HOMEWORK #4
Solutions

1A. The equations are given below:

- Linear (log odds) form:

$$\log \frac{P(\text{exer3})}{1-P(\text{exer3})} = -0.5186 - 0.4268\ \text{exer3}$$

- Nonlinear (probability) form:

$$P(\text{exer3}) = \frac{\exp(-0.5186 - 0.4268\ \text{exer3})}{1+\exp(-0.5186 - 0.4268\ \text{exer3})} = \frac{1}{1+\exp(0.5186 + 0.4268\ \text{exer3})}$$

1B. The coefficient denoted `cons_` in the output is the intercept of the regression model. This gives the log odds of hypertension (`hypt` = 1) for an individual reporting exercising less than 3 times / week (i.e. `exer3` = 0). The coefficient for `exer3` is the log odds ratio comparing the odds of hypertension in individuals exercising ≥ 3 times weekly to the corresponding odds in those who report less exercise.

1C. The exponentiated coefficient and 95% confidence interval (0.653; 0.501, 0.851) are shown in the Stata output below:

```
. dis exp(-.4267843 )
.6526043

. dis exp(-.6918335)
.50065727

. dis exp(-.1617351)
.85066652
```

1D. The estimated odds ratio for `exer3` is 0.653, indicating that women in HERS who exercise at the higher level have approximately 35% lower odds of hypertension than women reporting less exercise.

1E. As discussed in lecture, the *z* in the output refers to the Wald statistic for the null hypothesis that the true coefficient for `exer3` is zero, and is calculated as the ratio of the regression coefficient to its standard error. The *P*-value for this test is 0.002, indicating rejection of the null hypothesis using the conventional 5% significance level. We can restate the null hypothesis and test results equivalently in terms of the odds ratio for `exer3`, and conclude that the true odds ratio is less than one at the same level of significance. Similarly, the 95% confidence limits for the true coefficient (odds ratio) exclude the null value of zero (one), corroborating the Wald test *P*-value and adding information in the HERS sample about the degree of uncertainty about the true odds ratio.

2A. The requested representations are given by the two equations below:

- Linear (log odds) form : $\log\frac{P(\text{bmi})}{1-P(\text{bmi})} = -1.7421 + 0.0376\ \text{bmi}$
- Nonlinear (probability) form : $P(\text{bmi}) = \frac{\exp(-1.7421 + 0.0376\ \text{bmi})}{1+\exp(-1.7421 + 0.0376\ \text{bmi})}$

2B. As before, the coefficient denoted `cons_` is the intercept of the regression model. This gives the log odds of hypertension (`hypt` = 1) for an individual with a body mass index of zero (`bmi` = 0). This is clearly not of practical interest, and a more interpretable result would be based on a model fitted to the *centered* value of `bmi` (i.e. `bmi` – mean(`bmi`)). In this case, the intercept term is interpreted as the log odds ratio for an individual with average BMI. This model is shown below (where `bmic` represents the centered version of `bmi`):

```
. logistic hypt bmic, coef

Logistic regression                                     Number of obs   =       1055
                                                        LR chi2(1)      =       9.35
                                                        Prob > chi2     =     0.0022
Log likelihood = -667.76922                             Pseudo R2       =     0.0070

------------------------------------------------------------------------------
        hypt |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        bmic |   .0376171    .012309     3.06   0.002     .0134918    .0617423
       _cons |  -.6934261    .065649   -10.56   0.000    -.8220958   -.5647564
------------------------------------------------------------------------------
```

The intercept term now corresponds to an outcome odds of exp(-0.6934261) = 0.50. (i.e. the odds of hypertension among individuals with average BMI is 1 in 2.) This odds estimate can be converted to an outcome probability (*P*) using the following relationship:

$\text{odds} = \frac{P}{1-P};\ P = \frac{\text{odds}}{1+\text{odds}}$

Notice that the coefficient (and corresponding odds ratio) for `bmi` are the same in both models. Thus, the original model suffices if interest is limited to the effect on outcome risk of changes in BMI. The coefficient for `bmi` (in either model) represents the change in the log odds of hypertension associated with a one-unit (in kg/m^2) increase in BMI.

2C. The exponentiated coefficient and 95% confidence interval (`1.0383336`; `1.0135832`, `1.0636882`) are shown in the Stata output below:

```
. dis exp(.0376171)
1.0383336
. dis exp(.0134918)
1.0135832
. dis exp(.0617423)
1.0636882
```

2D. The estimated odds ratio for `bmi` is 1.04, indicating that a one-unit increase in `bmi` (in

kg/m^2) is associated an approximate 4% increase in odds of hypertension.

