender +
##     gender * duration, data = heart_data_log)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.6331  -1.0276  -0.0902   0.9795   4.3658
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        6.2176155  0.6391235   9.728  < 2e-16 ***
## ERvisits           0.2230079  0.0302710   7.367 4.47e-13 ***
## comp_bin           2.8748623  2.9942653   0.960   0.3373
## age               -0.0275914  0.0108861  -2.535   0.0115 *
## gender            -2.2856362  1.3007360  -1.757   0.0793 .
## duration           0.0055696  0.0006136   9.076  < 2e-16 ***
## ERvisits:comp_bin -0.1391906  0.1058228  -1.315   0.1888
## ERvisits:gender   -0.1253428  0.0531661  -2.358   0.0186 *
## comp_bin:age      -0.0081987  0.0500355  -0.164   0.8699
## comp_bin:gender    0.6529043  0.6770563   0.964   0.3352
## comp_bin:duration -0.0020781  0.0026232  -0.792   0.4285
## age:gender         0.0331379  0.0222942   1.486   0.1376
## gender:duration    0.0026648  0.0012870   2.070   0.0387 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.759 on 775 degrees of freedom
## Multiple R-squared:  0.2653, Adjusted R-squared:  0.2539
## F-statistic: 23.32 on 12 and 775 DF,  p-value: < 2.2e-16
```

```r
fit_mlr_add_1 = lm(log_tolcost ~ ERvisits + comp_bin + age + gender + duration
                          + ERvisits*gender + gender*duration ,data = heart_data_log)
summary(fit_mlr_add_1)
```

```
##
## Call:
## lm(formula = log_tolcost ~ ERvisits + comp_bin + age + gender +
##     duration + ERvisits * gender + gender * duration, data = heart_data_log)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.7024  -1.0390  -0.0979   0.9692   4.3403
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      5.8185494  0.5577731  10.432  < 2e-16 ***
## ERvisits         0.2075835  0.0293277   7.078 3.26e-12 ***
## comp_bin         1.4554186  0.2814426   5.171 2.96e-07 ***
## age             -0.0195064  0.0094059  -2.074   0.0384 *
## gender          -0.4283036  0.3081114  -1.390   0.1649
## duration         0.0053540  0.0005993   8.934  < 2e-16 ***
## ERvisits:gender -0.1121332  0.0528909  -2.120   0.0343 *
## gender:duration  0.0031348  0.0012617   2.485   0.0132 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.761 on 780 degrees of freedom
## Multiple R-squared:  0.2591, Adjusted R-squared:  0.2524
## F-statistic: 38.96 on 7 and 780 DF,  p-value: < 2.2e-16
```

```r
strat_female = heart_data_log %>% filter(gender==0)
strat_male = heart_data_log %>% filter(gender==1)
fit_f = lm(log_tolcost ~ ERvisits + comp_bin + age + duration, data = strat_female) %>% broom::tidy()
fit_m = lm(log_tolcost ~ ERvisits + comp_bin + age + duration, data = strat_male) %>% broom::tidy()

fit_m
```

```
## # A tibble: 5 x 5
##   term        estimate std.error statistic      p.value
##   <chr>          <dbl>     <dbl>     <dbl>        <dbl>
## 1 (Intercept)  3.97      1.42        2.80  0.00564
## 2 ERvisits     0.0904    0.0551      1.64  0.103
## 3 comp_bin     1.91      0.696       2.75  0.00660
## 4 age          0.00566   0.0243      0.233 0.816
## 5 duration     0.00813   0.00142     5.74  0.0000000410
```

```r
fit_f
```

```
## # A tibble: 5 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)  6.24     0.578        10.8  5.97e-25
## 2 ERvisits     0.211    0.0270        7.81 2.51e-14
## 3 comp_bin     1.31     0.298         4.39 1.33e- 5
## 4 age         -0.0269   0.00983      -2.74 6.34e- 3
```

```
## 5 duration     0.00541  0.000550      9.84 2.70e-21
```

```r
fit_m_s = lm(log_tolcost ~ comp_bin + duration, data = strat_male)
fit_m_l =lm(log_tolcost ~ ERvisits + comp_bin + age + duration, data = strat_male)
anova(fit_m_s, fit_m_l)
```

```
## Analysis of Variance Table
##
## Model 1: log_tolcost ~ comp_bin + duration
## Model 2: log_tolcost ~ ERvisits + comp_bin + age + duration
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    177 855.21
## 2    175 841.69  2    13.516 1.4051 0.2481
```

