1. (10 pts) For a general linear regression problem with p covariates (including intercept and $p-1$ additional covariates) and sample size n, the regression model can be written as $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$, where $\mathbf{e} \sim N(0, \sigma^2 I_{n\times n})$, and $\mathbf{X}$ is full rank.

   (a) (4 pts) What are the dimensions of matrix/vector of $\mathbf{y}$, $\mathbf{X}$, $\boldsymbol{\beta}$, and $\mathbf{e}$? What is the rank of $\mathbf{X}$? Please explain why $\text{cov}(\mathbf{e}) = \sigma^2 I_{n\times n}$ implies the assumptions of independence and homogeneity.

$y_{n\times 1}$ $\quad X_{n\times p}$ $\quad e_{n\times 1}$ $\quad \text{rank}(X) = p$

$\text{cov}(e) = \sigma^2 \begin{pmatrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \end{pmatrix} \Rightarrow \begin{cases} \text{independence since off diagonals are } 0 \\ \text{homogeneity since } Var(y_i) = \sigma^2 \text{ for all } i \end{cases}$

   (b) (4 pts) Derive the least squares estimates: $\hat{\boldsymbol{\beta}} = (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}(\mathbf{X}^\mathsf{T}\mathbf{y})$ by minimizing the least squares objective function, i.e., to minimize $(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$.

See lecture note

   (c) (4pts) Calculate $E(\hat{\boldsymbol{\beta}})$ and $\text{cov}(\hat{\boldsymbol{\beta}})$.

See lecture note

2. (8pts) Now assume $\mathbf{e} \sim N(0, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma} = \text{cov}(\mathbf{e})$ and $\boldsymbol{\Sigma}$ is positive definite. Given the eigen-value decomposition of $\boldsymbol{\Sigma} = \mathbf{V}\boldsymbol{\Gamma}\mathbf{V}'$, where $\boldsymbol{\Gamma}$ is a diagonal matrix and $\mathbf{V}$ is an orthonormal matrix such as $\mathbf{V}\mathbf{V}' = \mathbf{V}'\mathbf{V} = I_{n\times n}$, we define $\boldsymbol{\Sigma}^{-1/2} = \mathbf{V}\boldsymbol{\Gamma}^{-1/2}\mathbf{V}'$. Let $\tilde{\mathbf{y}} = \boldsymbol{\Sigma}^{-1/2}\mathbf{y}$, $\tilde{\mathbf{X}} = \boldsymbol{\Sigma}^{-1/2}\mathbf{X}$, and $\tilde{\mathbf{e}} = \boldsymbol{\Sigma}^{-1/2}\mathbf{e}$. We consider a linear regression problem $\tilde{\mathbf{y}} = \tilde{\mathbf{X}}\boldsymbol{\alpha} + \tilde{\mathbf{e}}$.

by matrix multiplication for diagonal matrix

(a) (2 pts) Please show $\text{cov}(\tilde{\mathbf{e}}) = I_{n\times n}$.

$$\text{Cov}(\tilde{e}) = \Sigma^{-\frac{1}{2}} \Sigma \Sigma^{-\frac{1}{2}}$$

$$= V\Gamma^{-\frac{1}{2}}\underline{V'V}\Gamma\underline{V'V}\Gamma^{-\frac{1}{2}}V'$$

$$= V\underline{\Gamma^{-\frac{1}{2}}\Gamma\Gamma^{-\frac{1}{2}}}V' = VV' = I$$

(b) (2 pts) For a linear regression problem $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$, the formula for least squares estimates is: $\hat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}(\mathbf{X}^{\mathsf{T}}\mathbf{y})$. Please use this formula to calculate the least squares estimates of regression coefficients $\hat{\boldsymbol{\alpha}}$ for the regression model $\tilde{\mathbf{y}} = \tilde{\mathbf{X}}\boldsymbol{\alpha} + \tilde{\mathbf{e}}$, in terms of $\mathbf{X}$, $\mathbf{y}$ and $\boldsymbol{\Sigma}$.

