Topic: Regression and model diagnostics

Note: *Start each problem with the starter code shared in the repository. Feel free to use scipy reference or any other source for syntax, but invest time in following the logic of the programs.* *https://docs.scipy.org/doc/scipy/reference/stats.html*

Important note for submissions:

Each time you generate a plot, save it with a filename like a7_task1_fig1_yourInitials.png (assignment 7 → task 1 → figure 1). Upload all your figures to the figs_results directory in your GitHub repo.

Generate a pdf using all figures and your remarks. Upload this pdf on your repository. The pdf will be graded.

Regression analysis and model diagnostics

1. Setup:

You have experimental data collected by varying mix proportions and curing age of concrete specimens, then measuring **compressive strength** (MPa). Each trial sets component amounts (Cement, Fly Ash, Water, etc.) and **Age**; the goal is to build a regression model to predict compressive strength at untested mix/age combinations.

Regression serves several purposes:

- predict the outcome and its expected dispersion;
- rank predictor importance (is Fly Ash more important or water, in which range); and
- reveal interaction effects (does one variable matter only when another is high?).

Model diagnostics, on the other hand, addresses:

- Outliers: answers questions like: “Is this observation an extreme value or an error?” test regression assumptions (normality, constant variance); and
- compare candidate models (fit vs parsimony).

During this assignment, we will work on a dataset from Kaggle (a data-science competition platform where data scientists and ML engineers compete to solve data science challenges). The dataset is saved on the assignment repository. Link to original data.

The dataset has 8 key attributes: Cement, Blast Furnace Slag, Fly Ash, Water, Superplasticizer, Coarse Aggregate, Fine Aggregate, Age

Target attribute: **Compressive Strength**.

2. Background Concepts:

2.1 Fit with individual variables

Fit a linear model per predictor to gauge its marginal effect:

$$y \approx \beta_0 + \beta_1 x \qquad \text{(where } y\text{= strength, } x\text{= one predictor).}$$

Interpret β_0(baseline), slope β_1(change in mean y per unit x), and p-value (evidence against $\beta_1 = 0$). Repeat for each predictor to get a quick ranking before multivariable modeling.

2.2 Q-Q plot (normality check)

A Q-Q plot compares ordered residuals to theoretical normal quantiles. If residuals are roughly linear on the Q-Q plot, the normality assumption for errors is plausible. Deviations (heavy tails, curvature) indicate non-normality and suggest alternatives (transformations).

2.3 Error / residual plots

Residuals are defined as:

$$\varepsilon_i = y_i - \hat{y}_i$$

Plot ε_i versus fitted values $\hat{y}_i$.

Residuals scattered randomly around zero with no pattern is a good sign. A pattern indicates model misspecification (missing nonlinearity or interaction) or heteroscedasticity (variance changes with level). We can use residuals versus each predictor to localize problems.

2.4 Outliers and influence

Outliers are observations with large residuals. They can influence the regression by changing coefficients significantly. Ask: “Is this an extreme physical specimen or measurement error?” Cook’s distance is used to spot outliers. Domain knowledge is often used to decide the presence outliers. If points are removed, it is crucial to show how estimates and CI change due to removing outliers.

2.5 Interactions and higher-order terms

Interactions terms: if effect of x_1 on response (y) depends on the value (or level) of another predictor, say x_2, interaction terms are used:

$$y \approx \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2.$$

Quadratic terms: to capture curvature, quadratic terms such as $\beta_2 x^2$ are used.

Including these interaction and quadratic terms can reveal “hidden” importance, i.e., a variable that is weak marginally may be crucial in interaction.

2.6 Constant variance: homo- vs heteroscedasticity

Homoscedasticity: A model is called homoscedastic if the residuals have a constant variance.

$$\mathrm{Var}(\varepsilon_i) = \sigma^2, \text{ a constant.}$$

Heteroscedasticity: variance depends on predictors. Heteroscedasticity is detected using residual plots or tests (Breusch-Pagan). Some remedies include transform response (log), weighted least squares, or robust standard errors (not covered today).

2.7 Measures of fit: R^2, AIC, BIC

R^2 measures explained variance;

Adjusted R^2 penalizes extra predictors.

