Wage + β_0 + β_1 educ + β_2 expert + ... + u

3 Testing other hypotheses about β_j

- Most common H_0 is $H_0 : \beta_j = 0$. ↙ test whether x_j has a statistically significant impact on y

t or z test

- However, we can test other types of hypotheses. For example, $H_0 : \beta_j = a_j$ where a_j is a constant number.

$$t_{d.f.} = \frac{\hat{\beta}_j - a_j}{s.e.\ a_j} = \frac{\text{estimate - hypothesized value}}{s.e.}$$

d.f. = n-K-1

Example : test whether 1 point increased in wife's income corresponds to 1 point increased in husband's income

husband income = $\hat{\beta}_0 + \hat{\beta}_1$ (Wife's income) + other factors.

$H_0 : \beta_1 = 1$
$H_a : \beta_1 \neq 1$

» we test hypothesis about the population parameters ONLY
Therefore, the β_j in the H_0, H_a cannot be $\hat{\beta}_j$

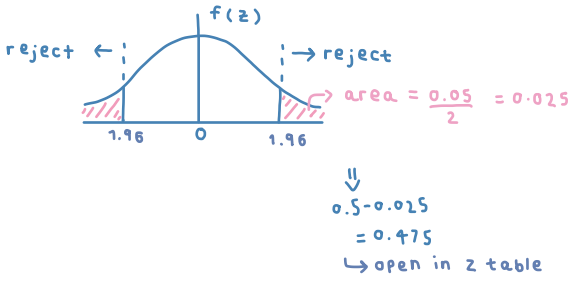
$$t_{\text{wife-income}} = \frac{\hat{\beta}_1 - 1}{s.e\ \hat{\beta}_1}$$

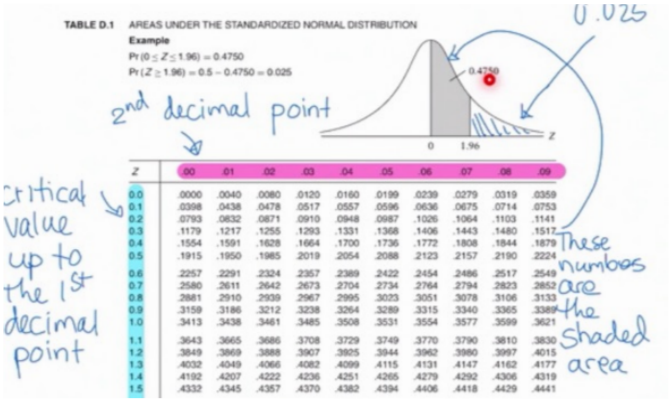
» This is a 2-tailed test
» The equal sign "=" is always with H_0!

Suppose d.f = 250 ⇒ use z-table

If we want to test the H_0 using a 5% significance level, then we reject H_0 if

$$t_{wife\_income} > \underline{\quad 1.96 \quad} \quad or$$
$$t_{wife\_income} < \underline{\quad -1.96 \quad}$$





4 Testing Hypotheses about a Single Linear Combination of the Parameter

Consider

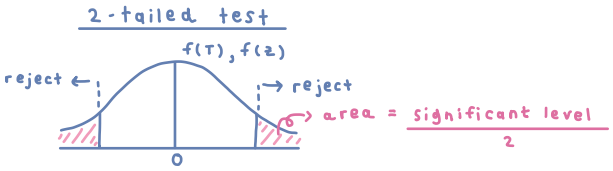
$$\log(wage) = \beta_0 + \beta_1 jc + \beta_2 univ + \beta_3 \exp er + u$$

where jc = number of years attending a two-year college
$univ$ = number of years at a four-year college
$\exp er$ = months in the workforce.
We want to test whether $\beta_1 = \beta_2$. → if the returns from 1 more year of education at a junior college is the same as that of the university

$H_0: \beta_1 = \beta_2 \rightarrow H_0: \beta_1 - \beta_2 = 0$

against

$H_a: \beta_1 \neq \beta_2 \rightarrow H_a: \beta_1 - \beta_2 \neq 0$



$$t = \frac{(\hat{\beta}_1 - \hat{\beta}_2) - 0}{s.e.(\hat{\beta}_1 - \hat{\beta}_2)}$$

we compute this t-statistic and compare with the critical value

where $s.e.(\hat{\beta}_1 - \hat{\beta}_2) = \sqrt{Var(\hat{\beta}_1 - \hat{\beta}_2)}$

$$= \sqrt{Var(\hat{\beta}_1) + Var(\hat{\beta}_2) - 2\,cov(\hat{\beta}_1, \hat{\beta}_2)}$$

