# Assignment 8

Instructions (the same as for Assignment 1): Make an R Notebook and in it answer question below. When you are done, hand in on Quercus the *output* from Previewing (or Knitting) your Notebook. Do *not* hand in the Notebook itself. You want to show that you can (i) write code that will answer the questions, (ii) run that code and get some sensible output, (iii) write some words that show you know what is going on and that reflect your conclusions about the data. Your goal is to convince the grader that you *understand* what you are doing: not only doing the right thing, but making it clear that you know *why* it's the right thing.

Do *not* expect to get help on this assignment. The purpose of the assignments is for you to see how much *you* have understood. You will find that you also learn something from grappling with the assignments. The time to get help is after you watch the lectures and work through the problems from PASIAS, via tutorial and the discussion board. The only reason to contact the instructor while working on the assignments is to report something missing like a data file that cannot be read.

You have 4 hours to complete this assignment after you start it.

My solutions to this assignment, with extra discussion, will be available after everyone has handed in their assignment.

1. If you are driving, and you hit the brakes, how far do you travel before coming to a complete stop? Presumably this depends on how fast you are going. Knowing this relationship is important in setting speed limits on roads. For example, on a very bendy road, the speed limit needs to be low, because you cannot see very far ahead, and there could be something just out of sight that you need to stop for.

   Data were collected for a typical car and driver, as shown in http://ritsokiguess.site/STAC32/stopping.csv. These are American data, so the speeds are miles per hour and the stopping distances are in feet.

   (a) Read in and display (probably all of) the data.

   **Solution:**

   The usual:

```
my_url <- "http://ritsokiguess.site/STAC32/stopping.csv"
stopping <- read_csv(my_url)
```

```
## Parsed with column specification:
## cols(
##   speed = col_double(),
##   distance = col_double()
## )
```

```
stopping
```

```
## # A tibble: 8 x 2
##   speed distance
##   <dbl>    <dbl>
## 1     0        0
## 2    10       20
## 3    20       50
```

```
## 4    30        95
## 5    40       150
## 6    50       220
## 7    60       300
## 8    70       400
```

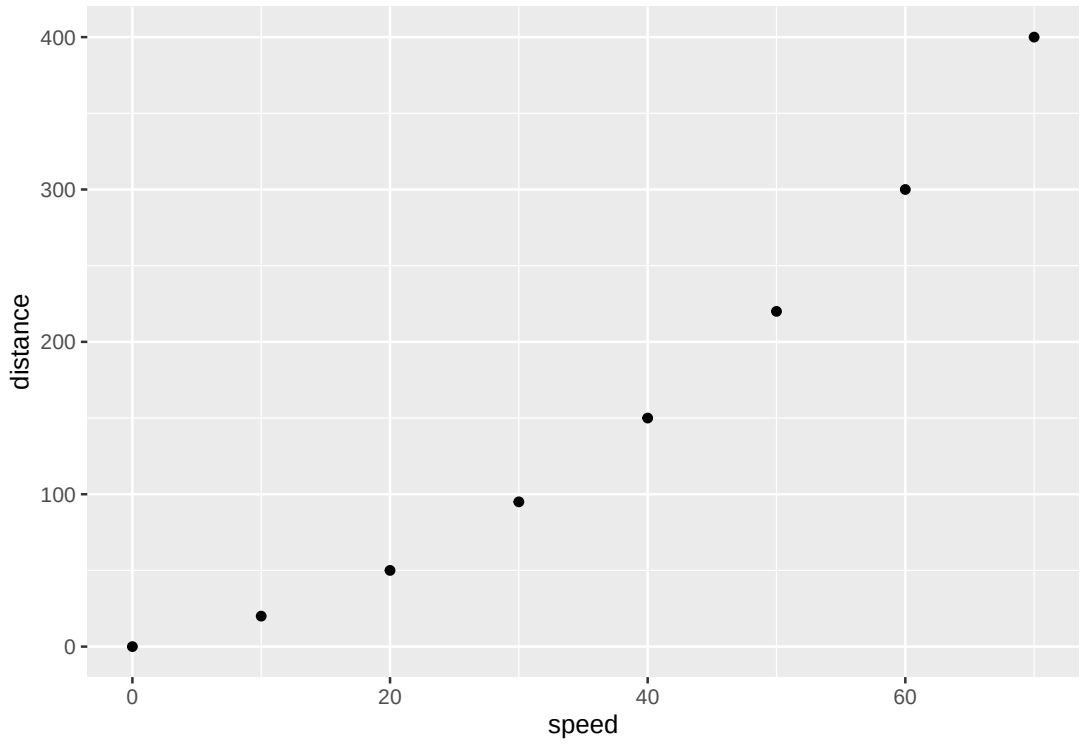
There are only eight observations.