2E. The estimated odds ratio for a 30-unit increase in `bmi` is given below:

```
. dis exp(.0376171*30)
3.0910567
```

2F. The derivation is given below, comparing the log odds of an individual with `bmi` = 50 to one with `bmi` = 20:

$$\log\frac{P(50)}{1-P(50)} - \log\frac{P(20)}{1-P(20)} = (-1.7421 + 0.0376 \times 50) - (-1.7421 + 0.0376 \times 20)$$

$$= 0.0376 \times 30 = 1.128; \quad \exp(1.128) = 3.09$$

2G. The desired "relative risk" is obtained by plugging the estimated coefficients and values for `bmi` directly into the expression for the outcome probability:

$$\frac{P(50)}{P(20)} = \frac{\frac{\exp(-1.7421 + 0.0376 \times 50)}{1 + \exp(-1.7421 + 0.0376 \times 50)}}{\frac{\exp(-1.7421 + 0.0376 \times 20)}{1 + \exp(-1.7421 + 0.0376 \times 20)}} = \frac{0.5344}{0.2709} = 1.973$$

Since we aren't given the necessary variability information about the coefficients in the logistic output, we can't compute a confidence interval for the estimated relative risk. A confidence would require an application of the `nlcom` command, or an alternative like bootstrap sampling. Here's the `nlcom` estimate:

```
. nlcom ((exp(_b[_cons ] + _b[bmi]*50)/(1+exp(_b[_cons ] + _b[bmi]*50)))) /
((exp(_b[_cons ] + _b[bmi]*20)/(1+exp(_b[_cons ] + _b[bmi]*20))))

       _nl_1:  ((exp(_b[_cons ] + _b[bmi]*50)/(1+exp(_b[_cons ] + _b[bmi]*50)))) /
((exp(_b[_cons ] + _b[bmi]*20)/(1+exp(_b[_cons] + _b[bmi]*20))))

------------------------------------------------------------------------------
        hypt |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _nl_1 |   1.973107   .3929321     5.02   0.000     1.202974     2.74324
------------------------------------------------------------------------------
```

Problem 3

Here is the Stata output for the desired model:

```
. logistic sta i.crn i.typ i.can i.loc i.agecat i.syscat
note: 1.loc != 0 predicts success perfectly
      1.loc dropped and 5 obs not used


Logistic regression                               Number of obs   =        195
                                                  LR chi2(10)     =      62.06
                                                  Prob > chi2     =     0.0000
Log likelihood = -60.742305                       Pseudo R2       =     0.3381

------------------------------------------------------------------------------
         sta | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       1.crn |   2.840999   2.165964     1.37   0.171     .6375571    12.65969
       1.typ |   52.63315   69.20641     3.01   0.003     3.999717     692.611
       1.can |   13.62877   13.28441     2.68   0.007     2.017263     92.0769
             |
         loc |
           1 |          1  (empty)
           2 |   16.38254   15.17629     3.02   0.003     2.665963     100.672
             |
      agecat |
           2 |   4.240071   3.244213     1.89   0.059     .9464375    18.99566
           3 |   5.588847   4.478942     2.15   0.032     1.161885    26.88322
           4 |   13.20398   9.938337     3.43   0.001     3.020121    57.72781
             |
      syscat |
           2 |   .7691673   .4733606    -0.43   0.670     .2302368    2.569608
           3 |   1.059359   .6293668     0.10   0.923     .3306305    3.394246
           4 |   .0768537   .0754065    -2.62   0.009     .0112329     .525821
             |
       _cons |   .0011421   .0018182    -4.26   0.000     .0000504    .0258683
------------------------------------------------------------------------------
```

3A. The interpretation of the estimated odds ratios is as follows:

1. The odds ratio for `crn` compares the odds of dying for individuals with a history of chronic renal failure with the corresponding odds for those with no such history. This estimate is adjusted for the other predictors in the model.

2. The odds ratio for `typ` compares the odds of death for individuals admitted for emergencies relative to the corresponding odds for elective admissions, also adjusted for the presence of the other included predictors.

3B. The predictor `loc` has three levels, yet only one odds ratio was estimated. Stata's warning message about this appears in the output above. An explanation is apparent when viewing the cross-tabulation of the outcome `sta` and the predictor `loc`:

```
. tab sta loc

    vital |
 status (1 | level of consciousness at admis
   = died) |         0          1          2 |     Total
-----------+---------------------------------+----------
         0 |       158          0          2 |       160
         1 |        27          5          8 |        40
-----------+---------------------------------+----------
     Total |       185          5         10 |       200
```

Since all five individuals with level 1 of `loc` were observed to die, the corresponding odds are infinite. Therefore, a log odds ratio can't be estimated for the logistic model (because an infinite OR doesn't make any sense, and there are no data to provide a realistic estimate) so these individuals are removed from the model fit. One way to retain these data in the model would be to collapse the categories of `loc` into two. The disadvantage with this approach is that collapsing categories may change the interpretation of other aspects of the model. In this case, it may be best to go with the default output, and note that five individuals couldn't be used.

3C. The trend test is given below in Stata 11 syntax, referring the table in Chapter 4:

```
. test -2.agecat+3.agecat+3*4.agecat==0

( 1)  - [sta]2.agecat + [sta]3.agecat + 3*[sta]4.agecat = 0

           chi2(  1) =   11.55
         Prob > chi2 =    0.0007
```

The results are consistent with a significant linear trend in the log mortality odds with increasing age.

