BIOSTAT 200B HW5

Part A.

1. For the regression of logdocper on logpopdens bedp1000 hsgrad poverty unemp pcinck, provide an interpretation of each of the regression coefficients.

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	-2.98003	0.31428	-9.48	<.0001	0
logpopdens	1	0.04832	0.01573	3.07	0.0023	1.56293
bedp1000	1	0.15378	0.00854	18.01	<.0001	1.36750
hsgrad	1	0.01711	0.00332	5.16	<.0001	2.53581
poverty	1	0.04025	0.00518	7.76	<.0001	2.73125
unemp	1	-0.02622	0.00815	-3.22	0.0014	1.70162
pcinck	1	0.06515	0.00556	11.72	<.0001	2.38561

The regression coefficients of the regression model are shown above.

The coefficient of **logpopdens** is 0.04832, which means when the natural logarithm of population density increases one unit, the natural logarithm of doctors per capita times 1000 will increase by 0.04832. That is to say, for any 1% increase in the population density increase, the expected ratio of doctors per capita times 1000 will be $(1.01)^{0.04832} = 1.00048$ (0.048% increase in doctors per capita).

The coefficient of **bedp1000** is 0.15378, which means when hospital beds per capita times 1000 increases one unit, the natural logarithm of doctors per capita times 1000 will increase by 0.15378. That is to say, for a 1000 increase in hospital beds per capita, we expect to see about 16.62% increase in doctors per capita times 1000, since $e^{0.15378} = 1.1662$.

The coefficient of **hsgrad** is 0.01711, which means when the percent of adult population (25 years and older) who completed 12 or more years of school increases one unit, the natural logarithm of doctors per capita times 1000 will increase by 0.01711. That is to say, for one percent increase in adult population (25 years and older) who completed 12 or more years of school, we expect to see about 1.73% increase in doctors per capita times 1000, since $e^{0.01711} = 1.0173$.

The coefficient of **poverty** is 0.04025, which means when the percent of 1990 population with income below poverty level increases one unit, the natural logarithm of doctors per capita times 1000 will increase by 0.04025. That is to say, for one percent increase in population with income below poverty level, we expect to see about 4.11% increase in doctors per capita times 1000, since $e^{0.04025} = 1.0411$.

The coefficient of **unemp** is -0.02622, which means when the percent of 1990 labor force that was unemployed increases one unit, the natural logarithm of doctors per capita times

1000 will decrease by 0.02622. That is to say, for one percent increase in labor force that was unemployed, we expect to see about 2.59% decrease in doctors per capita times 1000, since $e^{-0.02622} = 0.9741$.

The coefficient of **pcinck** is 0.06515, which means when the per capita income divided by 1000 increases one unit, the natural logarithm of doctors per capita times 1000 will increase by 0.06515. That is to say, for a 1000 increase in the per capita income, we expect to see about 6.73% increase in doctors per capita times 1000, since $e^{0.06515} = 1.0673$.

The intercept is -2.98003, which means when all the other predictor variables is set to 0, the mean natural logarithm of doctors per capita times 1000 is -2.98003. (doctors per capita times 1000 is $e^{-2.98003} = 0.0508$)

2. For the predictor pcinck, verify that the SE of the regression coef is equal to the formula involving R^2_j given in lecture, ie, find each of the quantities in the SE and calculate the SE, verifying it is equal to the SE given in the SAS output for the regression.

The SE of the regression coefficient for the predictor **pcinck** is 0.00556.

According to the lecture,

$$Var(\hat{\beta}_j) = \frac{1}{(1-R_j^2)} \frac{\sigma^2}{\sum_{i=1}^{n}(x_{ij}-\bar{x}_j)^2} = \frac{1}{(1-R_j^2)} \frac{\sigma^2}{(n-1)s_j^2}$$

$$Var(\hat{\beta}_j) = \frac{1}{(1-0.5808)} * \frac{0.0937}{(440-1)*4.059192^2}$$

$$Var(\hat{\beta}_j) = SE^2 = \ 0.00003090114$$

$$SE = \ 0.00555887938 \approx 0.00556$$

```
proc reg data=cdi;
      model pcinck =    logpopdens bedp1000 hsgrad
poverty unemp;
run;
proc means data=cdi std;
  var pcinck;
run;
```

The MEANS Procedure

Analysis Variable : pcinck
Std Dev
4.0591920

The REG Procedure
Model: MODEL1
Dependent Variable: logdocper

Number of Observations Read	440
Number of Observations Used	440

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	95.24259	15.87376	169.42	<.0001
Error	433	40.57094	0.09370		
Corrected Total	439	135.81353			

