

STAC67 Homework 1: Due Friday, September 20th

Total possible points: 70

Name:

Student Number :

1. The goal of this exercise is to determine the factors that influenced income, in dollars, in 1971. We consider dataset **Prestige**. It contains several predictor variables that researchers want to use to predict the expected income level for a particular earner in the 70s. In particular, we are interested in the following variables:

- **income** Average income of incumbents, dollars, in 1971.
- **education** Average education of occupational incumbents, years, in 1971.
- **women** Percentage of incumbents who are women.
- **prestige** Pineo-Porter prestige score for occupation, from a social survey conducted in the mid1960s

To download the data set, first install the `car` package by using the function `install.packages('car')`. To open the library use the the command `library('car')`, and then call the `Prestige` dataset. **Solutions without R-code will be assigned a mark of zero.**

- (a) (5 points) Carry out the simple linear regression in R for the predictor variable *prestige*. Write down the fitted regression line.

- (b) (2 points) Give the value of the estimated mean income with a Pineo-Porter prestige score of 46.83.

- (c) (3 points) Interpret the intercept and slope terms in the practical context of the data being analysed.
- (d) (3 points) What is the coefficient of determination for this model (i.e R^2). What does this say about the relationship between X and Y ?
- (e) (3 points) Based on the fitted regression model, the data considered, and your answer to part (d), would you say that it is worth it for researchers to use the Pineo-Porter prestige scores to track the income of incumbents in 1971? Explain.
- (f) (4 points) In the dataset there is one observation for which incumbent had a Pineo-Porter score of 68.8, with an annual income of 12,351 dollars. Compute the predicted estimate for a Pineo-Porter score of 68.8. Has the model underpredicted or overpredicted the Pineo-Porter score relative to what is observed? What is this difference called? Should this difference be 0? Explain.

2. The least squares estimator of $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$ can be expressed as a linear function of Y_i .

- (a) (10 points) Let $\hat{\beta}_0'$ be another unbiased estimator of β_0 , say, $\hat{\beta}_0' = \sum_{i=1}^n c_i Y_i$ where $c_i = a_i + b_i$. Show that $\text{Var}(\hat{\beta}_0) \leq \text{Var}(\hat{\beta}_0')$.

- (b) (10 points) The least squares estimator of β_0 is given as $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$. Find $\text{Cov}(\hat{\beta}_0, \hat{\beta}_1)$.
When is $\text{Cov}(\hat{\beta}_0, \hat{\beta}_1) = 0$?

3. Consider the simple linear regression model without intercept

$$Y_i = \beta X_i + \epsilon_i, i = 1, \dots, n,$$

where the random errors ϵ_i are independent and identically distributed as a normal distribution with mean 0 and variance σ^2 . In this exercise, we will apply maximum likelihood to estimate the parameters β and σ^2 in our model.

- (a) (4 points) Show that the **log-likelihood** function $\mathcal{L}(\beta, \sigma^2 | Y_1, \dots, Y_n)$ is

$$\mathcal{L}(\beta, \sigma^2 | Y_1, \dots, Y_n) = -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n (Y_i - \beta X_i)^2$$

Show all the details of the proof.

- (b) (4 points) Give the maximum likelihood estimators of β . Hint: Write down the normal equation; solve the system of equation in order to obtain the maximum likelihood estimators of β .

- (c) (4 points) Find the least squares estimator $\hat{\beta}$ in the model in Equation (1). Write down the least squares criterion for this model, give the normal equation(s), and show all the details of your work. Are you surprised by the estimator? Explain.

4. The response variable Y is the mortality due to skin cancer (number of deaths per 10 million people) and the predictor variable X is the latitude (degrees North) at the center of each of 49 states in the U.S. (The data were compiled in the 1950s, so Alaska and Hawaii were not yet states. And, Washington, D.C. is included in the data set even though it is not technically a state). We consider dataset `latitude`.

Call:

```
lm(formula = Mort ~ Lat, data = dat)
```

Residuals:

Min	1Q	Median	3Q	Max
-38.972	-13.185	0.972	12.006	43.938

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	(a)	23.8123	16.34	< 2e-16 ***
Lat	-5.9776	0.5984	(b)	3.31e-13 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: (c) on 47 degrees of freedom

Multiple R-squared: (d), Adjusted R-squared: 0.673

F-statistic: 99.8 on 1 and 47 DF, p-value: 3.309e-13

```
> anova(model)
```

Analysis of Variance Table

Response: Mort

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Lat	(e)	(f)	36464	99.797	3.309e-13 ***
Residuals	47	17173	365		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Some of the numerical values have been replaced by letters (a) to (f). Find the values of the numbers that have been replaced by letters. Give your answers near the corresponding letters below. Show all of your work.

- (a) (10 points) Some values have been replaced with numbers. Fill in those values. Show your work for this part.

(A) _____ (B) _____ (C) _____

(D) _____ (E) _____ (F) _____

(G) _____ (H) _____ (I) _____
(J) _____

(b) (5 points) Use the t-test to test the following

$$\begin{cases} H_0 : \beta_1 = 0 \\ H_1 : \beta_1 \neq 0 \end{cases}$$

at the level of significance 5%. Give the value of the test statistic and conclude within the context of the study.

(c) In simple linear regression, we can also use the confidence interval to test

$$\begin{cases} H_0 : \beta_1 = 0 \\ H_1 : \beta_1 \neq 0 \end{cases}$$

(3 points) Determine what is the conclusion using the confidence interval (reject or do not reject H_0) without doing any computations. Justify.