Week 1: Simple Regression, Training Exercise 1.5

Coursera/Erasumus U., Econometric Methods and Applications

Anthony Nguyen

Questions

In Lecture 1.5, we applied simple regression for data on winning times on the Olympic 100 meter (athletics). We computed the regression coefficients a and b for two trend models, one with a linear trend and one with a nonlinear trend. In a test question, you created forecasts of the winning times for both men and women in 2008 and 2012. Of course, you can also forecast further ahead in the future. In fact, it is even possible to predict when men and women would run equally fast, if the current trends persist.

(a) Show that the linear trend model predicts equal winning times at around 2140.

(b) Show that the nonlinear trend model predicts equal winning times at around 2192.

(c) Show that the linear trend model predicts equal winning times of approximately 8.53 seconds.

(d) Comment on these outcomes and on the underlying regression models.

Answers

(a) Show that the linear trend model predicts equal winning times at around 2140.
For the sake of the exercise, we'll compute the linear regression "by hand", using the formulas:

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

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

First, we can solve for solve for b and a for the men:

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

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

Next, we repeat the steps to get b and a for the women:

```
## [1] -0.06292857
```

```
## [1] 11.6061
```

Using algebra, we can find what year the men's time and women's time will be equal by setting the two equations equal to one another and solivng for X where:

$$Y_{men} = 10.386 - 0.38G$$

$$Y_{women} = 11.606 - .063G$$

$$10.386 - .038G = 11.606 - .063G$$

Solving for G gives us the game number: **48.8**

To convert game number to year, we can use: $1948 + (G - 1) * 4 = Year$

Subbing in our G value, we get: $1948 + (48.8 - 1) * 4 = 2139.2$, or year 2140

(b) Show that the nonlinear trend model predicts equal winning times at around 2192.

For part (b), we can model the data along a non-linear trend using the following formula:

$$W_i = \gamma e^{\beta G_i}$$

Which we can re-write in linear form as:

$$log(W_i) = \alpha + \beta G + \varepsilon_i$$

(with G_i and $\alpha = log(\gamma)$)

To calculate these non-linear values, we need to first take the log of the win times for both men and women, and then we can deduce the new α and β values as before.

```
## [1] -0.003755949
```

```
## [1] 2.340604
```

```
## [1] -0.005612877
```

```
## [1] 2.451648
```

Re-written in linear form, our previous linear equations become:

$$Y_{men} = 2.341 - 0.0038G$$

$$Y_{women} = 2.452 - 0.0056G$$

$$2.341 - 0.0038G = 2.452 - 0.0056G$$

Solving for G here gives us the game number: **61.67**

To convert game number to year, we can use: $1948 + (G - 1) * 4 = Year$

Subbing in our G value, we get: $1948 + (61.67 - 1) * 4 = 2190.68$, or Olympic year 2192

(c) Show that the linear trend model predicts equal winning times of approximately 8.53 seconds.