Note that in Stata 12-13 we could also use the new contrast command as shown in the book:

```
. contrast q(1).agecat, noeffects

Contrasts of marginal linear predictions

Margins      : asbalanced
------------------------------------------------
             |         df        chi2     P>chi2
-------------+----------------------------------
      agecat |          1       11.55     0.0007
------------------------------------------------
```

3D. The following `lincom` statement asks for the desired OR:

```
. lincom 3.agecat - 2.agecat

 ( 1)  - [sta]2.agecat + [sta]3.agecat = 0

------------------------------------------------------------------------------
         sta | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   1.318102   .8756019     0.42   0.678     .3585127    4.846114
```

Refitting the model using the 2nd category of age as baseline would be another way to get this result. The following output is edited to include only the estimated ORs for `agecat`:

```
. logistic sta i.crn i.typ i.can i.loc ib2.agecat i.syscat

      agecat |
          1  |   .2358451   .1804525    -1.89   0.059     .0526436    1.056594
          3  |   1.318102   .8756019     0.42   0.678     .3585127    4.846114
          4  |   3.114093   1.864163     1.90   0.058     .9633494    10.06652
```

3E. The Stata code on the following page refits the full model again (suppressing output), saves the log-likelihood in the location `LF`, then fits each of the nested models excluding the predictors of interest (again suppressing output) and performs the requested LR tests.

```
. quietly logistic sta i.crn i.typ i.can i.loc i.agecat i.syscat

. estimates store LF

. quietly logistic sta i.crn i.typ i.can i.loc i.syscat

. lrtest LF

Likelihood-ratio test                                 LR chi2(3)  =     14.96
(Assumption: . nested in LF)                          Prob > chi2 =    0.0018

. quietly logistic sta i.crn i.typ i.can i.loc i.agecat

. lrtest LF

Likelihood-ratio test                                 LR chi2(3)  =     10.86
(Assumption: . nested in LF)                          Prob > chi2 =    0.0125
```

The first test result is for the contribution of `agecat`, the second for `syscat`. (Recall that the LR test evaluates the contribution of the variable(s) left out of the full model.)

O1. Possible Stata code for the requested program is reproduced below:

```
* Example simulation to investigate non-collapsibility of an OR
* estimated from a logistic model for a binary outcome Y,
* dependent on two binary predictors X1 and X2 that are
* independent (i.e. not confounders of each other's effects
* on Y). The program below takes the sample size "n" as well as
* regression coefficients (b0, b1 & b2) for the logistic model
* relating the outcome y to predictors x1 & x2 as arguments

program define lreg
args n b0 b1 b2
drop _all
set obs `n'

* generate a binary random variable x1 with prevalence 0.5
gen x1 = rbinomial(1,0.5)

* generate an indep. continuous random variable x2 with
* mean 0, SD 1
gen x2 = rnormal(0,1)

* generate a binary random variable y dependent on x1 & x2
* via a logistic model with coefficients b0, b1 & b2
gen px = 1/(1 + exp(-(`b0' + `b1'*x1 + `b2'*x2)))
gen y = rbinomial(1,px)

* fit adjusted model, save adjusted effect of x1
logistic y x1 x2, coef
matrix rb = r(table)
scalar badj = rb[1,1]

* fit marginal model, save b1 & b2 coefficients
logistic y x1, coef
matrix rb = r(table)
scalar bmar = rb[1,1]

end
```

The output below shows a few runs of the program using the `simulate` command with 1000 repetitions. The created variables `_sim_1` and `_sim_2` contain the simulated values for the conditional and marginal log odds ratios, respectively, so the `summarize` command is used to provide the mean values over the 1000 repetitions for each simulation.

```
. simulate badj bmar, nodots reps(1000): lreg 400 -0.85 0.7 -0.1

      command:  lreg 400 -0.85 0.7 -0.1
       _sim_1:  badj
       _sim_2:  bmar


. summarize _sim_1 _sim_2

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
      _sim_1 |      1,000    .7097447     .212815   .1003962    1.42337
      _sim_2 |      1,000    .7059824    .2119548   .1236945   1.420839

. simulate badj bmar, nodots reps(1000): lreg 400 -0.85 0.7 -0.5

      command:  lreg 400 -0.85 0.7 -0.5
       _sim_1:  badj
       _sim_2:  bmar


. summarize _sim_1 _sim_2

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
      _sim_1 |      1,000    .7003013    .2079314   .0898541   1.349149
      _sim_2 |      1,000    .6616466    .2002114   .0433823   1.309007

. simulate badj bmar, nodots reps(1000): lreg 400 -0.85 0.7 -0.7

      command:  lreg 400 -0.85 0.7 -0.7
       _sim_1:  badj
       _sim_2:  bmar


. summarize _sim_1 _sim_2

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
      _sim_1 |      1,000    .7201719      .22069   .0398158   1.361675
      _sim_2 |      1,000    .6489487    .2062387  -.0381939   1.267067
```

The results indicate that for stronger effects of x_2, the discrepancy between the conditional and marginal log odds ratios for x_1 increases (i.e. the non-collapsibility effect becomes more apparent). Since x_1 & x_2 are generated independently, this effect can't be explained by confounding.