Biostatistics 200B Midterm Exam Winter 2018

Name: David Levy

1. Data were collected from 102 male and 100 female athletes whose sports were either track, swimming or soccer. The variables include the following. Some descriptive statistics by sex are also provided.

Variable	Description
Hg	Hemoglobin concentration, g/dL
Bfat	Percent body fat
Hc	Hematocrit, volume percentage
RCC	Red blood cell count, cells/μL
WCC	White blood cell count, cells/μL
Sex	Sex (0 = male, 1 = female)
Track	Sport is track (1 = Yes, 0 = No)
Swimming	Sport is swimming (1 = Yes, 0 = No)

TABLE 1. Descriptive Statistics and Pearson Correlation Coefficients by Sex

Men

Sex=0

Variable	N	Mean	Std Dev	Minimum	Maximum
Hg	102	15.55294	0.93424	13.50000	19.20000
Bfat	102	9.25088	3.18469	5.63000	19.94000
Hc	102	45.65000	2.56846	40.30000	59.70000
RCC	102	4.98647	0.39229	4.00000	6.72000
WCC	102	7.22108	1.89960	3.90000	14.30000

Pearson Correlation Coefficients

	Hg	Bfat	Hc	RCC	WCC
Hg	1.00000	0.10175	0.89496	0.70445	0.06896
Bfat	0.10175	1.00000	0.11157	-0.19909	0.37997
Hc	0.89496	0.11157	1.00000	0.78597	0.11436
RCC	0.70445	-0.19909	0.78597	1.00000	-0.02421
WCC	0.06896	0.37997	0.11436	-0.02421	1.00000

Women

Sex=1

Variable	N	Mean	Std Dev	Minimum	Maximum
Hg	100	13.56000	0.92332	11.60000	15.90000
Bfat	100	17.84910	5.45289	8.07000	35.52000
Hc	100	40.48200	2.62465	35.90000	47.10000
RCC	100	4.40450	0.32090	3.80000	5.33000
WCC	100	6.99400	1.69544	3.30000	13.30000

Pearson Correlation Coefficients

	Hg	Bfat	Hc	RCC	WCC
Hg	1.00000	-0.13299	0.90343	0.77678	0.20142
Bfat	-0.13299	1.00000	-0.19405	-0.13925	0.11867
Hc	0.90343	-0.19405	1.00000	0.85220	0.19898
RCC	0.77678	-0.13925	0.85220	1.00000	0.24054
WCC	0.20142	0.11867	0.19898	0.24054	1.00000

(a) (15 points) Calculate $\hat{\beta}_0$ and $\hat{\beta}_1$, the estimated intercept and slope for the simple linear regression of Hg (response) on Bfat (predictor) if we fit the model only to the data for men.

$$\hat{\beta}_1 = r\frac{S_Y}{S_X} = (0.10175)\left(\frac{0.93424}{3.18469}\right) = 0.02985$$ ✓

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1\bar{X} = (15.55294) - (0.02985)(9.25088) = 15.27681$$

(b) (12 points) Regressing Hg on Hc, Bfat, RCC, Track, Swimming, and interactions between Bfat and Track and between Bfat and Swimming for the full sample (men and women) produces the output below. Three items in the output are missing and indicated by blanks. Fill in three missing values. −4

$RSS = TSS - RegSS = 373.11109 - 32.78648$

TABLE 2

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model(or Regression) SS	7	340.32461	48.61780	287.68	<.0001
Error (or Residual) SS	194	32.78648	0.16900		
Total SS	201	373.11109			

$MSE = \frac{RSS}{n-(k+1)} = \frac{32.78648}{202-8}$

$t = \frac{\hat{\beta}}{SE(\hat{\beta})} \Rightarrow \hat{\beta} = t\,SE(\hat{\beta}) = (-1.26)(0.00663)$

Variable	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	-0.34944	0.45524	-0.77	0.4437
Hc	0.35380	0.01843	19.20	<.0001
Bfat	-------	0.00663	-1.26	0.2085
RCC	-0.03706	0.14469	-0.26	0.7981
track	0.59049	0.22669	2.60	0.0099
swimming	-0.26955	0.24461	-1.10	0.2718
bfatXtrack	-0.08062	0.02341	-3.44	0.0007
bfatXswimming	0.01375	0.01979	0.69	0.4881

