Statistics 305a
Homework 1, due Thursday, September 24, 2020 by 12pm.

This problem set is partly aimed at getting you going with R. The coursework webpage points you to useful sources for learning R; the first few chapters of Dalgaard "Introductory Statistics with R" or of Venables and Ripley would also be helpful. All questions with a bold **R** *next to the question number can be solved in groups up to size three. For such groups a single writeup can be turned in, but make sure to indicate who the three are. All datasets used in the homework assignments can be found on canvas in the directory* `Datasets`.

1. Read chapters 1 and 2 of Weisberg. This should be revision for most of you, and will be a quick read/browse. This will be helpful for at least one of the questions below.

2. Assume that $y_i = \beta_0 + \beta_1 x_i + \epsilon_i,\ i = 1, \ldots, n$, and assume the x_i are fixed (non-random), and that the ϵ_i are uncorrelated, each having mean 0 and variance σ^2. Let $\tilde{x}_i = (x_i - \bar{x})/s_x,\ i = 1, \ldots, n$ be the standardized versions of the x_i, where $\bar{x}$ is the sample mean, and s_x is the sample standard deviation of the x_i (scaled by $1/n$ rather than $1/(n-1)$). Suppose we now fit a simple linear regression model of y_i on $\tilde{x}_i$ by least squares.

   (a) Explain the "scaled by ..." comment in parenthesis above.

   (b) What are the least squares estimates $\hat{\alpha}_0$ and $\hat{\alpha}_1$ for the intercept and slope (in their simplest form) for these transformed $\tilde{x}_i$.

   (c) Show the relationship between this slope estimate and the sample correlation coefficient between y_i and x_i.

   (d) What are the sampling variances of each of these estimates, and their sampling covariance?

   (e) Can you use these estimates to obtain LS estimates for the linear regression model with x_i and y_i instead of $\tilde{x}_i$ and y_i? How?

   (f) Show how you can convert the variances for the former to the latter.

3. *Regression through the origin.* Occasionally, a model in which the intercept is known apriori to be zero might more appropriate. This model is

$$y_i = \beta_1 x_i + \varepsilon_i, \quad i = 1, \ldots, n.$$

We assume that the errors are i.i.d. $(0, \sigma^2)$.

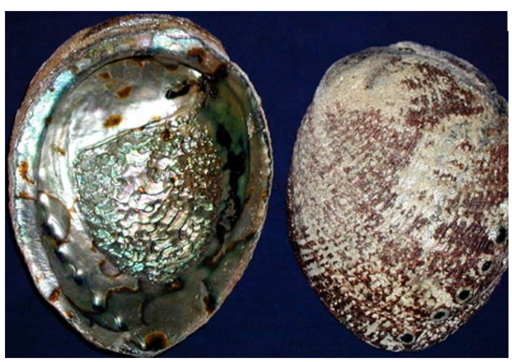
   (a) Derive the least squares estimate of β_1. Show that it is unbiased for β_1. What is its variance? Derive an expression for $\hat{\sigma}^2$. How many degrees of freedom does it have?

   (b) How would you test whether $\beta_1 = 0$? Give all the details.

4. **R** Write an R function to fit a simple linear regression with or without intercept. The function should thus start:

```
linear.regression <- function(x, y, intercept=TRUE)
...
```

This says the default is that there is an intercept, but allows for the user to enter `FALSE` instead. Have the function return as named components the coefficients, their estimated covariance matrix, the fitted values, the residuals, and the R-squared statistic. Use the ability of R to perform simple matrix and vector operations when writing your function; i.e. no `for()` loops! You may not use built-in functions such as `lm` or `lmfit`. Aim for elegance—only the most elegant solution will earn full points.



5. **R** The datafile `abalone_305.csv` can be found on the class web page in the `Datasets` section. The variable `Rings` is age of the animal, measured by cutting its shell, and counting rings (like tree-ring dating, and just as tedious). The variable `Length` is the length of the animal, and is thought to be a good surrogate for age.

(a) Use your R function in (3) to fit the no-intercept regression model to the data (X=Length), to produce $\hat{\beta}_1$, $\hat{\sigma}^2$, and a 95% confidence interval for β_1.

(b) Plot the data (making sure the point $(0,0)$ is in the plot - `xlim` and `ylim`). Include your fitted line (`abline`), and using dotted lines show the confidence interval for the slope.

(c) Include the fitted line when an intercept is included in the linear model.

(d) Include as well a 95% prediction interval for $\widehat{\text{Rings}}$ (hint: read those chapters). Explain what this is. In light of what you see, comment about the validity of the assumptions used in this model.

6. **R** Bootstrap confidence intervals — (e.g. Efron and Tibshirani, "An Introduction to the Bootstrap", Chapman and Hall, 1993).

We use the pivotal "t-statistic"

$$T = \frac{\hat{\beta} - \beta}{\hat{\text{s.e.}}(\hat{\beta})}$$

to obtain confidence intervals for a coefficient β in linear regression (simple or multiple). In doing so we assume the x_i are fixed, the linear model is correct, and that the errors $\epsilon_i,\ i = 1, \ldots, n$ are i.i.d $N(0, \sigma^2)$. If all this holds, T has a t-distribution and we can use its quantiles to make confidence statements. Here β is the true (unknown) coefficient for a linear model fit to the population F, and $\hat{\beta}$ the estimate from the *Empirical* distribution $\hat{F}$ (i.e. the sample).

In reality none of these assumptions need be reasonable, but we still may want to fit a linear model, and get confidence intervals for the parameters. The bootstrap method avoids some of these assumptions, and tries to *estimate* the sampling distribution of T. We proceed as follows. A bootstrap sample is obtained by drawing a random sample of size n from $\hat{F}$, the empirical distribution of our data (with replacement). Each such bootstrap sample gives us a "new" training sample, to which the same model can be fit. The bth such sample produces a bootstrap realization

$$T_b^* = \frac{\hat{\beta}_b^* - \hat{\beta}}{\hat{\text{s.e.}}(\hat{\beta}_b^*)}.$$

This is done B times (e.g. $B = 1000$), and the quantiles of the bootstrap distribution T^* are used in place of those of the t-distribution.

Write an R-function to compute a 95% bootstrap confidence interval for the slope β for the abalone data. Redo your plot in the previous question, showing just the line through the origin, and the original confidence interval. Include your bootstrap interval (using a different color). What did you learn here? In what way does the bootstrap distribution differ from the original t-distribution?

You may not use bootstrap packages — i.e. you must do this from scratch. There are several ways to do this elegantly in R. Only elegant solutions will receive full marks