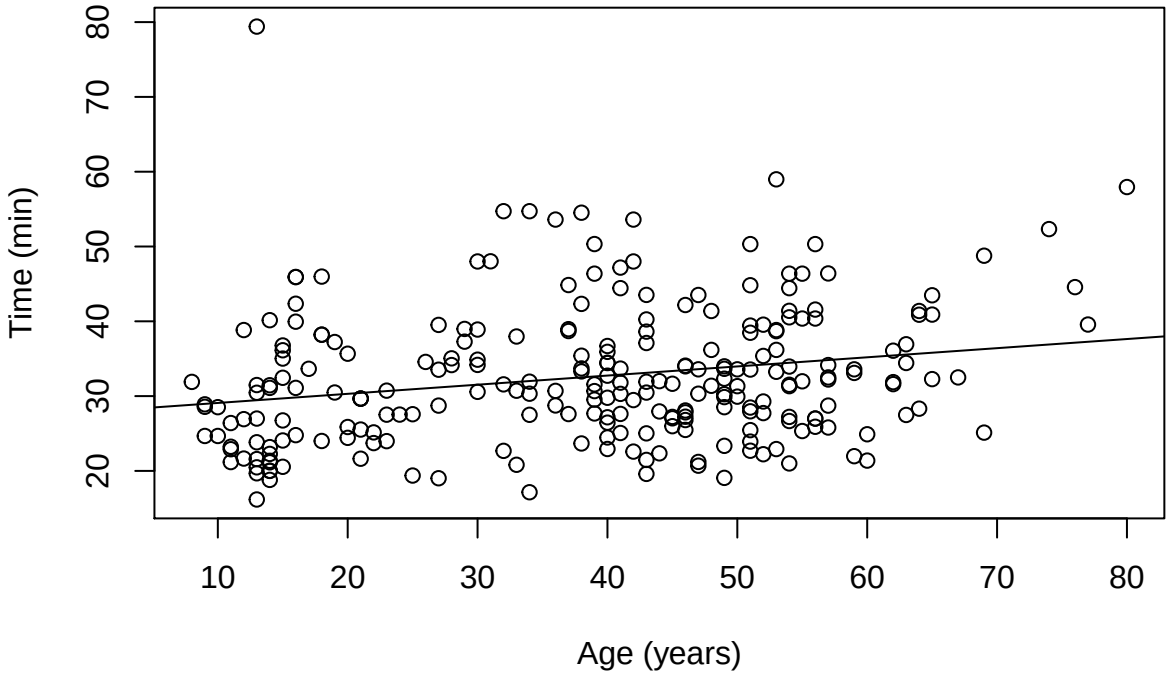
5K Running Times vs. Age

Math 222 - Workshop



The graph above compares results for 248 runners in a 5k race that took place in California in 2013 with their ages. The average race time was 32.55 minutes with a standard deviation of 9.07. The average age of the runners was 38.4 with a standard deviation of 16.89. The correlation between age and race time was 0.2274.

1. Find a formula for the least squares regression line in this example.

2. Make an ANOVA table for this example.

3. What is the residual standard error?

5. On average, older runners take longer to finish the race. Make a 95% confidence interval for the slope of the regression line using the standard error for slope $SE_{b_1} = \frac{s}{s_x\sqrt{n-1}}$.

6. Make a 95% confidence interval for the average time it would take 20 year old runners to complete the race. Use the standard error for average y-values $SE_{\hat{\mu}} = s\sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{s_x^2(n-1)}}$.

7. Are there any outliers in the data? Would removing these outliers make the association more or less significant?

Here is the output of the `summary()` function applied to this linear model:

```
##
## Call:
## lm(formula = Time ~ Age, data = myData)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -14.866  -6.162  -1.216   5.326  49.947
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 27.86561    1.39798  19.933  < 2e-16 ***
## Age          0.12208    0.03333   3.663 0.000306 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.85 on 246 degrees of freedom
## Multiple R-squared:  0.05171,    Adjusted R-squared:  0.04786
## F-statistic: 13.41 on 1 and 246 DF,  p-value: 0.0003057
```

8. As you can see, the F-statistic has a very low p-value. What does that mean?

9. What are the assumptions that we need to make in order to safely use these inference techniques (confidence intervals, F-test for association). How would you check each one?