Assignment 2

ST5227 Applied Statistical Learning
Qi Shuoli E1124031 A0274285W

Classification and Generalized Linear Regression with Probabilistic Models

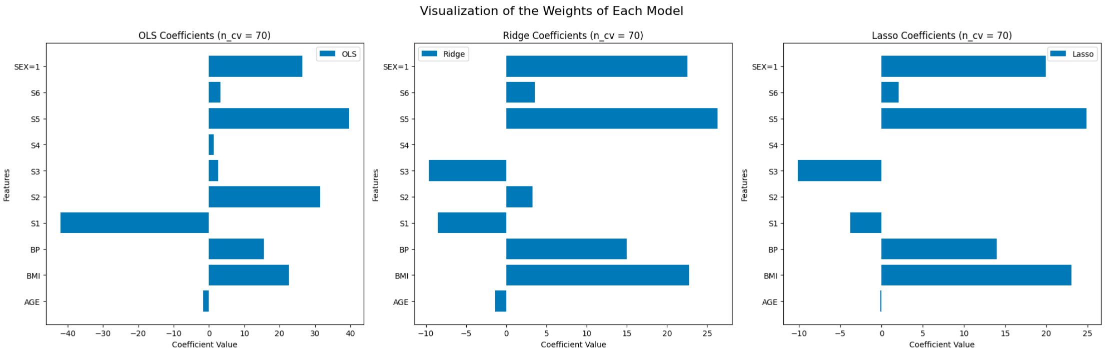
Problem 1

I standardize the predictors before using ridge regression and Lasso. Then, I increased the number of cross-validation splits to 70 for a more robust analysis rerun the experiment with an increased number of cross-validation splits and analyze your results using hypothesis tests paired t-test to select the penalty parameter λ and compare the predictive performance of the models.

the average squared prediction error of LinearRegression: 55.6181
the average squared prediction error of RidgeCV: 55.7450
the average squared prediction error of LassoCV: 55.7366

When I run this, Lasso wins. consequently, after implementing three models, I want to test whether one model consistently outperforms another, or whether this result is just noise.

Explore the weights for each of the models by visualizing them. When I run this, the weights for the three models look slightly different. Lasso does set some of the coefficients to zero. OLS has positive S3 coefficients, but Ridge coefficients are negative.



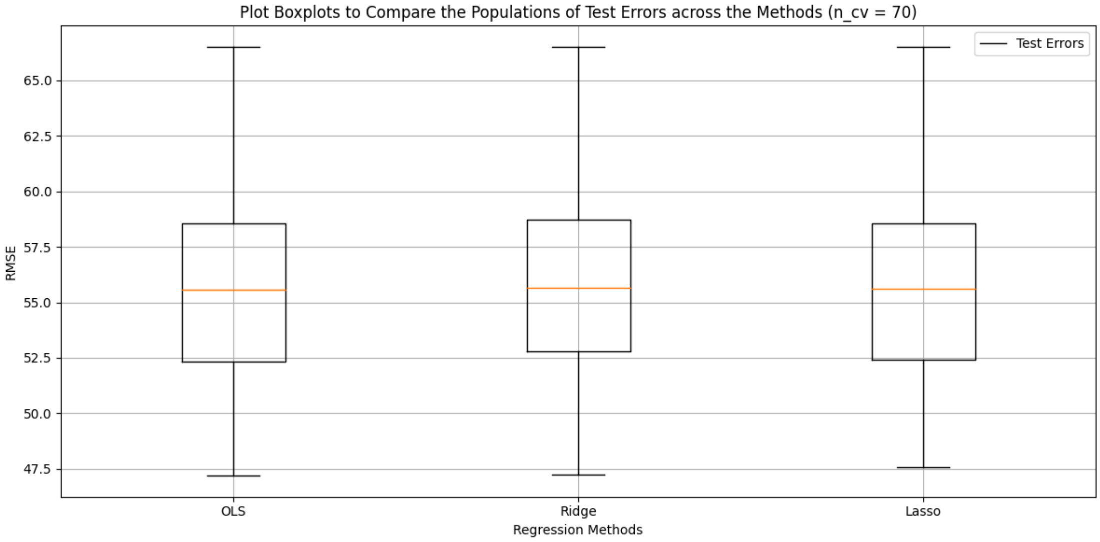
Take a look at some penalty parameters that are selected by the `RidgeCV` and `LassoCV` methods. The following parameters are those found with the last test set:

Ridge penalty found by CV: 11.1120
Lasso penalty found by CV: 1.0111

Consider the noise over the test sets. Plot boxplots to compare the populations of test errors across the methods.