$$\hat{\alpha} = (\tilde{X}^T\tilde{X})^{-1}\tilde{X}^T\tilde{Y}$$

$$= (X^T\Sigma^{-\frac{1}{2}}\Sigma^{-\frac{1}{2}}X)^{-1} X^T\Sigma^{-\frac{1}{2}}\Sigma^{-\frac{1}{2}}Y$$

$$= (X^T\Sigma^{-1}X)^{-1} X^T\Sigma^{-1}Y$$

(c) (4pts) Calculate $E(\hat{\boldsymbol{\alpha}})$ and $\text{cov}(\hat{\boldsymbol{\alpha}})$.

$$E(\hat{\alpha}) = (X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}X\beta = \beta$$

$$\text{Cov}(\hat{\alpha}) = (X^T\Sigma^{-1}X)^{-1}X^T\underline{\Sigma^{-1}\Sigma}\,\Sigma^{-1}X(X^T\Sigma^{-1}X)^{-1}$$

$$= (X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}X(X^T\Sigma^{-1}X)^{-1}$$

$$= (X^T\Sigma^{-1}X)^{-1}$$

3. (20pts) We are interested in data collected by the Environmental Protection Agency (EPA) at the Health Effects Research Laboratory at UNC: Chapel Hill. One hundred seventy young adult males received a battery of pulmonary function tests. Fit a model with average forced vital capacity (FVC) (in ml) as the outcome and height, weight, body mass index (BMI=$\frac{\text{weight (kg)}}{(\text{height (m)})^2}$), body surface area, age, average treadmill elevation, average treadmill speed, temperature, barometric pressure, and humidity as predictors.

   (a) (5pts) To assess for possible co-linearity in the covariates, we perform PCA on the correlation matrix of this data. As shown in the following output, the 10-th eigen-value is very small, which means a particular

Eigenvalue decomposition of the Correlation Matrix

Eigenvalues of the Correlation Matrix

	Eigenvalue	Difference	Proportion	Cumulative
1	2.98457157	0.95976513	0.2985	0.2985
2	2.02480644	0.36097943	0.2025	0.5009
3	1.66382702	0.63279205	0.1664	0.6673
4	1.03103496	0.06376972	0.1031	0.7704
5	0.96726525	0.20929379	0.0967	0.8672
6	0.75797146	0.20748316	0.0758	0.9429
7	0.55048830	0.53278371	0.0550	0.9980
8	0.01770459	0.01584173	0.0018	0.9998
9	0.00186286	0.00139531	0.0002	1.0000
10	0.00046755		0.0000	1.0000

Eigenvectors

		Prin1	Prin2	Prin3	Prin4
height	Height (cm)	0.429342	0.036906	0.384653	-.240983
weight	Weight (kg)	0.562292	-.092679	-.095943	0.026718
bmi		0.340930	-.147689	-.443854	0.239531
area	Body Surface Area (M**2)	0.566969	-.047500	0.092208	-.079656
age	Age (years)	0.084799	-.102998	-.198442	-.321636
avtrel	Average Treadmill Elevation (deg)	-.116240	-.026373	0.497176	0.182497
avtrsp	Average Speed of Treadmill (mph)	0.144346	0.094273	0.571216	0.156073
temp	Air Temperature (deg C)	0.084996	0.675223	-.123262	0.099538
barm	Barometric Pressure (mmHg)	0.070861	-.175834	-.013681	0.832373
hum	Relative Humidity %	0.089569	0.677455	-.095377	0.116735