Root MSE	0.30610	R-Square	0.7013
Dependent Mean	0.58353	Adj R-Sq	0.6971
Coeff Var	52.45639		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	-2.98003	0.31428	-9.48	<.0001	0
logpopdens	1	0.04832	0.01573	3.07	0.0023	1.56293
bedp1000	1	0.15378	0.00854	18.01	<.0001	1.36750
hsgrad	1	0.01711	0.00332	5.16	<.0001	2.53581
poverty	1	0.04025	0.00518	7.76	<.0001	2.73125
unemp	1	-0.02622	0.00815	-3.22	0.0014	1.70162
pcinck	1	0.06515	0.00556	11.72	<.0001	2.38561

We can get R_j^2 by regress other predictor variables on the predictor **pcinck**, which is 0.5808 and shown in the ANOVA table beside.
Form the ANOVA table of the original regression model, we get σ^2 equals to MSE (0.0937).
To acquire the standard deviation of **pcinck**, we can do proc means in SAS and get the result 4.0591920.

The REG Procedure
Model: MODEL1
Dependent Variable: pcinck

Number of Observations Read	440
Number of Observations Used	440

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	4201.31045	840.26209	120.27	<.0001
Error	434	3032.11001	6.98643		
Corrected Total	439	7233.42046			

Root MSE	2.64319	R-Square	0.5808
Dependent Mean	18.56148	Adj R-Sq	0.5760
Coeff Var	14.24016		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	1.08942	2.71334	0.40	0.6882
logpopdens	1	1.50256	0.11514	13.05	<.0001
bedp1000	1	0.15008	0.07337	2.05	0.0414
hsgrad	1	0.13296	0.02792	4.76	<.0001
poverty	1	-0.40379	0.04036	-10.01	<.0001
unemp	1	0.17885	0.06986	2.56	0.0108

3. In general, would you expect VIFs to increase or decrease as more predictors are added to a model? Provide a justification for your answer. Illustrate by providing some example regressions using the CDI dataset.

I expect VIF to increase as more predictors are added to a model. Since $VIF_j = \frac{1}{1 - R_i^2}$, we can expect VIF_j increase while R_j^2 is high, which is high only when the other predictors can explain most of the variance in the regression on variable X_j. The more predictors are added to a model, the more variance in the regression on variable X_j can be explained by those predictors. That is to say, when more predictors are added to a model, R_j^2 will be higher due to the higher possibility that those predictors are correlated and can explain each other, and VIF_j will increase consequently.

```
proc reg data=cdi;
      model logdocper =    logpopdens /vif;
run;
```

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	-0.77646	0.12220	-6.35	<.0001	0
logpopdens	1	0.22824	0.02013	11.34	<.0001	1.00000

```
proc reg data=cdi;
      model logdocper =    logpopdens bedp1000 pcinck /vif;
```

run;

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	-1.45360	0.09244	-15.73	<.0001	0
logpopdens	1	0.05622	0.01669	3.37	0.0008	1.48601
bedp1000	1	0.17589	0.00834	21.08	<.0001	1.10297
pcinck	1	0.05712	0.00465	12.28	<.0001	1.41043

proc reg data=cdi;

model logdocper = logpopdens bedp1000 hsgrad bagrad poverty unemp pcinck crmp1000 pop pop18 pop65 totalinc /vif;

run;

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	-2.26170	0.41397	-5.46	<.0001	0
logpopdens	1	0.00950	0.01602	0.59	0.5535	1.99481
bedp1000	1	0.15038	0.00860	17.48	<.0001	1.70840
hsgrad	1	0.00157	0.00378	0.41	0.6787	4.05862
bagrad	1	0.02540	0.00431	5.90	<.0001	6.26396
poverty	1	0.01706	0.00607	2.81	0.0052	4.61291
unemp	1	-0.01119	0.00787	-1.42	0.1559	1.95175
pcinck	1	0.04033	0.00885	4.56	<.0001	7.43660
crmp1000	1	0.00147	0.00070906	2.08	0.0384	1.79836
pop	1	3.250915E-7	2.004061E-7	1.62	0.1055	83.94670
pop18	1	0.01341	0.00511	2.62	0.0090	2.64898
pop65	1	0.02095	0.00463	4.52	<.0001	1.97266
totalinc	1	-0.00001148	0.00000961	-1.19	0.2328	88.42120

As the example we shown above, the VIF for the first model is 1 because there is only one predictor variable. When there are three predictor variables in the second model, we get those VIFs to be less than 1.5, since those variables are not highly correlated; thus, R^2 is not high and so is VIF. We put a lot of predictors variables in the third model, where we can notice that some corresponding VIFs are really high, which shows collinearity in this regression model. This is because we add too many correlated variables in this model, and those variables could be considered to be removed because they can be explained by other predictor variables.