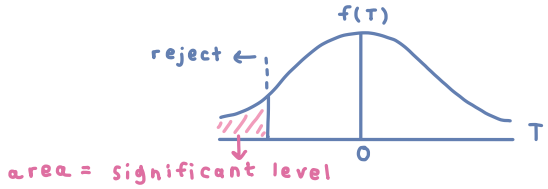
Not a very straight forward to calculate
⇒ we use a variable transformation trick!
⇒ see notes!

another possible hypothesis test (one-tailed alternative)

$H_0: \beta_1 = \beta_2 \rightarrow H_0: \beta_1 - \beta_2 = 0$

$H_a: \beta_1 < \beta_2 \rightarrow H_a: \beta_1 - \beta_2 < 0$

* It is assumed that β_1 would not be more than β_2 (returns to a 2-year college would never be more than returns to university education)



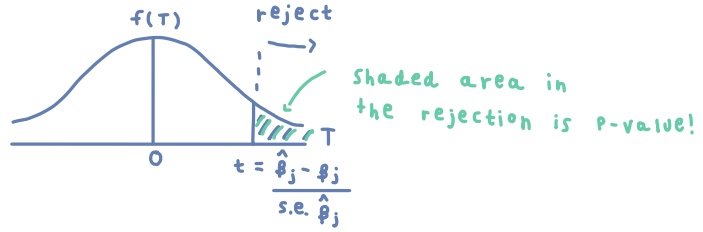
$$t = \frac{(\hat{\beta}_1 - \hat{\beta}_2) - 0}{s.e.(\hat{\beta}_1 - \hat{\beta}_2)}$$

Then, go to the extra note **

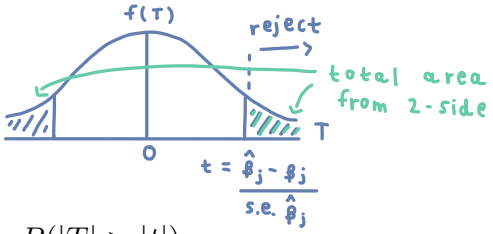
5 Computing p-Values for t-Tests

- What is the significance level given the computed t-statistics?

1-tailed



2-tailed



- p-value : $P(|T| > |t|)$

T = t-distributed random variable with d.f. = n-k-1
t = computed t-statistics

>> p-value = probability that a random T value will be greater (in the + term) than our t in the Ho test.

PLOYPCHY

BE26

Extra note

* In class excercise

Consider the multiple regression model, assume MLR 1-6 are satisfied

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u$$

You would like to test the $H_0: \beta_1 - 3\beta_2 = 1$

1st) Write the t-statistic for testing H_0

$$t = \frac{(\hat{\beta}_1 - 3\hat{\beta}_2) - 1}{s.e.(\hat{\beta}_1 - 3\hat{\beta}_2)}$$

2nd) Define $\theta_1 = \hat{\beta}_1 - 3\hat{\beta}_2$ >> $H_0: \theta_1 = 1$, $H_a: \theta_1 \neq 1$

$$t = \frac{\hat{\theta}_1 - 1}{s.e.(\hat{\theta}_1)}$$ >> we need our regression to have θ_1 in it. So, STATA or OLS estimation will automatically give $\hat{\theta}_1$ & $s.e.\hat{\theta}_1$

Now, $\hat{\beta}_1 = \hat{\theta}_1 + 3\hat{\beta}_2$

or, $\hat{\beta}_1 = \theta_1 + 3\hat{\beta}_2$

sub in the main regression and get

$$Y = \beta_0 + (\theta_1 + 3\beta_2)x_1 + \beta_2 x_2 + \beta_3 x_3 + u$$

$$= \beta_0 + \theta_1 x_1 + 3\beta_2 x_1 + \beta_2 x_2 + \beta_3 x_3 + u$$

$$= \beta_0 + \theta_1 x_1 + \beta_2(x_2 + 3x_1) + \beta_3 x_3 + u$$

>> Now, the explanatory variables are going to be x_1, $x_2 + 3x_1$, and x_3

>> we can calculate $t = \frac{\hat{\theta}_1 - 1}{s.e.\hat{\theta}_1}$

Example 1: $H_0 : \beta_j \geq 0, \quad H_a : \beta_j < 0, \quad$ d.f.$= 140$.

