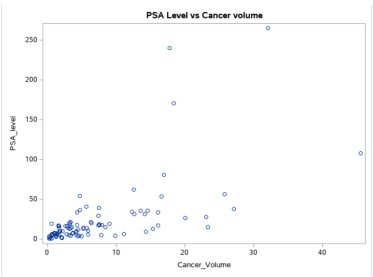
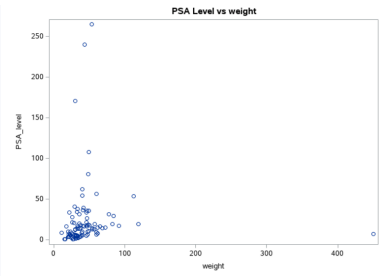
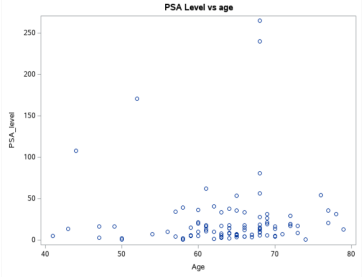
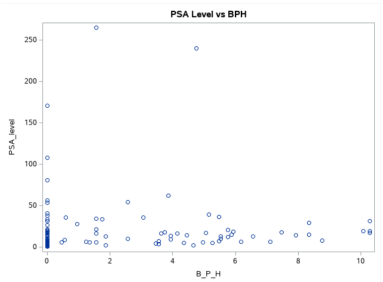
1: We can see that most of the predictors do not have a precise linear relationship. However, the predictors that may be useful with the predicting PSA level are:

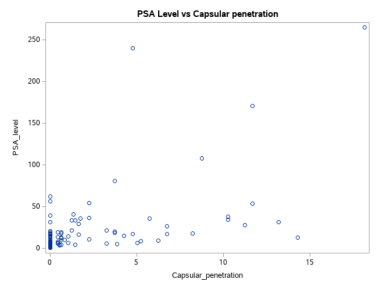
- Cancer_Volume (psa level increases as cancer volume goes up)
- Capsular penetration (psa level increases with some extent as capsular penetration)





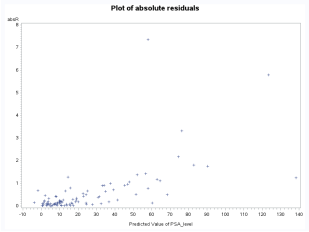






2: Does the model follow all the assumptions?

- We can see that residuals increase as our predicted value increases. This will lead to problems with constant variance assumption moving forward

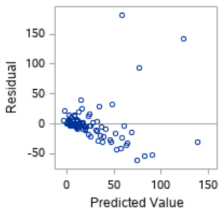


- Linearity assumption:
  - We will use the F-test and lack of fit test to see if linearity holds in the model. From the tests below we can see that our P-value < 0.05. Hence, we can conclude that at least one of the predictors has a linear relationship with PSA level. Thus, linearity holds.

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	67617	13523	13.37	<.0001
Error	91	92055	1011.58897		
Lack of Fit	91	92055	1011.58897	.	.

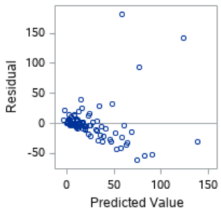
- Independence of errors:
  - We can see that errors are independent because our residual vs predicted plot does not show any signs of repetitive patterns i.e time sequence. Thus, independence of error holds.



- Constant variance assumption:
  - We will look at 2 things to identify constant variance assumption:
    - Brown-Forsythe Test:
      - H0: Homoscedasticity is present (the residuals are distributed with equal variance) vs HA: Heteroscedasticity is present
      - Since our p-value = 0.0017 < 0.05. We will reject HO and conclude that Heteroscedasticity is present.

Brown and Forsythe's Test for Homogeneity of R Variance ANOVA of Absolute Deviations from Group Medians					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Group	1	10.1971	10.1971	10.49	0.0017
Error	95	92.3231	0.9718		

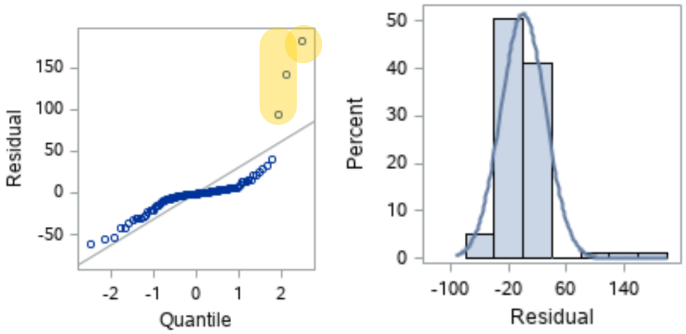
    - Residual vs Fitted Plot: We can see that our points are not centered around 0, and do not show a 'rectangular shape'. Therefore, we can conclude that our residuals do not follow constant variance assumption.



