Olympic Data and What it can Tell Us

Parker Joncus, Evan De Broux, Margaret Hoeller
pjoncus@iit.edu, edebroux@iit.edu, mhoeller@hawk.iit.edu

33.3% 33.3% 33.3%

Math 484

November 2018

Contents

1 Abstract

The Olympics are an exciting opportunity to see a snapshot of the height of sports performance over a range of events. Although the lineup of events is changeable, some events have persisted throughout the entire modern games. One of these events is track and field. Although the same events have occurred every four years since 1896, records in these events continue to be smashed each games. How long will we continue to do so? It is highly improbable for humanity to reach a zero second 100M Dash, but is a five second 100M Dash possible? At some point, without exterior help, there must be a minimum time or maximum human performance. We explore this hypothesis using statistical analysis, and find that the answer varies depending on event and the quality of the data gathered.

2 Introduction

Every four years, countries all over the world gather to compete in the Summer Olympics. Athletes train their entire lives to qualify, and winning an event is a point of personal and national pride. First begun in 776 BC, the games continued for twelve centuries before being interrupted by Christianity's fervor for destroying pagan practices. Restarted in the globalization of 1896 they keep the original spirit of competition and cooperation.

In addition to serving as a symbol of unity across the globe, the Olympics offer a unique opportunity to spectate a wide range of the highest level of sports at the same event. This has been used to collect data on the best human performance across many events, and this data can be used to test hypotheses about the height of human ability. Analysis of this data can help us understand the probability of decreasing the fastest time for distance, or what factors can influence a country's performance in the games.

Our group explored questions with the Olympic data available. The first of these questions is if we are still improving in each of the events. It seems as though there are records that are broken every year, while other records will stand for many years. If we are continuing to get better year after year, then it is expected that every record will eventually be broken. We expect that we are improving over the years as techniques, strength, and agility are perfected and since sports have become a very popular over the years. This then leads to the next question, for how long will we be able to continue breaking records and what is the limit on the events participated in. It is obvious that there is a limit of 0 seconds for timed events, but what is the fastest a person can ever run in an event? We hope to find a model that fits the data well and will converge to a value such that we can call that value our theoretical optimal result.

While exploring the data we plan on using many different models and transformations on the data to compare how the transformations and models perform. We are interested in their performance for the current data, short term predictions, and long term predictions. We can then use these performances to see how relevant or realistic the models are.

3 Data Sources

We utilized data from:

Athlete Performance	Track and Field Times and Scores from 1896-2016[3]	The Guardian
Athlete Biometric Data	120 years of Olympic History: Athletes and Results[1]	The Guardian

The data can also be obtained from our Github repository at:
https://github.com/parkerjoncus/Olympic-Results-Regression

4 Analysis Process

There is a wealth of data on Olympic medals, some stretching back all the way to the first modern games in 1896. We explored the question: At what rate is the best performance improving? Are there events in which we are unlikely to beat the current Olympic record?

4.1 Data Collection and Description

The dataset consists of two types of track and field events: running events and field events. Performance in running events is measured in seconds and in field events are measured in meters. We can compile the podium finishes from the Olympic data for both the men's and women's track and field events. We decided only to attempt to model 5 popular events: the 100 meter dash, the 200 meter dash, the 1500 meter run, shot put, and the long jump.

For each event at each game, we found: the names of the athletes who won a medal [3] [1], their performance in seconds or meters [3], and their height, weight, and age [1].

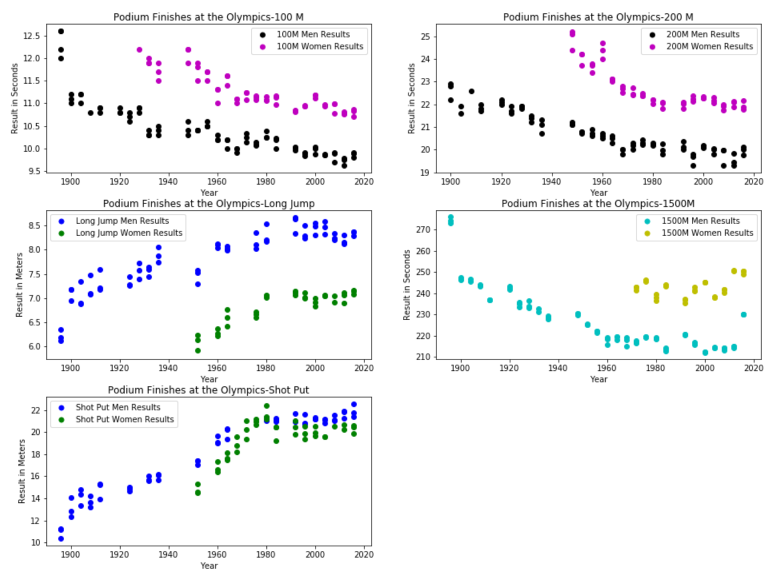


Figure 1: Scatter plots of medal earning results from the Olympics in the 100 Meter Run, 200 Meter Run, 1500 Meter Run, Shot Put, and Long Jump.

Two particular trends are apparent from the scatter and correlation plots of the data (See Appendix, Figure 8, 24, 25, 26). For the running events, with the exception of the women's 1500 meter run, there is a clear negative relationship between time and results, with the possible horizontal asymptotes appearing in each of these events. For the field events there is a clear positive relationship between time and results, again with possible horizontal asymptotes appearing as upper bounds. There also appear to be some curvilinear relationships between time and result for some events as well. It is also very clear that Year is the most important predictor, and so since we care more about how we perform over time, we can create models just based on the Year.

4.1.1 Simple Linear Regression

For every single simple linear regression model, each $\hat{\beta}_1$ estimated using ordinary least squares method yielded significant results for a linear association between time and results with error controlled at a $\alpha = 0.05$ level, with an exception for the women's 1500 meter run.

With the exception of Women's 1500M and Shot Put, simple linear models had fair R^2 values indicating a reasonable capture of the response. However, there do appear to be some curvilinear relationships between the predictor variable and the response based on some of the plots, so it is definitely worthwhile to run lack-of-fit tests on the simple linear regression models. See Figure 2, plots of the regressed equations, in the appendix.

Figure 2: Adjusted-R^2 Value for Linear Model on Performance Data

	100M Dash	200M Dash	1500M Run	Shot Put	Long Jump
Men	0.732	0.864	0.677	0.929	0.788
Women	0.758	0.678	0.081	0.396	0.726

4.1.2 Simple Linear Regression – Diagnostics

In order to determine whether there is a curvilinear response, we tested the constant error variance with a Breusch-Pagan Test and the normality of the residuals with normal probability plots. As mentioned above, we also want to test the adequacy of our simple linear models. To test the adequacy of the model, we also conducted a Lack-of-Fit test.

4.1.3 Breusch-Pagan Tests for Constant Error Variance

A Breusch-Pagan test is to test form departures in constancy of the error terms. This test is carried out as follows:

We regress the function:

$$log(\sigma_i^2) = \gamma_0 + \gamma_1 x_{i1} \tag{1}$$

$$H_0 : \gamma_1 = 0 \tag{2}$$

$$H_1 : \gamma_1 \neq 0 \tag{3}$$

Test statistic:

$$\frac{SSR^*/2}{(SSE/n)^2} \tag{4}$$

SSR^* is the regression sum of squares for the log regression of the residuals. The critical value for the rejection region would be $\chi^2_{\alpha/2;1}$, and we tested at an $\alpha = 0.05$ level.

Figure 3: P-Values of Breusch-Pagan Tests for Linear Model on Performance Data

	100M Dash	200M Dash	1500M Run	Shot Put	Long Jump
Men	6.601e-4	0.8201	0.2545	0.2529	0.0449
Women	0.0212	0.0048	0.1016	0.0073	0.0739

At a significance level of $\alpha = 0.05$, the Men's 100M , the Women's 100M, the Women's 200M, the Men's Long Jump, and the Women's Shot Put do not have constant error variance. This implies that some factors are missing from our model.[2]

4.1.4 Normal Probability Plots

A normal probability plot of the residuals plots the residual against its expected value under the assumption of normality. A nearly linear plot would indicate that the error terms, and thus the response is normal, but plots that do not appear to be linear suggest that the distribution of the error terms is non-normal.

Based on the normal probability plots listed as Figure 10 in the appendix, there are departures from normality, some of which are major departures from the normal assumption, like in the Men's 100 meter dash, the Women's 200 meter dash, the Men's and Women's 1500 meter run, and the Men's long jump.

4.1.5 F-tests for Lack of Fit

Given our concerns about non-constant error terms and non-normal error terms, we tested the simple linear models for lack of fit using the F-test for lack of fit.

Since we have repeat observations at each independent predictor variable, we are able to run an F test for lack of fit. The F-test for lack of fit is carried out as follows:

$$H_0 : E(Y) = \beta_0 + \beta_1 X \tag{5}$$

$$H_1 : E(Y) \neq \beta_0 + \beta_1 X \tag{6}$$

The test statistic here would be:

$$F^* = \frac{SSLF}{c-2} / \frac{SSPE}{n-c} = \frac{MSLF}{MSPE} \tag{7}$$

where c is the number of years with replicates and $n = \sum_{j=1}^{c} n_j$ and n is the number of observations.

Our decision rule is:

$$F^* \leq F(1-\alpha; c-2, n-c) \tag{8}$$

then we conclude that our simple linear model is appropriate.

$$F^* > F(1-\alpha; c-2, n-c) \tag{9}$$

then we conclude that our simple linear model is not appropriate for our data. We controlled the error at $\alpha = 0.05$.

Each lack of fit test led us to the conclusion that the simple linear models we used to predict results were inadequate for predicting the results of these selected track and field events in the Olympic finals. Combining this with the fact that the error terms may not be constant and/or normally distributed, we chose to investigate the possibility of using higher order terms to better fit the data and use Box-Cox transformations of the response variable in order to correct the skewness of the distributions of error terms, non-constant error variance, or non-linearity of the regression function we are trying to use.

Figure 4: F-test for Lack of Fit on Performance Data - p-values

	100M Dash	200M Dash	1500M Run	Shot Put	Long Jump
Men	<2.2e-16	2.518e-08	<2.2e-16	1.47e-10	1.948e-09
Women	8.394e-09	4.959e-13	1.765e-14	7.022e-10	2.686e-06

4.1.6 Higher Order Multiple Regression – Subset Regression and Diagnostics

We quickly realized that linear regression is not the best option for predicting times or distances in the Olympics. When plotting the times/distances of the individual events,it can easily be seen that there is a curve to the data points. This is a good indication that we should try higher order of input. Lets assume that the data is of the form:

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 \tag{10}$$

We chose to go up to cubic order because we do not want to create a model that over fits and curves up and down and up and down. We also did not want to assume that every event was of the same form, so we used step regression to select the best model up to cubic for each individual event.

Each event used selected a different type of model. Some selected the full model, while others selected just cubic or just squared. In all of the events, the adjusted r-squared values increased as compared to the linear models. This is a good indication that the models fit the data much better.

We then plotted (Figures 29-31) and summarized the data and found that the Q-Q plots were still not very good for most of the events, so in order to try to fix this we wanted to try a Box-Cox transformation.

4.1.7 Higher Order Multiple Regression – Box-Cox transformation

Due to the skew in most of our models based on their normal probability plots, unequal error variances, and the lack of fit of the simple linear regression functions, it was worthwhile to investigate a possible transformation of the observed values of the response variable to try to better predict performance results. The transformed response variable, say Y', has form:

$$Y' = Y_i^{\lambda} = \beta_0 + \beta_1 x_{i1} + \epsilon_i \tag{11}$$

Where β's are found using maximum likelihood estimates, and the error terms are independently and normally distributed with mean 0 and variance σ^2. Y' has the equivalent transformation of form:

$$Y' = \frac{Y^{\lambda} - 1}{\lambda}, \lambda \neq 0 \tag{12}$$

$$Y' = log(Y), \lambda = 0 \tag{13}$$

We first found the λ's for each event that maximizes the log-likelihood of the transformation. Trivially each λ is different for the different events since the data is much different. We then created a model from the transformed data using the corresponding λ and found that the adjusted r-squared value was also higher than the linear regression models. This is as expected since the data has a curve to it and the Box-Cox allows for a model with a curve.

4.1.8 Logarithmic Transformation

The last type of transformation we wanted to try was a log transformation. We decided to do this because when the data is plotted for the running events, it has a decreasing curve that levels off. We also can assume that there is some limit for each event. For example we can assume that nobody will be able to run a mile in 10 seconds in the near future. Therefore, we wanted to have a function that as it goes to infinity, its limit goes to some constant. Therefore, we are assuming the data is of the form:

$$Y = e^{-\beta X} + \beta_0 \tag{14}$$

This is mainly just for the running events. The throwing events will have models that not great fits since their data is increasing over time rather than decreasing like the running events. The hope is to find a model that fits well and so as $X \to \infty, Y \to \beta_0$. Therefore we can hypothesis that β_0 is the fastest a person will ever be able to run.