(b) Make a suitable plot of the data.

**Solution:**

Two quantitative variables means a scatterplot. Stopping distance is the outcome, so that goes on the $y$-axis:

```
ggplot(stopping, aes(x=speed, y=distance)) + geom_point()
```




(c) Describe any trend you see in your graph.

**Solution:**

It's an upward trend, but not linear: the stopping distance seems to increase faster at higher speeds.

(d) Fit a linear regression predicting stopping distance from speed. (You might have some misgivings about doing this, but do it anyway.)

**Solution:**

Having observed a curved relationship, it seems odd to fit a straight line. But we are going to do it anyway and then critique what we have:

```
stopping.1 <- lm(distance~speed, data=stopping)
summary(stopping.1)
```

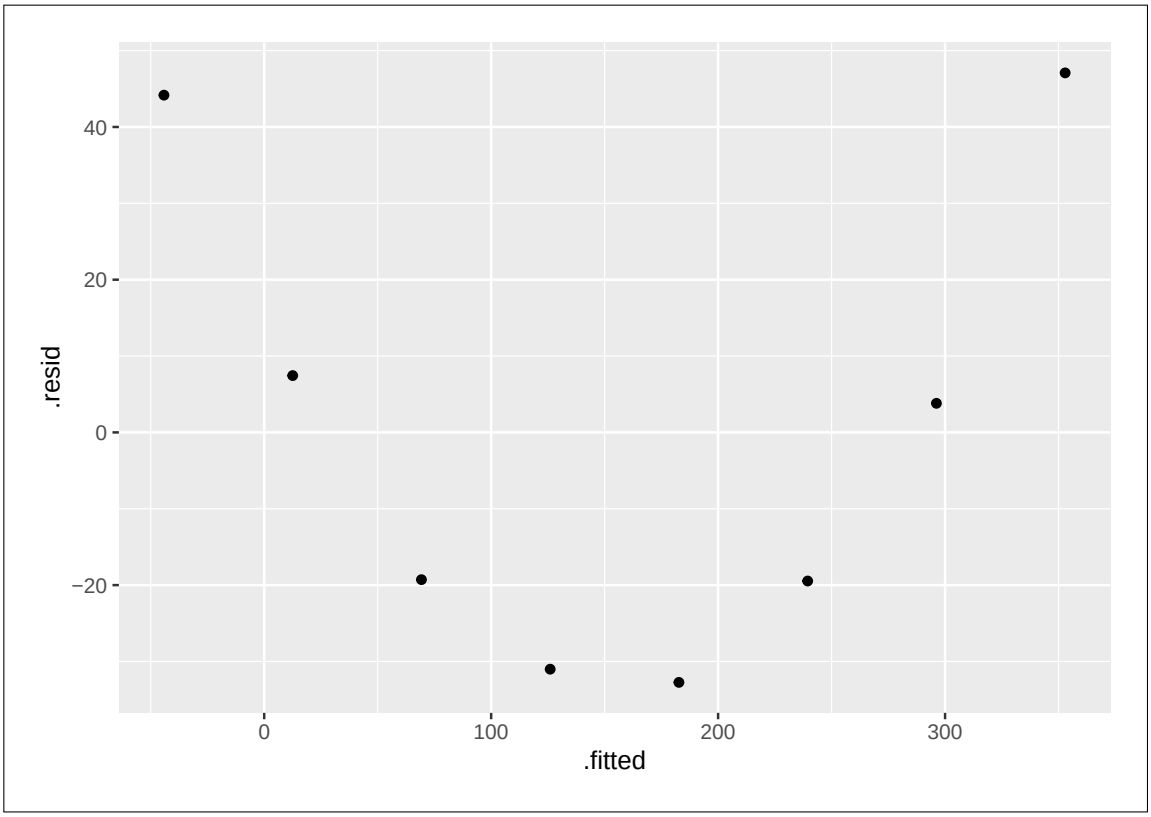
```
##
## Call:
## lm(formula = distance ~ speed, data = stopping)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -32.738 -22.351  -7.738  16.622  47.083
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -44.1667    22.0821   -2.00   0.0924 .
## speed         5.6726     0.5279   10.75 3.84e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 34.21 on 6 degrees of freedom
## Multiple R-squared:  0.9506, Adjusted R-squared:  0.9424
## F-statistic: 115.5 on 1 and 6 DF,  p-value: 3.837e-05
```

