Table of Contents

Introduction

In the world of swimming, there is a commonly accepted notion of "the perfect swimmer body." Coaches, spectators, and swimmers alike use the phrase when referring to the top notch swimmers by saying, "Of course Michael Phelps is fast, he has the perfect swimmer's body." Typically what is meant by this saying is that the athlete is tall, slender around with waist with broad shoulders, and has disproportionately long arms. Well if long arms help a swimmer pull faster, it could be thought that long feet make the swimmer kick faster. To see if the "perfect swimmer body" actually indicates a swimmer's speed, I set up a project to test this theory.

I have been a swim coach for 5 years now, so access to swimmer measurements was not hard to come by. I coach swimmers ages 4-18 (typically referred to as age group swimmers) and ranging in all abilities. I decided to set up a project where I measured height, wingspan, shoulder circumference (around the shoulders at armpit level), waist circumference, foot length, how many years each swimmer has participated in competitive swimming, and their 50 yard freestyle time. Every variable was measured in centimeters, besides years of swimming which was measured in years. If this was a swimmer's first year, it was considered 1 year of swimming. Collecting the data was not done in the best manner, since I had to choose swimmers whose parents I knew would be comfortable with me taking these measurements. Therefore, I do not have a random sample of swimmers, but the ages and range of ability still vary greatly. I ended up getting measurements from nineteen females ranging in ages 7-17, and nineteen males ranging in ages 6-15. I knew if I analyzed all of the ages for each gender together, that a trend would occur because as much as a 17 year old female is taller than a 7 year old female, she would also be faster, naturally, because of strength and coordination that is developed through growing up. The same can be said for the males. Therefore, I split up both genders into two age groups: 10 and under, and 11 and up. I was able to collect the following data:

10 and under males:

Swimmer	Age	Height	Wingspan	Shoulder Circumference	Waist	Foot length	Years of swimming	50 yard freestyle time
1	6	111.4	113.7	68.0	44.9	18.1	1	68.00
2	7	119.7	119.7	73.3	54.0	18.3	4	39.55
3	7	124.1	121.4	74.8	55.5	18.5	3	51.36
4	8	122.6	120.0	73.7	52.8	19.2	2	54.50
5	9	134.3	132.6	82.4	56.9	21.4	4	42.93
6	9	135.9	133.6	84.6	60.8	22.5	2	50.22
7	9	134.3	134.5	79.8	54.5	20.0	5	29.91
8	10	144.5	142.4	90.1	57.5	21.7	4	31.88
9	10	143.5	146.8	95.9	71.4	22.3	5	46.80

11 and up males:

Swimmer	Age	Height	Wingspan	Shoulder Circumference	Waist	Foot length	Years of swimming	50 yard freestyle time
10	11	155.6	152.9	88.3	62.2	25.0	3	36.73
11	12	175.3	177.1	108.0	73.4	26.5	1	29.55
12	12	157.8	161.9	91.3	64.4	25.1	9	29.21
13	12	156.2	152.1	87.0	66.5	24.2	9	31.84
14	12	152.1	153.0	88.0	59.4	22.6	7	30.64
15	12	144.5	144.1	83.9	59.5	21.5	10	35.78
16	13	158.8	161.3	92.0	62.0	24.4	8	31.29
17	15	181.9	178.6	114.3	70.5	26.8	2	27.77
18	15	177.5	179.1	117.0	77.0	26.6	5	26.06
19	15	172.1	171.0	116.4	79.4	25.0	1	27.65

10 and under females:

Swimmer	Age	Height	Wingspan	Shoulder Circumference	Waist	Foot Length	Years of Swimming	50 yard freestyle time
20	7	129.2	124.2	75.1	52.5	20.5	2	81.52
21	8	124.5	113.9	68.9	52.0	18.9	3	58.39
22	8	127.0	122.7	76.6	57.7	20.3	4	55.86
23	9	142.9	136.9	76.7	56.5	22.4	2	42.85
24	9	129.5	126.0	79.5	53.8	20.5	3	55.58
25	9	141.9	146.0	87.4	64.0	23.5	2	53.60
26	9	142.2	144.7	83.6	58.6	22.5	2	49.46
27	8	140.0	142.3	78.5	54.8	20.5	2	54.75