Run a statistical (hypothesis) test that two of these sets of scores significantly differ using the `scipy.stats.ttest_rel()` method to run a paired t-test, which tests whether the means of two paired samples differs significantly.

OLS average test error: 55.6181
Ridge average test error: 55.7450
Lasso average test error: 55.7366



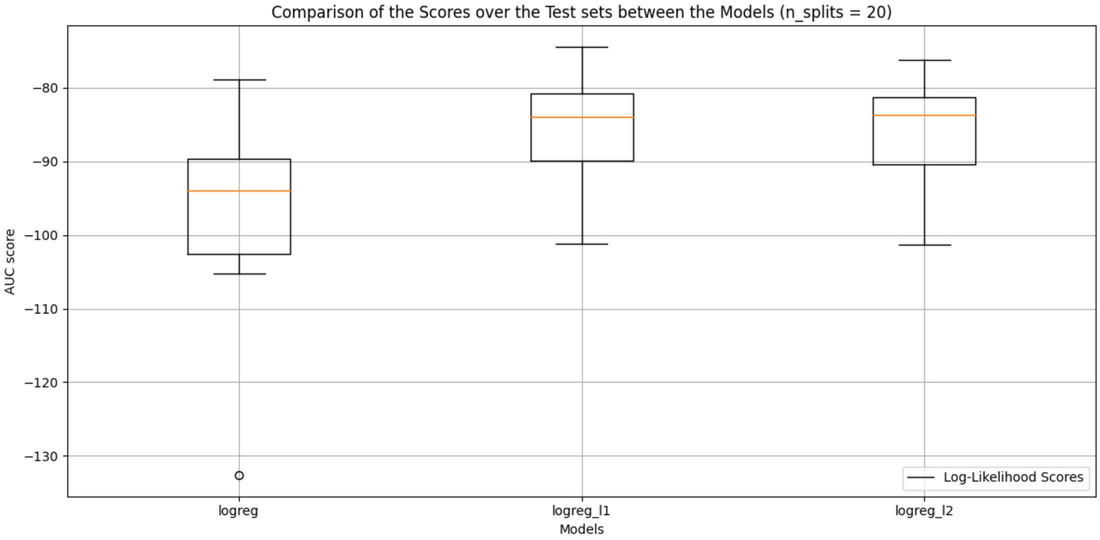
p-value of paired t-test between OLS and Ridge: 0.0273
p-value of paired t-test between OLS and Lasso: 0.0384
p-value of paired t-test between Lasso and Ridge: 0.8324

The p-values of look significant for p-value of paired t-test between OLS and Ridge as well as p-value of paired t-test between OLS and Lasso. So I would say there is good evidence that one is consistently outperforming the other. Considering the mean errors, the OLS should be slightly outperform others.

Problem 2

Define the log-likelihood functions for both classification and count outcomes. Repeat both the classification experiment (predicting ’DEFAULT’) and the count regression experiment (predicting ’CREDIT’) using the log-likelihood as a performance metric.

Visualize a comparison of the scores over the test sets between the models in each case.

