Assignment ON3

Elmer V Villanueva

Due at 3:55 PM on Monday 16 March 2020

SET YOUR WORKING DIRECTORY!

```
setwd("C:/Users/1/Dropbox/00 - Working Folder/Teaching/DPH112/2019-2020/Week ON3")
```

Data

```
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, 1380.4,575.4)
TUBERCULIN <- data.frame(OLD, NEW)
str(TUBERCULIN)
```

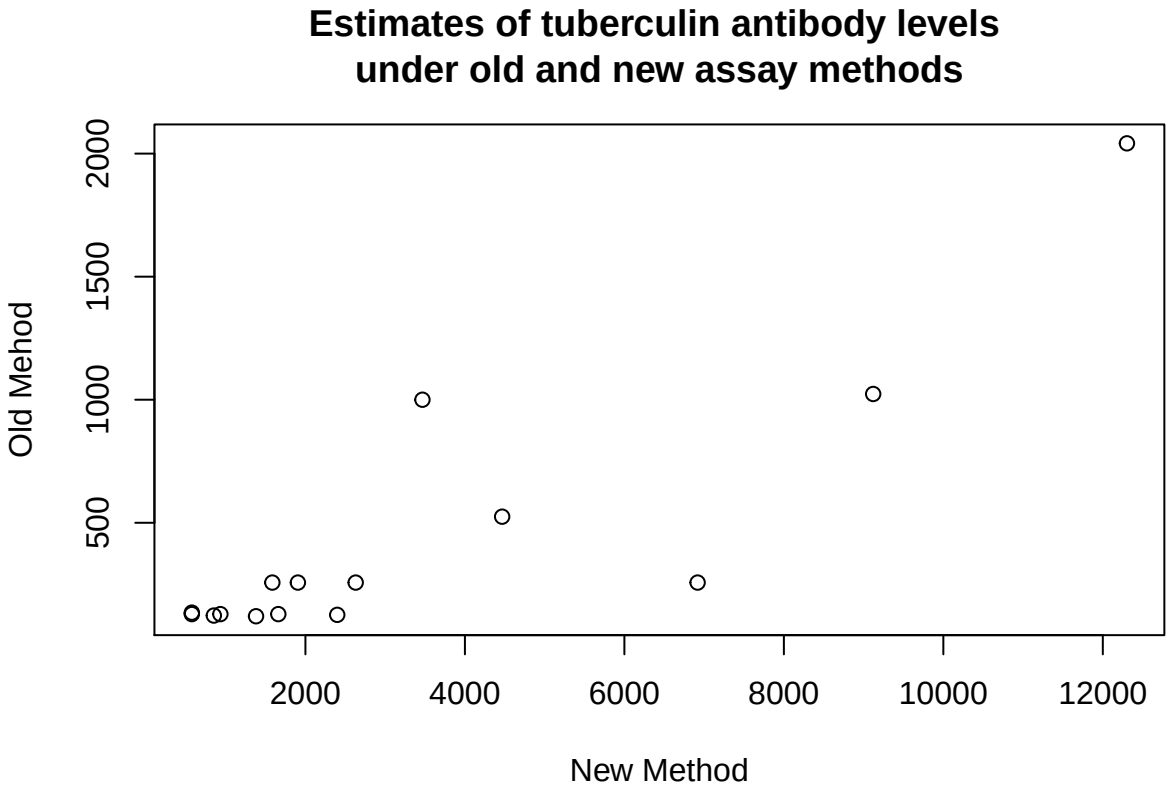
```
## 'data.frame':    15 obs. of  2 variables:
##  $ OLD: num  2042 257 525 257 129 ...
##  $ NEW: num  12303 6918 4467 1585 933 ...
```

```
head(TUBERCULIN)
```

```
##      OLD     NEW
## 1 2041.7 12302.7
## 2  257.0  6918.3
## 3  524.8  4466.8
## 4  257.0  1584.9
## 5  128.8   933.3
## 6  128.8  1659.6
```

Remember to visualise your data. ALWAYS DO THIS.

```
plot(OLD ~ NEW, data = TUBERCULIN,
     main = "Estimates of tuberculin antibody levels \nunder old and new assay methods",
     ylab = "Old Mehod",
     xlab = "New Method")
```



Questions

1. (10 marks) Confirm the relationship between correlation and simple linear regression as described in the notes. Calculate the slope estimate using the two methods described.

