**STA 3180 Fall 2024**

**LAB 4: Polynomial Regression**

**Set-up:** The Kentucky Derby is an annual horse race run at Churchill Downs in Louisville, Kentucky, on the first Saturday in May. The race is known as the “Most Exciting Two Minutes in Sports” and is the first leg of horse racing’s Triple Crown.

**Dataset:** The dataset **KDerby** contains information on each running of the Kentucky Derby since 1875. The Derby was first run at a distance of 1.5 miles but in 1896 the distance changed to its current 1.25 miles.

1. Create a labeled scatterplot of the winning times (*Time,* in seconds,on the vertical axis) against the Year (horizontal axis). Describe any patterns you see. Are there any unusual observations? How do you suggest proceeding with the analysis of how the performance of the horses has changed over the years?

The scatterplot shows that winning times have decreased over the years. This could be attributed to improved horse breeding, training, and possibly even track conditions. Additionally, the decrease in winning times seems to have slowed/plateaued. When it comes to unusual observations, there seem to be some, which may be attributed to factors such as weather conditions or “off days” with individual horse races. Moreover, given the pattern of marginal declining times, it may be worth fitting a nonlinear model to capture the curvature of the model (which is already seen by the scatterplot).

2. A new variable was created, included in the dataset, that calculates the winning horse’s speed (in miles per hour) taking into consideration track length called “Speed”. Produce a scatterplot of the speeds of the winning horses (in miles per hour) vs year. What is the overall pattern in these speeds over the years? Why does this pattern make sense in this context? Are there any unusual observations now? If so, identify by name. Does it make sense to fit a linear model to these data?

The scatterplot of Speed vs. Year shows an increase in winning speeds over the years. Similarly to the winning times scatterplot, the rate of increase in speed has slowed/plateaued in more recent years. Additionally, the pattern is consistent with advancements in horse breeding, training, etc. There is also potential that speeds increased due to the derby distance being shortened in 1896, so the horses would appear to be faster. Like the winning time data, there are also some outliers, with a quadratic or polynomial model being more appropriate for the marginal differences in speeds.

3. Fit a quadratic model to predict speed from year and year2. Include a residual analysis for the model. Does the model appear to capture the curvature in the data? Provide justification in your answer.

The residual analysis seems to capture the curvature in the data. This is because the residuals are randomly scattered around zero. Therefore, a quadratic or more complex model seems appropriate to capture the data.

4. Provide an interpretation of the intercept of this model. Does it make sense in this context? Explain.

The intercept (-9.888e+02) represents the predicted winning speed at year 0 (before the start of the Kentucky Derby in 1875). Therefore, it doesn’t make sense in this context as the intercept is based on extrapolating back to a time when data isn’t provided and thus lacks meaningful interpretation.

5. Write out the prediction equation for the quadratic model. Is the coefficient of year2 positive or negative and what does that imply about the relationship between speed and year as year increases?

Speed = -9.888045e+02 ​+ 1.030260e+00​\*Year - 2.587208e-04\* Year²

The coefficient of Year² is negative, which implies that while speeds increase over time, they have plateaued in more recent years (which aligns with marginal returns in performance improvements).

6. We expect year2 and year to be related to each other, but for the range of data values at which we are looking, is the relationship linear? Produce a scatterplot of year2 vs year. What do you learn?

The scatterplot of Year² vs. Year appears to be perfectly linear. This is expected as the quadratic term is simply the square of the Year variable and is directly related.

***Note: Standardizing the variable before applying a polynomial model moves the curved part of the association between the linear and quadratic terms into the middle of our explanatory variable region, reducing the linear association between the linear and quadratic terms. The variable StdYear is included in the dataset.***

7. Produce a scatterplot of standardized year (StdYear) vs (StdYear)2. Is the association still linear?

The scatterplot of StdYear vs. StdYear² shows a perfect quadratic relationship and reduces the collinearity between the two terms.

8. Fit a new model using Std Year and (StdYear)2 to predict winning speeds. What does the plot of the standardized quadratic model look like? Interpret the intercept of this model.

Speed = 36.33966 + 0.96192\*StdYear - 0.45017\*(StdYear)2

The intercept of the standardized model (36.33966) represents the predicted speed when StdYear = 0. This interpretation makes more sense contextually, as the intercept now is not an extrapolated value outside the data range.

9. Answer each of the following questions. Make sure that you not only answer the question, but also make it clear how you are making your decision.

a) Is the standardized quadratic model useful in explaining variation in winning speeds?

The standardized quadratic model is useful in explaining variation in winning speeds, explains a significant portion of the variance, and produces residuals that are randomly distributed without pattern. This is because the R2 value is high (0.7522) and the p-value (2.2e-16) is significant at alpha = 0.05.

b) Is the standardized quadratic model statistically significant?

The model is statistically significant as the p-value is 2.2e-16. This indicates that both StdYear and (StdYear)2 are important predictors of winning speed.

c) What is your prediction for 2019 using the standardized model? Do you find this prediction appropriate?

The predicted speed for 2019 is 36.65162. This prediction seems to be appropriate as the speed in 2018 was 36.2318841 and isn’t marginally different than the predicted speed.

10. Now fit a cubic model to the data. Does the cubic model explain significantly more variation in the speeds than the quadratic model? How are you deciding? Is the cubic term significant?

The cubic model doesn’t explain significantly more variation in speeds than the quadratic model. Using the ANOVA test doesn’t significantly reduce the residual sum of squares or produce significant term p-values below the typical significance threshold of alpha = 0.05. Thus, the cubic model is not better at capturing more complex trends in the data than the quadratic model.