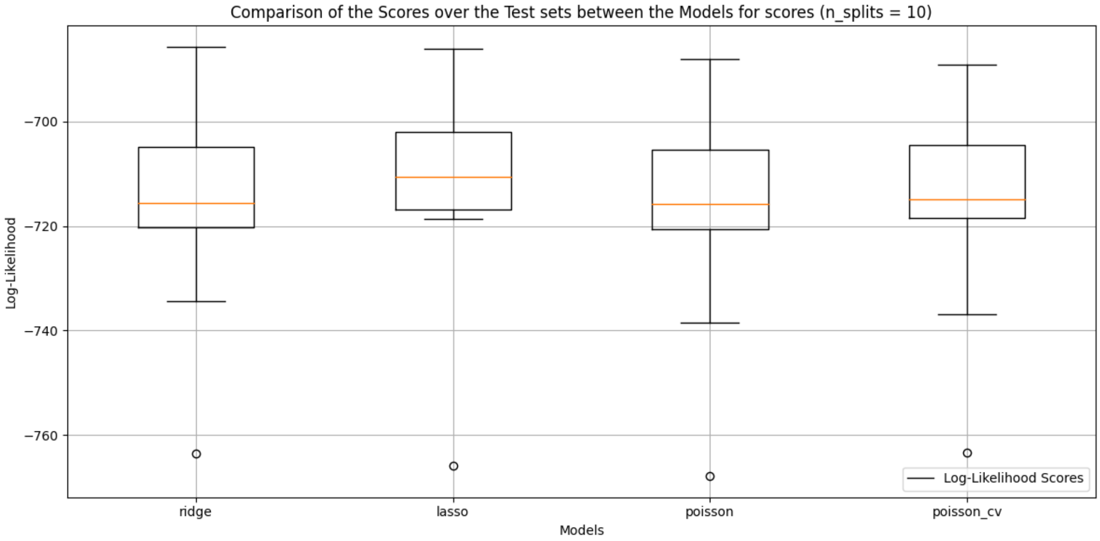


p-value of Wilcoxon signed-rank test between regularization (logreg) and regularization (logreg_l1): 0.000004
p-value of Wilcoxon signed-rank test between regularization (logreg_l1) and regularization (logreg_l2): 0.595819
p-value of Wilcoxon signed-rank test between regularization (logreg) and regularization (logreg_l2): 0.000002
p-value of Paired T-test between regularization (logreg) and regularization (logreg_l1): 0.000056
p-value of Paired T-test between regularization (logreg_l1) and regularization (logreg_l2): 0.797080
p-value of Paired T-test between regularization (logreg) and regularization (logreg_l2): 0.000069

The p-values of Wilcoxon signed-rank test and the paired t-test comparing the log-likelihood scores of these models suggest significant differences in performance between models without regularization (logreg) and those with regularization (logreg_l1 and logreg_l2).

Comparisons between L1 and L2 regularization models (logreg_l1 vs. logreg_l2) showed no significant difference in one of the tests, suggesting that the choice between L1 and L2 regularization may not significantly affect the model performance in this specific context, or the differences are not captured by the log-likelihood metric used.

The significant p-values indicate that the regularization technique employed can have a statistically significant effect on the logistic regression model's predictive performance, as measured by log-likelihood. This highlights the importance of selecting the appropriate regularization method based on the data and the specific problem being addressed.



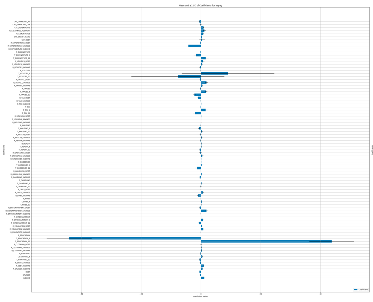
```
p-value of Wilcoxon signed-rank test between ridge and lasso: 0.105469
p-value of Wilcoxon signed-rank test between ridge and poisson: 0.048828
p-value of Wilcoxon signed-rank test between ridge and poisson_cv: 0.625000
p-value of Wilcoxon signed-rank test between lasso and poisson: 0.009766
p-value of Wilcoxon signed-rank test between lasso and poisson_cv: 0.037109
p-value of Wilcoxon signed-rank test between poisson and poisson_cv: 0.048828
---------------------
p-value of paired t-test between ridge and lasso: 0.081512
p-value of paired t-test between ridge and poisson: 0.042480
p-value of paired t-test between ridge and poisson_cv: 0.703408
p-value of paired t-test between lasso and poisson: 0.023075
p-value of paired t-test between lasso and poisson_cv: 0.056825
p-value of paired t-test between poisson and poisson_cv: 0.052897
```

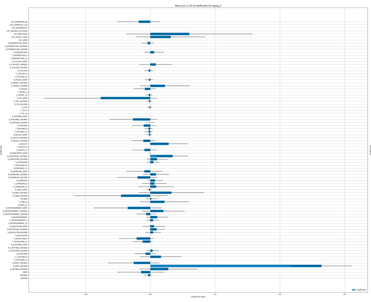
The p-values of the Wilcoxon signed-rank test and the paired t-test comparing the log-likelihood scores of these models suggest significant differences in performance between Ridge and Poisson, Lasso and Poisson.

