1. A sample of size four has \((x, y)\) equal to \((2, 4), (0, 0), (2, 2),\) and \((0, 2)\). We wish to estimate the model \(y = \beta_1 + \beta_2 x_2 + u\). We assume \(u \sim N[0, \sigma^2]\) and independent.

(a) Using the matrix algebra formula for the OLS estimator, compute the OLS coefficients \(\hat{\beta}_1\) and \(\hat{\beta}_2\) for these data.

(b) Compute the standard error for \(\hat{\beta}_2\) for these data.

(c) Compute \(R^2\) for these data, using any method you like (matrix algebra is not required).

(d)(i) Give the Stata command(s) that you would use to manually read the data for this example into Stata.

(ii) Give the Stata command that you would use to compute \(\hat{\beta}_1\) and \(\hat{\beta}_2\).

2. Consider the multiple regression model

\[
y = Z\gamma + v,
\]

where \(y\) is an \(N \times 1\) vector, \(Z\) is an \(N \times k\) matrix of constants, \(\gamma\) is a \(k \times 1\) vector and \(v\) is an \(N \times 1\) vector with \(v \sim N[0, \sigma^2I]\).

Consider the estimator \(\tilde{\gamma}\) that minimizes

\[
Q(\gamma) = v'Av,
\]

for nonsingular symmetric positive definite constant matrix \(A\).

(a) Show that \(\tilde{\gamma} = (Z'AZ)^{-1}Z'Ay\).

(b) Derive the variance of \(\tilde{\gamma}\).

(c) Derive \(E[\tilde{v}\tilde{v}']\), where \(\tilde{v} = y - Z\tilde{\gamma}\) and state whether \(E[\tilde{v}\tilde{v}'] = \sigma^2I\).
3. Consider the estimator

$$
\hat{\beta} = (Z'X)^{-1}Z'y
$$

in the multiple regression model

$$
y = X\beta + u,
$$

where \( y \) is an \( N \times 1 \) vector, \( X \) and \( Z \) are \( N \times k \) matrices, \( \beta \) is a \( k \times 1 \) vector and \( u \) is an \( N \times 1 \) vector.

We assume that the regressors \( X \) and \( Z \) are random with components \( x_i \) and \( z_i \) that are iid.

We assume that the error components \( u_i \) are iid with mean 0 and variance \( \sigma^2 \) and that \( u \) is independent of \( X \) or \( Z \).

For parts (a)-(d) there is no need to explicitly use laws of large numbers and central limit theorems. Instead state explicitly any additional assumptions that you have made.

(a) Show whether or not \( \hat{\beta} \) is unbiased for \( \beta \).
Your answer here should account for \( X \) and \( Z \) being stochastic.

(b) Show whether or not \( \hat{\beta} \) is consistent for \( \beta \).

(c) Obtain the limit distribution of \( \sqrt{N}(\hat{\beta} - \beta) \).

(d) Give the asymptotic distribution for \( \hat{\beta} \).
Your answer should depend only on \( \sigma^2, Z \) and \( X \).

(e) Suppose we have the same setup as above, except now \( k = 1 \) so there is just one regressor. Apply a law of large numbers to obtain the probability limit of \( N^{-1}\sum_i z_iu_i \).
Provide as much detail as possible, including verifying that a LLN applies.

(f) Suppose we have the same setup as above, except now \( k = 1 \) so there is just one regressor. Apply a central limit theorem to obtain the limit distribution of \( N^{-1/2}\sum_i z_iu_i \).
Provide as much detail as possible, including verifying that a CLT applies.

4. Consider testing \( q \) linear restrictions of the form

$$
H_0 : R\beta = r
$$

in the linear regression model.

(a) Suppose we have a model with four regressors (the first of which is an intercept). Give \( R \) and \( r \) in each of the following situations

(i) \( H_0 : \beta_2 = \beta_3 + 2 \).
(ii) \( H_0 : \beta_2 = 0, \beta_3 = 0, \beta_4 = 0 \).

