STAT GU4205/GR5205 (Section 004) Linear Regression Models

Fall 2019

Midterm II

Write your name and UNI in the spaces provided above.
Do not turn over this page until instructed to do so.

Name:

UNI:

If you are in 4205, write Yes (otherwise, leave it blank):

Instructions: Write your name and UNI in the spaces provided above. Do not turn over this page until instructed to do so. You have 80 minutes to complete this examination. You are permitted one letter size cheat sheet (can write on both sides). No other outside material or assistance is permitted. There are 9 pages to this exam, which is worth a total of 73 (for students in 5205) and 68 (for students in 4205).

Please sign below to indicate your agreement with the Columbia College Honor Code, whether or not you are a student of Columbia College.

I affirm that I will not plagiarize, use unauthorized materials, or give or receive illegitimate help on assignments, papers, or examinations. I will also uphold equity and honesty in the evaluation of my work and the work of others. I do so to sustain a community built around this Code of Honor.

Signature:

1. (18 points) True/False. Write the answers on the left of the questions.

   (a) Letting R_1^2 denote the coefficient of determination for the regression `y` $\sim$ `x1`, and R_2^2 denote that for `y` $\sim$ `x2`, and R_{12}^2 that for `y` $\sim$ `x1 + x2`, we must have $R_{12}^2 \geq \max\{R_1^2, R_2^2\}$.

   (b) Letting $\hat{\sigma}_1^2$ denote the estimated residual variance for the regression `y` $\sim$ `x1`, and $\hat{\sigma}_2^2$ that for `y` $\sim$ `x2`, and $\hat{\sigma}_{12}^2$ that for `y` $\sim$ `x1 + x2`, it is necessarily the case that $\hat{\sigma}_{12}^2 \geq \max\{\hat{\sigma}_1^2, \hat{\sigma}_2^2\}$.

   (c) Researchers wishing to study the relationship between cholesterol and patient height and weight consider a regression model with mean function $\mathbb{E}(Y|X = x) = \beta_0 + \beta_1 x$, where Y = LDL cholesterol in mg/dL, and X = BMI = weight in kg /(height in m)^2; in the terminology of this course, BMI is the predictor, and height and weight are the two regressors.

   (d) Given n observations from a model with the mean function

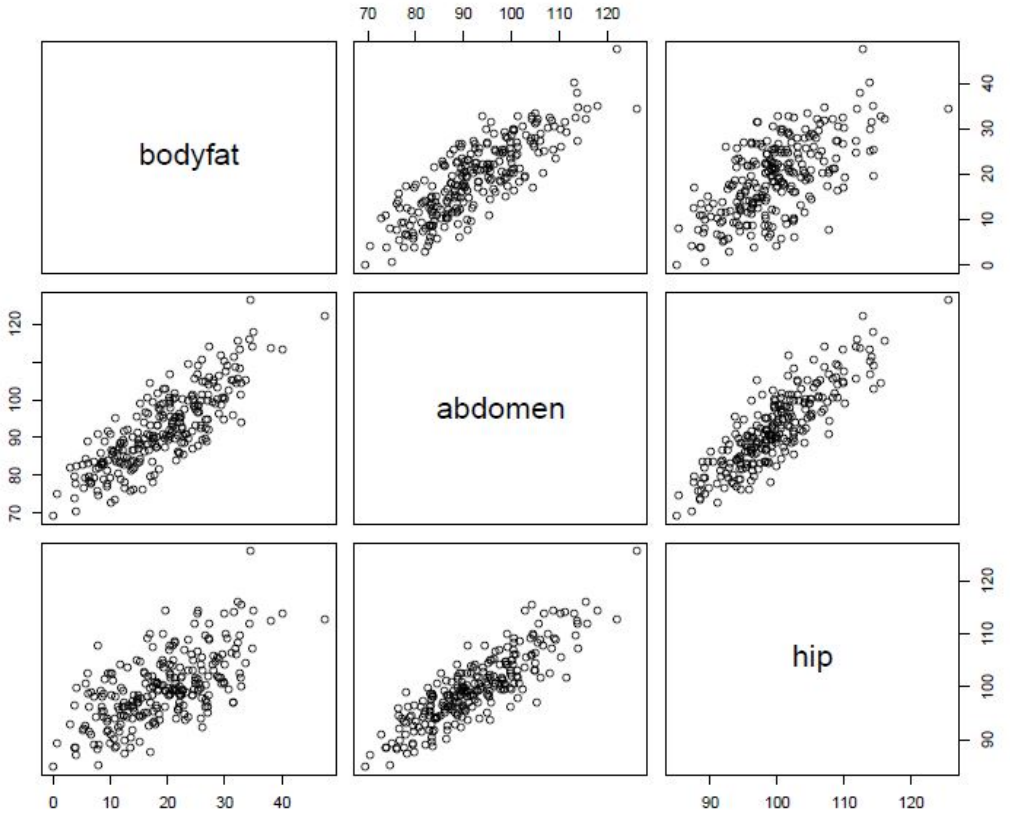
   $$\mathbb{E}(Y|X_1 = x_1, \ldots, X_p = x_p) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p,$$

   the vector of responses $\boldsymbol{Y} = (Y_1, \ldots, Y_n)'$ can be written $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where $\boldsymbol{X}$ is $n \times p$, $\boldsymbol{\beta}$ is $p \times 1$, and $\mathbb{E}(\boldsymbol{\varepsilon}) = \boldsymbol{0}$.

