1. Perform a Deming regression using the provided data set (HW2\_data1.txt), and compare the results with a regular regression analysis.

Note: yy is the observed outcome, and xx is the observed predictor, and x\_star is the “unobserved” true x value. Your analysis should be based on the observed data (xx and yy), and x\_star could be used to evaluate the effectiveness of Deming regression in taking errors-in-variable into consideration.

|  |  |  |
| --- | --- | --- |
| Group | b\_0 | b\_1 |
| xx w/o errors | 3.02214 | 0.34279 |
| xx w errors | 2.614046 | 0.437296 |

A line graph with dots and lines

Description automatically generated

The Deming Regression is effective in evaluating errors in variables because the parameter estimates of xx with and without errors are similar. The ratio (r value) is small which suggests errors are not large. Plotting the OLS and Deming model shows that the OLS model appears to fit the data better.

> # Declare Libs

> library(car)

> library(dplyr)

> library(tidyverse)

> library(MethComp)

>

>

> # Question 1: Perform Deming Regression on HW2\_data1.txt + Regular regression

>

> ## import data file

>

> data1 <- read.table("C:/Users/arobe/OneDrive/Desktop/DataSP24/HW2\_data1.txt",

+ header=T)

> summary(data1)

yy x\_star xx

Min. :-0.3595 Min. :0.3806 Min. :-4.340

1st Qu.: 2.5577 1st Qu.:2.1675 1st Qu.: 1.703

Median : 3.9833 Median :4.0995 Median : 3.612

Mean : 4.5023 Mean :4.4732 Mean : 4.318

3rd Qu.: 6.4713 3rd Qu.:6.8285 3rd Qu.: 7.305

Max. : 9.2622 Max. :9.7919 Max. :14.261

>

> # Variables: yy, xx, and x\_star

> ## xx is observed values, x\_star is unobserved true values

> xx <- data1$xx

> yy <- data1$yy

> x\_star <- data1$x\_star

>

>

> d1dem <- Deming(x=xx, y=yy);d1dem #y\*=a+bx\*

Intercept Slope sigma.xx sigma.yy

2.614046 0.437296 1.869558 1.869558

> d1lm <- lm(yy~xx, data=data1);d1lm#y=a+bx or y+a+b(x\*+eta)

Call:

lm(formula = yy ~ xx, data = data1)

Coefficients:

(Intercept) xx

3.0221 0.3428

> summary(d1lm)

Call:

lm(formula = yy ~ xx, data = data1)

Residuals:

Min 1Q Median 3Q Max

-3.5570 -1.8159 -0.1074 1.6292 4.1999

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 3.02214 0.42062 7.185 3.83e-09 \*\*\*

xx 0.34279 0.07195 4.764 1.79e-05 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.005 on 48 degrees of freedom

Multiple R-squared: 0.321, Adjusted R-squared: 0.3069

F-statistic: 22.7 on 1 and 48 DF, p-value: 1.792e-05

> anova(d1lm)

Analysis of Variance Table

Response: yy

Df Sum Sq Mean Sq F value Pr(>F)

xx 1 91.218 91.218 22.696 1.792e-05 \*\*\*

Residuals 48 192.921 4.019

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> var(d1lm$residuals)

[1] 3.937159

>

> d1lms <- lm(yy~x\_star, data=data1);d1lms #y=a+bx\* or y=a+b(x-eta)

Call:

lm(formula = yy ~ x\_star, data = data1)

Coefficients:

(Intercept) x\_star

0.7797 0.8322

> summary(d1lms)

Call:

lm(formula = yy ~ x\_star, data = data1)

Residuals:

Min 1Q Median 3Q Max

-2.1081 -0.5208 0.1142 0.6016 1.4356

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 0.77968 0.25215 3.092 0.00331 \*\*

x\_star 0.83221 0.04847 17.170 < 2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9104 on 48 degrees of freedom

Multiple R-squared: 0.86, Adjusted R-squared: 0.8571

F-statistic: 294.8 on 1 and 48 DF, p-value: < 2.2e-16

> var(d1lms$residuals)

[1] 0.8119049

> anova(d1lms)

Analysis of Variance Table

Response: yy

Df Sum Sq Mean Sq F value Pr(>F)

x\_star 1 244.356 244.356 294.82 < 2.2e-16 \*\*\*

Residuals 48 39.783 0.829

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

>

> eta <- xx-x\_star

> sd(eta)

[1] 3.232759

> r <- sd(eta)/sd(xx);r # r is not large

[1] 0.8121891

1. For the following matrices and vectors, ***manually*** conduct the indicated operations (you can use software to verify your results if you want)
2. Find the determinants of and.

Note the relationships across columns in C or D. Particularly, any individual column in C cannot be obtained as a linear combination of other columns in C. However, the third column of D can be obtained as the average of first two columns (or in other words, ½\*first column + ½\*second column). This illustrates the situation when a matrix has a zero determinant.

