DDS-8555 Assignment 2: Build Regression Models

Todd DeLozier

2026-02-22

1 Executive Summary

This report addresses all three required components of Assignment 2 for Predictive Analytics Week 2. The statistical results and implementation details are produced in the accompanying Jupyter notebooks and are referenced here to support interpretation, methodological transparency, and reproducibility. Part 1 and Part 2 are implemented in `Todd_DeLozier_Assignment_2_Python_Code_Parts_1_and_2.ipynb`. Part 3 is implemented in `Todd_DeLozier_Assignment_2_Python_Code_Part_3.ipynb`. The analysis emphasizes the core inferential assumptions of ordinary least squares regression, namely linearity in parameters, independence of errors, collinearity diagnostics as a practical identifiability concern, and distributional structure of residuals rather than raw variables, consistent with best practice recommendations (Ernst & Albers, 2017).

2 Part 1: Conceptual Question 3 from ISLR Python

The conceptual task considers a linear model for starting salary, measured in thousands of dollars, as a function of five predictors: grade point average, intelligence quotient, an indicator for educational level, and two interaction terms. The fitted coefficients supplied by the textbook problem define the prediction function. In the notebook, the model is written explicitly and evaluated numerically as a correctness check (see `Todd_DeLozier_Assignment_2_Python_Code_Parts_1_and_2.ipynb`, cells titled “Conceptual Q3 sanity checks” and “Difference between College and High School”).

The interpretation question hinges on the interaction between grade point average and educational level. Because the model contains a product term between grade point average and level, the effect of level is not constant across values of grade point average. Under indicator coding where Level equals 1 for college and 0 for high school, the conditional difference between predicted college salary and predicted high school salary, holding grade point average and intelligence quotient fixed, is the level coefficient plus the interaction coefficient multiplied by grade point average. With the supplied coefficients, the difference reduces to 35 minus 10 times grade point average. This expression implies that college graduates are predicted to earn more than high school graduates when grade point average is below 3.5, and to earn less when grade point average exceeds 3.5. Therefore, the correct conceptual choice is the option stating that high school graduates earn more than college graduates provided that grade point average is high enough, because the negative interaction causes the college advantage to diminish with increasing grade point average and eventually reverse.

The prediction component requests the salary of a college graduate with an intelligence quotient of 110 and a grade point average of 4.0. Substituting the provided values and Level equals 1 yields a prediction of 137.1, interpreted as $137,100 in the units of the problem. This value is computed directly in the notebook to avoid algebraic transcription error (see `Todd_DeLozier_Assignment_2_Python_Code_Parts_1_and_2.ipynb`, cell output `137.1`).

The final conceptual statement asserts that a small interaction coefficient implies little evidence of interaction. This statement is false as written because evidentiary strength is not determined by coefficient magnitude in isolation. The substantive meaning of a coefficient depends on the measurement scale of the variables and the range over which the product term varies. Even a small interaction coefficient can yield a material effect when the product term spans a wide range, and statistical evidence depends on uncertainty, sample size, and the fitted standard error. More broadly, interaction terms encode model structure that can fundamentally change interpretation even when the coefficient is numerically modest. Simulation evidence in the methodological literature shows that omitting an interaction when it exists can bias estimated simple effects, sometimes reversing their signs, particularly under certain correlation and noise regimes (Rimpler et al., 2025). The conceptual lesson is that interaction relevance must be evaluated with respect to scale, uncertainty, and theoretical meaning, rather than with a single heuristic about coefficient magnitude.

3 Part 2: Applied Question 10 from ISLR Python using the Carseats dataset

The applied task fits multiple regression models predicting Carseats sales from price and two categorical predictors. The implementation uses an ordinary least squares model specified as Sales regressed on Price, Urban, and US, with Urban and US treated as indicator coded categorical variables by the modeling framework (see `Todd_DeLozier_Assignment_2_Python_Code_Parts_1_and_2.ipynb`, the cell printing the OLS summary for “Sales ~ Price + Urban + US”).

The full model yields an R squared of 0.239 with an adjusted R squared of 0.234. The price coefficient is negative, approximately minus 0.0545, and is statistically significant at conventional thresholds. Interpreted in the units of the dataset, holding Urban status and US status constant, a one unit increase in price is associated with a decrease of about 0.0545 units of sales. The US indicator coefficient is approximately 1.2006 and is statistically significant. Holding price and Urban status constant, stores in the United States are predicted to have about 1.20 higher sales than stores outside the United States. The Urban indicator coefficient is near zero, approximately minus 0.0219, and is not statistically distinguishable from zero with a p value of 0.936, implying no evidence that Urban status is associated with sales after controlling for the other predictors in this specification.

