

STAC67H3F: Regression Analysis

Assignment 2

Fall 2026

Released Wednesday October 7, 2026 · Due **Wednesday October 21, 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 6 must be done in R Markdown.** Submit the knitted output, with your code and its output visible. Use `ggplot2` for all figures. Question 6 uses the function you write in Question 1.
 - Derivations and calculations (Questions 2 to 5) 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

Testing for constant variance when the errors are not normal.

The *Brown–Forsythe test* checks whether the error variance changes with the level of x . Split the residuals into two groups at the median of x ; in each group j , compute the absolute deviations $d_{ij} = |e_{ij} - \tilde{e}_j|$ from the group median $\tilde{e}_j$; then compare the two mean deviations $\bar{d}_1$ and $\bar{d}_2$ with a two-sample t statistic using the pooled variance, on $n - 2$ degrees of freedom. The classical F test of equal variances instead compares s_1^2/s_2^2 , the ratio of the residual variances in the two groups, with an $F(n_1 - 1, n_2 - 1)$ distribution. In R it is performed with `var.test`.

Throughout, use the model $Y_i = 10 + 2x_i + \varepsilon_i$, $i = 1, \dots, 40$, with fixed design points $x_i = i$. Group 1 contains the observations with x_i below the median of x , and group 2 the others. Begin your R Markdown file with `set.seed(your student number)`.

- (a) Write an R function `bf_test(e, x)` that returns the Brown–Forsythe statistic t_{BF} , its degrees of freedom and its two-sided p -value. Use only basic functions such as `median`, `mean`, `var` and `pt`: do not use `t.test` or a package that implements the test. Simulate one sample with $N(0, 2^2)$ errors, fit the regression with `lm`, and show the output of `bf_test` applied to its residuals.

- (b) In this part the error variance is constant, so the null hypothesis of both tests is true. For each of the three error distributions

- $N(0, 2^2)$, Student t_3 (`rt(40, 3)`), Exponential(1) - 1 (`rexp(40) - 1`),

simulate $K = 2000$ samples in each case, fit the regression to each, and record whether each test rejects at the 5% level (with both `var.test` and `bf.test`). Apply `var.test` to the residuals of the two groups. Give a table of the two rejection rates for each distribution (averaging the K samples), and comment: which test keeps its level?

- (c) Now let the errors be independent $N(0, (0.1x_i)^2)$, so that their spread grows with x . Recall that the *power* of a test is the probability that it rejects H_0 when H_0 is false. Estimate the power of each test at the 5% level by the proportion of $K = 2000$ samples in which it rejects. Based on (b) and (c), which test would you use to check for constant variance in practice, and why?

Question 2

Let $\mathbf{Z} = (Z_1, Z_2, Z_3)'$ be a random vector with expectation and covariance matrix

$$E(\mathbf{Z}) = \begin{bmatrix} 2 \\ -1 \\ 4 \end{bmatrix} \quad \text{and} \quad \sigma^2\{\mathbf{Z}\} = \begin{bmatrix} 4 & 2 & 1 \\ 2 & 9 & -3 \\ 1 & -3 & 4 \end{bmatrix}.$$

Recall that for a matrix $\mathbf{A}$ of constants (meaning that $\mathbf{A}$ is not random, but fixed), $E(\mathbf{AZ}) = \mathbf{A}E(\mathbf{Z})$ and $\sigma^2\{\mathbf{AZ}\} = \mathbf{A}\sigma^2\{\mathbf{Z}\}\mathbf{A}'$.

- (a) Let $W_1 = Z_1 + Z_2 + Z_3$ and $W_2 = Z_1 - 2Z_3$. Write $\mathbf{W} = (W_1, W_2)'$ as $\mathbf{W} = \mathbf{AZ}$ for some matrix $\mathbf{A}$ to explicit, and find $E(\mathbf{W})$ and $\sigma^2\{\mathbf{W}\}$.
- (b) Find $\text{Corr}(W_1, W_2)$.
- (c) Find the constant c for which $U = Z_2 - cZ_1$ is uncorrelated with Z_1 , and compute $\text{Var}(U)$.

Question 3

The hat matrix.