$\hat{\beta}_1 = \frac{\sum(X_i-\bar{X})(Y_i-\bar{Y})}{\sum(X_i-\bar{X})^2} \rightarrow (S_X)^2(n-1)$

1. Regress Y on all other variables
2. Regress this variable on all other
3. Regress (1&2) residuals of (1) on residuals of (2) and we'll get the parameter

(c) (10 points) What is the value of R^2 for the model in Table 2?

$$R^2 = 1 - \frac{RSS}{TSS} = 1 - \left(\frac{32.78648}{373.11109}\right) = 0.91$$ ✓

(d) (12 points) In Table 2, the F value is 287.68 and its p-value is < .0001. What hypothesis test does this F statistic correspond to? What do you conclude based on it? −1

The hypothesis test that corresponds with the F-statistic is the null H_0 that all of the regression coefficients (parameters) except intercept have a true value of zero. We have sufficient evidence at any reasonable confidence level to reject this null hypothesis in favor of the alternative, that at least one of the parameters does not have a true value of zero.

(e) (12 points) Provide a quantitative interpretation of the regression coefficient for the interaction term bfatXtrack using language that a non-statistician could understand.

The regression coefficient in question tells us that those who run track have a 0.08602 g/dL lower increase in hemoglobin concentration per additional % body fat, controlling for all of the other variables in the model (holding them constant). Compared to soccer

-2

(f) (14 points) When the model in Table 2 is fit without the two interactions terms, the residual sum of squares is 34.94960. Conduct a test that the regression coefficients for the two interaction terms are both equal to zero. (Provide the value of the test statistic and its DF under the null. You do not need to get a p-value.) If the null hypothesis is rejected, what would you conclude? If the null is not rejected, what would you conclude?

We are essentially performing a partial F-test. Full model is found in table, previous page, reduced is where the two coefficients on the interaction terms are equal to zero.

$$F = \frac{\left(\frac{\Delta RSS}{\Delta df}\right)}{MSE_{Full}} = \left(\frac{(34.94960 - 32.78648)}{2}\right) \Big/ 0.16900 = 6.40 \sim F_{2,194} \text{ under } H_0$$

✓

-2

If rejected ---?

If null is not rejected, we conclude that regression coefficients of interaction terms = 0. There is no interaction in the model with those other variables included.

(g) (13 points) Regressing natural log-transformed body fat on hemoglobin (Hg) yields the following output:

	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	5.07397	0.29073	17.45	<.0001
Hg	-0.17661	0.01987	-8.89	<.0001

Provide a quantitative interpretation of the relationship between body fat and hemoglobin level and body fat using language that a non-statistician could understand.

✓

$e^{-0.17661} = 0.838$

In general, when we natural log transform Y, the effect of a one unit increase in X is to multiply the conditional mean of Y by $e^{\hat{\beta}_1}$. So our interpretation is that for every 1 g/dL increase in hemoglobin concentration, we would expect a 16.2% reduction in body fat. As a side note, this seems less sensible than using body fat as the X variable. I will leave as is because of how the problem was presented.

2. (12 points) Suppose that we have data that follow the model $\mathbf{Y} = \mathbf{X}\beta + \epsilon$, $E(\epsilon) = \mathbf{0}$, $COV(\epsilon) = \sigma^2\mathbf{I}$. Under what condition on the design matrix $\mathbf{X}$ can we obtain unique least squares estimates of the regression parameters, $\hat{\beta}$, and why is this condition necessary?

-10

We are essentially solving a system of equations that we obtained by taking partial derivatives of the model (as a function of β) with respect to each component of β. We can extend from models with 1 or 2 predictors to say that our design matrix must have the same number of columns as β has rows, so that we have a system with the same number of unknowns and equations. Otherwise, some least squares estimates of the regression parameters will be functions rather than distinct values.

-8