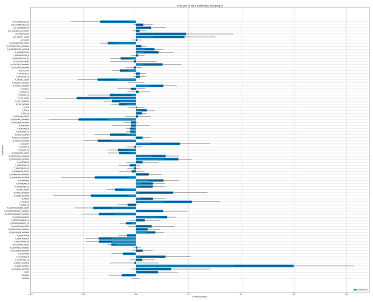
Comparisons between Ridge and Lasso, Ridge and Poisson showed no significant difference in one of the tests, suggesting that the choice between Ridge and Lasso, Ridge and Poisson may not significantly affect the model performance in this specific context or the differences are not captured by the log-likelihood metric used.

The significant p-values indicate that the regularization technique employed can have a statistically significant effect on the logistic regression model's predictive performance, as measured by log-likelihood. This highlights the importance of selecting the appropriate regularization method based on the data and the specific problem being addressed.

Problem 3





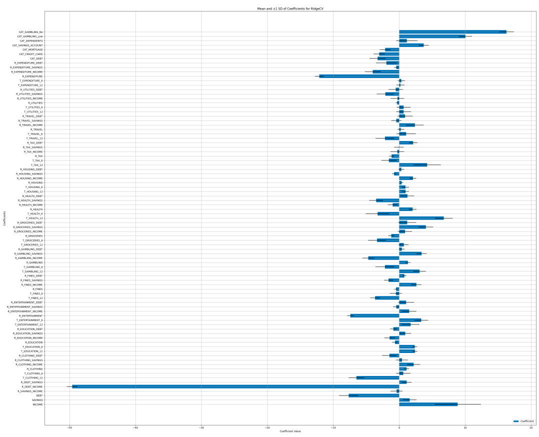


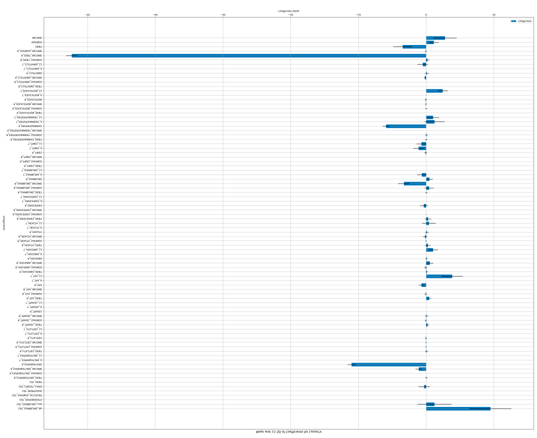
Here are the plots of horizontal bar plot with ± one standard deviation bar for default.

According to the results from Logistic Regression, R_EXPENDITURE_SAVINGS, T_UTILITIES_6, T_UTILITIES_12, T_EDUCATION_6, T_EDUCATION_12 has a negative or positive high coefficient. Among these features, R_EXPENDITURE_SAVINGS, T_EDUCATION_6, and T_EDUCATION_12 have low standard deviations. Consequently, R_EXPENDITURE_SAVINGS, T_EDUCATION_6, and T_EDUCATION_12 have relatively higher coefficients and magnitudes, and higher certainty in the case.

According to the results from Logistic Regression (l1), only R_DEBT_INCOME has a negative or positive high coefficient and also low standard deviations. Consequently, R_DEBT_INCOME has relatively higher coefficients and magnitudes, and higher certainty in the case.