Consider simple linear regression in matrix form, $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, with $E(\boldsymbol{\varepsilon}) = \mathbf{0}$ and $\sigma^2\{\boldsymbol{\varepsilon}\} = \sigma^2\mathbf{I}$, where $\mathbf{X}$ is the $n \times 2$ design matrix whose i th row is $(1, x_i)$, and $\mathbf{X}'\mathbf{X}$ is invertible. Let

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}, \quad \hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{H}\mathbf{Y} \quad \text{with} \quad \mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}', \quad \mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}} = (\mathbf{I} - \mathbf{H})\mathbf{Y}.$$

Recall that for constant, non-random matrices $\mathbf{A}$ and $\mathbf{B}$, $\sigma^2\{\mathbf{AY}\} = \mathbf{A}\sigma^2\{\mathbf{Y}\}\mathbf{A}'$, and the matrix of covariances between the entries of $\mathbf{AY}$ and those of $\mathbf{BY}$ is $\sigma\{\mathbf{AY}, \mathbf{BY}\} = \mathbf{A}\sigma^2\{\mathbf{Y}\}\mathbf{B}'$.

- (a) Show that $\mathbf{H}$ is symmetric and idempotent (meaning $\mathbf{H}\mathbf{H} = \mathbf{H}$), and that $\mathbf{H}\mathbf{X} = \mathbf{X}$. Deduce that $\mathbf{I} - \mathbf{H}$ is symmetric and idempotent, and that $(\mathbf{I} - \mathbf{H})\mathbf{X} = \mathbf{0}$.
- (b) Show that $\mathbf{X}'\mathbf{e} = \mathbf{0}$. Write out the two equations that $\mathbf{X}'\mathbf{e} = \mathbf{0}$ contains, and say what each one tells you about the residuals.

- (c) Show that $\sigma^2\{e\} = \sigma^2(\mathbf{I} - \mathbf{H})$ and that $\sigma\{e, \hat{\mathbf{Y}}\} = \mathbf{0}$. If the errors are also normal, what does the second result imply about the residuals and the fitted values? Why does this support plotting residuals against fitted values to check the model?
- (d) Using $(\mathbf{X}'\mathbf{X})^{-1} = \frac{1}{S_{xx}} \begin{bmatrix} \sum_i x_i^2/n & -\bar{x} \\ -\bar{x} & 1 \end{bmatrix}$, show that the i th diagonal entry of $\mathbf{H}$ is $h_{ii} = 1/n + (x_i - \bar{x})^2/S_{xx}$. Deduce that $\sum_i h_{ii} = 2$ and that $\text{Var}(e_i) = \sigma^2(1 - h_{ii})$. Which residuals have the smallest variance?

Question 4

Apple trees.

In an urban orchard, x is the trunk circumference of an apple tree, measured 1 m above the ground (in cm), and Y is the weight of apples it produced over the season (in kg). For 16 trees, with circumferences from 30 to 88 cm, the following matrices were obtained, where $\mathbf{X}$ is the design matrix of the simple linear regression of Y on x .

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} 16 & 967 \\ 967 & 63921 \end{bmatrix}, \quad \mathbf{X}'\mathbf{Y} = \begin{bmatrix} 734 \\ 49968 \end{bmatrix}, \quad \mathbf{Y}'\mathbf{Y} = 39842.$$

You may use R as a calculator and to obtain quantiles, but show the formulas you use, and carry full precision through intermediate steps.

