

STAC67H3F: Regression Analysis

Assignment 1

Fall 2026

Released Friday September 18, 2026 · Due **Friday October 2, 2026, 23:59**, on Crowdmark

Instructions

- This is **individual work**. You may discuss general ideas with classmates, but what you submit must be your own.
 - **Generative AI tools may not be used to complete this assignment**, as stated in the course syllabus. You are accountable for every line you submit, and should be able to explain any of it.
 - **Late submissions are not accepted.**
 - **Questions 1 and 7 must be done in R Markdown.** Submit the knitted output, with your code and its output visible. Use `ggplot2` for all figures.
 - Derivations (Questions 2 to 6) may be typed or handwritten; handwritten pages must be clearly legible scans. Show your work: an answer without justification receives no credit.
 - Upload each question to its own slot on Crowdmark.
 - As stated in the syllabus, a subset of the questions is marked in detail and the remainder is checked for completion. Answer every part fully.
-

Question 1

Least squares when the errors have heavy tails.

Consider the simple linear regression model

$$Y_i = 4 + 1.5x_i + \varepsilon_i, \quad i = 1, \dots, 20,$$

with fixed design points $x_i = i$, and errors $\varepsilon_1, \dots, \varepsilon_{20}$ independent with a Student t_q distribution. Recall that if $T \sim t_q$, then $E(T) = 0$ for $q > 1$ (the mean does not exist for $q = 1$), and $\text{Var}(T) = q/(q - 2)$ for $q > 2$ (the variance is infinite for $q \leq 2$).

Begin your R Markdown file with `set.seed(your student number)`.

- Write an R function `slope_t(q)` that simulates one sample from this model with t_q errors (use `rt`) and returns the least squares slope computed from the formula $\hat{\beta}_1 = \sum_i (x_i - \bar{x})Y_i / S_{xx}$. Do not use `lm`. Show the output of three calls to `slope_t(5)`.
- Using $K = 10,000$ simulated slopes for each of $q = 1, 2, 5$ and 30:
 - Give a table with, for each q : the mean of the simulated slopes, the simulated mean squared error (the average of $(\hat{\beta}_1 - 1.5)^2$), the theoretical variance σ^2/S_{xx} where it is finite, and the proportion of simulations with $|\hat{\beta}_1 - 1.5| > 1$. Compare the simulated mean squared error with the theoretical value, and comment.

- Draw boxplots of $\hat{\beta}_1 - 1.5$ for the four values of q , using `coord_cartesian(ylim = c(-1, 1))`.
- (c) Describe what happens for $q = 2$ and for $q = 1$. For each, which assumption behind the Gauss–Markov theorem fails, and what does that imply about the guarantee the theorem gives?

Question 2

Let $X \sim \text{Uniform}(0, 2)$ and $\varepsilon \sim N(0, 4)$ be independent, and define

$$Y = 3X + \varepsilon.$$

- Find $E(Y)$ and $\text{Var}(Y)$.
- Find $\text{Cov}(X, Y)$ and the correlation $\text{Corr}(X, Y)$.
- Show that $\text{Cov}(\varepsilon, \varepsilon^2) = 0$. Are ε and ε^2 independent? Justify your answer, and explain what it tells you about what a correlation coefficient measures.

Question 3

Regression through the origin.

In some applications the response must be zero when the predictor is zero, for example the fuel used by a vehicle against the distance it travels. Consider the model

$$Y_i = \beta x_i + \varepsilon_i, \quad i = 1, \dots, n,$$

where the x_i are known constants, not all zero, and the errors ε_i are uncorrelated with $E(\varepsilon_i) = 0$ and $\text{Var}(\varepsilon_i) = \sigma^2$.

- Derive the least squares estimator b of β .
- Show that b can be written as $b = \sum_{i=1}^n c_i Y_i$ for constants c_i , and identify the c_i . Use this to show that b is unbiased and that $\text{Var}(b) = \sigma^2 / \sum_{i=1}^n x_i^2$.
- Let $e_i = Y_i - b x_i$ be the residuals. Show that $\sum_{i=1}^n x_i e_i = 0$. Then **construct your own example of a data set with two observations** for which $\sum_{i=1}^n e_i \neq 0$.
- Now assume the errors are independent $N(0, \sigma^2)$ with σ^2 unknown. Write down the likelihood function of the sample $Y_1, \dots, Y_n$, and find the maximum likelihood estimators of β and σ^2 . Is the MLE of β the same as b ?
- Suppose the true model actually has an intercept, $Y_i = \beta_0 + \beta x_i + \varepsilon_i$ with $\beta_0 \neq 0$, but you still use b from part (a). Find $E(b)$ and the bias of b .
- Compare $\text{Var}(b)$ with $\text{Var}(\hat{\beta}_1) = \sigma^2 / S_{xx}$, the variance of the slope estimator in the model with an intercept. Which is smaller? What do parts (e) and (f) together suggest about the decision to leave the intercept out?