```r
fit_m_s %>% broom::tidy()
```

```
## # A tibble: 3 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)  4.60      0.284       16.2  9.77e-37
## 2 comp_bin     1.94      0.694        2.79 5.79e- 3
## 3 duration     0.00848   0.00139      6.09 6.85e- 9
```

The overall model is significant at 0.05 level with adjusted R square of 0.2454, all five variables are significant at level of 0.05 when donot consider interaction. The MLR moedl is: $log(total\ cost) = 5.8 + 0.17ERvisit + 1.53 * comp\_bin - 0.02 * age - 0.32gender + 0.006 * duration$. But when add in all possible interactions involve categorical variables, only Ervisits:gender, duration:gender and age are significant at level of 0.05. By removing nonsignificant interactions from the MLR, comp_bin become significant again. The final MLR should be stratified by gender. For male, the ERvisit and age are no longer significant at level of 0.05, the anova prefer small model. The MLR moedl for male is: $log(total\ cost\_male) = 4.6+1.9*comp\_bin+0.0085*duration$. For female, all variables are significant at level of 0.05, The MLR moedl for female is: $log(total\ cost\_female) = 6.4 + 0.21 * ERvisit + 1.3 * comp\_bin - 0.027 * age + 0.0054 * duration$.

ii) Compare the SLR and MLR models. Which model would you use to address the investigator's objective and why? (2p)

I would use the MLR models, in that the main outcome of 'total cost' is influenced by different facters in patients of different gender. 'number of emergency room (ER) visits' is not a significant predictor of the main outcome of 'total cost' in male patients but a significant predictor in female patients. While other significant predictors duration and complication are only included in the MLR model. The reletion between ERvisits and total cost could be largely due to the factor of gender but not ERvisit itself. The SLR result is biased.

Problem 3 (15p)

A hospital administrator wishes to test the relationship between 'patient's satisfaction' (Y) and 'age', 'severity of illness', and 'anxiety level' (data 'PatSatisfaction.xlsx'). The administrator randomly selected 46 patients, collected the data, and asked for your help with the analysis.

```r
stf_data = readxl::read_excel("PatSatisfaction.xlsx") %>% janitor::clean_names()
```

a) Create a correlation matrix and interpret your initial findings. (2p)

```r
round(cor(stf_data),3)
```

```
##              safisfaction    age severity anxiety
## safisfaction        1.000 -0.787   -0.603  -0.645
## age                -0.787  1.000    0.568   0.570
## severity           -0.603  0.568    1.000   0.671
```

```
## anxiety             -0.645  0.570    0.671   1.000
```

All three factors are inversely correleted with satisfaction but age is the most closely one. Age, severity and anxiety are pairwise correlated with severity and anxiety the most closely one.

b) Fit a multiple regression model and test whether there is a regression relation. State the hypotheses, decision rule and conclusion. (3p)

```r
fit_stf_1 = lm(safisfaction ~ age + severity + anxiety, data = stf_data)
summary(fit_stf_1)
```

```
##
## Call:
## lm(formula = safisfaction ~ age + severity + anxiety, data = stf_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -18.3524  -6.4230   0.5196   8.3715  17.1601
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 158.4913    18.1259   8.744 5.26e-11 ***
## age          -1.1416     0.2148  -5.315 3.81e-06 ***
## severity     -0.4420     0.4920  -0.898   0.3741
## anxiety     -13.4702     7.0997  -1.897   0.0647 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.06 on 42 degrees of freedom
## Multiple R-squared:  0.6822, Adjusted R-squared:  0.6595
## F-statistic: 30.05 on 3 and 42 DF,  p-value: 1.542e-10
```

Hypotheses: H0: all beta = 0, no linear relation; H1: at least one beta not 0. The result show at significant 0.05 only age is a significant variable, but p value for anxiety is 0.065, try to remove severity:

```r
fit_stf_2 = lm(safisfaction ~ age + anxiety, data = stf_data)
summary(fit_stf_2)
```

```
##
## Call:
## lm(formula = safisfaction ~ age + anxiety, data = stf_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -19.4453  -7.3285   0.6733   8.5126  18.0534
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 145.9412    11.5251  12.663 4.21e-16 ***
## age          -1.2005     0.2041  -5.882 5.43e-07 ***
## anxiety     -16.7421     6.0808  -2.753  0.00861 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.04 on 43 degrees of freedom
## Multiple R-squared:  0.6761, Adjusted R-squared:  0.661
## F-statistic: 44.88 on 2 and 43 DF,  p-value: 2.98e-11
```

Both age and anxiety are significant when remove severity from the MLR, the adjusted R-square in creased form 0.6595 to 0.661. Then MLR model is :$Safisfaction = 145.9 - 1.2 * age - 16.7 * anxiety$

c) Show the regression results for all estimated coefficients with 95% CIs. Interpret the coefficient and 95% CI associated with 'severity of illness'. (5p)

