Y3874726

Red color: used to indicate the caption of figures and tables
Blue color: used to indicate stata code commands.
Brown color: used to indicate stata comments.

Abbreviations used
CI : Confidence interval
SD: Standard Deviation
Cook’s D: Cook’s Distance

Y3874726

Part A

1(a)

(i)

Diastolic blood pressure is a continuous variable in the given dataset, hence simple linear regression or multiple regression , depending on the number of independent variables, may be used to model it as an outcome.

From the histogram we can see that it is fairly normally distributed. Even though the tabulate command returns the categories, these are quite high in number (12) to be considered as categorical.

```
histogram dbp, start(0) width(1) norm frequency
```

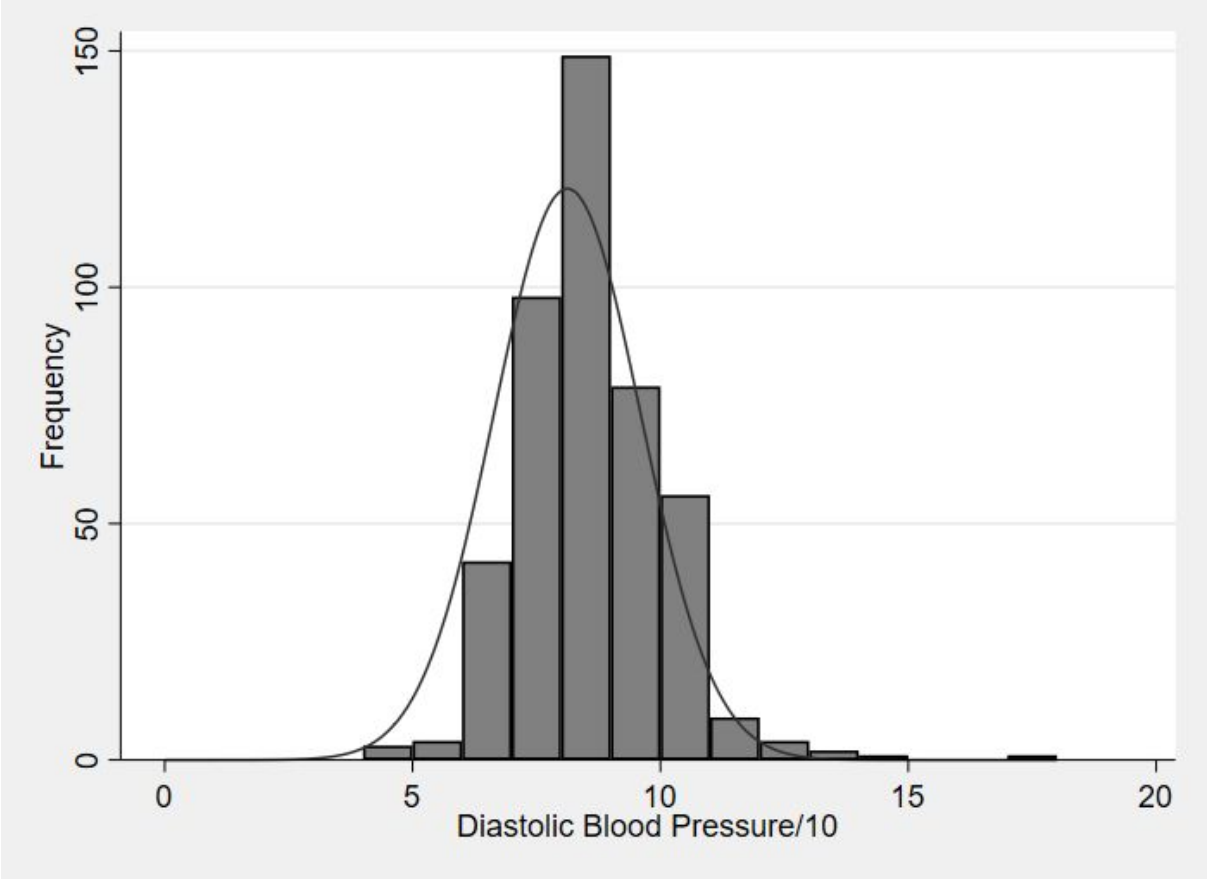

Fig. 1: Distribution of Diastolic blood pressure in the data.

```
tabulate dbp
```

```
. tabulate dbp
```

Diastolic Blood Pressure/10	Freq.	Percent	Cum.
4	3	0.67	0.67
5	4	0.89	1.56
6	42	9.38	10.94
7	98	21.88	32.81
8	149	33.26	66.07
9	79	17.63	83.71
10	56	12.50	96.21
11	9	2.01	98.21
12	4	0.89	99.11
13	2	0.45	99.55
14	1	0.22	99.78
18	1	0.22	100.00
Total	448	100.00	

Fig. 2: Snapshot showing diastolic blood pressure categories.

(ii)

Modelling strategy

DBP= Continuous variable.

Since scientific research in the field of blood pressure has shown that blood pressure may be related to weight, height, age we will use these explanatory variables for modelling DBP. Moreover, healthcare utilisation may also affect the diastolic blood pressure, so we use that also as an explanatory variable. This is because people who visit hospitals may be more likely to know about their blood pressure and may use medication/lifestyle change on discovering some problem. Furthermore, the stage of disease and treatment could also play some role in explaining the variation in DBP, hence we also include these to prevent confounding that may be caused in their absence.One may also use systolic blood pressure, after ascertaining the relationship between dbp and sbp.

Explanatory Variable Used = *wt_in*, *age*, *hospital_visits*, *treatment*, *stage*, *sbp*.

The following model is fit

```
regress dbp sbp wt_in age hospitalvisits treatment stage sbp
```

I would use a hierarchical regression model, in which the sequence of explanatory variable addition is dependent on the strength of relationship between the outcome and explanatory variable. Knowledge of this strength is obtained from prior studies. Thus, the order could be : sbp, age, wt_in, hospitalvisits, stage. It may change depending on the knowledge gathered.

1(b)

(i)

regress dbp wt_in age sbp

```
. regress dbp wt_in age sbp
```

Source	SS	df	MS	Number of obs	=	445
				F(3, 441)	=	112.24
Model	422.333606	3	140.777869	Prob > F	=	0.0000
Residual	553.113585	441	1.25422582	R-squared	=	0.4330
				Adj R-squared	=	0.4291
Total	975.447191	444	2.19695313	Root MSE	=	1.1199

dbp	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
wt_in	.01504	.0040334	3.73	0.000	.0071128	.0229671
age	-.0300252	.0074641	-4.02	0.000	-.0446949	-.0153555
sbp	.3739664	.0224917	16.63	0.000	.3297623	.4181706
_cons	3.420958	.7043912	4.86	0.000	2.036577	4.805338

Fig. 3: Snapshot showing results of regression on diastolic blood pressure by weight index, age and systolic blood pressure.

The total observations are 448, of which 445 were used to model this. F(3,441) was highly significant statistically with p<0.0001 indicating that atleast one of the explanatory variables is a statistically significant predictor of dbp. R-squared value is 43% and adjusted R-squared is 42.9%. This indicates that the fitted model is able to explain close to 43% of the variation in dbp which is quite a good figure. The model appears to be doing a satisfactory job. All three predictors are statistically significant.

1. Age (-0.03, p<0.0001)
2. Wt_in (0.015, p<0.0001)
3. Sbp (0.374, p<0.0001)
4. Constant (3.42, p<0.0001)

Thus, the obtained regression equation is:

dbp= 3.42 + (0.015 *wt_in) - (0.03 * age) + (0.374 * sbp)

For every one unit increase in weight index (wt_in) keeping other variables constant, diastolic blood pressure increases by 0.015 * 10 units = 0.15 on actual scale. This is because dbp is diastolic blood pressure divided by 10. Similarly, for every one unit increase in age (adjusting for other variables) which is years in this case, diastolic blood pressure decreases by 0.03 * 10 = 0.3 units. Keeping wt_in and age constant, one unit increase in sbp

causes 0.374 unit increase in dbp, or 3.74 unit increase in the actual diastolic blood pressure.

(ii)

10 point decrease in sbp

dbp_initial=3.42 + (0.015 *wt_in) - (0.03 * age) + (0.374 * sbp)

dbp_final=3.42 + (0.015 *wt_in) - (0.03 * age) + (0.374 * (sbp-10))

dbp_final-dbp_initial=0.374 *(sbp-10) - 0.374 * sbp = .374 (-10)

dbp_final-dbp_initial= -3.74

We can obtain estimate using lincom command

```
//_b[sbp] indicates the value of coefficient of sbp. In this case it //is 0.374.
lincom _b[sbp]*(-10)
```

```
. lincom _b[sbp]*(-10)

 ( 1)  - 10*sbp = 0
```

dbp	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
(1)	-3.739664	.2249166	-16.63	0.000	-4.181706	-3.297623

Fig. 4: Snapshot showing the influence of decrease in sbp by 10 units on dbp.

The output is -3.73 95% CI (-4.18, -3.30), p<0.0001.

From the output, we can see that dbp is estimated to decrease by 3.73 units when sbp decreased by 10 units. On the actual scale it means that when the systolic blood pressure decreases by 100 units (because sbp = systolic blood pressure divided by 10), diastolic pressure decreases by 37.3 units. This decrease can be as large as 41.8 units or as small as 33 units. We see that there exists a positive correlation between systolic blood pressure and diastolic blood pressure; as one increases the other increases and vice versa. The p-value of these estimates is highly significant which is expected because sbp is a statistically significant predictor of dbp as seen in part (i).

(iii)

Linearity assumption of dbp on sbp

To check the linearity assumption of dbp on sbp while taking into consideration other variables we need to plot the component plus residual plot. This is because our interest centers on the *partial relationship* between dbp and sbp (controlling for age, wt_in) and

not on the *marginal relationship* between dbp and sbp (ignoring age, wt_in). So, we plot component plus residual plot of the model versus sbp.

```
cprplot sbp, lowess mcolor("black") msymbol(0) msize(vsmall)
lcolor("purple") mlcolor("yellow") lwidth(vthick)
```

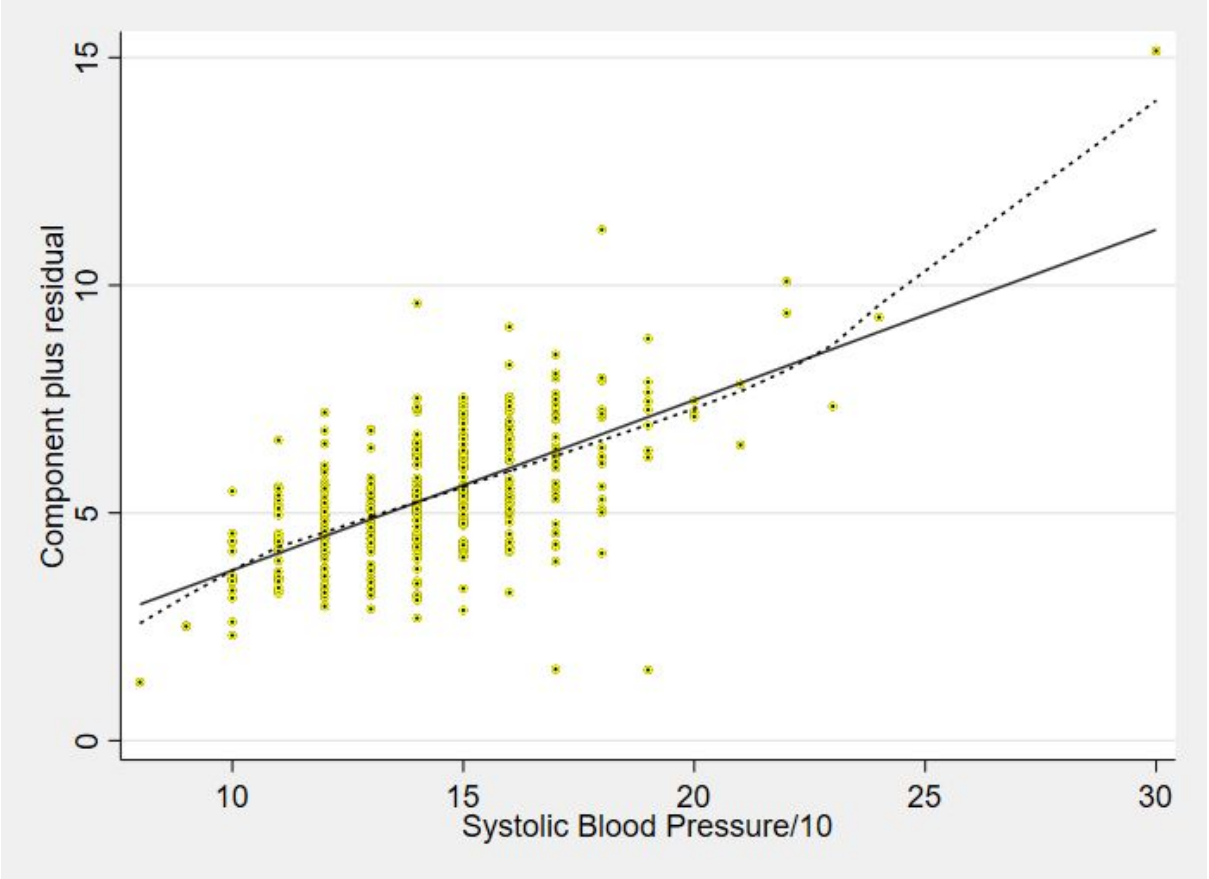

Fig. 5: Component-plus-residual plot for systolic blood pressure. The solid lines are for lowess smooths and the broken lines are for linear least-square fits.

From the plot, it is clear that the relationship between sbp and dbp is linear but it tends to abruptly increase in slope at higher values of sbp deviating from the pattern seen for the smaller values of sbp (this could be due to the outlier present at corner of the plot with sbp=30). The relationship is monotone in that dbp tends to increase with sbp, the lowess smooth also indicates a straight line, however the increase is not uniform for all values of sbp. No departure from linearity is observed.

(iv)

Diagnostic test.

We will test for Outliers, influential points and points with high leverage.

