Assignment 1 Key

Note that the results of Questions 1-3 do not change based on sample size since they relate to the truth.

The true model here is

$$logit(\mathbb{E}[Y|X_1 = x_1, X_2 = x_2]) = -2.0 + 2.0 * x_1 - 2.25x_2$$

Question 1

Note that the joint distribution $P(Y, X_1, X_2) = P(Y|X_1, X_2)P(X_1, X_2)$ and $P(X_1, X_2) = P(X_2|X_1)P(X_1)$. Thus $P(Y, X_1, X_2) = P(Y|X_1, X_2)P(X_2|X_1)P(X_1)$.

- X_1 is randomly simulated from a uniform distribution between 0 and 5.

  ```
  gen X1=5*runiform()
  ```

- $X_2 \mid X_1$ is simulated as $0.5 * X_1$ plus an error from a normal distribution with mean 0 and standard deviation 2, or $dist(X_2 \mid X_1) = N(0.5 * X_1, 2)$:

  ```
  gen X2=0.5*X1+rnormal(0,2)
  ```

- $Y \mid X_1, X_2$ is Bernoulli with $P(Y = 1 \mid X_1, X_2)$ defined by the logistic regression:

  $$Y \sim Bernoulli(p(x_1, x_2)) = P(Y = 1|X_1 = x_1, X_2 = x_2) = \frac{1}{1 + e^{-(-2.0+2.0*x_1-2.25*x_2)}}$$

  ```
  scalar b0 = -2
  scalar b1 = 2.0
  scalar b2 = -2.25
  gen logitPY = b0+b1*X1+b2*X2
  gen PY=1/(1+exp(-logitPY))
  gen Y = rbinomial(1, PY)
  ```

Question 2

Plug the values $X_1 = 0, X_2 = 1$ into the true regression equation and use the expit function to isolate the mean value of Y given covariates. Thus the predicted value of Y is $\mathbb{E}(Y|X_1 = 0, X_2 = 1) = \frac{1}{1+\exp(-(-2.0+2.0*0-2.25*1))} = 0.014$, or

```
. display 1/(1+exp(-(b0+b1*0+b2*1)))
.01406363
```

Question 3

We take the difference between the logit (log-odds) when $X_1 = x_1 + .5$ and $X_1 = x_1$ in the true regression.

$$logit(\mathbb{E}[Y|X_1 = x_1 + .5, X_2 = x_2]) - logit(\mathbb{E}[Y|X_1 = x_1, X_2 = x_2]) = \log\left(\frac{\mathbb{P}_{x_1+.5}/(1-\mathbb{P}_{x_1+.5})}{\mathbb{P}_{x_1}/(1-\mathbb{P}_{x_1})}\right)$$

$$= \{b0 + b1 * (x_1 + 0.5) + b2 * x_2\} - \{b0 + b1 * (x_1) + b2 * x_2\} = 0.5 * b_1 = 1$$

$$OR(x_1 + 0.5, x_1) = exp(1) = 2.7$$

Question 4

Note that your answers will vary based on the observations that were generated by your simulation so do not be alarmed if your numbers do not match!

Note that for the regression part, $E(Y|X_1, X_2)$, we do not need to specify the distribution of (X_1, X_2), since our parameter of interest (the coefficients) do not depend on the distribution of X_1, X_2. We just need to run logistic regression.

Simulation, n=100

```
. logit Y X1 X2


Logistic regression                             Number of obs   =        100
                                                LR chi2(2)      =      77.87
                                                Prob > chi2     =     0.0000
Log likelihood = -30.198525                     Pseudo R2       =     0.5632

------------------------------------------------------------------------------
           Y |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          X1 |   1.610765   .3853198     4.18   0.000     .8555519    2.365978
          X2 |  -1.727101   .3538075    -4.88   0.000    -2.420551   -1.033651
       _cons |   -2.14513   .7563672    -2.84   0.005    -3.627582   -.6626774
------------------------------------------------------------------------------
```

Using the estimates of the betas from logistic regression we have

$$\hat{E}(Y|X_1 = x_1, X_2 = x_2) = \frac{1}{1 + \exp(-(-2.14 + 1.61 * x_1 - 1.73 * x_2))}$$

The predicted value at $X_1 = 0, X_2 = 1$ is:

$$\hat{E}(Y|X_1 = x_1, X_2 = x_2) = \frac{1}{1 + \exp(-(-2.14 + 1.61 * 0 - 1.73 * 1))} = 0.020$$

or using Stata,

```
. matrix b = get(_b)
. matrix list b

b[1,3]
             Y:          Y:          Y:
            X1          X2       _cons
y1   1.6107648  -1.7271006  -2.1451298
. display 1/(1+exp(-(b[1,3]+b[1,1]*0+b[1,2]*1)))
.02038759
```

We can use *lincom* command to get the OR as follows:

```
. lincom 0.5*X1, or

 ( 1)  .5*[Y]X1 = 0

------------------------------------------------------------------------------
           Y | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   2.237552   .4310865     4.18   0.000     1.533842    3.264116
------------------------------------------------------------------------------
```

You just repeat these steps for the sample size of $n = 500$.

Question 5

- Coef: The estimated log-odds ratio for a unit increase in X_2, holding X_1 constant.
- Std. Err: the standard error, or the estimated standard deviation of $\hat{b}_2$.
- z: The z-statistic calculated as $z = \frac{\hat{b}_2}{se(\hat{b}_2)}$
- $P > |z|$: The two-sided p-value, or $P(|Z| > z)$, where $Z \sim N(0, 1)$.
- 95%CI: Assuming $\hat{b}_2$ is normally distributed, this interval has the property that, in repeated experiments of this kind, 95% of intervals constructed in this manner will contain the true value (b_2).

Question 6

The results for $n = 100$ are above. For $n = 500$ we get:

```
. logit Y X1 X2

Logistic regression                               Number of obs   =        500
                                                  LR chi2(2)      =     443.49
                                                  Prob > chi2     =     0.0000
Log likelihood = -124.42877                       Pseudo R2       =     0.6406

------------------------------------------------------------------------------
           Y |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          X1 |    1.94138    .212514     9.14   0.000      1.52486      2.3579
          X2 |  -2.372624   .2398227    -9.89   0.000    -2.842668    -1.90258
       _cons |  -1.759943   .3491303    -5.04   0.000    -2.444226    -1.07566
------------------------------------------------------------------------------
```

Just looking at the estimate of b_1 and the SE we get:

- for $n = 100$, $hatb_1 = 1.61(SE = 0.38)$.
- for $n = 500$, $hatb_1 = 1.95(SE = 0.21)$.

Thus, as expected, as the sample size goes up or precision about b_1 increases. Also, because our estimator (logistic regression) is unbiased, since the model we used is indeed the true model that generated the data, as sample size gets larger, the estimate will get (on average across repeated experiments) closer to the truth.