Practice Questions: Multiple Regression with Categorical Predictors

A sample of 43 active women (red) and 44 active men (green) who exercise at a local gym was used to study differences between men and women in physical activity. Each individual ran on a treadmill of increasing tilt until exhaustion. The duration in seconds (Duration_Sec) of each effort was recorded. In addition, the study recorded the maximum rate of oxygen use (VO_2 max) and the maximum observed heart rate (HR max, in beats per minute), as well as the Age (in years) and body mass index (BMI). Questions follow the output.

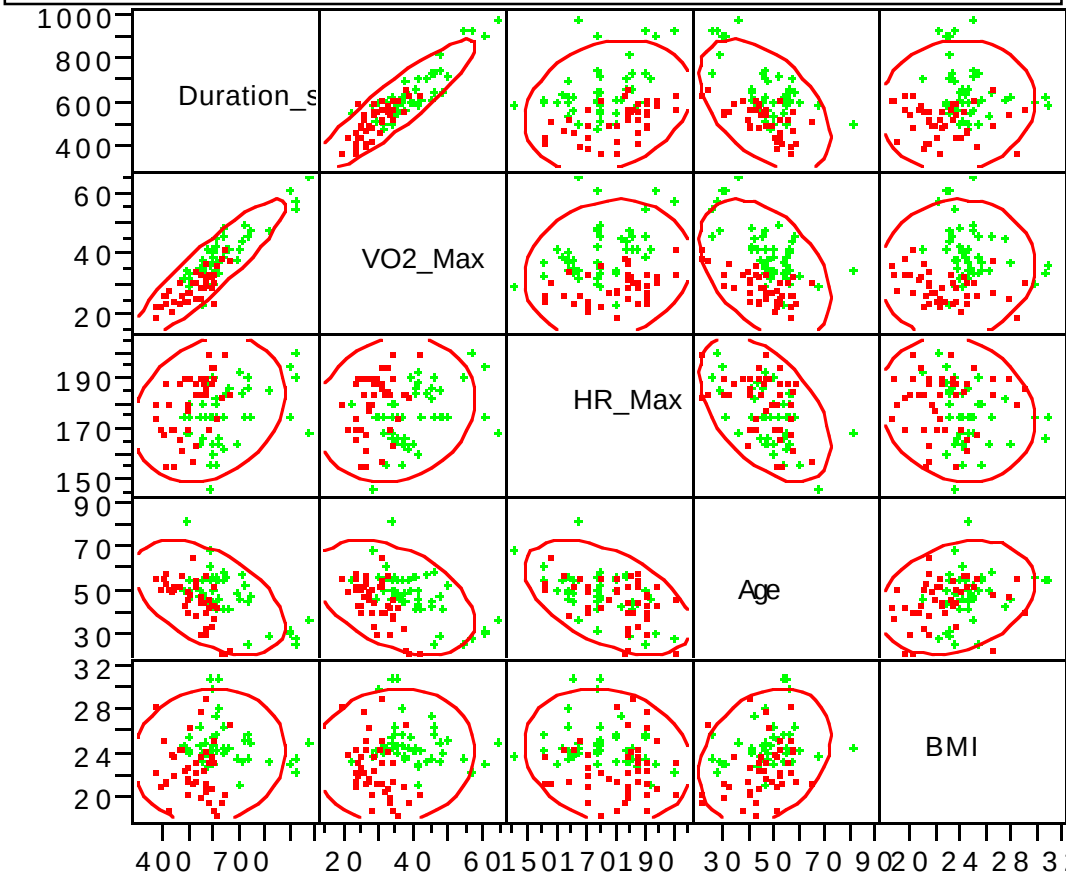
Summary of Duration_Sec, by Sex

Level	Number	Mean	Std Dev
female	43	500	81
male	44	650	110

Correlations

Variable	Duration_Sec	VO2_Max	HR_Max	Age	BMI
Duration_Sec	1.00	0.91	0.24	-0.49	0.08
VO2_Max	0.91	1.00	0.14	-0.42	0.07
HR_Max	0.24	0.14	1.00	-0.50	-0.16
Age	-0.49	-0.42	-0.50	1.00	0.27
BMI	0.08	0.07	-0.16	0.27	1.00

Scatterplot Matrix



MODEL 1

Response: Duration_Sec

RSquare	0.833
Root Mean Square Error	51.2
Observations	87

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-51.49	87.02	-0.59	0.56
Sex[female]	-3.22	7.75	-0.42	0.68
Sex[male]	3.22	7.75	0.42	0.68
VO2_Max	11.58	0.82	14.08	0.00
HR_Max	1.29	0.53	2.45	0.02

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	3	1088844	362948	138.36
Error	83	217726	2623	**Prob>F**
C Total	86	1306570		<.0001

(1) If we were to regress *Duration* on *VO2Max*, how much variation would be explained in the simple regression?