Upon completion of transforming the data, we found that the log transformation generally had a better adjusted r-squared than the linear models, but not as great as the polynomial or Box-Cox transformation.

4.1.9 Comparing all the Different Transformations

Now that we have many different models based on the different types of transformations, we need to compare their performance. We decided to use adjusted r-squared as our measurement metric to compare the models.

Figure 5: Men's event model comparison adjusted-R^2

	100M Dash	200M Dash	1500M Run	Shot Put	Long Jump
Linear	0.7323	0.8641	0.6774	0.9291	0.7883
Polynomial	0.8506	0.9025	0.8600	0.9569	0.8512
Box-Cox	0.8794	0.8678	0.7443	0.9421	0.8249
Logarithmic	0.7611	0.8671	0.6960	0.8983	0.7666

Figure 6: Women's event model comparison adjusted-R^2

	100M Dash	200M Dash	1500M Run	Shot Put	Long Jump
Linear	0.7584	0.6777	0.0807	0.3962	0.7257
Polynomial	0.7865	0.8775	0.4064	0.8414	0.8614
Box-Cox	0.7982	0.7598	0.0704	0.3553	0.7259
Logarithmic	0.7642	0.6864	0.0796	0.4076	0.7213

See Table 5 and Table 6. Based on the adjusted r-squared values, we can see that all of the events except for 100M Men and 100M Women have the polynomial model has the best model. For 100M Men and 100M Women, the best model is the Box-Cox transformation model. This tells us what models fit the data we trained on the best.

We also wanted to know how the different models performed for future events and as time went on. We chose to predict the results of the 2020 Olympics (next Olympics) to see how the models performed short term and to predict the results of the 3000 Olympics to see how the models performed long term. The results for the short term for all models seemed to be pretty realistic. The linear is very optimistic since it is always getting better while the others can be a little optimistic. For the long term, the linear and polynomial models perform terribly and are not very realistic at all. The linear models can become negative for the running events and extremely high for the throwing events which is impossible. The polynomial can go a little crazy and will also become negative or highly positive values that are impossible to achieve. The log transformation models are generally pretty stable for the long term. It can be optimistic in the the times are really fast in the future, but they are at least somewhat realistic. Last, the Box-Cox transformation seems to perform the best for long term. It will typically gradually increase for throwing and gradually decrease for running events. The times and distances it give are pretty realistic and possibly achievable. The trends of the models can also be seen in Figures 21-23.

4.1.10 Biometric Predction

Given the small size of the dataset for each model, we cleaned and explored options for adding more observations or predictors. We found a World Bank dataset of population statistics from 1960-2016, and a dataset gathered by The Guardian on biometric data such as athlete's age, height, and weight. Unfortunately, we were unable to integrate the information in an effective way. Predictions including this new data were not significantly better.

Figure 7: Standardized Adjusted-R^2 Value for Full Linear Model on Biometric Data

	100M Dash	200M Dash	1500M Run	Shot Put	Long Jump
Men	0.742	0.870	0.688	0.933	0.749
Women	0.756	0.647	0.072	0.400	0.701

Despite some improvements in the normality assumption of error terms in some of our models the based on the normal probability plots of the multiple linear regression, shown in Figure 27 of the appendix, the some of residuals actually increased from the simple linear regression models to the multiple linear regression models as shown in Figure 28 of the appendix. This included with the fact that most of the physical characteristic regression coefficients are insignificant at individual levels may indicate that a subset of the predictor variables selected for predicting the response.

5 Conclusion

Each of the five events for both genders responded to modeling in different ways, some more realistically than others.

5.1 Men's and Women's 100M

Although the raw data appears strongly linear, there is non-normality in the normal probability plot of a first order linear model, and the adjusted R^2 is maximized after in the Box-Cox transform. The Box-Cox also has a reasonable long-term estimate at around 8 seconds for the Men's 100M Dash. This is potentially unrealistic as this would require an average sustained speed of approximately 28 mph, a speed usually only obtained when elite sprinters hit top speed. See Tables 5, and 6, Figures 8, 10, 21.

5.2 Men's and Women's 200M

Again, the raw data seems linear, but a first order linear model deviates from the normal probability plot. The polynomial fit has a much larger adjusted R^2. However, the polynomial fit has an impossible long term-prediction, and therefore should not be used past the training data. The Box-Cox once again has the best long-term prediction, but again would require unsustainable speeds by human beings in order to reach the predicted times on the Men's side. The Women's prediction of just over 20 seconds seems plausible, as that is only eclipsed currently by only some of the world's best male sprinters. See Tables 5, and 6, Figures 8, 10, 21.

5.3 Men's 1500M

Although the raw data seems linear, the normal probability plot of the simple linear regression was demonstrated non-normality. The polynomial regression fit the data with the highest Adjusted R^2, but had an impossible long-term estimation. The Box-Cox has a stable long-term prediction but the apparent predicted times hundreds of years into the future have already been eclipsed. This points to a lack data points to fit for our regression model. See Table 5, Figures 8, 10, 22.

5.4 Women's 1500M

The data for this event was particularly sub-optimal, in that there was few data and what was there had less of a pattern than the rest of the events. This could indicate that women have achieved peak performance in the 1500M Run. However, a more plausible explanation of the erratic behavior of the data could the strategic nature of the 1500M Run where the pace of race is generally determined by the "rabbit" who sets the pace for the first two laps and then the leaders of the race. This would lead to more erratic response variables which require more data to determine the underlying trend. With the considerable lack of data we analyzed, the results may not actually capture a trend of improving results in the 1500M Women's. See Table 6, Figures 8, 10, 21.

5.5 Men's Shot Put

The Men's Shot Put had very high adjusted R^2 for each of the different models. All except for logarithmic were above .9. This is most likely because the logarithmic assumption does not work well for the throwing events. The assumption we made for this event will give a negative β_1 so that the model is of the form $Y = e^{\beta_1 X} + \beta_0$. This is not the form of the data and we should have instead used the form $Y = log(\beta_1 X) + \beta_0$. We did not do this since we were using applying models to each of the datasets iteratively, and we should have split the data up into running events and throwing events and then applied the models iteratively over the split event types. If we were to do this again, we would definitely split the events and try this transformation to see how the logarithmic transformation performed.

It can also be seen that the polynomial model had the best adjusted R^2, but looking at the plot of the performance over time (Figure 21-23), the polynomial model is not realistic for the distant future. For this event, the box cox transformation model is the most realistic with a shot put distance of 62.23 meters in the year 3000. This is highly unlikely since it is 3 times larger than the current word record, however it is definitely more probable than the other model predictions.

5.6 Women's Shot Put

Women's Shot Put had considerably lower adjusted-R^2 values for the linear, Box-Cox transformed, and Logarithmic transformed models compared to our polynomial model. Based on the simple linear regression residuals and the normal probability plots seen in figures 8 and 10, it is pretty clear that we have non-constant variances that we have to deal with. The Box-Cox plot in Figure 20 suggests that a power family transformation with $\lambda = 4$ or 5 would be the best transformation. Thus, it is sensible that the polynomial regression with third order terms performs the best out of all the potential models based on Table 6.

5.7 Men's and Women's Long Jump

Both Men's and Women's Long Jump have decent predictions based on the adjusted-R^2, but the residuals and the normal probability plots of the simple linear models indicate some pattern to the residuals shown in Figures 9 and 10. This can explain why the polynomial data fits most closely. Again, the only reasonable long term estimate was the Box-Cox model which only predicts slight improve over the next few hundred years 22.

5.8 Overall

This analysis highlights that while the polynomial has a higher adjusted R^2 for most of the models, it is very unpredictable and unreliable for the distant future. The most reliable and realistic models were the Box-Cox transformation models. The Box-Cox transformation models had one of the higher adjusted R^2 and when plotting the trend from 2020 to 3000, we could see that it predicted more realistic results than the other models. Therefore, the Box-Cox models are the best models to use.

We also can answer the question that we are getting better in each event as time goes on, excluding the 1500M Women's. This can be seen in the linear models. Since we are getting better there has to be some limit. While we did not find the limits for each event explicitly, we were able to predict the results in the distant future, and as years go on, the change in these results change less and seem to somewhat converge. We also do not believe that the convergence would be anything realistic, in that the convergence value will not be actually achievable and there should be a larger convergence value.

If we were to do more with this project, there are many different things that we would try differently. First, we would want to continue to try to find the limit explicitly for each event. We would do this by looking at more data sources and collecting more data to train better models. We would also be more in depth with our cleaning of the data to try to find errors or missing data. We would also try different models and transformations to see how they would perform. For example, we would be sure to transform the throwing events correctly. Last, we would be sure to split our data into a testing data set and a training data set. We would most likely split the data by a certain year so that we are predicting the 'future', which we already know. This would help test how reliable and accurate our models are to actually predict data in the near future rather than just seeing if it is reasonable. The reasons we did not do this for this project, is that we did not have much data, so splitting the data would decrease how much we could train on and we did not care as much about the near future as much as we cared about the distant future.

Overall, the only conclusion we can make with certainty is that we need more data. Although Olympic performance continues to improve, there are multiple events that are predicted to plateau at a level not far above the current best. Tokyo 2020 will provide the next opportunity to improve our data.

References

[1] The Guardian. 120 years of olympic history: athletes and results. kaggle datasets download -d heesoo37/120-years-of-olympic-history-athletes-and-results. Accessed: 2016-11-01.

[2] J. Neter, M. H. Kutner, C. J. Nachtsheim, and W. Wasserman. *Applied Linear Statistical Models*. Irwin, 1996.

[3] Jay Ravaliya. Olympic track and field results.

6 Appendix

6.1 Plots

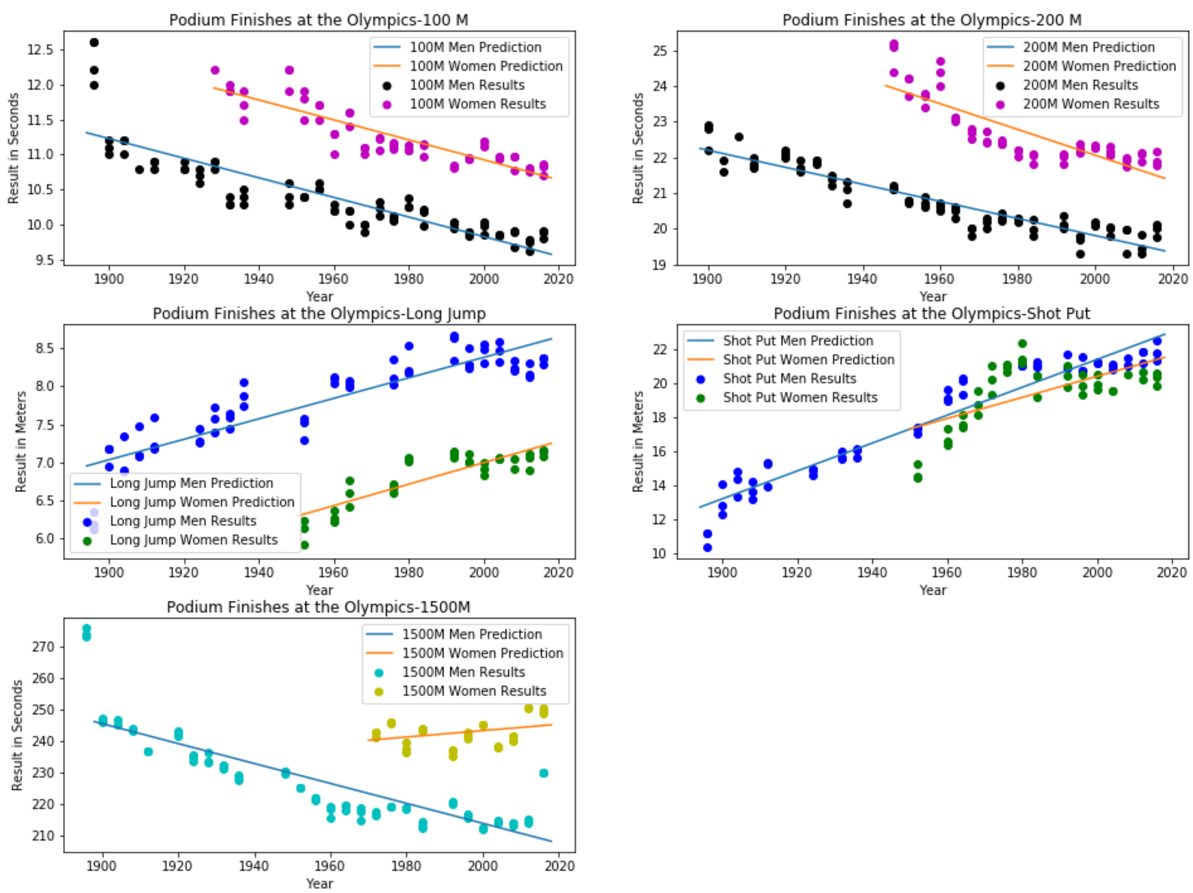


Figure 8: Plots of simple linear regression models for time vs. results for the 100 Meter Run, 200 Meter Run, 1500 Meter Run, Shot Put, and Long Jump.

