STAC67H3F: Regression Analysis

Assignment 3

Fall 2026

Released Wednesday November 4, 2026 · Due **Friday November 20, 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 5 must be done in R Markdown.** Submit the knitted output, with your code and its output visible. Use `ggplot2` for all figures. Question 5 uses the function you write in Question 1.
- Derivations and calculations (Questions 2 to 4) 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 coefficients when the predictors are correlated.

In this question, you study by simulation how the t and F tests behave when two predictors are correlated. The data come from the model $Y_i = 1 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i$, $i = 1, \ldots, 30$, with ε_i independent $N(0, 1)$. We always fit it with an intercept, so that $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)'$ and there are $30 - 3 = 27$ error degrees of freedom.

Begin your R Markdown file with `set.seed(`*your student number*`)`, then copy this code, which builds the predictors:

```
n  <- 30
x1 <- as.numeric(scale(1:n))
z  <- as.numeric(scale(resid(lm(cos(1:n) ~ x1))))
design <- function(r) data.frame(x1 = x1, x2 = r * x1 + sqrt(1 - r^2) * z)
```

Here `scale` subtracts the mean of a vector and divides by its standard deviation, and `z` is a fixed vector that is uncorrelated with `x1`. For any r with $|r| < 1$, the two columns of `design(r)` therefore have mean 0 and variance 1, and their sample correlation is exactly r.

The predictors stay fixed: only the errors change from one sample to the next. For example, one sample with $r = 0.9$ and $(\beta_1, \beta_2) = (0.3, 0.3)$, and its fit, are obtained with:

```
d <- design(0.9)
d$y <- 1 + 0.3 * d$x1 + 0.3 * d$x2 + rnorm(n)
fit <- lm(y ~ x1 + x2, data = d)
```

(a) Write an R function `glh_test(fit, K, m)` that takes a fitted model, a matrix `K` and a vector `m`, and returns the statistic F^* of $H_0 : \boldsymbol{K}'\boldsymbol{\beta} = \boldsymbol{m}$, its two degrees of freedom and its p-value. The matrix `K` is the matrix $\boldsymbol{K}'$ of the lecture: one row per restriction, and one column per coefficient. Use only basic functions such as `coef`, `vcov`, `solve`, `df.residual` and `pf`: do not use a package that implements the test.

Then check your function on the sample above, writing `K` and `m` in each case:

- For $H_0 : \beta_1 = \beta_2 = 0$, compare with the last line of `summary(fit)`.
- For $H_0 : \beta_1 = \beta_2$, the reduced model has the single predictor $x_1 + x_2$. Create it with `d$s <- d$x1 + d$x2`, fit `reduced <- lm(y ~ s, data = d)`, and compare with `anova(reduced, fit)`, which gives the same test from two fits (Lecture 15).

(b) There are six cases: r in $\{0,\ 0.9,\ 0.99\}$, and (β_1, β_2) in $\{(0.3, 0.3),\ (0.3, -0.3)\}$. In all of them, $H_0 : \beta_1 = 0$ and $H_0 : \beta_1 = \beta_2 = 0$ are false, so the proportion of samples in which a test rejects estimates its power. For each case, simulate $N = 2000$ samples as above and fit the regression to each (`replicate` is useful, as in Assignment 2). For each sample `f`, record:

- whether the t test of $H_0 : \beta_1 = 0$ rejects at the 5% level, using its p-value, `summary(f)$coefficients["x1", 4]`;
- whether the F test of $H_0 : \beta_1 = \beta_2 = 0$ rejects at the 5% level, using the p-value from `glh_test`;
- the value of $\hat{\beta}_1$, which is `coef(f)[["x1"]]`.

Give a table with one row per case and these columns: r, β_2, the rejection rate of the t test, the rejection rate of the F test, the mean of $\hat{\beta}_1$, the standard deviation of $\hat{\beta}_1$, and its theoretical standard deviation. By the formula $\mathrm{Var}(\hat{\beta}_1) = \sigma^2/(S_{11}(1 - r_{12}^2))$ of Lecture 15, the latter is $1/\sqrt{29(1 - r^2)}$.

Then answer these questions. Is $\hat{\beta}_1$ unbiased in all six cases? How does its standard deviation change with r, and does it agree with the theory? What happens to the rejection rate of the t test as r grows? And to the rejection rate of the F test, in each of the two settings for (β_1, β_2)?

(c) The rejection rates of the F test in (b) can also be computed exactly, using a distribution that we have not seen in class: the noncentral F distribution, defined below.