suppose the calculated $t_{\hat{\beta}_j} = -2.75$

- From the z-table, the value -2.75 corresponds to area = _ _ _ _ _ _ _ _ _ _ _ _ _.
- Thus, p-value = _ _ _ _ _ _ _ _ _ _ _.
- Would we reject H_0 if we use the significance level = 5%?

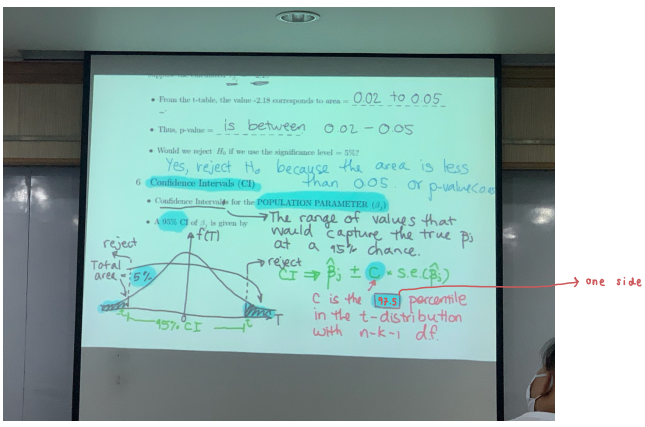
Example 2: $H_0 : \beta_j = a_j, \quad H_a : \beta_j \neq a_j, \quad$ d.f.$= 18$.

suppose the calculated $t_{\hat{\beta}_j} = -2.18$

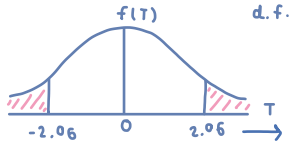
- From the t-table, the value -2.18 corresponds to area = _ _ _ _ _ _ _ _ _ _ _ _ _.
- Thus, p-value = _ _ _ _ _ _ _ _ _ _ _.
- Would we reject H_0 if we use the significance level = 5%?

6 Confidence Intervals (CI)

- Confidence Intervales for the POPULATION PARAMETER (β_j)
- A 95% CI of β_j is given by



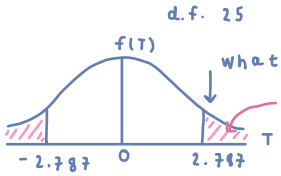
Example 1: **95% CI**



value of 97.5^{th} percentile in t_{25} distribution

The 95% CI for $\hat{\beta}_j = [\hat{\beta}_j$

Example 2: **99% CI**



what percentile? 99.5 or area $1 - 0.005 = 0.995$

area $= \frac{1\%}{2} = \frac{0.01}{2} = 0.005$

The 95% CI for $\hat{\beta}_j = [\hat{\beta}_j -$

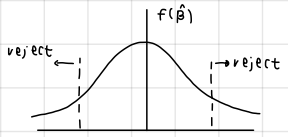
Inference → Hypothesis testing about "β" the true parameter

wage = $\beta_0 + \beta_1 educ + \beta_2 exp + ... + u$

we want to test about the true impact (β) of each x variables (educ, experience) on the dependent variable (Y)

BUT we don't know what the true β are. So, we use $\hat{\beta}$ (estimator)

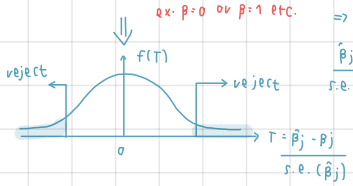
and s.e.($\hat{\beta}$) to test the hypothesis



1) Test if β = same number

e.g. $\beta_j = 0$ → x_j has no impact on y

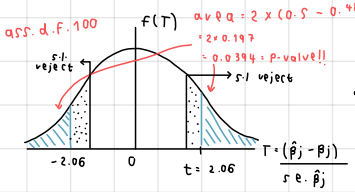
$\beta_j = 1$ → 1 unit ↑ in x_j correspond to 1 unit ↑ in y



⇒ t-test ☆ How??