(2) What information can you glean from the scatterplot matrix that is not apparent in the correlation summary?

(3) Based on the fit of MODEL 1, if a woman increases her maximum heart rate by 10 beats per minute while retaining her maximum VO_2, then how can she expect her time on the treadmill to change? Give a confidence interval. How would the interval differ for a man?

(4) Based on the results reported in MODEL 1, how much longer (or shorter) can women on average stay on the treadmill relative to men? If there is no difference, say so; if there is a difference, report it with a confidence interval.

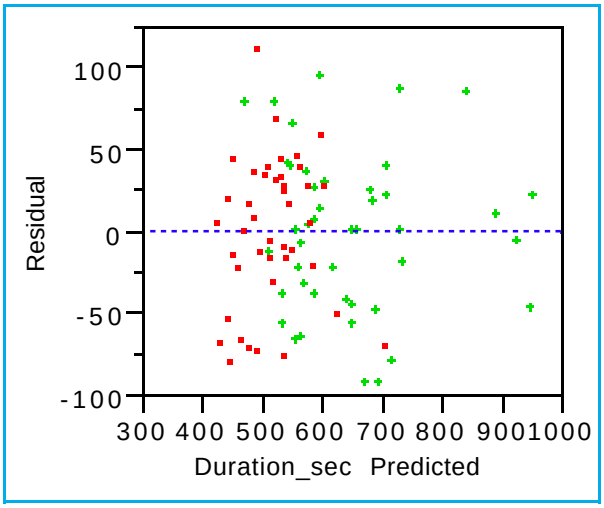
(5) Suppose that we repeated this regression analysis, but with four times as many men and four times as many women in the study (an independent sample from the same population). Then the model fit to this larger data set would reliably show which of the following characteristics?

(a) A much smaller value for the RMSE.
(b) A substantially increased value for R^2.
(c) A substantially decreased value for R^2.
(d) A negative slope for the *Sex* categorical predictor.
(e) A larger, more significant overall F-ratio.

(6) A local fitness coach who trains tri-athletes uses predictions from MODEL 1 to screen athletes before he accepts them for training. The predicted endurance of one candidate was 405 seconds, whereas that of another was 425 seconds. Assuming both candidates are women with comparable max VO_2 and max HR, does this model imply that the athlete with the lower time is in significantly better shape?

(7) Does MODEL 1 explain significantly more variation in duration time on the treadmill than a regression that uses only *Sex* and VO_2 as predictors?

(8) The following plot shows the residuals from MODEL 1. Men appear as small green crosses +, and women are small red dots. The points on the right side of the plot are predominantly green crosses. Does this plot indicate a problem with the model?



(9) How should the analysis in Model 1 be extended/continued? For example, what potential features of the model are ignored in this analysis?

(1) The correlation of VO_2 max with the response is 0.91, so this single predictor would explain 0.91^2% of the variation in duration.

(2) With color coding, we can see that women have less duration than men (red points are lower on the duration axis). Women also have smaller values of VO_2, consistent with the high correlation between duration and VO_2. A few outlying men are seen with very large duration.

(3) An increase of 10 in max HR produces an expected change of 10(1.29) seconds. The SE is 0.53 so the CI for a change of 1 in max HR is $[1.29 \pm 2(.53)] = [0.23, 2.35]$. Thus the interval for a change of 10 is [2.3, 23.5]. The interval for a man is the same since the model lacks an interaction term.

(4) The coefficient for the categorical term *Sex* is half the difference in the fitted intercepts, but this effect is not significant in Model 1.

(5) We could only be assured of a more significant overall fit as measured by the F-ratio. More data, as we have seen in class, does not improve the fit as measured by RMSE and R^2. On the other hand, the F-ratio is very sensitive to the number of observations as reflected in the approximation

$$F - ratio \approx \frac{R^2}{1-R^2} \times \frac{\#\, observations}{\#\, slopes}$$

(6) Since the athlete with the lower time does not stay on so long, she is doing worse (longer times are good in this context). Since the RMSE of the fitted model is 51.2 sec, the difference in performance of these two women is slight given the accuracy of this model.

(7) The t-ratio for max HR in this multiple regression measures the significance of the addition of this predictor to the fit; thus, this model explains significantly more variation. You don't need a partial F for this one – the t-ratio is the partial F when just one predictor is involved.

(8) The crosses at the right of the plot indicate larger predicted times for the men in the analyzed data. The plot does not, however, indicate a lack of constant variance. The data are simply sparse on the right side of the plot.

(9) Add some interactions between Sex and the other predictors.