Running for the Gold

We will examine the relationship between year ($x$, the independent variable) and the winning time for the 100 meter dash ($y$, the dependent variable) during the modern Olympic era, 1896 to present. The data for 1896 - 1996 is shown below.

<table>
<thead>
<tr>
<th>YEAR</th>
<th>1896</th>
<th>1900</th>
<th>1904</th>
<th>1906</th>
<th>1908</th>
<th>1912</th>
<th>1920</th>
<th>1924</th>
<th>1928</th>
<th>1932</th>
<th>1936</th>
<th>1948</th>
<th>1952</th>
</tr>
</thead>
<tbody>
<tr>
<td>TIME</td>
<td>12.0</td>
<td>11.0</td>
<td>11.0</td>
<td>11.2</td>
<td>10.8</td>
<td>10.8</td>
<td>10.8</td>
<td>10.6</td>
<td>10.8</td>
<td>10.3</td>
<td>10.3</td>
<td>10.3</td>
<td>10.4</td>
</tr>
</tbody>
</table>

|------|------|------|------|------|------|------|------|------|------|------|------|

1. Below is a scatterplot of the data. Describe the relationship between winning times and year. What is the sign of the correlation coefficient for this data? **In general the gold medal winning times are getting faster over the years. Correlation is negative.**

The 5-number summary for this bivariate data are as follows:

$\bar{x} = 1945.92$, $s_x = 32.34$, $\bar{y} = 10.46$, $s_y = 0.513$, $r = -0.886$

**In Excel:** Data > Data Analysis > Descriptive Statistics;
correlation: Insert Function: `=correl(`

**In Excel:** put least squares line on scatterplot: click on scatterplot, in menu bar click Layout > Trendline > More Trendline Options;

$y = -0.014x + 37.81$  
$R^2 = 0.784$
2. Write the equation of the least squares prediction line:
\[
\hat{y} = 37.81 - 0.014x \quad (b_1 = \frac{s_y}{s_x}; \ b_0 = \hat{y} - b_1 \bar{x})
\]
interpret slope: For each additional year the gold medal winning time decreases by 0.014 seconds.

3. Summarize the assumptions that are made when using the linear model
\[
y = \beta_0 + \beta_1 x + \epsilon
\]
to model the relationship between \(x\) and \(y\). \(\epsilon_i \sim iid \ N(0, \sigma)\) for all \(x\)

In Excel: least squares line output: \(Data > Data Analysis > Regression;\)

SUMMARY OUTPUT

Regression Statistics
Multiple R 0.885826655
R Square 0.784688863
Adjusted R Square 0.774901994
Standard Error 0.243511037
Observations 24

ANOVA
\[
\begin{array}{lcccr}
\text{df} & \text{SS} & \text{MS} & \text{F} & \text{Significance F} \\
\text{Regression} & 1 & 4.754348085 & 4.754348085 & 80.17771523 & 8.65637E-09 \\
\text{Residual} & 22 & 1.304547748 & 0.059297625 & & \\
\text{Total} & 23 & 6.058895833 & & & \\
\end{array}
\]

Coefficients  Standard Error    t Stat     P-value            Lower 95%    Upper 95%    Lower 90.0%    Upper 90.0%
Intercept 37.81707888   3.055301418 12.377   2.18804E-11    31.4807         44.1533        32.5706           43.0634
yr -0.014057        0.001569901 -8.954    8.65637E-09    -0.0173         -0.0108  -0.01675            -0.0113

4. Is the slope \(\beta_1\) significantly different from 0? (that is, is "year", the independent variable \(x\), useful for predicting \(y\), the winning time in the olympic 100m dash?)

\(H_0 : \beta_1 = 0 \quad H_A : \beta_1 \neq 0\)

Test Statistic: \(t = \frac{b_1}{SE(b_1)} = -8.954 \text{ from output}\)

P-value=\(P(t_{22} \leq -8.954) + P(t_{22} \geq 8.954) = 0.00000000865637 \text{ from output.}\)
CONCLUSION: reject \(H_0\) and conclude that the slope is not 0; slope is negative.

5. What is a 95% confidence interval for the unknown value of the slope \(\beta_1\)?

\(b_1 \pm t_{.025,22} \cdot SE(b_1) \quad \text{where} \ \ SE(b_1) = \frac{s_\epsilon}{\sqrt{n-1 \cdot s_x}} = 0.001569 \ldots \text{ from output}\)

\((-0.0173, -0.0108) \text{ from output.}\)

6. What is the sum of the squares of the residuals?

\[
\sum_{i=1}^{24} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{24} y_i^2 - b_0 \sum_{i=1}^{24} y_i - b_1 \sum_{i=1}^{24} x_i y_i = 1.3045 \ldots \text{ from output}\]

7. a. Predict the winning time in the 100 meter dash in the 2008 Olympics.

TREND FUNCTION: \(=\text{trend}(\)
\[\hat{y}_{2008} = b_0 + b_1 (2008) = 37.81 - 0.014(2008) = 9.698 \text{ seconds}\]
b. Calculate a 95% confidence interval for the mean winning time $\mu_{2008}$ in the 2008 Olympics. $SE(\hat{\mu}_{2008}) = \sqrt{SE^2(b_1)*(1945.92-2008)^2 + \frac{s^2}{24}}$; $t_{22}^* = 2.0739$

$$\hat{y} \pm 2.0739 \sqrt{0.00157^2*(1945.92-2008)^2 + \frac{0.0593}{24}}$$

$$= 9.698 \pm 2.0739*0.1094 = 9.698 \pm .227 = (9.471, 9.925)$$

c. Calculate a 95% prediction interval for the winning time $y_{2008}$ of an individual runner in the 2008 Olympics.

$$SE(\hat{y}_{2008}) = \sqrt{SE^2(b_1)*(1945.92-2008)^2 + \frac{s^2}{24} + s^2_{e}}$$

$$\hat{y} \pm 2.0739 \sqrt{0.00157^2*(1945.92-2008)^2 + \frac{0.0593}{24} + 0.0593}$$

$$= 9.698 \pm 2.0739*0.2670 = 9.698 \pm .554 = (9.144, 10.252)$$

8. Shown below is a plot of the residuals against the independent variable $x$ (year). Do you notice any patterns? The outlier at upper left is 1896. Starting in the 1970's, all the residuals are positive, which means that in later years the least squares line is predicting winning times that are too fast. The 1896 point in the scatterplot is pulling the left end of the least squares line UP, thus pushing DOWN the line at the right end. The decreasing trend of the least squares line cannot continue indefinitely because we reach the physical limits of human ability to run fast. Should probably throw out the 1896 point.
9. Shown below is a normal probability plot of the residuals. Are the assumptions in question 3 reasonable? **Normal probability plot looks pretty good; high point at right corresponds to 1896. Normal assumption for residuals looks reasonable.**