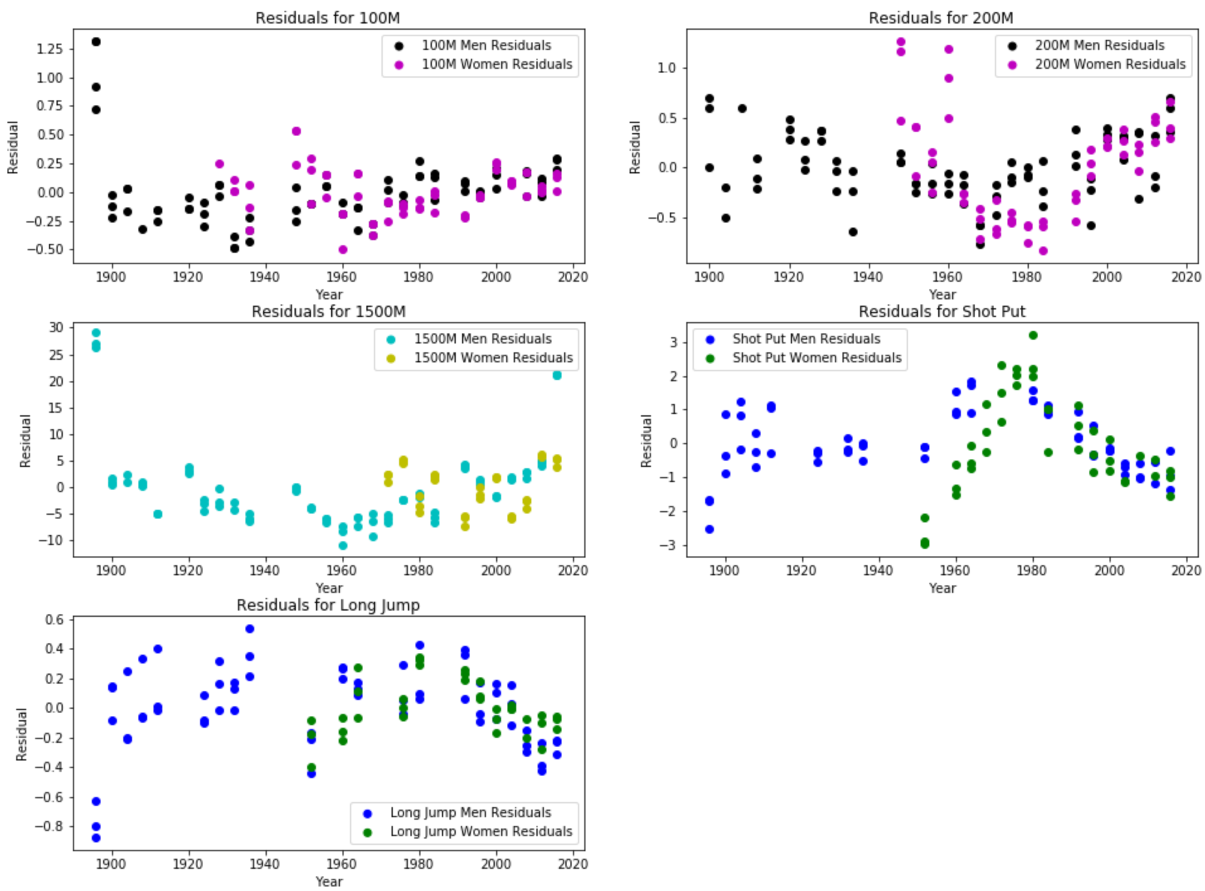


Figure 9: Plots of the Residuals for the simple linear regression models for time vs. results for the 100 Meter Run, 200 Meter Run, 1500 Meter Run, Shot Put, and Long Jump.

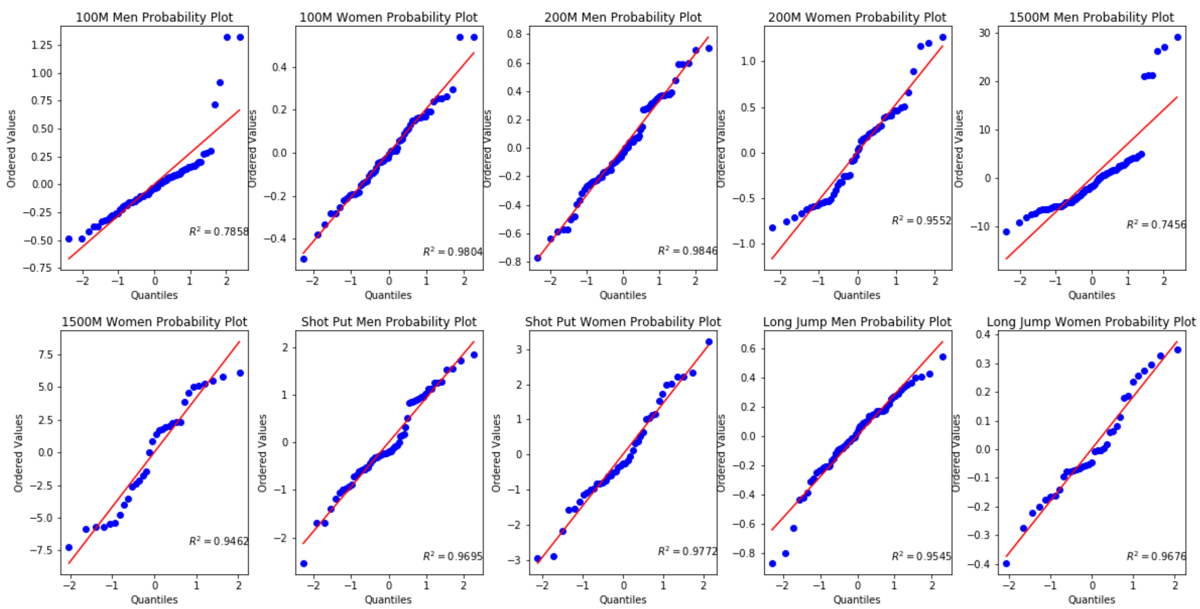


Figure 10: Normal Probability plots for each of our simple linear models.

