

STAC67H3F: Regression Analysis

Case Study Project

Fall 2026

1 The project

The project is the opportunity to apply, on a real dataset, the models and methods of this course: from a question you want to answer, through a multiple regression model, to its validation and its interpretation.

The project is **optional**. As explained in the syllabus, your final grade is computed both with and without it, and you receive the higher of the two. If you start the project and stop, you are simply assessed without it.

2 Groups and datasets

- You work in a **group of two to four** students.
- You choose your dataset from the **nine approved datasets** listed at the end of this document. You cannot use another dataset.
- **To find partners**, answer the poll on Piazza: it is permanent and not anonymous, so everyone can see which datasets each student is ready to work on (multiple-choice possible). Update your answer as your plans change.
- **To declare your group**, sign up on Quercus (People, then Groups), where each group corresponds to a dataset: joining a group and claiming its dataset at the same time. All members of your group sign up in the same group. You may switch until the deadline.
- **Deadline: Friday 6 November**. After this date, groups and datasets are final.

3 What the project must contain

1. **A question and an objective**. State clearly what you want to explain or predict, and why it matters.
2. **A descriptive analysis of the data**. *Careful*: this part must only contain what is useful for the rest of the project. It is not a list of every possible plot, but the description that motivates your modelling choices (distribution of the response, relations with the predictors, unusual observations, data issues).
3. **A multiple linear regression model**, built from the question and the descriptive analysis. It should use both quantitative and qualitative predictors, and, where they are justified, transformations, polynomial terms or interactions.
4. **A model selection step**: explain how you arrived at your final model, with criteria and procedures seen in class.

5. **A validation step:** residual analysis, outlying and influential observations, multicollinearity, and the remedial measures you took.
6. **An interpretation** of the final model in the context of the data, with the answer to your question and the limitations of your analysis.

Careful!

- Every modelling choice must be justified, and the final model must be validated. It is not required that the model fits the data well, but the validation must be carried out properly and its conclusions stated honestly. The grade does not depend on how well the model fits.
- If you use a tool that was not covered in the lectures or tutorials, you must explain it in your report, and you must be able to answer questions about it (there will be).

4 The report

1. There is one report per group, and **the report mark is common to all members**.
2. The report presents, explains and interprets the elements and models listed above. Support what you say with well-chosen figures and tables. Introduce the dataset, state the question clearly, and justify the tools and models you use.
3. **All figures must be made with the ggplot2 package.**
4. The report is written in R Markdown and handed in as a **PDF or HTML** file only. Any other format will not be accepted.
5. **The code must not appear in the report.**
6. The report is at most **10 pages**. What appears after the limit won't be read, but additional figures and tables can be added to an appendix.
7. With the report, you also hand in your **R Markdown file** and **the dataset**. In your code, the data must be read with the file name only, without any folder path: your code must knit without any change on my side.
8. You also hand in a short **contribution statement**, agreed by all members, saying what each member did.
9. As for the assignments, **generative AI tools may not be used** to write the report or the slides.
10. **Deadline: Friday 27 November, 23:59**, on Quercus. No late report is accepted.

The report mark takes into account the quality of the writing and its clarity, the justification of the modelling choices and their validation, and the choice and interpretation of the figures and tables. The detailed rubric is posted with this document.

5 The checkpoint

The checkpoint gives you feedback on your first results and on what is expected in the final report.

1. Hand in a short report of **at most 3 pages** (PDF or HTML, code not shown), with your R Markdown file.
2. It must contain:
 - your question, in a few sentences;
 - a descriptive analysis of the data, targeted at that question;
 - a first multiple linear regression, with its residual analysis.
3. **Deadline: Friday 13 November, 23:59**, on Quercus.

6 The presentation

1. Each group presents its project in **five minutes**, in the week of 30 November. The schedule will be posted once the groups are known.
2. Every member presents a part of the talk.
3. Your talk should not be addressed to the instructor only, but to your classmates: it must be concise and clear, and explain what you did to students who have not seen your data.
4. Upload your slides as a **PDF** to Quercus by **Monday 30 November, 23:59**.
5. The presentation mark can differ between members of a group. It takes into account the clarity and pedagogical quality of the talk, the respect of the time, and the answers to the questions.

