Economics 471: Econometrics

Department of Economics, Finance and Legal Studies
University of Alabama

Course Packet

The purpose of this packet is to show you one particular dataset and how it is used in practice with the methods taught in lecture. The data used in the packet are available on the course website

https://dhenderson.people.ua.edu/uploads/1/2/3/4/123460056/1600.xls

A brief description of the variables is available on the course website

https://dhenderson.people.ua.edu/uploads/1/2/3/4/123460056/eh3733_des.pdf

You should be able to replicate all of the results using the econometric package, Gretl (http://gretl.sourceforge.net/). This should help bridge the material from class with the computer program and at the same time help you with the assigned problem sets. Brief explanations on how to obtain the various tables and figures within the Gretl window framework are given within the packet.

The academic paper from which this subset of data was collected can be found at

http://dx.doi.org/10.1111/j.1368-423X.2008.00244.x

You are not required to read or understand the material in the paper, but you may be interested in the motivation for choosing this particular data. You must be on campus to access the information in the above link. A related paper of interest is here:

http://dx.doi.org/10.1016/j.econedurev.2011.03.011

This packet is designed to be a tool which will help you understand the difficult material in the course. You are expected to bring this to each lecture as I will often refer to the results given in the packet. Questions, comments and suggestions are always appreciated.

Descriptive Statistics

Sample Average

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

Sample Standard Deviation

$$\hat{\sigma}_y = \sqrt{\hat{\sigma}_y^2} = \left(\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2\right)^{1/2}$$

	Mean	Median	Minimum	Maximum
testscores	52.435	52.770	31.920	71.930
homework	0.64754	0.50000	0.00000	5.7500
hrsofclass	3.9930	4.1667	0.00000	9.1667
classsize	23.342	24.000	1.0000	89.000
prevtestscores	52.656	51.600	34.380	77.200

	Std. Dev.	C.V.	Skewness	Ex. kurtosis
testscores	9.5666	0.18245	-0.078154	-0.97185
homework	0.39249	0.60611	5.0171	51.337
hrsofclass	0.99277	0.24863	-1.7804	4.2670
classsize	7.0785	0.30325	1.4664	10.747
prevtestscores	9.9067	0.18814	0.33088	-0.77401

	5% perc.	95% perc.	IQ range	Missing obs.
testscores	36.674	67.320	15.165	0
homework	0.25000	1.0000	0.25000	0
hrsofclass	0.91667	5.0000	0.83333	0
classsize	12.000	33.000	8.0000	0
prevtestscores	38.220	70.060	15.660	0

Gretl Steps:

View → Summary Statistics → Include: testscores, homework, hrsofclass, classsize, and prevtestscores → Ok → Show full statistics → Ok

Conditional Expectation

$$E(Y|X = x)$$

testscores if sex=0 = 52.896
testscores if sex=1 = 52.003

Interpretation: Tenth grade boys (sex=0) scored higher, on average, than 10th grade girls

Gretl Steps:

Click on testscores → Sample → Restrict, based on criteria → Enter boolean condition for selection cases: sex=0
After observations are dropped, right click on testscores → summary statistics

Then, click on testscores → Sample → Restore full range and repeat for sex=1

Sample Covariance

$$\hat{\sigma}_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

$$= \frac{1}{3733-1}\sum_{i=1}^{3733}(x_i - 0.647544)(y_i - 52.43538) = 0.6177\ldots$$

Sample covariance between homework and testscores = 0.617928

Gretl Steps:

Open new script (in bottom "ribbon", second from left) → enter code: q1_d= cov(testscores,homework) → highlight and run (control+R)

Sample Correlation Coefficient

$$\hat{\rho}_{XY} = \frac{\hat{\sigma}_{XY}}{\hat{\sigma}_X \hat{\sigma}_Y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(\sum_{i=1}^{n}(x_i - \bar{x})^2)^{1/2}(\sum_{i=1}^{n}(y_i - \bar{y})^2)^{1/2}}$$

$$= \frac{\sum_{i=1}^{3733}(x_i - 0.647544)(y_i - 52.43538)}{(\sum_{i=1}^{3733}(x_i - 0.647544)^2)^{1/2}(\sum_{i=1}^{n}(y_i - 52.43538)^2)^{1/2}} = 0.16457$$

```
corr(testscores, homework) = 0.16457221
Under the null hypothesis of no correlation:
 t(3731) = 10.1913, with two-tailed p-value 0.0000
```

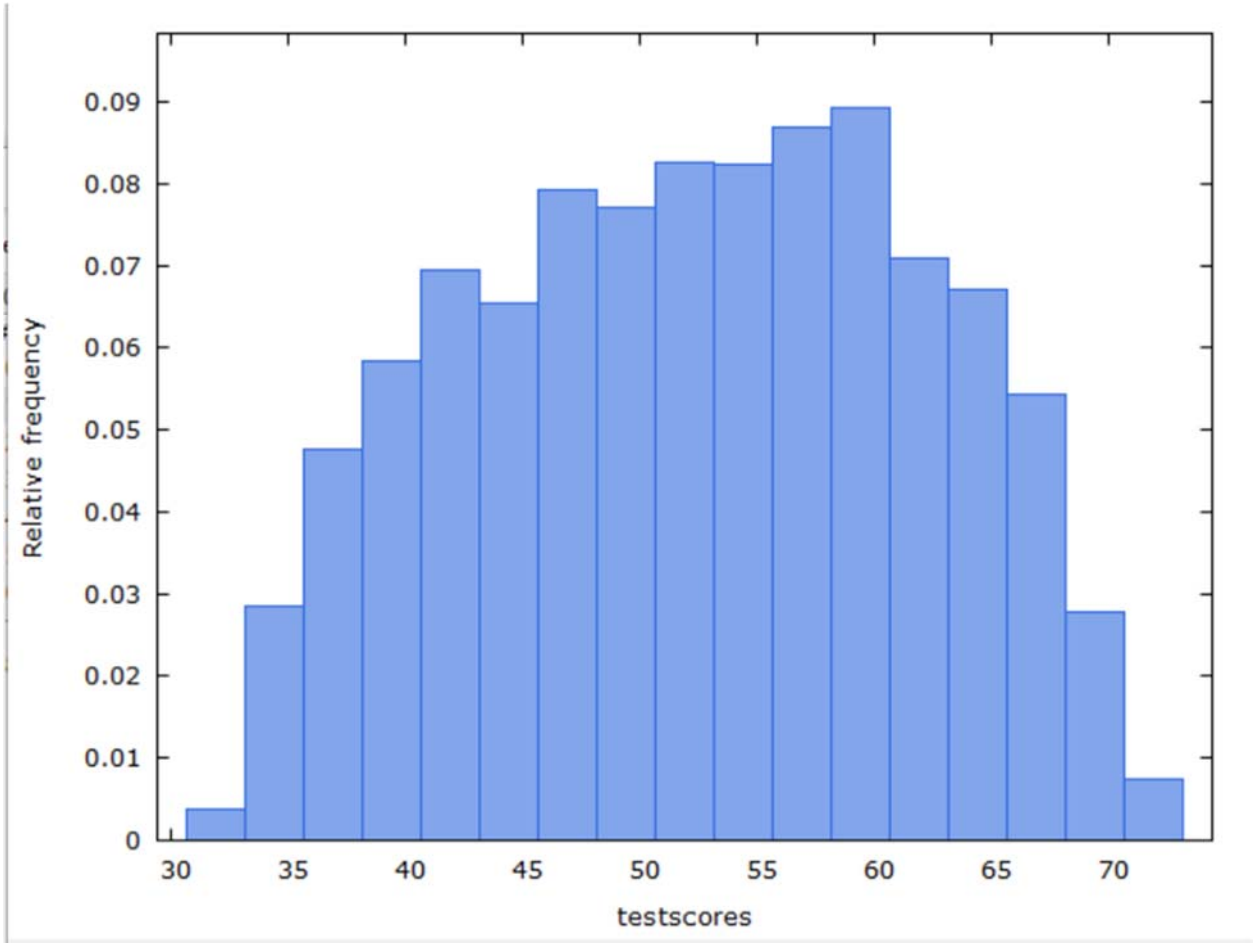
Interpretation: Both the covariance and correlation measure the linear dependence between the variables.