Biostat 200B HW5

Part B. Least squares using matrix algebra

1. $I-H = I-X(X'X)^{-1}X'$

$(I-H)' = (I-X(X'X)^{-1}X')'$

$= I' - (X')'[(X'X)^{-1}]'X'$

$= I - X(X'X)^{-1}X' = I-H \Rightarrow I-H$ is symmetric

$(I-H)(I-H) = I^2 - IH - HI + H^2$

$= I - 2H + H^2$

$= I - 2H + [X(X'X)^{-1}X'][X(X'X)^{-1}X']$

$= I - 2H + X(X'X)^{-1}\underbrace{X'X}_{I}(X'X)^{-1}X'$

$= I - 2H + X(X'X)^{-1}X'$

$= I - 2H + H = I - H \Rightarrow I-H$ is idempotent

2. $I - \frac{1}{n}J = I - \begin{bmatrix} \frac{1}{n} & \cdots & \frac{1}{n} \\ \vdots & \ddots & \vdots \\ \frac{1}{n} & \cdots & \frac{1}{n} \end{bmatrix}$

$(I-\frac{1}{n}J)' = I' - \frac{1}{n}J'$

$= I - \frac{1}{n}J$, $\because I$ is symmetric and $J = \begin{bmatrix} 1 & \cdots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \cdots & 1 \end{bmatrix}$ is also symmetric

$\Rightarrow I - \frac{1}{n}J$ is symmetric

$(I-\frac{1}{n}J)(I-\frac{1}{n}J) = I^2 - \frac{1}{n}JI - \frac{1}{n}JI + \frac{1}{n^2}J^2$

$= I - \frac{2}{n}J + \frac{1}{n^2}J^2$

$= I - \frac{2}{n}J + \frac{1}{n}\cdot(\frac{1}{n}J^2)$ → $J^2 = \begin{bmatrix} n & \cdots & n \\ \vdots & \ddots & \vdots \\ n & \cdots & n \end{bmatrix}$ $\therefore \frac{1}{n}J^2 = J$

$= I - \frac{2}{n}J + \frac{1}{n}J = I - \frac{1}{n}J \Rightarrow I - \frac{1}{n}J$ is idempotent

3. (a) $E(Y) = E(X\beta + \varepsilon) = E(X\beta) + E(\varepsilon)$

$= X\beta + 0 = X\beta$ ($\because \underset{n\times 1}{X\beta}$ is a vector of constants)

$Var(Y) = Var(X\beta + \varepsilon) = Var(\varepsilon) = \sigma^2 I$

(b) The normal equation for least squares estimation of regression parameters:

$$\begin{cases} n\hat{\beta}_0 + \hat{\beta}_1 \Sigma X_i = \Sigma Y_i \\ \hat{\beta}_0 \Sigma X_i + \hat{\beta}_1 \Sigma X_i^2 = \Sigma X_i Y_i \end{cases}$$

The MLR least squares estimation of regression parameters shown by matrices:

$$X'X\hat{\beta} = X'Y$$

We can get this equation from $Q = (Y-X\beta)'(Y-X\beta)$

$$Q = (Y'-\beta'X')(Y-X\beta)$$

$$= Y'Y - \underbrace{\beta'X'Y}_{1\times 1} - \underbrace{Y'X\beta}_{1\times 1} + \beta'X'X\beta$$

($\because \beta'X'Y$ and $Y'X\beta$ are constant, $(\beta'X'Y)' = Y'X\beta = \beta'X'Y$)

$$= Y'Y - 2\beta'X'Y + \beta'X'X\beta$$ as expression for sum of squared residuals

Take partial derivatives with each β_j, set equal to 0 and solve for β_j's

$$\frac{\partial Q}{\partial \beta} = \begin{bmatrix} \frac{\partial Q}{\partial \beta_0} \\ \vdots \\ \frac{\partial Q}{\partial \beta_{p-1}} \end{bmatrix} \quad \frac{\partial Y'Y}{\partial \beta} = 0_{p\times 1}; \quad \frac{\partial \beta'X'Y}{\partial \beta} = X'Y_{p\times 1}; \quad \frac{\partial \beta'X'X\beta}{\partial \beta} = 2X'X\beta_{p\times 1}$$

$$\Rightarrow \frac{\partial Q}{\partial \beta} = -2X'Y + 2X'X\beta \xrightarrow{\text{set } 0} X'Y = X'X\beta$$

