

---

**Problem 1.** Linear Regression, Prediction and Inference

Date file: TermLifeData.csv

Variable	description
MARSTAT	marital status 0 = other, 1 = married, 2 = living together
EDUCATION	years of education
NUMHH	number of household members
INCOME	annual income
FACE	amount of insurance purchased

We wish to investigate factors driving demand for term life insurance based on this data set. Let's focus on those who have purchased at least \$50,000 of term life insurance.

[a] Note that FACE is highly skewed. Apply the optimal box-cox transformation to FACE using *sklearn.preprocessing.PowerTransformer()*. Please do NOT standardize the transformed output (i.e. use `standardize = False` )

Present the before- and after- transformation histograms in a single side-by-side chart similar to the one in the lecture side. Please display the corresponding skewness (to 3 decimal places) and the optimal  $\lambda$  in the chart.

[b] Let FACEbc denote the transformed FACE as in [a]. Fit the following regression model

$$\text{FACEbc} = \beta_0 + \beta_1 \text{EDUCATION} + \beta_2 \text{NUMHH} + \beta_3 \ln(\text{INCOME}) + \beta_4 \text{MAR0} + \beta_5 \text{MAR2} + \epsilon$$

where  $\text{MAR}k$  is the dummy variable for  $\text{MARSTAT} = k, k = 0, 1, 2$ .

Let your script report  $\hat{\beta}$ ,  $\hat{\sigma}$  and  $R^2$ .

[c] Test the hypothesis (at the 5% significance level) that marital status is not an important driver for term life insurance demand. That is  $H_0 : \beta_4 = \beta_5 = 0$ .

Let your script report the value of the test statistic, the corresponding critical value and whether  $H_0$  is rejected or not.

[d] Predict the amount of insurance that would be purchased by a married man, with a BS in Actuarial Science degree (16 years of education), a wife and two kids, and earning \$120,000 a year.

Let your code report the predicted amount of insurance (in \$) and the 95% prediction interval.

You may find the following resources useful.

Power Transform

<https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.PowerTransformer.html>

Skew

<https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.skew.html>

F distribution

<https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.f.html>

t distribution

<https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.t.html>