According to the results from Logistic Regression (l2), CAT_GAMBLING_No, CAT_CREDIT_CARD, R_EXPENDITURE_INCOME, T_EXPENDITURE_12, R_TRAVEL_INCOME, R_TAX_DEBT, R_HOUSINGG_SACINGS, R_HEALTH_INCOME, R_GROCERIES_INCOME, T_FINES_6, T_EDUCATION_6, T_EDUCATION_12, R_DEBT_INCOME have relatively higher coefficients and magnitudes, and higher certainty in the case. Consequently, CAT_GAMBLING_NO CAT_CREDIT_CARD, R_EXPENDITURE_INCOME, T_EXPENDITURE_12, R_TRAVEL_INCOME, R_TAX_DEBT, R_HOUSINGG_SACINGS, R_HEALTH_INCOME, R_GROCERIES_INCOME, T_FINES_6, T_EDUCATION_6, T_EDUCATION_12, R_DEBT_INCOME have relatively higher coefficients and magnitudes, and higher certainty in the case.





Model: RidgeCV

Coefficients with high mean and low std:

Coefficient: T_EDUCATION_12, Mean: 2.3701, Std: 0.3534

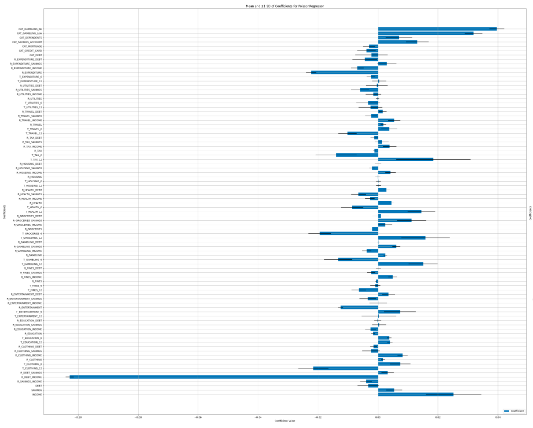
Coefficient: T_EDUCATION_6, Mean: 2.3631, Std: 0.3379

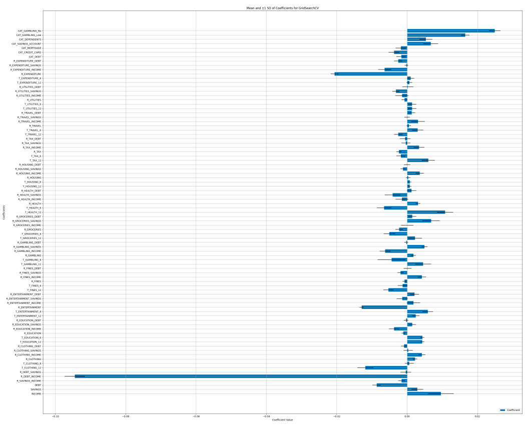
Coefficient: R_HOUSING_INCOME, Mean: 2.0858, Std: 0.4850

Model: LassoCV

Coefficients with high mean and low std:

Coefficient: T_EDUCATION_6, Mean: 0.0201, Std: 0.0270





Model: PoissonRegressor

Coefficients with high mean and low std:

Coefficient: T_EDUCATION_12, Mean: 0.0040, Std: 0.0009

Coefficient: R_HEALTH, Mean: 0.0046, Std: 0.0008

Model: GridSearchCV

Coefficients with high mean and low std:

Coefficient: T_EDUCATION_12, Mean: 0.0043, Std: 0.0006

Coefficient: T_EDUCATION_6, Mean: 0.0043, Std: 0.0006

Coefficient: R_GAMBLING_SAVINGS, Mean: 0.0050, Std: 0.0008

Coefficient: R_HEALTH, Mean: 0.0031, Std: 0.0006

Here are the plots of horizontal barplot with ± one standard deviation bar for default.

According to the results from RidgeCV, T_EDUCATION_12, T_EDUCATION_6, and R_HOUSING_INCOME have relatively higher coefficients and magnitudes, and higher certainty in the case. Consequently, T_EDUCATION_12, T_EDUCATION_6, and R_HOUSING_INCOME have relatively higher coefficients and magnitudes, and higher certainty in the case.