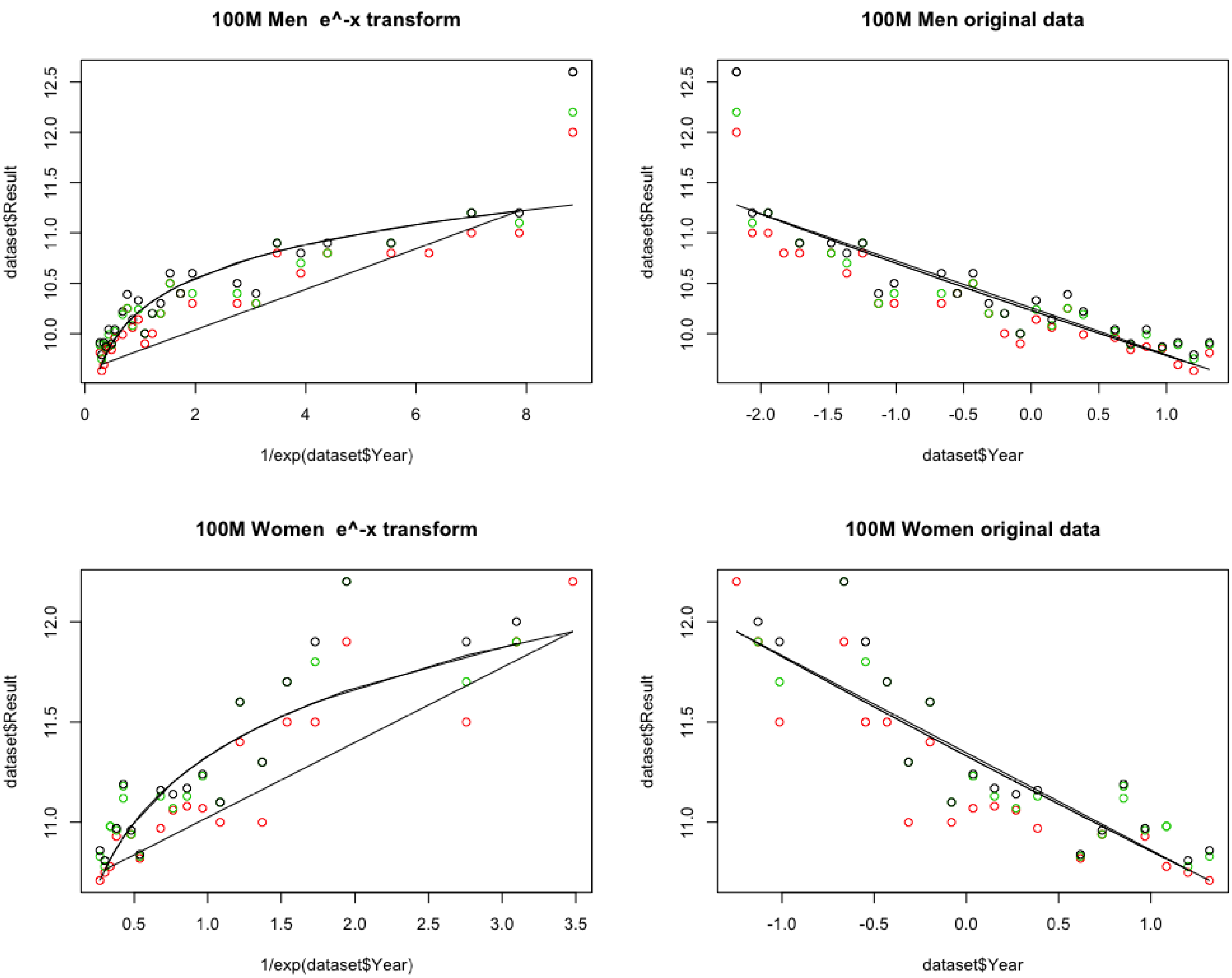


Figure 11: Logarithmic transformed models for 100M Dash

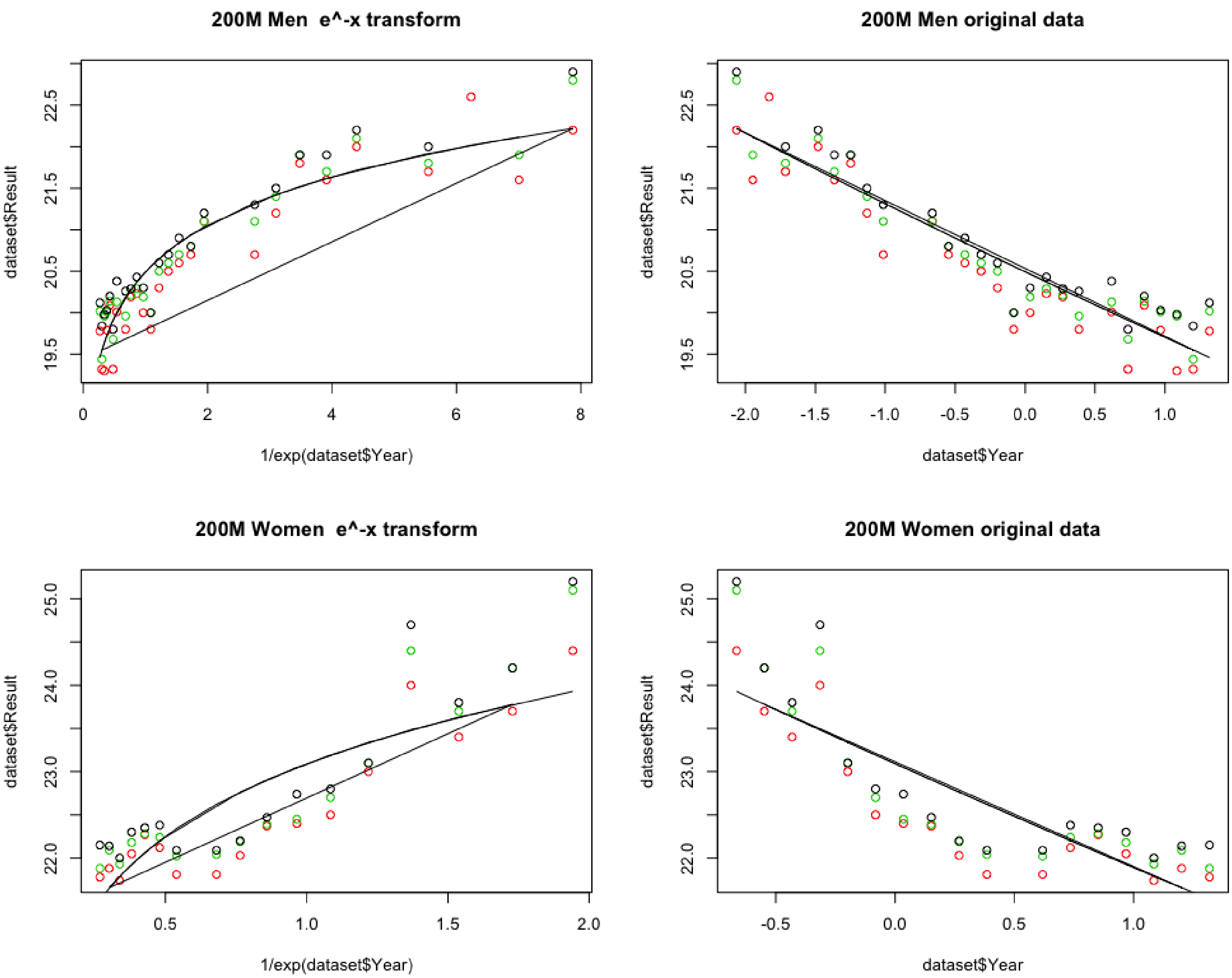


Figure 12: Logarithmic transformed models for 200M Dash

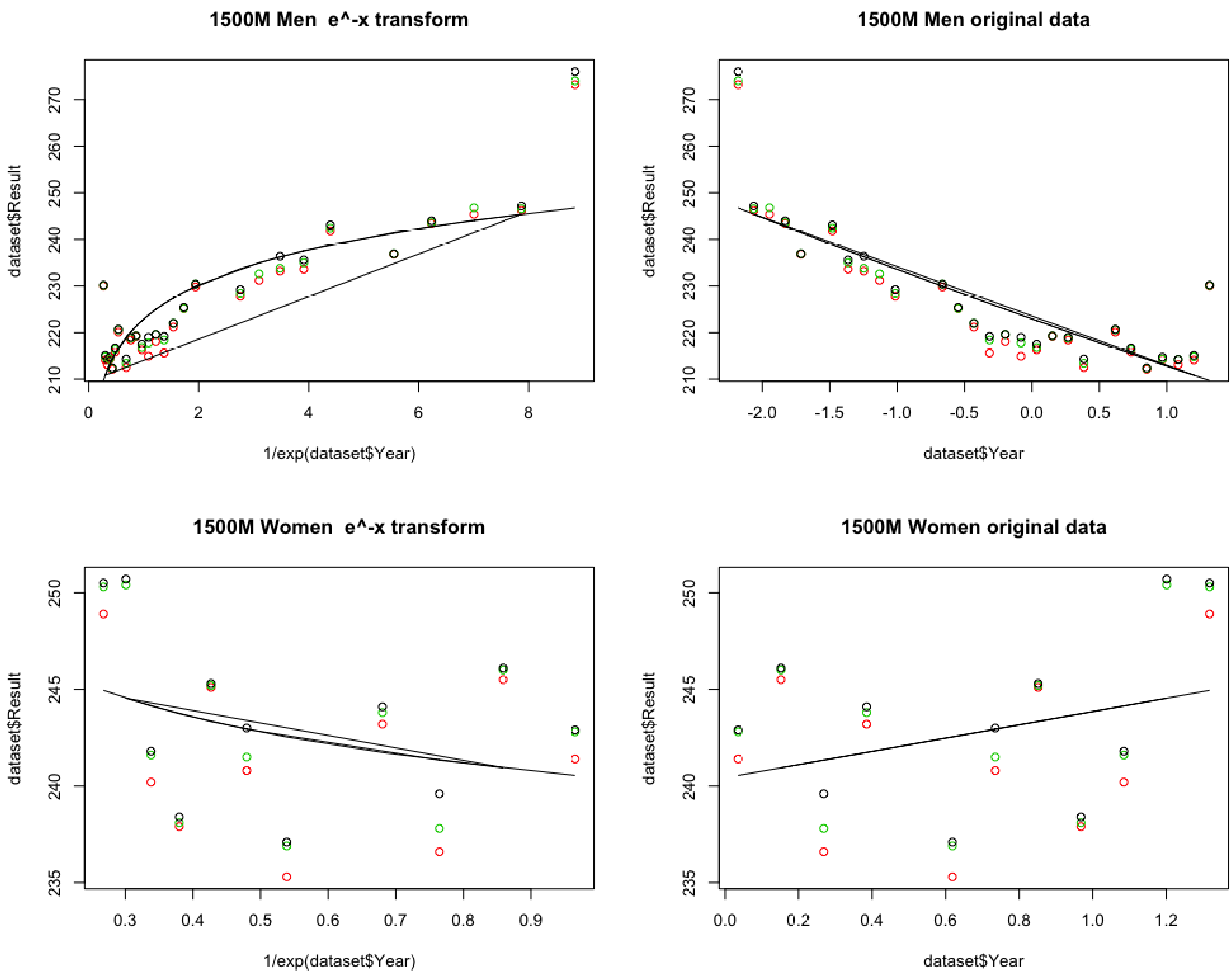


Figure 13: Logarithmic transformed models for 1500M Run

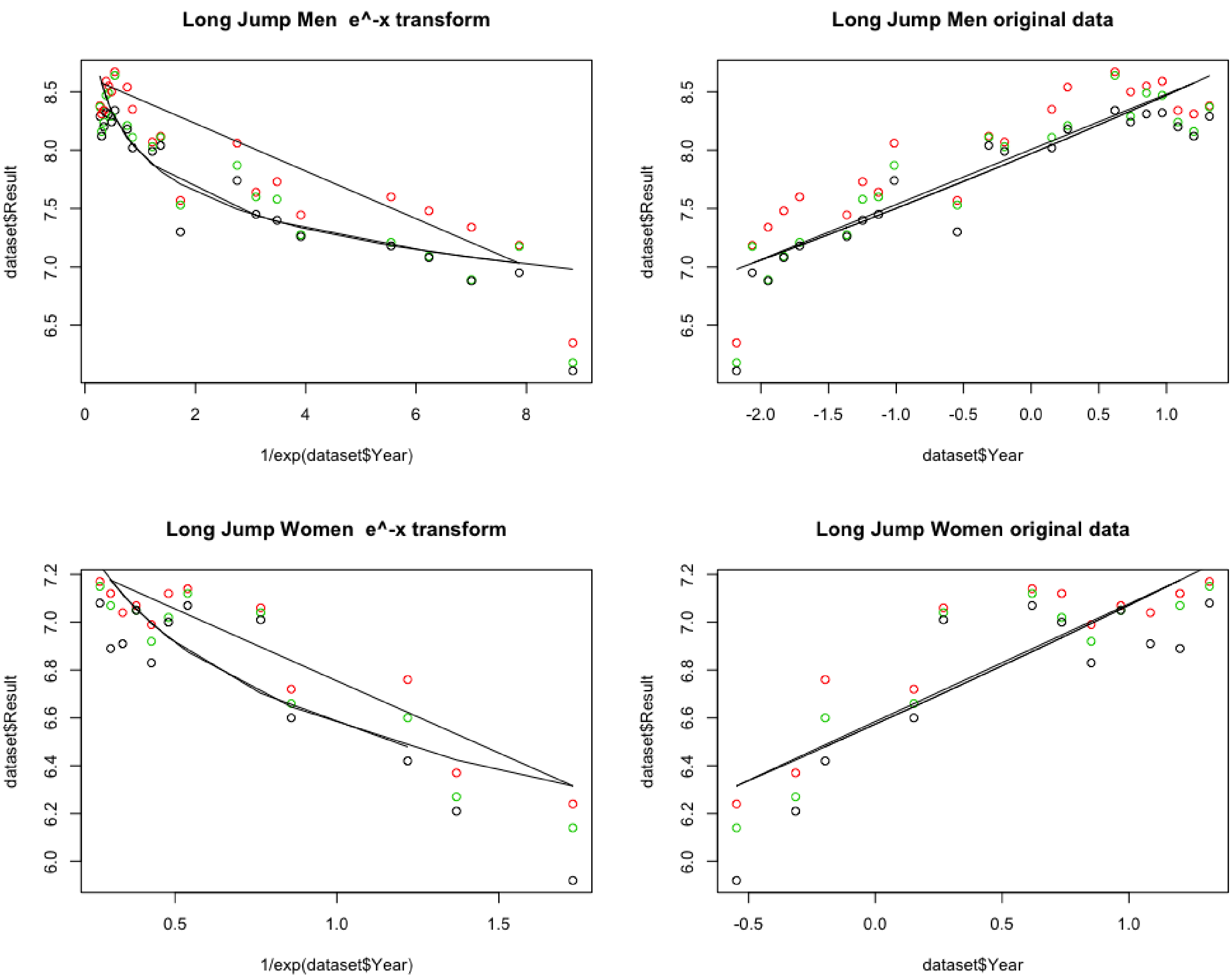


Figure 14: Logarithmic transformed models for Long Jump