```
confint(fit_stf_1, level = 0.95)
```

```
##                  2.5 %      97.5 %
## (Intercept) 121.911727 195.0707761
## age          -1.575093  -0.7081303
## severity     -1.434831   0.5508228
## anxiety     -27.797859   0.8575324
```

The Intercept is the expectation of mean value of Safisfaction without considering predictors, we are 95% confident that it is in (121.911727, 195.0707761), with age increase by 1, expectation of mean value of Safisfaction will decrease by 1.14, 95% condifent in (-1.575093, -0.7081303), with the anxiety level increase by 1, expectation of mean value of Safisfaction will decrease by 13.5, 95% condifent in(-27.797859, 0.8575324).

The 95% CI associated with 'severity of illness' contain 0, indicate is is not significant, for it could change the satisfaction to either directions.

d) Obtain an interval estimate for a new patient's satisfaction when Age=35, Severity=42, Anxiety=2.1. Interpret the interval. (2p)

```
new = data.frame(age=35, severity = 42, anxiety = 2.1)
predict.lm(fit_stf_1, newdata = new, interval = "prediction")
```

```
##        fit      lwr      upr
## 1 71.68332 50.06237 93.30426
```

```
predict.lm(fit_stf_2, newdata = new, interval = "prediction")
```

```
##        fit      lwr      upr
## 1 68.76642 48.22251 89.31033
```

We are 95% confident that the new patient's satisfaction is between (50.06237, 93.30426) if all three variables are considered as predictors, or 95% confident that the new patient's satisfaction is between (48.22251 89.31033) if only Age and Anxiety are considered as predictors.

e) Test whether anxiety level can be dropped from the regression model, given the other two covariates are retained. State the hypotheses, decision rule and conclusion. (3p)

Hypotheses: H0: model without 'anxiety level' is the same as model with 'anxiety level' (the beta related to anxiety is 0); H1: model without 'anxiety level' and model with 'anxiety level' are different (the beta related to anxiety is not 0).

```
fit_stf_1 = lm(safisfaction ~ age + severity + anxiety, data = stf_data)
fit_stf_3 = lm(safisfaction ~ age + severity, data = stf_data)
summary(fit_stf_1)
```

```
##
## Call:
## lm(formula = safisfaction ~ age + severity + anxiety, data = stf_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -18.3524  -6.4230   0.5196   8.3715  17.1601
##
## Coefficients:
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 158.4913    18.1259   8.744 5.26e-11 ***
## age          -1.1416     0.2148  -5.315 3.81e-06 ***
## severity     -0.4420     0.4920  -0.898   0.3741
## anxiety     -13.4702     7.0997  -1.897   0.0647 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.06 on 42 degrees of freedom
## Multiple R-squared:  0.6822, Adjusted R-squared:  0.6595
## F-statistic: 30.05 on 3 and 42 DF,  p-value: 1.542e-10
```

```r
summary(fit_stf_3)
```

```
##
## Call:
## lm(formula = safisfaction ~ age + severity, data = stf_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -17.1662  -8.5462  -0.4595   7.1342  17.2364
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 156.6719    18.6396   8.405 1.27e-10 ***
## age          -1.2677     0.2104  -6.026 3.35e-07 ***
## severity     -0.9208     0.4349  -2.117   0.0401 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.36 on 43 degrees of freedom
## Multiple R-squared:  0.655,  Adjusted R-squared:  0.6389
## F-statistic: 40.81 on 2 and 43 DF,  p-value: 1.16e-10
```

```r
anova(fit_stf_3, fit_stf_1)
```

```
## Analysis of Variance Table
##
## Model 1: safisfaction ~ age + severity
## Model 2: safisfaction ~ age + severity + anxiety
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1     43 4613.0
## 2     42 4248.8  1    364.16 3.5997 0.06468 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

If the other two covariates are retained, at 95% level of confidence we cannot reject the null hypothesis, the model with or without anxiety are not significantly different. We should keep the model with less variables, drop 'anxiety level'. Adjusted R-square changed from 0.6595 to 0.6389, by only 3%.

Conclusion: 'anxiety level' can be dropped.