- Errors normality assumption:
  - We will look at 2 things to identify if errors follow normal distribution:
    - Kolmogorov-Smirnov test (because sample size > 50):
      - H0: Normality is present vs HA: Normality is not present
      - Since our P-value < 0.0100 indicates that we can reject H0 and conclude that errors does not follow normal distribution

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.575483	Pr < W	<0.0001
Kolmogorov-Smirnov	D	0.277087	Pr > D	<0.0100

    - Normality Q-Q plot and histogram:
      - We can see that our qq-plot does not follow a linear line, and most points are right skewed as they're above the line. The histogram is also right skewed and does not show any instances of normality. Hence, our errors are not normally distributed

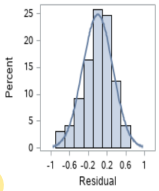


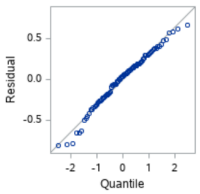
3: Applying transformation to fix the assumptions:

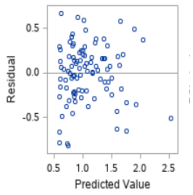
- Log10 transformation- we will apply the log10 transformation because as we can see our data is right skewed, so we have to shrink larger values.
    - After applying the transformation our residuals plot follows normal distribution (0.15 > 0.05), and we can see that constant variance assumption is met (0.67 > 0.05)

Tests for Normality

Test	Statistic		p Value	
Shapiro-Wilk	W	0.978001	Pr < W	0.1028
Kolmogorov-Smirnov	D	0.073345	Pr > D	>0.1500







Brown and Forsythe's Test for Homogeneity of R Variance
ANOVA of Absolute Deviations from Group Medians

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Group	1	0.0574	0.0574	0.18	0.6753
Error	95	30.8865	0.3251		

4: Deciding on what variables to keep in the model:

- We will use Type 3 error, and drop each predictor if its F-statistic < 1 and p-value > 0.05
    - HO: Bk = 0 vs HA: Bk != 0
    - We will set each predictor = 0, then fit the model to test if (Bk|all other predictors) is a good fit. If it is a good fit (P-value < 0.05, F-value > 1), we will keep it in the model, otherwise it will be removed from the model. Then we will repeat the same step for each predictor.

Test for Cancer Volume

The REG Procedure
Model: MODEL1

Test 1 Results for Dependent Variable PSA_level_log

Source	DF	Mean Square	F Value	Pr > F
Numerator	1	2.50842	22.56	<.0001
Denominator	89	0.11121		

Test for Age

The REG Procedure
Model: MODEL1

Test 1 Results for Dependent Variable PSA_level_log

Source	DF	Mean Square	F Value	Pr > F
Numerator	1	0.00410	0.04	0.8478
Denominator	90	0.11068		

Test for Weight

The REG Procedure
Model: MODEL1

Test 1 Results for Dependent Variable PSA_level_log

Source	DF	Mean Square	F Value	Pr > F
Numerator	1	0.06157	0.56	0.4564
Denominator	90	0.11004		

Test for BPH

The REG Procedure
Model: MODEL1

Test 1 Results for Dependent Variable PSA_level_log

Source	DF	Mean Square	F Value	Pr > F
Numerator	1	1.35846	12.40	0.0007
Denominator	91	0.10951		

Test for Capsular Penetration

The REG Procedure
Model: MODEL1

Test 1 Results for Dependent Variable PSA_level_log

Source	DF	Mean Square	F Value	Pr > F
Numerator	1	0.07399	0.68	0.4132
Denominator	91	0.10951		

- Example: We set Cancer volume = 0, then retained all other predictors in the model. (Cancervolume|age, BPH, Capsular Penetration) and got P-value of 0.0001<0.05, and F-value of 22.56 >1. In the next step, we set age=0, then retained all other predictors in model. (age|Cancervolume, BPH, Capsular Penetration) and got p-value of 0.84 and f-value of 0.04. So, age will be dropped from the model.
- Final reduced model predictors:
    - Cancer Volume, BPH