Question 4

Consider simple linear regression with an intercept, fitted by least squares to $(x_1, Y_1), \dots, (x_n, Y_n)$. Recall that the least squares estimates are

$$\hat{\beta}_1 = \frac{S_{xY}}{S_{xx}}, \quad \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x}, \quad \text{with} \quad S_{xx} = \sum_i (x_i - \bar{x})^2, \quad S_{xY} = \sum_i (x_i - \bar{x})(Y_i - \bar{Y}), \quad S_{YY} = \sum_i (Y_i - \bar{Y})^2,$$

and the fitted values are $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$. Write r for the sample correlation between x and Y , s_x and s_Y for the sample standard deviations, $\text{SSE} = \sum_i (Y_i - \hat{Y}_i)^2$ and $\text{MSE} = \text{SSE}/(n - 2)$.

- (a) Show that $\sum_i (x_i - \bar{x})(Y_i - \bar{Y}) = \sum_i x_i Y_i - n\bar{x}\bar{Y}$.
- (b) Show that $\hat{\beta}_1 = r s_Y / s_x$.
- (c) Show that $\text{SSE} = S_{YY} - \hat{\beta}_1^2 S_{xx}$.
- (d) Using $\text{se}(\hat{\beta}_1) = \sqrt{\text{MSE}/S_{xx}}$ and parts (b) and (c), show that

$$\frac{\hat{\beta}_1}{\text{se}(\hat{\beta}_1)} = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}.$$

Question 5

Medieval cathedrals.

For 25 medieval English cathedrals, x is the length of the cathedral and Y the height of its nave, both in feet. The lengths range from 182 to 611 feet and the heights from 45 to 103 feet. The following summary statistics were obtained.

$$\begin{array}{lll} n = 25 & \sum x_i = 10637 & \sum Y_i = 1869 \\ S_{xx} = 286432.24 & S_{YY} = 5366.56 & S_{xY} = 25056.88 \end{array}$$

You may use R to obtain quantiles and p -values, but show the formulas you use.

- (a) Find the least squares estimates $\hat{\beta}_0$ and $\hat{\beta}_1$, and write down the fitted regression line.
- (b) Find MSE, and the standard errors of $\hat{\beta}_0$ and $\hat{\beta}_1$.
- (c) Construct a 90% confidence interval for β_1 , and interpret it in context.
- (d) A historian claims that nave height increases by 15 feet for every additional 100 feet of length. Test this claim at the 5% level: state the hypotheses, the test statistic and its degrees of freedom, and your conclusion.
- (e) Interpret $\hat{\beta}_0$. Does it have a meaningful interpretation here? Explain.

Question 6

The Gauss–Markov theorem.

In simple linear regression with uncorrelated errors of mean zero and common variance σ^2 , the least squares slope is $\hat{\beta}_1 = \sum_i k_i Y_i$ with $k_i = (x_i - \bar{x})/S_{xx}$. Let $\tilde{\beta}_1 = \sum_i c_i Y_i$ be any other linear estimator, with constants c_i .

- (a) Show that $\tilde{\beta}_1$ is unbiased for all β_0, β_1 if and only if $\sum_i c_i = 0$ and $\sum_i c_i x_i = 1$.
- (b) Suppose $\tilde{\beta}_1$ is unbiased, and write $d_i = c_i - k_i$. Show that $\sum_i k_i d_i = 0$, and deduce that

$$\text{Var}(\tilde{\beta}_1) = \text{Var}(\hat{\beta}_1) + \sigma^2 \sum_i d_i^2 \geq \text{Var}(\hat{\beta}_1).$$

Question 7

Ozone and temperature.

The file `ozone.txt`, posted with this assignment, contains measurements taken on 112 days in the summer of 2001 in Rennes, France. We use two of its variables:

- `maxO3`: the maximum ozone concentration of the day, in $\mu\text{g}/\text{m}^3$;
- `T12`: the temperature at noon, in $^\circ\text{C}$.

Read the data with `read.table("ozone.txt", header = TRUE)`; the other columns are not used. Consider the model $\text{maxO3} = \beta_0 + \beta_1 \text{T12} + \varepsilon$, with the usual assumptions and normal errors.

- Draw a scatterplot of `maxO3` against `T12` with the fitted least squares line, and label both axes with their units. Describe the relationship you see.
- Compute $\hat{\beta}_0$ and $\hat{\beta}_1$ in R directly from their formulas, and check that they agree with `lm`. Interpret $\hat{\beta}_1$ in context. Is $\hat{\beta}_0$ meaningful here?
- Test at the 1% level whether higher noon temperatures are associated with higher daily ozone maxima. State the hypotheses, compute the t statistic and the p -value in R using `pt` (not by reading them from `summary`), and conclude. Then compute a 95% confidence interval for β_1 from its formula using `qt`.
- Using the fitted model, estimate $P(|\hat{\beta}_1 - \beta_1| > 0.5)$. State the distribution you use, and explain why your answer is only an approximation.