The fitted equation can be written using indicator variables. Let I(US=Yes) be 1 for US stores and 0 otherwise, and let I(Urban=Yes) be 1 for Urban stores and 0 otherwise. The fitted equation from the notebook is:

Sales = 13.0435 + 1.2006 * I(US=Yes) - 0.0219 * I(Urban=Yes) - 0.0545 * Price.

Hypothesis testing of the model coefficients uses the null that each coefficient equals zero, evaluated through the standard t tests produced by the OLS summary. Price and US reject the null at alpha 0.05, while Urban does not. This empirical pattern motivates the reduced model, which removes Urban and refits Sales on Price and US (see `Todd_DeLozier_Assignment_2_Python_Code_Parts_1_and_2.ipynb`, the cell printing the OLS summary for “Sales ~ Price + US”).

The reduced model yields essentially identical explanatory performance, with R squared remaining 0.239 and adjusted R squared increasing slightly to 0.235. The inference is that Urban contributes no measurable explanatory value in this model once Price and US are included, and removing it preserves fit while improving parsimony. This is consistent with the principle that, when the inferential target is an interpretable model, variables without evidentiary support can be excluded when theory does not compel their inclusion, reducing unnecessary variance in estimates.

Confidence intervals were computed for the reduced model coefficients. The 95 percent interval for the US effect is approximately [0.6915, 1.7078], and the 95 percent interval for the price effect is approximately [-0.0648, -0.0442] (see `Todd_DeLozier_Assignment_2_Python_Code_Parts_1_and_2.ipynb`, the confidence interval output table). These intervals quantify uncertainty in the direction and magnitude of the associations under the model assumptions.

Diagnostics were evaluated using standard residual and influence methods. The notebook constructs a residuals versus fitted plot and a normal quantile plot of residuals, and computes leverage and Cook distance maxima. The maximum leverage is approximately 0.0433 and the maximum Cook distance is approximately 0.0261, values that do not suggest a small set of cases dominating the fitted solution. The residual plots show no pronounced structure, supporting the adequacy of the linear functional form in this restricted predictor set, and the quantile plot suggests approximate residual normality with only mild tail deviation. These diagnostics support the claim that model assumptions are not grossly violated in a way that would obviously invalidate inference, while also reinforcing the broader methodological point that assumptions pertain to residual behavior and design, not to marginal normality of the raw variables (Ernst & Albers, 2017).