7 The project mark

The project is marked out of 100:

Part	Points	Mark
Checkpoint	8	common to the group
Report	72	common to the group
Presentation	20	individual

Each instruction of this document that is not respected costs 5 points on the project mark (for instance: code visible in the report, a figure not made with `ggplot2`, a missing R Markdown file, or code that does not knit).

8 Deadlines

Friday 6 November	Group formed and dataset claimed on Quercus
Friday 13 November, 23:59	Checkpoint
Friday 27 November, 23:59	Report, R Markdown file, dataset, contribution statement
Monday 30 November, 23:59	Slides (PDF)
Week of 30 November	Presentations

Missing the group deadline or the checkpoint means that you are assessed without the project, as explained in the syllabus.

Approved datasets

Nine datasets are available. Each has at least 100 observations, a quantitative response and both quantitative and qualitative predictors. The data files (CSV) are posted on Quercus, and the variables of each dataset are described in the pages that follow. The response suggested below is a starting point: you may choose another quantitative response in the same dataset, as long as you justify it.

Dataset	Size	Suggested response and predictors
Boston housing	506 tracts	<i>Median home value</i> of census tracts of the Boston area (1970) against crime, zoning, pollution, rooms, age of the housing, distances, access to highways, tax, pupil-teacher ratio and the Charles River.
Diamonds	1078 (+ 215)	<i>Price</i> of a diamond against its carat, cut, colour, clarity and dimensions. A separate test set of 215 diamonds is provided, to assess predictions.
Life expectancy	183 countries	<i>Life expectancy</i> in 2014 against health, economic and schooling indicators and development status, from WHO data.
London bike sharing	730 days	Daily <i>number of rentals</i> against temperature, humidity, wind, hours of rain and snow, weather, holidays, weekends and season (2015 and 2016).
NBA players	490	A measure of scoring (for instance points per minute) against height, weight, position and experience, 2014-15 season.
Ontario wages	3987	Hourly <i>wage</i> against education, age, sex and language, from the 1994 Survey of Labour and Income Dynamics.
Pokémon	721	A statistic of a Pokémon (for instance its attack) against its other statistics, types, generation, legendary status and stage of evolution. One row per species.
Professors' salaries	397	Nine-month <i>salary</i> of professors in a US college (2008-09) against rank, discipline, years since PhD, years of service and sex.
Used cars	494	<i>Price</i> (in euros) of second-hand cars against make, mileage, year, engine displacement and power, transmission, fuel, doors and seats.

Variables of each dataset

Each file is read in R with `read.csv("file.csv")`. “Values” gives the range of each quantitative variable and the categories of each qualitative one, computed from the file itself.

Boston housing

File: `boston_housing.csv`. 506 census tracts of the Boston area, 1970. Source: Harrison and Rubinfeld (1978), distributed in the R package `MASS` (data `Boston`).

Variable	Description	Values
<code>crim</code>	Per capita crime rate of the town	0.00632 to 88.98
<code>zn</code>	Proportion of residential land zoned for lots over 25,000 sq. ft	0 to 100
<code>indus</code>	Proportion of non-retail business acres in the town	0.46 to 27.74
<code>chas</code>	1 if the tract bounds the Charles River, 0 otherwise	0 to 1
<code>nox</code>	Nitrogen oxides concentration (parts per 10 million)	0.385 to 0.871
<code>rm</code>	Average number of rooms per dwelling	3.561 to 8.78
<code>age</code>	Proportion of owner-occupied units built before 1940	2.9 to 100
<code>dis</code>	Weighted mean of distances to five Boston employment centres	1.13 to 12.13
<code>rad</code>	Index of accessibility to radial highways	1 to 24
<code>tax</code>	Full-value property tax rate per \$10,000	187 to 711
<code>prratio</code>	Pupil-teacher ratio of the town	12.6 to 22
<code>lstat</code>	Percentage of the population of lower status	1.73 to 37.97
<code>medv</code>	Median value of owner-occupied homes (in \$1000s)	5 to 50

Diamonds

File: `diamonds.csv`. `diamonds.csv` (1078 diamonds) and `diamonds_test.csv` (215 other diamonds, with the same variables, to assess predictions). Source: a random subset of the data diamonds of the R package `ggplot2`.