(b) Suppose we have an estimator \( \tilde{\beta} \sim \mathcal{N}[\beta, \Sigma] \).
Derive the chisquared form of the Wald test statistic of \( H_0 : R\beta = r \).
5. Consider the following Stata input and output

```stata
program simols, rclass
    drop _all
    set obs 5
    gen x = 1 + 2*invnorm(uniform())
gen y = 1 + 2*x + (invnorm(uniform()))^2 - 1
    regress y x
    return scalar theta = _b[x]
end
simulate "simols" theta=r(theta), reps(1000)
summarize, detail
```

with results

```
. summarize, detail
    r(theta)

-------------------------------------------------------------
Percentiles Smallest
1%  .3500591  -.1988292
5%  1.289173  -.1873094
10% 1.534989  -.0204161  Obs 1000
25% 1.823596  .0668301  Sum of Wgt. 1000
50% 2.006446  Mean 2.016591
    Largest Std. Dev. .6211563
75% 2.177897  4.594079 Variance .3858351
90% 2.460031  4.857336 Skewness 6.850154
95% 2.763496  4.769485 Kurtosis 130.4659
99% 3.622856  13.72522

(a) Give the data generating process in as much detail as possible.
(b) Does the simulation setup correspond to regressors fixed in repeated samples or does it correspond to random sampling of both y and x? Explain. [This is tricky. You need to think hard about this.]
(c) Are the simulation results what you expect? Explain in as much detail as possible.
```
6. Consider the attached Stata output that uses data on family income and food expenditure from the 1998 Consumer Expenditure Survey.

**Dependent Variable**

Food = Annual family expenditure on food in 1998 in dollars.

**Regressors**

Income = Annual family income in 1998 in dollars.

Incomesq = Square of Income.

Famsize = number of members in family (averaged over year so can be non-integer)

Drural = 1 if live in rural (non-urban) area and 0 if urban

(a) Are the regressors jointly statistically significant at level 0.05? Explain.

(b) Does food expenditure appear to be linear in income? Perform two appropriate statistical significance tests at level 0.05.

(c) Test at level 0.05 the claim that one more family member is associated with a $700 increase in food expenditures. State clearly details of the test and your conclusion.

[ASIDE: I had meant to ask this as a one-sided test question.]

```
summarize food income incomesq famsize drural
```

```
<table>
<thead>
<tr>
<th>Variable</th>
<th>Obs</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>food</td>
<td>700</td>
<td>4897.343</td>
<td>2855.653</td>
<td>585</td>
<td>20765</td>
</tr>
<tr>
<td>income</td>
<td>700</td>
<td>41118.78</td>
<td>33096.46</td>
<td>150</td>
<td>192123.1</td>
</tr>
<tr>
<td>incomesq</td>
<td>700</td>
<td>2.78e+09</td>
<td>4.69e+09</td>
<td>22500</td>
<td>3.69e+10</td>
</tr>
<tr>
<td>famsize</td>
<td>700</td>
<td>2.469429</td>
<td>1.414257</td>
<td>1</td>
<td>9</td>
</tr>
<tr>
<td>drural</td>
<td>700</td>
<td>.1285714</td>
<td>.3349643</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>
```

```
regress food income incomesq famsize drural
```

```
<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>Number of obs = 700</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>2.4890e+09</td>
<td>4</td>
<td>622261100</td>
<td>F( 4, 695) = 134.68</td>
</tr>
<tr>
<td>Residual</td>
<td>3.2111e+09</td>
<td>695</td>
<td>4620327.67</td>
<td>Prob &gt; F = 0.0000</td>
</tr>
<tr>
<td>Total</td>
<td>5.7002e+09</td>
<td>699</td>
<td>8154752.7</td>
<td>R-squared = 0.4367</td>
</tr>
<tr>
<td></td>
<td>Adj R-squared = 0.4334</td>
<td>Root MSE = 2149.5</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
```

```
food   | Coef. | Std. Err. | t   | P>|t|  | [95% Conf. Interval] |
-------|-------|-----------|-----|------|----------------------|
income | 0.0425326 | 0.0071027 | 5.99 | 0.000 | 0.0285873 0.0564779 |
incomesq| -5.47e-08 | 4.87e-08 | -1.12 | 0.262 | -1.50e-07 4.09e-08 |
famsize | 805.701 | 61.51642 | 13.10 | 0.000 | 684.9207 926.4813 |
drural | -14.3438 | 246.752 | -0.06 | 0.954 | -498.8126 470.125 |
_cons  | 1313.055 | 212.3991 | 6.18 | 0.000 | 896.0338 1730.076 |
```