Extra: note that R-squared is actually really high. We come back to that later.

(e) Plot the residuals against the fitted values for this regression.

**Solution:**

```
ggplot(stopping.1, aes(x=.fitted, y=.resid)) + geom_point()
```




(f) What do you learn from the residual plot? Does that surprise you? Explain briefly.

**Solution:**

The points on the residual plot form a (perfect) curve, so the original relationship was a curve. This is exactly what we saw on the scatterplot, so to me at least, this is no surprise.

(Make sure you say *how you know* that the original relationship was a curve from looking at the residual plot. Joined-up thinking. There are *two* ways we know that the relationship is a curve. Get them both.)

(g) What is the actual relationship between stopping distance and speed, according to the physics? See if you can find out. Cite any books or websites that you use: that is, include a link to a website, or give enough information about a book that the grader could find it.

**Solution:**

I searched for "braking distance and speed" and found the first two things below, that seemed to be useful. Later, I was thinking about the fourth point (which came out of my head) and while searching for other things about that, I found the third thing:

- a British road safety website, that says "The braking distance depends on how fast the vehicle was travelling before the brakes were applied, and is proportional to the square of the initial speed."
- the Wikipedia article on braking distance, which gives the actual formula. This is the velocity squared, divided by a constant that depends on the coefficient of friction. (That

is why your driving instructor tells you to leave a bigger gap behind the vehicle in front if it is raining, and an even bigger gap if it is icy.)
- an Australian math booklet that talks specifically about braking distance and derives the formula (and the other equations of motion).
- also, if you have done physics, you might remember the equation of motion $v^2 = u^2 + 2as$, where $u$ is the initial velocity, $v$ is the final velocity, $a$ is the acceleration and $s$ is the distance covered. In this case, $v = 0$ (the car is stationary at the end), and so $-u^2/2a = s$. The acceleration is negative (the car is slowing down), so the left side is, despite appearances, positive. There seems to be a standard assumption that deceleration due to braking is constant (the same for all speeds), at least if you are trying to stop a car in a hurry.

These are all saying that we should add a speed-squared term to our regression, and then we will have the relationship exactly right, according to the physics.

Extra: Another way to measure how far you are behind the vehicle in front is time. Many of the British "motorways" (think 400-series highways) were built when I was young, and I remember a TV commercial that said "Only a Fool Breaks the Two Second Rule".[1] In those days (the linked one is from the 1970s[2]), a lot of British drivers were not used to going that fast, or on roads that straight, so this was a way to know how big a gap to leave, so that you had time to take evasive action if needed. The value of the two-second rule is that it works for any speed, and you don't have to remember a bunch of stopping distances. (When I did my theory test, I think I worked out and learned a formula for the stopping distances that I could calculate in my head. I didn't have to get very close since the test was multiple-choice.)

(h) Fit the relationship that your research indicated (in the previous part) and display the results. Comment briefly on the R-squared value.