$\therefore$ if $(X'X)^{-1}$ exists: $\hat{\beta} = (X'X)^{-1}X'Y$

$$E(\hat{\beta}) = E[(X'X)^{-1}X'Y]$$

$$= E[(X'X)^{-1}X']E(Y) \quad (\because (X'X)^{-1}X' \text{ is a constant matrix})$$

$$= (X'X)^{-1}X' \cdot E(Y)$$

$$= (X'X)^{-1}X'X\beta$$

$$= \underbrace{X^{-1}(X')^{-1}X'X}_{I}\beta = \beta$$

$$Var(\hat{\beta}) = \sigma^2[(X'X)^{-1}X'Y] =$$

$$= (X'X)^{-1}X'\sigma^2(Y)[(X'X)^{-1}X']' \quad (\because (X'X)^{-1}X' \text{ is a constant matrix})$$

$$= (X'X)^{-1}X'\sigma^2 I X(X'X)^{-1}$$

$$= \sigma^2(X'X)^{-1}\underbrace{X'X(X'X)^{-1}}_{X'XX^{-1}(X')^{-1}=I}$$

$$= \sigma^2(X'X)^{-1}I = \sigma^2(X'X)^{-1}$$

(c) $\hat{Y} = X\hat{\beta} = X(X'X)^{-1}X'Y$

$$= HY$$

$$E(\hat{Y}) = E(HY) = E(X(X'X)^{-1}X'X\beta)$$
$$= E(X\beta) = X\beta \quad (\because X\beta \text{ is a vector of constants})$$
$$Var(\hat{Y}) = Var(HY)$$
$$= H\sigma^2(Y)H'$$
$$= H\sigma^2 I H' \quad (\because H' = H \text{ \& } HH = H \therefore HH' = H)$$
$$= \sigma^2 H$$

4. $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$

(a)
$$X_{2n\times 2} = \begin{bmatrix} 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 1 & 1 \\ \vdots & \vdots \\ 1 & 1 \end{bmatrix} \begin{matrix} \} n \\ \} n \end{matrix}$$

(b)
$$X'X = \underset{2\times 2}{} \begin{bmatrix} 1 & \cdots & 1 \\ 0 & \cdots & 0 \ 1 \cdots 1 \end{bmatrix}_{2\times 2n} \begin{bmatrix} 1 & 0 \\ \vdots & \vdots \\ 1 & 1 \end{bmatrix}_{2n\times 2} = \begin{bmatrix} 2n & n \\ n & n \end{bmatrix}_{2\times 2}$$

$$(X'X)^{-1} = \frac{1}{\begin{vmatrix} 2n & n \\ n & n \end{vmatrix}} \begin{bmatrix} n & -n \\ -n & 2n \end{bmatrix} = \begin{bmatrix} \frac{1}{n} & -\frac{1}{n} \\ -\frac{1}{n} & \frac{2}{n} \end{bmatrix}_{2\times 2}$$

$$X'Y = \begin{bmatrix} 1 & \cdots & 1 \\ 0 & \cdots & 0 \ 1 \cdots 1 \end{bmatrix} \begin{bmatrix} u_1 \\ \vdots \\ u_n \\ v_1 \\ \vdots \\ v_n \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} u_i + \sum_{i=1}^{n} v_i \\ \sum_{i=1}^{n} v_i \end{bmatrix}_{2\times 1}$$

(c)
$$\hat{\beta} = (X'X)^{-1}X'Y = \begin{bmatrix} \frac{1}{n} & -\frac{1}{n} \\ -\frac{1}{n} & \frac{2}{n} \end{bmatrix} \begin{bmatrix} \sum_{i=1}^{n} u_i + \sum_{i=1}^{n} v_i \\ \sum_{i=1}^{n} v_i \end{bmatrix} = \begin{bmatrix} \frac{1}{n}\sum_{i=1}^{n} u_i + \frac{1}{n}\sum_{i=1}^{n} v_i - \frac{1}{n}\sum_{i=1}^{n} v_i \\ -\frac{1}{n}\sum_{i=1}^{n} u_i - \frac{1}{n}\sum_{i=1}^{n} v_i + \frac{2}{n}\sum_{i=1}^{n} v_i \end{bmatrix} = \begin{bmatrix} \bar{u} \\ \bar{v} - \bar{u} \end{bmatrix}_{2\times 1}$$

$\hat{\beta}_0 = \bar{u}$, $\hat{\beta}_1 = \bar{v} - \bar{u}$

$\hat{\beta}_0$ means when x_i is set to be 0 (i.e. patients are females), the mean of outcome variables is $\bar{u}$; when x_i is set to be 1 (i.e. patients are males), the expected mean of outcome variables is $\bar{v} - \bar{u}$, which is shown as $\hat{\beta}_1$ here.