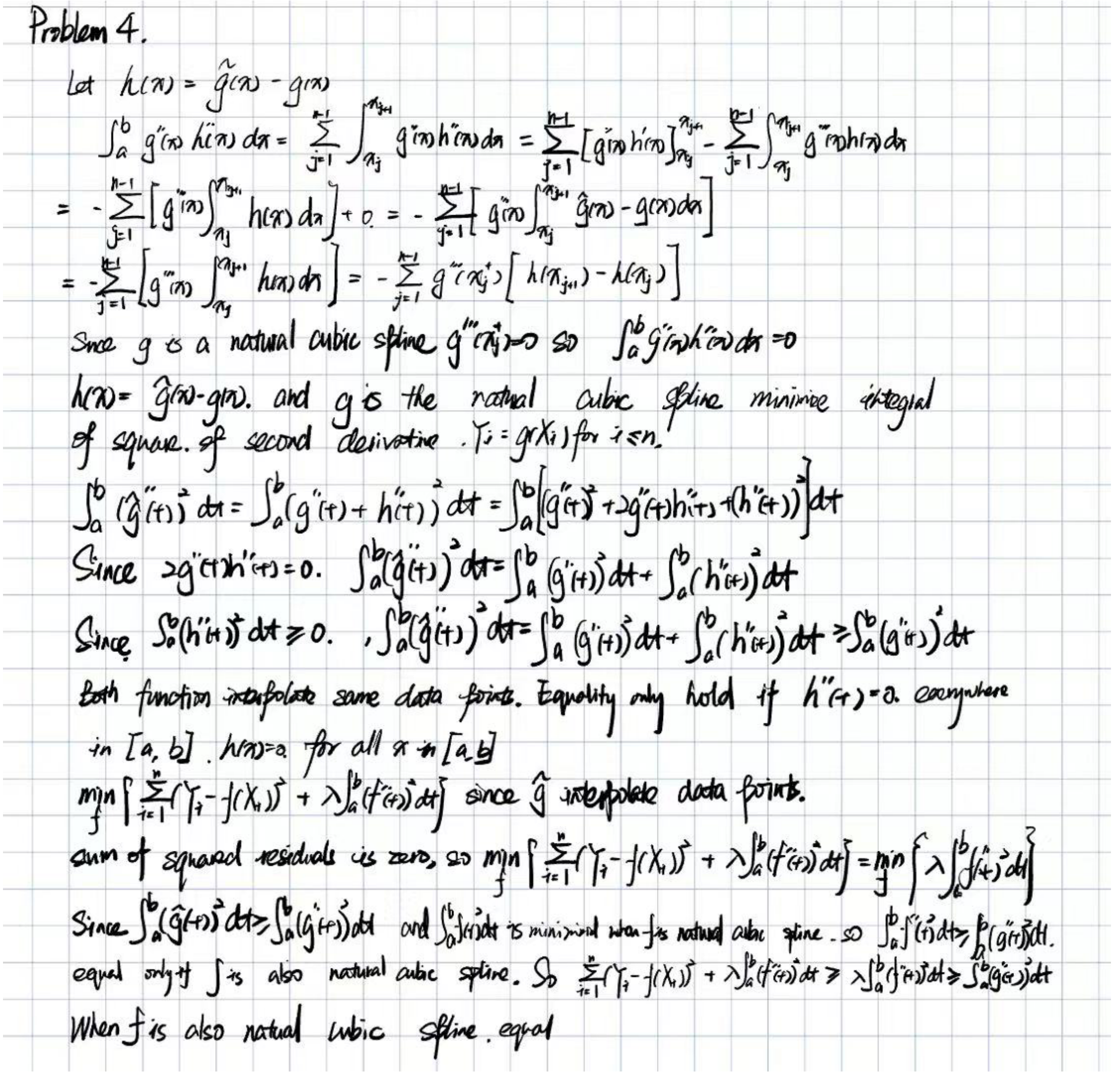
According to the results from LassoCV, T_EDUCATION_6 has a negative or positive high coefficient and also low standard deviations. Consequently, T_EDUCATION_6 has relatively higher coefficients and magnitudes, and higher certainty in the case.

According to the results from PoissonRegressor, T_EDUCATION_12, and R_HEALTH have a negative or positive high coefficient and also low standard deviations. Consequently, T_EDUCATION_12, and R_HEALTH have relatively higher coefficients and magnitudes, and higher certainty in the case.

