Q: Does xb represent the fitted values?

A: The xb option in a predict command is used to calculate the fitted values after a regression.

Q: I was also confused by the instruction 'use the cprplot command to check the functional form of the explanatory variables in your model'.

A: You need to use "cprplot varname", where varname is one of the explanatory variables in your model, in order to check whether the relationship between the outcome and this variable is actually linear or needs a transformation. If the relationship is linear then the residuals should hover around a line with a slope equal to that of the regression slope associated with varname.

Q: Could you remind me what value we hope for to show an absence of collinearity when using the 'vif' command? Is it less than 1?

A: Unfortunately, there is no theoretical ground for what the threshold value should be to judge that VIF is “high.”. Recall VIF(explanatory variable A) = 1/(1- Rsquared of regression of A on all other explanatory variables in the model). So VIF is always greater than or equal to 1; its inverse tolerance is always less than or equal to one (Tolerance is the proportion of variance in A not accounted for by other variables in the model). VIF and tolerance are equal to 1 if there is no linear relationship between A and the other explanatory variables, that is when Rsquared of regression of A on all other explanatory variables in the model is 0. Big values of VIF are always a concern. As a rule of thumb, VIF > 4or 5 suggests multi-collinearity; VIF > 10 is strong evidence that collinearity is affecting the regression coefficients.

Q: what are we looking for when we plot cook's distance and leverage values?

A: We are mainly looking at points that are away from the bulk of the points as they might influence our estimates unduly. They suggest whether we have to investigate these points further and see how they affect our model.

Q: To remove a data point, use the ***drop*** command, e.g. *drop in 51

A: You can also use "if" in your regression command so that you exclude the point without having to drop it from the data set.

Q: Sample Assessment Document 1 Question 3: Am I supposed to have included interactions? If so which ones?

A: Not really, however, if you strongly suspect that two variables interact, you might want to check for that. However, in an exam setting make sure that you answer the main questions before attempting something that has not been asked explicitly. It might earn you some points but make sure it is not on the expense of other questions.

Q: Pe(0.1) - I know this means entry - but how does that work?

A: The way it works is that the statistical package will only add the variable to a model if its P-value is less than or equal to 0.1 when controlling for the other variables.

Q: Sample Assessment Document 1 Question 1:Cprplot - should these be done for each explanatory variable and are we looking for linear relationships?

A: Yes, usually they are done for each continuous explanatory variable to check for the functional form of the variable at hand.

Q: When I got home I found the other example of the likelihood-ratio test which was on Page 4 and 5 of the log of the regressionII file you created. This example had the nested model as A and the more complex model as B, whereas it was the other way round in the logistic regression example. So can I just check again how to interpret the results? If the p value of the likelihood ratio test is < 0.05 is this evidence for using the more complex model in preference to the one nested in it? And is this the case whether the more complex model is model A or model B?

A: I could not find exactly which output but no worries here is an example

```
. xi: logit cvd  age i.sex
. est store A
. xi: logit cvd  i.sex*age
. est store B
. lrtest A B
```

```
Likelihood-ratio test                          LR chi2(1)   =      2.32
(Assumption: A nested in B)                    Prob > chi2 =    0.1281
```

```
. lrtest B A

Likelihood-ratio test                                 LR chi2(1)  =      2.32
(Assumption: A nested in B)                           Prob > chi2 =    0.1281
```

In the above A is the smaller model and B is the bigger model (A is nested in B). With STATA being so clever these days (it was not always), it does not matter whether you use (lrtest A B) or (lrtest B A). For an example, see above. STATA will guess which one is nested in the other (based on the likelihood value) and will state that in its output (Assumption: A nested in B).

So based on the above example since P –value > 0.05, it means that the bigger model is not doing a better job than the smaller model and we will favour the smaller model in this case. In some other cases, even though there might not be statistical evidence for using the bigger model we might want to stick to it. For example, if in the literature age has been found to be a significant predictor you will want it to be in your model regardless of its statistical significance.

If the P-value of the lrtest is < 0.05 then this is evidence for using the more complex model (the bigger one).

Q: If we have our stata commands in an open word doc in the exam are we allowed to copy and paste them straight into stata in the exam?

A: I can see no reason why you should not be able to do so, however, remember that the name of the variables will definitely be different.

Q: A few questions relating to the second practical log.
- For the interaction question 4 how do you tell from the outputs (other than the graphs) that there is an interaction, because the regression output for .regress y x gp looks very similar to the interaction regression output .regress y x gp xgp? -

You can look at the P-value associated with xgp term since there is only one interaction term. You can also use (test xgp) after the regress command and check its P-value. This is particularly useful if you have more than two categories for the categorical variable, say 3. Then you can use (test xgp_2 xgp_3) which will enable you test whether the interaction between x and the variable gp as a whole is significant or not. Yet another way can be done using lrtest, so you save the estimates for the model without the interaction term say in A (est store A) and those with the interaction say in B (est store B) then you can compare the two using (lrtest A B).