- Now, we will test to see we dropped the valid predictors. (Compare final model with initial model): Since P-value > 0.05, we will keep the reduced model.
    - Capsular penetration:

Test 1 Results for Dependent Variable PSA_level_log

Source	DF	Mean Square	F Value	Pr > F
Numerator	1	0.15020	1.14	0.2880
Denominator	91	0.13147		

, Age:

Test 2 Results for Dependent Variable PSA_level_log

Source	DF	Mean Square	F Value	Pr > F
Numerator	1	0.04952	0.38	0.5409
Denominator	91	0.13147		

Weight:

Test 3 Results for Dependent Variable PSA_level_log

Source	DF	Mean Square	F Value	Pr > F
Numerator	1	0.03566	0.27	0.6038
Denominator	91	0.13147		

- My prediction had Capsular penetration and Cancer volume as the useful predictors, however after applying type 3 sum of square test, we see that the useful predictors are Cancer volume and BPH. This shows how misleading plots can be, so it is always useful to use statistical tests to verify your findings. On the other hand, It is also useful to look at plots to identify trends in the data. So, doing both is the best choice when selecting predictors.

5: Final model analysis: (See code)

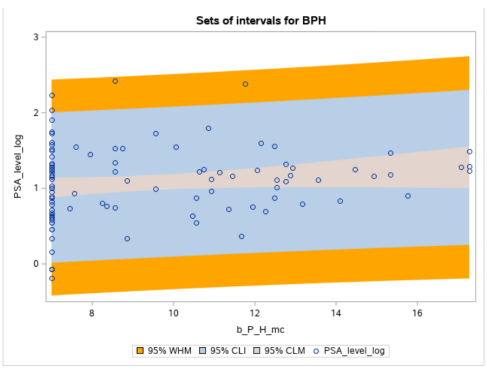
- Coefficient of multiple determination = 0.4928
- Coefficient of multiple correlation = SQRT(0.4928) = 0.70
- Coefficient of Partial correlation:
    - (PSA_level_log Cancer_Volume|BPH|)= 0.69
    - (PSA_level_log BPH| Cancer_Volume|)= 0.32
- Coefficient of Partial determination:
    - (PSA_level_log Cancer_Volume|BPH|)^2= 0.47
    - (PSA_level_log BPH| Cancer_Volume|)^2= 0.10

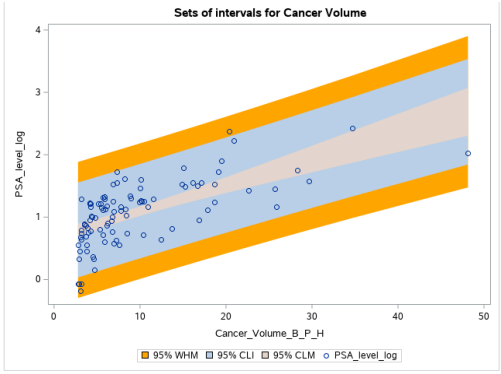
We can see that the alternative interpretation in terms of coefficient of simple determination holds by fitting appropriate models, as we get the same coefficient of Partial determination as cancer volume. (See code)

Root MSE	0.35869	**R-Square**	0.4799
Dependent Mean	-1.1446E-16	**Adj R-Sq**	0.4745
Coeff Var	-3.13384E17		

6: Interval estimates:

- We will add the mean of one predictor to the other predictor, and then plot the interval of that wrt predictor. (see code for more details).





7: 95% confidence ellipsoid:

- H0: BK = 0, k = 0,1,2 vs HA: BK != 0, (at least one B0, B1, B2 != 0)
  - , $\left\{\underset{\sim}{\beta} : (\underset{\sim}{\beta} - \underset{\sim}{b})'(X'X)(\underset{\sim}{\beta} - \underset{\sim}{b}) \leq pS^2F_{1-\alpha}\right\}$
  - (X`X) =

Model Crossproducts X'X X'Y Y'Y			
Variable	Intercept	Cancer_Volume	B_P_H
Intercept	97	678.8722	245.8683
Cancer_Volume	678.8722	10713.587545	1415.2681032
B_P_H	245.8683	1415.2681032	1505.2589539