- (a) Find $(\mathbf{X}'\mathbf{X})^{-1}$ using the formula for the inverse of a 2×2 matrix, and then $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$. Write down the fitted regression line, and interpret $\hat{\beta}_1$ in context.
- (b) Using $\text{SST} = \mathbf{Y}'\mathbf{Y} - \frac{1}{n}(\sum_i Y_i)^2$ and $\text{SSE} = \mathbf{Y}'\mathbf{Y} - \hat{\beta}'\mathbf{X}'\mathbf{Y}$, construct the ANOVA table (sources, degrees of freedom, sums of squares, mean squares and F statistic), and find R^2 . Test $H_0 : \beta_1 = 0$ against $H_a : \beta_1 \neq 0$ at the 5% level using the F statistic.
- (c) Find the estimated variance-covariance matrix $\mathbf{s}^2\{\hat{\beta}\} = \text{MSE}(\mathbf{X}'\mathbf{X})^{-1}$, and the standard errors of $\hat{\beta}_0$ and $\hat{\beta}_1$.
- (d) For trees with a circumference of 60 cm, construct a 95% confidence interval for the mean yield, and a 95% prediction interval for the yield of one such tree. You may use $\mathbf{s}^2\{\hat{Y}_h\} = \mathbf{X}'_h \mathbf{s}^2\{\hat{\beta}\} \mathbf{X}_h$ with $\mathbf{X}_h = (1, 60)'$, which equals $\text{MSE}(1/n + (x_h - \bar{x})^2/S_{xx})$.

Question 5

Working on the log scale.

Suppose that $\log Y = \beta_0 + \beta_1 x + \varepsilon$ with $\varepsilon \sim N(0, \sigma^2)$, where $\log$ is the natural logarithm.

- (a) Show that, at a given x , the median of Y is $\exp(\beta_0 + \beta_1 x)$ and the mean of Y is $E(Y) = \exp(\beta_0 + \beta_1 x + \sigma^2/2)$. Deduce that increasing x by one unit multiplies both the median and the mean of Y by e^{β_1} .

Hint: if $V \sim N(\mu, \tau^2)$, then $E(e^V) = e^{\mu + \tau^2/2}$.

- (b) Let (L, U) be a 95% prediction interval for $\log Y_{h(\text{new})}$ at x_h , and (L^*, U^*) a 95% confidence interval for $E(\log Y_h)$. Explain why (e^L, e^U) is a 95% prediction interval for $Y_{h(\text{new})}$, but (e^{L^*}, e^{U^*}) is not a 95% confidence interval for $E(Y_h)$. What quantity does (e^{L^*}, e^{U^*}) cover with 95% confidence instead?

Question 6

Ozone, continued.

In Assignment 1 you fitted the model $\text{max03} = \beta_0 + \beta_1 \text{T12} + \varepsilon$ to the 112 days in `ozone.txt`, posted again with this assignment. We now check that model. Recall that `max03` is the maximum ozone concentration of the day (in $\mu\text{g}/\text{m}^3$) and `T12` the temperature at noon (in $^\circ\text{C}$).

- (a) Fit the model. Plot the residuals against the fitted values, and draw a normal Q-Q plot of the residuals. Then, at the 5% level, test for constant variance with your function `bf_test` from Question 1, splitting at the median of `T12`, and test for normality with `shapiro.test`. Which assumptions of the model look reasonable, and which do not?
- (b) Use `boxcox` from the `MASS` package to choose a power transformation of `max03`. For each λ on a grid, it computes the log-likelihood $\ell(\lambda)$ of the model for the transformed response, and $\hat{\lambda}$ is the value that maximizes it. A 95% confidence interval for λ follows from this result: if λ is the true value, the statistic

$$2[\ell(\hat{\lambda}) - \ell(\lambda)]$$

has approximately a χ_1^2 distribution. The interval is the set of values of λ that a test at the 5% level based on this statistic would not reject. Compute $\hat{\lambda}$ and this interval in R, using a grid with steps of 0.01, and say which simple transformation they suggest. Then fit the regression of $\log(\text{max03})$ on `T12`, and repeat the checks of part (a).

- (c) Interpret the slope of the log model as a percentage change in ozone per degree (see Question 5). Then compute 95% prediction intervals for `max03`, in $\mu\text{g}/\text{m}^3$, on days with `T12 = 15` and `T12 = 30`, from the original model and from the log model. Compare their widths, and explain the difference using your residual plots.
- (d) The variable `max03v`, which is not in the model, is the maximum ozone concentration of the *previous* day, in $\mu\text{g}/\text{m}^3$. Draw a scatterplot of the residuals of the log model against `max03v`, and add to it the least squares line of the residuals on `max03v` (for example with `geom_smooth(method = "lm")`). What do you see, and which assumption of the model does it call into question?