**Solution:**

Add a squared term in `speed`:

```
stopping.2 <- lm(distance~speed+I(speed^2), data=stopping)
summary(stopping.2)
```

```
##
## Call:
## lm(formula = distance ~ speed + I(speed^2), data = stopping)
##
## Residuals:
##       1       2       3       4       5       6       7       8
## -1.04167  0.98214  0.08929  1.27976 -0.44643 -0.08929 -2.64881  1.87500
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.041667   1.429997   0.728    0.499
## speed       1.151786   0.095433  12.069 6.89e-05 ***
## I(speed^2)  0.064583   0.001311  49.267 6.51e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.699 on 5 degrees of freedom
```

```
## Multiple R-squared:  0.9999, Adjusted R-squared:  0.9999
## F-statistic: 2.462e+04 on 2 and 5 DF,  p-value: 1.039e-10
```

The R-squared now is basically 1, so that the model fits very close to perfectly.

Extra: you probably found in your research that the distance should be just something times speed squared, with no constant or linear term. Here, though, we have a significant linear term as well. That is probably just chance, since the distances in the data look as if they have been rounded off. With more accurate values, I think the linear term would have been closer to zero.

If you want to go literally for the something-times-speed-squared, you can do that. This doesn't quite work:

```
stopping.3x <- lm(distance~I(speed^2), data=stopping)
summary(stopping.3x)
```

```
##
## Call:
## lm(formula = distance ~ I(speed^2), data = stopping)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.7327  -3.4670   0.6761   6.2323   8.4513
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 14.732704   4.362859   3.377   0.0149 *
## I(speed^2)   0.079796   0.001805  44.218 8.96e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.514 on 6 degrees of freedom
## Multiple R-squared:  0.9969, Adjusted R-squared:  0.9964
## F-statistic:  1955 on 1 and 6 DF,  p-value: 8.958e-09
```

because it still has an intercept in it. In R, the intercept is denoted by `1`. It is always included, unless you explicitly remove it. Some odd things start to happen if you remove the intercept, so it is not a good thing to do unless you know what you are doing. The answers here have some good discussion. Having decided that you *are* going to remove the intercept, you can remove it the same way as anything else (see `update` in the multiple regression lecture) with "minus". I haven't shown you this, so if you do it, you will need to cite your source: that is, say where you learned what to do:

```
stopping.3 <- lm(distance~I(speed^2)-1, data=stopping)
summary(stopping.3)
```

```
##
## Call:
## lm(formula = distance ~ I(speed^2) - 1, data = stopping)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.6123  -0.7859  10.5314  15.5314  19.2141
##
## Coefficients:
```

```
##            Estimate Std. Error t value Pr(>|t|)
## I(speed^2) 0.084207   0.001963   42.89 9.77e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.42 on 7 degrees of freedom
## Multiple R-squared:  0.9962, Adjusted R-squared:  0.9957
## F-statistic:  1840 on 1 and 7 DF,  p-value: 9.772e-10
```

The R-squared is still extremely high, much higher than for the straight line. The coefficient value, as I said earlier (citing Wikipedia), depends on the coefficient of friction; the stopping distances you see typically are based on a dry road, so you have to allow extra distance (or time: see above) if the road is not dry.

(i) Somebody says to you "if you have a regression with a high R-squared, like 95%, there is no need to look for a better model." How do you respond to this? Explain briefly.

**Solution:**

An example of a regression with an R-squared of 95% is the straight-line fit from earlier in this question. This is an example of a regression that fits well but is not appropriate because it doesn't capture the form of the relationship.

In general, we are saying that no matter how high R-squared is, we might still be able to improve on the model we have. The flip side is that we might not be able to do any better (with another data set) than an R-squared of, say, 30%, because there is a lot of variability that is, as best as we can assess it, random and not explainable by anything.

Using R-squared as a measure of absolute model quality is, thus, a mistake. Or, to say it perhaps more clearly, asking "how high does R-squared have to be to indicate a good fit?" is asking the wrong question. The right thing to do is to concentrate on getting the form of the model right, and thereby get the R-squared as high as we can for that data set (which might be very high, as here, or not high at all).

# Notes

1. This is perhaps not a commercial so much as a public safety message.

2. There are some typical British cars of the era in the commercial.