11 and up females:

Swimmer	Age	Height	Wingspan	Shoulder Circumference	Waist	Foot Length	Years of Swimming	50 yard freestyle time
28	11	169.9	161.6	100.3	72.2	25.0	1	35.95
29	11	157.5	155.8	90.4	62.9	24.6	5	35.32
30	12	153.7	149.2	96.5	69.6	23.0	5	36.09
31	14	168.3	168.8	116.6	89.2	26.4	12	29.05
32	14	167.6	165.7	100.4	65.4	22.6	5	29.63
33	14	163.8	157.5	90.3	62.0	22.6	6	38.67
34	14	167.6	161.5	96.5	63.8	25.1	1	32.15
35	14	170.8	172.0	109.0	77.6	26.5	5	28.68
36	15	161.3	165.9	101.5	73.2	24.0	4	34.92
37	17	164.8	165.1	99.1	66.2	24.4	1	33.06
38	17	172.7	178.8	104.4	75.5	25.8	11	28.07

The rest of this paper will focus on each data set individually and running statistical analyses involving multiple linear regression models. However, to understand the significance of the models and how they're formed, one would need to be aware of certain statistical definitions and methods. Therefore, these will be discussed in the next section.

Statistical Methods Used

To understand multiple regression models, one must first understand simple linear regression models. These models will always be in the form $y = \beta_0 + \beta_1 x + \varepsilon$ where:

- y is the dependent or response variable
- x is the independent or predictor variable
- ε is the random error component
- β_0 is the y-intercept of the line
- β_1 is the slope of the line

The model above is set up so that it fits the entire population. However, this is hardly ever known so we take a sample instead and set up a similar model based on this sample:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

This formula yields the least-squares line where $\hat{\beta}_0$ and $\hat{\beta}_1$ are estimates of the population parameters based off our sample and can be calculated as follows:

$$\hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

This introduces even more notation that needs to be discussed. Pearson's correlation coefficient (denoted by r) is a measure of the strength and direction of the linear relationship between x and y. This value always falls between -1 and 1. The closer the absolute value of r is to one, the stronger the relationship is between x and y. The formula for this correlation coefficient is:

$$r = \frac{SS_{xy}}{\sqrt{SS_{xx} SS_{yy}}}$$

where $SS_{xy} = \sum(x - \bar{x})(y - \bar{y})$ is "the sum of the cross products," $SS_{xx} = \sum(x - \bar{x})^2$ is "the sum of the squared deviation in x" and , $SS_{yy} = \sum(y - \bar{y})^2$ is "the sum of the squared deviation in y."

Also, r^2 is known as the coefficient of determination and represents the proportion of the total sample variability that is explained by the linear relationship between x and y. For a model to be considered a "good" prediction, it is best for r^2 to be close to one. This way we know the simple linear regression model explains most of the variability within the sample.

Now that we have a formula to predict a variable, how can we determine whether or not the predictor variable is a useful aid in predicting our dependent variable? Here one could use a hypothesis test set us as follows:

$$H_0: \beta_1 = 0$$

$$H_1: \beta_1 \neq 0$$

Essentially, if $\beta_1 = 0$ then for all values of x we would predict one constant value of y. So we assume this is true, and then reject if we have enough evidence to show that this is not true, and hence x is a useful predictor of y. To test this hypothesis, we would use a t-distribution with a t-statistic of

$$t = \frac{\hat{\beta}_1}{S_\varepsilon \Big/ \sqrt{SS_{xx}}}$$

where $S_\varepsilon = \sqrt{\frac{SS_{yy} - \hat{\beta}_1 SS_{xy}}{n-2}}$ and n is the number of observations in the sample. The p-value will be the area under the curve to the right and left of the value $\pm t$. This area represents the probability of observing what was true in the sample if the null hypothesis is true for the population. Therefore, we want this value to be small, meaning if $\beta_1 = 0$ then there is hardly any chance we should see this event happening, which leads us to believe β_1 is most likely not equal to 0. For this paper, a p-value of less that 10% or 0.10 will be used in order for an event to be called "statistically significant."