The two equations used to estimate the slope are

$$b_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

and

$$b_1 = r\frac{s_y}{s_x}$$

We can simplify the first equation by recognising that the numerator is derived from the covariance between x and y

$$Cov(x, y) = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n - 1}$$

and the denominator is derived from the variance of x

$$var(x) = \frac{\sum (x_i - \bar{x})^2}{n - 1}.$$

If you take the ratio of the two, the factor $(n - 1)$ is eliminated. Thus,

```
cov(OLD, NEW)
```

```
## [1] 1617418
```

```
var(NEW)
```

```
## [1] 12018864
```

```
cov(OLD, NEW)/var(NEW)
```

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

$$b_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{1617418}{12018864} = 0.13$$

The second equation requires the estimation of the standard deviations of x and y and the correlation between them.

```
cor(OLD, NEW)
```

```
## [1] 0.8679609
```

```
sd(OLD)
```

```
## [1] 537.5147
```

```
sd(NEW)
```

```
## [1] 3466.823
```

```
cor(OLD, NEW) * sd(OLD) / sd(NEW)
```

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

$$b_1 = r\frac{s_y}{s_x} = 0.8679609 \times \frac{537.5147}{3466.823} = 0.013$$

2. (10 marks) Regress the readings from the old method on the readings from the new method. Report the SST, SSR and SSE.

```
TUBERCULIN.LM <- lm(OLD ~ NEW, data = TUBERCULIN)
anova(TUBERCULIN.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 SST, SSR and SSE are 4,044,909, 3,047,257 and 997,652, respectively.

3. (5 marks) Estimate the variance of the model. The variance is estimated as 76,742.
4. (5 marks) Report the F statistic and its degrees of freedom. $F(1, 13) = 39.708$
5. (10 marks) Report the p-value and interpret it without recourse to statistical significance. The p-value is 2.74×10^{-5}. Under the null hypothesis, we would observe at least as extreme a result as reported in about once every 35,000 trials.
6. (10 marks) Calculate r^2 and confirm that the two methods of calculation described in the notes are the equivalent.

The two ways to calculate the coefficient of determination are (1) as the ratio between SSR and SST and (2) as the square of the coefficient of correlation.

$$r^2 = \frac{SSR}{SST} = \frac{3047257}{4044909} = 0.7534$$

$$r^2 = (r)^2 = (0.8679609)^2 = 0.7534$$

7. (6 marks) Estimate the mean value of tuberculin antibodies under the old method if the new method gives results of 200, 500 and 750.

```
predict(TUBERCULIN.LM, data.frame(NEW = c(200, 500, 750)))
```

```
##         1         2         3
##  5.308973 45.680955 79.324275
```

The mean values under the old method are expected to be 5.31, 45.68 and 79.32, respectively.

8. (6 marks) Construct an appropriate interval for each of your point estimates in (7) above.

The appropriate interval for the *mean* value is the confidence interval.

```
predict(TUBERCULIN.LM, data.frame(NEW = c(200, 500, 750)),
        interval = "confidence")
```