Let $\mu_i = E(Y_i) = 1 + \beta_1 x_{i1} + \beta_2 x_{i2}$ be the mean response of observation i, and $\bar{\mu} = \frac{1}{n}\sum_i \mu_i$ the average of these means. By the theorem on quadratic forms (Lecture 14), $SSE/\sigma^2 \sim \chi^2_{27}$ and, independently of it, SSR/σ^2 has a noncentral χ^2 distribution with 2 degrees of freedom and noncentrality parameter $\Omega = \sum_i (\mu_i - \bar{\mu})^2/(2\sigma^2)$. We use these two facts in

$$F^* = \frac{(SSR/\sigma^2)/2}{(SSE/\sigma^2)/27}.$$

If $H_0 : \beta_1 = \beta_2 = 0$ is true, then $\Omega = 0$ and $F^* \sim F(2, 27)$, as in the lecture. If H_0 is false, then $\Omega > 0$, and the numerator is no longer a usual χ^2 variable. We say that a ratio of this form, with a noncentral χ^2 variable in the numerator, has a *noncentral F distribution*, here with 2 and 27 degrees of freedom and noncentrality parameter Ω.

The test rejects when $F^* > F(0.95;\ 2,\ 27)$, so its power is the probability of this event when F^* has the noncentral F distribution. In R:

```
crit  <- qf(0.95, 2, 27)
power <- 1 - pf(crit, 2, 27, ncp = 2 * Omega)
```

Here `qf(0.95, 2, 27)` is the critical value of the test, the 0.95 quantile of the usual $F(2, 27)$ distribution. Note that R's `ncp` is 2Ω, not Ω, as for `dchisq` in Lecture 13.

(i) For each of the six cases, compute in R the vector of the μ_i, then Ω (here $\sigma^2 = 1$), then the power. Add Ω and the power to your table from (b), and compare the power with the simulated rejection rate of the F test.

(ii) Show that, for this design, $\sum_i (\mu_i - \bar{\mu})^2 = 29\,(\beta_1^2 + \beta_2^2 + 2r\beta_1\beta_2)$. Start by finding $\bar{\mu}$. This formula is valid because of the properties of the predictors stated at the start of the question: say where you use each of them. Check it against your values of Ω.

(iii) Simplify this expression for $(\beta_1, \beta_2) = (0.3, 0.3)$ and for $(\beta_1, \beta_2) = (0.3, -0.3)$. As r approaches 1, x_{i2} gets close to x_{i1}: what happens to the μ_i in each setting? Use this to explain why the power of the F test behaves so differently in the two settings.

Question 2

What "holding the others fixed" means.

Lecture 10 showed on an example that a coefficient can be obtained from a regression of residuals on residuals. We now prove this. Write $\boldsymbol{X} = [\,\boldsymbol{X}_1 \mid \boldsymbol{x}\,]$, where $\boldsymbol{X}_1$ has $p - 1$ columns, including the column of ones, and $\boldsymbol{x}$ is the column of one predictor, with $\boldsymbol{X}$ of full column rank. The full fit is $\boldsymbol{Y} = \boldsymbol{X}_1\hat{\boldsymbol{\beta}}_1 + \boldsymbol{x}\hat{\beta} + \boldsymbol{e}$, where $\hat{\beta}$ is the estimated coefficient of $\boldsymbol{x}$ and $\boldsymbol{e}$ the vector of residuals. Let $\boldsymbol{H}_1 = \boldsymbol{X}_1(\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'$ be the hat matrix of the model without $\boldsymbol{x}$, and $\boldsymbol{M} = \boldsymbol{I} - \boldsymbol{H}_1$. Recall from Lecture 11 that $\boldsymbol{M}$ is symmetric and idempotent, and that $\boldsymbol{M}\boldsymbol{X}_1 = \boldsymbol{0}$. Let $\boldsymbol{u} = \boldsymbol{M}\boldsymbol{x}$ and $\boldsymbol{v} = \boldsymbol{M}\boldsymbol{Y}$: these are the residuals of the regressions of $\boldsymbol{x}$ and of $\boldsymbol{Y}$ on $\boldsymbol{X}_1$.

(a) Use $\boldsymbol{X}'\boldsymbol{e} = \boldsymbol{0}$ to show that $\boldsymbol{M}\boldsymbol{e} = \boldsymbol{e}$ and $\boldsymbol{u}'\boldsymbol{e} = 0$.

(b) Multiply the full fit by $\boldsymbol{M}$ to show that $\boldsymbol{v} = \hat{\beta}\,\boldsymbol{u} + \boldsymbol{e}$. Deduce that $\hat{\beta} = \boldsymbol{u}'\boldsymbol{v}/\boldsymbol{u}'\boldsymbol{u} = \boldsymbol{u}'\boldsymbol{Y}/\boldsymbol{u}'\boldsymbol{u}$. In words: $\hat{\beta}$ is the slope of which regression, and what are the residuals of that regression?

(c) Show that $\mathrm{Var}(\hat{\beta}) = \sigma^2/\boldsymbol{u}'\boldsymbol{u}$. Let $S = \sum_i (x_i - \bar{x})^2$, and let R_x^2 be the R^2 of the regression of $\boldsymbol{x}$ on $\boldsymbol{X}_1$. Explain why $\boldsymbol{u}'\boldsymbol{u} = S(1 - R_x^2)$, and conclude that $\mathrm{Var}(\hat{\beta}) = \sigma^2/(S(1 - R_x^2))$. What does this become with two predictors? When is the variance inflation factor $1/(1 - R_x^2)$ of Lecture 14 equal to 1?

(d) Show that $\boldsymbol{v}'\boldsymbol{v} = \hat{\beta}^2\,\boldsymbol{u}'\boldsymbol{u} + \boldsymbol{e}'\boldsymbol{e}$. Identify $\boldsymbol{v}'\boldsymbol{v}$ and $\boldsymbol{e}'\boldsymbol{e}$ as the SSE of two models, and conclude that the extra sum of squares is $SSR(\boldsymbol{x} \mid \boldsymbol{X}_1) = \hat{\beta}^2\,\boldsymbol{u}'\boldsymbol{u}$. Deduce, without using the formula of the general linear hypothesis, that the F^* for $H_0 : \beta = 0$ is the square of the t^*.

Question 3

Radio or online?

A furniture retailer records, for $n = 30$ months, its sales Y, its spending on radio advertising x_1 and on online advertising x_2 (all in thousands of dollars), and its average discount x_3 (in percent). Its advertising budget is split between the two channels, so x_1 and x_2 are strongly correlated ($r = 0.951$). For the model $Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i$, with independent $N(0, \sigma^2)$ errors,

$$\boldsymbol{X'X} = \begin{bmatrix} 30 & 508 & 509 & 252 \\ 508 & 9656 & 9596 & 4185 \\ 509 & 9596 & 9637 & 4229 \\ 252 & 4185 & 4229 & 2676 \end{bmatrix}, \qquad \boldsymbol{X'Y} = \begin{bmatrix} 3416 \\ 61260 \\ 61322 \\ 29309 \end{bmatrix}, \qquad \boldsymbol{Y'Y} = 404462.$$

You may use R as a calculator and to obtain quantiles, but show the formulas you use.

(a) Find $\hat{\boldsymbol{\beta}}$, $SSE = \boldsymbol{Y'Y} - \hat{\boldsymbol{\beta}}'\boldsymbol{X'Y}$, MSE and the estimated variance-covariance matrix $\boldsymbol{s^2\{\hat{\beta}\}}$. Give the standard errors of $\hat{\beta}_1$ and $\hat{\beta}_2$, and the correlation between these two estimates.

(b) Test $H_0 : \beta_1 = 0$ and $H_0 : \beta_2 = 0$ separately at the 5% level. Then compute F^* for $H_0 : \beta_1 = \beta_2 = 0$ from the formula of the general linear hypothesis, and conclude at the 5% level. Explain why these results do not contradict each other.

(c) Give 95% confidence intervals for $\beta_1 + \beta_2$ and for $\beta_2 - \beta_1$. Say what each one measures for the retailer's advertising budget, and explain why their widths are so different.

(d) The marketing manager claims that radio and online dollars are equally effective, and that each extra percentage point of discount brings in exactly one thousand dollars of sales. Write this claim as $H_0 : \boldsymbol{K'\beta} = \boldsymbol{m}$, compute F^* from the formula, and conclude at the 5% level.

Question 4

Sum and difference.

Consider the model $Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i$ with $\boldsymbol{X'X}$ invertible, and let $s_i = x_{i1} + x_{i2}$ and $d_i = x_{i1} - x_{i2}$. We write the same model with these two variables, $Y_i = \gamma_0 + \gamma_1 s_i + \gamma_2 d_i + \varepsilon_i$, and call $\boldsymbol{Z}$ its design matrix.

(a) Find the 3×3 invertible matrix $\boldsymbol{C}$ such that $\boldsymbol{Z} = \boldsymbol{XC}$. Show that $\boldsymbol{Z}$ and $\boldsymbol{X}$ have the same hat matrix, so that the two fits have the same fitted values, residuals and SSE. Show that $\hat{\boldsymbol{\gamma}} = \boldsymbol{C}^{-1}\hat{\boldsymbol{\beta}}$, and express γ_1 and γ_2 in terms of β_1 and β_2. What does the t test of $H_0 : \gamma_2 = 0$ test, in terms of β_1 and β_2?

(b) Suppose from now on that $\sum_i x_{i1} = \sum_i x_{i2} = 0$, $\sum_i x_{i1}^2 = \sum_i x_{i2}^2 = S$ and $\sum_i x_{i1} x_{i2} = rS$, with $|r| < 1$. Find $\boldsymbol{Z'Z}$. Show that $\hat{\gamma}_1$ and $\hat{\gamma}_2$ are uncorrelated, and find their variances. Deduce $\mathrm{Var}(\hat{\beta}_1)$, and check that it agrees with Question 2(c). As $r \to 1$, which combination of β_1 and β_2 can be estimated precisely, and which cannot? How does this explain your results in Questions 1(b) and 3(c)?

(c) With the same assumptions, show that the variance of the fitted mean at (x_{h1}, x_{h2}), with $s_h = x_{h1} + x_{h2}$ and $d_h = x_{h1} - x_{h2}$, is

$$\mathrm{Var}(\hat{Y}_h) = \sigma^2 \left[\frac{1}{n} + \frac{s_h^2}{2S(1+r)} + \frac{d_h^2}{2S(1-r)} \right].$$

Now take $r = 0.95$, and $a > 0$ small enough that a and $-a$ both lie well inside the observed range of each predictor. Compute the sum of the last two terms in the brackets at the point (a, a) and at the point $(a, -a)$: how many times larger is it at $(a, -a)$? Since x_1 and x_2 are strongly positively correlated, which of the two points follows the pattern of the data? Explain why $(a, -a)$ is a case of hidden extrapolation (Lecture 12), even though each of its coordinates lies within the range of its predictor.

Question 5

Ozone: correlated temperatures in real data.

We use the 112 days in `ozone.txt`, posted again. As in Assignment 2, the response is log(`maxO3`). The predictors are the temperatures `T9`, `T12` and `T15` at 9:00, 12:00 and 15:00 (in °C); `Ne12`, the cloud cover at noon (in eighths of the sky); `Vx12`, the east-west component of the wind at noon; and `maxO3v`, the maximum ozone of the previous day. The full model is

$$\log(\texttt{maxO3}) = \beta_0 + \beta_1 \texttt{T9} + \beta_2 \texttt{T12} + \beta_3 \texttt{T15} + \beta_4 \texttt{Ne12} + \beta_5 \texttt{Vx12} + \beta_6 \texttt{maxO3v} + \varepsilon,$$

and it is fitted with:

```
oz <- read.table("ozone.txt", header = TRUE)
full <- lm(log(maxO3) ~ T9 + T12 + T15 + Ne12 + Vx12 + maxO3v, data = oz)
```

(a) We first look at the temperatures one at a time. Check the assumptions of the full model with a plot of the residuals against the fitted values and a normal Q-Q plot. From `summary(full)`, which of the three temperatures are significant at the 5% level? Give the correlation matrix of the three temperatures, using `cor`. Then compute the variance inflation factor of `T12`: regress `T12` on the five other predictors, let R_k^2 be the R^2 of this regression, and compute $1/(1 - R_k^2)$ (Lecture 15). Use these results to explain why `T12` is not significant here, although temperature was strongly related to ozone in Assignment 2.

(b) We now look at the temperatures together, using `glh_test`. The coefficients are $\beta_0, \ldots, \beta_6$, so `K` has 7 columns. Write `K` and `m`, and test at the 5% level:

- $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$, which says that temperature plays no role;
- $H_0 : \beta_1 = \beta_2 = \beta_3$, which says that temperature matters only through the sum `Tsum` = `T9` + `T12` + `T15`.

The second hypothesis gives this reduced model:

```
oz$Tsum <- oz$T9 + oz$T12 + oz$T15
reduced <- lm(log(maxO3) ~ Tsum + Ne12 + Vx12 + maxO3v, data = oz)
```

Check your second F^* with `anova(reduced, full)`. Compare the standard error and the t value of `Tsum` in `summary(reduced)` with those of the three temperatures in (a). In two or three sentences, say what the data tell us about temperature and ozone, and what they cannot tell us.

(c) Finally, we use the full model to estimate the mean response on two days, both with `Ne12` = 4, `Vx12` = 0 and `maxO3v` = 90. On day A, the temperature rises during the day, with (`T9`, `T12`, `T15`) = (18, 22, 24). On day B, it falls, with (26, 20, 18). Each of these temperatures lies within the range of its variable in the data. This code gives 95% confidence intervals for $E(\log(\texttt{maxO3}))$ on the two days:

```
days <- data.frame(T9 = c(18, 26), T12 = c(22, 20), T15 = c(24, 18),
                   Ne12 = 4, Vx12 = 0, maxO3v = 90)
predict(full, days, interval = "confidence")
```

Compare the widths of the two intervals. They use the same MSE and the same t quantile, so the difference comes only from $\boldsymbol{X}_h'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}_h$. To see why it differs, draw the scatterplot of `T15` against `T9` for the 112 days, and add the points of days A and B in another colour. Where does each day lie with respect to the cloud of data? Explain why day B is a case of hidden extrapolation, and link your answer to Question 4(c).