4 Part 3: Kaggle regression with the Abalone dataset

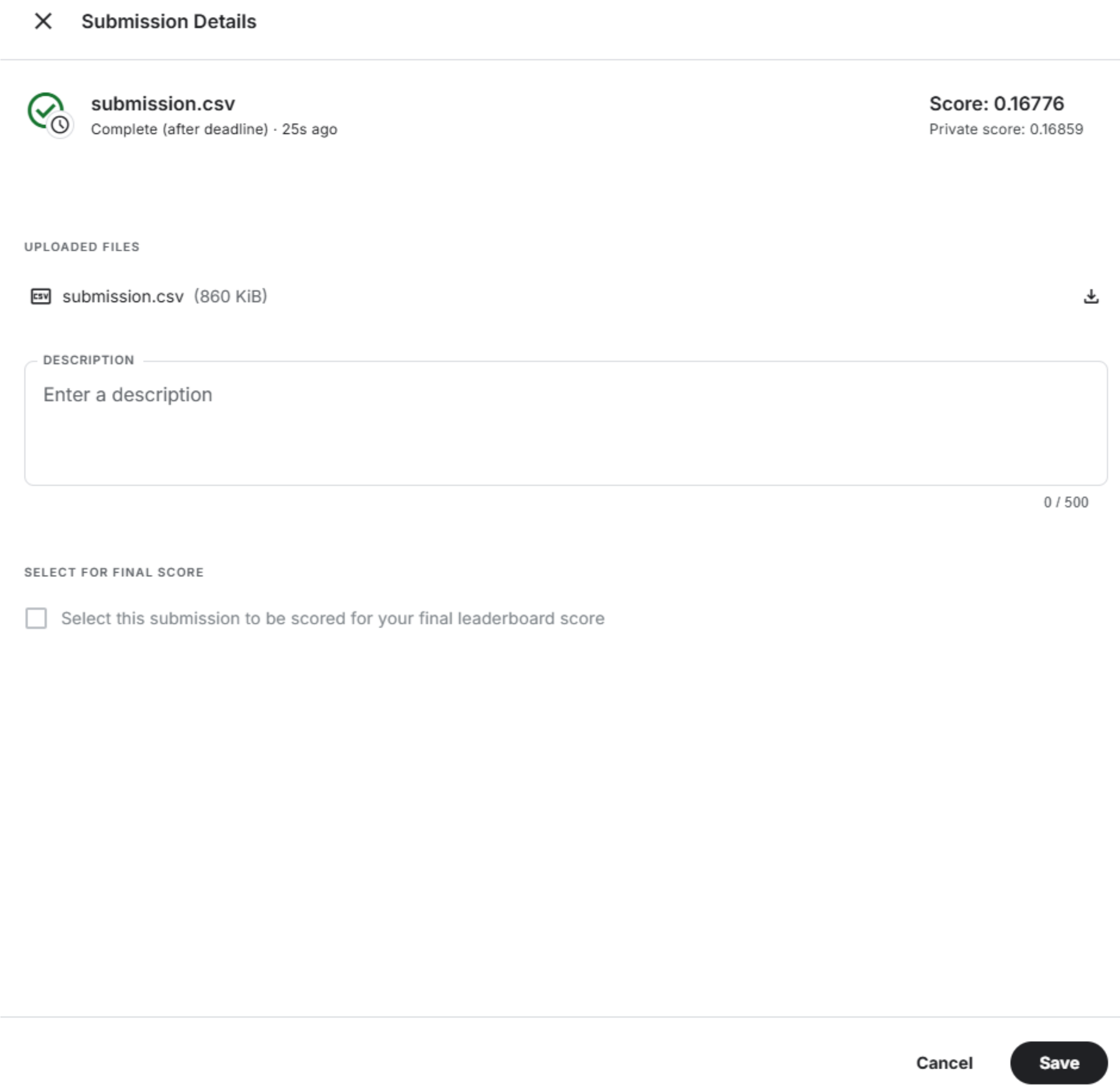
The Kaggle component uses the Regression with Abalone Dataset competition data to build and submit at least two regression models. The submitted work is implemented in `Todd_DeLozier_Assignment_2_Python_Code_Part_3.ipynb`. The notebook loads the Kaggle `train.csv` and `test.csv`, identifies the target as Rings, one hot encodes the Sex variable, and fits multiple models under a consistent preprocessing pipeline.

Three models are trained and evaluated using a holdout validation split: ordinary linear regression, ridge regression, and a random forest regressor. On the validation set, linear regression attains an RMSE of 2.0237 with R squared of 0.6013. Ridge regression is similar, with RMSE 2.0245 and R squared 0.6010. The random forest model improves predictive performance, achieving RMSE 1.8920 with R squared 0.6515, and is therefore selected as the best model by validation RMSE (see

`Todd_DeLozier_Assignment_2_Python_Code_Part_3.ipynb`, cells printing “Linear Regression RMSE”, “Ridge RMSE”, and “Random Forest RMSE”, followed by the cell printing “Best model: RandomForest”).

After model selection, the best model is refit on the full training data and used to generate predictions for the held out Kaggle test set. The notebook writes a `submission.csv` containing id and predicted Rings values and prints the resolved path where that file was written, along with the first rows of the submission file (see `Todd_DeLozier_Assignment_2_Python_Code_Part_3.ipynb`, final cell that prints “Wrote:” and displays the submission head). Evidence of successful Kaggle submission is included as a screenshot at the path provided in the assignment assets.

Kaggle submission proof screenshot:



Submission Details

submission.csv
Complete (after deadline) · 25s ago

Score: 0.16776
Private score: 0.16859

UPLOADED FILES

submission.csv (860 KiB)

DESCRIPTION
Enter a description

0 / 500

SELECT FOR FINAL SCORE

□ Select this submission to be scored for your final leaderboard score

4.1 Assumption investigations for the Abalone regression work

Because the random forest is not an ordinary least squares model, classical residual assumptions used for OLS inference are not directly applicable to that predictive estimator. For that reason, the notebook evaluates assumptions using a linear style model, ridge regression, on the validation split. This separation aligns with the distinction between explanatory modeling and predictive modeling. Predictive models can improve accuracy without preserving interpretability of coefficients, while explanatory models prioritize unbiased and stable parameter interpretation under assumptions (Shmueli, 2010; Rimpler et al., 2025).

Linearity and homoscedasticity are assessed through the residuals versus fitted plot produced from ridge regression residuals. The residual scatter is centered near zero with no strong nonlinear structure. Some widening in the tails is typical in large datasets, and does not alone imply invalidity. Normality of residuals is assessed via a quantile plot. With a validation sample size exceeding eighteen thousand, minor tail deviations are expected and the practical implication for inference is limited by asymptotic normality of estimators, provided other assumptions are not severely violated (Ernst & Albers, 2017).