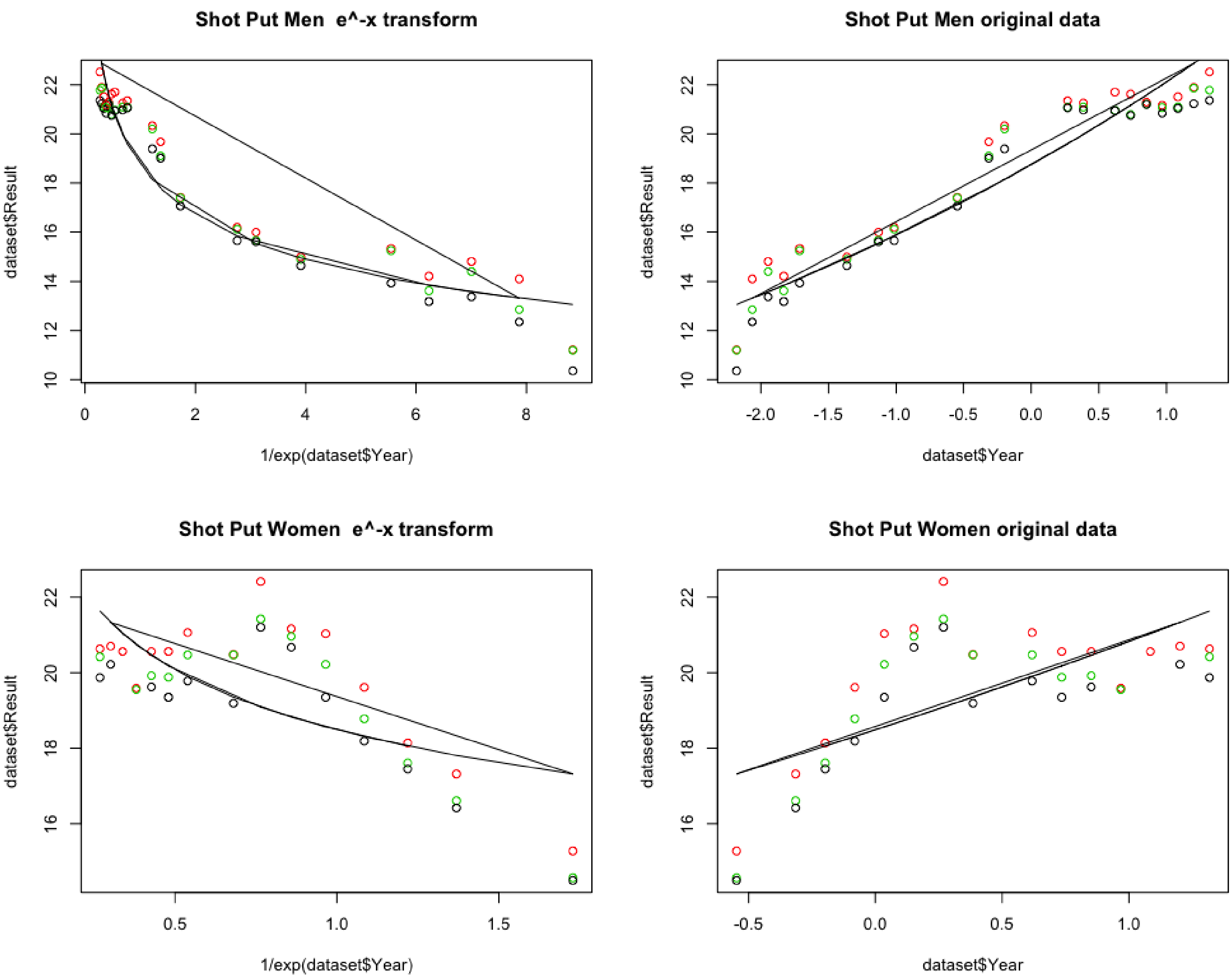


Figure 15: Logarithmic transformed models for Shot Put

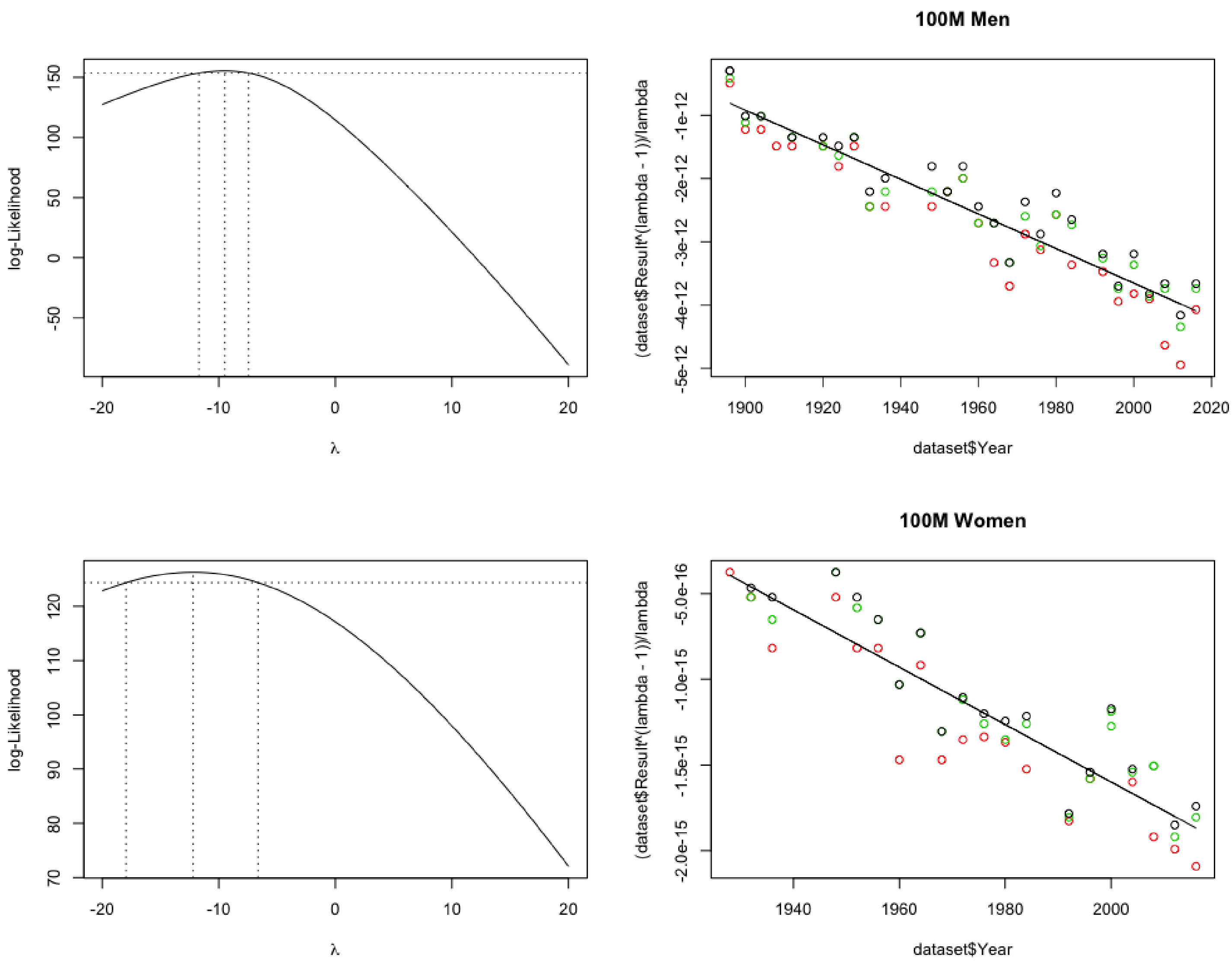


Figure 16: Box-Cox Transformed Models for 100M Dash

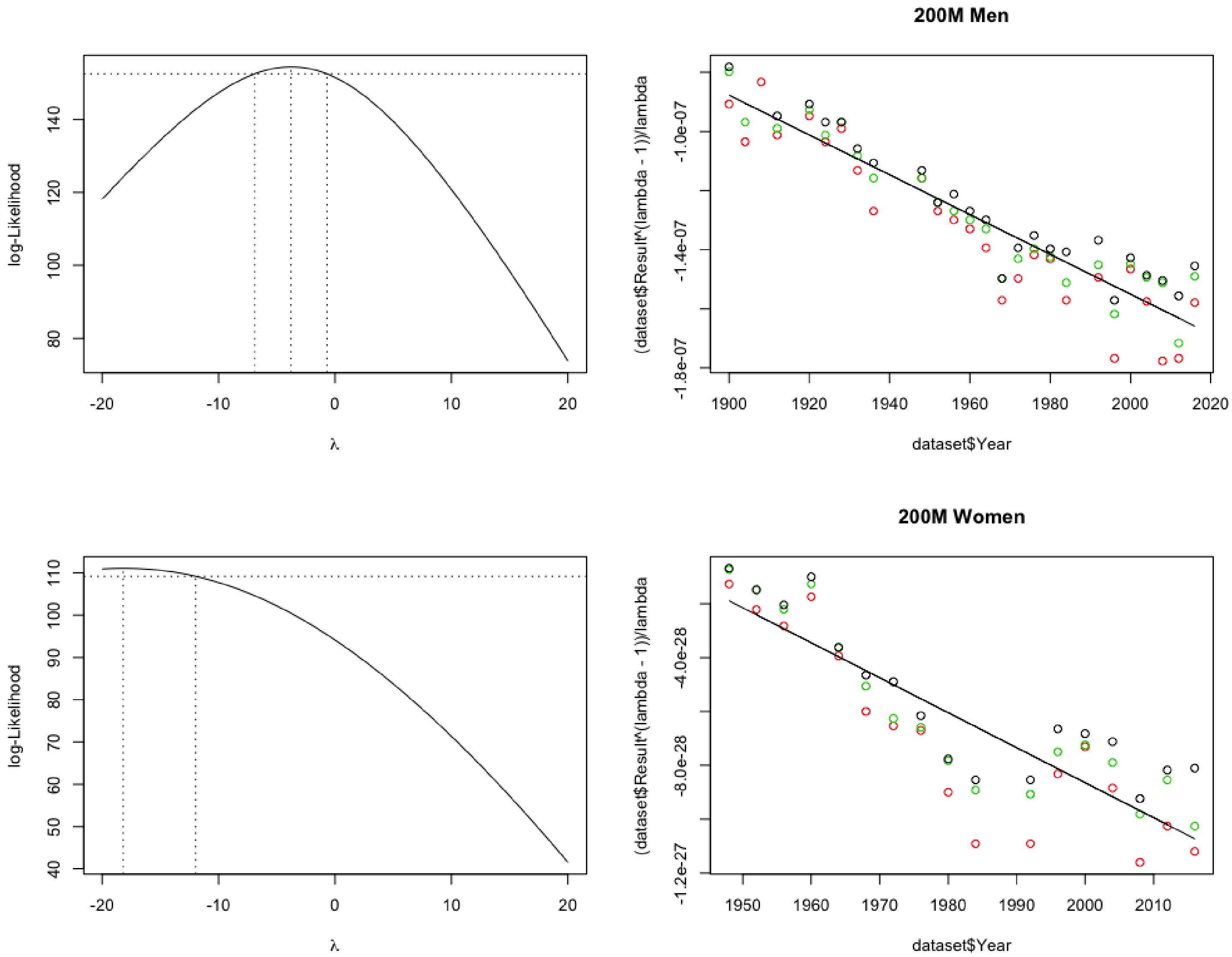


Figure 17: Box-Cox Transformed Models for 200M Dash

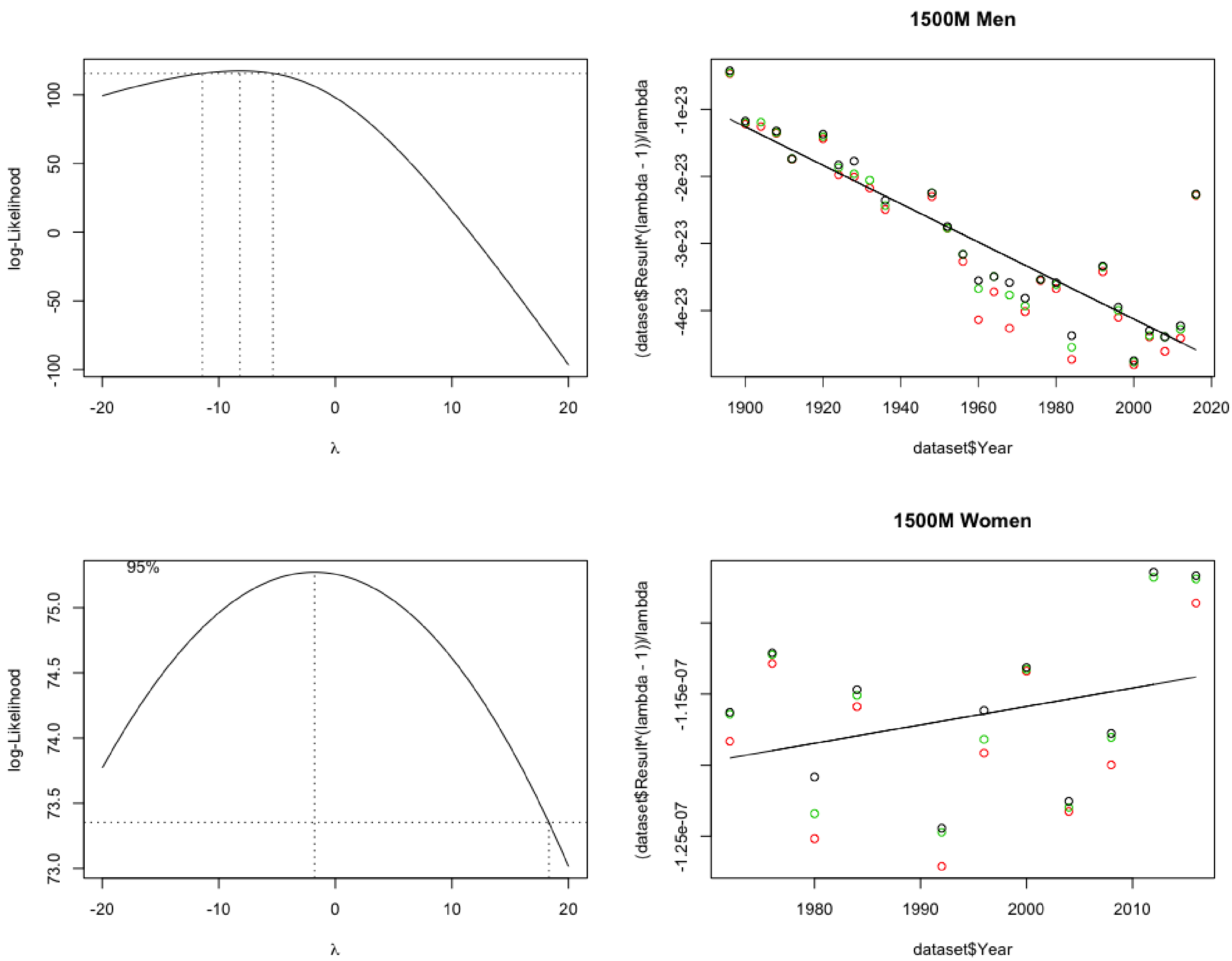


Figure 18: Box-Cox Transformed Models for 1500M Run

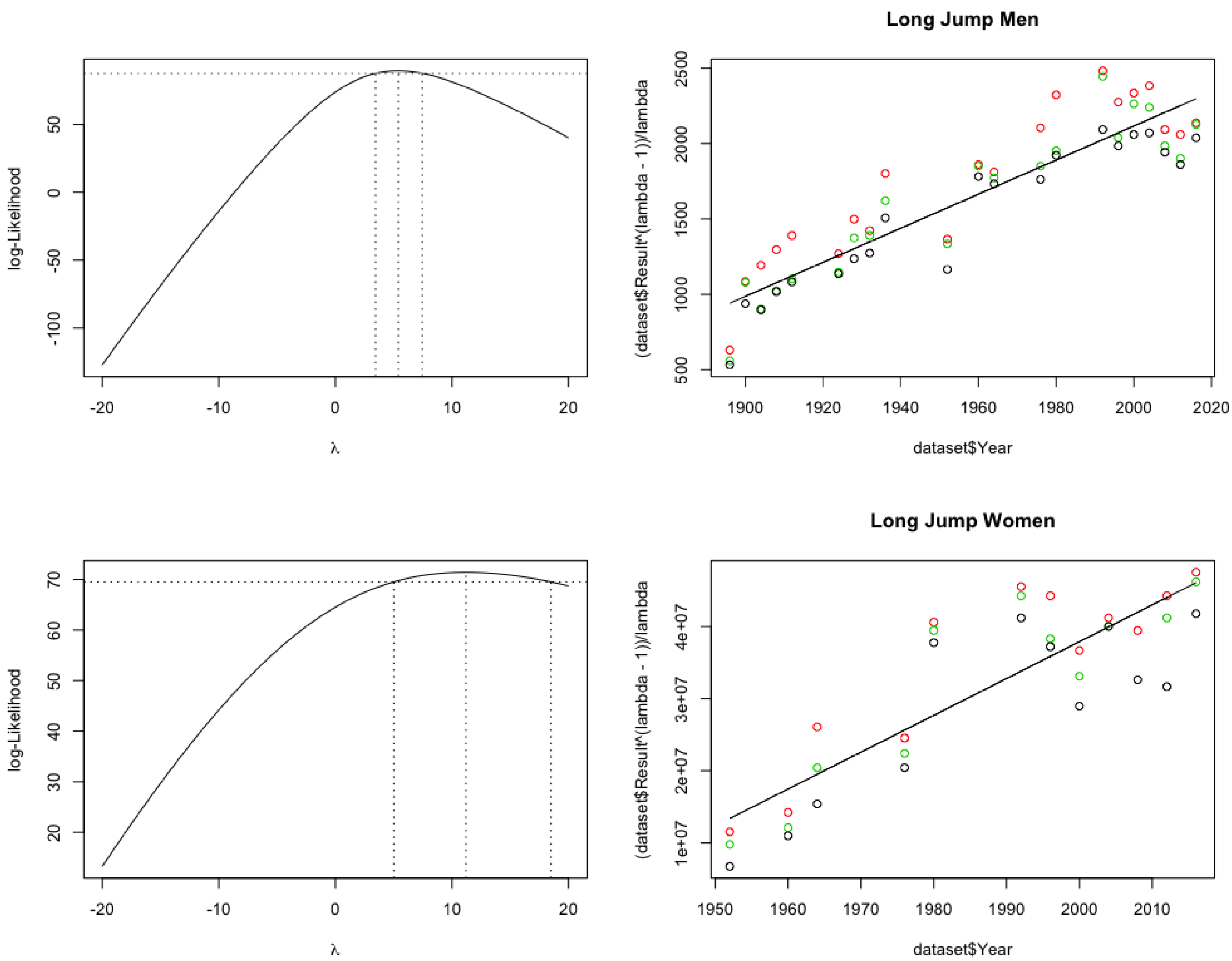


Figure 19: Box-Cox Transformed Models for Long Jump

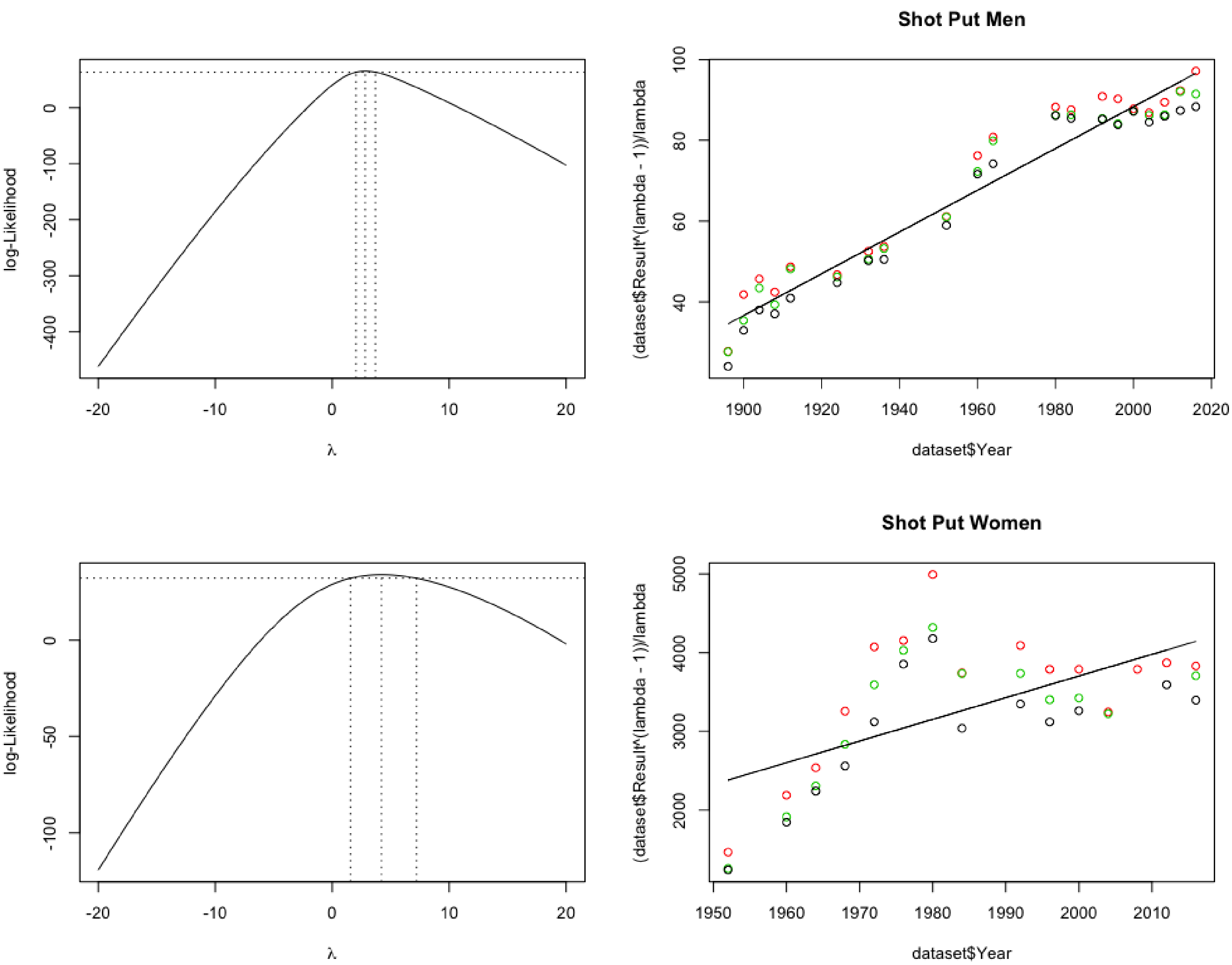


Figure 20: Box-Cox Transformed Models for Shot Put

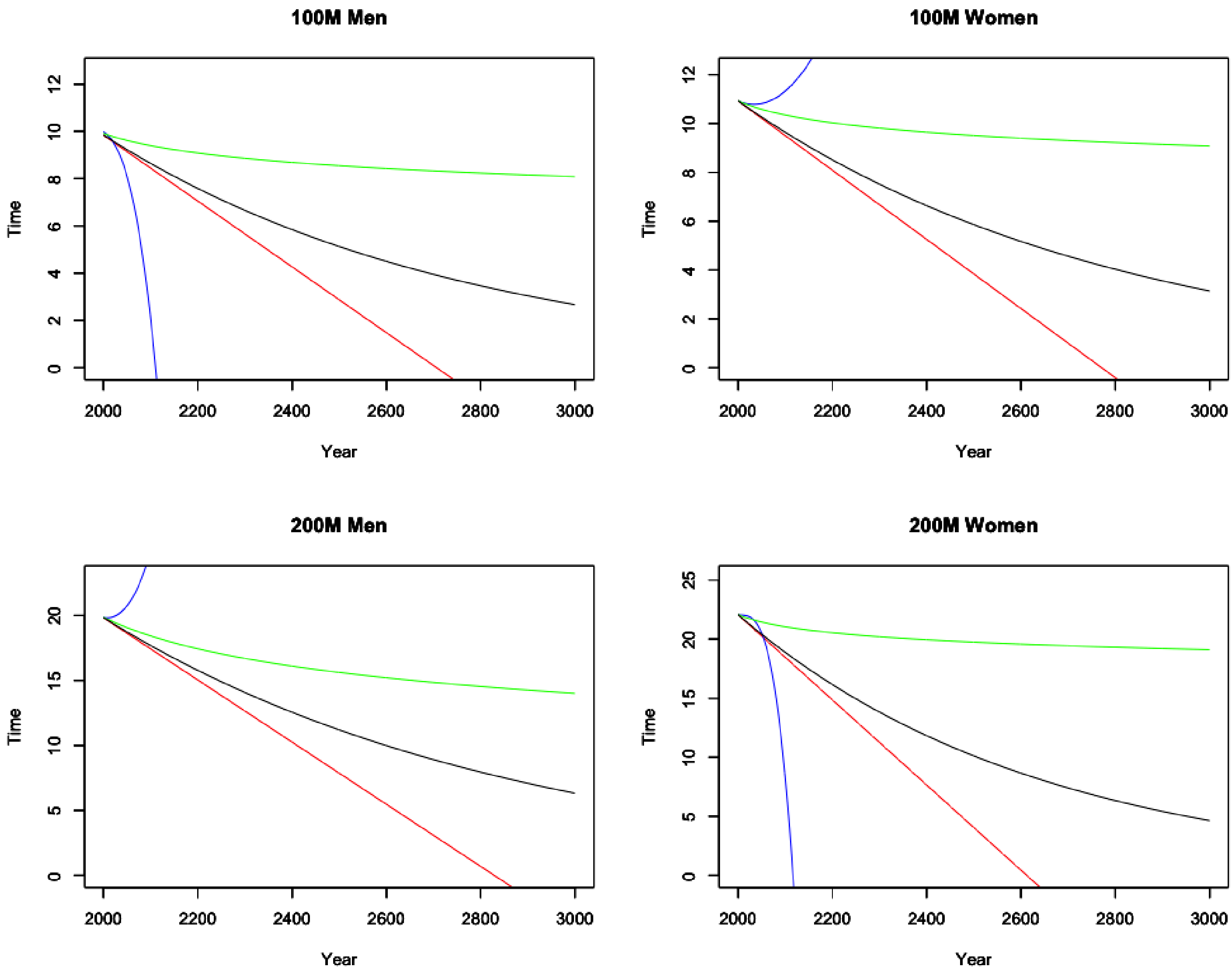


Figure 21: Predictions of Event Results for 100M Dash and 200M Dash. The blue line represents the polynomial model, green represents the Box-Cox transformed models, red is the linear model, and black is the logarithmic transformed model

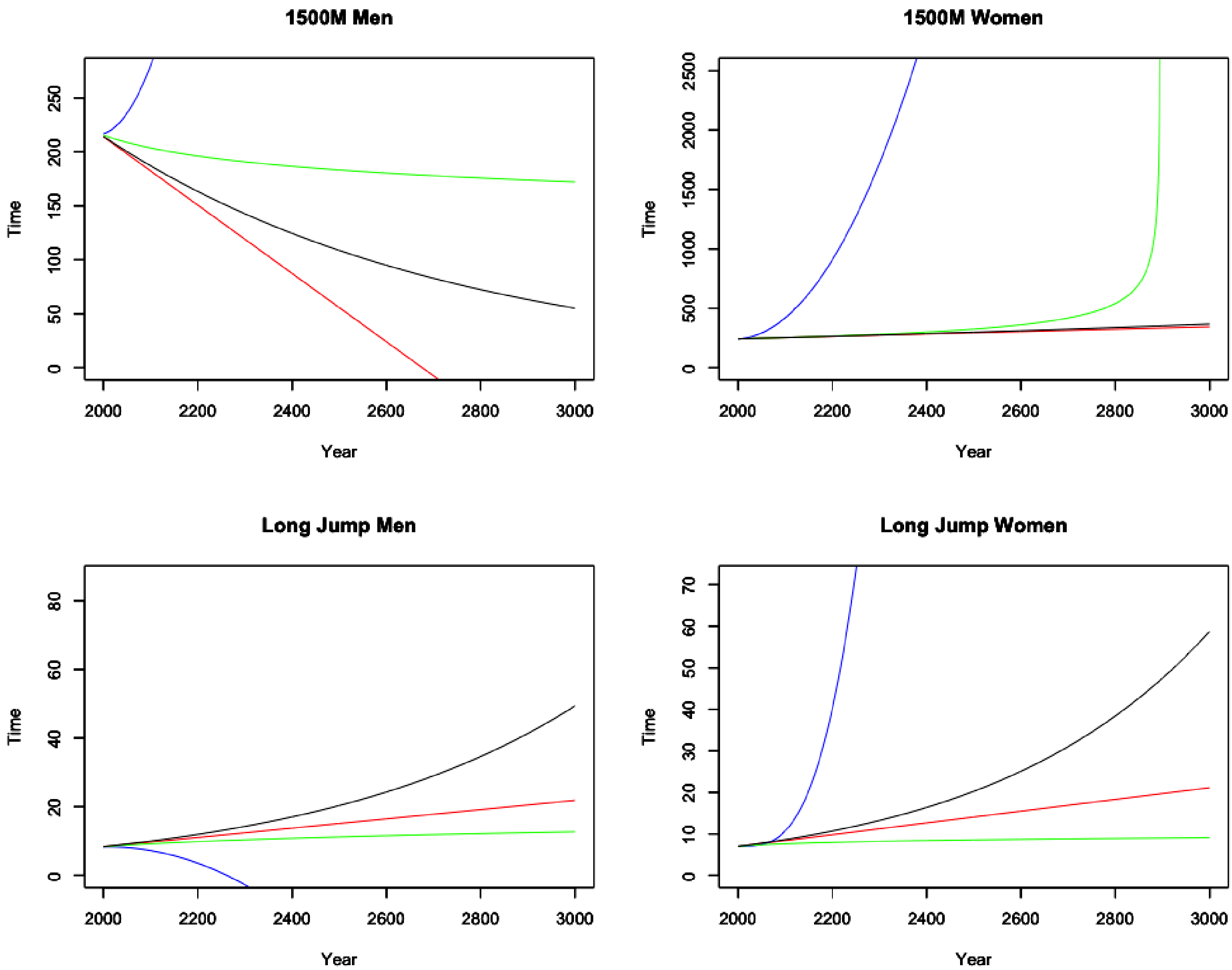


Figure 22: Predictions of Event Results for 1500M Run and the Long Jump. The blue line represents the polynomial model, green represents the Box-Cox transformed models, red is the linear model, and black is the logarithmic transformed model

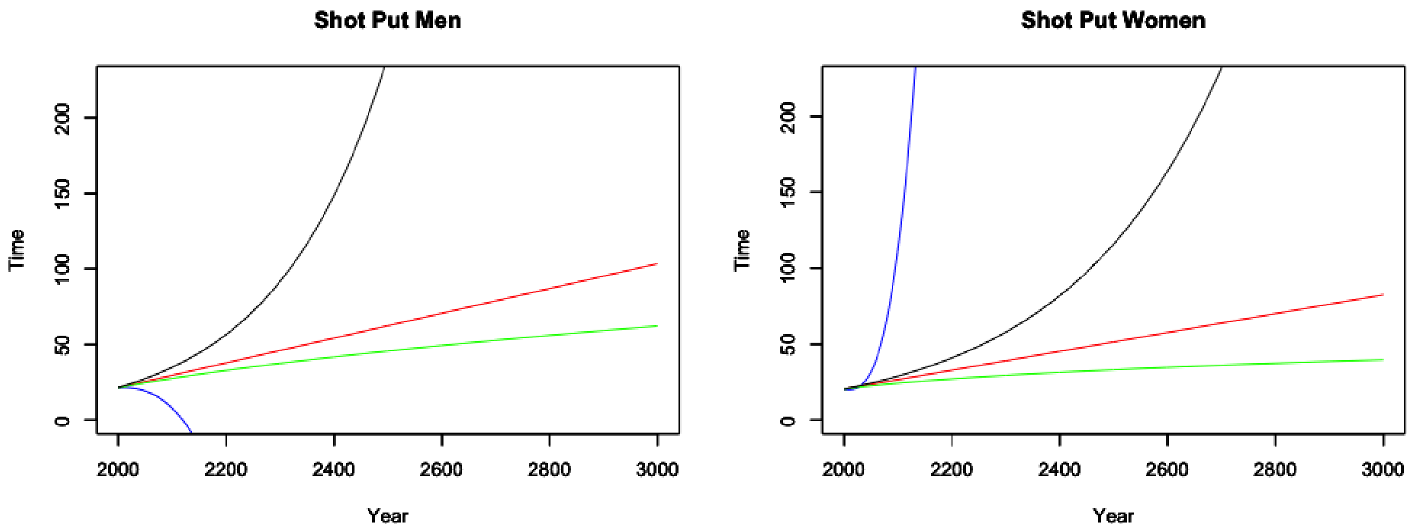


Figure 23: Predictions of Event Results for the Shot Put. The blue line represents the polynomial model, green represents the Box-Cox transformed models, red is the linear model, and black is the logarithmic transformed model

Figure 24: Correlation Plots for the numerical data in the 100M Dash and 200M Dash.

Figure 25: Correlation Plots for the numerical data in the 1500M Run and Long Jump.

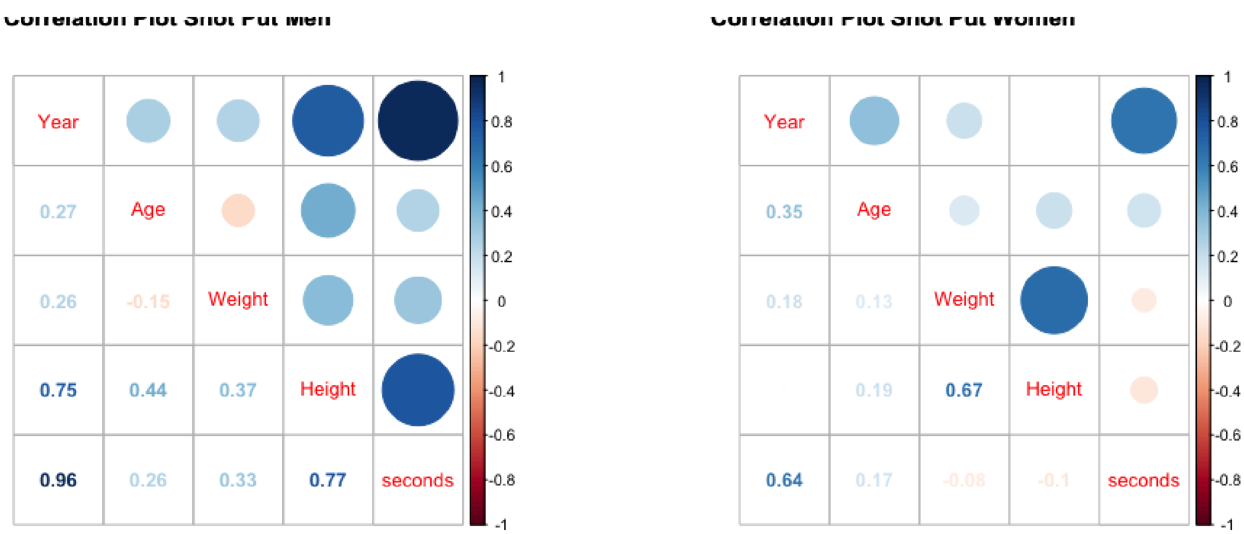


Figure 26: Correlation Plots for the numerical data in the Shot Put.

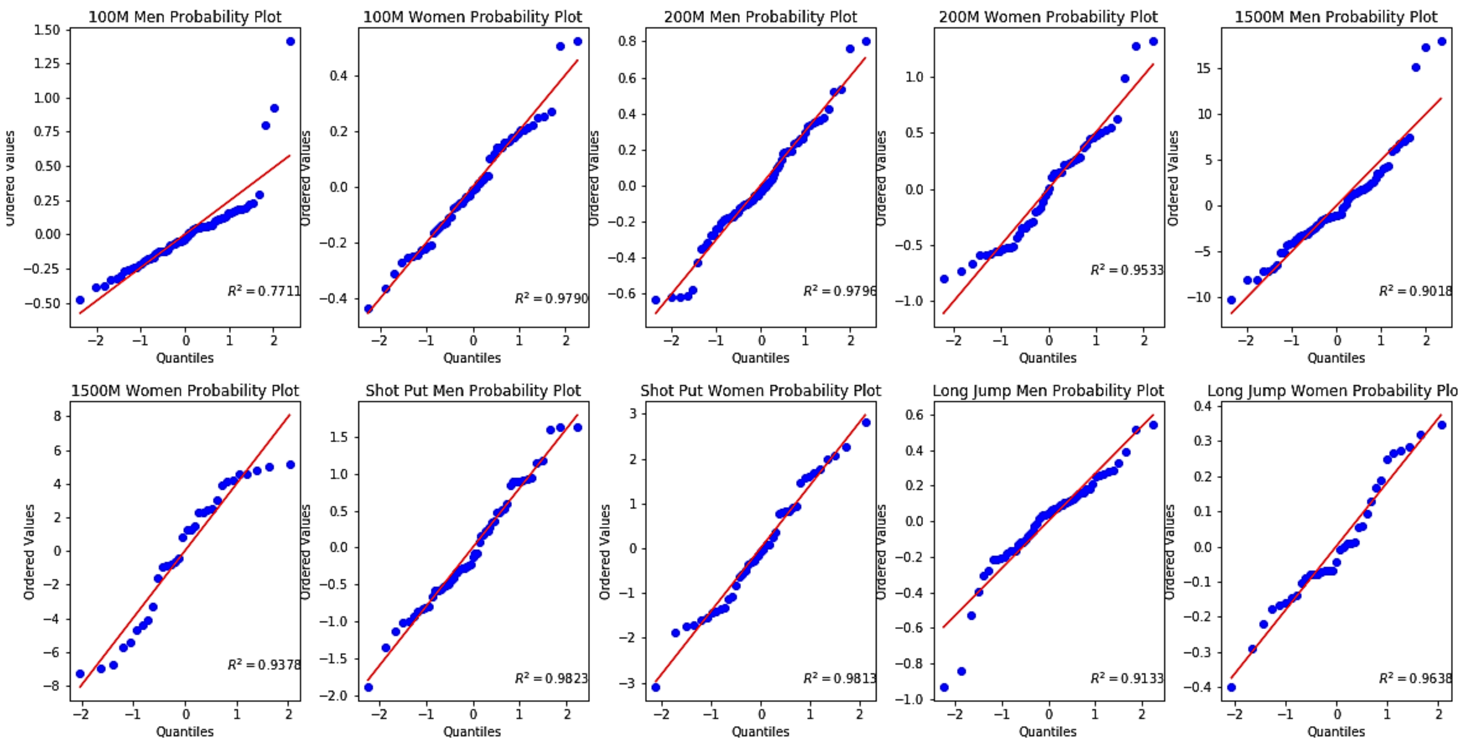


Figure 27: Normal Probability plots of the first order Biometric Data Model.

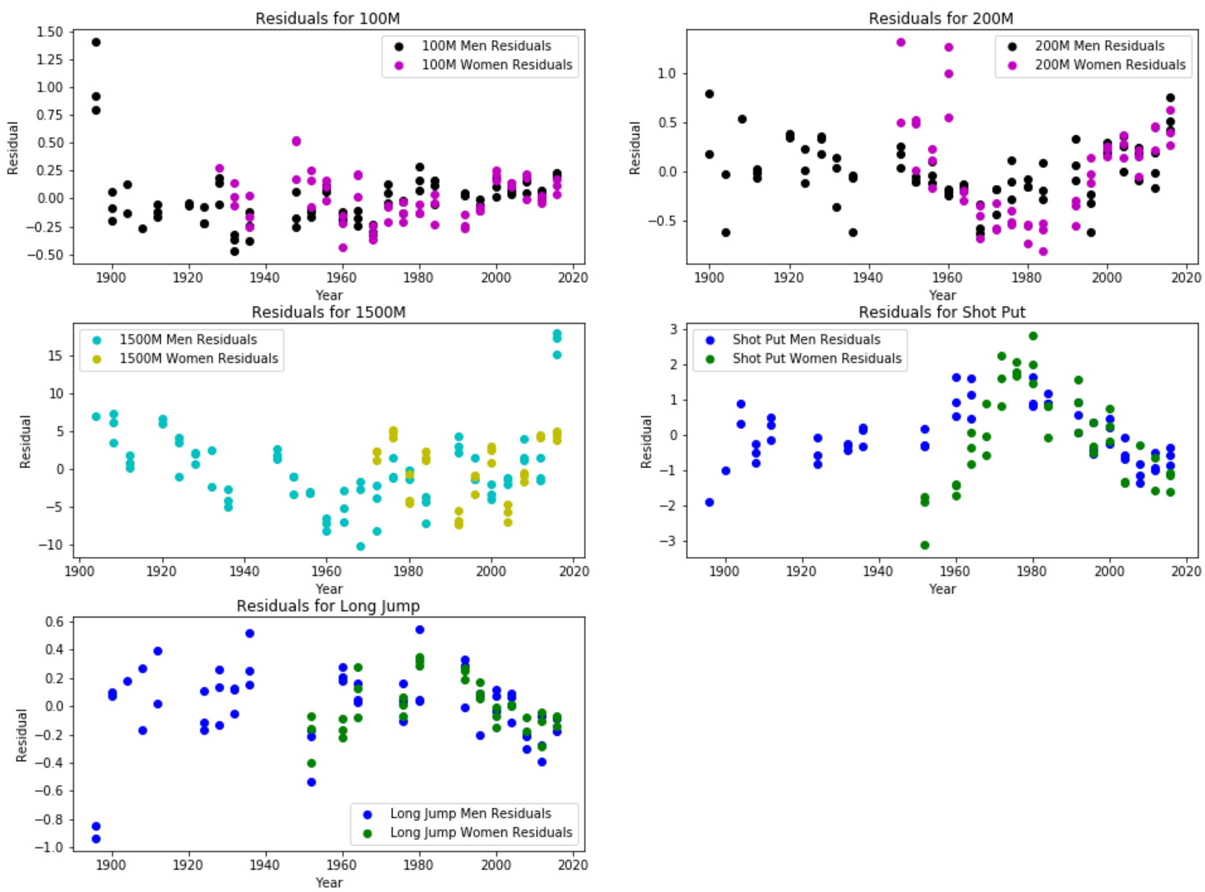


Figure 28: Plots of the first order Biometric Data Model residuals against Year.

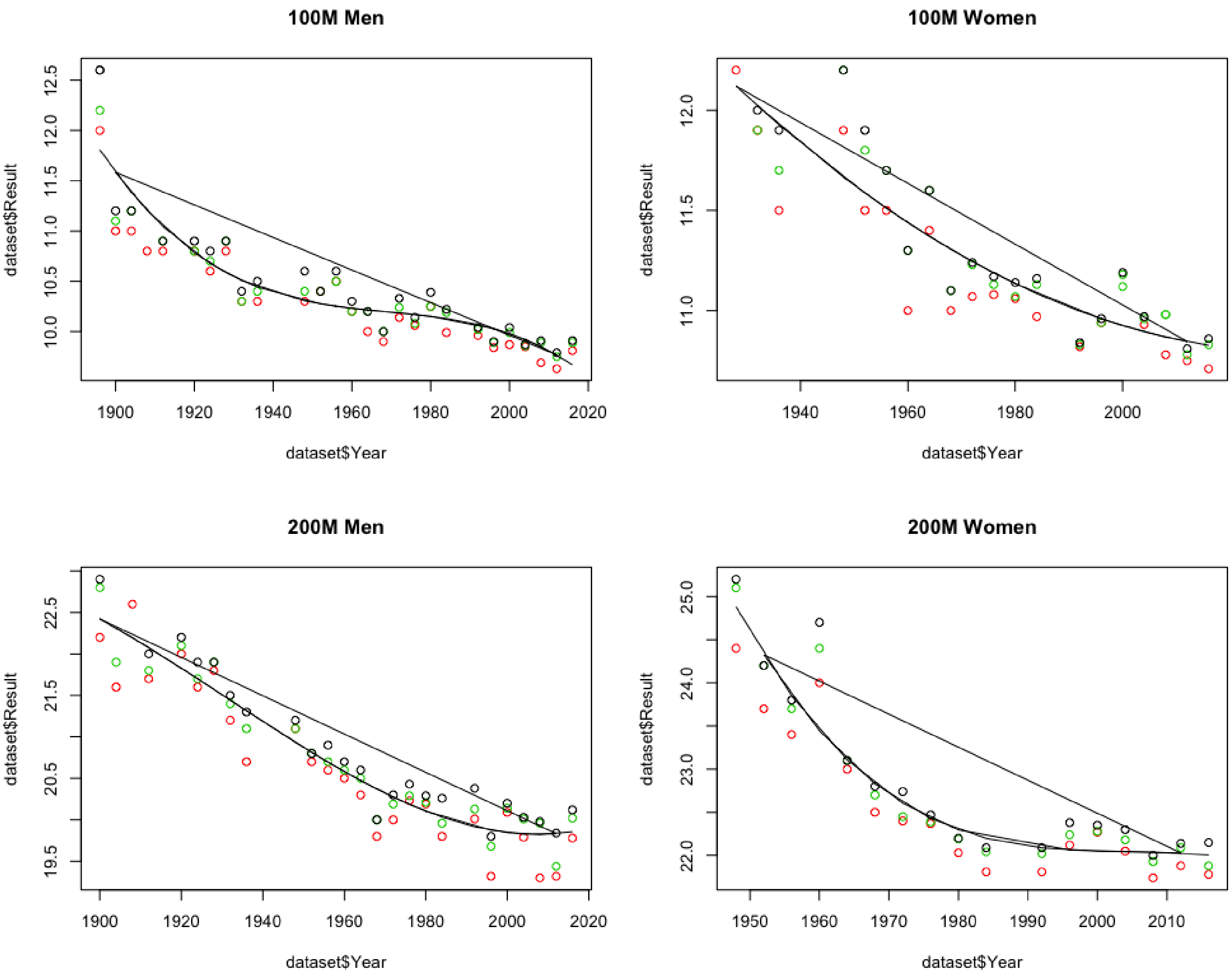


Figure 29: Polynomial Regression Plots with regression lines for the 100M Dash and 200M Dash.

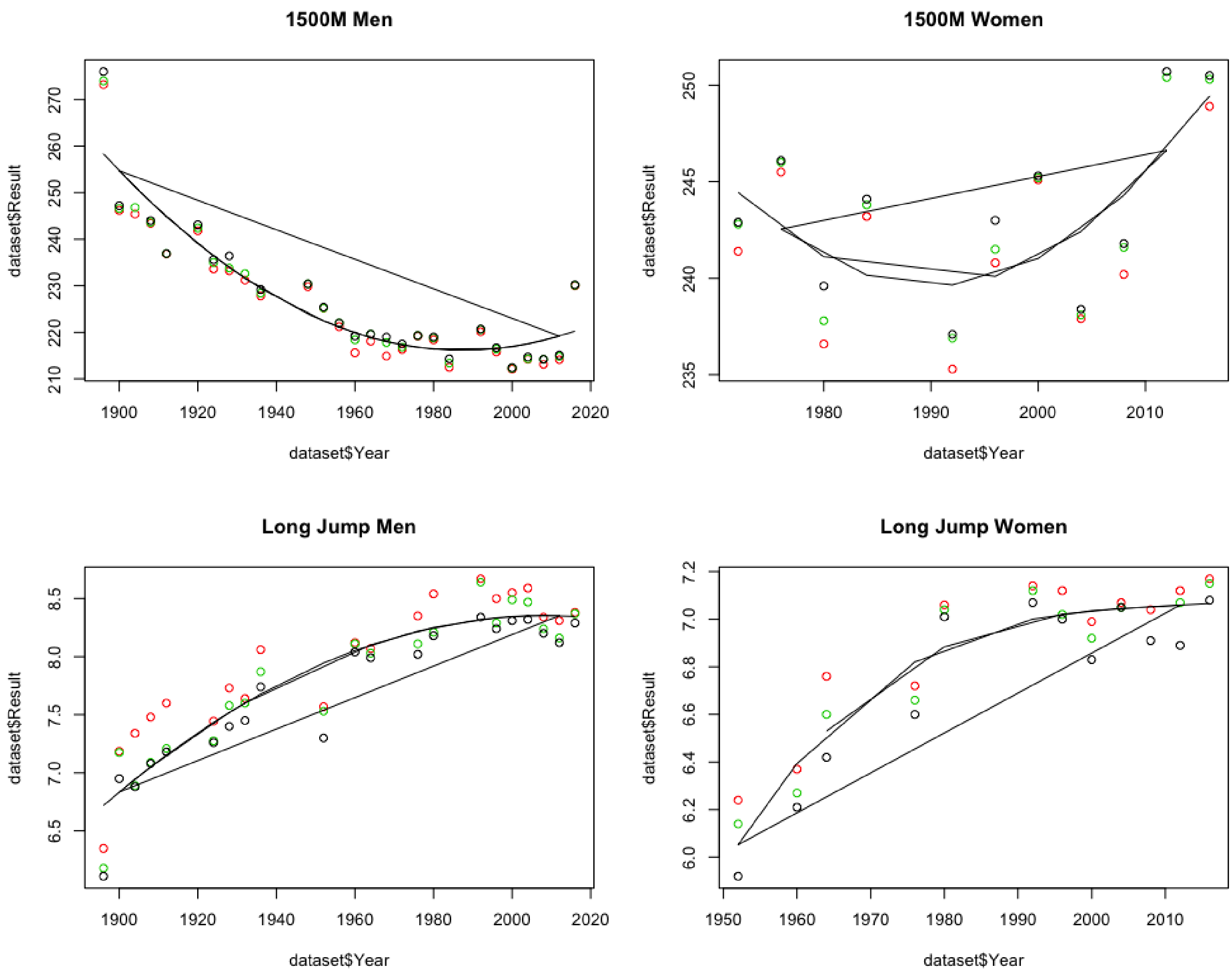
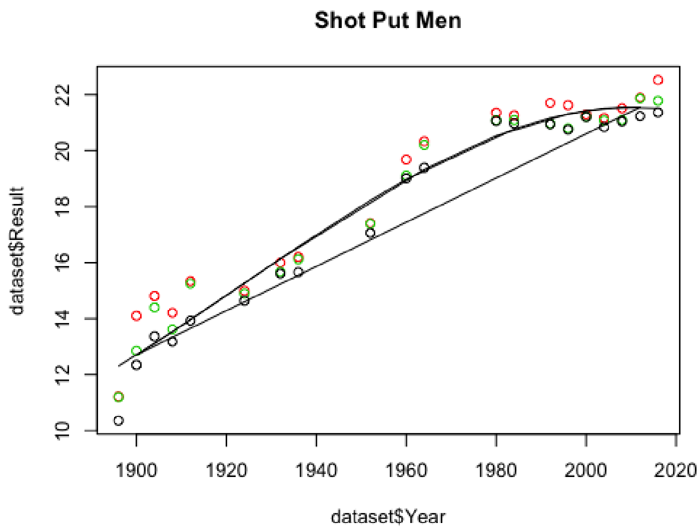


Figure 30: Polynomial Regression Plots with regression lines for the 1500M Run and Long Jump.



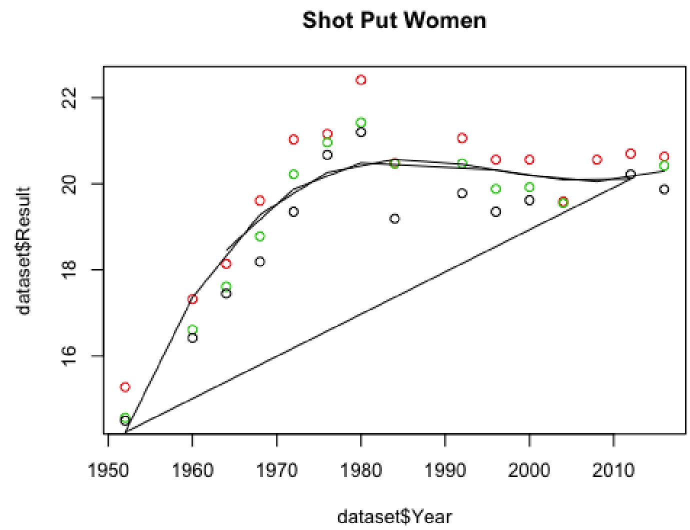


Figure 31: Polynomial Regression Plots with regression lines for the Shot Put.