Assessing Leverage: Hat Values

```
// Predict the leverage values for the fitted model
```

```
predict dbp_hat, leverage
// Calculating the average leverage for the fitted regression.
//average hat = (n+1)/k, where n is the number of regressors
without constant and k is the number of points used to fit the
model.
scalar h_avg=(3+1)/(445)
//Computing 2 times average hat value.
scalar h2=2*h_avg
//Computing thrice the average hat value
scalar h3=3*h_avg
//label those points which have hat-values more than twice the
average hat values
gen lb=ID if dbp_hat>=2*h_avg
// Plotting leverage for each point in the dataset
scatter dbp_hat ID, msymbol(o) mcolor("black") mlabel(lb) ||,
yline( `=h3'   , lwidth(thin) lcolor("red")  ) yline( `=h2'   ,
lwidth(thin) lcolor("green")  )
```

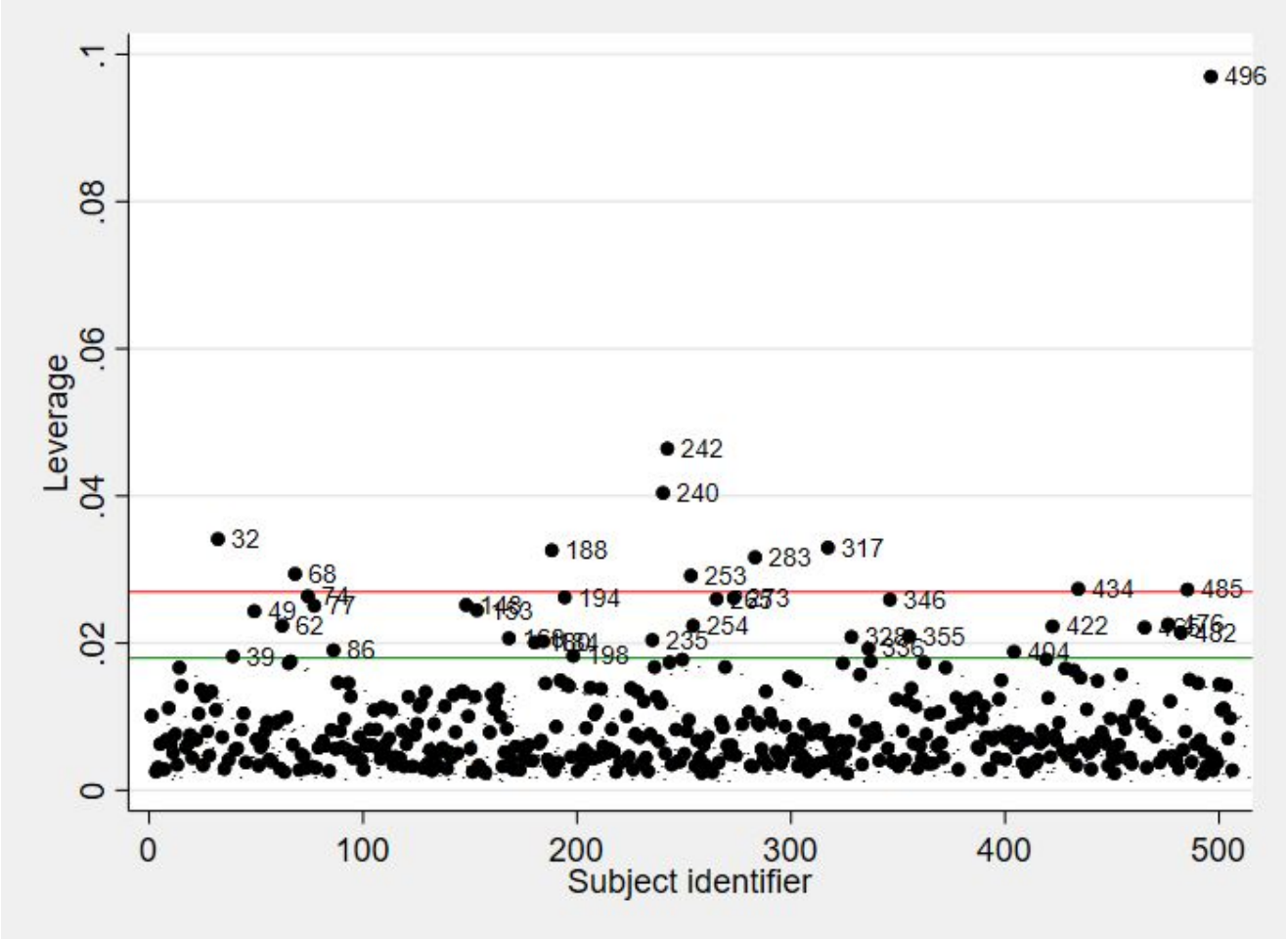

Fig. 6: An Index plot for hat values in the data set for regression of diastolic blood pressure on age, weight index and systolic blood pressure. The red horizontal line indicates 3 * average hat value and the green line indicates 2* average hat value.

From the plot, we can see that data point 496 shows the highest leverage (leverage =0.097, average hat = 0.0089. This person has the following features (age = 76, wt_in=144, sbp=30, dbp=8). The sbp of this person (30) stands out from the rest (mean sbp = 14.30). It may be playing some role in giving it high leverage. Other 8 data points, as labelled in the graph exceed thrice the average hat value (above the red line). Thus, these are the points with highest leverage. It is not clear whether they influence the model or not. This will become clear as the diagnostic proceeds. But one thing is clear that observations with high leverage values have an unusual combination of explanatory variables.

Detecting Outliers: studentized residuals

We use studentized residuals to see if there are any outliers with high residuals.

```
//obtain the studentized residuals
predict rstud, rstudent
//standard deviation of the studentized residuals computed using
//summarize(rstud) and r(sd)
scalar sd_rstud=1.0082317
//upper limit of studentized residuals = 2 * standard deviation of
rstud
scalar sd2=2*sd_rstud
//lower limit of studentized residuals =-2 * standard deviation of
rstud
scalar sd3=-2*sd_rstud
//label points which lie beyond 95% limits for studentized
residuals
gen lb2=ID if rstud>=sd2 | rstud <=sd3
//plot the studentized residuals labeled by ID
scatter rstud ID, mcolor("black") msize(small) mlabel(lb2) ||,
yline(`=sd2', lwidth(thin) lcolor("red")) yline(`=sd3',
lwidth(thin) lcolor("red"))
//get absolute value of studentized residuals
gen abs_rstud=abs(rstud)
//sort descending order the absolute value of studentized
residuals
gsort -abs_rstud
```

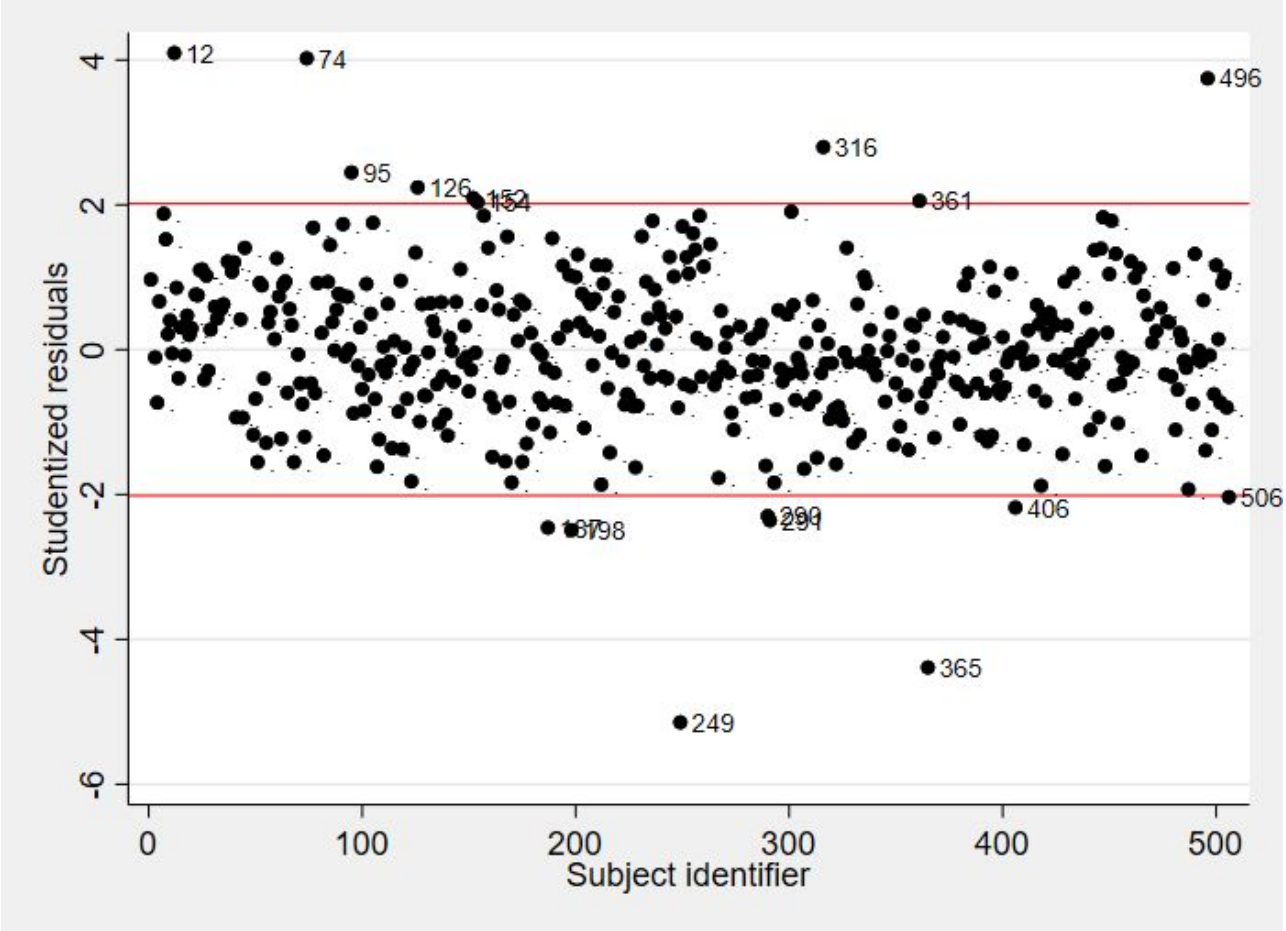

Fig. 7: Distribution of studentized residuals for regression of diastolic blood pressure on age, weight index and systolic blood pressure. Red horizontal lines indicate the +2SD and -2SD values of the studentized residuals.

From the plot of studentized residuals versus subject identifiers we can see that the labelled points exceed twice the standard deviation of the studentized residuals, indicating the possibility of outliers. Point 496 again stands out (studentized residual = 3.744, mean studentized residual = 0.0002) and point 249, 365, 2, 74 exceed way beyond the 95% CI interval expected of the studentized residuals.

For point 496, E(studentized)=3.744. Here n-k-2 = 445-3-2=440. And the associated Bonferroni p-value is = 2* 445* Pr(t(440) >3.744) = 0.09, showing that it is slightly unusual to observe a studentized residual this big in the sample. Nevertheless the p-value is not significant.

Furthermore, we have 4 points which exceed in their values of studentized residuals.

1. ID = 249, rstud=5.145. Associated Bonferroni p-value is = 2* 445* Pr(t(440) >5.145) = 0.00017

2. ID= 365, rstud = 4.387. Associated Bonferroni p-value is = 2* 445* Pr(t(440) >4.387) = 0.0064
3. ID=12 , rstud=4.09. Associated Bonferroni p-value is = 2* 445* Pr(t(440) >4.09) = 0.022
4. ID=75, rstud=4.02. Associated Bonferroni p-value is = 2* 445* Pr(t(440) >4.02) =0.030.

All these 4 points show that it is unusual to see these studentized residuals and thus, these points qualify as the outliers in the given dataset.

Measuring Influence

Observations that combine high leverage with a large studentized residual exert substantial influence on the regression coefficients. Cook's distance provides this estimate of influence.

```
//Computing the cook's distance for every observation
predict cook, cooksd
//Ploting studentized residuals versus leverage values
//Observation points weight provided by their cook's distance.
scatter rstud dbp_hat [w=cook], msymbol(circle_hollow)  || scatter rstud dbp_hat, msymbol(i) mlabel(lb) ||, legend(off) yline(2 -2) xline(`=h2' `=h3', lwidth(vthin) lcolor("red"))
```

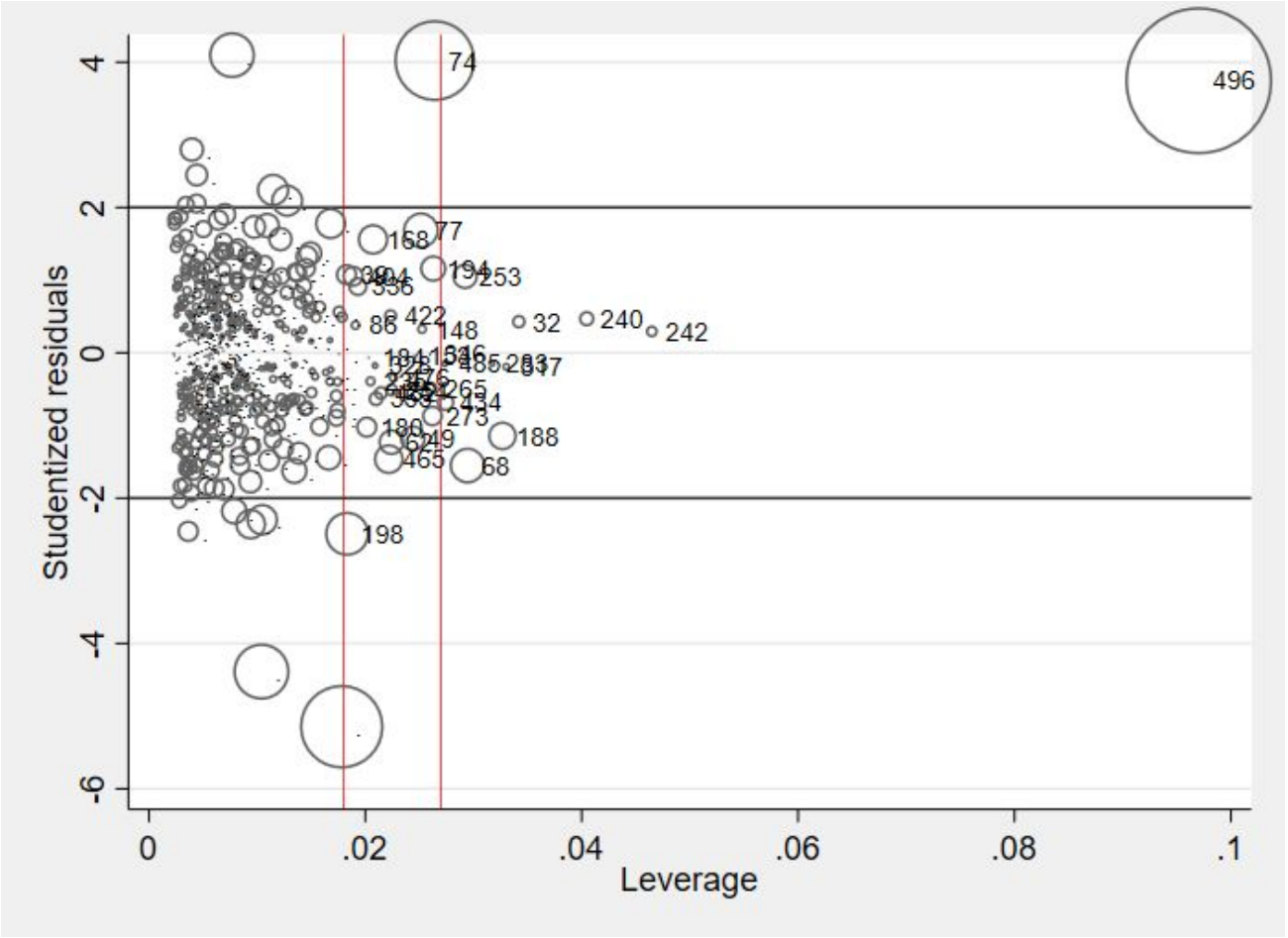

Fig. 8: Bubble plot for Cook's Distance, studentized residuals and hat values for the regression of diastolic blood pressure on age, weight index, age and systolic blood pressure. Horizontal black lines indicate the +2SD and -2SD for studentized residuals. Vertical red lines indicate the 2*average hat and 3*average hat value obtained in the data. Observation 496 has high leverage and high studentized residual. Observation 74 has high studentized residual but relatively low leverage. Observation 242 has high leverage but relatively low studentized residual. Observation 496 exerts considerable influence on the model estimates.