According to the results from GridSearchCV, T_EDUCATION_12, T_EDUCATION_6, R_GAMBLING_SAVINGS, and R_HEALTH have a negative or positive high coefficient and also low standard deviations. Consequently, T_EDUCATION_12, T_EDUCATION_6, R_GAMBLING_SAVINGS, and R_HEALTH have relatively higher coefficients and magnitudes, and higher certainty in the case.

Nonlinear Regression with Splines

Problem 4



Problem 4.

Let $h(x) = \tilde{\hat{g}}(x) - g(x)$

$$\int_a^b g''(x)\, h''(x)\, dx = \sum_{j=1}^{n-1} \int_{x_j}^{x_{j+1}} g''(x) h''(x)\, dx = \sum_{j=1}^{n-1} \left[g''(x) h'(x) \right]_{x_j}^{x_{j+1}} - \sum_{j=1}^{n-1} \int_{x_j}^{x_{j+1}} g'''(x) h'(x)\, dx$$

$$= -\sum_{j=1}^{n-1} \left[g'''(x) \int_{x_j}^{x_{j+1}} h'(x)\, dx \right] + 0 = -\sum_{j=1}^{n-1} \left[g'''(x) \int_{x_j}^{x_{j+1}} \hat{g}'(x) - g'(x)\, dx \right]$$

$$= -\sum_{j=1}^{n-1} \left[g'''(x) \int_{x_j}^{x_{j+1}} h'(x)\, dx \right] = -\sum_{j=1}^{n-1} g'''(x_j^+) \left[h(x_{j+1}) - h(x_j) \right]$$

Since g is a natural cubic spline $g'''(x_j^+) = 0$ so $\int_a^b g''(x) h''(x)\, dx = 0$

$h(x) = \hat{g}(x) - g(x)$. and g is the natural cubic spline minimize integral of square. of second derivative. $Y_i = g(X_i)$ for $i \leq n$.

$$\int_a^b (\hat{g}''(t))^2\, dt = \int_a^b (g''(t) + h''(t))^2\, dt = \int_a^b \left[(g''(t))^2 + 2g''(t)h''(t) + (h''(t))^2 \right] dt$$

Since $2g''(t)h''(t) = 0$. $\int_a^b (\hat{g}''(t))^2\, dt = \int_a^b (g''(t))^2\, dt + \int_a^b (h''(t))^2\, dt$

Since $\int_a^b (h''(t))^2\, dt \geq 0$. , $\int_a^b (\hat{g}''(t))^2\, dt = \int_a^b (g''(t))^2\, dt + \int_a^b (h''(t))^2\, dt \geq \int_a^b (g''(t))^2\, dt$

Both function interpolate same data points. Equality only hold if $h''(t) = 0$. everywhere in $[a, b]$. $h(x) = 0$ for all x in $[a, b]$

$\min_f \left\{ \sum_{i=1}^n (Y_i - f(X_i))^2 + \lambda \int_a^b (f''(t))^2\, dt \right\}$ since $\hat{g}$ interpolate data points.

sum of squared residuals is zero, so $\min_f \left\{ \sum_{i=1}^n (Y_i - f(X_i))^2 + \lambda \int_a^b (f''(t))^2\, dt \right\} = \min_f \left\{ \lambda \int_a^b f''(t)^2\, dt \right\}$

Since $\int_a^b (\hat{g}''(t))^2\, dt \geq \int_a^b (g''(t))^2\, dt$ and $\int_a^b f''(t)^2\, dt$ is minimized when f is natural cubic spline. so $\int_a^b f''(t)^2\, dt \geq \int_a^b (g''(t))^2\, dt$. equal only if f is also natural cubic spline. So $\sum_{i=1}^n (Y_i - f(X_i))^2 + \lambda \int_a^b (f''(t))^2\, dt \geq \lambda \int_a^b f''(t)^2\, dt \geq \int_a^b (g''(t))^2\, dt$