  - $(\underset{\sim}{\beta} - \underset{\sim}{b})'(X'X)(\underset{\sim}{\beta} - \underset{\sim}{b})$ =

	A_1	A_2	A_3
1	0.66486	0.04389	0.04122

*

	B_1
1	104.421812184
2	979.911679455954
3	287.630889067206

= 124.29

  - P = 3, S^2 = 0.13003, F(0.95, 3, 94) = 2.701
    - So, $(\underset{\sim}{b} - \underset{\sim}{\beta})'(X'X)(\underset{\sim}{b} - \underset{\sim}{\beta}) / pS^2$ = 124.29 / 3(0.13003) = 318.61
    - Since, 318 > 2.70. We will conclude that at least one of the BK is not equal to 0

8: 95% confidence Bonferroni:

- We will get the Bonferroni interval for each estimator using this formula:

$$b_0 \pm t_{(1-\alpha_0/2),n-p}\ S\sqrt{(X'X)^{-1}_{00}}$$
$$b_1 \pm t_{(1-\alpha_1/2),n-p}\ S\sqrt{(X'X)^{-1}_{11}}$$
$$\vdots$$
$$b_{p-1} \pm t_{(1-\alpha_{p-1}/2),n-p}\ S\sqrt{(X'X)^{-1}_{(p-1)\times(p-1)}}$$

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

X'X Inverse, Parameter Estimates, and SSE			
Variable	Intercept	Cancer_Volume	B_P_H
Intercept	0.0281864751	-0.001344907	-0.003339463
Cancer_Volume	-0.001344907	0.0001707482	0.0000591363
B_P_H	-0.003339463	0.0000591363	0.001154203

, T(1-0.05/6, 97-3) = 2.43

- B =

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	0.66486	0.06054	10.98	<.0001
Cancer_Volume	1	0.04389	0.00471	9.31	<.0001
B_P_H	1	0.04122	0.01225	3.36	0.0011

