Week 1: Simple Regression, Training Exercise 1.3

Coursera/Erasumus U., Econometric Methods and Applications

Anthony Nguyen

Questions

Dataset `TrainExer13` contains the winning times (W) of the Olympic 100-meter finals (for men) from 1948 to 2004. The calendar years 1948-2004 are transformed to games (G) 1-15 to simplify computations. A simple regression model for the trend in winning times is $W_i = \alpha + \beta G_i + \varepsilon_i$.

(a) Compute a and b, and determine the values of R^2 and s.

(b) Are you confident on the predictive ability of this model? Motivate your answer.

(c) What prediction do you get for 2008, 2012, and 2016? Compare your predictions with the actual winning times.

Answers

(a) Compute a (intercept) and b (slope), and determine the values of R^2 and s.

$$b = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

```
#we can use the "Year" column for `x`.
xbar <- mean(TrainExer13$Game)
xbar
```

```
## [1] 8
```

```
ybar <- mean(TrainExer13$`Winning time men`)
ybar
```

```
## [1] 10.082
```

```
b <- sum((TrainExer13$Game - xbar)*(TrainExer13$`Winning time men` - ybar)) / sum((TrainExer13$Game - x

b
```

```
## [1] -0.038
```

$$a = \bar{y} - b\bar{x}$$

```
a <- ybar - b*xbar

a
```

```
## [1] 10.386
```

Next, to calculate R^2 and s, we can use the following formulas:

$$e_i = y_i - a - bx_i$$

R^2 = (sum of squares explained)/(sum of squares total), or:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} e^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

```
R2 <- 1 - (sum(e^2) / sum((TrainExer13$`Winning time men`- ybar)^2))
R2
```

```
## [1] 0.6733729
```

And finally for s, the *residual standard error*, we can use:

$$s^2 = \frac{1}{n-2}\sum_{i=1}^{n} e_i^2$$

```
s2 <- (1/(length(e)-2))*(sum(e^2))
s2
```

```
## [1] 0.01508615
```

```
s <- sqrt(s2)
s
```

```
## [1] 0.1228257
```

(b) Are you confident on the predictive ability of this model? Motivate your answer.

I would not be very confident about this model's ability to predict times, as I don't think that the year of the race is a good explanatory variable for faster winning times. Undoubtedly there are more factors that go into this.

(c) What prediction do you get for 2008, 2012, and 2016? Compare your predictions with the actual winning times.

Based our model, we can predict the winning times in subsequent olympic years using the forumula:

$$y = a + bx$$

```
y2008 <- a + b*(16)
y2012 <- a + b*(17)
y2016 <- a + b*(18)

y2008
```

```
## [1] 9.778
```

```
y2012
```

```
## [1] 9.74
```

```
y2016
```

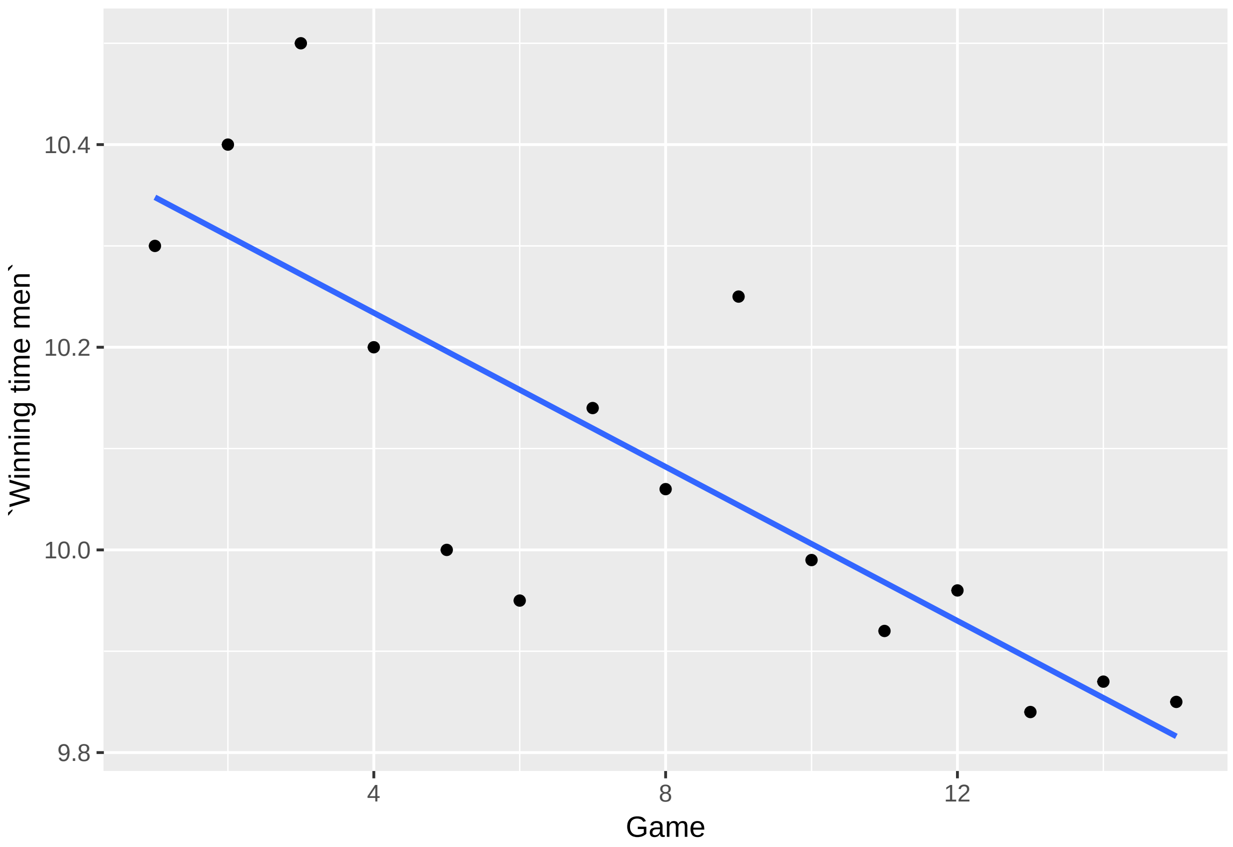
```
## [1] 9.702
```

A quick search on the internet, and we find that the actual results for the event in those years was:

2008: 9.69
2012: 9.63
2016: 9.81

Just to check, let's run the regression using the baseR `lm` function.

```
##
## Call:
## lm(formula = `Winning time men` ~ Game, data = TrainExer13)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -0.208 -0.048 -0.016  0.032  0.228
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 10.38600    0.06674 155.623  < 2e-16 ***
## Game        -0.03800    0.00734  -5.177 0.000178 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1228 on 13 degrees of freedom
## Multiple R-squared:  0.6734, Adjusted R-squared:  0.6482
## F-statistic:  26.8 on 1 and 13 DF,  p-value: 0.0001781
```


10.4
10.2
10.0
9.8
`Winning time men`
4
8
12
Game