Variable	Description	Values
<code>carat</code>	Weight of the diamond (carats)	0.23 to 4
<code>cut</code>	Quality of the cut, from Fair (worst) to Good, Very Good, Premium and Ideal (best)	Fair, Good, Ideal, Premium, Very Good
<code>color</code>	Colour, from D (best) to J (worst)	D, E, F, G, H, I, J
<code>clarity</code>	Clarity, from I1 (worst) to SI2, SI1, VS2, VS1, VVS2, VVS1 and IF (best)	I1, IF, SI1, SI2, VS1, VS2, VVS1, VVS2
<code>depth</code>	Total depth percentage, 100 times z over the mean of x and y	56.3 to 71
<code>table</code>	Width of the top of the diamond relative to its widest point (%)	52 to 68
<code>price</code>	Price (US dollars)	364 to 18,790
<code>x</code>	Length (mm)	3.9 to 10.01
<code>y</code>	Width (mm)	3.95 to 9.94
<code>z</code>	Depth (mm)	0 to 6.31

Variable	Description	Values
----------	-------------	--------

Life expectancy

File: `life_expectancy_2014.csv`. 183 countries, year 2014. Source: the dataset “Life Expectancy (WHO)” on Kaggle, compiled from the World Health Organization’s Global Health Observatory and United Nations data.

Variable	Description	Values
<code>country</code>	Country	183 values
<code>status</code>	Development status	Developed, Developing
<code>life_expectancy</code>	Life expectancy at birth (years)	48.1 to 89
<code>adult_mortality</code>	Probability of dying between 15 and 60 years, per 1000	1 to 522
<code>infant_deaths</code>	Infant deaths, per 1000 population	0 to 957
<code>alcohol</code>	Alcohol consumption per capita (litres of pure alcohol)	0.01 to 15.19 (1 missing)
<code>percentage_expenditure</code>	Expenditure on health as a percentage of GDP per capita	0 to 19,480
<code>hepatitis_b</code>	Hepatitis B immunization among 1-year-olds (%)	2 to 99 (10 missing)
<code>measles</code>	Reported measles cases, per 1000 population	0 to 79,560
<code>bmi</code>	Average body mass index of the population	2 to 77.1 (2 missing)
<code>under_five_deaths</code>	Deaths under five years of age, per 1000 population	0 to 1,200
<code>polio</code>	Polio immunization among 1-year-olds (%)	8 to 99
<code>total_expenditure</code>	Government health expenditure, as a percentage of total government expenditure	1.21 to 17.14 (2 missing)
<code>diphtheria</code>	Diphtheria, tetanus and pertussis immunization among 1-year-olds (%)	2 to 99
<code>hiv_aids</code>	Deaths due to HIV/AIDS, per 1000 live births (ages 0 to 4)	0.1 to 9.4
<code>gdp</code>	Gross domestic product per capita (US dollars)	12.28 to 119,200 (28 missing)
<code>population</code>	Population	41 to 1,294,000,000 (41 missing)
<code>thinness_10_19</code>	Prevalence of thinness among children aged 10 to 19 (%)	0.1 to 26.8 (2 missing)
<code>thinness_5_9</code>	Prevalence of thinness among children aged 5 to 9 (%)	0.1 to 27.4 (2 missing)

Variable	Description	Values
income_composition	Human Development Index in terms of income composition of resources (0 to 1)	0.345 to 0.945 (10 missing)
schooling	Expected years of schooling	4.9 to 20.4 (10 missing)

London bike sharing

File: london_bikes_daily.csv. 730 days from 4 January 2015 to 3 January 2017. Source: the dataset “London bike sharing” on Kaggle, built from Transport for London open data and weather records. The original hourly records were summed into days.

Variable	Description	Values
date	Date	730 values
count	Number of bike-share rentals that day (the response)	4,869 to 72,500
hours_recorded	Number of hours recorded that day (24 when complete)	9 to 24
temp	Mean temperature (degrees C)	1.25 to 27.38
temp_feels	Mean “feels like” temperature (degrees C)	-2.4 to 27.44
humidity	Mean humidity (%)	46.92 to 98.69
wind_speed	Mean wind speed (km/h)	2.77 to 41.9
rain_hours	Number of hours with rain or a thunderstorm	0 to 17
snow_hours	Number of hours with snow	0 to 6
weather	Most frequent weather of the day	broken clouds, clear, cloudy, few clouds, rain, snow
holiday	1 if a public holiday	0 to 1
weekend	1 if a Saturday or Sunday	0 to 1
season	Season	autumn, spring, summer, winter
weekday	Day of the week	Friday, Monday, Saturday, Sunday, Thursday, Tuesday, Wednesday
month	Month	12 values

NBA players

File: `nba_players.csv`. 490 players, totals for the 2014-15 regular season.

Variable	Description	Values
<code>name</code>	Player	490 values
<code>games</code>	Games played	1 to 83
<code>minutes</code>	Minutes played	3 to 2,981
<code>points</code>	Points scored	0 to 2,217
<code>fgm</code>	Field goals made	0 to 659
<code>fga</code>	Field goals attempted	0 to 1,471
<code>three_pm</code>	Three-point shots made	0 to 286
<code>three_pa</code>	Three-point shots attempted	0 to 646
<code>ftm</code>	Free throws made	0 to 715
<code>fta</code>	Free throws attempted	0 to 824
<code>assists</code>	Assists	0 to 838
<code>experience</code>	Seasons played before 2014-15 (0 for rookies)	0 to 19 (68 missing)
<code>height</code>	Height (cm)	172.5 to 222.5 (68 missing)
<code>weight</code>	Weight (kg)	72.45 to 162 (68 missing)
<code>position</code>	Position: C, PF, PG, SF or SG	C, PF, PG, SF, SG (68 missing)

Ontario wages

File: `ontario_wages.csv`. 3987 workers in Ontario, 1994. Source: the Survey of Labour and Income Dynamics, distributed in the R package `carData` (data `SLID`); rows with a missing value removed.

Variable	Description	Values
<code>wages</code>	Composite hourly wage rate (dollars)	2.3 to 49.92
<code>education</code>	Years of schooling	0 to 20
<code>age</code>	Age (years)	16 to 69
<code>sex</code>	Sex	Female, Male
<code>language</code>	Language	English, French, Other

Pokémon

File: `pokemon.csv`. 721 Pokémon of the first six generations.

Variable	Description	Values
<code>number</code>	Number of the species in the Pokédex	1 to 721
<code>name</code>	Name	721 values
<code>type1</code>	Primary type	18 values
<code>type2</code>	Secondary type (missing when there is none)	18 values (371 missing)
<code>hp</code>	Hit points	1 to 255
<code>attack</code>	Attack	5 to 165
<code>defense</code>	Defense	5 to 230
<code>sp_attack</code>	Special attack	10 to 154
<code>sp_defense</code>	Special defense	20 to 230
<code>speed</code>	Speed	5 to 160
<code>generation</code>	Generation (1 to 6)	1 to 6
<code>legendary</code>	TRUE for a legendary Pokémon	TRUE, FALSE
<code>evolution_stage</code>	Stage of evolution (1, 2 or 3), or legendary	1, 2, 3, legendary

Professors' salaries

File: `professors_salaries.csv`. 397 professors of a college in the United States, 2008-09. Source: the R package `carData` (data `Salaries`).

Variable	Description	Values
<code>rank</code>	Academic rank: assistant, associate or full professor	AssocProf, AsstProf, Prof
<code>discipline</code>	A for theoretical departments, B for applied departments	A, B
<code>yrs.since.phd</code>	Years since the PhD	1 to 56
<code>yrs.service</code>	Years of service	0 to 60
<code>sex</code>	Sex	Female, Male
<code>salary</code>	Nine-month salary (US dollars)	57,800 to 231,500

Used cars

File: `used_cars.csv`. 494 second-hand cars. The variables match the Kaggle dataset “Classified Ads for Cars” (online advertisements in the Czech Republic and Germany).

Variable	Description	Values
<code>maker</code>	Make of the car	27 values
<code>mileage</code>	Mileage (km)	0 to 398,000
<code>manufacture_year</code>	Year of manufacture	1,990 to 2,015
<code>engine_displacement</code>	Engine displacement (cm ³)	599 to 5,967
<code>engine_power</code>	Engine power (kW)	33 to 335
<code>transmission</code>	Transmission: manual or automatic	auto, man
<code>door_count</code>	Number of doors	2 to 5
<code>seat_count</code>	Number of seats	2 to 9
<code>fuel_type</code>	Fuel	diesel, gasoline
<code>price_eur</code>	Price (euros)	366.4 to 107,000