The plot shows that point 496 combines high leverage with high studentized residuals, thus gives the highest impact on the model coefficients. The rest of the points either have high leverage or high studentized residuals such as 74 (high studentized residual but not gross deviation from leverage) and 240, 242 which have high leverage but studentized residuals within the 95% limit of studentized residuals.
We look at *dfbetas* to further understand the influence of points on the model coefficients.

```
//Computing dfbeta for age predictor
dfbeta age
//Computing dfbeta for wt_in predictor
dfbeta wt_in
//Computing dfbeta for sbp predictor
dfbeta sbp
//plotting the scatter plots and saving them to combine later
scatter _dfbeta_1  ID, mlabel(lb) saving(a)
scatter _dfbeta_2  ID, mlabel(lb) saving(b)
scatter _dfbeta_3  ID, mlabel(lb) saving(c)
gr combine a.gph b.gph c.gph, ycommon
```

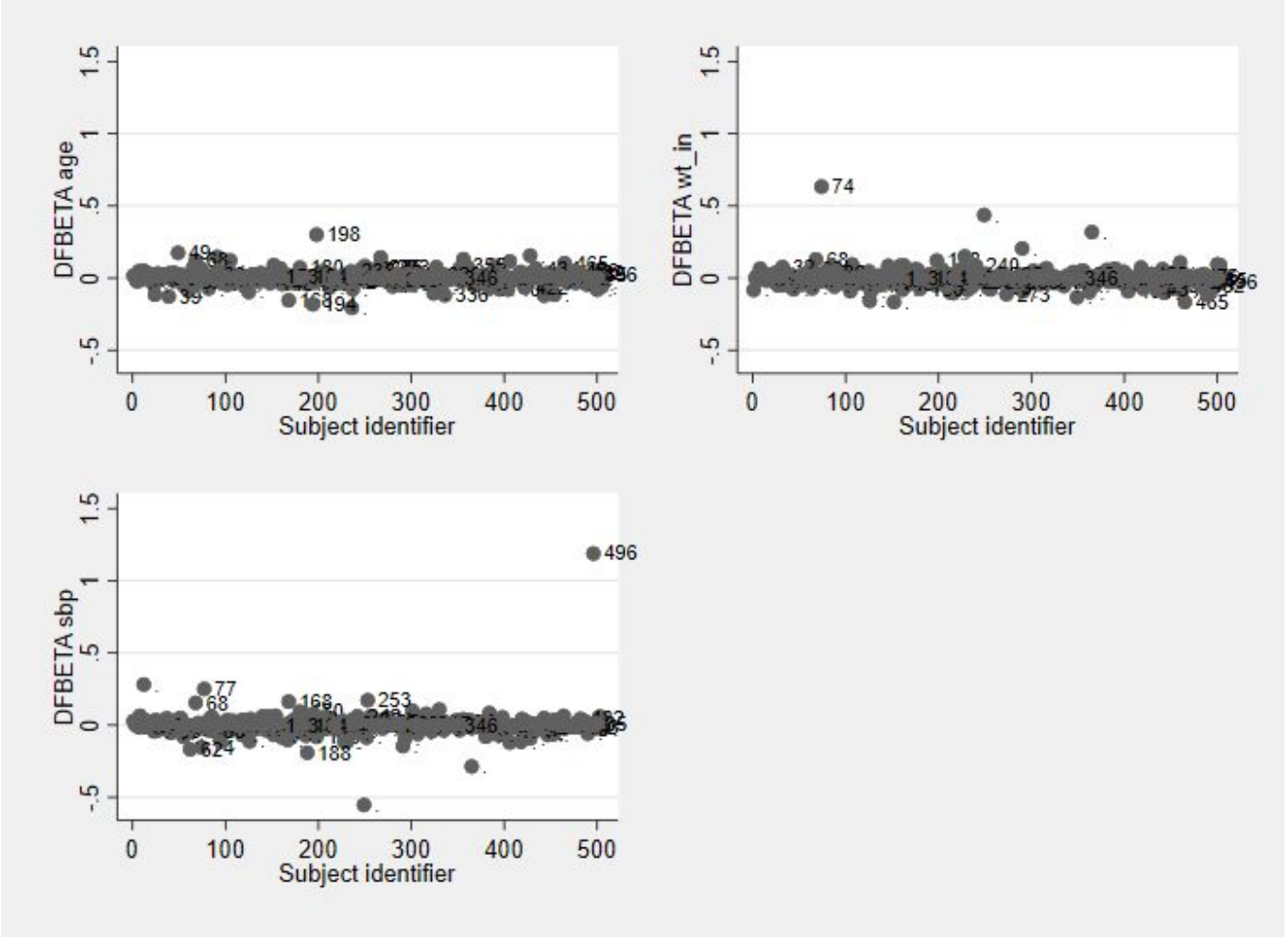

Fig. 9: DFBETAS plot for the regression model on diastolic blood pressure, age, weight index and systolic blood pressure. From figure at the bottom, observation 496 considerably influences the beta coefficient for systolic blood pressure.

From the plot of dfbetas we can see clearly that point 496 influences the coefficient of sbp (dfbeta >1). This we had noted before that its sbp was quite unusual and not in line with the other values (sbp=30). Either it could be a plausible value, that

the person has high systolic blood pressure and is quite hypertensive because his/her sbp exceeds 180 (18 on sbp scale).

(v)

```
Bootstrap regression
bootstrap, bca reps(150) : regress dbp wt_in age sbp
```

```
Linear regression                               Number of obs     =        445
                                                Replications      =        150
                                                Wald chi2(3)      =     178.88
                                                Prob > chi2       =     0.0000
                                                R-squared         =     0.4330
                                                Adj R-squared     =     0.4291
                                                Root MSE          =     1.1199
```

dbp	Observed Coef.	Bootstrap Std. Err.	z	P>|z|	Normal-based [95% Conf.	Interval]
wt_in	.01504	.0047893	3.14	0.002	.005653	.0244269
age	-.0300252	.0065856	-4.56	0.000	-.0429327	-.0171176
sbp	.3739664	.0334714	11.17	0.000	.3083638	.4395691
_cons	3.420958	.7880113	4.34	0.000	1.876484	4.965431

Fig. 10: Snapshot showing results of bootstrap regression of diastolic regression on age, weight index and systolic blood pressure without bias correction.

The bootstrap resulting regression is fairly similar to that obtained when the normality of residual assumption is taken, except that the pvalue for age is not less than 0.001, it is equal to 0.002. The coefficient values are however, similar. The r-squared and adjusted r-squared values are close to each other , 43%. Thus, this is similar to what we had obtained earlier and may be interpreted in the similar manner (Section 1(b) (i))

```
estat bootstrap, bca
```

```
. estat bootstrap, bca

Linear regression                               Number of obs     =        445
                                                Replications      =        150

------------------------------------------------------------------------------
             |   Observed               Bootstrap
         dbp |       Coef.       Bias    Std. Err.  [95% Conf. Interval]
-------------+----------------------------------------------------------------
        wt_in |   .01503998  -.0005415   .00478934     .007796    .0271383 (BCa)
          age |  -.03002517  -.0005735    .0065856    -.042433   -.0153459 (BCa)
          sbp |   .37396641   .0045549   .03347135    .3181163     .436284 (BCa)
        _cons |   3.4209576   .0303538   .78801133    1.788736    4.783499 (BCa)
------------------------------------------------------------------------------
(BCa)  bias-corrected and accelerated confidence interval
```

Fig. 11: Snapshot showing results of bootstrap regression of diastolic regression on age, weight index and systolic blood pressure with bias correction.

For the bias corrected and accelerated bootstrap regression confidence intervals are slightly different from those obtained using the multiple linear regression in that a right shift in the lower limit of CI is observed for each variable but this does not grossly affect the coefficient estimates.
The bootstrap regression fairly resembles the results of the regression obtained before.

(vi)

Adding stage to the model

```
regress dbp wt_in age sbp stage
```

```
. regress dbp wt_in age sbp stage
```

Source	SS	df	MS	Number of obs	=	445
				F(4, 440)	=	84.24
Model	423.036337	4	105.759084	Prob > F	=	0.0000
Residual	552.410854	440	1.25547921	R-squared	=	0.4337
				Adj R-squared	=	0.4285
Total	975.447191	444	2.19695313	Root MSE	=	1.1205

dbp	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
wt_in	.0147319	.0040564	3.63	0.000	.0067596	.0227043
age	-.0304623	.0074907	-4.07	0.000	-.0451843	-.0157403
sbp	.3738253	.0225037	16.61	0.000	.3295972	.4180534
stage	-.0811799	.1085072	-0.75	0.455	-.2944367	.132077
_cons	3.762411	.8396188	4.48	0.000	2.112249	5.412572

Fig. 12: Snapshot showing results of regression of diastolic regression on age, weight index, systolic blood pressure and stage.

The p-value for stage in the obtained model is 0.455 indicating that it is not a statistically significant predictor of dbp and does not help in explaining the variation in dbp in the given dataset. The R-squared value is 43.37% and the adjusted r-squared is 42.85%. In the model without stage these values were 43.3% and 42.9% respectively. Thus, no major change in R-squared is observed, in fact the model has been penalized by a reduced R-squared value.

Model 1

```
regress dbp wt_in age sbp
estat ic
```

```
. estat ic
```

Akaike's information criterion and Bayesian information criterion

Model	N	ll(null)	ll(model)	df	AIC	BIC
.	445	-806.0505	-679.819	4	1367.638	1384.03

Note: BIC uses N = number of observations. See [R] BIC note.

Fig. 13: Snapshot showing AIC for regression of diastolic blood pressure on weight index, age, systolic blood pressure.

Model 2

```
regress dbp wt_in age sbp stage
estat ic
```

```
. estat ic

Akaike's information criterion and Bayesian information criterion
```

Model	N	ll(null)	ll(model)	df	AIC	BIC
.	445	-806.0505	-679.5361	5	1369.072	1389.563

Note: BIC uses N = number of observations. See [R] BIC note.

Fig. 14: Snapshot showing AIC for regression of diastolic blood pressure on weight index, age, systolic blood pressure and stage.

Looking at the AIC of model without stage, the value of AIC is 1367.64 whereas for Model 2 it is 1369. Thus, the second model has been penalized because the addition of stage does not contribute much to the model fit. The BIC value for the Model 2 is also higher than that for Model 1 (1389.56 vs 1384.03). Thus, by both R-squared criteria and the Information Criteria (AIC and BIC), addition of the variable stage does not improve the model fit. Hence, we conclude that it is better not to add to the complexity of the model by adding a variable which contributes nothing substantial to the model fit.

1(c)

(i)

tabulate stage treatment, summarize(hospitalvisits)

```
tabulate stage treatment, summarize(hospitalvisits)
```

Means, Standard Deviations and Frequencies
of Health Care utilisation durimg the last month

	treatment arm				
Stage	placebo	0.2 mg es	1.0 mg es	5.0 mg es	Total
3	.359375	1.1290323	.35820896	.89285714	.67068273
	.87952346	2.4990744	.9801282	2.1965144	1.7813958
	64	62	67	56	249
4	2.0222222	.90697674	1.122449	1.2272727	1.320442
	3.1872537	1.5401601	2.1275245	2.8273057	2.515965
	45	43	49	44	181
Total	1.0458716	1.0380952	.68103448	1.04	.94418605
	2.2948627	2.1524754	1.607586	2.4860012	2.1432718
	109	105	116	100	430

Fig. 15: Snapshot showing cross-tabulated summary of hospital visits by stage and treatment type in the dataset.

1. From the table we can see that out of 448 data points, data for this classification is only available for 430 respondents.
2. Patients belong to stage 3 and 4 and for each treatment group, a higher number of people belong to stage 3 than stage 4.
3. A higher proportion of people are found in Stage 3 (249, 57.9%) as compared to that in Stage 4.
4. Health care utilization during the last month is maximum among Stage 4 patients undergoing placebo treatment (2.02 visits) and minimum is for those in stage 3 undergoing placebo treatment (0.34 visits). These two groups also show the highest and least variation in health care utilization respectively (3.18 visits in Stage 4 placebo vs. 0.88 visits in Stage 3 placebo)
5. People in stage 4 taking 1 or 5 mg of estrogen report higher health care utilization than those in stage 3 (1.12 visits vs .36 visits in 1 mg estrogen and 1.32 visits vs .67 visits in 5 mg estrogen) whereas for those taking 0.2 mg estrogen, stage 3 report higher health care utilization.
6. Average health care utilization during the past month is higher among patients in stage 4 (1.32 visits vs. 0.67 in Stage 3. This group's health care utilization is also more variable than that for Stage 3 (2.5 visits vs. 1.78 visits way from mean number of visits). Thus, they vary more in their health care utilization behaviour.
7. In the treatment arm, a higher proportion of people are taking 1 mg of estrogen (116) and least for 5mg estrogen (100). A fairly equal distribution of treatments is seen.
8. Overall it is clear that patients at higher stages of cancer are using more medicine and report higher health care utilization.

(ii)

Model Representation for basic understanding.
hospitalvisits= f(age, wt_in)

Hospital visits is a continuous variable with discrete non-negative integer values in the data as can be seen from its distribution using histogram and summary statistics using tabulate command in stata. The distribution is highly right skewed and the hospital visits actually tell us the count of the times a patient has visited hospital in the last month (the time period/ time

frame involved). Thus, it appears as a count data, moreover, the *qqplot* for hospital visits does not indicate normality at all. In this case where we have a non-normal and a highly right skewed distribution with count data, we would use Poisson Regression.

```
histogram hospitalvisits, start(0) width(0.5) norm frequency
```

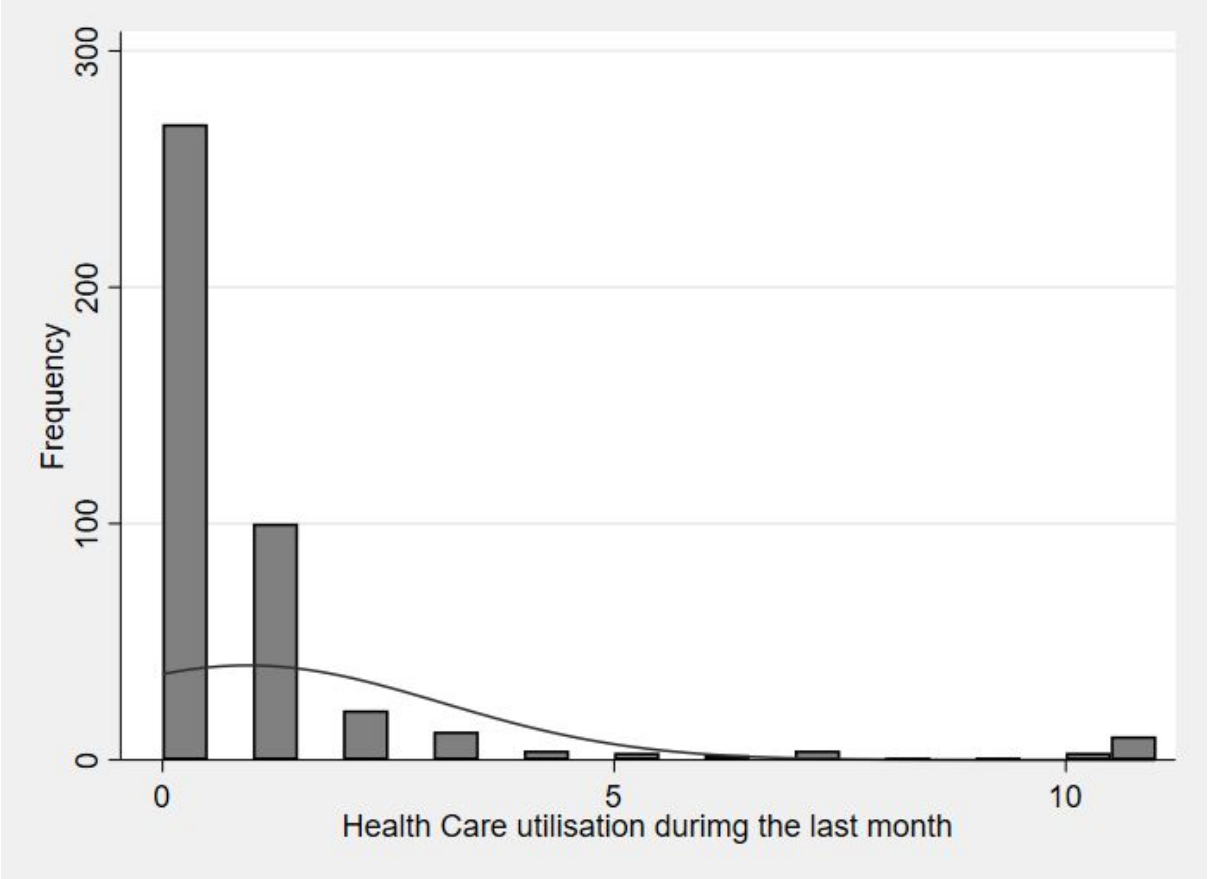

Fig. 16: Distribution of health care utilization during the last month in the given dataset.

```
qnorm hospitalvisits
```