- B0: (0.518, 0.811)
  - $0.66486 + 2.43 \cdot \sqrt{0.13003} \cdot \sqrt{0.028}$ = 0.811
  - $0.66486 - 2.43 \cdot \sqrt{0.13003} \cdot \sqrt{0.028}$ = 0.518
- Cancer Volume : (0.03, 0.05)
  - $0.04389 + 2.43 \cdot \sqrt{0.13003} \cdot \sqrt{0.0001707}$ = 0.05
  - $0.04389 - 2.43 \cdot \sqrt{0.13003} \cdot \sqrt{0.0001707}$ = 0.03
- BPH: (0.01144, 0.07)
  - $0.04122 + 2.43 \cdot \sqrt{0.13003} \cdot \sqrt{0.001154293}$ = 0.07
  - $0.04122 - 2.43 \cdot \sqrt{0.13003} \cdot \sqrt{0.001154293}$ = 0.01144
- The coefficients are significant since they're all > 0.

```
filename medical "/home/u62280666/project2/Prostate(2).dat";

DATA C;
INFILE medical;
INPUT ID PSA_level Cancer_Volume weight Age B_P_H S_V_I Capsular_penetration Gleason_score;
RUN;

/* Question 1: ScatterPlots*/
PROC SGPLOT data=C;
scatter y=PSA_level x=Cancer_Volume;
TITLE 'PSA Level vs Cancer volume';
RUN;

PROC SGPLOT data=C;
scatter y=PSA_level x=weight;
TITLE 'PSA Level vs weight';
RUN;

PROC SGPLOT data=C;
scatter y=PSA_level x=Age;
TITLE 'PSA Level vs age';
RUN;

PROC SGPLOT data=C;
scatter y=PSA_level x=B_P_H;
TITLE 'PSA Level vs BPH';
RUN;

PROC SGPLOT data=C;
scatter y=PSA_level x=Capsular_penetration;
TITLE 'PSA Level vs Capsular penetration';
RUN;

/* Question 2: */
PROC REG Data=C;
MODEL PSA_level = Cancer_Volume weight Age B_P_H  Capsular_penetration  / lackfit;
OUTPUT OUT=D RSTUDENT=R PREDICTED=P;
*PLOT RSTUDENT.*footage RSTUDENT.*land RSTUDENT.*rooms RSTUDENT.*baths;
RUN;

PROC PLOT Data=D HPERCENT=50 VPERCENT=50; /* Residual plot (for university edition) */
plot R*(Cancer_Volume weight Age B_P_H  Capsular_penetration);
RUN;

DATA D; SET D;
absR = abs(R); /* save absolute value of residuals */
RUN;

PROC GPLOT DATA = D;
PLOT absR*P; /* absolute residuals vs fitted values to check homogeneity assumption */
TITLE "Plot of absolute residuals";
RUN;

/* Checking for constant variance assumption */
/* Median = 13.30 */
PROC UNIVARIATE DATA=C;
VAR PSA_level ;
RUN;

DATA D; SET D;
Group = (PSA_level>13.33000);
RUN;

/* Brown Forsythe (or Levene) Test for homogeneity
H0: Homoscedasticity is present (the residuals are distributed with equal variance) vs HA: Heteroscedasticity is present
Since our p-value = 0.0017 > 0.05. We will accept HO and conclude that Homoscedasticity is present. */
PROC GLM Data=D;
                class Group;
                model R=Group;
                means Group / hovtest=BF; /*BF can be replaced with Levene to perform Levene Test*/
run;

/* Identifying normality assumption */ :
PROC UNIVARIATE DATA=D NORMAL PLOT;
```

```
VAR R;
RUN;

/* Question 3 */
 /* Transforming the variable to follow the normality assumption */
/* Applying the transformation */
DATA C; SET C;
PSA_level_log = log10(PSA_level);
RUN;

PROC REG Data = C;
MODEL PSA_level_log = Cancer_Volume weight Age B_P_H S_V_I Capsular_penetration Gleason_score;
OUTPUT OUT=C1 RSTUDENT=R;
RUN;

PROC UNIVARIATE DATA=C1 NORMAL PLOT; /* Check normality of Studentized residuals */
VAR R;
RUN;

/* Median = 13.30 */
PROC UNIVARIATE DATA=C;
VAR PSA_level_log ;
RUN;

DATA C1; SET C1;
Group = (PSA_level_log>1.12483);
RUN;

/* Brown Forsythe (or Levene) Test for homogeneity
H0: Homoscedasticity is present (the residuals are distributed with equal variance) vs HA: Heteroscedasticity is present
Since our p-value = 0.2267 > 0.05. We will accept HO and conclude that Homoscedasticity is present. */
PROC GLM Data=C1;
                class Group;
                model R=Group;
                means Group / hovtest=BF; /*BF can be replaced with Levene to perform Levene Test*/
run;

/* Question 4 */
/* Deciding what variables to keep in model */

/* Type 1 and Type 3 error */
PROC GLM DATA=C ;
MODEL PSA_level_log = Cancer_Volume weight Age B_P_H Capsular_penetration ;
RUN;

PROC REG DATA = C;
MODEL PSA_level_log = Cancer_Volume weight Age B_P_H Capsular_penetration;

/* F-value = 22.56, P-value < 0.05. Hence, we can conclude that cancer_Volume cannot be dropped from the model */
TEST Cancer_Volume=0;
TITLE 'Test for Cancer Volume';

PROC REG DATA = C;
MODEL PSA_level_log = Cancer_Volume weight B_P_H Capsular_penetration;