Eigenvectors

	Prin5	Prin6	Prin7	Prin8	Prin9	Prin10
height	-.120483	-.361837	0.222180	0.018428	0.521355	0.375921
weight	-.012319	0.158716	0.071437	0.004011	-.639418	0.475397
bmi	0.092604	0.520638	-.117929	0.006137	0.557078	0.060451
area	-.061800	-.035176	0.137407	-.017784	-.093401	-.792758
age	0.856166	-.289687	-.149123	-.013509	-.001675	-.003181
avtrel	0.442067	0.455127	0.550162	0.007661	-.000498	-.001474
avtrsp	0.112098	0.166298	-.760874	0.019013	-.010800	0.001191
temp	0.096489	-.017527	0.055784	0.706187	-.007909	-.015994
barm	0.121960	-.502455	0.060375	0.005988	0.003385	-.001295
hum	0.087996	-.009950	0.046292	-.707073	0.009011	0.017050

linear combination of the covariates has small variance. Which linear combination it is? Explain why is it possible that this combination has small variance? Could this PCA captures co-linearity between intercept and other covariates? and why?

approximatly 0.4 height + 0.5 ~~weight~~ weight
— 0.8 area

No intercept effect has been remoded from correlation matrix since it remode mean values

(b) (4pts) Consider a linear regression model with all the covariates. Let $\beta = (\beta_0, \beta_1, ..., \beta_{10})^T$ be the intercept and the regression coefficients for height, weight, bmi, area, age, avtrel, avtrsp, temp, barm, and hum, respectively. Test the hypothesis: H_0: $\beta_1 = \beta_2 = 2\beta_4$ using general linear hypothesis. Please write down C and θ_0 so that the test can be written $C\beta = \theta_0$, and please write down the formula of test-statistic while denoting the data matrix for intercept and the 10 covariates by $\mathbf{X}$, and denoting the residual variance of this linear regression model by $\hat{\sigma}^2$.

$$C = \begin{pmatrix} 0 & 1 & -1 & 0 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & 0 & -2 & 0 & \cdots & 0 \end{pmatrix} \qquad \theta_0 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

$$F = \frac{(\hat{\theta} - \theta_0)\, M^{-1} (\hat{\theta} - \theta_0) / 2}{\hat{\sigma}^2}$$

$$M = C(X'X)^{-1}C'$$

(c) (7pts) After a few rounds of testing, we decide to have final model without area, temp, hum, and barm.

i. (2pts) Based on the following ANOVA table, what is the R^2? Please show your calculation and you may round those numbers to simplify the calculation.

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	49984918	8330820	20.81	<.0001
Error	163	65242013	400258		
Corrected Total	169	115226931			

Root MSE	632.65927	R-Square	
Dependent Mean	5335.43235	Adj R-Sq	0.4130
Coeff Var	11.85769		

$$R^2 \approx \frac{50}{115}$$

ii. (2pts) Based on the following t-table, if we test whether the regression coefficient for age is 0 by added last test, what is the value of F-statistic, and what are the degrees of freedom?

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	4899.55068	13356	0.37	0.7142
height	Height (cm)	1	-32.70393	75.89989	-0.43	0.6671
weight	Weight (kg)	1	119.42970	92.38958	1.29	0.1980
bmi		1	-286.38260	297.80536	-0.96	0.3377
age	Age (years)	1	27.86381	13.61020	2.05	0.0422
avtrel	Average Treadmill Elevation (deg)	1	51.50263	37.65313	1.37	0.1733
avtrsp	Average Speed of Treadmill (mph)	1	755.16631	379.56118	1.99	0.0483

$$F = (2.05)^2$$

$$df = (1, 163)$$

iii. (3pts) Based on this reduced model with 6 covariates, which characteristics are associated with the best (largest) FVC?

Shorter heavier, low bmi, older, higher Avtrel and higher avtrsp

(d) (4pts) In the diagnosis of this model, we detect a few data points as outliers based on either leverage or cook's distance. Please explain what are the difference of leverage and cook's distance.

high leverage means outlier in X

large Cook's distance means high influence on regression

4. (20pts) Consider a linear regression problem to study the association between the physical activity of 12 mice vs. environment (0 for standard environment and 1 for enriched one) and dosage of a drug (with dosage 0, 1, and 2).

observation	activity	environment	drug
1	102	0	0
2	97	0	0
3	102	0	1
4	82	0	1
5	108	0	2
6	111	0	2
7	95	1	0
8	100	1	0
9	106	1	1
10	110	1	1
11	118	1	2
12	116	1	2

(a) (4pts) First consider a linear model with two covaraites:

$$E(\texttt{activity}) = b_0 + b_1\texttt{environment} + b_2\texttt{dose}$$

If we write the above model by a matrix form: $\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e}$, what are the meanings of $\mathbf{y}$, $\mathbf{X}$ and $\mathbf{e}$, and what are their dimensions?

$y_{12\times 1}$ $X_{12\times 3}$ $e_{12\times 1}$

$\uparrow$ $\uparrow$ $\uparrow$

response covariate error

(b) (4pts) Please calculate the correlation between two variables: environment and drug. For added in-order test, would the p-values for environment and drug remain the same for two orders: environment followed by drug; and drug followed by environment?

$$\mathrm{Cor}(\underset{X}{en}, \underset{Y}{drug}) = E(XY) - E(X)E(Y)$$

$$= \frac{\sum X_i y_i}{12} - \frac{\sum X_i}{12} \frac{\sum y_i}{12} = 0$$

(c) (8pts) Given the following regression coefficient estimates and type III ANOVA table.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	92.958	4.000	23.240	2.41e-09
enviorment	7.167	4.276	1.676	0.1281
drug	7.375	2.619	2.816	0.0202

Response: activity

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
enviorment	1	154.08	154.08	2.8088	0.12806
drug	1	435.12	435.12	7.9321	0.02017
Residuals	9	493.71	54.86		

Please test the null hypothesis $H_0: b_1 = b_2 = 0$ using (1) general linear hypothesis testing and (2) comparison of the sum squares of two models. Write down your test statistic, its asymptotic distribution and the degree of freedom. You should plug in the numbers into your formula of test statistic but do not need to calculate it. If you need $(\mathbf{X}'\mathbf{X})^{-1}$. Simply use $(\mathbf{X}'\mathbf{X})^{-1}$ rather than the actual numbers.

$\theta = C b$

(1) GLH $\quad C = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \quad \theta_0 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$

$$F = \frac{(\theta - \theta_0)' M^{-1} (\theta - \theta_0) / 2}{\hat{\sigma}^2} \qquad df = (2, 9)$$

$$M = C (X'X)^{-1} C'$$

(2) Model 1 $\quad y \sim \beta_0$

Model 2 $\quad y \sim \beta_0 + \beta_1 en + \beta_2 drug$

$$F = \frac{(SS_2 - SS_1)/2}{SSE_2 / 9} = \frac{(154.08 + 435.12)/2}{493.71 / 9} \qquad df = (2, 9)$$

Model $y \sim \beta_0 + \beta_1 \text{ drug}$ $\quad R_1^2 = \frac{453}{154+453+494}$

Mode 2 $y \sim \beta_0 + \beta_1 \text{ en} + \beta_2 \text{ drug}$ $\quad R_2^2 = \frac{453+154}{154+453+494}$

(d) (4pts) What is the R^2 of a smaller model with intercept and drug? What is the R^2 of a larger model with intercept, environment, and drug? Feel free to use approximations in your calculation. Then if we double the sample size from 12 to 24, while assuming the R^2 of these two models remain the same, what would be the F-statistic to test the null hypothesis that the regression coefficient for environment is 0.

$$F = \frac{(CSS_2 - CSS_1)/1}{SSE_2/(n-p)}$$

SSY = sum squares of Y

$$= \frac{(CSS_2 - CSS_1)}{(SSY - CSS_2)(n-p)}$$

$$= \frac{R_2^2 - R_1^2}{(1 - R_2^2)/(n-p)}$$

$$\frac{F_{new}}{F_{old}} = \frac{n_{new} - p}{n_{old} - p} = \frac{24-3}{12-3} = \frac{21}{9}$$