```
##         fit       lwr      upr
## 1  5.308973 -207.9207 218.5387
## 2 45.680955 -158.2580 249.6199
## 3 79.324275 -117.2818 275.9303
```

The estimated results are tabulated below:

Table. Estimated mean and 95% confidence interval tuberculin antibody levels under the old method for specific results under the new method.

Estimate under NEW method	Mean	95% confidence interval
200	5.31	(-207.92, 218.54)
500	45.68	(-158.26, 249.62)
750	79.32	(-117.28, 275.93)

9. (12 marks) Interpret each of the three intervals in (8) above.

We are 95% confident that the estimated mean value of tuberculin antibody under the old method is between -207.92 and 218.54 if the new method results in 200 ng/ml.

We are 95% confident that the estimated mean value of tuberculin antibody under the old method is between -158.26 and 249.62 if the new method results in 500 ng/ml.

We are 95% confident that the estimated mean value of tuberculin antibody under the old method is between -117.28 and 275.93 if the new method results in 700 ng/ml.

10. (10 marks) Construct a properly formatted scatterplot showing the individual data points, the regression line, and 90% confidence and prediction intervals for the data.

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

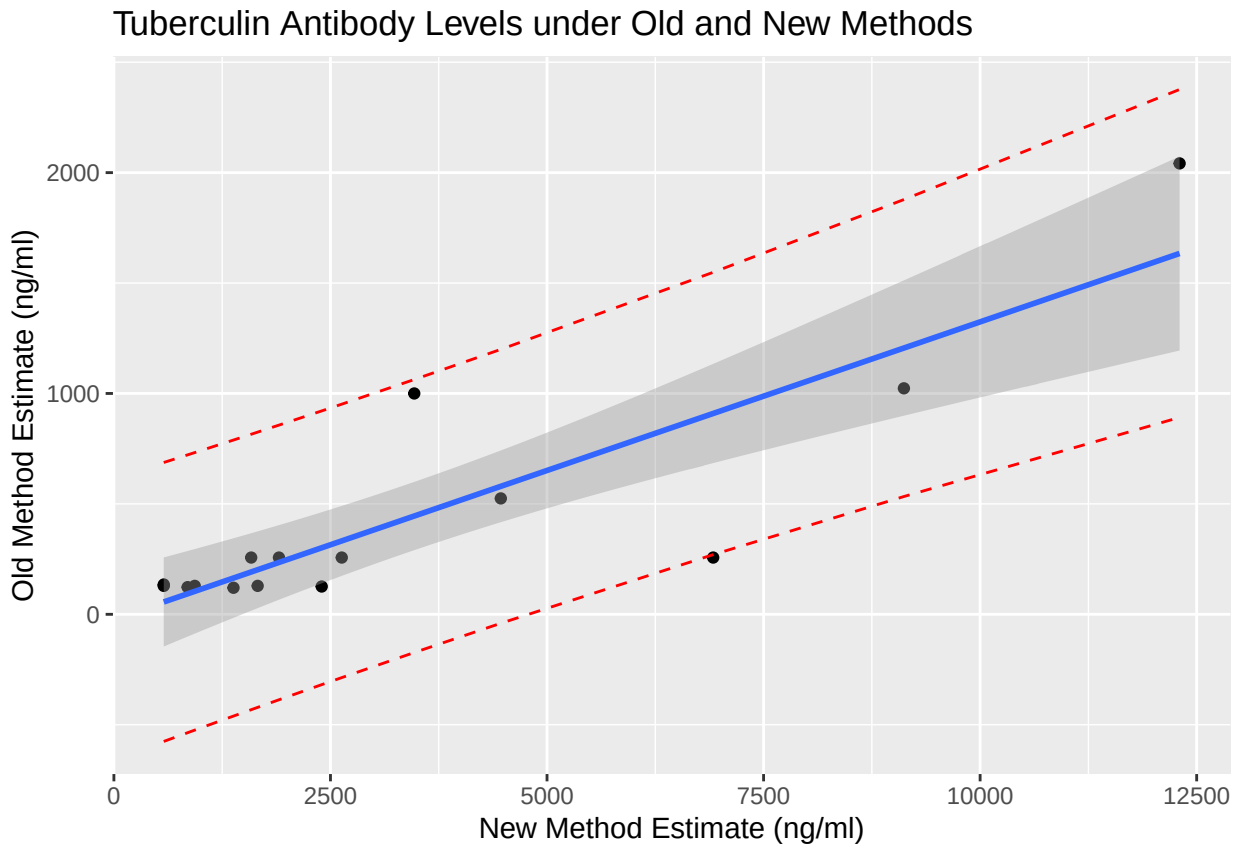
```
## Loading required package: ggplot2
```

```
library(ggplot2)

TUBERCULIN.PRED <- predict(TUBERCULIN.LM, interval = "prediction")
```

```
## Warning in predict.lm(TUBERCULIN.LM, interval = "prediction"): predictions on current data refer to
```

```r
TUBERCULIN2 <- cbind(TUBERCULIN, TUBERCULIN.PRED)

G <- ggplot(data = TUBERCULIN2, aes(y=OLD, x=NEW)) +
  geom_point(fill = 3) +
  stat_smooth(method = lm) +
  geom_line(aes(y = lwr), color = "red", linetype = "dashed") +
  geom_line(aes(y = upr), color = "red", linetype = "dashed")
G <- G + ggtitle("Tuberculin Antibody Levels under Old and New Methods") +
  ylab("Old Method Estimate (ng/ml)") +
  xlab("New Method Estimate (ng/ml)")
G
```

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



THE END