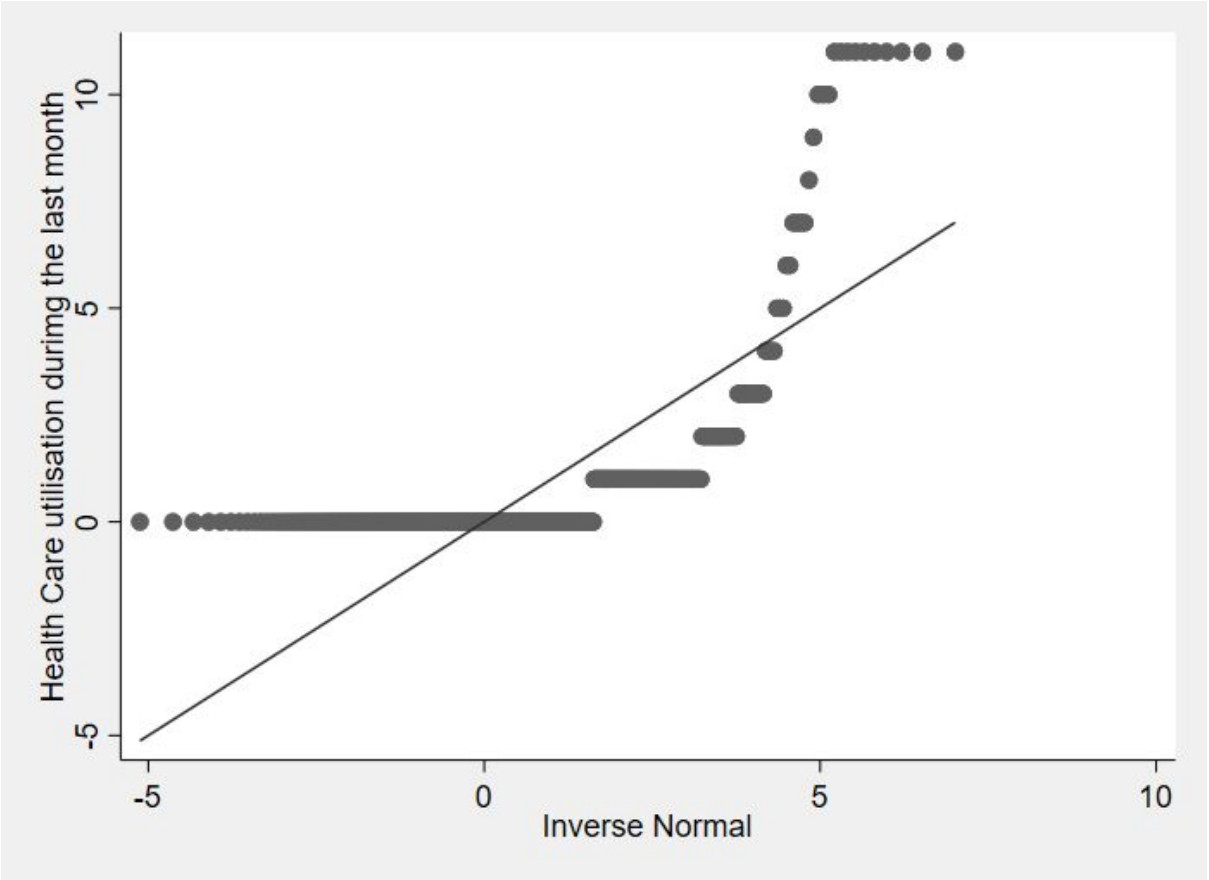

Fig. 17: QQ plot for the health care utilization during the last month in the given dataset.

One problem is the high number of zeroes in the data for hospital visits (62.56%), thus even though the Poisson distribution is the suitable method. Log(0) does not exist, so in this case, we might need to use the *Zero-inflated Poisson Regression Method*. This is not the part of the course so I will write the usual poisson commands to achieve this result.

First I will fit the constant only model

```
//Null model
poisson  hospitalvisits
```

Then I will add first variable age

Model 1

```
//Model with age
poisson  hospitalvisits age
```

I am following stepwise regression modelling. After obtaining this model, I will test the fit of this model as compared to the null or the baseline model using residual deviance statistics which will be provided by the results of these models.

Deviance = -2 * loglikelihood(null model) - (- 2 * loglikelihood (model 1))

I will note down the reduction in the deviance and see whether the model with age is better than the null model.
Then I will go forward and add the second variable wt_in.
Model 2

```
//Model with age and wt_in
poisson  hospitalvisits age wt_in
```

After executing this model, I will look at the deviance of this model and compare it with that of the Model 2 to see how good the fit of the model is using the residual deviance of the models.
I can also use goodness of fit tests to compare the baseline, model 1 and model 2 using estat gof and lrtest command after each model fitting.

After model fit, I will analyse the coefficients both on log scale and actual scale using lincom command.
Residual analysis will be performed to check for outliers or unusual observation using the following commands. Thus, the commands are as below:

```
poisson  hospitalvisits
poisson  hospitalvisits age
poisson  hospitalvisits age wt_in
predict hospitalvisit_fit
//Deviance Residual
Glm hospitalvisits age wt_in, family(poisson)
predict dev_glm, deviance
predict std_dev_glm, deviance standardized
Scatter std_dev_glm hospitalvisit_fit
//Pearson Residual
Predict pr_glm, pearson
predict std_pr_glm, pearson standardized
scatter std_pr_glm hospitalvisit_fit
```

The commands in deviance residual and pearson residual will help me in gauging the nature of residuals and find out whether any outlier exists and whether or not the assumption of linear residuals for deviance residuals is violated.

After this I will also run negative binomial poisson regression but I am not sure if such a process exists for Zero-inflated poisson regression. This will be important too if the assumption of equality of mean and variance is violated in the dataset.

Y3874726

Y3874726

PART B

2(a)

Outcome = death. 0 = alive, 1 = dead

(i)

a) Treatment is a categorical variable in the dataset with 4 categories (Placebo, 0.2 mg estrogen, 1 mg estrogen and 5mg estrogen)

b) Wt_in is a continuous variable.

Close to 67% of the respondents in the data are dead.
Since death is a binary variable with values 0 and 1, we use logistic regression to model it as a function of treatment and wt_in.

```
xi: logit death_all i.treatment wt_in
```

```
. xi: logit death_all i.treatment wt_in, or
i.treatment       _Itreatment_1-4     (naturally coded; _Itreatment_1 omitted)

Iteration 0:   log likelihood = -282.00094
Iteration 1:   log likelihood = -272.63274
Iteration 2:   log likelihood =   -272.583
Iteration 3:   log likelihood =   -272.583

Logistic regression                             Number of obs     =        446
                                                LR chi2(4)        =      18.84
                                                Prob > chi2       =     0.0008
Log likelihood =   -272.583                     Pseudo R2         =     0.0334

------------------------------------------------------------------------------
   death_all | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
_Itreatment_2 |   .9860024   .3028213    -0.05   0.963     .5400782    1.800111
_Itreatment_3 |   .4216097   .1199166    -3.04   0.002     .2414388     .736231
_Itreatment_4 |   .8275075   .2509995    -0.62   0.532     .4566523    1.499541
        wt_in |   .9806457   .0074913    -2.56   0.011     .9660724    .9954388
        _cons |   19.54692    15.8093     3.68   0.000     4.005322     95.3936
------------------------------------------------------------------------------
Note: _cons estimates baseline odds.
```

Fig. 18: Snapshot showing the fitted logistic model coefficients for regression of death status on treatment type, weight index.

```
//Command used to fit the logistic model
xi: logit death_all i.treatment wt_in, or
//Test the overall significance of treatment in the logistic model
test _Itreatment_2 _Itreatment_3 _Itreatment_4
```

```
. xi: logit death_all i.treatment wt_in, or
i.treatment       _Itreatment_1-4     (naturally coded; _Itreatment_1 omitted)

Iteration 0:   log likelihood = -282.00094
Iteration 1:   log likelihood = -272.63274
Iteration 2:   log likelihood =   -272.583
Iteration 3:   log likelihood =   -272.583

Logistic regression                               Number of obs     =        446
                                                  LR chi2(4)        =      18.84
                                                  Prob > chi2       =     0.0008
Log likelihood =   -272.583                       Pseudo R2         =     0.0334

------------------------------------------------------------------------------
   death_all | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
_Itreatment_2 |   .9860024   .3028213    -0.05   0.963     .5400782    1.800111
_Itreatment_3 |   .4216097   .1199166    -3.04   0.002     .2414388     .736231
_Itreatment_4 |   .8275075   .2509995    -0.62   0.532     .4566523    1.499541
        wt_in |   .9806457   .0074913    -2.56   0.011     .9660724    .9954388
        _cons |   19.54692    15.8093     3.68   0.000     4.005322    95.3936
------------------------------------------------------------------------------
Note: _cons estimates baseline odds.
```

Fig. 19: Snapshot showing the fitted logistic model Odds Ratio for regression of death status on treatment type, weight index.

Both models one for coefficient and the other providing odds ratio is run.

The obtained equation is

logit(p)=2.97 -0.0195 * wt_in - (.0140965) * (Treatment=2)
-(.8636754) * (Treatment=3)-(.1893371) * (Treatment=4)

The odds ratio are

Variable	**Odds Ratio**[1]	**95%CI**	**P value**
Treatment			
Placebo[2]	1		p=0.0043
0.2 mg estrogen	0.986	0.54, 1.80	p=0.963
1 mg estrogen	0.42	0.24, 0.74	p=0.002
5 mg estrogen	0.83	0.46, 1.50	p=0.532
Weight Index	0.98	0.966, 0.995	p=0.011
Constant	19.54	4, 95.39	p<0.0001

1 indicates adjusted for the other variables in the table.
2 indicates the reference category

Table 1: Adjusted odds-ratio for the effects of treatment type and weight index on death status in the dataset.

From the table above we can see that, adjusted for weight index, the treatment with 0.2mg of estrogen (OR=0.986 95%CI(0.54, 1.80), p=0.963) and 5 mg of estrogen (OR=0.83 95%CI(0.46, 1.50), p=0.532) are not statistically significant and have very little impact on the death/alive status of the respondent as compared to the placebo treatment. Whereas , compared to the placebo group, those who receive 1 mg of estrogen have lower odds of death (OR=0.42, 95%CI (0.24, 0.74),p=0.002). Thus, this indicates that 1 mg of estrogen is reduced with lower odds of diabetes. Adjusting for treatment, a unit increase in weight index reduces the odds of death to 0.98 of the initial odds (OR=0.98 95%CI(0.966, 0.995), p=0.011).

(ii)

Reliable estimates of the parameter

To look at how many parameters may be estimated reliably, we look at model diagnostics.

Cook's distance gives an indicator of influence of observation on the estimated parameters.

```
xi: logit death_all i.treatment wt_in
//Predict the cook's distance analogue in logistic regression
predict influence_stat, dbeta
//Predict the probabilities from the fitted model
predict probability, pr
predict num, number
//Plot dbeta values versus the fitted probabilities
scatter influence_stat probability, mlabel(num) mcolor("black")
msize(vsmall)
```

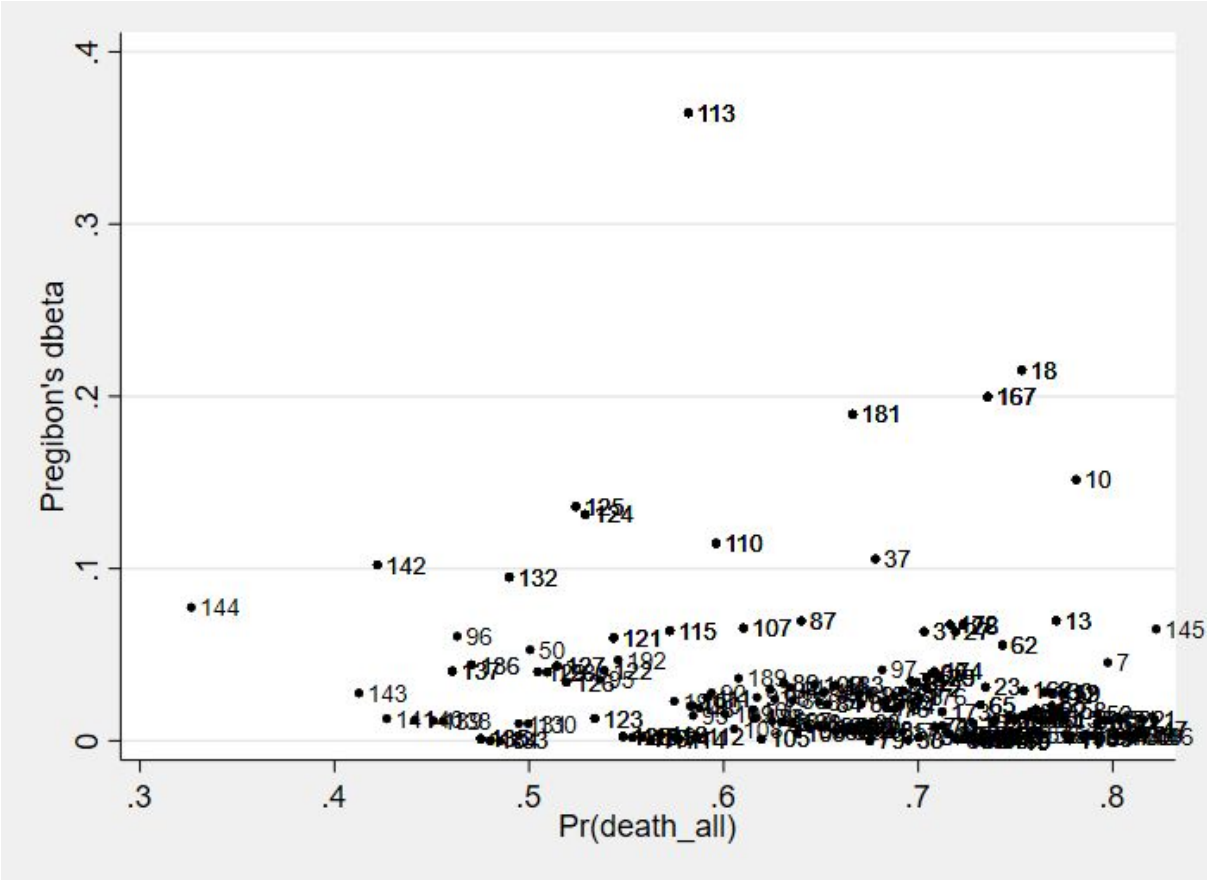

Fig. 20: Distribution of cook's distance for the logistic regression of death status on treatment type and weight index.

Based on the figure above we can say that the covariate pattern 113 exerts some influence on the estimated coefficients of the model we have fitted.

