Lecture :

Discuss HW#1

Discuss Binge drinking research

Econometrics: The quantitative measurement and analysis of economic, business and sometimes social phenomena.

Three major uses
- Description
- Hypothesis testing (theory testing)
- Forecasting

Intro to Econometrics
- Metrics of economists
- Different then other disciplines, special tools

Intro to regression

A technique to explain the movements in the dependent variable (Endogenous, Y), by movements in the independent (explanatory, exogenous, X variable).

Wages = F(education, experience, tenure....)

Faculty Wages= F(Discipline, Rank, Gender, Years, ….)

Understanding of micro econ (score) = F(study time, instructor, interest, ability....)

The dependent variable must be ratio/interval (continuous)

Regression analysis can find correlation, not causation. Causation requires theory.

Simple linear regression

$$Y = \beta_0 + \beta_1 X$$

where β_0 is the intercept or constant

and β_1 is the slope coefficient, or marginal effect of a one unit change of X on Y

Linear in coefficients versus linear in variables.

$Y = \beta_0 + \beta_1 X$ Linear in both

$Y = \beta_0 + \beta_1 X^2$ Not linear in variables

$Y = \beta_0 + X^{\beta_1}$ Not linear in coefficients

$Y = e^{\beta_0} X_1^{\beta_1}$ Not linear in coefficients (chapter 7)

Regression analysis requires that the estimated equation be linear in the coefficients.

The dependent variable *must* be ratio/interval (continuous) (there are some caveats)

General form $f(Y) = \beta_0 + \beta_1 f(X)$

The Stochastic Error Term

There is always some variation in Y that can't be explained. Example (performance in micro)

1. Omitted variables
2. Measurement error
3. Incorrect functional form
4. Random chance

so we add a term to our equation

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

Two parts, deterministic, and stochastic (random)

$$E(Y \mid X) = \beta_0 + \beta_1 X$$

Expanded notation

$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$ where $i = (1..n)$ and indexes individual observations

so

$$Y_1 = \beta_0 + \beta_1 X_1 + \varepsilon_1$$
$$Y_2 = \beta_0 + \beta_1 X_2 + \varepsilon_2$$
....
$$Y_n = \beta_0 + \beta_1 X_n + \varepsilon_n$$

```
REGRESSION
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS R ANOVA
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT salary
  /METHOD=ENTER market.
```

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.407[a]	.166	.164	11585.82899

a. Predictors: (Constant), market

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	18096.994	3288.009		5.504	.000
	market	34545.219	3424.333	.407	10.088	.000

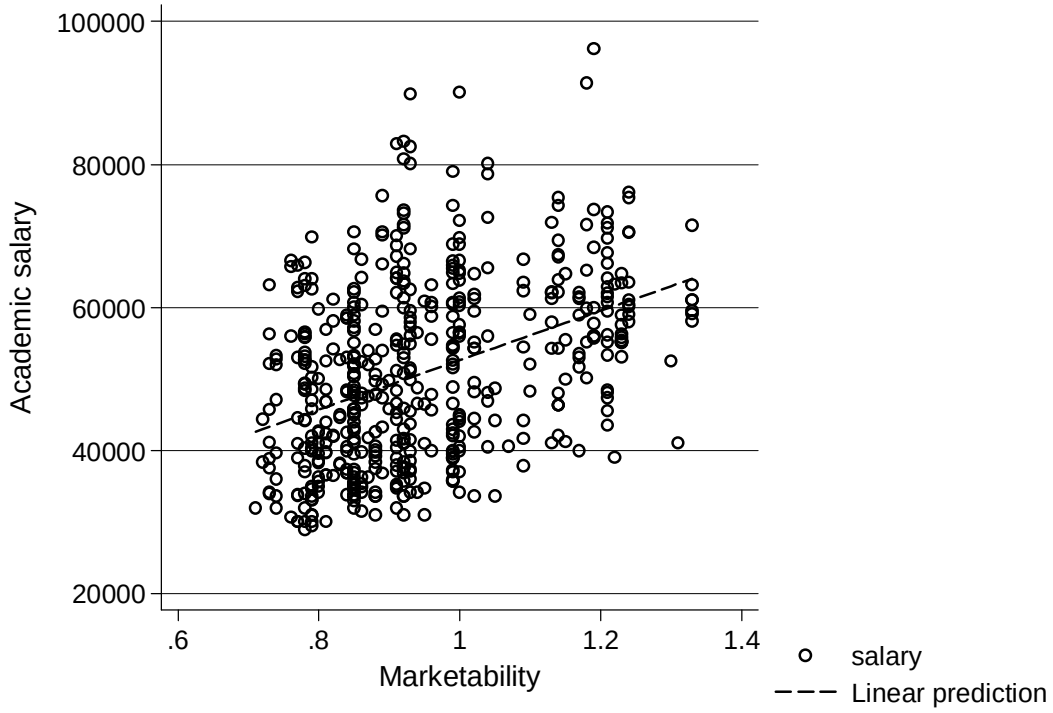
a. Dependent Variable: salary

```
. regress salary market
```

Source	SS	df	MS
Model	**1.3661e+10**	**1**	**1.3661e+10**
Residual	**6.8726e+10**	**512**	**134231433**
Total	**8.2387e+10**	**513**	**160599133**

Number of obs = **514**
F(1, 512) = **101.77**
Prob > F = **0.0000**
R-squared = **0.1658**
Adj R-squared = **0.1642**
Root MSE = **11586**

salary	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
market	**34545.22**	**3424.333**	**10.09**	**0.000**	**27817.75**	**41272.69**
_cons	**18096.99**	**3288.009**	**5.50**	**0.000**	**11637.35**	**24556.64**


100000
80000
60000
40000
20000
Academic salary
.6
.8
1
1.2
1.4
Marketability
salary
Linear prediction

Lecture 9:

Again the multivariate representation is

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_k x_{ki} + \varepsilon_i$$

Again the $\beta' s$ represent the partial effects of the x

Constant is a junk collector, so that the residuals sum to zero. Be careful about making inferences on the value of the constant

$$\sum \varepsilon_i^{\,2} = \sum (y_i - \hat{y}_i)^2 = \sum \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{1i} - \hat{\beta}_2 x_{2i}\right)^2$$

Minimizing by differentiating with respect to the betas and solving them simultaneously yields the normal equations.
http://en.wikibooks.org/wiki/Econometric_Theory/Normal_Equations_Proof
(Note: There is a mistake in the derivation of the above. The solution is correct, but an n appears in front of the alpha a few equations early.)

Where the solutions for the multivariate case are given here:

$$\hat{\beta}_1 = \frac{\sum y x_1 \sum x_2^{\,2} - \sum y x_2 \sum x_1 x_2}{\sum x_1^{\,2} x_2^{\,2} - (\sum x_1 x_2)^2}$$

$$\hat{\beta}_2 = \frac{\sum y x_2 \sum x_1^{\,2} - \sum y x_1 \sum x_1 x_2}{\sum x_1^{\,2} x_2^{\,2} - (\sum x_1 x_2)^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}_1 - \hat{\beta}_2 \bar{x}_{21}$$

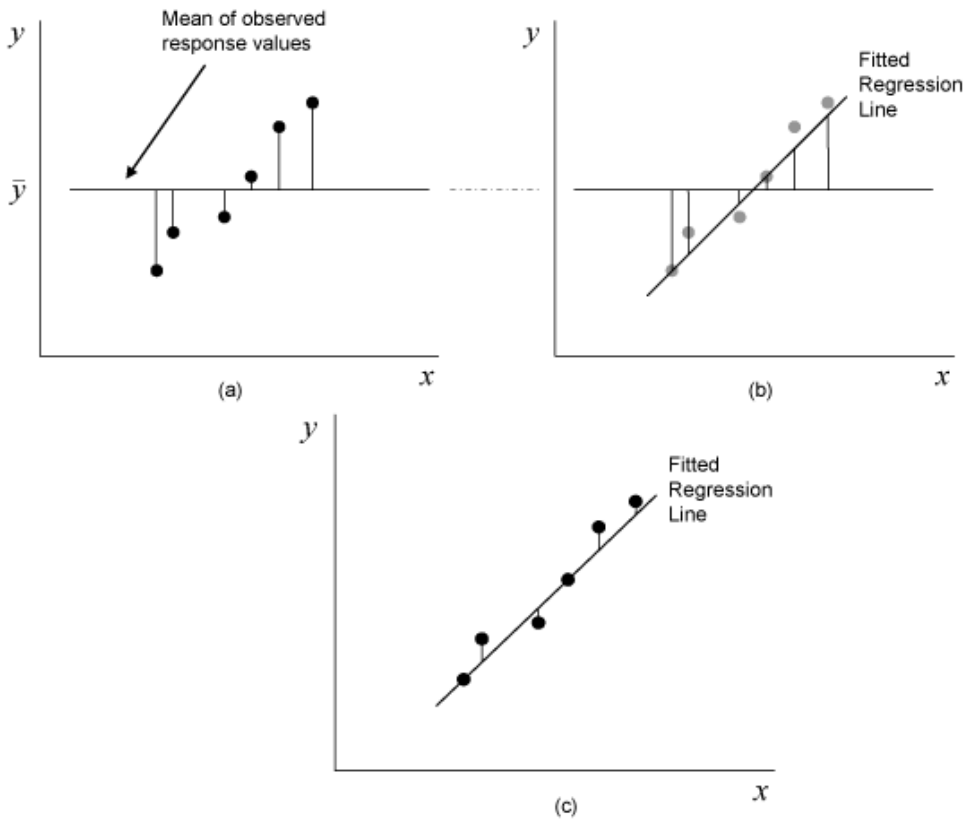
where the lower case letters immediately above represent deviations form their mean
$x_1 = x_{1i} - \bar{x}_1$ and $x_2 = x_{2i} - \bar{x}_2$

Evaluating the quality of a regression.

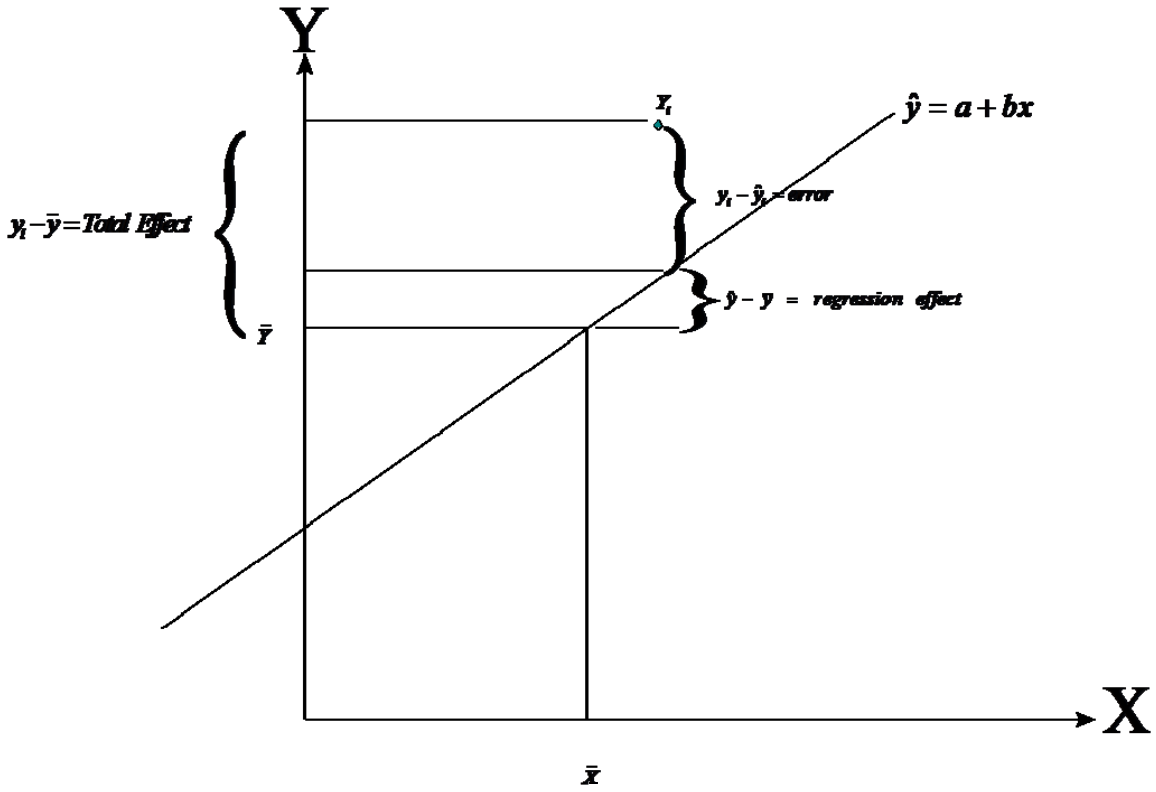
Spend time before running the regression thinking about the expected output.

1. Is the estimated equation supported by the theory?
2. How well does it fit the data?
3. Is the dataset reasonable large and accurate?
4. Is OLS the best estimator for this case?
5. How well do estimates match your prediction?
6. Any important omitted variables?
7. Has the most logical functional form been used?
8. Is the regression free from other econometric problems?

Describing the fit:


y
Mean of observed
response values
$\bar{y}$
x
(a)
y
Fitted
Regression
Line
x
(b)
y
Fitted
Regression
Line
x
(c)

Decomposition of Effects



Total, explained and residual sum of squares.

TSS, ESS, RSS

$TSS = \sum (y_i - \bar{y})^2$ deviation of observation from mean (picture in upper left)

which can be decomposed into two parts

TSS= ESS+RSS

$$\sum (y_i - \bar{y})^2 = \sum (\hat{y}_i - \bar{y})^2 + \sum (y_i - \hat{y}_i)^2$$

The explained portion (ESS), from the fitted line to the mean (this is represented by the solid vertical lines in the upper right hand picture.

The Residual or unexplained portion is depicted in the lowest picture. From the fitted line to the observation.

R^2 (R squared) coefficient of determination

R^2 = (ESS/TSS) = 1-(RSS/TSS)

$$1-\frac{\sum e_i^{\,2}}{\sum (y_i-\bar{y})^2}$$

$$0 \le R^2 \le 1$$

Be careful when comparing time series vs cross section. R squared in the .9 range is common for time series and unheard almost unheard of on cross sectional analysis.

What happens when you add an explanatory variable? TSS doesn't change, but ESS goes up. So we would always want to add a variable, but then the degrees of freedom fall.

Degrees of freedom reflect the reliability of our estimates.

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

we are estimating 2 coefficients degrees of freedom = observations-2. n-2.
More generally

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_k x_{ki} + \varepsilon_i$$

we are estimating k+1 coefficients so degrees of freedom = n-(k+1)

we can use this information to "penalize" the inclusion of an additional variable to better reflect the tradeoff.

NOTE: We cannot estimate the model if there are negative DOF. We effectively have less information than coefficients to estimate. The solution is not unique. n>k+1 is a requirement

Adjusted R^2 or sometimes referred to as r-bar squared

$$\overline{R}^2 = 1-\frac{RSS/(n-(k+1))}{TSS/(n-1)}$$

by a simple rearrangement we get

$$\overline{R}^2 = 1-\left\{(1-R^2)\frac{(n-1)}{(n-(k+1))}\right\}$$

note as k rises so does the penalty, whether or not it offsets the increase in R squared will impact R bar squared.

Note Adjusted R squared (sometimes called R bar Squared) can be less then 0, but it is bounded above by 1.

Appropriate and inappropriate uses of R bar squared

```
COMMENT lets run our first regression.
REGRESSION
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS R /*I've removed the ANOVA from the default */
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
```

```
  /DEPENDENT salary
  /METHOD=ENTER market.
```

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.407[a]	.166	.164	11585.82899

a. Predictors: (Constant), market

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	18096.994	3288.009		5.504	.000
	market	34545.219	3424.333	.407	10.088	.000

a. Dependent Variable: salary

```
COMMENT lets run our second regression adding yearsdg.
REGRESSION
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS R /*I've removed the ANOVA from the default */
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT salary
  /METHOD=ENTER market yearsdg.
```

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.824[a]	.680	.678	7187.88271

a. Predictors: (Constant), yearsdg, market

Coefficients[a]

Model	Unstandardized Coefficients	Standardized Coefficients	t	Sig.

		B	Std. Error	Beta		
1	(Constant)	-1685.118	2153.797		-.782	.434
	market	39630.458	2131.883	.467	18.589	.000
	yearsdg	979.458	34.221	.719	28.622	.000

a. Dependent Variable: salary

Lecture 10: October 5

LAB

Review appropriate and inappropriate use of R_squared

Theory should still drive the inclusion of variables and the fit in terms of expected coefficient signs. You should not dump variables into the regression solely to increase R_squared.

Forecasting water demand in So Cal

Insert example from book here

Using regression analysis in research

1. Review literature and develop theoretical model
2. Specify model: select variables and specify functional form
3. Hypothesize coefficient signs
4. Collect data
5. Estimate and evaluate Equation
6. Document results

1. Review literature

Look for theoretical model to test
Look for previous empirical work to use as a basis for further research with additional data, different country, different time period, different data.
Look for models that may exclude potential important variables.

Use database to search for articles. Econ lit is a good start

Lecture 11: October 10

2. Specify the theoretical model

Select the dependent variable
Select the independent variables and how they are measured.
Select the functional form of the variables
Select the form of the error

Mistakes here result in what is known as **specification error**.

Explain dummy variables.

Dummy variables sometimes called indicator variables take the value of one if the observation has the attribute of interest and zero otherwise.

3. Hypothesize coefficient signs

performance in ECO 110 = F (variables,....)

4. Collect data

How you measure the data is important.

Time series data: what is the frequency or periodicity? Quarterly, monthly, annual? All variables must be measured over the same time span.

Aggregation bias

When looking at cross sectional data, the variable should be measured for the unit of observation. If the dependent variable is different for different states, you don't want to include a variable that is measured for the entire country.

More data More better...use all available data.

Units of measure matter only for the scale of the coefficient, it doesn't matter for its sign or statistical significance.

5. Estimate and evaluate Equation

a. Use OLS as a first pass
b. look at the data again
c. Evaluate..be careful of fixup, problems arise from errors in variables

6. Document results

Example Woodys
Example CLL Paper

Lecture 12:

The Classical Assumptions

I The regression model is linear in the coefficients, is correctly specified, and has an additive error term

II. The error has mean zero $E(\varepsilon_i) = 0$

III. All included variables are uncorrelated with the error term $E(\varepsilon_i x_j) = 0 \forall i, j$

IV. Observations on the error term are uncorrelated with each other, they are independent. $E(\varepsilon_i \varepsilon_j) = 0 \forall i \neq j$

V. The error term has constant variance $E(\varepsilon_i^2) = \sigma^2$

VI. No explanatory variable is a linear function of another variable, i.e. no perfect colliniearity

VII. The error term is normally distributed.

I-V classical error term

I-V, VII classical normal error term

I The regression model is linear in the coefficients, is correctly specified, and has an additive error term

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_k x_{ki} + \varepsilon_i$$

Y=XB+E

Central Limit Theorem

The mean of an iid random variable will tend to be normally distributed if their number is large enough.

Omitted variables

Sampling distribution of Beta_hat

Beta_hat are normally distributed if you use OLS and the errors are normally distributed.

Each sample will produce an estimate. If we resample, and calculate OLS estimates many times we will have the sampling distribution of beta_hat. We want the mean of the sampling distribution to equal the true coefficient.

$E(\hat{\beta}_k) = \beta_k$ then we have an unbiased estimator

if an estimator produces a distribution of Beta_hat not centered around the true value then it is a biased estimator

Properties of the Variance

Beta_hat variance should be as small as possible

as sample size increases variance in estimator decreases

as errors increase so does the variance in beta_hat

$$S.E.\left(\hat{\beta}_1\right) = \sqrt{\frac{\sum e_i^{\,2} / (n-3)}{\sum (x_{1i} - \bar{x}_1)\left(1 - r_{12}^{\,2}\right)}}$$

where $r_{12} = r_{21} = \dfrac{\sum (x_1 - \bar{x}_1)(x_2 - \bar{x}_2)}{\sqrt{\sum (x_1 - \bar{x}_1)^2 \sum (x_2 - \bar{x}_2)^2}}$

as n increases

Betas are normally distributed

Monte Carlo Experiment

1. Assume True model and error distribution
2. Select (fix) values for independent variables
3. Select estimating technique
4. Create sample by drawing an error from the specific distribution being used and combining it with the x values according to the pre-specified “true” model to generate y values for a specific sample size.
5. Calculate coefficient estimates using specified technique
6. evaluate results
7. Repeat 5,000 or 10,000 times collecting the coefficient estimates, plot them, giving you an empirical estimate if the sampling distribution of the estimator.
8. Sensitivity analysis. Choose different error distribution or different values for the x’s

Lecture 13:
Lecture 14:
Gauss Markov Theorem

Given assumptions I-VI
OLS is minimum variance among all linear unbiased estimators

Efficient unbiased smallest variance

Given all 7 assumptions OLS
1. Unbiased
2. Min variance
3. Consistent
4. normally distributed

t-test test one coefficient versus F-test which is a joint test of all coefficients

T-test of slope coefficient.

HO: Beta=0
Ha: Beta ~=0

$$t_k = \frac{\left(\hat{\beta}_k - \beta_{Ho}\right)}{S.E.\left(\hat{\beta}_k\right)}$$

degrees of freedom = n-(k+1)

Critical value (Tcrit) for T with large degrees of freedom at the 5% level is 1.96

confidence interval = $\hat{\beta}_k + t_{crit}(S.E.(\hat{\beta}_k))$

Don't misuse t-scores. They are only a test of statistical significance, not economic importance

F-test
HO: Beta_1=Beta_2=..=Beta_k=0
HA: HO not true

$$F = \frac{ESS/k}{RSS/(n-(k+1))}$$

Examples:

From Before:

```
COMMENT lets run our second regression adding yearsdg.
REGRESSION
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS R /*I've removed the ANOVA from the default */
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT salary
  /METHOD=ENTER market yearsdg.
```

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.824[a]	.680	.678	7187.88271

a. Predictors: (Constant), yearsdg, market

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	5.599E10	2	2.799E10	541.813	.000[a]
	Residual	2.640E10	511	5.167E7		
	Total	8.239E10	513			

a. Predictors: (Constant), yearsdg, market

b. Dependent Variable: salary

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-1685.118	2153.797		-.782	.434
	market	39630.458	2131.883	.467	18.589	.000
	yearsdg	979.458	34.221	.719	28.622	.000

a. Dependent Variable: salary

A t-test of the slope coefficients for the previous regression would go as follows.

For the coefficient on the market variable

$$t_k = \frac{\left(\hat{\beta}_k - \beta_{HO}\right)}{S.E.\left(\hat{\beta}_k\right)}$$

T= (39630.458-0)/(2131.883) = 18.589 Which is greater than 1.96 so reject HO

For the coefficient on the yearsdg variable

T= (979.458-0)/(34.221) = 28.622 Which is greater than 1.96 so reject HO

Remember the F Test

$$F = \frac{ESS/k}{RSS/(n-(k+1))}$$

F=(5.599E10/2)/(2.640E10/(513-(2+1))) = 541.83

Lecture 15: October 22
EXAM I Chapter 16, 1-5