Gretl Steps:

Select testscores and homework → right click → Correlation matrix

Histograms, PDFs, and CDFs

Histogram

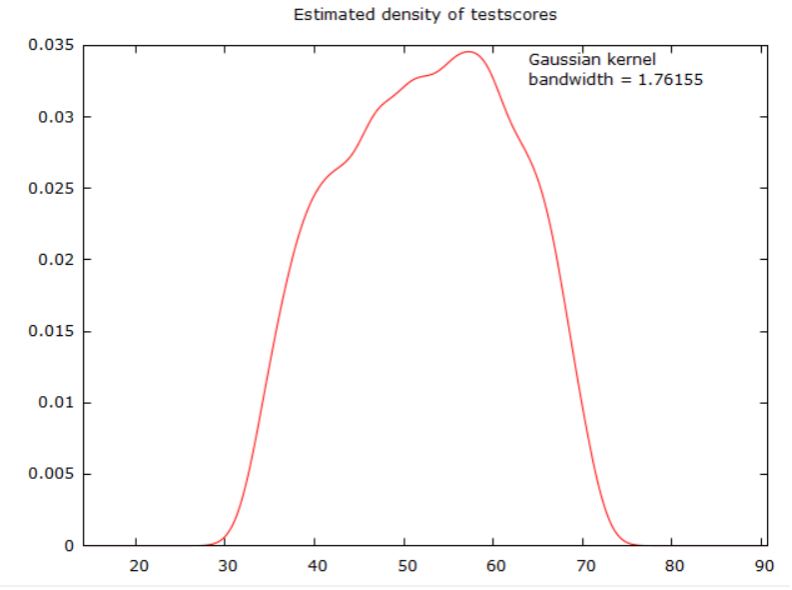


Gretl Steps:

Right click on testscores → Frequency distribution → Change Number of Bins to the desired number of "blocks" (in this case, 17) → Ok

Probability Density Function (pdf)

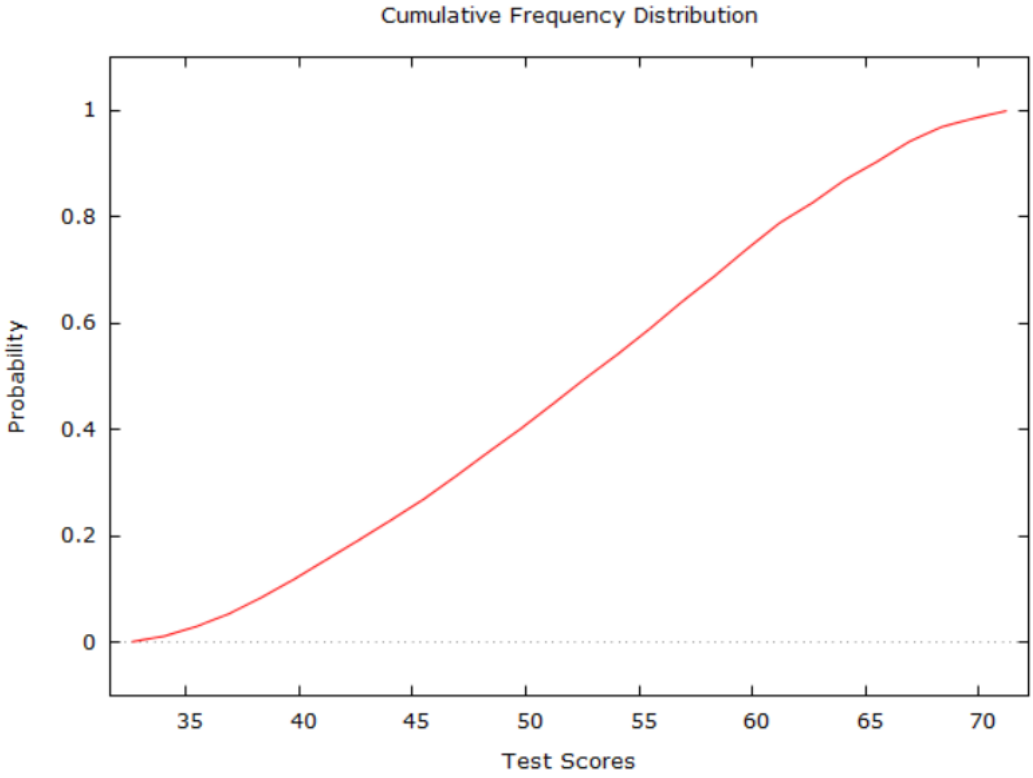
$$f(x_j) = p_j, j = 1, 2, \ldots, k$$



Gretl Steps:

Select testscores → Variable → Estimated density plot → Gaussian kernel with bandwidth adjustment factor = 1.06 → Ok

Cumulative Distribution Function (CDF)



Gretl Steps:

From PDF output box →Save interval midpoint and the cumulative frequency as new variables → View → Graph Specified Vars → X-Y Scatter → X-Axis: interval, Y-Axis: cumulative frequency → Ok → Edit → Lines → Change Points to Lines

Simple Linear Regression Model (Regression of *y* on *x*)

$$y_i = \alpha + \beta x_i + u_i$$

$$\hat{y}_i = \hat{\alpha} + \hat{\beta} x_i$$

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})y_i}{SST_x}$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

Where $SST_x = \sum_{i=1}^{n}(x_i - \bar{x})^2$ is the total variation in x.

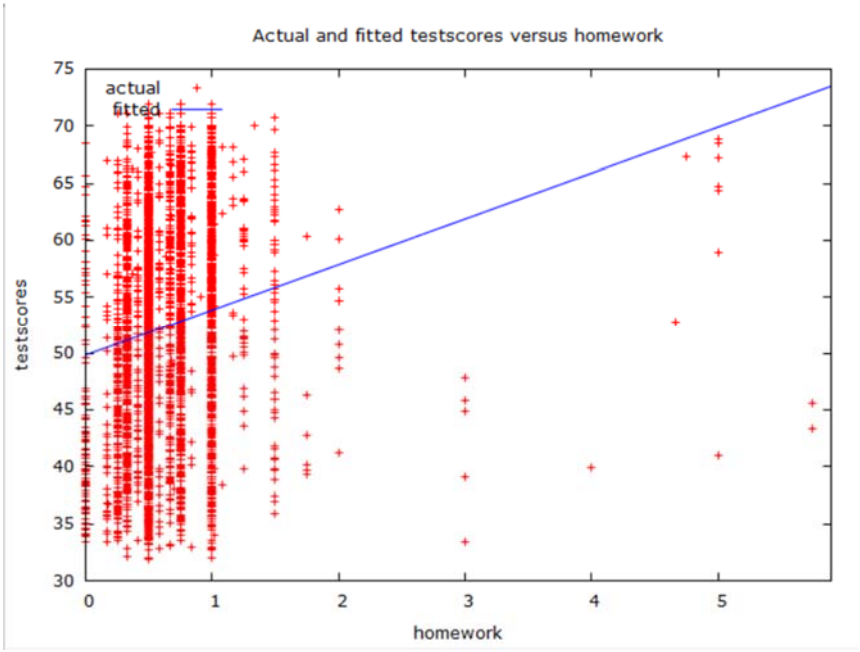
```
Model 1: OLS, using observations 1-3733
Dependent variable: testscores

              coefficient   std. error   t-ratio    p-value
  ---------------------------------------------------------
  const        49.8379       0.298027    167.2     0.0000    ***
  homework      4.01135      0.393603     10.19    4.47e-024 ***

Mean dependent var   52.43538   S.D. dependent var   9.566599
Sum squared resid    332301.3   S.E. of regression   9.437423
R-squared            0.027084   Adjusted R-squared   0.026823
F(1, 3731)           103.8635   P-value(F)           4.47e-24
Log-likelihood      -13675.30   Akaike criterion     27354.60
Schwarz criterion    27367.05   Hannan-Quinn         27359.03
```

Interpretation: The constant, 49.8379, indicates that if a student does no homework, they would earn an expected score of 49.8379 on the test. The homework coefficient indicates that each additional hour of homework increases the expected test score by 4.01135 points.

Gretl Steps:

Model→ Dependent Variable: testscores → Regressors: const, homework → Ok

Plot



Gretl Steps:

In previous model "dialog box" → Graphs → Fitted, actual plot → Against homework

Standard Error of $\widehat{\beta}$

$$se(\hat{\beta}) = \frac{\hat{\sigma}}{\sqrt{SST_x}} = \frac{\hat{\sigma}}{(\sum_{i=1}^{n}(x_i - \bar{x})^2)^{1/2}}$$

$$= \frac{9.437423}{\left(\sum_{i=1}^{3733}(x_i - 0.647544)^2\right)^{1/2}} = 0.3936$$

The standard error for the slope parameter is .393603.