$\frac{\hat{\beta}_j - \beta_j}{s.e.(\hat{\beta}_j)} \sim t\ d.f.$

Significant level = total area in the rejection region



- suppose, we calculate a t-statistic = $\frac{\hat{\beta}_j - \beta_j}{s.e.(\hat{\beta}_j)}$ = 5.78
- suppose, we are testing $H_0: \beta_j = 0$, $H_a: \beta_j \neq 0$ (↳ 2-tailed test)
- p-value = total shaded area

p-value = significant level which we will reject the H_0 or prob. that we will reject H_0

*if p-value < significance level ⇒ reject H_0 !!

7 Testing Multiple Linear Restrictions: The F-test

Suppose the model is specified by

$$\begin{aligned} y &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u \\ H_0 &: \beta_1 = 0 \text{ and } \beta_2 = 0 \\ H_1 &: H_0 \text{ is not true} \end{aligned}$$

We can use the F-test to test this type of "multiple hypotheses".

1. Our full model is called the "unrestricted" model (ur). Suppose it can be expressed as:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + ... + \beta_k x_k + u$$

2. The model which takes out x (which we think its associated $\beta = 0$) is called the restricted model (r).

3. Some useful facts

4. Other ways to calculate the F-statistics:

Example: Suppose we are interested in understanding the determinant of a baseball player's salary.

$salary$	= season salary
$years$	= years in major leagues
$gamesyr$	= games per year in the league
$bavg$	= career batting average
$hrunsyr$	= homeruns per year
$rbisyr$	= runs batted in per year

- the unrestricted model (ur) is defined by

```
. regress log_salary years gamesyr bavg hrunsyr rbisyr
```

Source	SS	df	MS	Number of obs =	353
				F(5, 347) =	117.06
Model	308.989208	5	61.7978416	Prob > F =	0.0000
Residual	183.186327	347	.527914487	R-squared =	0.6278
				Adj R-squared =	0.6224
Total	492.175535	352	1.39822595	Root MSE =	.72658

log_salary	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
years	.0688626	.0121145	5.68	0.000	.0450355	.0926898
gamesyr	.0125521	.0026468	4.74	0.000	.0073464	.0177578
bavg	.0009786	.0011035	0.89	0.376	-.0011918	.003149
hrunsyr	.0144295	.016057	0.90	0.369	-.0171518	.0460107
rbisyr	.0107657	.007175	1.50	0.134	-.0033462	.0248776
_cons	11.19242	.2888229	38.75	0.000	10.62435	11.76048

- the restricted model (r) is defined by

```
. regress log_salary years gamesyr
```

Source	SS	df	MS	Number of obs =	353
				F(2, 350) =	259.32
Model	293.864058	2	146.932029	Prob > F =	0.0000
Residual	198.311477	350	.566604221	R-squared =	0.5971
				Adj R-squared =	0.5948
Total	492.175535	352	1.39822595	Root MSE =	.75273

log_salary	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
years	.071318	.012505	5.70	0.000	.0467236	.0959124
gamesyr	.0201745	.0013429	15.02	0.000	.0175334	.0228156
_cons	11.2238	.108312	103.62	0.000	11.01078	11.43683

Now, our H_0 and H_a becomes

8 How the Hypothesis Testing is done in Practice

1. Check the values of $t-statistic$ reported by the statistical software (i.e. STATA, SPSS, SAS)

   $\Rightarrow$ These $t-statistics$ are to test $H_0 : \beta_i = 0$

   $\Rightarrow$ If the d.f. > 30, then when $t > 1.96$, we can reject H_0

   $\Rightarrow$ **When $t > 1.96$**, we can say that β_i is **statistically significant** at 5% level. (value of $\beta_i \neq 0$)

   $\Rightarrow$ **When $t < 1.96$** we can say that β_i is **not statistically significant** at 5% level.

   $\Rightarrow$ If $t < 1.96$ we can drop x_i from the model

   $\Rightarrow$ After we drop x_i, we estimate the new regression function and obtain a new set of $\hat{\beta}$.