Instead of doing all of these calculations by hand, Excel will calculate them for us by producing an ANOVA table (Analysis of Variance Table). The bottom portion of the ANOVA table will label the rows with intercept and the X variable and the columns will be labeled Coefficients, Standard Error, t Stat, and P-value in order respectively. The coefficient column will give us $\hat{\beta}_0$ in the intercept row and $\hat{\beta}_1$ in the X row. Each row will also have a p-value. If the P-value is below 0.10, we consider that variable to be significant in predicting the y-variable. Also, from this table we can form the least square line by using $\hat{\beta}_0$ and $\hat{\beta}_1$ that is computed.

The top portion of the ANOVA table gives us data about the regression model as a whole. This becomes more important when we start adding more predictor variables. Regression models that include more than one independent variable are called regression models and follow the form

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$$

where:

- y is the dependent variable
- $x_1, x_2, \cdots, x_k$ are the independent variables
- β_i determines the contribution of the independent variable x_i

Multiple regression is very similar to simple linear regression except now we introduce an F-distribution which we use to determine the usefulness of the entire model instead of just one variable at a time. Here we can set up our hypothesis test for usefulness as

$$H_0: \beta_1 = \beta_2 = \cdots = \beta_k = 0$$

$$H_1: At\ least\ one\ \beta_i \neq 0\ for\ some\ 1 \leq i \leq k$$

Then we would use the F-statistic to calculate the p-value just like we did with the t-statistic before. The F-statistic and resulting p-value (labeled Significance F) can be found in the top portion of the ANOVA table. The top portion of the ANOVA table will have 3 rows: Model, Error, and Total. There will be 5 other columns, but our main focus will be on the F Value column and p-value. These are the cells we will use to determine how useful the regression model is.

The lower portion of an ANOVA table for multiple regression will be only slightly different by having more than one X-variable listed. All of the predictor variables will be shown with the value of $\hat{\beta}_i$ and the resulting t stat and p-value for each predictor variable.

To determine a multiple regression model for the swimmers, I first ran an analysis through excel that used all 7 predictor (independent) variables, and the one dependent variable. Essentially, I created a multiple regression equation of the following model:

$$\begin{aligned} 50\ yard\ freestyle\ time \\ = \hat{\beta}_0 + \hat{\beta}_1(age) + \hat{\beta}_2(height) + \hat{\beta}_3(wingspan) + \hat{\beta}_4(shoulder\ circumference) \\ + \hat{\beta}_5(waist) + \hat{\beta}_6(foot\ length) + \hat{\beta}_7(years\ of\ swimming) \end{aligned}$$

For each group I ran this analysis on, the model failed to be useful in predicting a swimmer's 50 freestyle time. Thinking there must be a better model out there, I used a method referred to as "Backward Elimination." Using this, I would eliminate the variable with the highest (least statistically significant) p-value, and run the regression analysis again. If the Significance F value is statistically significant (below 0.10) and each variable's p-value is statistically significant, then this is the final model used. If not, I again eliminate the variable with the highest p-value and run the regression analysis again. I continue doing this until all variables in the model show statistical significance.

