BIOSTATISTICS 767 (Spring 2019)
Homework 1

1. The table below gives summary statistics for lung function measurements in random samples from three groups of subjects; non-smokers, light smokers and heavy smoker. Let Y_{ij} denote the lung function measure for the j-th subject in the i-th group. Assume that $Y_{ij} \sim \mathrm{N}(\mu_i, \sigma^2), i = 1, 2, 3, j = 1, 2, \cdots, n_i$.

i	Group	n_i	Mean	SD
1	Non-smoker	21	3.78	1.79
2	Light smoker	21	3.23	1.86
3	Heavy smoker	21	2.59	1.82

Show your calculations. If you use a computer, simply state that and describe **briefly** what you did using common statistical language. Do **not** give any computer code. Note: The sample standard deviation (SD) above was computed using the formula with $n - 1$ in the denominator.

(a) Carry out an ANOVA test of the hypothesis that the lung function means (theoretical population means) in the three groups are all equal. State your hypothesis clearly. Show your calculations and provide the ANOVA table, the test statistic and any associated quantities, and compute the p-value.

(b) Use least-squares to estimate parameters in the model

$$\mu_i = \alpha_1 + \alpha_2 i.$$

Provide the estimates $(\hat{\alpha}_1, \hat{\alpha}_2, \hat{\sigma}^2)$. Also provide standard error estimates for $\hat{\alpha}_1$ and $\hat{\alpha}_2$.

(c) Consider the model:

$$\mu_i = \beta_1 + \beta_2 i + \beta_3 i^2.$$

Test the hypothesis $H_0 : (\beta_2, \beta_3) = (0, 0)$ against its complement. Report the test statistic and any associated quantities, and compute the p-value. (Note: H_0 states that both β_2 and β_3 are zero. It is sometimes written as $H_0 : \beta_2 = \beta_3 = 0$).

2. (Practice with linear combinations) Consider the usual simple regression model $\mathrm{E}[Y_i] = \beta_1 + \beta_2 x_i$, with outcome vector $Y = (Y_1, Y_2, Y_3, Y_4)^\top$ and covariate x given as: $x_1 = 1, x_2 = 1, x_3 = 2, x_4 = 4$. It is also assumed that the 4 outcomes are uncorrelated and $\mathrm{var}(Y_1) = \mathrm{var}(Y_2) = \mathrm{var}(Y_3) = \mathrm{var}(Y_4) = \sigma^2$ (unknown). The model parameters will be estimated by ordinary least squares (OLS).

Do this problem entirely by hand using only a simple calculator. Do not use a computer and do not use matrix operations.

(a) Express $\hat{\beta}_1$ explicitly as a linear function of Y. That is, write $\hat{\beta}_1$ explicitly as $\hat{\beta}_1 = a_1 Y_1 + a_2 Y_2 + a_3 Y_3 + a_4 Y_4$, giving the a_i's numerical values (to 3 significant digits).

(b) Given the information: $\mathrm{E}[Y_1] = \mathrm{E}[Y_2] = 1, \mathrm{E}[Y_3] = 2, \mathrm{E}[Y_4] = 8$, compute the expected value of $\hat{\beta}_1$.

(c) Suppose that Y_1, Y_2, Y_3, Y_4 are uncorrelated (i.e. $\mathrm{corr}(Y_i, Y_j) = 0, i \neq j$), and $\mathrm{var}(Y_1) = 1, \mathrm{var}(Y_2) = 2, \mathrm{var}(Y_3) = 4, \mathrm{var}(Y_4) = 5$. Compute the variance of $\hat{\beta}_1$.

(d) Repeat the last two parts for $\hat{\beta}_2$.

(e) Using the linear representations above, compute $\mathrm{cov}(\hat{\beta}_1, \hat{\beta}_2)$.

(f) Define the *fitted values* $\hat{\mu}_i = \hat{\beta}_1 + \hat{\beta}_2 x_i$, and the residuals $R_i = Y_i - \hat{\mu}_i$. Express $\hat{\mu}_3$ as a linear function of Y. Compute the expected value of $\hat{\mu}_3$.

(g) Express R_3 as a linear function of Y. Compute $\mathrm{E}[R_3]$, $\mathrm{var}(R_3)$ and $\mathrm{E}[R_3^2]$.