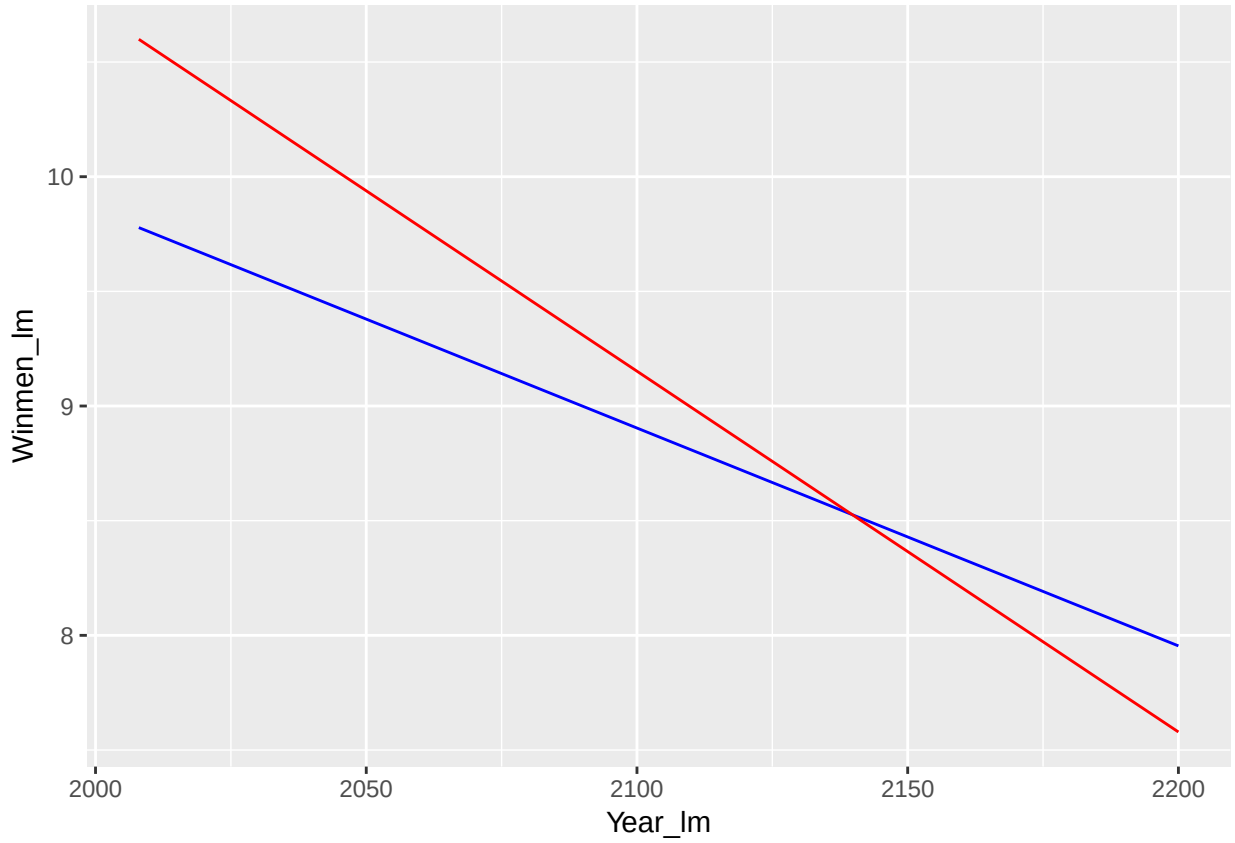
As we alread calculated the G value (48.8) in part A for when the times converge, we can just plug that back into the original equation to get the exact winning time that year.
$Y_i = a + bG$
$Y_i = 10.386 - 0.038 * 48.8$
$Y_i = 8.5316$

(d) Comment on these outcomes and on the underlying regression models.

The nonlinear model gives different predictions in the long-run.

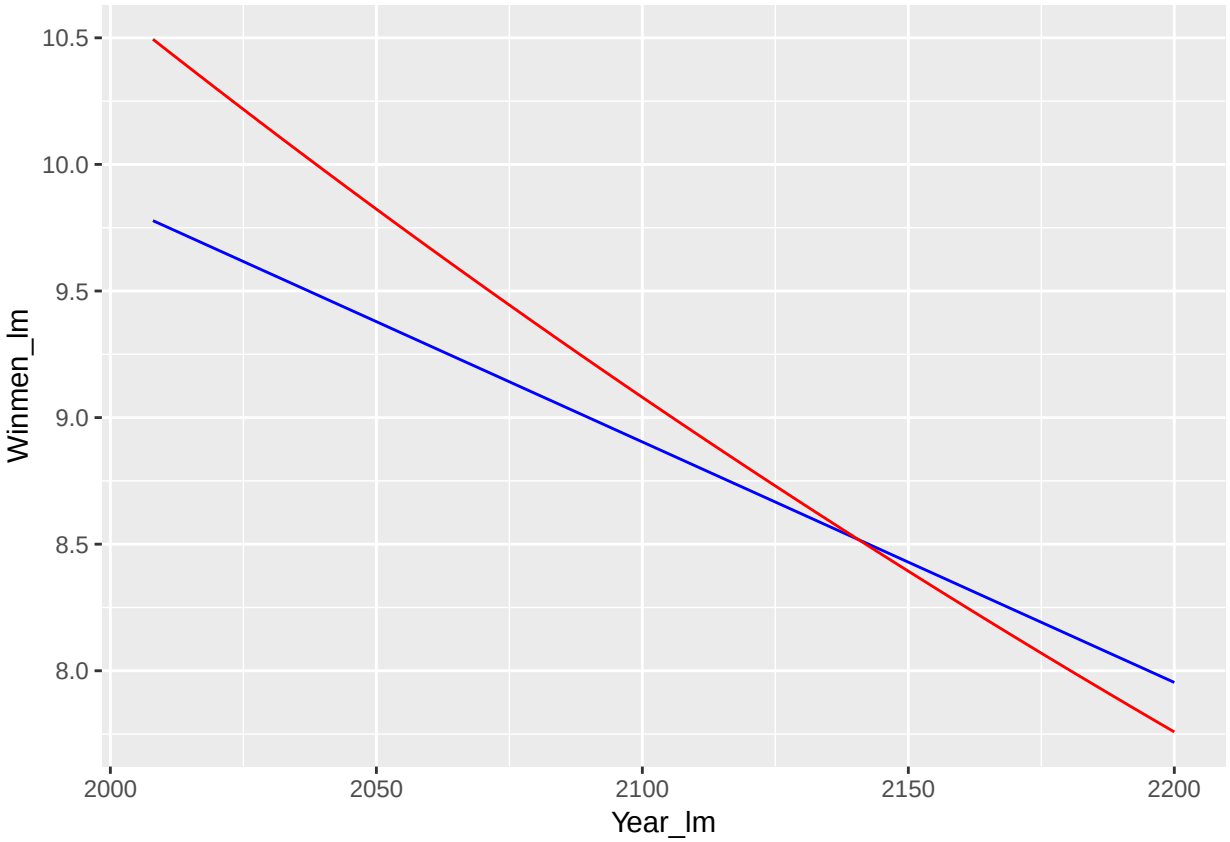
We can also construct a table with all of the predicted values from the time of our dataset (up to 2008), extending to 2200 to verify our results.

```
##    Game_lm Year_lm Winmen_lm Winwomen_lm   Time_diff
## 34      49    2140     8.524    8.522595 0.001404762
```