/*F-value = 0.57, P-value > 0.05. Hence, we can conclude that weight can be dropped from the model*/
TEST weight=0;
TITLE 'Test for Weight';

PROC REG DATA = C;
MODEL PSA_level_log = Cancer_Volume Age B_P_H  Capsular_penetration;

/*F-value = 0.06, P-value > 0.05. Hence, we can conclude that age can be dropped from the model*/
TEST Age=0;
TITLE 'Test for Age';

PROC REG DATA = C;
MODEL PSA_level_log = Cancer_Volume B_P_H Capsular_penetration;

/*F-value = 12.40, P-value < 0.05. Hence, we cannot conclude that bph cannot be dropped from the model*/
TEST B_P_H=0;
TITLE 'Test for BPH';

/* Comparing full with reduced model */
PROC REG DATA = C;
MODEL PSA_level_log = Cancer_Volume weight Age B_P_H Capsular_penetration;
```

```
TEST Capsular_penetration=0;
TEST Age=0;
TEST weight=0;
TITLE 'Comparing full with reduced model';

/* Question:5 */
/*partial correlation coefficient = 0.6927
 Coefficient of partial determination = 0.6927^2 = 0.47 */
proc corr data=C;
var PSA_level_log Cancer_Volume;
partial B_P_H;
run;

/*partial correlation coefficient = 0.32784
 Coefficient of partial determination = 0.32784^2 = 0.10 */
proc corr data=C;
var PSA_level_log B_P_H;
partial Cancer_Volume;
run;

/*We can see we get the same coefficient of simple determination = 0.47*/
PROC REG Data = C;
MODEL PSA_level_log = B_P_H;
OUTPUT OUT=C4 r=R1;
RUN;

/* Verifying the alternative interpretation */
PROC REG Data = C;
MODEL Cancer_Volume = B_P_H;
OUTPUT OUT=C3 r=R2;
RUN;

DATA D; MERGE C4 C3;

proc reg data = D;
model R1=R2;
run;


/* Question 6: */
/* Finding mean of B_P_H and Cancer_volume */

/* Mean = 2.53472474 */
PROC UNIVARIATE DATA=C;
VAR B_P_H ;
RUN;

/* Mean = 6.99868247 */
PROC UNIVARIATE DATA=C;
VAR Cancer_Volume ;
RUN;

/* mean of other predictor added to predcitor */
DATA C; SET C;
b_P_H_mc = B_P_H + 6.99868247;
RUN;

DATA C; SET C;
Cancer_Volume_B_P_H = Cancer_Volume + 2.53472474;
RUN;

/*B_P_H*/
PROC REG Data=C;
MODEL PSA_level_log = b_P_H_mc /clm cli;
output out=D6 predicted=pr stdi=se lclm=lwr_ci uclm=upr_ci lcl=lwr_pi ucl=upr_pi;
run;

/* Calculating simultaneous confidence bands for the entire regression line */
data D6; set D6;
WHU=pr+(sqrt(finv(0.95,3,97-3)*3)*se);
WHL=pr-(sqrt(finv(0.95,3,97-3)*3)*se);
run;

proc print data=D6(obs=6);
Title "Lower and Upper Bounds for CI, PI, and Working Hoteling ";
```

```
run;

/* Sorting the column to avoid inconsistency in plot */
proc sort data = D6;
By b_P_H_mc;
Run;

proc sgplot data=D6;

 band x=b_P_H_mc lower=WHL upper=WHU /
       fillattrs=(color=orange)
       legendlabel="95% WHM" name ="band3";

  title "Sets of intervals for BPH";

  band x=b_P_H_mc lower=lwr_pi upper=upr_pi /
       legendlabel="95% CLI" name="band1";

  band x=b_P_H_mc lower=lwr_ci upper=upr_ci /
       fillattrs=GraphConfidence2
       legendlabel="95% CLM" name="band2";


  scatter x=b_P_H_mc y=PSA_level_log;
RUN;

/* Cancer volume */
PROC REG Data=C;
MODEL PSA_level_log = Cancer_Volume_B_P_H /clm cli;
output out=D6 predicted=pr stdi=se lclm=lwr_ci uclm=upr_ci lcl=lwr_pi ucl=upr_pi;
run;

/* Sorting the column to avoid inconsistency in plot */
proc sort data = D6;
By Cancer_Volume_B_P_H;
Run;

/* Calculating simultaneous confidence bands for the entire regression line */
data D6; set D6;
WHU=pr+(sqrt(finv(0.95,3,97-3)*3)*se);
WHL=pr-(sqrt(finv(0.95,3,97-3)*3)*se);
run;

/* Plotting the Confidence Intervals */
proc sgplot data=D6;
    band x=Cancer_Volume_B_P_H lower=WHL upper=WHU /
       fillattrs=(color=orange)
       legendlabel="95% WHM" name ="band3";

  title "Sets of intervals for Cancer Volume";

  band x=Cancer_Volume_B_P_H lower=lwr_pi upper=upr_pi /
       legendlabel="95% CLI" name="band1";

  band x=Cancer_Volume_B_P_H lower=lwr_ci upper=upr_ci /
       fillattrs=GraphConfidence2
       legendlabel="95% CLM" name="band2";
  scatter x=Cancer_Volume_B_P_H y=PSA_level_log;
RUN;


/*Question 7: */
/* Getting X'X from the code, then calculating everything in the equation*/
proc reg DATA = C;
MODEL PSA_level_log = Cancer_Volume B_P_H/XPX;
run;

/* Question 8: */
/* Getting Inverse from the code, then calculating everything in the equation*/
proc reg Data = C;
MODEL PSA_level_log = Cancer_Volume B_P_H/ I;
run;
```