   (e) Consider the multiple linear regression model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$, where $\mathbb{E}(\boldsymbol{e}) = \boldsymbol{0}$ and $\mathrm{Var}(\boldsymbol{e}) = \sigma^2 \boldsymbol{I}$. As long as $\boldsymbol{X}$ has full column rank, the least squares estimates $\hat{\boldsymbol{\beta}}$ are unique and satisfy $\mathbb{E}(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$ and $\mathrm{Var}(\hat{\boldsymbol{\beta}}) = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$; if in addition that the error term is multivariate normal, then the sampling distribution of $\hat{\boldsymbol{\beta}}$ is multivariate normal as well.

   (f) If the effect of the predictor variable X_2 differ depending on the value of the predictor variable X_1 (and the effect of X_1 varies depending on the value of X_2), then the two predictor variables are said to have an interaction effect; one common approach to modeling interaction is to include a product term $(X_1 X_2)$ as a regressor.

2. (10 points) Measurements were made on n = 251 men in order to relate the percentage of body fat determined by underwater weighing (`bodyfat`), which is inconvenient and costly to obtain, to abdomen circumference in cm (`abdomen`) and hip circumference in cm (`hip`), both recorded using only measuring tape.



Partial `summary.lm()` R output for `bodyfat` $\sim$ `abdomen + hip`:

```
> summary(m3)

Call:
lm(formula = bodyfat ~ abdomen + hip)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -25.92937    5.02802  -5.157 5.14e-07
abdomen       0.86159    0.05549  15.528  < 2e-16
hip          -0.34637    0.08709  -3.977 9.16e-05

Residual standard error: 4.583 on 248 degrees of freedom
Multiple R-squared:  0.6993, Adjusted R-squared:  0.6968
F-statistic: 288.3 on 2 and 248 DF,  p-value: < 2.2e-16
```

(a) (5 points) Briefly (30 to 40 words max) summarize the information in the scatterplot matrix.

(b) (5 points) Do you think adjusted R-squared is a better metric for evaluating the fitness of multiple regression models than multiple R-Squared? Briefly explain why? In addition, why is the adjusted R-squared is smaller than multiple R-Squared?

3. (15 points) A study was conducted to develop predictive equations for lean body weight, a measure of men's health. Specifically, measurements were made on $n = 251$ men in order to relate the percentage of bodyfat determiend by underwater weighing (`bodyfat`) to abdomen circumference in cm (`abdomen`), and hip circumference in cm (`hip`). The fitted mean function is

$$\hat{\mathbb{E}}(\texttt{bodyfat}|\texttt{abdomen, hip}) = -25.93 + 0.86\ \texttt{abdomen} - 0.35\ \texttt{hip}.$$

(a) (5 points) Give the coefficients in the estimated mean function if abdomen and hip circumference had been measured in inches. You are reminded that one inch equals 2.54 cm.

(b) (5 points) Carefully interpret the estimated coefficient of `hip` in the multiple regression model (in cm).

(c) (5 points; students in 5205 only) Define the variables

$$\texttt{mean} = (\texttt{abdomen} + \texttt{hip})/2 \text{ and } \texttt{diff} = \texttt{abdomen} - \texttt{hip}$$

where `abdomen` and `hip` are both measured in cm. Give the estimated coefficients in the mean function

$$\mathbb{E}(\texttt{bodyfat}|\texttt{abdomen, hip}) = \beta_0 + \beta_1\ \texttt{mean} + \beta_2\ \texttt{diff}.$$

4. (30 points) Consider the multiple linear regression with response `body fat percentage`, predictor variables `triceps skin fold thickness` (X_1), `thigh circumference` (X_2) and `midarm circumference` (X_3). The following ANOVA table is obtained:

Source	SS	df	MS	F
Regression (SSR(X_1, X_2, X_3))	396.98	_____	_____	
SSR(X_1)	352.27	_____	_____	
SSR($X_2|X_1$)	33.17	_____	_____	
SSR($X_3|X_1, X_2$)	_____	_____	_____	_____
Error	98.41	16	_____	
Total	495.39	19		

(a) (5 points) Fill in the above ANOVA table. In the above table,for the F column, only fill in the F statistic needed in part (d).

(b) (5 points) What is the extra sum of squares by adding X_3 to the model given X_1 and X_2 are already in the model?

(c) (5 points) Briefly explain how extra sum of squares can be understood from error sum of squares and regression sum of squares perspectives?

(d) (10 points) Test if X_3 (midarm circumference) can be dropped from the model at significance level 0.01? Given $F(0.99, 1, 16) = 8.543$. Please state the null and alternative hypothesis.

(e) (5 points) What would be the proper alternative hypothesis for the test of the null hypothesis $H_0 : \beta_2 = \beta_3 = 0$? Please also find the value of the test statistic.