2. We can also perform other hypothesis testings of interest.

   e.g. $H_0 : \beta_i = \beta_j$

   or $H_0 : \beta_i = 5$ etc.

   or perform an F-test for testing multiple linear restrictions

3. Usually, in economics, the estimation results are reported using this form

Dependent Variable: log(*salary*)			
Independent Variables	**(1)**	**(2)**	**(3)**
log(*sales*)	.224 (.027)	.158 (.040)	.188 (.040)
log(*mktval*)	——	.112 (.050)	.100 (.049)
profmarg	——	−.0023 (.0022)	−.0022 (.0021)
ceoten	——	——	.0171 (.0055)
comten	——	——	−.0092 (.0033)
intercept	4.94 (0.20)	4.62 0.25)	4.57 (0.25)
Observations *R*-squared	177 .281	177 .304	177 .353

Multiple Regression Analysis : Further Issues

1 Data scaling on OLS statistics

When we change the unit of measurment of a variable, the value of estimators would chance accordingly. For example

$$\widehat{bweght} = \widehat{\beta}_0 + \widehat{\beta}_1 cigs + \widehat{\beta}_2 fa\min c,$$

where
$bwght =$ child birth weight, in grams.
$cigs =$ number of cigarettes smoked by the mother while pregnant, per day.
$fa\min c =$ annual family income, in thousands of dollars.

2 More on functional forms

- Logarithmic Functional Form

$$\log(y) = \beta_0 + \beta_1 \log(x_1) + \beta_2 x_2 + u$$

- Models with Quadratics

Example : Effects of Pullution on Housing Prices

$$\log(price) = \beta_0 + \beta_1 \log(nox) + \beta_2 \log(dist) + \beta_3 rooms + \beta_4 room^2 + \beta_5 stratio + u$$

where

price = housing price

nox = level of pollution

dist = distance from downtown

rooms = number of rooms

stratio = average student per teacher ratio

The estimation result is given by

```
regress  lprice lnox dist rooms rooms_sq stratio
```

Source	SS	df	MS		
Model	51.4933152	5	10.298663	Number of obs =	506
Residual	33.0889098	500	.06617782	F(5, 500) =	155.62
				Prob > F =	0.0000
				R-squared =	0.6088
				Adj R-squared =	0.6049
Total	84.582225	505	.167489554	Root MSE =	.25725

lprice	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
lnox	-.9767545	.0995938	-9.81	0.000	-1.172429	-.7810806
dist	-.0321972	.0094013	-3.42	0.001	-.050668	-.0137264
rooms	-.5528032	.1612965	-3.43	0.001	-.8697056	-.2359007
rooms_sq	.0624697	.0124867	5.00	0.000	.0379368	.0870025
stratio	-.0486667	.0058131	-8.37	0.000	-.0600879	-.0372455
_cons	13.59154	.5650901	24.05	0.000	12.4813	14.70178

Consider the effect of "room"

What wold be the % change in price when the number of room increases from 5 to 6?

3 Models with Interaction Terms

Consider

$$price = \beta_0 + \beta_1 sqrft + \beta_2 bdrms + \beta_3 sqrft \times bdrms + \beta_4 bthrms + u$$

where
$price$ = housing price
$sqrft$ = house size (square feet)
$bdrms$ = number of bedrooms
$bthrms$ = number of bathrooms

4 More on the Goodness-of-Fit and Selection of Regressors

- Adding more regressors ALWAYS improve fit

Using adjusted R-squared to choose between non-nested models (one model is not a subset of another).

Consider Model 1

$$\begin{aligned} \widehat{salary} &= \underset{(223.90)}{830.63} \underset{(0.0089)}{+0.0163 sales} \underset{(11.08)}{+19.63 roe} \\ n &= 209, \quad R^2 = 0.029, \ \bar{R}^2 = 0.020 \end{aligned}$$

Consider Model 2

$$\begin{aligned} \widehat{\log(salary)} &= \underset{(0.29)}{4.36} \underset{(0.033)}{+0.2751 \log(sales)} \underset{(0.004)}{+0.0179 roe} \\ n &= 209, \quad R^2 = 0.282, \ \bar{R}^2 = 0.275 \end{aligned}$$

Multiple Regression Analysis with Qualitative Information:

1 Outline

- Describing qualitative information
- Using a single dummy independent variable
- Using dummy variables for multiple categories
- Interactions involving dummy variables
- A binary dependent variable (Y variable): The linear probability model

2 Describing Qualitative Information

- "Female" and "Married" are qualitative variable.
- We arbitarily assign a dummy variable to decribe them.