Note: All statistical definitions, concepts, and equations, should be credited to the online notes from Paul Holmes' STAT 6315 class. See the reference page for more details. Also, if the reader would like further explanations of concepts, please consult the following textbook:

Introduction to the Practice of Statistics (7th Edition) by Moore, McCabe & Craig (ISBN-13: 978-1-4292-4032-1)

Males Ages 10 and Under

To determine a multiple regression model for the 10 and under males, I first ran an analysis through excel that used all 7 predictor (independent) variables, and the one dependent variable. Essentially, this follows the model

$$50\ yard\ freestyle\ time = \hat{\beta}_0 + \hat{\beta}_1(age) + \hat{\beta}_2(height) + \hat{\beta}_3(wingspan) + \hat{\beta}_4(shoulder\ circumference) + \hat{\beta}_5(waist) + \hat{\beta}_6(foot\ length) + \hat{\beta}_7(years\ of\ swimming)$$

where β_0 is the constant intercept and each $\hat{\beta}_1, \hat{\beta}_2, \cdots, \hat{\beta}_7$ are slope coefficients for each independent variable.

Excel produced the following ANOVA table:

Multiple R	0.95804
R Square	0.91783
Adjusted R Square	0.34267
Standard Error	9.56116
Observations	9.00000

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	7.00000	1021.1551	145.8793	1.5958	0.5454
Residual	1.00000	91.4158	91.4158		
Total	8.00000	1112.5710			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	89.23278	137.7734	0.6477	0.6341	-1661.3439	1839.8095
Age	-0.41282	11.9811	-0.0345	0.9781	-152.6473	151.8217
Height	-1.76945	2.0229	-0.8747	0.5425	-27.4732	23.9343
Wingspan	1.23682	3.1220	0.3962	0.7599	-38.4317	40.9054
Shoulder Circumference	0.29902	3.5928	0.0832	0.9471	-45.3518	45.9498
Waist	0.76151	1.5770	0.4829	0.7136	-19.2765	20.7995
Foot length	-0.57391	7.4224	-0.0773	0.9509	-94.8848	93.7370
Years of swimming	-7.51347	6.2439	-1.2033	0.4414	-86.8502	71.8233

Using the backwards elimination method I removed the following variables one by one (in order): age, shoulder circumference, foot length, and wingspan. This leaves the following variables in the model: Height, waist, and years of swimming. The following ANOVA table showed my final regression analysis:

Regression Statistics	
Multiple R	0.9296
R Square	0.8642
Adjusted R Square	0.7827
Standard Error	5.4970
Observations	9

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	3	961.4858	320.4953	10.6064	0.0132
Residual	5	151.0852	30.2170		
Total	8	1112.5710			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	96.7689	26.3950	3.6662	0.0145	28.9183	164.6195
Height	-0.6917	0.3124	-2.2142	0.0777	-1.4947	0.1113
Waist	1.0852	0.4644	2.3369	0.0666	-0.1085	2.2789
Years of swimming	-6.5972	1.8340	-3.5972	0.0156	-11.3117	-1.8827

This gives us the following model to predict a 10 and under male swimmer's 50 yard freestyle time:

$$50\ yard\ freestyle\ time = 96.7689 - 0.6917(height) + 1.0852(waist) - 6.5972(years\ of\ swimming)$$

Examining these coefficients we see that (holding all else constant):

- For each 1cm increase in height, we predict the swimmer to be 0.6917 seconds faster in the 50 freestyle
- For each 1cm increase in waist, we predict the swimmer to be 1.0852 seconds slower in the 50 freestyle
- For each 1 additional year of swimming, we predict the swimmer to be 6.5972 seconds faster in the 50 freestyle

Overall, our model is statistically significant at the <2% level, which falls below the 10% mark and 86.42% of the variability in the sample is explained through this model.

Males 11 and Up

To determine a multiple regression model for the 11 and up males, I first ran an analysis through excel that used all 7 predictor (independent) variables, and the one dependent variable. Essentially, this follows the model

$$50\ yard\ freestyle\ time = \hat{\beta}_0 + \hat{\beta}_1(age) + \hat{\beta}_2(height) + \hat{\beta}_3(wingspan) + \hat{\beta}_4(shoulder\ circumference) + \hat{\beta}_5(waist) + \hat{\beta}_6(foot\ length) + \hat{\beta}_7(years\ of\ swimming)$$

where β_0 is the constant intercept and each $\hat{\beta}_1, \hat{\beta}_2, \cdots, \hat{\beta}_7$ are slope coefficients for each independent variable.

Excel produced the following ANOVA table:

Regression Statistics	
Multiple R	0.9448
R Square	0.8927
Adjusted R Square	0.5170
Standard Error	2.3935
Observations	10

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	7	95.2979	13.6140	2.3765	0.3279
Residual	2	11.4573	5.7286		
Total	9	106.7552			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	88.9478	25.9534	3.4272	0.0756	-22.7205	200.6161
Age	-1.4453	2.3535	-0.6141	0.6017	-11.5716	8.6810
Height	0.0155	0.4865	0.0318	0.9775	-2.0779	2.1088
Wingspan	-0.5773	0.5123	-1.1271	0.3768	-2.7814	1.6267
Shoulder Circumference	0.4713	0.7700	0.6121	0.6028	-2.8417	3.7843
Waist	-0.3503	0.4445	-0.7880	0.5133	-2.2630	1.5624
Foot length	1.1899	1.7106	0.6956	0.5586	-6.1704	8.5502
Years of swimming	-0.0609	0.7406	-0.0823	0.9419	-3.2475	3.1257

Using the backwards elimination method I removed the following variables one by one (in order): height, years of swimming, foot length, waist, shoulder circumference, and age. This left only wingspan in the model. This meant that a simple linear regression model was the best fit for the data I obtained. The following ANOVA table showed my final regression analysis:

Regression Statistics	
Multiple R	0.8372
R Square	0.7009
Adjusted R Square	0.6635
Standard Error	1.9980
Observations	10

ANOVA

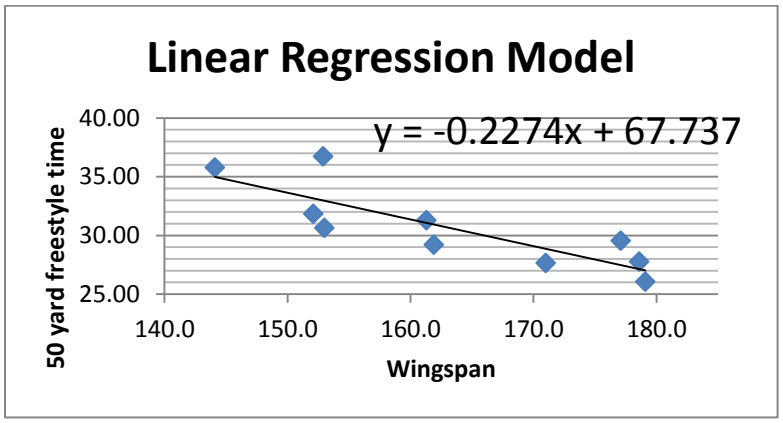
	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	74.8196	74.8196	18.7427	0.0025
Residual	8	31.9355	3.9919		
Total	9	106.7552			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	67.7368	8.5893	7.8862	0.0000	47.9298	87.5438
Wingspan	-0.2274	0.0525	-4.3293	0.0025	-0.3485	-0.1063

This gives us the following model to predict a 10 and under male swimmer's 50 yard freestyle time:

$$50\ yard\ freestyle\ time = 67.7368 - 0.2274(wingspan)$$

Examining the slope of this line we see that for each 1cm increase in wingspan, we predict the swimmer to be 0.2274 seconds faster in the 50 freestyle. This model would produce the following graph along with a scatterplot of the data.



Overall, our model is statistically significant at the <1% level, which falls way below the 10% mark and 70.09% of the variability in the sample is explained through this model.

Females 10 and Under

To determine a multiple regression model for the 10 and under males, I first ran an analysis through excel that used all 7 predictor (independent) variables, and the one dependent variable. Essentially, this follows the model

$$50\ yard\ freestyle\ time = \hat{\beta}_0 + \hat{\beta}_1(age) + \hat{\beta}_2(height) + \hat{\beta}_3(wingspan) + \hat{\beta}_4(shoulder\ circumference) + \hat{\beta}_5(waist) + \hat{\beta}_6(foot\ length) + \hat{\beta}_7(years\ of\ swimming)$$

where β_0 is the constant intercept and each $\hat{\beta}_1, \hat{\beta}_2, \cdots, \hat{\beta}_7$ are slope coefficients for each independent variable.

Excel produced the following ANOVA table:

SUMMARY OUTPUT

Regression Statistics	
Multiple R	1
R Square	1
Adjusted R Square	65535
Standard Error	0
Observations	8

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	7	878.1851	125.4550	#NUM!	#NUM!
Residual	0	0.0000	65535.0000		
Total	7	878.1851			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	216.7821	0.0000	65535.0000	#NUM!	216.7821	216.7821
Age	-12.1284	0.0000	65535.0000	#NUM!	-12.1284	-12.1284
Height	-0.6819	0.0000	65535.0000	#NUM!	-0.6819	-0.6819
Wingspan	-0.6564	0.0000	65535.0000	#NUM!	-0.6564	-0.6564
Shoulder Circumference	1.8289	0.0000	65535.0000	#NUM!	1.8289	1.8289
Waist	-0.7086	0.0000	65535.0000	#NUM!	-0.7086	-0.7086
Foot Length	1.6533	0.0000	65535.0000	#NUM!	1.6533	1.6533
Years of Swimming	-7.3799	0.0000	65535.0000	#NUM!	-7.3799	-7.3799

As you can see, Excel was producing #NUM! for all p-values, and for the F-statistic and Significance of F value. In short, this shows that some sort of error within the data set is keeping Excel from being able to calculate what is being asked. After some further exploration, I found that my sample size was too small for the amount of variables I was including. However, I have a suspicion that height and wingspan might be dependent, and hence not really adding two separate variables to the model. (I.e. wingspan and height might be adding the same information to the model.) To test this idea, I ran a Chi-Square analysis between height and wingspan. Essentially, a Chi-Square distribution will inform me if the two variables are in fact dependent on each other.

For this analysis, I needed the data to be put into a contingency table, and hence I needed the data to be categorical. For logistic reasons, I took the range of data I had for height, divided it by two, and made two categories: short and long where the interval for short was 124.5cm-133.7cm and the interval for long was 133.8cm-143.0cm. I did the same thing for wingspan and decided that the interval for short was 113.9cm-130cm and the interval for long was 130.1cm-146.2cm. Then I was able to form the following contingency table:

Wingspan	Height		
	Small	Long	Total
Small	4	0	4
Long	0	4	4
Total	4	4	8

For this analysis, we will set up a hypothesis test as follows:

$$H_0: Wingspan\ and\ Height\ are\ independent\ variables$$

$$H_1: Wingpsan\ and\ Height\ are\ dependent\ variables$$

Our test statistic for this analysis will be

$$\chi^2 = \sum \frac{\left[n_{ij} - E(n_{ij})\right]^2}{E(n_{ij})}$$

where *i* and *j* refer to the row and column number respectively, and $E(n_{ij}) = \frac{(row\ total)\times(column\ total)}{total\ sample\ size}$.

In this situation, $E(n_{11}) = E(n_{12}) = E(n_{21}) = E(n_{22}) = \frac{4\times4}{8} = 2$. Then,

$$\chi^2 = \frac{(4-2)^2}{2} + \frac{(0-2)^2}{2} + \frac{(0-2)^2}{2} + \frac{(4-2)^2}{2} = 2 + 2 + 2 + 2 = 8$$

Using 8 as our test statistic and a degrees of freedom of (number of rows – 1) x (number of columns – 1) = (2-1) x (2-1) = 1, we get a p-value of 0.0047 which is below 0.1, so we reject the null hypothesis. Hence, wingspan and height are dependent on each other, and there is no need to include both variables in the model. To determine which variable to delete, I ran a simple linear regression analysis on both variables

separately and determined that height had a significance level of 0.149419 and wingspan had a significance level of 0.248615. Since wingspan was least statistically significant, I deleted that one and ran a multiple regression analysis and received the following ANOVA table:

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.9999
R Square	0.9998
Adjusted R Square	0.9983
Standard Error	0.4685
Observations	8

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	6	877.9656	146.3276	666.6838	0.0296
Residual	1	0.2195	0.2195		
Total	7	878.1851			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	268.7939	6.3892	42.0701	0.0151	187.6116	349.9761
Age	-11.0340	0.3880	-28.4402	0.0224	-15.9637	-6.1044
Height	-1.6231	0.0620	-26.1894	0.0243	-2.4105	-0.8356
Shoulder Circumference	1.1058	0.0700	15.8067	0.0402	0.2169	1.9947
Waist	-0.9182	0.1330	-6.9016	0.0916	-2.6087	0.7723
Foot Length	3.8625	0.4260	9.0661	0.0699	-1.5508	9.2759
Years of Swimming	-7.1607	0.5686	-12.5940	0.0504	-14.3853	0.0638

Surprisingly, this showed that every remaining independent variable was statistically significant at the 10% level, so the regression model to predict a 10 and under female's 50 yard freestyle time is:

$$\begin{aligned} 50\ yard\ freestyle\ time \\ &= 268.7939 - 11.0340(age) - 1.6231(height) \\ &+ 1.1058(shoulder\ circumference) - 0.9182(waist) + 3.8625(foot\ length) \\ &- 7.1607(years\ of\ swimming) \end{aligned}$$

Examining these coefficients we see that (holding all else constant):

- For a one year increase in age, we predict the swimmer to be 11.0340 seconds faster in the 50 freestyle.

- For a 1cm increase in height, we predict the swimmer to be 1.6231 seconds faster in the 50 freestyle.
- For a 1cm increase in the shoulder circumference, we predict the swimmer to be 1.1058 seconds slower in the 50 freestyle.
- For a 1cm increase in the waist, we predict the swimmer to be 0.9182 seconds faster in the 50 freestyle.
- For a 1cm increase in the foot length, we predict the swimmer to be 3.8625 seconds slower in the 50 freestyle.
- For a one year increase in years of swimming, we predict the swimmer to be 7.1607 seconds faster.

Some of these coefficients do seem strange to me. In particular, I do not expect a bigger foot size to make a swimmer slower. These unusual responses could be due to a small sample size. Therefore, I would recommend further studies be done with a larger sample of swimmers.

Overall, our model is statistically significant at the <3% level, which falls below the 10% mark and 99.98% of the variability in the sample is explained through this model.

Females 11 and Up

To determine a multiple regression model for the 11 and up females, I first ran an analysis through excel that used all 7 predictor (independent) variables, and the one dependent variable. This follows the model

$$50\ yard\ freestyle\ time = \hat{\beta}_0 + \hat{\beta}_1(age) + \hat{\beta}_2(height) + \hat{\beta}_3(wingspan) + \hat{\beta}_4(shoulder\ circumference) + \hat{\beta}_5(waist) + \hat{\beta}_6(foot\ length) + \hat{\beta}_7(years\ of\ swimming)$$

where β_0 is the constant intercept and each $\hat{\beta}_1, \hat{\beta}_2, \cdots, \hat{\beta}_7$ are slope coefficients for each independent variable.

Excel produced the following ANOVA table:

Regression Statistics	
Multiple R	0.9793
R Square	0.9590
Adjusted R Square	0.8633
Standard Error	1.3328
Observations	11

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	7	124.6111	17.8016	10.0210	0.0424
Residual	3	5.3293	1.7764		
Total	10	129.9404			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	89.9869	12.3581	7.2816	0.0053	50.6578	129.3161
Age	0.3728	0.3612	1.0321	0.3779	-0.7768	1.5224
Height	0.2021	0.1581	1.2779	0.2912	-0.3011	0.7052
Wingspan	-0.1890	0.1695	-1.1148	0.3462	-0.7285	0.3505
Shoulder Circumference	-1.0816	0.2430	-4.4505	0.0211	-1.8550	-0.3082
Waist	0.9961	0.2441	4.0808	0.0266	0.2193	1.7730
Foot Length	-0.9819	0.5336	-1.8402	0.1630	-2.6799	0.7162
Years of Swimming	-0.4724	0.1824	-2.5896	0.0811	-1.0529	0.1081

Here we see that the overall model produced is in fact statistically significant since Significance F = 0.0424. However, I would prefer each variable to also be statistically significant so using the backwards elimination method I removed the following variables one by one (in order): age, wingspan, and height. This left the following variables in the model: shoulder circumference, waist, foot length, and years of swimming. The following ANOVA table showed my final regression analysis:

Regression Statistics	
Multiple R	0.9658
R Square	0.9327
Adjusted R Square	0.8879
Standard Error	1.2071
Observations	11

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	4	121.1976	30.2994	20.7939	0.0012
Residual	6	8.7428	1.4571		
Total	10	129.9404			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	96.4917	8.2991	11.6267	0.0000	76.1845	116.7989
Shoulder Circumference	-0.9834	0.1531	-6.4255	0.0007	-1.3580	-0.6089
Waist	0.8897	0.1624	5.4796	0.0015	0.4924	1.2871
Foot Length	-1.0263	0.3908	-2.6264	0.0392	-1.9825	-0.0702
Years of Swimming	-0.4977	0.1410	-3.5306	0.0124	-0.8426	-0.1528

This gives us the following model to predict an eleven and up female swimmer's 50 yard freestyle time:

$$50\ yard\ freestyle\ time = 96.4917 - 0.9834(shoulder\ circumference) + 0.8897(waist) - 1.0263(foot\ length) - 0.4977(years\ of\ swimming)$$

Examining these coefficients we see that (holding all else constant):

- For each 1cm increase in shoulder circumference, we predict the swimmer to be 0.9834 seconds faster in the 50 freestyle
- For each 1cm increase in waist, we predict the swimmer to be 0.8897 seconds slower in the 50 freestyle
- For each 1 cm increase in footlength, we predict the swimmer to be 1.0263 seconds faster in the 50 freestyle

- For each one additional year of swimming, we predict the swimmer to be 0.4977 seconds faster in the 50 freestyle

Overall, our model is statistically significant at the <1% level, which falls way below the 10% mark and 93.27% of the variability in the sample is explained through this model.

References

Holmes, P. *Course Notes 2* [PDF document]. Retrieved from the University of Georgia's eLearning Commons website: https://www.elc.uga.edu/webct/urw/lc3278721206011.tp3400373671041//RelativeResourceManager?contentID=3400379816041

Holmes, P. *Course Notes 3* [PDF document]. Retrieved from the University of Georgia's eLearning Commons website: https://www.elc.uga.edu/webct/urw/lc3278721206011.tp3400373671041//RelativeResourceManager?contentID=3400379835041

Holmes, P. *Course Notes 4* [PDF document]. Retrieved from the University of Georgia's eLearning Commons website: https://www.elc.uga.edu/webct/urw/lc3278721206011.tp3400373671041//RelativeResourceManager?contentID=3400379878041