Problem 1 (Chess ratings)

The ChessRatingComparison.csv file on the class website contains self-reported chess ratings compiled by Chess.com. The dataset has ratings from Chess.com, USCF (United States Chess Federation) and FIDE (the World Chess Federation). Higher ratings indicate stronger chess players. Bullet and blitz are different versions of fast chess.

The goal of this problem is to determine the relationship between the other variables and the USCF regular rating. For this problem remove the timestamp and username variables.

(a) Produce a pairwise scatterplot of the whole dataset, and describe the visual relationships between the variables. Be sure to put the USCF regular rating as the top row.

(b) Write down a simple statistical model for the relationship between USCF regular rating and Chess.com live standard rating. Using the full dataset, fit your model, providing OLS estimates of your regression parameters. Additionally report 95% confidence intervals, t-statistics and relevant p-values. Are your estimated parameters significantly different from zero?

(c) Provide some relevant diagnostic plots for your model. Discuss the results – is the model adequate, is there evidence of outliers, are there high leverage data points, etc.?

(d) If there are outliers in the data, remove them and redo part (b) (providing all of the same estimates/discussion as in part (b)). Provide a plausible explanation for why particular data values outliers?

(e) Interpret your parameter estimates and results from part (b)/(d) in words.

(f) If your professor were a chess player, his Chess.com standard rating would probably be about 1600 (a long time ago). What would the average 1600-rated player on Chess.com rate as in the USCF rating system (be sure to include confidence bounds)? Additionally predict your professor’s USCF rating (with confidence bounds).

(g) How would you decide which of the other rating systems provides best predictions for USCF regular ratings? Fit some relevant models and provide an defensible explanation for your solution.
Problem 2 (Relationship between fitted values and residuals)

If \( y = (y_1, \ldots, y_n)^T \) are a set of observations with corresponding design matrix \( X \), show that the fitted values and estimated residuals are orthogonal vectors. That is, show

\[
v_1 = X(X^TX)^{-1}X^Ty
\]

and

\[
v_2 = y - X(X^TX)^{-1}X^Ty
\]

are orthogonal. (That is, take the dot product of \( v_1 \) and \( v_2 \) and show it is zero).

Problem 3 (When do we reject \( H_0 \)?)

The standard t-test for testing \( H_0 : \beta_1 = \beta^* \) vs. \( H_A : \beta_1 \neq \beta^* \) rejects if

\[
\left| \frac{\hat{\beta}_1 - \beta^*}{SE(\hat{\beta}_1)} \right| > t_{n-2}(1 - \alpha/2)
\]

where \( t_{n-2}(1 - \alpha/2) \) is the \( t_{n-2} \)-quantile corresponding to significance level \( \alpha \).

If the sample size is \( n \) and we wish to reject at the \( \alpha = 0.05 \) level under assumptions A1, find all values of \( \beta^* \) that will result in a rejection of \( H_0 \). Does this region look familiar?

Problem 4 (Grad students only: Gauss-Markov Theorem)

Suppose \( (x_1, y_1), \ldots, (x_n, y_n) \) are observations from the multiple linear regression model

\[
Y = \beta_0 + \sum_{i=1}^{p} \beta_i X_i + \varepsilon.
\]

Working under assumption set A3, show that the OLS estimator \( \hat{\beta}_{OLS} \) for \( \beta = (\beta_0, \beta_1, \ldots, \beta_p)^T \) is the best linear unbiased estimator (BLUE) in the sense that \( \text{Var}(\hat{\beta}_{OLS}) \leq \text{Var}(\hat{\beta}') \) for any other unbiased linear estimator \( \hat{\beta}' \). We call an estimator linear if it is of the form \( \sum_{i=1}^{n} w_i y_i \) for some weights \( \{w_i\} \).

(Hint: variance matrices are nonnegative definite; assume that the other weights take the form \( (X^TX)^{-1}X^T + D \) for some matrix \( D \), and show \( \text{Var}(\hat{\beta}') = \text{Var}(\hat{\beta}_{OLS}) + \text{constant} \times DD^T \) where \( DD^T \) is also nonnegative definite).