We will also look at the change in Pearson chi-square statistic and change in deviance when an observation is deleted to see the influence of observations on the model estimates.

```
//Predict hosmer lemeshow influence statistic dx2
predict hosmer_influence, dx2
scatter hosmer_influence probability, mlabel(num) mcolor("black")
msize(vsmall)
```

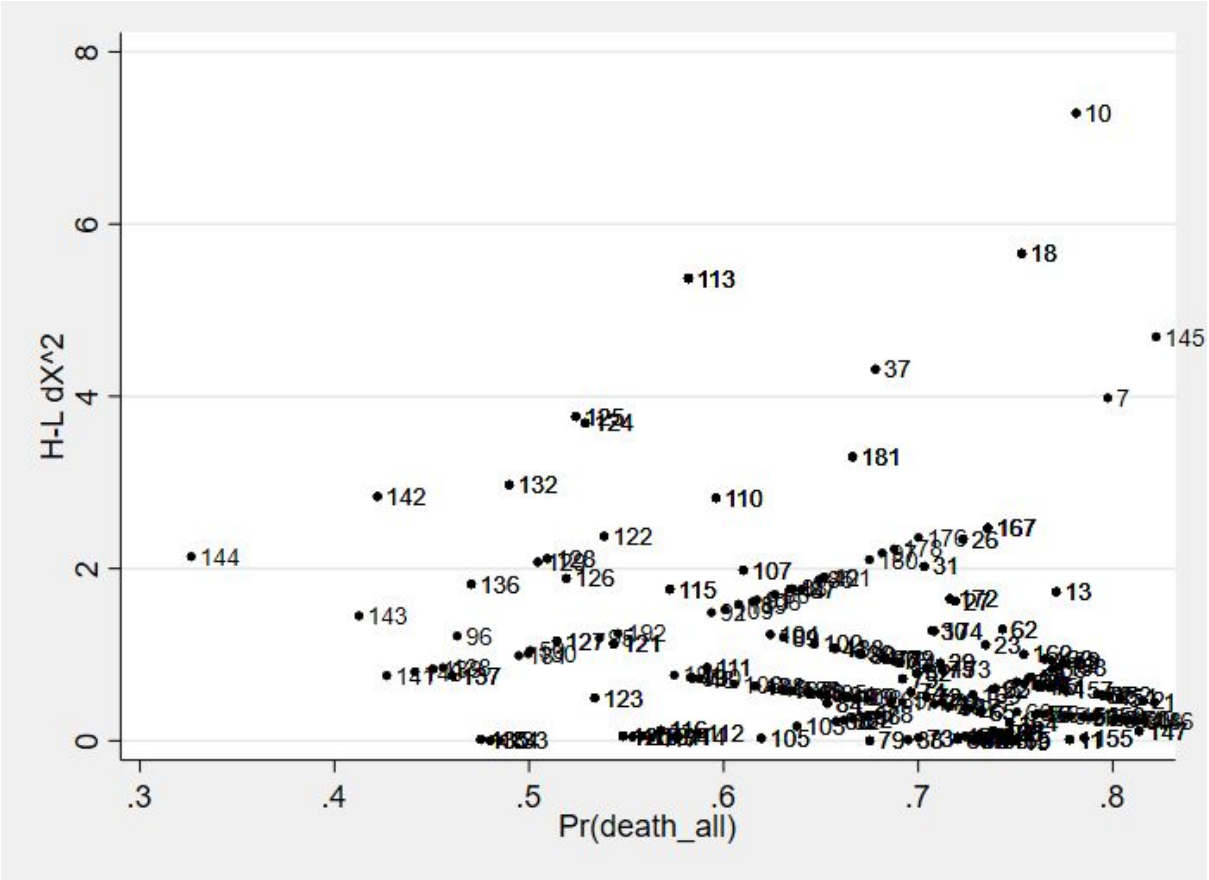

Fig. 21: Distribution of Hosmer Lemeshow influence statistic for logistic regression of death status on treatment type, weight index in the data.

From figure 20 above we can see that observation 10, 18 stand out and 113 appears again at mid level of Pr(death).

```
//Predicting hosmer lemeshow Ddeviance
predict hosmer_deviance, ddeviance
scatter hosmer_deviance probability, mlabel(num) mcolor("black")
msize(vsmall)
```

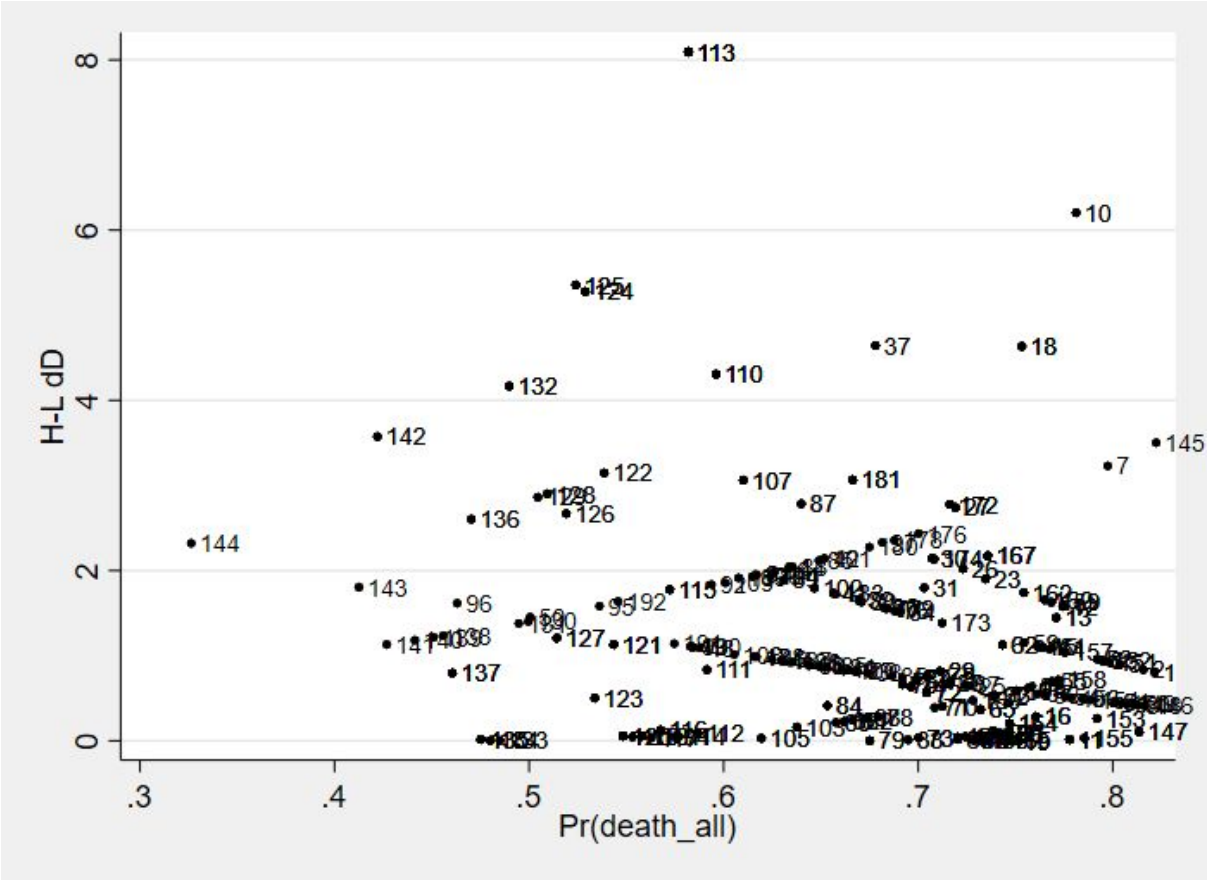

Fig. 22: Distribution of Hosmer Lemeshow influence statistic (DDeviance) for logistic regression of death status on treatment type, weight index in the data.

Again from the plot showing difference of log-likelihood when observation is deleted, observation 113 and 10 seem to be showing high influence on the parameter estimates.

Based on these results, it appears that the estimates of parameters are being influenced by the observation 113 and may need to be explored further using dfbetas (not covered in this course for the logistic regression). Given the situation, we can only estimate two parameters reliably (treatment effect = 3 and wt_index).

(iii)

To estimate the discrimination of the fitted model, I will use

a) Confusion matrix at normal cut off
b) Confusion matrix at optimal cut-off based on examination of sensitivity and specificity relationship in the model for different cut-off values for detecting death.
c) ROC curve
d)

Confusion matrix at normal cut off

The data for death vs alive is fairly balanced (67.26% are dead), hence it is suitable to use the methods stated above to assess discrimination.

At the cut-off of .50 , i.e. death is detected for probability equal to or more than 0.50, we see that Sensitivity is excellent (96%) but the specificity is extremely poor (12.33%). This means that the test is assessing dead cases quite rightly but it is predicting alive cases as dead quite frequently. The discriminatory power is compromised for alive cases for this cut off. To rectify it, we look at the optimal cut off using the code given in the next part.

```
. estat classification

Logistic model for death_all

                 True
Classified |         D          ~D |     Total
-----------+----------------------+----------
         + |       288         128 |       416
         - |        12          18 |        30
-----------+----------------------+----------
     Total |       300         146 |       446

Classified + if predicted Pr(D) >= .5
True D defined as death_all != 0
-------------------------------------------------
Sensitivity                     Pr( +| D)   96.00%
Specificity                     Pr( -|~D)   12.33%
Positive predictive value       Pr( D| +)   69.23%
Negative predictive value       Pr(~D| -)   60.00%
-------------------------------------------------
```

Fig. 23: Snapshot showing confusion matrix obtained for a cut-off 0f 0.50 for logistic regression of death status on weight index and treatment type.

Confusion matrix at optimal cut-off

```
quietly lsens, genprob(cutoff) gensens(sens) genspec(spec) nograph
//obtain difference between sensitivity and specificity
gen difference= sens-spec
//get absolute difference
gen mod_diff=abs(difference)
//sort difference absolute values from lowest to highest
sort mod_diff
//take cut-off corresponding to lowest difference and put in below
//command
estat classification, cutoff(.7001389596778)
```

The above code provides us the cut off of 0.70 where both sensitivity and specificity intersect or differ minimally from each other.
For this case we observe that sensitivity has decreased (56%) but specificity has increased (60.27%). The discriminatory power of the model overall has become more balanced now, even though the ability to predict dead cases has reduced by 40 percentage points, which may not be acceptable given it is more important to predict the risk of death in a scenario such as this.

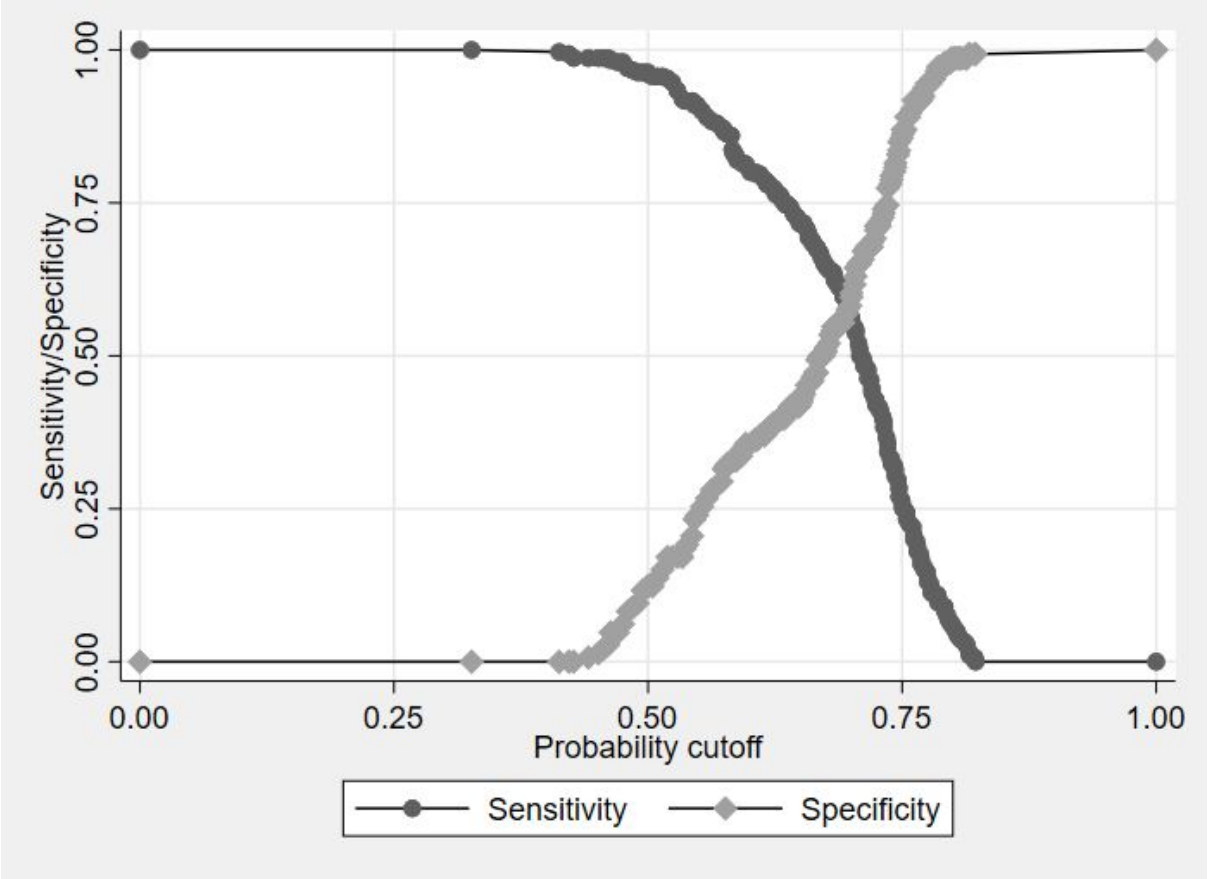

Fig. 24: Plot showing the relationship between sensitivity and specificity at different cut-off values. The point of intersection optimises the two.

```
. estat classification, cutoff(.7001389596778)

Logistic model for death_all

              True
Classified |         D          ~D |      Total
-----------+--------------------------+-----------
         + |       168          58 |        226
         - |       132          88 |        220
-----------+--------------------------+-----------
     Total |       300         146 |        446

Classified + if predicted Pr(D) >= .700139
True D defined as death_all != 0
-------------------------------------------------
Sensitivity                     Pr( +| D)   56.00%
Specificity                     Pr( -|~D)   60.27%
Positive predictive value       Pr( D| +)   74.34%
Negative predictive value       Pr(~D| -)   40.00%
-------------------------------------------------
```

Fig. 25: Snapshot showing confusion matrix obtained for a cut-off of 0.70 for logistic regression of death status on weight index and treatment type.

ROC Curve

The area under the ROC curve for this model is only 62% which indicates not a very good amount of discriminatory power of the model. In other words, for a given pair of dead and alive people, it will be able to tell only 62% of the time correctly their actual status.

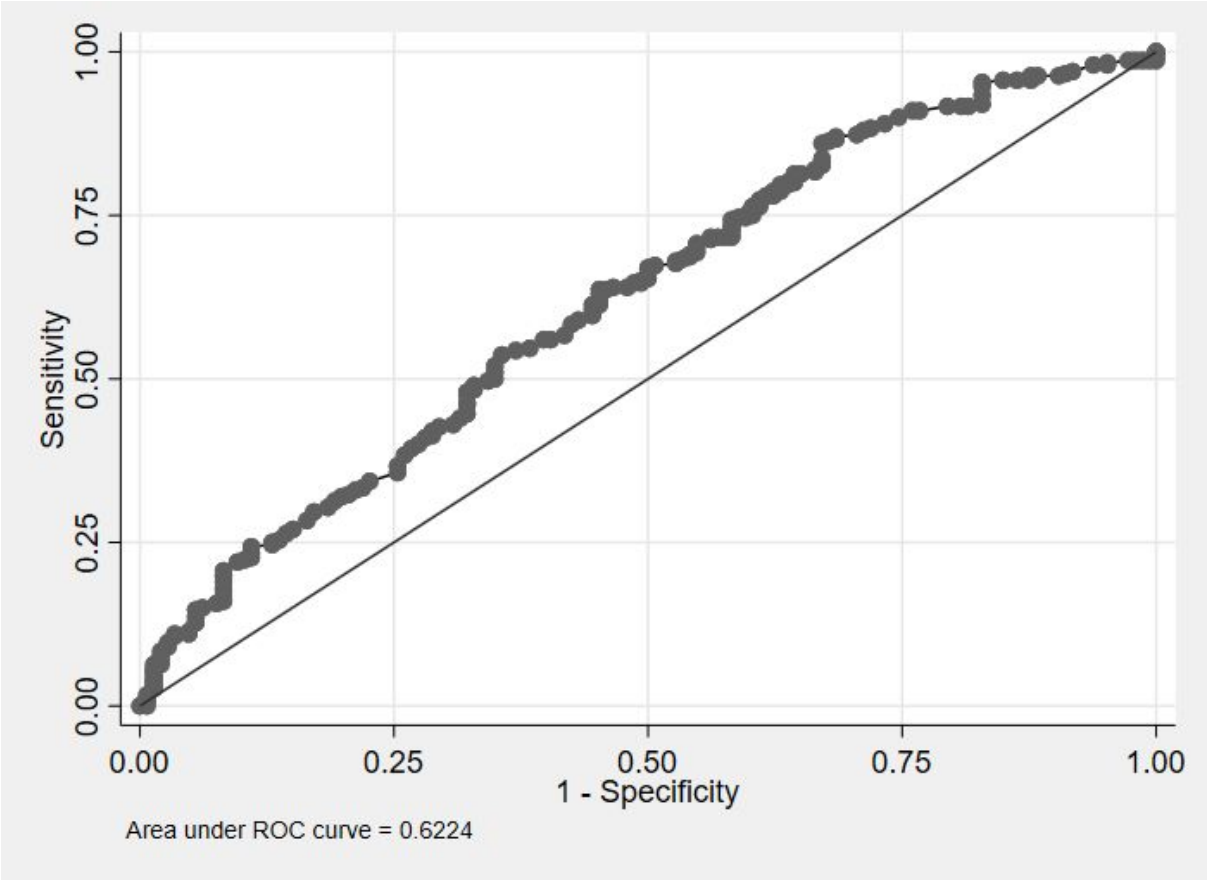

Fig. 26: Receiver-operating characteristic curve obtained for the logistic regression of death status on weight index and treatment types in the given dataset.

(iv)

Interaction term: treatment and stage controlling for age.

```
//logistic regression model with interaction term
logit death_all wt_in i.treatment##i.stage
```

There are 4 treatment groups and 2 stages. A total of 8 groups will be formed.

Logistic regression

Number of obs	=	446
LR chi2(8)	=	26.03
Prob > chi2	=	0.0010
Pseudo R2	=	0.0461

Log likelihood = -268.98806

death_all	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
wt_in	-.0180225	.007764	-2.32	0.020	-.0332396	-.0028054
treatment						
0.2 mg estrogen	.2543843	.3792112	0.67	0.502	-.4888561	.9976247
1.0 mg estrogen	-.6811504	.3569071	-1.91	0.056	-1.380676	.0183747
5.0 mg estrogen	.1066591	.383146	0.28	0.781	-.6442932	.8576115
4.stage	1.063981	.4884062	2.18	0.029	.1067222	2.021239
treatment#stage						
0.2 mg estrogen#4	-.7943393	.6673325	-1.19	0.234	-2.102287	.5136083
1.0 mg estrogen#4	-.6089792	.6160527	-0.99	0.323	-1.81642	.598462
5.0 mg estrogen#4	-.8866001	.651604	-1.36	0.174	-2.16372	.3905202
_cons	2.448413	.8397584	2.92	0.004	.8025168	4.094309

Fig. 27: Snapshot showing the results of logistic regression on death status, weight index, treatment type, stage and interaction between treatment type and stage.

The overall logistic regression may be written as:

Log(odds)=constant+ wt_in + T2 +T3+T4+ S4 + T2 * S4 + T3 * S4+ T4* S4

Log(odds) =constant +
b0* wt_in+
b1* S4 +
b2* T2 +
b3* T3 +
b4* T4 +
b12 * T2 * S4+
b13 * T3 * S4+
b14 * T4 * S4

Where T2, T3, T4 indicates Treatments 2, 3,4 respectively and S4 represents stage 4. Treatment 1 = Placebo and Stage 3 are the respective reference categories.

b0 represents the coefficient for wt_in and b1 for stage 4, b2, b3, b4 are the coefficients for treatment groups 2, 3, 4 respectively.

Log(Odds_1)	Placebo + Stage 3	Const+ b0*wt_in
Log(Odds_2)	Placebo + Stage 4	Const + b0*wt_in + b1
Difference	Log(Odds_2)-Log(Odds_1)	b1

Table 2: Comparing Log odds of death for patients on Placebos in Stage 3 and Stage 4.

Adjusted for weight index, in the case of placebo, those in stage 3 have exp(b1) = exp(1.06) = 2.897 times the odds of death of those in stage 3. Thus, they have much higher odds of death, 95% CI being (1.11 to 7.54), not including 1 means statistically significant difference in odds.

This can be obtained using the command

lincom _b[4.stage], or

```
. lincom _b[4.stage], or

 ( 1)  [death_all]4.stage = 0
```

death_all	Odds Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
(1)	2.897884	1.415345	2.18	0.029	1.112625	7.547674

Fig. 28: Snapshot showing the odds ratio stage 4 versus stage 3 in placebo group.

-----Next Case----

Log(Odds_1)	Treatment =2 and stage 3	Const+ b0*wt_in + b2
Log(Odds_2)	Treatment =2 and stage 4	Const+ b0*wt_in + b2 + b1 + b12
Difference	Log(Odds_2)-Log(Odds_1)	b1+b12

Table 3: Comparing Log odds of death for patients on 0.2mg of estrogen in Stage 3 and Stage 4.

Adjusted for weight index, among those who are taking 0.2 mg of estrogen, the odds of diabetes in stage 4 is exp(b1+b12) =exp(1.06-.7943393) = 1.309 times the odds of death of those in stage 3. Thus, they have higher odds of death, 95% CI being 0.536 to3.19, which includes 1, hence the result is not statistically significant, but the higher upper value of CI indicates that difference may exist.

lincom _b[4.stage]+_b[2.treatment#4.stage], or

Y3874726

```
( 1)  [death_all]4.stage + [death_all]2.treatment#4.stage = 0
```

death_all	Odds Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
(1)	1.309495	.5962242	0.59	0.554	.5364688	3.196415

Fig. 29: Snapshot showing the odds ratio stage 4 versus stage 3 among those who take 0.2mg of estrogen.

-----Next Case----

Log(Odds_1)	Treatment = 3 and stage 3	Const+ b0*wt_in + b3
Log(Odds_2)	Treatment =3 and stage 4	Const+ b0*wt_in + b3 + b1 + b13
Difference	Log(Odds_2)-Log(Odds_1)	b1+b13

Table 4: Comparing Log odds of death for patients on 1 mg of estrogen in Stage 3 and Stage 4.

Adjusted for weight, among those who take 1 mg of treatment the odds of death in stage 4 patients is exp(1.063981-.6089792) = 1.576 times the odds of those in stage 3. The 95% CI is (0.754 , 3.29) which includes 1 indicating that this ratio is not statistically significant.

lincom _b[4.stage]+_b[3.treatment#4.stage], or

```
. lincom _b[4.stage]+_b[3.treatment#4.stage], or

 ( 1)  [death_all]4.stage + [death_all]3.treatment#4.stage = 0
```

death_all	Odds Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
(1)	1.576176	.5927487	1.21	0.226	.7542153	3.293927

Fig. 30: Snapshot showing the odds ratio stage 4 versus stage 3 among those who take 1mg of estrogen.

-----Next Case----

Log(Odds_1)	Treatment = 4 and stage 3	Const+ b0*wt_in + b4
Log(Odds_2)	Treatment = 4 and stage 4	Const+ b0*wt_in + b4 + b1 + b14
Difference	Log(Odds_2)-Log(Odds_1)	b1+b14

Table 5: Comparing Log odds of death for patients on 5 mg of estrogen in Stage 3 and Stage 4.

Adjusted for weight, among those who take 1 mg of treatment the odds of death in stage 4 patients is exp(1.063981 -.8866001) =

1.194 times the odds of those in stage 3. The 95% CI is (0.512 , 2.78) which includes 1 indicating that this ratio is not statistically significant.

```
lincom _b[4.stage]+_b[4.treatment#4.stage], or
```

```
. lincom _b[4.stage]+_b[4.treatment#4.stage], or

 ( 1)  [death_all]4.stage + [death_all]4.treatment#4.stage = 0

------------------------------------------------------------------------------
   death_all | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   1.194086   .5156673     0.41   0.681     .5122041    2.783735
------------------------------------------------------------------------------
```

Fig. 31: Snapshot showing the odds ratio stage 4 versus stage 3 among those who take 5mg of estrogen.

These results are consistent with the p-values of the interaction terms of treatment and stage, none of them were less than 0.05. Moreover, the p-value for individual treatments (without interaction of stage) also failed to show any statistical significance in the given dataset, except for 1mg estrogen treatment which showed marginal statistical significance (p=0.056). Overall, adjusting for weight, difference in the log odds of death is found significant only for the placebo group.

(v)

Effect on death among those with 0.2mg estrogen and those on placebo on stage 4 of disease.

Log(Odds_1)	Treatment =2 and stage 4	Const+ b0*wt_in + b2 + b1 + b12
Log(Odds_2)	Placebo + Stage 4	Const + b0*wt_in + b1
Difference	Log(Odds_1)-Log(Odds_2)	b2+b12

Table 6: Comparing Log odds of death for patients on 0.2 mg of estrogen and placebo belonging to Stage 4.

Adjusted for weight, the ratio of odds (Odds_1/Odds_2) is exp(.2543843-.7943393) = 0.583 . Thus, among stage 4 patients, the odds of death among those who take 0.2mg of estrogen is 0.58 times the odds of death among those taking placebo. The 95% CI is (0.198, 1.71). The confidence interval contains 1, hence the effect does not seem to be statistically significant , however the wide range of CI indicates that deeper analysis may be required.

```
lincom _b[2.treatment]+_b[2.treatment#4.stage], or
```

```
 ( 1)  [death_all]2.treatment + [death_all]2.treatment#4.stage = 0
```

death_all	Odds Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
(1)	.5827745	.3201225	-0.98	0.326	.198577	1.710299

Fig. 32: Snapshot showing the odds ratio Treatment = 2 (0.2 mg of estrogen) versus Placebo in stage 4.

(vi)

Log(Odds_1)	Treatment =4 and stage 4	Const+ b0*wt_in + b4 + b1 + b14
Log(Odds_2)	Treatment =2 and stage 4	Const+ b0*wt_in + b2 + b1 + b12
Difference	Log(Odds_1)-Log(Odds_2)	b4+b14-b2-b12

Table 7: Comparing Log odds of death for patients on 5 mg of estrogen and 0.2 mg of estrogen belonging to Stage 4.

Adjusted for weight, the ratio of odds (Odds_1/Odds_2) is exp(.1066591-.8866001 - (.2543843 -.7943393)) = 0.7866. Thus, among those in stage 4, those who are on 5 mg of estrogen the odds of death is 0.79 times the odds of death among those who take 0.2 mg of estrogen. They seem to be having lower odds of death. The corresponding CI is (0.30, 2.05) which includes 1, indicating that this difference is not statistically significant. This is further indicated by the p value which is 0.624, much higher than 0.05.

```
lincom _b[4.treatment]+_b[4.treatment#4.stage]-_b[2.treatment]-_b[2.treatment#4.stage], or
```

```
. lincom _b[4.treatment]+_b[4.treatment#4.stage]-_b[2.treatment]-_b[2.treatment#4.stage],
> or

 ( 1)  - [death_all]2.treatment + [death_all]4.treatment - [death_all]2.treatment#4.stage
       + [death_all]4.treatment#4.stage = 0
```

death_all	Odds Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
(1)	.7866389	.3848176	-0.49	0.624	.3015604	2.051996

Fig. 33: Snapshot showing the odds ratio Treatment = 2 (0.2 mg of estrogen) versus Treatment =4 (5mg of estrogen) in stage 4.

(vii)

Fit of the model

The model fitted is

```
logit death_all wt_in i.treatment##i.stage
```

To assess the fit, we use

1. Compare it with the model without interaction.
2. Goodness of fit using Hosmer-Lemeshow method.

Compare it with the model without interaction.

```
//Fitting logistic model without interaction term
xi: logit death_all wt_in i.treatment i.stage
//Store estimates of model fitted in previous command
est store A
//Fitting logistic model with interaction term
logit death_all wt_in i.treatment##i.stage
//Store estimates of model fitted in previous command
est store B
//Log-likelihood test between model estimates A and B
lrtest A B
```

Model A is nested in Model B, thus it is a more parsimonious model. The null hypothesis for `lrtest` is that the parsimonious model is a better fit. The p-value for the test is 0.53 which indicates that we cannot reject the null hypothesis and that model A is a better fit than model B for this data; addition of interaction term does not improve the fit of the model. This is further substantiated by the Pseudo-R squared value. For the model without interaction (model A) the value is 4.2% but for that with interaction it is 4.6%; a marginal improvement.

Goodness of fit using Hosmer-Lemeshow method.

The null hypothesis tested in this test is, the model fits against the alternative hypothesis that it does not fit. The results of this analysis show that the p-value is 0.9446, Hosmer-Lemeshow chi-squared statistic is 2.83. A p-value higher than 0.05 indicates that the null hypothesis cannot be rejected and that the model fits well overall to the data.

```
logit death_all wt_in i.treatment##i.stage
estat gof, table group(10)
```

```
. estat gof, table group(10)

Logistic model for death_all, goodness-of-fit test

(Table collapsed on quantiles of estimated probabilities)
```

Group	Prob	Obs_1	Exp_1	Obs_0	Exp_0	Total
1	0.5048	20	21.3	26	24.7	46
2	0.5729	26	23.8	18	20.2	44
3	0.6226	27	26.5	17	17.5	44
4	0.6602	29	28.9	16	16.1	45
5	0.6876	31	30.4	14	14.6	45
6	0.7103	29	30.8	15	13.2	44
7	0.7291	29	32.4	16	12.6	45
8	0.7633	35	33.6	10	11.4	45
9	0.8139	36	34.5	8	9.5	44
10	0.8983	38	37.7	6	6.3	44

```
      number of observations =       446
            number of groups =        10
  Hosmer-Lemeshow chi2(8) =         2.83
                  Prob > chi2 =     0.9446
```

Fig. 34: Snapshot showing the results of Hosmer-Lemeshow Goodness of Fit test for logistic regression of death on treatment type, weight index, stage and interaction of treatment type and stage.

Furthermore at
Iteration 0: log likelihood = -282.00094
Iteration 4: log likelihood = -268.98806
The reduction in log likelihood = -268.98806 - (-282.00094) = 13.01. Thus the deviance is 2 * 13.01 = 26.02 which is quite less., degree of freedom = 8.

2 (b)

DBP VS SBP

```
twoway (scatter dbp sbp if death_all==0, msize(tiny) mcolor(red)
scale(0.8) yscale(titlegap(*10))  xscale(titlegap(*10))) (lfit
dbp sbp  if death_all==0, lwidth(medium) lcolor(red)
lpattern(solid)) (scatter dbp sbp if death_all==1, msize(tiny)
mcolor(blue) xtitle("Systolic Blood Pressure/10")
ytitle("Diastolic Blood Pressure/10")) (lfit  dbp sbp  if
death_all==1, lwidth(medium) lcolor(blue) lpattern(solid)),
legend( label(1 "Alive" )  label(3 "Dead"))
```

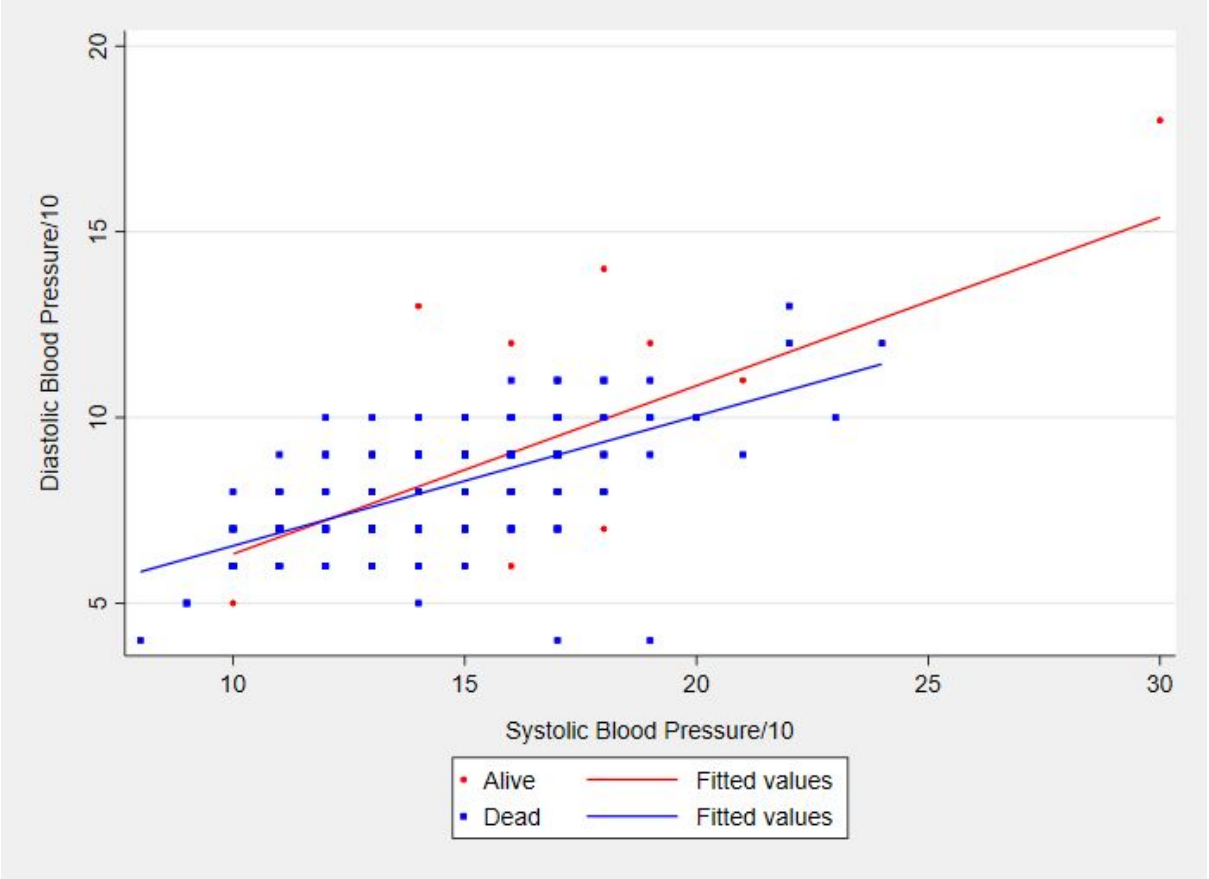

Fig. 35: Relationship between Diastolic Blood Pressure and Systolic Blood Pressure in the data.

Positive correlation seen among dead and alive for diastolic and systolic blood pressure. Data points with dead death status dominate the plot. Outlier seen for high sbp=30.

DBP VS AGE

```
twoway (scatter dbp age  if death_all==0, msize(tiny) mcolor(red)
scale(0.8)yscale(titlegap(*10))  xscale(titlegap(*10))) (lfit  dbp
age  if death_all==0, lwidth(medium) lcolor(red)
lpattern(solid))(scatter dbp age if death_all==1, msize(tiny)
mcolor(blue) xtitle("Age (in years)") ytitle("Diastolic Blood
Pressure/10")) (lfit  dbp age  if death_all==1, lwidth(medium)
lcolor(blue) lpattern(solid)), legend( label(1 "Alive" )  label(3
"Dead"))
```

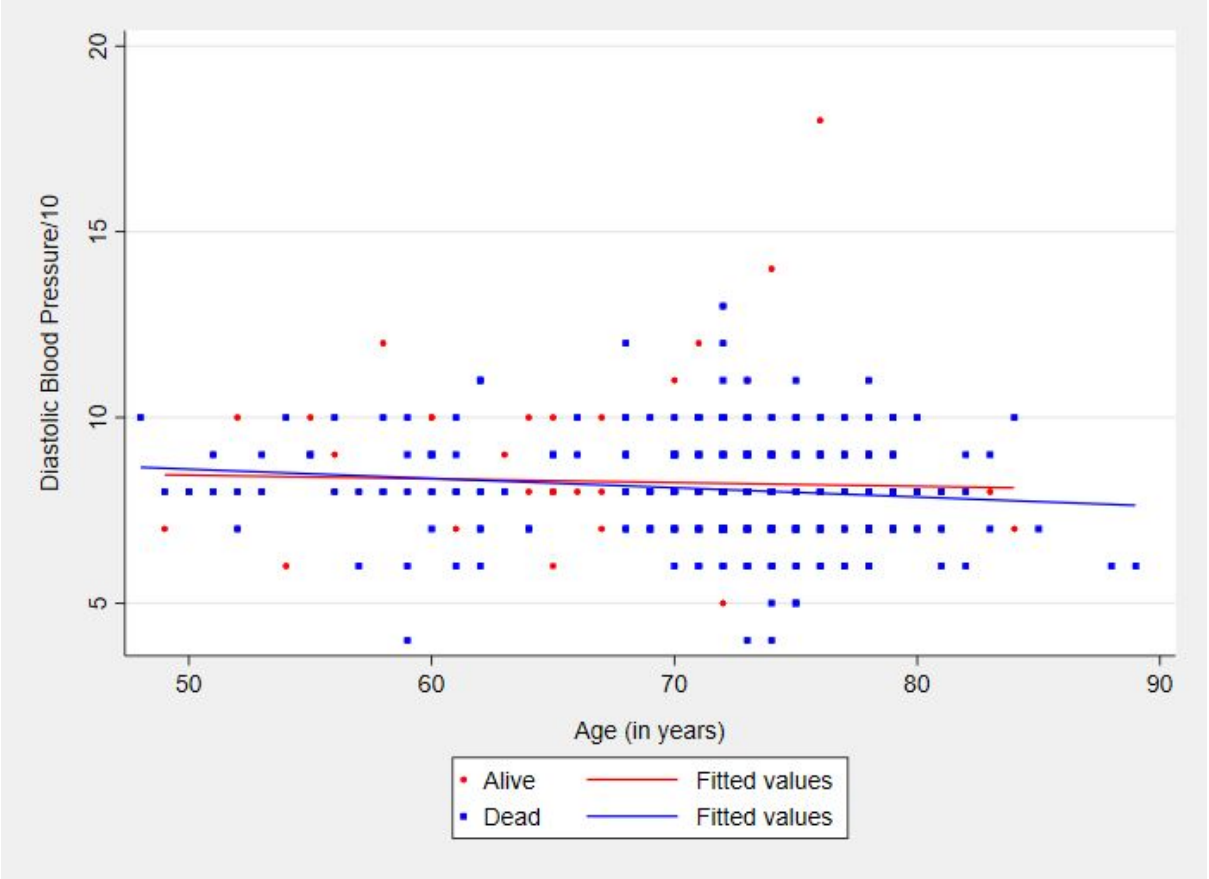

Fig. 36: Relationship between Diastolic Blood Pressure and Age in the data.

Most of the diastolic blood pressure values lie below 10 (100 on actual scale) acros all the age groups. Higher proportion of dead people is seen than alive, as indicated by blue data points.Almost no relationship seen among sbp and age, although slight negative correlation is indicated by the regression lines.

DBP VS WT_IN

```
twoway (scatter dbp wt_in if death_all==0, msize(tiny) mcolor(red)
scale(0.8)yscale(titlegap(*10))  xscale(titlegap(*10))) (lfit  dbp
wt_in  if death_all==0, lwidth(medium) lcolor(red)
lpattern(solid)) (scatter dbp wt_in if death_all==1, msize(tiny)
mcolor(blue) xtitle("Weight Index") ytitle("Diastolic Blood
```

```
Pressure/10")) (lfit  dbp wt_in  if death_all==1, lwidth(medium)
lcolor(blue) lpattern(solid)) , legend( label(1 "Alive" )  label(3
"Dead"))
```

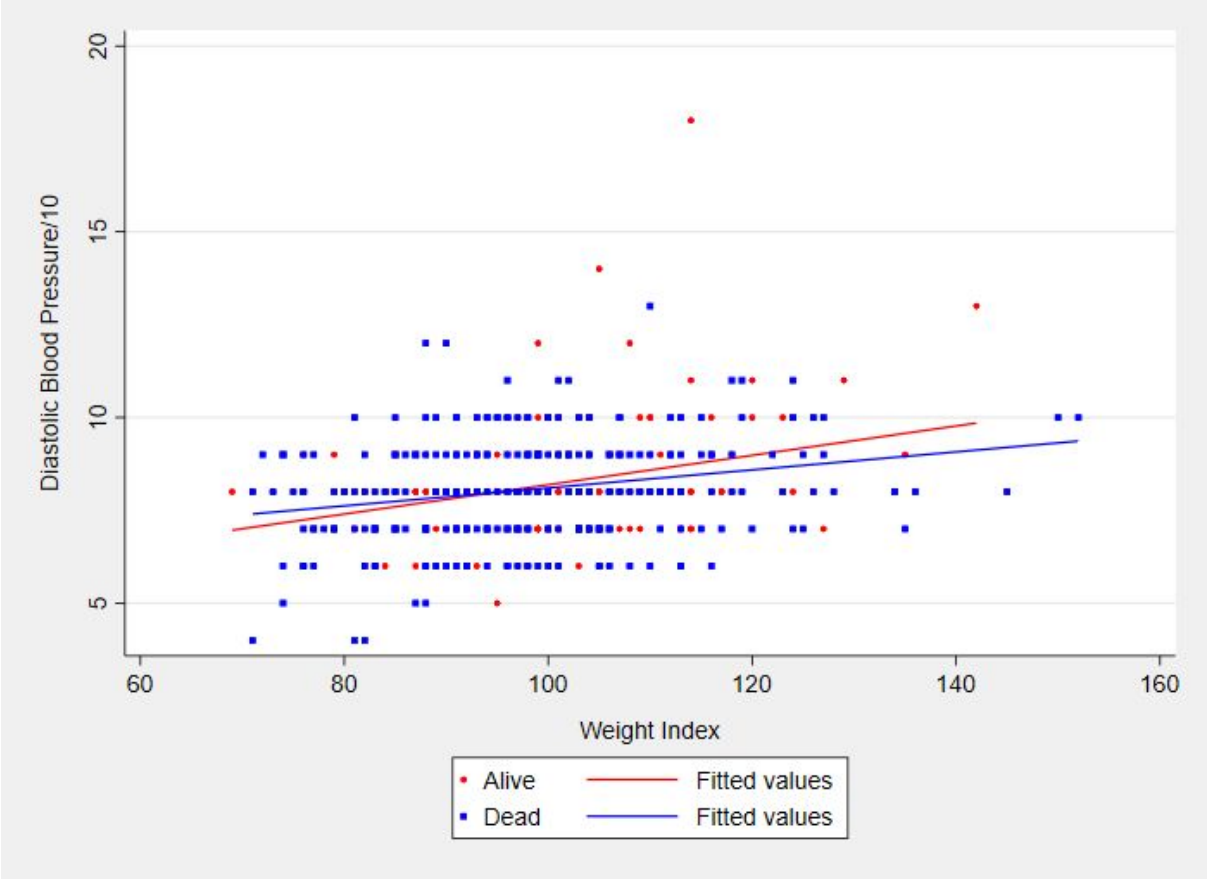

Fig. 37: Relationship between Diastolic Blood Pressure and Weight Index in the data.

For the range of weight index values, the diastolic blood seldom goes beyond 10 (100 in actual scale). Clusters among diastolic blood pressure are observed at different values. Positive correlation seen between diastolic blood pressure and weight index in both the groups, the slope for diastolic blood pressure in alive people is higher than that for dead.

SBP VS AGE

```
twoway (scatter sbp age if death_all==0, msize(tiny) mcolor(red)
scale(0.8)yscale(titlegap(*10))  xscale(titlegap(*10)))(lfit  sbp
age  if death_all==0, lwidth(medium) lcolor(red) lpattern(solid))
(scatter sbp age if death_all==1, msize(tiny) mcolor(blue)
xtitle("Age (in years)") ytitle("Systolic Blood Pressure/10"))
(lfit  sbp age  if death_all==1, lwidth(medium) lcolor(blue)
lpattern(solid)), legend( label(1 "Alive" )  label(3 "Dead"))
```

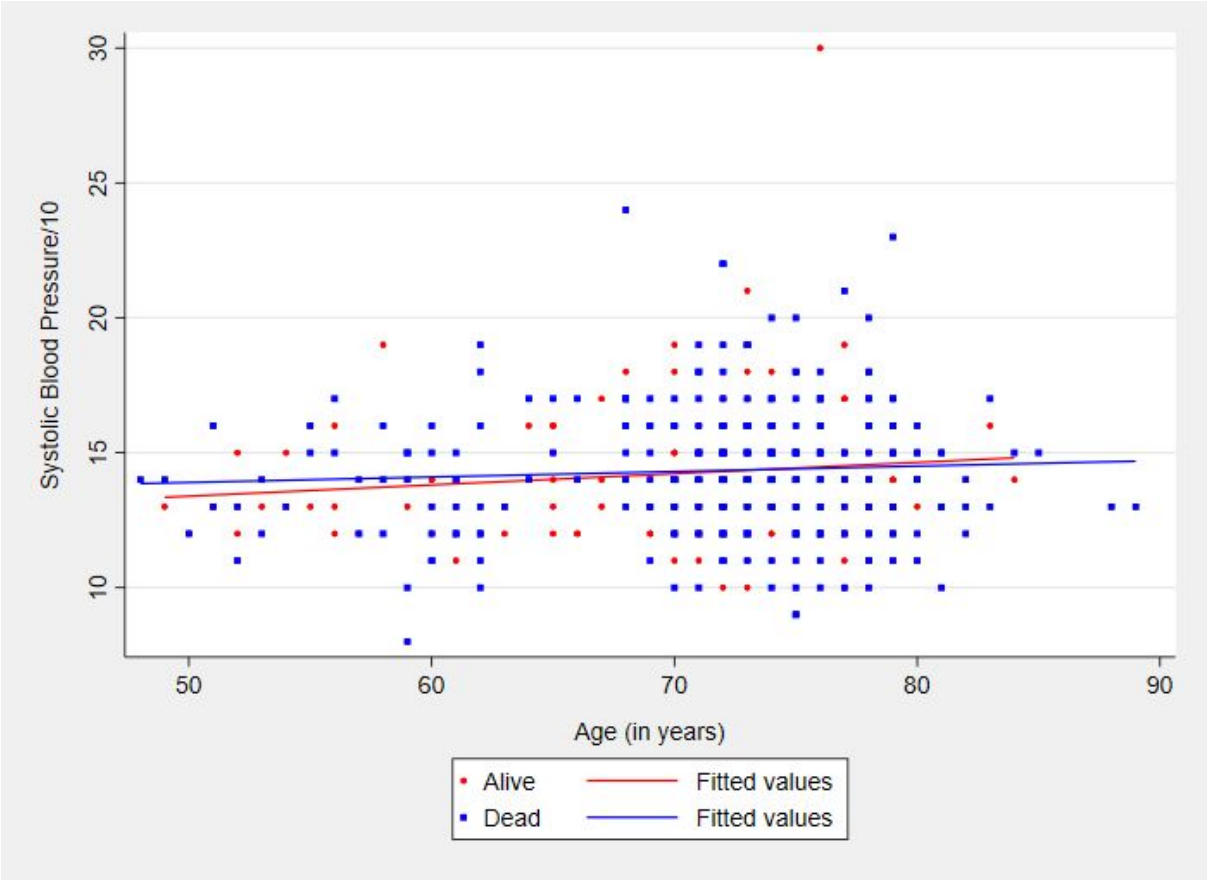

Fig. 38: Relationship between Systolic Blood Pressure and Age in the data.