Standard Error of Regression (SER)

$$\hat{\sigma} = \left(\frac{1}{n-k-1}\sum_{i=1}^{n}\hat{u}_i^2\right)^{\frac{1}{2}} = \left(\frac{SSR}{n-k-1}\right)^{\frac{1}{2}}$$

$$= \left(\frac{332301.3}{3733-2}\right)^{\frac{1}{2}} = 9.43742\ldots$$

The Standard Error of Regression, which is 9.437423, estimates the standard deviation in y after the effect of x has been accounted for.

Sum Squared Residuals (SSR)

$$SSR \equiv \sum_{i=1}^{n}\hat{u}_i^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y_i - \hat{\alpha} - \hat{\beta}x_i)^2$$

$$= \sum_{i=1}^{3733}(testscores - 49.83785 - 4.011346 * homework_i)^2 = 332301.3$$

The sum squared residuals, which is 332301.3, measures the unexplained variation between the data and the predicted values.

Explained Sum of Squares (SSE)

$$\boxed{SSE \equiv \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}$$

$$= \sum_{i=1}^{3733} (49.83785 + 4.011346 * homework_i - 52.43538)^2 = 9250.598$$

The explained sum of squares measures the explained variation between the predicted values and the mean.

Total Sum of Squares (SST)

$$\boxed{SST \equiv \sum_{i=1}^{n} (y_i - \bar{y})^2}$$

$$\sum_{i=1}^{3733} (y_i - 52.43538)^2 = 341551.898$$

The total sum of squares measures the total variation in y. It is the sum of the explained and unexplained variation.

$$\boxed{SST = SSE + SSR}$$

R-Squared

$$R^2 \equiv \frac{SSE}{SST} = 1 - \frac{SSR}{SST} = 1 - \left(\frac{SSR}{n}\right)\left(\frac{SST}{n}\right)$$

$$= \frac{SSE}{SST} = \frac{9250.598}{341551.898} = 0.027084 \ldots$$

R-squared, which is .027084, measures the proportion of sample variation in the dependent variable that is accounted for by the independent variable(s) in the model. This value never decreases, and usually increases, when adding more regressors.

Adjusted R-Squared

$$\bar{R}^2 \equiv 1 - \frac{\frac{SSR}{n-k-1}}{\frac{SST}{n-1}} = 1 - \frac{\hat{\sigma}^2}{\frac{SST}{n-1}} = 1 - \frac{(1-R^2)(n-1)}{n-k-1}$$

$$= 1 - \frac{(1-0.027084)(3733-1)}{3733-1-1} = 0.026823 \ldots$$

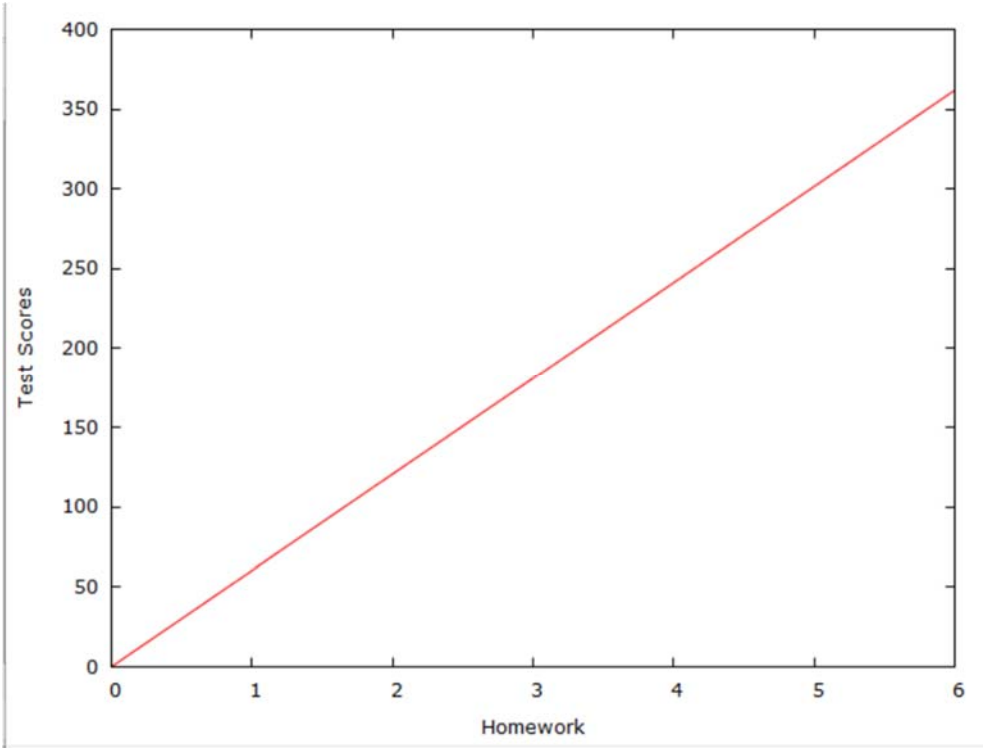
The adjusted R-squared, which is .026823, is similar to R-squared, but imposes a penalty for adding more dependent variables.

Regression through the Origin

$$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$$

$$testscores = \beta homework + u$$

```
Model 1: OLS, using observations 1-3733
Dependent variable: testscores

              coefficient   std. error   t-ratio   p-value
  ---------------------------------------------------------
  homework      60.3017      0.594503     101.4    0.0000  ***

Mean dependent var   52.43538   S.D. dependent var   9.566599
Sum squared resid     2822953   S.E. of regression   27.50306
Uncentered R-squared 0.733817   Centered R-squared  -7.265076
F(1, 3732)           10288.44   P-value(F)           0.000000
Log-likelihood      -17668.67   Akaike criterion     35339.34
Schwarz criterion    35345.56   Hannan-Quinn         35341.55
```



Interpretation: This model is used if we have reason to believe that the intercept is 0, e.g., if a student does no homework, then they would earn a 0 on the test. Because the coefficient for homework is 60.3, this indicates that each additional hour of homework boosts their expected

test score by 60.3 points. R-squared is meaningless when the intercept is excluded from the model.

Gretl Steps:

Model → Ordinary Least Squares → Dependent variable: testscores, Regressors: homework (const is NOT included) → Ok

For graph:
*Tools → Plot a Curve → formula: y=60.3017*x → x range = 5 → Ok*

Functional Forms with Logs

Model	Dependent Variable	Independent Variable	Interpretation of β
level-level	y	x	$\Delta y = \beta \Delta x$
level-log	y	log(x)	$\Delta y = (\beta/100)\%\Delta x$
log-level	log(y)	x	$\%\Delta y = (100\beta)\Delta x$
log-log	log(y)	log(x)	$\%\Delta y = \beta\%\Delta x$

Level-Log

$$\widehat{testscores} = \alpha + \beta \ln(classsize) + u$$

```
Model 2: OLS, using observations 1-3733
Dependent variable: testscores

                coefficient   std. error   t-ratio    p-value
  ------------------------------------------------------------
  const          43.9300       1.46569     29.97     4.87e-177 ***
  logclasssize    2.74231      0.469888     5.836     5.80e-09  ***

Mean dependent var   52.43538   S.D. dependent var   9.566599
Sum squared resid    338462.1   S.E. of regression   9.524505
R-squared            0.009046   Adjusted R-squared   0.008781
F(1, 3731)           34.05998   P-value(F)           5.80e-09
Log-likelihood      -13709.59   Akaike criterion     27423.17
Schwarz criterion    27435.62   Hannan-Quinn         27427.60
```

Interpretation: A one percent increase in class size increases expected test scores by 0.0274231 points.

Gretl Steps:

To create the new variable:
Right click on home screen → Define new variable → logclasssize=ln(classsize) → Ok

To create the model:
Model → Ordinary Least Squares → Dependent variable: testscores, Regressors: const, logclasssize → Ok

Log-Level

$$\ln(testscores) = \alpha + \beta classsize + u$$

```
Model 3: OLS, using observations 1-3733
Dependent variable: logtestscores

              coefficient   std. error    t-ratio   p-value
  ---------------------------------------------------------
  const       3.89087       0.0106293     366.1     0.0000   ***
  classsize   0.00219784    0.000435777     5.043   4.79e-07 ***

Mean dependent var   3.942173   S.D. dependent var   0.189056
Sum squared resid    132.4870   S.E. of regression   0.188440
R-squared            0.006772   Adjusted R-squared   0.006505
F(1, 3731)           25.43685   P-value(F)           4.79e-07
Log-likelihood       934.3802   Akaike criterion    -1864.760
Schwarz criterion   -1852.311   Hannan-Quinn        -1860.332
```

Interpretation: Each additional student added to a class will increase expected test scores by .2198%, which represents the semi-elasticity of test scores with respect to class size.

Gretl Steps:

Right click → Define new variable → logtestscores=ln(testscores) → Ok → Model → Ordinary Least Squares → Dependent variable: logtestscores, Regressors: const, classsize → Ok

Log-Log

$$\ln(testscores) = \alpha + \beta \ln(classsize) + u$$

```
Model 4: OLS, using observations 1-3733
Dependent variable: logtestscores

               coefficient   std. error   t-ratio    p-value
  ------------------------------------------------------------
  const          3.75648     0.0289361    129.8     0.0000    ***
  logclasssize   0.0598706    0.00927664     6.454   1.23e-010 ***

Mean dependent var   3.942173   S.D. dependent var   0.189056
Sum squared resid    131.9176   S.E. of regression   0.188035
R-squared            0.011041   Adjusted R-squared   0.010776
F(1, 3731)           41.65291   P-value(F)           1.23e-10
Log-likelihood       942.4203   Akaike criterion    -1880.841
Schwarz criterion   -1868.391   Hannan-Quinn        -1876.412
```

Interpretation: A one percent increase in class size will increase expected test scores by .059%. This can be interpreted as the elasticity of test scores with respect to class size.

Gretl Steps:

Right click → Define new variable → logclasssize=ln(classsize) → Ok
Right click → Define new variable → logtestscores=ln(testscores) → Ok

Model → Ordinary Least Squares → Dependent Variable: logtestscores, Regressors: const, logclasssize

Multiple Regression

$$testscores = \alpha + \beta_1 homework + \beta_2 classsize + \beta_3 hrsofclass + u$$

Model 5: OLS, using observations 1-3733
Dependent variable: testscores

	coefficient	std. error	t-ratio	p-value	
const	48.9805	0.820671	59.68	0.0000	***
homework	3.94193	0.392992	10.03	2.21e-023	***
classsize	0.0947537	0.0218489	4.337	1.48e-05	***
hrsofclass	-0.327947	0.155625	-2.107	0.0352	**

Mean dependent var	52.43538	S.D. dependent var	9.566599
Sum squared resid	330350.3	S.E. of regression	9.412200
R-squared	0.032796	Adjusted R-squared	0.032018
F(3, 3729)	42.14827	P-value(F)	8.86e-27
Log-likelihood	-13664.31	Akaike criterion	27336.61
Schwarz criterion	27361.51	Hannan-Quinn	27345.47

Interpretation: With 0 hours of homework, a class size of 0, and 0 hours of class, a student would be expected to score a 48.9805 on their test. Holding all else constant, each addition hour of homework would increase their expected test score by 3.94193 points, each additional one-unit increase in class size will increase their expected test score by 0.0947537 points, and each additional one-unit increase in hours of class will lower their expected test score by 0.327947 points.

Gretl Steps:

Model → Ordinary Least Squares → Dependent Variable: testscores, Regressors: const, homework, classsize, hrsofclass → Ok

Standard Error of $\widehat{\beta}$

$$se(\hat{\beta}_j) = \frac{\hat{\sigma}}{\sqrt{SST_j(1-R_j^2)}} = \frac{\hat{\sigma}}{\left((1-R_j^2)\sum_{i=1}^{n}(x_i-\bar{x})^2\right)^{1/2}}$$

Where $SST_j = \sum_{i=1}^{n}(x_{ij}-\bar{x}_j)^2$ for j = 1, 2, … , k and

R_j^2 is the R^2 from the regression of x_j on the other x′s

Quadratic Model

Estimated Slope

$$\frac{\partial \hat{y}}{\partial x} = \hat{\beta}_1 + 2\hat{\beta}_2 x$$

$$testscores = \alpha + \beta_1 homework + \beta_2 homework^2 + u$$

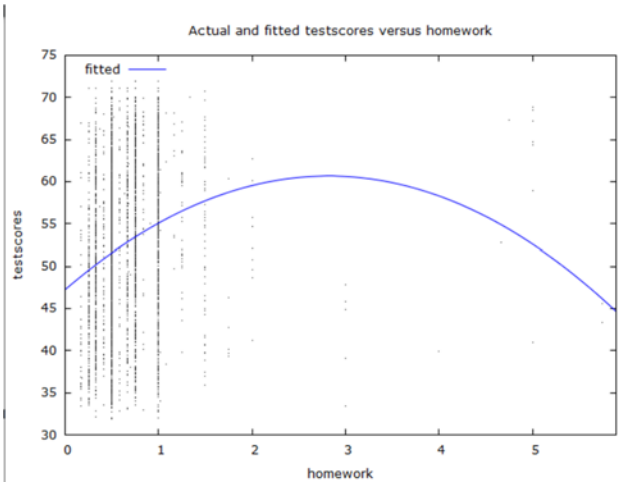
```
Model 6: OLS, using observations 1-3733
Dependent variable: testscores

              coefficient   std. error   t-ratio    p-value
  ---------------------------------------------------------
  const         47.2324      0.412152    114.6     0.0000    ***
  homework       9.53263     0.723923     13.17    9.69e-039 ***
  homeworksq    -1.69160     0.186972     -9.047   2.31e-019 ***

Mean dependent var   52.43538   S.D. dependent var   9.566599
Sum squared resid    325165.7   S.E. of regression   9.336797
R-squared            0.047976   Adjusted R-squared   0.047465
F(2, 3730)           93.98406   P-value(F)           1.51e-40
Log-likelihood      -13634.78   Akaike criterion     27275.56
Schwarz criterion    27294.24   Hannan-Quinn         27282.21
```

Interpretation: A curve that is concave down (e.g., has a negative square term) shows a relationship where each additional hour of homework continues to increase a student's expected test score up to a certain point, at which each additional hour of homework will actually lower their expected test score.

Gretl Steps:

Right click → Define new variable →homeworksq=(homework)^2
Model → Ordinary Least Squares → Dependent variable: testscores, Regressors: const, homework, homeworksq → Ok

Plot



Gretl Steps:

In previous model "dialog box" → Graphs → Fitted, actual plot → Against homework

Turning Point

$$x^* = \left|\frac{\hat{\beta}_1}{2\hat{\beta}_2}\right|$$

$$= \left|\frac{9.532628}{2(-1.691596)}\right| = 2.8176$$

Interpretation: The turning point represents the point at which the return to homework begins to diminish. For this data, if we restrict the sample to students who do more than 2.8176 hours of homework per day, there are only 17 sample points (out of 3733 total observations).

Standardized Beta Coefficients

$$\frac{(y_i - \bar{y})}{\hat{\sigma}_y} = \frac{\hat{\sigma}_1}{\hat{\sigma}_y}\hat{\beta}_1\frac{(x_{i1} - \bar{x}_1)}{\hat{\sigma}_1} + \frac{\hat{\sigma}_2}{\hat{\sigma}_y}\hat{\beta}_2\frac{(x_{i2} - \bar{x}_2)}{\hat{\sigma}_2} + \cdots + \frac{\hat{\sigma}_k}{\hat{\sigma}_y}\hat{\beta}_k\frac{(x_{ik} - \bar{x}_k)}{\hat{\sigma}_k} + \frac{\hat{u}_i}{\hat{\sigma}_y}$$

$$z_y = \hat{b}_1 z_1 + \hat{b}_2 z_2 + \cdots + \hat{b}_k z_k + \frac{\hat{u}_i}{\hat{\sigma}_y}$$

$$\text{Where } \hat{b}_j z_j = \frac{\hat{\sigma}_j}{\hat{\sigma}_y}\hat{\beta}_j \text{ for all j} = 1, 2, \ldots, \text{k}$$

$$\widehat{ztestscores} = \hat{b}_1 zhomework + \hat{b}_2 zhrsofclass + \hat{b}_3 zclasssize$$

```
Model 9: OLS, using observations 1-3733
Dependent variable: ztestscores

               coefficient   std. error    t-ratio    p-value
  ------------------------------------------------------------
  const          0.000000     0.0161029      0.0000   1.0000
  zhomework      0.161724     0.0161232     10.03     2.21e-023 ***
  zhrsofclass   -0.0340325    0.0161499     -2.107    0.0352    **
  zclasssize     0.00388743   0.000896386    4.337    1.48e-05  ***

Mean dependent var    0.000000   S.D. dependent var    1.000000
Sum squared resid     3609.604   S.E. of regression    0.983861
R-squared             0.032796   Adjusted R-squared    0.032018
F(3, 3729)            42.14827   P-value(F)            8.86e-27
Log-likelihood       -5234.157   Akaike criterion      10476.31
Schwarz criterion     10501.21   Hannan-Quinn          10485.17
```

Interpretation: For this regression, a one standard deviation increase in homework will result in a .1617 standard deviation increase in test scores, a one standard deviation increase in hours of class will result in a .034 standard deviation decrease in test scores, and a one standard deviation increase in class size will lead to a 0.00389 standard deviation increase in test scores. Additionally, the magnitude of the coefficients suggests which variable has the highest input. In this case, homework has the largest impact out of the three regressors on test scores.

Gretl Steps:

Right click → Define new variable →zhomework=(homework-mean(homework))/sd(homework)
→ Repeat for zhrsofclass, zclasssize

Model → Ordinary Least Squares → Dependent variable: ztestscores, Regressors: const, zhomework, zhrsofclass, zclasssize → Ok

Lagged Dependent Variable Model

$$testscores = \alpha + \beta_1 homework + \beta_2 prevtestscores + u$$

```
Model 10: OLS, using observations 1-3733
Dependent variable: testscores

                  coefficient   std. error   t-ratio    p-value
  ------------------------------------------------------------------
  const             8.01126      0.425281     18.84    1.09e-075 ***
  homework          0.884087     0.200035      4.420   1.02e-05  ***
  prevtestscores    0.832788     0.00792505  105.1     0.0000    ***

Mean dependent var   52.43538   S.D. dependent var   9.566599
Sum squared resid    83905.12   S.E. of regression   4.742855
R-squared            0.754342   Adjusted R-squared   0.754210
F(2, 3730)           5726.842   P-value(F)           0.000000
Log-likelihood      -11106.33   Akaike criterion     22218.66
Schwarz criterion    22237.34   Hannan-Quinn         22225.31
```

Interpretation: Including past test scores (from 8th grade) may help control for unobserved variables that may influence test scores in the 10th grade, such as ability or school quality. According to this model, increasing homework time by one hour will lead to a .88 point increase in expected test scores, while scoring one point higher on a test in 8th grade leads to a .83 point increase in expected test scores in 10th grade.

Gretl Steps:

Model → Ordinary Least Squares → Dependent Variable: testscores, Regressors: const, homework, prevtestscores

Dummy Variable Models

$$\widehat{testscores} = \hat{\alpha} + \hat{\beta} homework + \hat{\delta} sex$$

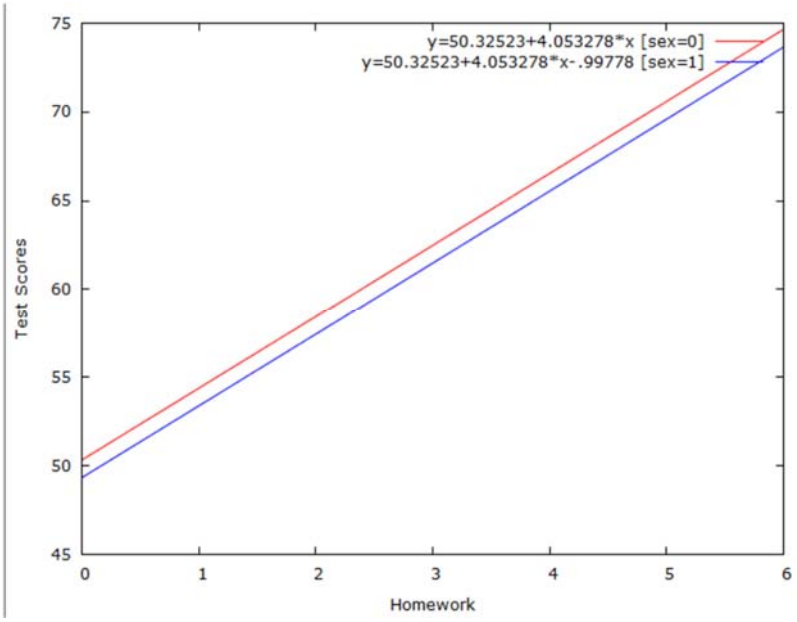
```
Model 11: OLS, using observations 1-3733
Dependent variable: testscores

             coefficient   std. error   t-ratio    p-value
  ---------------------------------------------------------
  const       50.3252       0.333700    150.8     0.0000    ***
  homework     4.05328      0.393321     10.31    1.42e-024 ***
  sex         -0.997780     0.308856     -3.231   0.0012    ***

Mean dependent var   52.43538   S.D. dependent var   9.566599
Sum squared resid    331374.2   S.E. of regression   9.425511
R-squared            0.029799   Adjusted R-squared   0.029278
F(2, 3730)           57.28137   P-value(F)           3.14e-25
Log-likelihood      -13670.08   Akaike criterion     27346.17
Schwarz criterion    27364.84   Hannan-Quinn         27352.81
```

Interpretation: The base group is male 10th graders. Being female will lower your expected test score by 0.998 points for any value of homework.

Gretl Steps:

Model → Ordinary Least Squares → Dependent Variable: testscores, Regressors: const, homework, sex

Plot



$$testscores = \alpha + \beta homework + \delta_1 sex + \delta_2 meduc2 + \delta_3 meduc3 + \delta_4 meduc4 + \delta_5 meduc5 + \delta_6 meduc6 + \delta_7 meduc7 + \delta_8 teachersex + u$$

```
Model 12: OLS, using observations 1-3733
Dependent variable: testscores

                 coefficient   std. error   t-ratio    p-value
  ------------------------------------------------------------
  const            45.4139       0.505328    89.87     0.0000    ***
  teachersex       -0.255157     0.294405    -0.8667   0.3862
  Dmothereduc_2      3.76932     0.459145     8.209    3.03e-016 ***
  Dmothereduc_3      5.78668     0.561344    10.31     1.38e-024 ***
  Dmothereduc_4      7.76384     0.631022    12.30     3.95e-034 ***
  Dmothereduc_5      8.84244     0.557380    15.86     6.78e-055 ***
  Dmothereduc_6     10.5801      0.676677    15.64     2.02e-053 ***
  Dmothereduc_7      7.35052     1.17792      6.240    4.86e-010 ***
  sex               -0.696597    0.293371    -2.374    0.0176    **
  homework           3.61142     0.373534     9.668    7.42e-022 ***

Mean dependent var   52.43538   S.D. dependent var   9.566599
Sum squared resid    297244.6   S.E. of regression   8.935327
R-squared            0.129724   Adjusted R-squared   0.127620
F(9, 3723)           61.66130   P-value(F)           9.0e-106
Log-likelihood      -13467.21   Akaike criterion     26954.42
Schwarz criterion    27016.67   Hannan-Quinn         26976.56

Excluding the constant, p-value was highest for variable 7 (teachersex)
```

Interpretation: This model contains 3 dummy variables, two of which are binary (sex and teachersex) and one with 7 groups (mother's education level). The base group for this model is a male tenth grader whose mother did not finish high school and currently has a male teacher (e.g., sex=0, teachersex=0, and mothereduc=1. The only different between the base group and the alternatives is the intercept.

The students with the highest expected test scores are 10th grade boys whose mothers have a masters degree and have a male teacher. The students with the lowest predicted test scores are 10th grade girls whose mothers dropped out of high school and currently have a female teacher. The difference between the predicted test scores of these two groups for any given level of homework is 11.532 points.

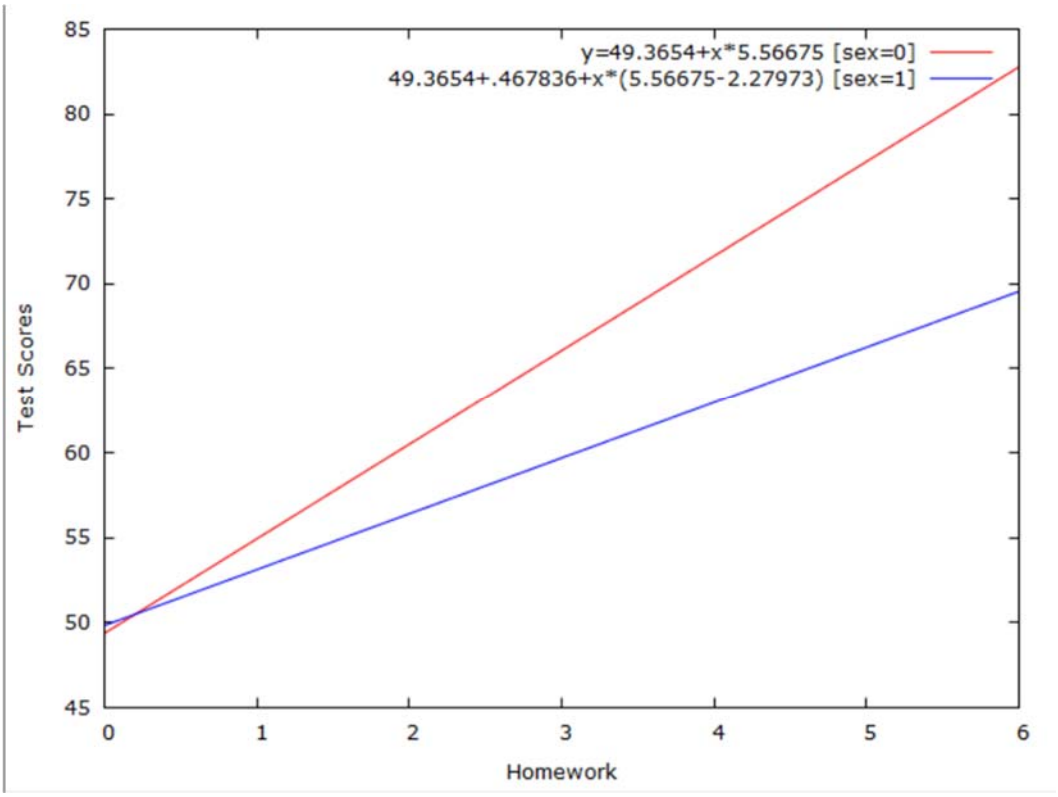
Gretl Steps:

Add → Dummies for Discrete Variables → mothereduc → encode all values → Model → Ordinary Least Squares → Dependent Variable: testscores, Regressors: const, homework, sex, teachersex, Dmothereduc_2, … ,Dmothereduc_7

Dummy Model with Interaction Terms

$$testscores = \alpha + \beta homework + \delta sex + \gamma homework * sex + u$$

```
Model 13: OLS, using observations 1-3733
Dependent variable: testscores

              coefficient   std. error   t-ratio     p-value
  ------------------------------------------------------------
  const         49.3654       0.483570    102.1      0.0000    ***
  homework       5.56675      0.677835      8.213    2.95e-016 ***
  sex            0.467836     0.617469      0.7577   0.4487
  hwrkandsex    -2.27973      0.831913     -2.740    0.0062    ***

Mean dependent var   52.43538   S.D. dependent var   9.566599
Sum squared resid    330708.2   S.E. of regression   9.417297
R-squared            0.031749   Adjusted R-squared   0.030970
F(3, 3729)           40.75739   P-value(F)           6.56e-26
Log-likelihood      -13666.33   Akaike criterion     27340.66
Schwarz criterion    27365.56   Hannan-Quinn         27349.51
```



Interpretation: There is a clear slope change and a relatively small change in the intercept. The expected return to homework is less for tenth grade girls than boys at all time periods spent on homework, except when hours spent on homework is very small.

Gretl Steps:

*Right click → Define new variable →hwrkandsex=homework*sex→ Model → Ordinary Least Squares → Dependent variable: testscores, Regressors: const, homework, hwrkandsex, sex → Ok*

Confidence Intervals

$$100(1-\alpha)\% \text{ Confidence Interval}$$
$$[\hat{\beta}_j \pm c_{\alpha/2} \cdot se(\hat{\beta}_j)]$$
$$or$$
$$[\hat{\beta}_j - c_{\alpha/2} se(\hat{\beta}_j), \hat{\beta}_j + c_{\alpha/2} se(\hat{\beta}_j)]$$

t(3729, 0.025) = 1.961

VARIABLE	COEFFICIENT	95% CONFIDENCE INTERVAL	
const	48.9805	47.3715	50.5895
homework	3.94193	3.17143	4.71243
hrsofclass	-0.327947	-0.633066	-0.0228274
classsize	0.0947537	0.0519168	0.137591

Interpretation: The confidence interval for each variable indicates that we have 95% confidence that the true population parameter is in this interval.

Gretl Steps

Model → Ordinary Least Squares → Dependent Variable: testscores, Regressors: const, homework, classsize, hrsofclass → Ok

In Model Output Box → Analysis → Confidence intervals for coefficients

T-test

OLS Model for test scores vs homework

```
Model 15: OLS, using observations 1-3733
Dependent variable: testscores

             coefficient   std. error   t-ratio    p-value
  ---------------------------------------------------------
  const       49.8379       0.298027    167.2     0.0000    ***
  homework     4.01135      0.393603     10.19    4.47e-024 ***

Mean dependent var   52.43538   S.D. dependent var   9.566599
Sum squared resid    332301.3   S.E. of regression   9.437423
R-squared            0.027084   Adjusted R-squared   0.026823
F(1, 3731)           103.8635   P-value(F)           4.47e-24
Log-likelihood      -13675.30   Akaike criterion     27354.60
Schwarz criterion    27367.05   Hannan-Quinn         27359.03
```

For Beta:

t=10.19 → reject the null

Interpretation: The p-value of 0.000 indicates that we can reject the null hypothesis with nearly 100% certainty. Additionally, the critical value for 99% confidence is $t = 2.576$, which the test statistic of $t = 10.19$ far exceeds.

For alpha:

t=167.2 → reject the null

Interpretation: The p-value of 4.47 x 10^-27 indicates that we can reject the null hypothesis with nearly 100% certainty. Additionally, the critical value for 99% confidence is $t = 2.576$, which the test statistic of $t = 167.2$ exceeds.

Testing Linear Combination of Parameters (Optional)
Single Restriction

$$H_0: \beta_1 = \beta_2$$
$$H_1: \beta_1 \neq \beta_2$$

$$t = \frac{\hat{\beta}_1 - \hat{\beta}_2}{se(\hat{\beta}_1 - \hat{\beta}_2)}$$
$$Var(\hat{\beta}_1 - \hat{\beta}_2) = Var(\hat{\beta}_1) + Var(\hat{\beta}_2) - 2Cov(\hat{\beta}_1, \hat{\beta}_2)$$
$$= \{[se(\hat{\beta}_1)]^2 - [se(\hat{\beta}_2)]^2 - 2s_{12}\}^{\frac{1}{2}}$$
Where s_{12} is the estimate of $Cov(\hat{\beta}_1, \hat{\beta}_2)$

Model: $testscores = \alpha + \beta_1 homework + \beta_2 classsize + \beta_3 hrsofclass + u$

In order to compute $se(\beta_1 - \beta_2)$, we could modify the test to state:

$H_0: \theta = \beta_1 - \beta_2 = 0$ against $H_1: \theta \neq 0$.

When we rearrange the model using $\beta_1 = \theta + \beta_2$, the model becomes

$$testscores = \alpha + \theta homework + \beta_2(homework + classsize) + \beta_3 hrsofclass + u$$

```
Model 3: OLS, using observations 1-3733
Dependent variable: testscores

                 coefficient   std. error   t-ratio    p-value
  ------------------------------------------------------------
  const           48.9805      0.820671     59.68     0.0000    ***
  homework         3.84718     0.394596      9.750    3.40e-022 ***
  hwrkclasssize    0.0947537   0.0218489     4.337    1.48e-05  ***
  hrsofclass      -0.327947    0.155625     -2.107    0.0352    **

Mean dependent var   52.43538   S.D. dependent var   9.566599
Sum squared resid    330350.3   S.E. of regression   9.412200
R-squared            0.032796   Adjusted R-squared   0.032018
F(3, 3729)           42.14827   P-value(F)           8.86e-27
Log-likelihood      -13664.31   Akaike criterion     27336.61
Schwarz criterion    27361.51   Hannan-Quinn         27345.47
```

Interpretation: This method bypasses the need to compute the covariance between $\widehat{\beta_1}$ and $\widehat{\beta_2}$We can apply these values to the original hypothesis test. The critical value from the *t*-distribution at a 99% confidence level in a one-tailed test is 2.326. With this specific test, we reject the null hypothesis when the t-statistic is greater than the absolute value of the critical value. Since the t-statistic, 9.750, is greater than 2.326, we reject the null hypothesis.

Gretl Steps:

Define New Variable → hwrkclasssize=homework+classsize → Ok
Model → Ordinary Least Squares → Dependent Variable: testscores; Regressors: const, homework, hwrkclasssize, hrsofclass → Ok

F-Test (Testing the Validity of the Regression)

$$H_0: \beta_1 = \beta_2 = \cdots = \beta_k$$
$$H_1: H_0 \text{ is false}$$

$$F \equiv \frac{(SSR_r - SSR_{ur})/q}{SSR_{ur}/(n-k-1)} = \frac{(R^2_{ur} - R^2_r)/q}{(1-R^2_{ur})/(n-k-1)}$$
$$q = \text{numerator degrees of freedom (number of restrictions)}$$
$$n-k-1 = \text{denominator degrees of freedom (number of parameters estimated in the unrestricted model)}$$
$$F \sim F_{q,n-k-1}$$

Restricted

$$testscores = \alpha + u$$

Unrestricted

$$testscores = \alpha + \beta_1 homework + \beta_2 classsize + \beta_3 hrsofclass + u$$

$$H_0: \beta_1 = \beta_2 = \beta_3 = 0$$

$$H_1: H_0 \text{ is false}$$

$$F = \frac{(341551.9 - 330350.3)/3}{330350.3/(3733-3-1)} = 42.136\ldots$$

In the case where the restricted model does not have a slope parameter, the R^2 value is zero for the restricted model. We can also compute the F-statistic using the following equation:

$$F = \frac{R^2/k}{(1-R^2)/(n-k-1)} = \frac{0.032796/3}{(1-0.032796)/(3733-3-1)} = 42.126\ldots$$

Where R^2 is the R^2 from the estimation of the unrestricted model.

F*=3.78 @ 99% confidence level

The F-statistic is greater than the critical value. Therefore, we reject the null hypothesis. At least one of the parameters is different than 0. We can conclude by saying that the inclusion of homework, class size, and hours of class explain some variation in the test scores.

Restricted

$$testscores = \alpha + \beta_1 homework + u$$

Unrestricted

$$testscores = \alpha + \beta_1 homework + \beta_2 classsize + \beta_3 hrsofclass + u$$

$$H_0: \beta_2 = \beta_3 = 0$$

$$H_1: H_0 \; is \; false$$

$$F = \frac{(332301.3 - 330350.3)/2}{330350.3/(3733 - 3 - 1)} = 11.008 \ldots$$

F*=4.61 @ 99% confidence level

In this test, we come to the same conclusion as before. Either β_2 or β_3 is different from zero, or β_2 and β_3 are different from 0. Thus, the inclusion of class size and hours of class help explain some variation in test scores.

Restricted

$$testscores = \alpha + u$$

Unrestricted

$$testscores = \alpha + \beta_1 homework + u$$

$$H_0: \beta_1 = 0$$

$$H_1: H_0 \; is \; false$$

$$F = \frac{(341551.9 - 332301.3)/1}{332301.3/(3733 - 1 - 1)} = 101.023367193$$

F*=6.63 @ 99% confidence level

Notes: when there is only one restriction, the F-statistic is the square of the *t*-statistic (10.19).

We reject the null hypothesis in this test as well. However, only this test tells us that homework is statistically significant. The inclusion of homework helps explain some variation in test scores.

Heteroskedasticity

Heteroskedasticity Robust Standard Errors for the OLS Parameter Estimates
White-Huber-Eicker Standard Errors

```
Model 19: OLS, using observations 1-3733
Dependent variable: testscores
Heteroskedasticity-robust standard errors, variant HC1

             coefficient   std. error   t-ratio    p-value
  ---------------------------------------------------------
  const       49.8379       0.396257    125.8     0.0000    ***
  homework     4.01135      0.586133      6.844   8.98e-012 ***

Mean dependent var   52.43538   S.D. dependent var   9.566599
Sum squared resid    332301.3   S.E. of regression   9.437423
R-squared            0.027084   Adjusted R-squared   0.026823
F(1, 3731)           46.83688   P-value(F)           8.98e-12
Log-likelihood      -13675.30   Akaike criterion     27354.60
Schwarz criterion    27367.05   Hannan-Quinn         27359.03
```

Interpretation: Since we are using OLS estimates, the *t*-statistic is still constructed and interpreted as before, but instead, the robust standard errors are used. The conclusions from the *t*-tests concerning α and β remain the same.

Gretl Steps:

Model → Ordinary Least Squares → Dependent Variable: testscores, Regressors: const, homework → Check box for robust standard errors → Ok

Heteroskedasticity Robust F-statistics

```
Model 23: OLS, using observations 1-3733
Dependent variable: testscores
Heteroskedasticity-robust standard errors, variant HC1

              coefficient   std. error   t-ratio    p-value
  ------------------------------------------------------------
  const        48.9805       0.879422     55.70     0.0000    ***
  homework      3.94193      0.581770      6.776    1.43e-011 ***
  hrsofclass   -0.327947     0.161880     -2.026    0.0429    **
  classsize     0.0947537    0.0225281     4.206    2.66e-05  ***

Mean dependent var   52.43538   S.D. dependent var   9.566599
Sum squared resid    330350.3   S.E. of regression   9.412200
R-squared            0.032796   Adjusted R-squared   0.032018
F(3, 3729)           23.72322   P-value(F)           3.32e-15
Log-likelihood      -13664.31   Akaike criterion     27336.61
Schwarz criterion    27361.51   Hannan-Quinn         27345.47
```

```
Test on Model 23:

  Null hypothesis: the regression parameters are zero for the variables
    homework, hrsofclass, classsize
  Test statistic: Robust F(3, 3729) = 23.7232, p-value 3.32182e-015
```

Interpretation: The F-test tests the below hypothesis, which is that all the coefficients of the regressors are equal to 0. The heteroskedasticity robust Wald statistic is a transformation of the heteroskedasticity robust F-statistic. Because the F-statistic is greater than the critical value, we have the same inference from the output.

$$\alpha + \beta_1 homework + \beta_2 classsize + \beta_3 hrsofclass + u.$$

$$H_0: \beta_1 = \beta_2 = \beta_3 = 0$$

$$H_1: H_0 \; is \; false$$

Gretl Steps:

Model → Ordinary Least Squares → Dependent Variable: testscores, Regressors: const, homework, classsize, hrsofclass → Ok

Wald test: Model output box → Tests → Omit variables → Omit homework, hrsofclass, classsize → Select Wald test, based on covariance matrix → Ok

White Test for Heteroskedasticity

$$\hat{u}^2 = \delta_0 + \delta_1 x_1 + \delta_2 x_2 + \delta_3 x_3 + \delta_4 x_1^2 + \delta_5 x_2^2 + \delta_6 x_3^2 + \varepsilon$$

```
White's test for heteroskedasticity (squares only)
OLS, using observations 1-3733
Dependent variable: uhat^2

                 coefficient   std. error    t-ratio     p-value
  ------------------------------------------------------------------
  const          138.675       13.1986       10.51      1.82e-025 ***
  homework         0.640897     7.19896       0.08903   0.9291
  hrsofclass     -13.3416        5.65928      -2.357     0.0185    **
  classsize       -2.21943       0.574597     -3.863     0.0001    ***
  sq_homework      9.66715       1.85734       5.205     2.05e-07  ***
  sq_hrsofclass    2.02197       0.878057      2.303     0.0213    **
  sq_classsize     0.0247272     0.00930605    2.657     0.0079    ***

  Unadjusted R-squared = 0.031392

Test statistic: TR^2 = 117.186213,
with p-value = P(Chi-square(6) > 117.186213) = 0.000000
```

Interpretation: The White test is based on the series of squared residuals from the OLS regression, where the F-statistic is computed from the R-squared of this regression. This approach uses up degrees of freedom quickly, which is important when your sample size is small.

The critical value for the F-statistic is 2.8 for a 99% Confidence Interval. Since the F-statistic is greater than the critical value (and the p-value of the F-statistic is 0), we can reject the null hypothesis that homoskedasticity holds.

Gretl Steps:

Model → Ordinary Least Squares → Dependent Variable: testscores, Regressors: const, homework, classsize, hrsofclass → Ok

Tests → Heteroskedasticity → White's Test (squares only)

$$\hat{u}^2 = \delta_0 + \delta_1 x_1 + \delta_2 x_2 + \delta_3 x_3 + \delta_4 x_1^2 + \delta_5 x_2^2 + \delta_6 x_3^2 + \delta_7 x_1 x_2 + \delta_8 x_1 x_3 + \delta_9 x_2 x_3 + \varepsilon.$$

```
White's test for heteroskedasticity
OLS, using observations 1-3733
Dependent variable: uhat^2

                 coefficient   std. error    t-ratio   p-value
  ---------------------------------------------------------------
  const          130.625       25.8991       5.044    4.79e-07 ***
  homework       -29.2015      22.6396      -1.290    0.1972
  hrsofclass     -12.7606       7.79527     -1.637    0.1017
  classsize       -0.795551     1.00041     -0.7952   0.4265
  sq_homework      9.87160      1.85979      5.308    1.17e-07 ***
  X2_X3           10.0176       4.54585      2.204    0.0276   **
  X2_X4           -0.475101     0.547422    -0.8679   0.3855
  sq_hrsofclass    2.11575      0.880713     2.402    0.0163   **
  X3_X4           -0.325467     0.216602    -1.503    0.1330
  sq_classsize     0.0279379    0.00949798   2.941    0.0033   ***

  Unadjusted R-squared = 0.033316

Test statistic: TR^2 = 124.367032,
with p-value = P(Chi-square(9) > 124.367032) = 0.000000
```

Interpretation: We can reject the null hypothesis since the F-statistic is greater than 2.8 and the p-value of the F-statistic equals 0.

Gretl Steps:

Model → Ordinary Least Squares → Dependent Variable: testscores, Regressors: const, homework, classsize, hrsofclass → Ok

Tests → Heteroskedasticity → White's Test

Alternative to the above model:

Estimate the model: $\hat{u}^2 = \delta_0 + \delta_1 \hat{y} + \delta_2 \hat{y}^2 + \varepsilon.$

```
Model 12: OLS, using observations 1-3733
Dependent variable: uhatsq

             coefficient   std. error   t-ratio    p-value
  ---------------------------------------------------------
  const      2162.53       338.727       6.384    1.93e-010 ***
  yhat        -79.6977      12.1550     -6.557    6.25e-011 ***
  yhatsq        0.764748     0.109032    7.014    2.74e-012 ***

Mean dependent var   88.49458   S.D. dependent var   93.98177
Sum squared resid    32219453   S.E. of regression   92.94043
R-squared            0.022562   Adjusted R-squared   0.022038
F(2, 3730)           43.04893   P-value(F)           3.29e-19
Log-likelihood      -22213.20   Akaike criterion     44432.40
Schwarz criterion    44451.07   Hannan-Quinn         44439.04
```

Interpretation: Use the R-squared from this regression to compute the F-statistic using:

$$F = \frac{R^2_{\hat{u}^2}/2}{\left(1 - R^2_{\hat{u}^2}\right)/(n-2-1)}$$

This approach uses fewer degrees of freedom, especially when more independent variables are added to the original model. The F-statistic is greater than 4.61 (the critical value), so we reject the null hypothesis that homoskedasticity holds. This can also be shown by the p-value of 3.29×10^{-19}

Gretl Steps:

Model → Ordinary Least Squares → Dependent Variable: testscores, Regressors: const, homework, classsize, hrsofclass → Ok → Save → Fitted Values → Name: yhat
Save → Squared residuals → Name: uhatsq

Right click → Define new variable → yhatsq=yhat^2 → Ok

Model → Ordinary Least Squares → Dependent Variable: uhatsq; Regressors: const, yhat, yhatsq

Feasible Generalized Least Squares (FGLS)

```
Model 11: WLS, using observations 1-3733
Dependent variable: testscores
Variable used as weight: hhat

              coefficient   std. error   t-ratio    p-value
  ---------------------------------------------------------
  const         49.1019      0.817838     60.04     0.0000    ***
  homework       3.24773     0.352037      9.226    4.60e-020 ***
  classsize      0.110150    0.0229076     4.808    1.58e-06  ***
  hrsofclass    -0.333496    0.154735     -2.155    0.0312    **

Statistics based on the weighted data:

Sum squared resid    10734658   S.E. of regression   53.65348
R-squared            0.029825   Adjusted R-squared   0.029044
F(3, 3729)           38.21190   P-value(F)           2.58e-24
Log-likelihood      -20161.74   Akaike criterion     40331.48
Schwarz criterion    40356.38   Hannan-Quinn         40340.34

Statistics based on the original data:

Mean dependent var   52.43538   S.D. dependent var   9.566599
Sum squared resid    330661.5   S.E. of regression   9.416633
```

Interpretation: *h(x)* is a function of the explanatory variables that determines the heteroskedasticity. Since it is hard to find such a function, we can model the function and then estimate the parameters in this model.

Gretl Steps:
Model → Ordinary Least Squares → Dependent Variable: testscores, Regressors: const, homework, classsize, hrsofclass → Ok → Save → Squared residuals → Name: uhatsq

Right click → Define new variable →lnuhatsq=ln(uhatsq)

Model → Ordinary Least Squares → Dependent Variable: lnuhatsq, Regressors: const, homework, classsize, hrsofclass → Ok → Save → Fitted values → Name: ghat

Right click → Define new variable →hhat=exp(ghat)

Model → Weighted Least Squares → Dependent Variable: testscores, Weight: hhat, Regressors: const, homework, classsize, hrsofclass → Ok

Ramsey's Regression Specification Error Test (RESET)

Model Misspecification (Neglected Nonlinearities)

Original Model:

$$y = \alpha + \beta_1 x_1 + \cdots + \beta_k x_k$$

Model with fitted values:

$$y = \alpha + \beta_1 x_1 + \cdots + \beta_k x_k + \delta_1 \hat{y}^2 + \delta_2 \hat{y}^3 + u$$

$$H_0: \delta_1 = \delta_2 = 0$$

$$H_1: H_0 \; is \; false$$

```
Auxiliary regression for RESET specification test
OLS, using observations 1-3733
Dependent variable: testscores

              coefficient     std. error    t-ratio    p-value
  ------------------------------------------------------------
  const        4721.57        668.949         7.058    2.00e-012 ***
  homework      633.507        88.1503        7.187    7.98e-013 ***
  hrsofclass    -52.4734        7.33525      -7.154    1.01e-012 ***
  classsize      15.1532        2.11897       7.151    1.03e-012 ***
  yhat^2         -2.66890       0.387805     -6.882    6.89e-012 ***
  yhat^3          0.0146725     0.00222195    6.603    4.59e-011 ***

Test statistic: F = 66.529184,
with p-value = P(F(2,3727) > 66.5292) = 4.08e-029
```

Interpretation: The F-statistic is 66.5292, which is greater than the critical value at a 99% confidence level, which allows us to reject the null hypothesis. This, in addition to a p-value of 4.08×10^{-29} is evidence that the functional form is misspecified.

Gretl Steps:

Model → Ordinary Least Squares → Dependent Variable: testscores, Regressors: const, homework, classsize, hrsofclass → Ok

Tests → Ramey's RESET → Squares and cubes → Ok

Test Against Nonnested Alternatives

Mizon and Richard Test

Two nonnested models:

1. $testscores = \alpha + \beta_1 homework + \beta_2 classsize + \beta_3 hrsofclass + \beta_4 \ln(\text{classsize}) + u$
2. $testscores = \alpha + \beta_1 homework + \beta_2 classsize + \beta_3 hrsofclass + \beta_4 homework^2 + \beta_5 classsize^2 + u$

The comprehensive model is:

$$testscores = \alpha + \gamma_1 homework + \gamma_2 classsize + \gamma_3 hrsofclass + \gamma_4 \ln(classize) + \gamma_5 homework^2 + \gamma_6 classsize^2 + u$$

$$H_0: \gamma_5 = \gamma_6 = 0$$

$$H_1: H_0 \; is \; false$$

We can form the F-statistic the same way as usual. The F-statistic is 39.15144. We reject the null hypothesis at a 99% confidence level. γ_5 and $\gamma_6$are jointly statistically significant. Either one or both of these parameters does not equal zero.

$$H_0: \gamma_4 = 0$$

$$H_1: \; H_0 \; is \; false$$

The F-statistic is 7.859. This means that $\gamma_4$is statistically significant at a 99% confidence level.

Davidson-MacKinnon Test

First estimate the original models

$$\widehat{testscores} = \hat{\alpha} + \hat{\beta}_1 homework + \hat{\beta}_2 classsize + \hat{\beta}_3 hrsofclass + \hat{\beta}_4 \ln(\text{classsize})$$

$$\widehat{\widehat{testscores}} = \hat{\alpha} + \hat{\beta}_1 homework + \hat{\beta}_2 classsize + \hat{\beta}_3 hrsofclass + \hat{\beta}_4 homework^2 + \hat{\beta}_5 classsize^2$$

Now estimate each model including the predicted value of the other model as a regressor

$$\widehat{testscores} = \hat{\alpha} + \hat{\beta}_1 homework + \hat{\beta}_2 classsize + \hat{\beta}_3 hrsofclass + \hat{\beta}_4 \ln(\text{classsize}) + \hat{\theta}\widehat{\widehat{testscores}}$$

$$\widehat{\widehat{testscores}} = \hat{\alpha} + \hat{\beta}_1 homework + \hat{\beta}_2 classsize + \hat{\beta}_3 hrsofclass + \hat{\beta}_4 homework^2 + \hat{\beta}_5 classsize^2 + \hat{\theta}\widehat{testscores}$$

The t-statistics are 8.557036 and 2.803395 for $\hat{\theta}$ in the two estimated models, respectively. In both instances, even at a 99% confidence level we reject the null hypothesis that $\theta = 0$. Therefore, neither model could be rejected. However, since the R-squared value for the second model is higher than that of the first, we can conclude that the second model is better. However, note that this need not be the "best possible model". It is just the best of the two being considered.