Akaike Information Criteria (AIC)/ Bayesian Information Criteria (BIC) trade off fit and complexity:

$$\mathrm{AIC} = 2k - 2\log\ L,$$

$$\mathrm{BIC} = k\log\ n - 2\log\ L.$$

AIC/BIC are used to prefer parsimonious models (ones with lesser number of variables) when predictive power is similar.

3. Objectives:

Build and diagnose regression models for predicting concrete compressive strength from mix components and age. You will fit ordinary least square (OLS) models, interpret coefficients and p-values, inspect residuals for normality and constant variance, detect outliers/influential points, compute fit metrics (R^2, AIC/BIC), and produce prediction intervals for new mix designs. The emphasis is both predictive accuracy and diagnostic reasoning: not merely obtaining coefficients but understanding when the model is trustworthy and what to do when assumptions fail.

4. Step-by-step tasks:

For following tasks, use the provided starter python file named `assm7_student.py` and work on each task by uncommenting the corresponding function in the `__main__`. Understand the sequence of provided code;

The provided code is error-free and works. However, you must:

- Prepare the report including findings from the code, and generated figures and tables.
- Look out for expanding the incomplete analysis in the starter code, you can do so by noticing which predictors are missing.
- Comment the code wherever you had to search to understand it.

Your submission is the PDF and updated python file.

Task 1. Quick sanity checks & exploratory plots

Subtasks:

- Load dataset (assign to df), show first 10 rows and variable names; report df.shape and dtypes.
- Compute basic summaries (mean, sd, min, max) for each predictor and target.
- Make 3 quick plots: histogram of compressive strength, scatter Cement vs Strength, scatter Water vs Strength.

Deliverable: one-page figure panel and one-line observations (e.g., skewness, obvious strong/weak predictors).

Note: keep an eye out for weak predictors that may become important with interaction.

Task 2. Single-predictor fits & ranking

Subtasks:

- For each predictor x_j, fit

$$y \approx \beta_0 + \beta_1 x_j.$$

Create a table with β_0, β_1, standard error, t-stat, p-value, and R^2.

- Make a summary table ranking predictors by their p-value and R^2.
- For a chosen predictor (strong or weak), plot residuals and Q–Q plot.

Deliverable: ranking table + example Q–Q plot.

Note: low p-value suggests predictive relevance but check effect size and R^2 to judge practical importance.

Task 3. Multivariable OLS, residual diagnostics

Subtasks:

- Fit multivariable linear model:

$$y \approx \beta_0 + \sum_j \beta_j x_j.$$

- Plot residuals vs fitted and residuals vs each key predictor. Produce Q-Q for residuals.

Deliverable: residual diagnostic plots and short diagnosis ("assumptions ok / heteroscedastic / non-normal").

Note: residual plots often expose missing nonlinear terms or interactions—this informs next tasks.

Task 4. Outliers & influence

Subtasks:

- Compute studentized residuals and Cook's distance.
- For a flagged observation: (a) inspect raw data row and experimental notes (hypothetical), (b) re-fit model without it and compare coefficients and R^2.

Deliverable: table of top five influencers and short decision (investigate / keep / exclude).

Note: In practice, your guiding question is "If removing one point changes the slope sign, is the point erroneous or influential physically?"

Task 5. Interactions, quadratic terms & ANOVA

Subtasks:

- Augment the baseline model with (a) pairwise interactions you suspect (e.g., Cement * Water), (b) quadratic terms for predictors where residuals show curvature.
- Use analysis of variance, ANOVA (F-test) to compare nested models (linear vs linear+interaction or vs quadratic). Report F-stat and p-value.

Deliverable: model comparison table (RSS, df, F, p-value) and short recommendation on whether to include interaction/quad terms.

Note: a marginally weak predictor may become significant when interacting with another. This reveals conditional effects.

Task 6. Heteroscedasticity remedies & prediction intervals

Subtasks:

- If heteroscedasticity detected under Task 3, try log-transform of y. Compare residual plots after remedy.
- Compute point predictions and **95% prediction intervals** for three new mix designs (choose representative mixes).

Deliverable: table of predictions + intervals and short rationale for chosen remedy.

Task 7. Model selection & cross-validation

Subtasks:

- Compute adjusted R^2, AIC, and BIC for candidate models (baseline, +interaction, +quad, transformed).

Deliverable: selection table and one-sentence justification of best model for prediction vs explanation.