(d)
$$H = X(X'X)^{-1}X' = \begin{bmatrix} 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 1 & 1 \\ \vdots & \vdots \\ 1 & 1 \end{bmatrix}_{2n\times 2} \begin{bmatrix} \frac{1}{n} & -\frac{1}{n} \\ -\frac{1}{n} & \frac{2}{n} \end{bmatrix}_{2\times 2} \begin{bmatrix} 1 & \cdots & 1 \\ 0 & \cdots & 0 \ 1 \cdots 1 \end{bmatrix}_{2\times 2n} = \begin{bmatrix} \frac{1}{n} & -\frac{1}{n} \\ \vdots & \vdots \\ \frac{1}{n} & -\frac{1}{n} \\ 0 & \frac{1}{n} \\ \vdots & \vdots \\ 0 & \frac{1}{n} \end{bmatrix}_{2n\times 2} \begin{bmatrix} 1 & \cdots & 1 \\ 0 & \cdots & 0 \ 1 \cdots 1 \end{bmatrix}_{2\times 2n}$$

$$H_{2n\times 2n} = \begin{bmatrix} \frac{1}{n} & \cdots & \frac{1}{n} & 0 & \cdots & 0 \\ \vdots & & \vdots & \vdots & & \vdots \\ \frac{1}{n} & \cdots & \frac{1}{n} & 0 & \cdots & 0 \\ 0 & \cdots & 0 & \frac{1}{n} & \cdots & \frac{1}{n} \\ \vdots & & \vdots & \vdots & & \vdots \\ 0 & \cdots & 0 & \frac{1}{n} & \cdots & \frac{1}{n} \end{bmatrix}_{2n\times 2n} \quad \begin{matrix} \} n \\ \\ \\ \} n \\ \\ \end{matrix}$$

(e)

$$\hat{Y} = HY = \begin{bmatrix} \frac{1}{n}_{n\times n} & O_{n\times n} \\ O_{n\times n} & \frac{1}{n}_{n\times n} \end{bmatrix}_{2n\times 2n} \begin{bmatrix} u_1 \\ \vdots \\ u_n \\ v_1 \\ \vdots \\ v_n \end{bmatrix}_{2n\times 1} = \begin{bmatrix} \frac{1}{n}\sum_{i=1}^{n} u_i \\ \vdots \\ \frac{1}{n}\sum_{i=1}^{n} u_i \\ \frac{1}{n}\sum_{i=1}^{n} v_i \\ \vdots \\ \frac{1}{n}\sum_{i=1}^{n} v_i \end{bmatrix} = \begin{bmatrix} \bar{u} \\ \vdots \\ \bar{u} \\ \bar{v} \\ \vdots \\ \bar{v} \end{bmatrix}_{2n\times 1}$$

(f) $SSE = e'e = (Y-\hat{Y})'(Y-\hat{Y})$

$$= \underset{1\times 2n}{\begin{bmatrix} u_1-\bar{u} & \cdots & u_n-\bar{u} & v_1-\bar{v} & \cdots & v_n-\bar{v} \end{bmatrix}} \begin{bmatrix} u_1-\bar{u} \\ \vdots \\ u_n-\bar{u} \\ v_1-\bar{v} \\ \vdots \\ v_n-\bar{v} \end{bmatrix}_{2n\times 1} = \sum_{i=1}^{n}(u_i-\bar{u}) + \sum_{i=1}^{n}(v_i-\bar{v})$$

$$MSE = s^2 = \hat{\sigma}^2 = \frac{SSE}{2n-2}$$

$$= \frac{\sum_{i=1}^{n}(u_i-\bar{u}) + \sum_{i=1}^{n}(v_i-\bar{v})}{2n-2}$$

$$= \left(\frac{\sum_{i=1}^{n}(u_i-\bar{u})}{n-1} + \frac{\sum_{i=1}^{n}(v_i-\bar{v})}{n-1} \right) \times \frac{1}{2}$$

$$= \frac{1}{2}(s_u^2 + s_v^2)$$

(g) H_0: $\mu_{female} = \mu_{male}$, assume that $\sigma^2_{female} = \sigma^2_{male}$

$$\hat{\sigma}^2 = \frac{(n-1)s_u^2 + (n-1)s_v^2}{(n-1)+(n-1)} = \frac{1}{2}(s_u^2 + s_v^2)$$

(h) The estimates for variance in (f) and (g) are the same, $\frac{1}{2}(s_u^2 + s_v^2)$

This is because we have the same sample size for each groups in this case.