1. Compute
2. Find the inverse of

Using the properties of a 2x2 matrix, we can find the inverse of B using the following:

1. In a diabetes study, 1123 subjects were recruited, and a number of clinical traits and information were collected, including (see data in the attached file “HW2\_data2.txt”, which is the same data set used for HW1):

Sex: male/female Age: age of the study subject

bmi: body mass index fbg: fasting blood glucose

fins: fasting insulin hba1c: hemoglobin A1c

tg: total glyceride tcho: total cholesterol

hdl: high density lipoprotein ldl: low density lipoprotein

Particularly, the investigators are interested in the effects of fbg (X1) and tg (X2) on hba1c (Y). In addition, hba1c > 6.5 is considered to be diabetic. So another question is whether the same model can be used to characterize the relationship between predictors and hba1c for all individuals, or whether two different models are needed: one for people with diabetes, and the other for those without diabetes. Now, you are assigned to analyze the data, and are asked to complete the following tasks.

Consider a multiple linear regression model between hba1c and fbg + tg.

1. Given that and (with the order of intercept, fbg, tg, in the model). Manually obtain estimates for the intercept, and coefficients for fbg and tg.
2. Manually find the variances for and , and their correlation coefficient. Check your results using software of your choice (providing that ).
3. For the two ways in entering the predictors into the model (fbg first, then tg, vs tg first, then fbg), summarize in a table for the sequential and partial sums of squares, respectively, for fbg and tg.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Summary of Model and ANOVA | | | | | | | |
| Source | DF | Parameter Estimate | SE | Type I SS | Type II SS | t-Value | Pr(>|t|) |
| Intercept |  | 2.47787 | 0.12337 |  |  | 20.085 | <2e-16 |
| fbg | 1 | 0.63302 | 0.01597 | 2796.47 | 2615.18 | 39.650 | <2e-16 |
| tg | 1 | 0.01685 | 0.04506 | 0.23 | 0.23 | 0.374 | 0.708 |
| Residuals | 1120 |  |  | 1863.11 | 1863.11 |  |  |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Summary of Model and ANOVA | | | | | | | |
| Source | DF | Parameter Estimate | SE | Type I SS | Type II SS | t-Value | Pr(>|t|) |
| Intercept |  | 2.47787 | 0.12337 |  |  | 20.085 | <2e-16 |
| tg | 1 | 0.01685 | 0.04506 | 181.53 | 0.23 | 0.374 | 0.708 |
| fbg | 1 | 0.63302 | 0.01597 | 2615.18 | 2615.18 | 39.650 | <2e-16 |
| Residuals | 1120 |  |  | 1863.11 | 1863.11 |  |  |

> # 3) find SS2 (partial SS) and SS1 (sequential SS)

> anova(d2lm) #SS1 Sequential SS for order 1

Analysis of Variance Table

Response: hba1c

Df Sum Sq Mean Sq F value Pr(>F)

fbg 1 2796.47 2796.47 1681.0876 <2e-16 \*\*\*

tg 1 0.23 0.23 0.1399 0.7085

Residuals 1120 1863.11 1.66

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> summary(d2lm)

Call:

lm(formula = hba1c ~ fbg + tg, data = data2)

Residuals:

Min 1Q Median 3Q Max

-4.3483 -0.6595 -0.1752 0.4311 5.6142

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 2.47787 0.12337 20.085 <2e-16 \*\*\*

fbg 0.63302 0.01597 39.650 <2e-16 \*\*\*

tg 0.01685 0.04506 0.374 0.708

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.29 on 1120 degrees of freedom

Multiple R-squared: 0.6002, Adjusted R-squared: 0.5995

F-statistic: 840.6 on 2 and 1120 DF, p-value: < 2.2e-16

> d2lmb <- lm(hba1c ~ tg + fbg, data=data2)

> anova(d2lmb) #SS1 Sequential SS for order 2

Analysis of Variance Table

Response: hba1c

Df Sum Sq Mean Sq F value Pr(>F)

tg 1 181.53 181.53 109.13 < 2.2e-16 \*\*\*

fbg 1 2615.18 2615.18 1572.10 < 2.2e-16 \*\*\*

Residuals 1120 1863.11 1.66

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> summary(d2lmb)

Call:

lm(formula = hba1c ~ tg + fbg, data = data2)

Residuals:

Min 1Q Median 3Q Max

-4.3483 -0.6595 -0.1752 0.4311 5.6142

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 2.47787 0.12337 20.085 <2e-16 \*\*\*

tg 0.01685 0.04506 0.374 0.708

fbg 0.63302 0.01597 39.650 <2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.29 on 1120 degrees of freedom

Multiple R-squared: 0.6002, Adjusted R-squared: 0.5995

F-statistic: 840.6 on 2 and 1120 DF, p-value: < 2.2e-16

>

> SS2 <- Anova(d2lm, type='II') #SS2 Partial SS

> as.data.frame(SS2)

Sum Sq Df F value Pr(>F)

fbg 2615.176783 1 1572.1019978 1.60181e-215

tg 0.232716 1 0.1398962 7.08455e-01

Residuals 1863.109392 1120 NA NA

> Anova(d2lmb, type='II') #SS2 is same for order 2

Anova Table (Type II tests)

Response: hba1c

Sum Sq Df F value Pr(>F)

tg 0.23 1 0.1399 0.7085

fbg 2615.18 1 1572.1020 <2e-16 \*\*\*

Residuals 1863.11 1120

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1