When f is also natural cubic spline. equal

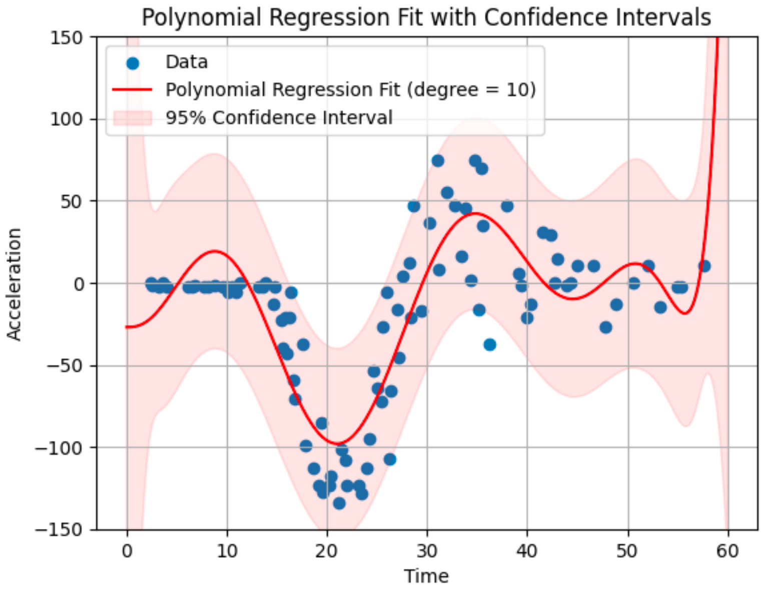
Problem 5

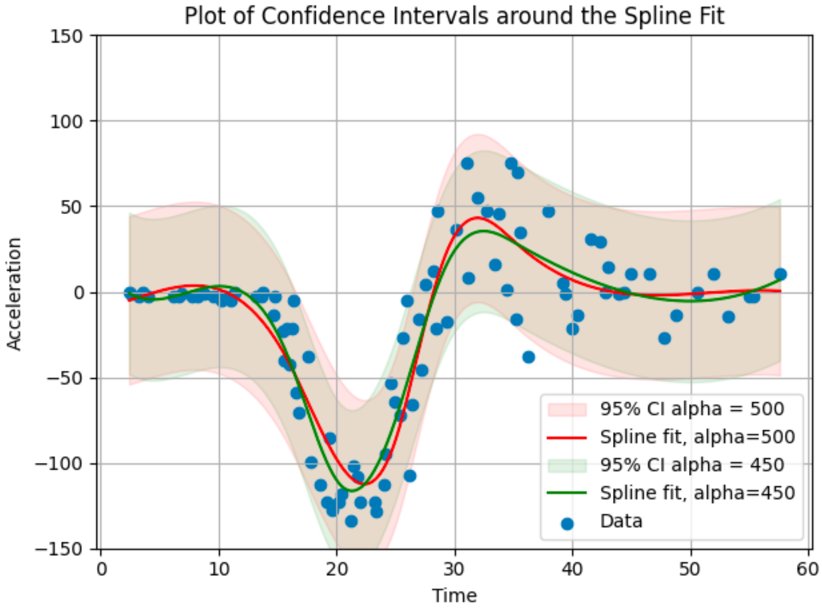
Let $f(X)$ denote your polynomial regression or spline model and assume that $Y = f(X) + \varepsilon$ where $\varepsilon \sim N(0, \sigma 2)$, for some parameter $\sigma^2 > 0$ to be estimated from the data, and use the results from Lecture 1:

$$\left[\hat{Y} - 2\sigma\sqrt{\mathbf{x}^T(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x}},\ \hat{Y} + 2\sigma\sqrt{\mathbf{x}^T(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x}} \right].$$

which will contain the true value $x^\top\beta$ with probability ≈ 95%

Here is the plot outcome of confidence intervals around the function $f(X)$:





This is the t-distribution's inverse cumulative distribution providing the critical value for a 95% confidence interval. Since a 95% CI captures the central 95% of the distribution, each tail contains 2.5%, and the cumulative probability up to the end of the right tail is 0.975. The second argument (df) specifies the degrees of freedom, which typically equals the number of observations minus the number of parameters estimated by the model.