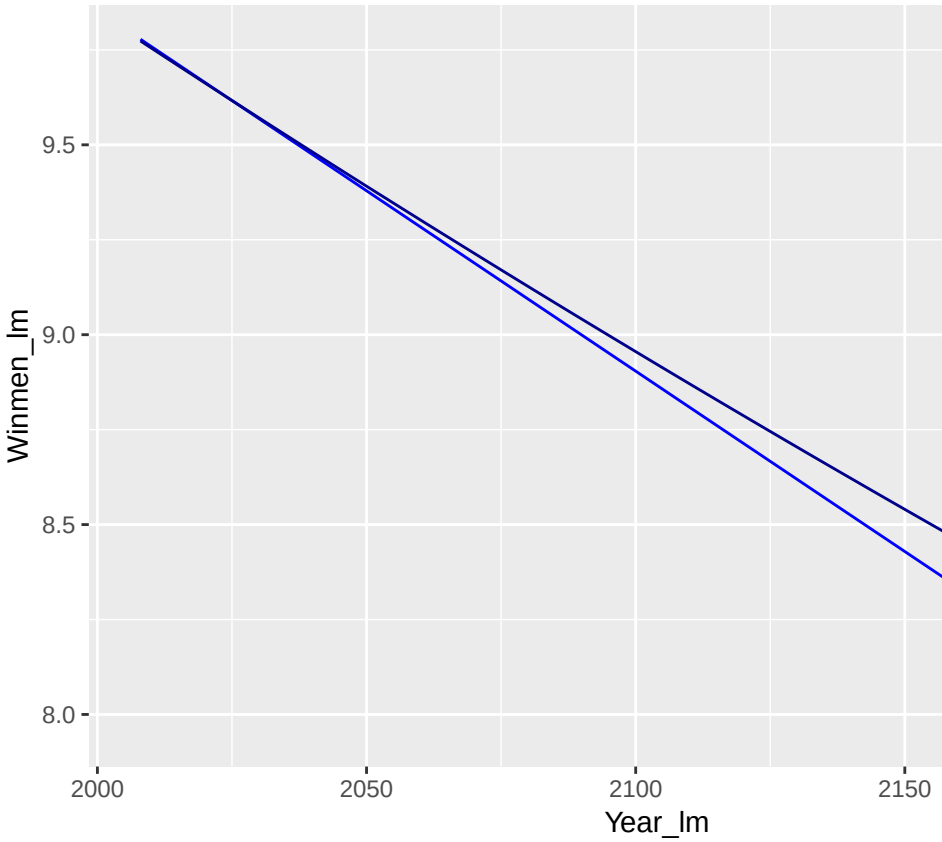


Finally, we can have a look to see which games will have closest times in the non-linear model.

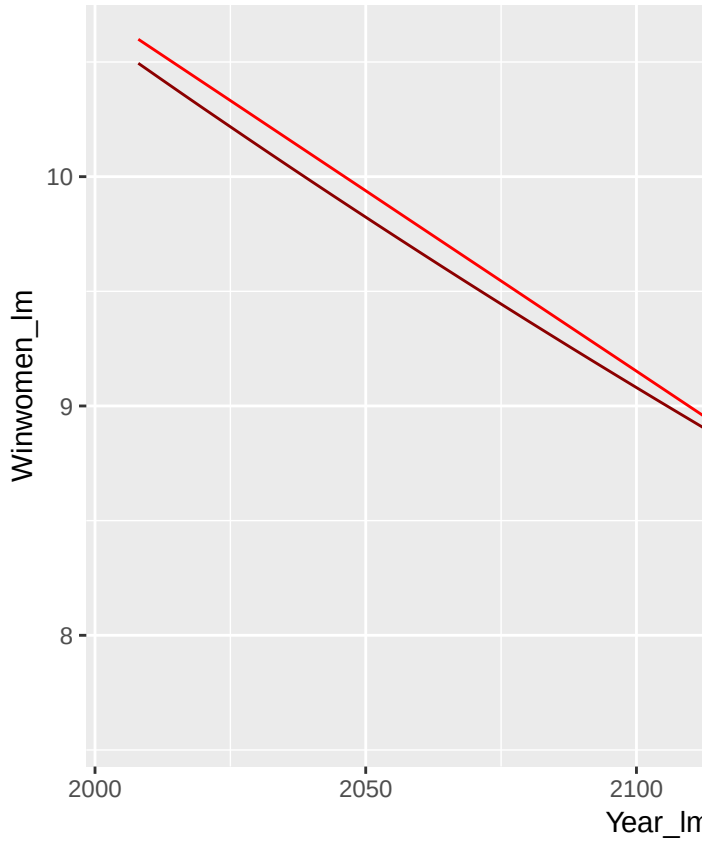
```
##    Game_lm Year_lm Winmen_lm Winwomen_lm  Time_diff Winmen_lm_log
## 30      45    2124     8.676     8.77431 0.09830952      8.753545
##    Winwomen_lm_log Time_diff_log
## 30        8.743861   0.009683984
```



We can also compare the lines for the standard lm model and the non-linear version by plotting them side by



side. Here is how the trend differs for men.



Here is the side-by-side comparison for the women's predictions.

And here are the numbers for the standard linear model, using baseR to calculate:

```
##
## Call:
## lm(formula = Winmen ~ Game, data = TrainExer15)
##
## 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
```

```r
lm_women <- lm(Winwomen~Game, data = TrainExer15)
summary(lm_women)
```

```
##
## Call:
## lm(formula = Winwomen ~ Game, data = TrainExer15)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.37388 -0.06327  0.01976  0.09562  0.35683
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 11.60610    0.11077 104.775  < 2e-16 ***
## Game        -0.06293    0.01218  -5.165 0.000182 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2039 on 13 degrees of freedom
## Multiple R-squared:  0.6724, Adjusted R-squared:  0.6472
## F-statistic: 26.68 on 1 and 13 DF,  p-value: 0.0001818
```