Q: A few questions relating to the second practical log. For question 1 relating to the confidence bands how do you tune the graph to change the colours and add the title in Stata version 9?

A: You can add the option (clcolor(---)) where --- is the colour of your choice, here cl stands for connect line, in the options of the lines that you want to have that particular colour say red. So here is the command

```
. scatter y x || line fit x || (line confup x, clcolor(red)) || (line confdn x,clcolor(red))
```

You can add a title by using (title("Scatter plot of Y and X with Fitted values and Confidence Bands")) in the options of the command, that is use (,).

```
. scatter y x || line fit x || (line confup x, clcolor(red)) || (line confdn x, clcolor(red)), title("Scatter plot of Y and X with Fitted values and  Confidence Bands")
```

Q: How to interpret the significance output on the Stata t-tests. I assume 'Ho' means the null hypothesis and 'Ha' is the alternative hypothesis which assumes there is a difference between the means.

A: Spot on.

Q: So, when I am trying to find out whether the t-test is significant do I use the following output and is this one tailed or 2?

```
Ha: diff != 0 Pr(|T| > |t|) = 0.0000
```

A: Yes, this is the 2-tailed one , !=0 means different than zero.

Q: And what do the other 2 outputs mean?

```
Ha: diff < 0               Ha: diff > 0
Pr(T < t) = 0.0000          Pr(T > t) = 1.0000
```

A: These two are the one-sided ones, if you are looking for difference in one direction; in most cases we are usually interested in the two-sided one.

Q: Problem with logistic output based on the following commands

```
. xi: logit cvd i.sex age, or nolog
. predict e6,xb
. predict see6, stdp
. gen ule6= e6+1.96*see6
. gen lle6 = e6-1.96*see6
. scatter e6 age if statcig3==0, msymbol (o) || line ule6 lle6 age if statcig3==0, sort
```

A: You have executed the following logisitc command "xi: logit cvd i.sex age, or nolog" So when you want to look at the fitted values you have to graph these values versus age and sex (now since sex is a categorical variable then you need one set of fitted values for each gender). You can plot these fitted values on the same graph or you can do separate graphs. So here is what you could do. First generate the fitted(predicted) values, the associated SEs, and the associated Conf. Bands as you did using

```
predict e6,xb
predict see6, stdp
gen ule6= e6+1.96*see6
gen lle6 = e6-1.96*see6
```

(PS you can call e6, see6, ule6 and lle6 anything you like)

A. Separate graphs for each sex

```
*  scatter e6 age if sex==1, msymbol (o) || line ule6 lle6 age
  if sex==1, sort title("Males")
*  scatter e6 age if sex==2, msymbol (x) || line ule6 lle6 age
  if sex==2, sort title("Females")
```

B. Graphs for both sexes (a bit more messy)

```
* scatter e6 age if sex==1, msymbol (o) || line ule6 lle6 age if
  sex==1, sort || scatter e6 age if sex==2, msymbol (x) || line ule6
  lle6 age if sex==2, sort lpattern(dash_dot dash_dot)
```

Q: I am practising STATA and am confused as to which is the correct command to check for homoscedasticity - is it the rvfplot or rvpplot? In the book "Presenting medical statistics from proposal to publication" it states to check homoscedasticity by plotting the residuals against the predicted values, which is what rvpplot command does, but I have rvfplot written down from our practical session.

A: Fitted and predicted values are usually used interchangeably and it is these that we want to plot against the residuals in order to check homoscedasticity

rvfplot *graphs a *residual-versus-fitted plot, a graph of the residuals against the fitted values.

If we want to check if there is a systematic relationship between the error variance and one of the predictors(that is independent variables/covariates) we plot residual-versus-predictor.

rvpplot graphs a ***residual-versus-predictor*** plot (independent variable plot or carrier plot), a graph of the residuals against the specified predictor.

Q: I've been trying to re run some of the practicals and produced this graph from week 5, low birth weight, smokers. Very peculiar conf intervals and I can't seem to get a title. Any ideas what I've done wrong? Extract from my log commands below

```
. predict e6,xb
. predict see6,stdp
. gen ule6 = e6+1.96*see6
. gen lle6 = e6-1.96*see6
. gen ule6 = e6+1.96*see6
. gen lle6 = e6-1.96*see6
. scatter e6 age if smoke == 0, msymbol(o) || line ule6 lle6 age if smoke ==0, xlabel
(10(5)45) title ("non-smokers") sort
invalid '"non-smokers'
r(198);
```

A: It looks to me that there is a space between title and (; they should be stuck together title(

```
. scatter e6 age if smoke == 0, msymbol(o) || line ule6 lle6 age if smoke ==0,  xlabel
(10(5)45) title("non-smokers") sort
option xlabel() not allowed
r(198);
```

A: it looks to me that there is a space between xlabel and “(“. They should be stuck together “xlabel(“.

