STATG006: Solutions to Exercise Sheet #4

The exercises in this sheet focus on linear regression. As before, we provide solutions, sometimes detailed, sometimes a sketch that should point you to the complete solution. Skeetches should not be taken at face value as the level of detail required for an exam answer.

1. Each p-value corresponds to testing the null hypothesis that each contributing covariate has no (linear) relationship with the sales outcome given the others. So for `Intercept`, `TV`, and `radio`, the small p-value means that there is evidence that these contributed to the behaviour of sales, while `newpaper` advertising did not (given advertising in other media). It cannot be concluded that spending money on newspaper advertising is useless, only that it seems not to contribute to sales when TV and radio are being used. It also does not mean that changing the way newspaper funds are used would continue not to show an effect on sales, only that the way done in the data does not seem to work.

2. For (a), a change of x_3 from 0 to 1 means a change of expectation from $50 + 20x_1 + 0.07x_2 + 35 \times 0 + 0.01x_1x_2 - 10x_1 \times 0 = 50 + 20x_1 + 0.07x_2 + 0.01x_1x_2$ to $50+20x_1+0.07x_2+35\times 1+0.01x_1x_2-10x_1\times 1 = 50+10x_1+0.07x_2+35+0.01x_1x_2$. So the difference in expectation between females and males is $35 - 10x_1$. If GPA is high enough (the scale is between 1 and 4), this will be negative, so the correct answer is iii.

   For (b), we plug-in the corresponding values and get 137100 dollars.

   For (c): this is of course false, as a small value is not the same as non-significant value (and vice-versa!). Why?

3. (a): the cubic regression has more freedom that the linear representation, which is a special case of the cubic representation (it is equivalent to the case where $\beta_2 = \beta_3 = 0$). So barring the very special case where the relation between Y and X is deterministic (and linear), the cubic variant should approximate the data better, with the consequent reduction of RSS.

   (b) This is different, as the test set is independent of the parameter estimates constructed out of the training data. It might be possible we have **overfitting** in the training set. Without further information, we cannot tell.

   (c) The explanation is identical to item (a) (except that in this case a linear deterministic relationship is ruled out).

(d) The explanation is similar to (b): even if the linear representation is not adequate, if variability is high and sample size is small, it is still possible that the cubic representation overfits the data and does worse at prediction time than the linear one.

4. Just write (using ISLR's notation of y_i, while in our slides we typically use $y^{(i)}$):

$$\hat{y}_i = x_i \frac{\sum_{i'=1}^{n} x_{i'} y_{i'}}{\sum_{i''=1}^{n} x_{i''}^2} = \sum_{i'=1}^{n} \left(\frac{x_i x_{i'}}{\sum_{i''=1}^{n} x_{i''}^2} \right) y_{i'}$$

from which the result follows. Linear regression is a specific case of a *linear smoother*, a method that reweights the training points in order to generate fitted/prediction values. The weights are not functions of the training outcomes $\{y_{i'}\}$, so the function is linear in training y even if the weights are non-linear functions of the training/test covariates (it is linear in the testing covariates, in this case - but in general the weights could be interpreted as linear functions of non-linear transformations of the inputs). One of the reasons for the difficulties/successes of neural networks is exactly the fact that they are not linear smoothers: the learn a non-linear mapping of the training outputs to the parameters (with the price of requiring computationally demanding fitting algorithms and not well-understood properties).

5. (a): By doing the algebra (which I will not work out the details, as I assume you will find it a matter of mechanical manipulation), we get $Cor(X, Y) = a/\sqrt{a^2}$. Now the tricky point would be to realize that this is equal to either 1 (if $a > 0$) or -1 (if $a < 0$). So correlations always lie in the interval $[-1, 1]$, with the extremes correponding to perfect linearly deterministic relationships of different or equal sign for the multiplicative coefficient a.

(b): The following is not the most straightforward algebraic manipulation. Starting by the definition of R^2 under the simplified assumption $\bar{x} = \bar{y} = 0$, we have

$$R^2 = \frac{\sum_i y_i^2 - \sum_i (y_i - \hat{y}_i)^2}{\sum_i y_i^2}$$

Now let's look at the numerator. First, it can be written as $2\sum_i y_i \hat{y}_i - \sum_i \hat{y}_i^2$. By plugging in

$$\hat{y}_i = \sum_{i'} a_{i'} y_{i'},$$

and

$$a_{i'} = \frac{x_i x_{i'}}{\sum_{i''} x_{i''}^2},$$

we get

$$2\sum_i y_i \hat{y}_i - \sum_i \hat{y}_i^2 = \frac{2}{\sum_{i''} x_{i''}^2} \sum_i \sum_{i'} x_i x_i' y_i y_{i'} - \frac{1}{(\sum_{i''} x_{i''}^2)} \sum_i \left(\sum_{i'} x_i x_{i'} y_{i'} \right)^2$$