Relationship is pretty random, higher data points in higher age range, indicating that old people (death status=dead) are more in the data set. Very low positive correlation seems to exist between age and systolic blood pressure.

SBP VS WT_IN

```
twoway (scatter sbp wt_in if death_all==0, msize(tiny) mcolor(red)
scale(0.9)yscale(titlegap(*10)) xscale(titlegap(*10))) (lfit sbp
wt_in if death_all==0, lwidth(medium) lcolor(red)
lpattern(solid)) (scatter sbp wt_in if death_all==1, msize(tiny)
mcolor(blue) xtitle("Weight Index") ytitle("Systolic Blood
Pressure/10") ) (lfit sbp wt_in if death_all==1, lwidth(medium)
lcolor(blue) lpattern(solid)), legend( label(1 "Alive" ) label(3
"Dead"))
```

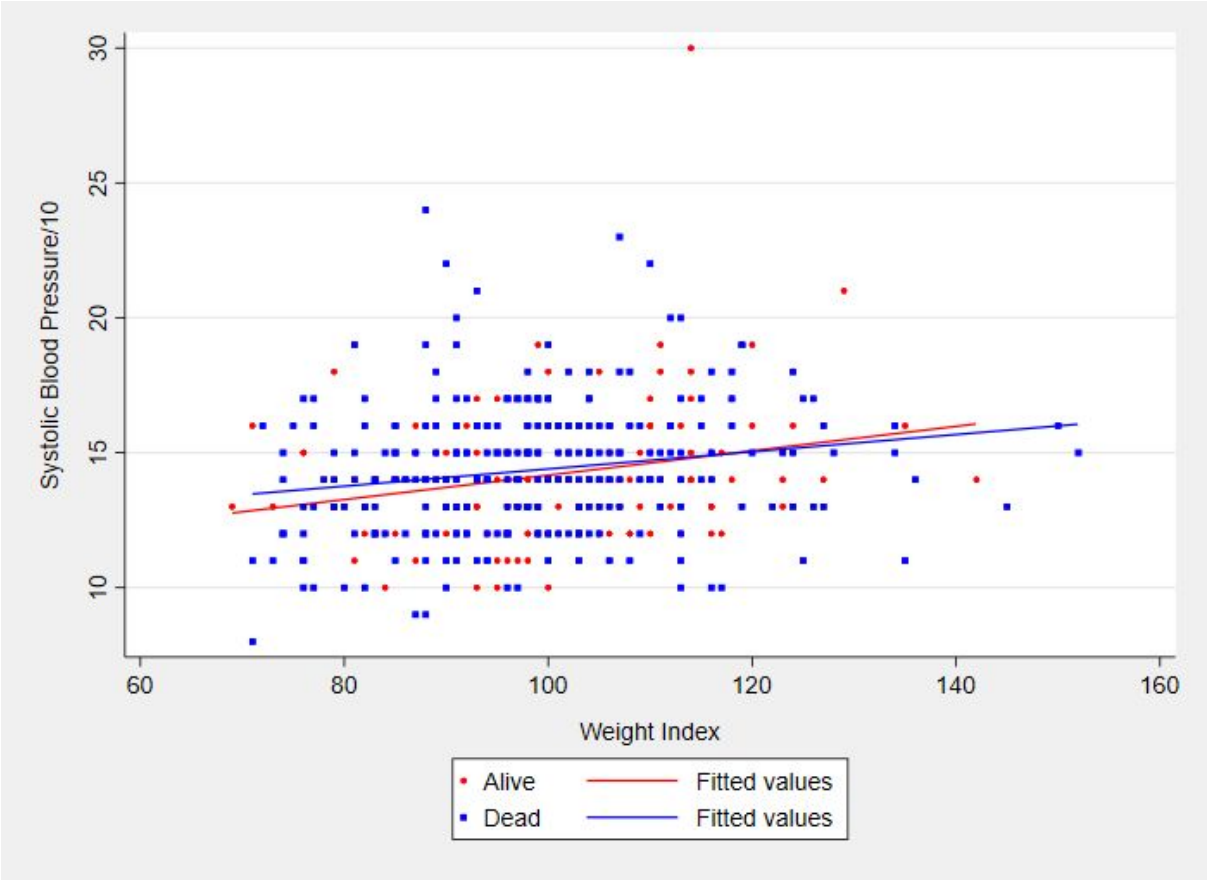

Fig. 39: Relationship between Systolic Blood Pressure and Weight Index in the data.

Slightly positive correlation observed between weight index and systolic blood pressure for both groups of death.

WT_IN VS AGE

```
twoway (scatter wt_in age  if death_all==0, msize(tiny)
mcolor(red) scale(0.8)yscale(titlegap(*10))
xscale(titlegap(*10))) (lfit  wt_in age  if death_all==0,
lwidth(medium) lcolor(red) lpattern(solid)) (scatter wt_in age  if
death_all==1, msize(tiny) mcolor(blue) ytitle("Weight Index")
xtitle("Age (in years)")) (lfit  wt_in age  if death_all==1,
lcolor(blue) lwidth(medium) lpattern(solid)),legend( label(1
"Alive" )  label(3 "Dead"))
```

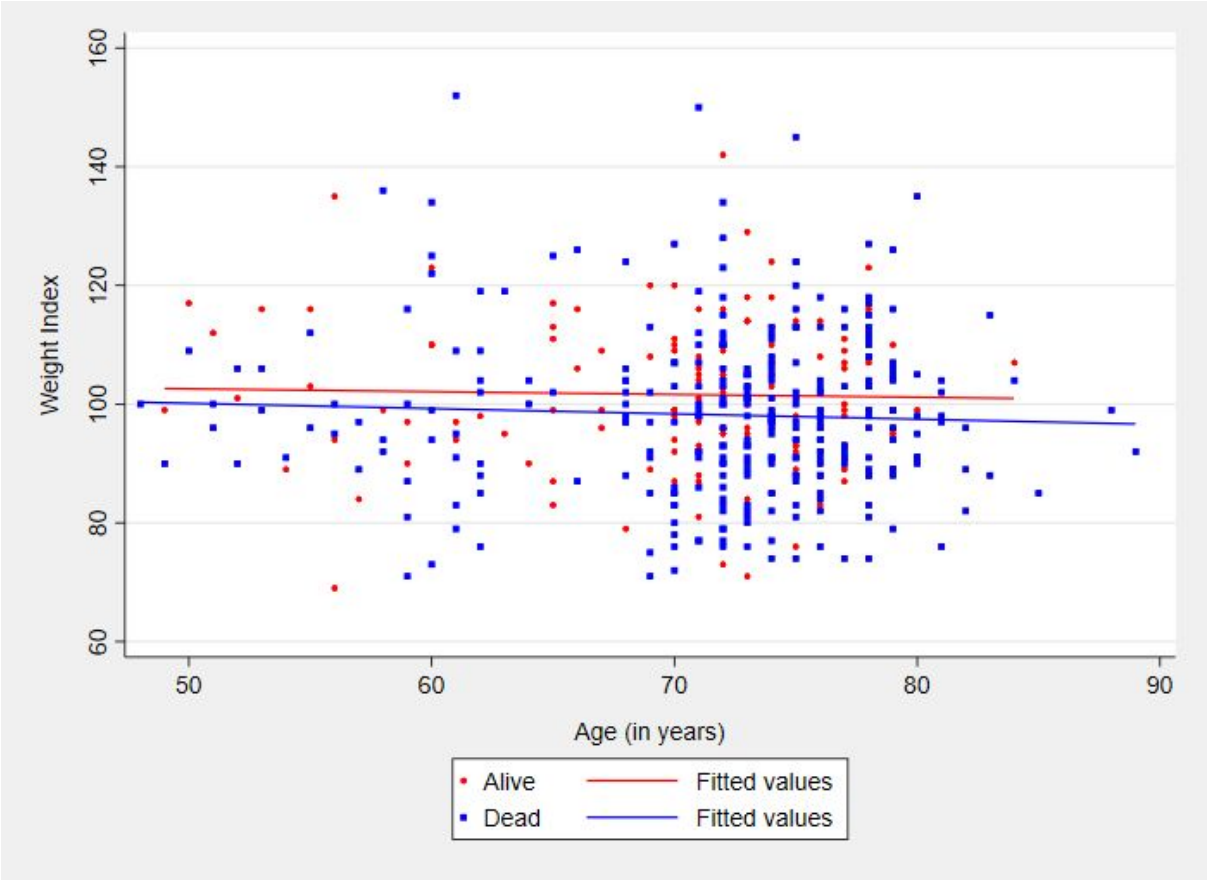

Fig. 40: Relationship between Weight Index and Age in the data.

Extremely low negative correlation, almost zero correlation between age and weight index observed for both death groups.

2(c)

(i)

Kaplan-Meier approach based survival curve is presented in Figure 1. Hence, this curve is known as the Kaplan-Meier Survival Curve.

(ii)

a) At the starting of the conception (Year = 0), the probability of not having prostate cancer is 100% among all the three groups of different conception methods. All groups are free from any case of prostate cancer.
b) As the years following conception increases, all three groups show a decrease in the probability of not having prostate cancer indicating that the risk of getting cancer increases.
c) Upto 18 years after child conception, the probability of not having prostate cancer is highest among those who conceived normally, followed by those with in vitro fertilization and least for those who conceived using sperm injection. This indicates that the risk of prostate cancer is quite high among this last group in the years following conception.
d) After 18 years of conception, the risk of not having prostate cancer becomes constant in the sperm-injection group, at 0.9870. This possibly indicates right censoring and no more cases were available beyond this point among those with sperm injection
e) The risk of not having prostate cancer continues to fall in the in-vitro fertilization group and goes till 0.98 , until a little after 20 years, where it becomes constant.
f) The decrease in the probability of not having risk cancer is quite slow for those who conceive normally and it does not go beyond 0.989, indicating that the risk for prostate cancer is pretty low for this group as compared to the other two groups.
g) The study went for 25 years after conception, and found no cases after 21 years.

(iii)

Note that : Probability of not having prostate cancer is the same as the probability of survival.

Variable Name General	**Variable STATA identifier**	**Description**
Follow up status	fol_pc_status	0 = No prostate cancer 1= Developed prostate cancer
Follow up time	time_to_event	Days since starting of the study
Time to event in years	time_to_event_years	Days/365.25
Conception Method	method_of_conception	0=Natural Conception 1=IVF 2=ICSI

Table 8: Table showing the variables assumed to explain the Survival Model for Prostate Cancer.

Commands used

```
//Compute time in year from the given time in days
gen time_to_event_years=time_to_event/365.25
//tell stata that the data is type time-to-event
stset time_to_event_years, failure(fol_pc_status)
//plot the basic survival curve
sts graph
//Detailed plot. This is a single command.
sts graph,  by(method_of_conception) legend( ring(1) position(12)
region(lstyle(none)) order(1 "Natural Conception" 2 "IVF" 3 "ICSI"
)) plot1(lcolor(blue)lwidth(medium)lpattern(solid))
plot2(lcolor(yellow)lwidth(medium)
lpattern(dash))plot3(lcolor(pink)lwidth(medium) lpattern(dash))
ytitle("Probability of not having prostate cancer") xtitle("Years
after child conception") ylabel(, angle(horizontal))
yscale(titlegap(*5))

//End of command
```

(iv)

In this figure 1, hazard ratio do not seem to be constant. Hence, Wilcoxon (Breslow) test for equality of survivor functions will be used to determine the difference between the three graphs.

This procedure does not require the assumption of the proportional hazards to be met, hence, may be used as an alternative to the Mantel-Haenszel statistic or log-rank test. Wilcoxon test yields more reliable results for the data with a constant hazard ratio as well. The null hypothesis is that the survival curves in the groups are the same. A p value of 0.002 (<0.05) indicates that the survival curves are not the same and there is a systematic difference between the survival probabilities in at least one group as compared to the other two. From the plot, it is clear that probability of prostate cancer is quite less among the normal conceived group than the other two throughout the 25 years after conception. Hence, this could be the reason behind the statistically significant p-value. The other two groups tend to intersect at a point in the curve, so equality seems to exist at one time-point between them, so the possibility of their being different is less.

```
// Command is
sts test method_of_conception, wilcoxon
```

(v)

Cox proportional hazards regression analysis will be used.
Log(hazard)=Log(baseline_hazard) + coef * method_of_conception
The outcome has already been set using the commend before:

```
stset time_to_event_years, failure(fol_pc_status)
```

Thus, for the regression command is

```
//Regression on method of conception
stcox i.method_of_conception
```

This will give us hazard ratio for difference conception method categories.
To obtain coefficients instead of the hazard ratios we use this command

```
//Regression without hazard ratios
stcox i.method_of_conception, nohr
```

(vi)

Analysis strategy

```
stcox i.method_of_conception
```

Since there are three methods of conception something like below will be returned

_t	Hazard Ratio	Std. Error, z	p, CI
Method of conception			
(Method =1)IVF	H1		P1

1.method_of_conception			
(Method=2)ICSI 2.method_of_conception	H2		P2

Table 9: Table showing the assumed result of a hypothetical Cox proportional hazards regression model.

After running the model, I will look at the log-likelihood of the developed model and compare it with that of the null model to understand whether or not the model is able to explain the pattern and shows reduced residual deviance as compared to the null deviance. Later, I will look at the value of H1 and H2 and their respective p values and confidence intervals. Using these I will be able to conclude whether or not the method of conception affects the hazard ratio, in this case the hazard of developing prostate cancer over a short period of time. This will help me in understanding how the hazard of prostate cancer differs in those who conceived using in-vitro fertilization (IVF) and sperm injection (ICSI) from those who conceived normally. Using the respective CIs of the H1 and H2, I will be able to understand the range of the hazard ratio and thus, the impact of these methods on hazard of developing prostate cancer in these groups.
Based on the results, I would also like to know the hazard ratio between IVF and ICSI, for that I will do this:
Log(hazard IVF)= Const+ H1
Log (hazard ICSI) =Const + H2
Log(hazard IVF)-Log(hazard ICSI) = H1-H2
Log(hazard IVF/hazard ICSI) = (H1-H2)
(hazard IVF/hazard ICSI) = exp(H1-H2)
To execute this, I will use
lincom _b[1.method_of_conception]-_b1[2.method_of_conception]
I would also like to test the assumptions of the model fitted so as to ascertain the reliability of my results. To do this, I will check
stphplot, by(method_of_conception) plot1opts(msymbol(i)) plot2opts(msymbol(i) lpattern(dash)) plot3opts(msymbol(i) lpattern(dot_dash)) nonegative

I will look at these plots to test the assumption of proportional hazards across the groups of conception methods by looking for an approximate parallel relationship between the three curves for the range of analysis time. To check the assumption more formally I will use
estat phtest

Y3874726

```
predict sch*, scaledsch
```

I will obtain two variables sch1 and sch2 corresponding to two regressors (two methods of conception).

```
scatter sch1 _t || lfit sch1 _t
scatter sch2 _t || lfit sch2 _t
```

These two plots should give zero slope for me to ascertain that the proportional hazard assumption is held.

--END----