Q: Incidentally I can't find how to save the log file into word, there only seems to be a default save as you close the file,is this right?.

A: You cannot save the log file as word from STATA. You need to close the log file first (as a .log that is in text format) then open Microsoft word. Once you open a Microsoft word. Go to the file menu and choose open. In the Files of Type field choose All Files (.*)and then choose your log file. Then save as word document from the File Menu of Microsoft Word.

Q: what's the difference between predict fit and predict e1 xb? I get the same values

A: In this case, they are the same thing since xb (should be predict e1, xb) is the default option. So what you get is the predicted(fitted) values under the model.

Q: What is the advantage of likelihood ratios over chi square, or odds ratios?

A: The likelihood ratio (LR) follows a chi-squared distribution with df equal to the difference of parameters of the two models. This means that one model should be contained in the other (nested models). In the case of contingency tables LR and Pearson chi-square behave pretty much similarly, in general. However, note that if one of the cells is zero you will not be able to compute the LR. Now when you are deciding on model selection, the LR enables you to compare whether a number of variables are needed or not, in contrast to the z-values of the individual variables that only tell you whether a variable is a significant predictor or not. This is particularly important when we are considering the statistical significance of categorical variables with more than two categories as the z-values will tell you whether an indicator variable is significant but not the whole set. As for the odds ratio, it gives you an estimate of the effect of a given variable rather than just statistical significance which can be inferred from the corresponding CI. So which one of these to use will certainly depend on the question you are trying to answer.

Q: what does lincom do?

A: lincom stands for linear combination so it uses the coefficients of the variables from the last regression model (linear, logistic, Poisson, ...) and finds the combined effect. For example, if in your logistic regression model you wanted to compute the effect of somebody whose married and male (assume that single is the reference group of marital and female is the reference group of gender), you cannot read it from the coefficients displayed as they give the effect of the variable of interest with the others being held constant (lincom overcomes this problem). Assume that the model is given by log(odds of disease) = 1.4 + 1.3 gender - 1.8 marital + error term

log(odds disease for a married male) = 1.4 + 1.3 - 1.8 when you engage lincom by asking lincom gender + marital it will substitute the coefficient of gender and that of marital (1.3-1.8) and if you ask it will compute an OR. The OR in this case would be comparing a married male to an unmarried female if you want to do that. One of its important uses is when you have an interaction term and you want to calculate ORs and corresponding CIs. A Stata example,

```
. xi: logit  cvd i.edlin2 i.alcon3
i.edlin2          _Iedlin2_2-9        (naturally coded; _Iedlin2_2 omitted)
i.alcon3          _Ialcon3_0-2        (naturally coded; _Ialcon3_0 omitted)

Iteration 0:   log likelihood = -443.18153
Iteration 1:   log likelihood = -438.43748
Iteration 2:   log likelihood = -438.39264
Iteration 3:   log likelihood = -438.39263

Logistic regression                               Number of obs   =        992
                                                 LR chi2(3)      =       9.58
                                                 Prob > chi2     =     0.0225
Log likelihood = -438.39263                       Pseudo R2       =     0.0108

------------------------------------------------------------------------------
        cvd |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
 _Iedlin2_9 |  -.4365988   .1772192    -2.46   0.014    -.7839422   -.0892555
 _Ialcon3_1 |     -.2849   1.076802    -0.26   0.791    -2.395394    1.825594
 _Ialcon3_2 |  -.2718079   .2298415    -1.18   0.237    -.7222889    .1786731
      _cons |  -1.347367   .1203371   -11.20   0.000    -1.583223    -1.11151
------------------------------------------------------------------------------

. lincom  _Iedlin2_9+ _Ialcon3_1

( 1)  _Iedlin2_9 + _Ialcon3_1 = 0

------------------------------------------------------------------------------
        cvd |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        (1) |  -.7214988   1.083411    -0.67   0.505    -2.844945    1.401947
------------------------------------------------------------------------------

. display -.4365988-.2849
-.7214988

. lincom  _Iedlin2_9+ _Ialcon3_1, or

( 1)  _Iedlin2_9 + _Ialcon3_1 = 0

------------------------------------------------------------------------------
        cvd | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
```

```
-------------+----------------------------------------------------------------
         (1) |    .4860232    .5265627    -0.67   0.505     .0581375    4.063103
------------------------------------------------------------------------------
```

Q: there seems to be lots of diagnostic tools for logistic regression - which is best?

A: They look at different things (goodness of fit/discrimination/outliers/influential points), we usually investigate all of them to see how the model is doing in general unless a specific question is asked.