Notice that the first term in the expression above can be written as

$$\begin{aligned}\frac{2}{\sum_{i''} x_{i''}^2}\sum_i\sum_{i'} x_i x_i' y_i y_{i'} &= \frac{2}{\sum_{i''} x_{i''}^2}\left(\sum_i x_i y_i\right)\left(\sum_{i'} x_{i'} y_{i'}\right)\\ &= \frac{2}{\sum_{i''} x_{i''}^2}\left(\sum_i x_i y_i\right)^2\end{aligned}$$

as symbols i and i' are exchangeable.

Let's have a closer look into the other abomination term found in the numerator:

$$\begin{aligned}\frac{1}{(\sum_{i''} x_{i''}^2)^2}\sum_i\left(\sum_{i'} x_i x_{i'} y_{i'}\right)^2 &= \frac{1}{(\sum_{i''} x_{i''}^2)^2}\sum_i x_i^2\left(\sum_{i'} x_{i'} y_{i'}\right)^2\\ &= \frac{1}{(\sum_{i''} x_{i''}^2)^2}(\sum_i x_i^2)\left(\sum_{i'} x_{i'} y_{i'}\right)^2\\ &= \frac{1}{(\sum_{i''} x_{i''}^2)}\left(\sum_i x_i y_i\right)^2\end{aligned}$$

where, once more, symbols i, i' and i'' are all exchangeable across equivalent expressions.

This means the numerator of R^2 boils down to

$$2\frac{1}{(\sum_{i''} x_{i''}^2)}\left(\sum_i x_i y_i\right)^2 - \frac{1}{(\sum_{i''} x_{i''}^2)}\left(\sum_i x_i y_i\right)^2 = \frac{1}{(\sum_{i''} x_{i''}^2)}\left(\sum_i x_i y_i\right)^2$$

and so R^2 is the same as the squared zero-average correlation coefficient

$$\frac{(\sum_i x_i y_i)^2}{(\sum_i x_i^2)(\sum_i y_i^2)}.$$

6. (a) The following will read the data (assuming it is reachable from whatever directory it is in), with the extra care of setting correctly which entries are missing:

```
> auto <- read.table("Auto.data", header = TRUE, na.strings = "?")
```

We then fit a linear regression model (an intercept is added automatically), and print its summary:

```
> auto_model <- lm(mpg ~ horsepower, data = auto)
> summary(auto_model)

Call:
lm(formula = mpg   horsepower, data = auto)
```

```
Residuals:
Min 1Q Median 3Q Max
-13.5710 -3.2592 -0.3435 2.7630 16.9240

Coefficients:
Estimate Std.  Error t value Pr(>|t|)
(Intercept) 39.935861 0.717499 55.66 <2e-16 ***
horsepower -0.157845 0.006446 -24.49 <2e-16 ***
---
Signif.  codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.'  0.1 ' ' 1

Residual standard error:  4.906 on 390 degrees of freedom
(5 observations deleted due to missingness)
Multiple R-squared:  0.6059, Adjusted R-squared:  0.6049
F-statistic:  599.7 on 1 and 390 DF, p-value:  < 2.2e-16
```

By the p-value of the F test ($H_0 : \beta_1 = 0$) the coefficient of horsepower is not zero (since there is only one covariate, this could be seen from the p-value of the t-test for the coefficient directly, the final line in the "`Coefficients`" section.

Concerning strenght of relationship. In terms of statistical evidence, the standard error is very small (assuming approximate Gaussianity, the confidence interval would be highly concentrated around the estimated value). In terms of practical significance, a R^2 of 0.60 is good news. To interpret whether the magnitude is 0.16 is relevant, we must understand what the units of the measurements are. In an informal, incomplete way (that uses no knowledge of the meaning of the measurements), plotting a histogram of horsepower and of mpg reveals that 0.15 units of mpg per 1 unit of horsepower is a non-trivial change, as the scale of the former is reasonably smaller than the scale of the latter.

A negative relation is clearly supported, at least in the linear approximation of the relationship. At a horsepower of 98, the estimated expected output is $39.9 - 0.15 \times 98 = 24.5$ miles per gallon. Getting confidence intervals (and prediction intervals) requires a more complicated formula we did not describe in detail in class. That is because the confidence interval is given by a function of the variance of $\hat{Y}^\star$, the expected outcome at $x^\star$. That is,

$$Var(\hat{Y}^\star) = Var(\hat{\beta}_0 + \hat{\beta}_1 x^\star).$$

The randomness, remember, is in $\hat{\beta}_0$ and $\hat{\beta}_1$. These two variables are **dependent**, as they came from the same training data. We will see later in Chapter 6 how we could calculate the variance of a function of multiple variables, but the formula itself is not necessary to understanding the meaning of this variance. Using function `predict` from R, we can calculate confidence intervals for the *expected* $\hat{Y}^\star$ as follows:

```
> predict(auto_model, data.frame(horsepower = 98), interval = "confidence")
```

```
fit lwr upr
1 24.46708 23.97308 24.96108
```

That is, with probability 95% (the default coverage in this function), the interval $[23.97, 24.96]$ will contain the *expected* miles per gallon achieved under cars with a horsepower of 98. On the other hand,

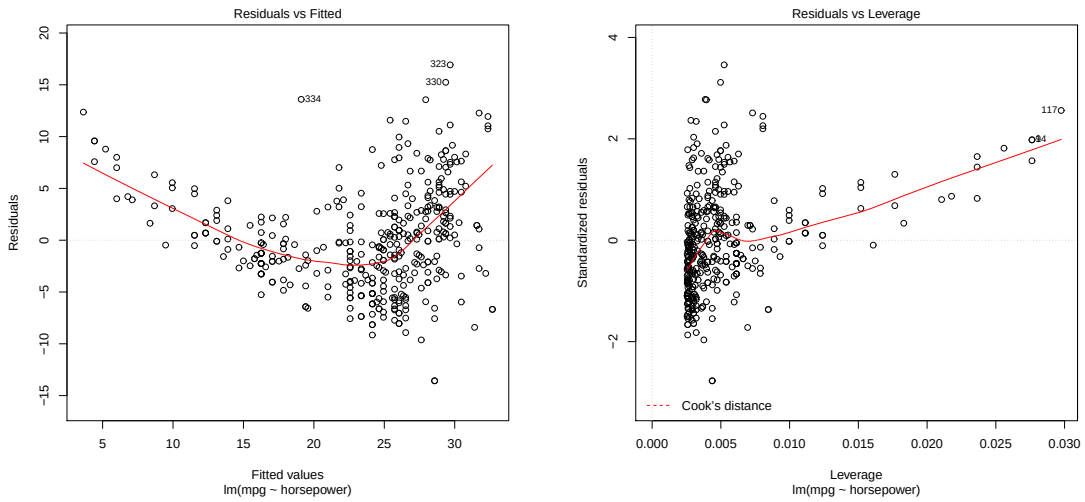
```
> predict(auto_model, data.frame(horsepower = 98), interval = "prediction")
fit lwr upr
1 24.46708 14.8094 34.12476
```

meas that, with probability 95%, the interval $[14.81, 34.12]$ will contain the miles per gallon achieved under a particular car with a horsepower of 98. Check that this makes sense: the residual standard error was reported as 4.906, and a 95% interval is obtained by an approximate distance of 1.96 standard deviations. So the standard deviation of the expected value estimate is $(24.96108 - 23.97308)/(2 * 1.96)$ which is approximately 0.252 (we can get this also from function `predict`, see its documentation). We can then verify that it is indeed true that $34.12 \approx 24.47 + 1.96 \times \sqrt{4.906^2 + 0.252^2}$).

(b) We can plot these two very easily with

`> plot(auto$horsepower, auto$mpg)` and verify the clear non-linearity there. Using `abline(auto_model)` will add a line that will emphasize the misfit at the low and high levels of horsepower. However, it is not unreasonable to say that for a range of values of horsepower ("middle range values"), the linear approximation is not too bad. Can you complement this analysis by fitting two separate linear models, one for horsepower less than 120, and one for horsepower greater than 120?

(c) When we run `plot(auto_model)`, we get plots like the ones below.



Notice that, in the residual vs. fitted plot there are strong dependencies between the two. The non-linear pattern is evident, and in particular outliers like point 334 stand out even more (that point is easy to spot in the plot of item (b)). Concerning leverage, there are several points of high leverage/high residual, which is evidence of trouble. Complement this analysis by fitting a linear model without some of the points of high leverage/high residual. What do you expect to see?

7. See `ex4.R`. Further interpretation is left to you.

8. See `ex4.R`. Further interpretation is left to you.