$$famale = \begin{cases} 1 & \text{if female} \\ 0 & \text{otherwise (or if male)} \end{cases}$$

$$married = \begin{cases} 1 & \text{if married} \\ 0 & \text{otherwise (of if single)} \end{cases}$$

TABLE 7.1

A Partial Listing of the Data in WAGE1.RAW

person	wage	educ	exper	female	married
1	3.10	11	2	1	0
2	3.24	12	22	1	1
3	3.00	11	2	0	0
4	6.00	8	44	0	1
5	5.30	12	7	0	1
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
525	11.56	16	5	0	1
526	3.50	14	5	1	0

3 Models with a single dummy independent variable

Consider

$$wage = \beta_0 + \delta_0 female + \beta_1 educ + u.$$

where

$$famale = \begin{cases} 1 & \text{if female} \\ 0 & \text{otherwise (or if male)} \end{cases}$$

In this case, the δ_0 notation is used to highlight the interpretation of the parameters multiplying dummy variables. In other cases, we can use any notation that is the most convenient.

5 Using dummy variables for multiple categories

Case 1 We can use many dummy variables in the same model

Consider a model which includes 2 dummy variables– $female$ and $married$.

$$\begin{aligned}\log(wage) &= \beta_0 + \delta_0 female + \delta_1 married + \beta_1 educ + \beta_2 \exp er + \beta_3 \exp er^2 \\ &\quad + \beta_4 tenure + \beta_5 tenure^2 + u.\end{aligned}$$

```
regress lwage female married educ exper expersq tenure tenursq
```

Source	SS	df	MS	Number of obs =	526
				F(7, 518) =	58.76
Model	65.6482326	7	9.37831895	Prob > F =	0.0000
Residual	82.6815188	518	.159616832	R-squared =	0.4426
				Adj R-squared =	0.4351
Total	148.329751	525	.28253286	Root MSE =	.39952

lwage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
female	-.2901838	.0361121	-8.04	0.000	-.3611279	-.2192396
married	.0529219	.0407561	1.30	0.195	-.0271456	.1329894
educ	.0791547	.0068003	11.64	0.000	.0657952	.0925143
exper	.0269535	.0053258	5.06	0.000	.0164907	.0374163
expersq	-.0005399	.0001122	-4.81	0.000	-.0007603	-.0003196
tenure	.0312962	.0068482	4.57	0.000	.0178426	.0447499
tenursq	-.0005744	.0002347	-2.45	0.015	-.0010355	-.0001134
_cons	.4177837	.0988662	4.23	0.000	.2235557	.6120116

Comments:

Consider a model which includes dummy variables for each gender/marital status combination– $marrmale$, $marrfem$ and $singfem$.

$$\begin{aligned}\log(wage) &= \beta_0 + \delta_0 marrmale + \delta_1 marrfem + \delta_3 singfem + \beta_1 educ + \beta_2 exper \\ &\quad + \beta_3 exper^2 + \beta_4 tenure + \beta_5 tenure^2 + u. \end{aligned} \tag{8.1}$$

```
regress lwage marrmale marrfem singfem educ exper expersq tenure tenursq
```

Source	SS	df	MS		
Model	68.3617623	8	8.54522029	Number of obs	= 526
Residual	79.9679891	517	.154676961	F(8, 517)	= 55.25
				Prob > F	= 0.0000
				R-squared	= 0.4609
				Adj R-squared	= 0.4525
Total	148.329751	525	.28253286	Root MSE	= .39329

lwage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
marrmale	.2126757	.0553572	3.84	0.000	.103923	.3214284
marrfem	-.1982676	.0578355	-3.43	0.001	-.311889	-.0846462
singfem	-.1103502	.0557421	-1.98	0.048	-.219859	-.0008414
educ	.0789103	.0066945	11.79	0.000	.0657585	.092062
exper	.0268006	.0052428	5.11	0.000	.0165007	.0371005
expersq	-.0005352	.0001104	-4.85	0.000	-.0007522	-.0003183
tenure	.0290875	.006762	4.30	0.000	.0158031	.0423719
tenursq	-.0005331	.0002312	-2.31	0.022	-.0009874	-.0000789
_cons	.3213781	.100009	3.21	0.001	.1249041	.5178521

Comments: