Di Zhen (1717719)

dizhen

2020.3.13

```r
OLD <- c(2041.7,257.0,524.8,257.0,128.8,128.8,1023.3,134.9,257.0,125.9,257.0,123.0,1000.0,120.2,128.8)
NEW <- c(12302.7,6918.3,4466.8,1584.9,933.3,1659.6,9120.1,575.4,2630.3,2398.8,1905.5,851.1,3467.4,1380.4
mydata <- data.frame(OLD, NEW)
```

Question 1

Using R, we can obtain b1 directly. b1 is calculated by $b_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$

The estimated slope parameter is 0.13

```r
mydata.LM <- lm(OLD~NEW,data = mydata)
summary(mydata.LM)
```

```
##
## Call:
## lm(formula = OLD ~ NEW, data = mydata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -652.41  -74.15   22.18   69.15  554.99
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -21.60568  101.69092  -0.212    0.835
## NEW           0.13457    0.02136   6.301 2.74e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 277 on 13 degrees of freedom
## Multiple R-squared:  0.7534, Adjusted R-squared:  0.7344
## F-statistic: 39.71 on 1 and 13 DF,  p-value: 2.739e-05
```

The slope parameter can be estimated from correlation coefficient:

$b_1 = r\frac{s_Y}{s_X}$

The estimated slope parameter is 0.13.

```r
corr <- cor(x = mydata$NEW, y = mydata$OLD)
sY <- sd(mydata$OLD)
sX <- sd(mydata$NEW)
b1 <- corr*sY/sX
b1
```

```
## [1] 0.1345733
```

These are two methods to calculate the slope estimate. $b_1 = r\frac{s_Y}{s_X}$ demonstrates the relationship between regression and correlation.

The relationship between regression and correlation can also be demonstrated by coefficient of determination. It can be calculated by squaring the coefficient of correlation.

In this model, r-squared is 0.7534. The model accounted for 75.34% of the variability in the data.

```
(cor(x = mydata$NEW, y = mydata$OLD))^2
```

```
## [1] 0.7533561
```

Question 2

```
mydata.LM <- lm(OLD~NEW,data = mydata)
# summary(mydata.LM)
anova(mydata.LM)
```

```
## Analysis of Variance Table
##
## Response: OLD
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## NEW        1 3047257 3047257  39.708 2.739e-05 ***
## Residuals 13  997652   76742
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The SSR is 3047257 on df = 1, SSE is 997652 on df = 13, SST = SSR + SSE = 4044909.

```
3047257 + 997652
```

```
## [1] 4044909
```

Question 3

MSE provides an estimate of the variance. MSE = SSE/df = 76742.

Question 4

```
mydata.LM <- lm(OLD~NEW,data = mydata)
summary(mydata.LM)
```

```
##
## Call:
## lm(formula = OLD ~ NEW, data = mydata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -652.41  -74.15   22.18   69.15  554.99
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -21.60568  101.69092  -0.212    0.835
## NEW           0.13457    0.02136   6.301 2.74e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 277 on 13 degrees of freedom
## Multiple R-squared:  0.7534, Adjusted R-squared:  0.7344
## F-statistic: 39.71 on 1 and 13 DF,  p-value: 2.739e-05
```

The F-statistic is 39.71 on 1 and 13 DF.

Question 5

The p-value is $2.74 * 10^{-5}$. If we assume that value of tuberculin antibodies level from new method is unrelated to value of tuberculin antibodies level from old method and we repeat the observation with sample 15 many many times, the probability of observing the model $y = 0.13x - 21.6$ or more extreme models is $2.74 * 10^{-5}$.

Question 6

Method 1. $r^2 = \frac{SSR}{SST}$

The r-squared is 0.7534.

```
3047257/4044909
```

```
## [1] 0.7533561
```

Method 2. r-squared can also be calculated by squaring the coefficient of correlation.

The r-suqared is 0.7534.

```
corr^2
```

```
## [1] 0.7533561
```

So that the two methods of calculation described in the notes are the equivalent

Question 7

```
predict(mydata.LM, data.frame(NEW = 200), interval = "confidence")
```

```
##        fit       lwr      upr
## 1 5.308973 -207.9207 218.5387
```

```
predict(mydata.LM, data.frame(NEW = 500), interval = "confidence")
```

```
##        fit      lwr      upr
## 1 45.68096 -158.258 249.6199
```

```
predict(mydata.LM, data.frame(NEW = 750), interval = "confidence")
```

```
##        fit       lwr      upr
## 1 79.32427 -117.2818 275.9303
```

For the results of 200, 500, 750 from new methods, the mean value of tuberculin antibodies level under the old methods are 5.31, 45.68, 79.32, respectively.

Question 8

1) We predict that the mean tuberculin antibodies concentration from old method is 5.31 with a 95% confidence interval of -207.9 to 218.5 if tuberculin antibodies concentration from new method is 200.
2) The mean tuberculin antibodies concentration from old method is 45.68 with a 95% confidence interval of -158.2 to 249.6 if tuberculin antibodies concentration from new method is 500.
3) The mean tuberculin antibodies concentration from old method is 79.32 with a 95% confidence interval of -117.3 to 275.9 if tuberculin antibodies concentration from new method is 750.

Question 9

1) If we get a tuberculin antibodies concentration of 200 from new method, we have 95% confidence that the limits of -207.9 and 218.5 contain the population mean tuberculin antibodies concentration from old method.
2) If we get a tuberculin antibodies concentration of 500 from new method, we have 95% confidence that the limits of -158.2 and 249.6 contain the population mean tuberculin antibodies concentration from old method.
3) If we get a tuberculin antibodies concentration of 750 from new method, we have 95% confidence that the limits of -117.3 and 275.9 contain the population mean tuberculin antibodies concentration from old method.

Question 10

```
if(!require(ggplot2)){install.packages("ggplot2")}
```

```
## Loading required package: ggplot2
```

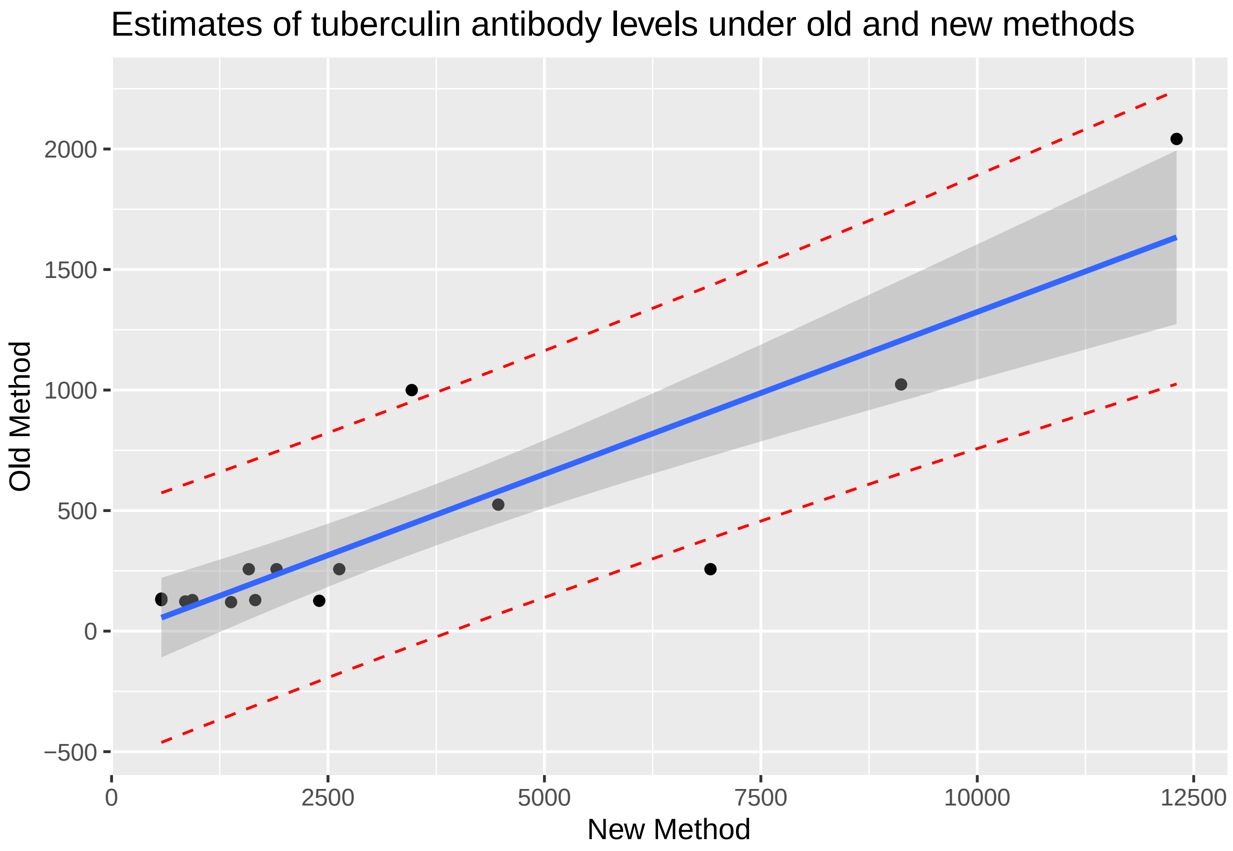
```
library(ggplot2)

mydata.PRED <- predict(mydata.LM, interval = "prediction",level=0.90)
```

```
## Warning in predict.lm(mydata.LM, interval = "prediction", level = 0.9): predictions on current data
```

```
mydata2 <- cbind(mydata, mydata.PRED)

G <- ggplot(data = mydata2, aes(y=OLD, x=NEW)) +
  geom_point() +
  stat_smooth(method = lm,level = 0.90) +
  geom_line(aes(y = lwr), color = "red", linetype = "dashed") +
  geom_line(aes(y = upr), color = "red", linetype = "dashed") +
  ylab("Old Method") +
  xlab("New Method") +
  ggtitle("Estimates of tuberculin antibody levels under old and new methods")
G
```


Estimates of tuberculin antibody levels under old and new methods
Old Method
New Method