Independence of errors is evaluated using the Durbin Watson statistic computed on validation residuals, which is approximately 2.0167, close to the value expected when serial autocorrelation is absent. This check is heuristic in this context because the Kaggle data are not time series, but it still provides a sanity check against gross correlation patterns in residual ordering.

Collinearity among numeric predictors is assessed through the correlation matrix and variance inflation factors. The notebook shows extremely high correlations among the various weight measurements and shell dimensions, and corresponding variance inflation factors that are large for several predictors. This does not automatically invalidate prediction, but it undermines coefficient stability and interpretability in linear models, which motivates ridge regression as a shrinkage method and also motivates the use of a nonlinear ensemble model for predictive competition performance. Collinearity also becomes especially consequential when interaction terms are under consideration, because misspecification and correlated predictors can bias simple effect estimates, reinforcing the need for theory driven model structure decisions (Rimpler et al., 2025).

Influence is evaluated by fitting a statsmodels OLS model on the encoded design matrix and computing leverage and Cook distance. The maximum leverage and Cook distance values are modest, and the table of the most influential cases indicates that no single observation appears to dominate the fit, even though a small subset of cases can still exert more influence than the median.

5 Integrated interpretation and modeling implications

Across all three parts, the shared technical theme is that regression is not simply the mechanical fitting of an equation, but an inferential framework whose trustworthiness depends on assumptions and on the match between theory, design, and functional form. A recurring misconception is to treat normality of variables as the requirement rather than normality of errors, which can cause inappropriate abandonment of linear regression, or misdirected diagnostic practice (Ernst & Albers, 2017). The work here follows the correct logic by inspecting residual diagnostics and influence measures rather than conducting distributional tests on raw predictors.

A second theme is that interaction terms and nonlinear structure are not optional decorations but encode substantive hypotheses. In Part 1, the interaction between grade point average and educational level changes the meaning of the level coefficient and can reverse comparative conclusions at high grade point average. Contemporary simulation work illustrates that omitting a true interaction can produce biased simple effects, even when a misspecified model can sometimes generalize comparably in out of sample prediction under certain noise regimes (Rimpler et al., 2025). This reinforces the doctoral level expectation that model specification must be justified with theory rather than chosen solely by convenience or superficial fit metrics.

A third theme is that applied regression often sits at the boundary of explanation and prediction. The Carseats analysis is explanatory in orientation, emphasizing coefficient interpretation, confidence intervals, and diagnostics. The Kaggle work is predictive in orientation, prioritizing validation RMSE and selecting a nonlinear model that improves predictive accuracy. This distinction is widely recognized in the methodological literature and frames why different model choices are appropriate under different goals (Rimpler et al., 2025; Shmueli, 2010). Moreover, empirical phenomena can follow functional forms that are not strictly linear across the range of a predictor, even when linear approximations work in segments, and model selection should consider saturating and nonlinear alternatives when theory and data support them (Wodarz et al., 2025).

6 References

Ernst, A. F., & Albers, C. J. (2017). Regression assumptions in clinical psychology research practice: A systematic review of common misconceptions. *PeerJ, 5*, e3323. https://doi.org/10.7717/peerj.3323

Rimpler, A., Kiers, H. A. L., & van Ravenzwaaij, D. (2025). To interact or not to interact: The pros and cons of including interactions in linear regression models. *Behavior Research Methods, 57*, 92. https://doi.org/10.3758/s13428-025-02613-6

Shmueli, G. (2010). To explain or to predict. *Statistical Science, 25*(3), 289-310.

Wodarz, M. N., Komarova, N. L., & Ma, T. (2025). The functional form of the association between K-12 student performance and household income in U.S. school districts. *PLOS ONE, 20*(9), e0329296. https://doi.org/10.1371/journal.pone.0329296

6.1 Appendix: Reproducibility notes

The complete code requested by the assignment is provided in the two submitted notebooks. Part 1 and Part 2 computations and outputs are in `Todd_DeLozier_Assignment_2_Python_Code_Parts_1_and_2.ipynb`. Part 3 computations, preprocessing pipelines, model evaluation, diagnostic plots, and the generation of `submission.csv` are in `Todd_DeLozier_Assignment_2_Python_Code_Part_3.ipynb`. If this report is knitted on the same machine as the course materials, the Kaggle proof image will render from the provided assets path. If the file is not present at knit time, the report will still record the expected file location for verification.