

$$wage = \beta_0 + \beta_1 educ + \beta_2 expert + \dots + u$$

66 6. Multiple Regression Analysis (Inference)

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.

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

= n-k-1

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

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

$$H_0 : \beta_1 = 1$$

$$H_a : \beta_1 \neq 1$$

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

Suppose $d.f. = 250 \Rightarrow$ use z-table

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

>> This is a 2-tailed test

>> The equal sign "=" is always with H_0 !

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

$$t_{\text{wife-income}} > \underline{\underline{1.96}} \quad \text{or}$$

$$t_{\text{wife-income}} < \underline{\underline{-1.96}}$$

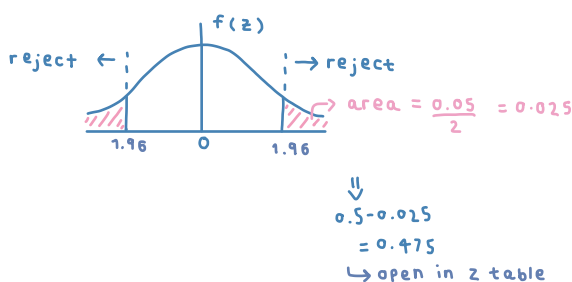


TABLE D.1 AREAS UNDER THE STANDARDIZED NORMAL DISTRIBUTION

Example
 $Pr(0 \leq Z \leq 1.96) = 0.4750$
 $Pr(Z \leq 1.96) = 0.5 + 0.4750 = 0.975$

2nd decimal point

critical value up to the 1st decimal point

These numbers are the shaded area

Z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.0000	.0040	.0080	.0120	.0160	.0199	.0239	.0279	.0319	.0359
0.1	.0398	.0438	.0478	.0517	.0557	.0596	.0636	.0675	.0714	.0753
0.2	.0793	.0832	.0871	.0910	.0948	.0987	.1026	.1064	.1103	.1141
0.3	.1179	.1217	.1255	.1293	.1331	.1368	.1406	.1443	.1480	.1517
0.4	.1554	.1591	.1628	.1664	.1700	.1736	.1772	.1808	.1844	.1879
0.5	.1915	.1950	.1985	.2019	.2054	.2088	.2123	.2157	.2190	.2224
0.6	.2257	.2291	.2324	.2357	.2389	.2422	.2454	.2486	.2517	.2549
0.7	.2580	.2611	.2642	.2673	.2704	.2734	.2764	.2794	.2823	.2852
0.8	.2881	.2910	.2939	.2967	.2995	.3023	.3051	.3078	.3106	.3133
0.9	.3159	.3186	.3212	.3238	.3264	.3289	.3315	.3340	.3365	.3389
1.0	.3413	.3438	.3461	.3485	.3508	.3531	.3554	.3577	.3599	.3621
1.1	.3643	.3665	.3686	.3708	.3729	.3749	.3770	.3790	.3810	.3830
1.2	.3849	.3869	.3888	.3907	.3925	.3944	.3962	.3980	.3997	.4015
1.3	.4032	.4049	.4066	.4082	.4099	.4115	.4131	.4147	.4162	.4177
1.4	.4192	.4207	.4222	.4236	.4251	.4265	.4279	.4292	.4306	.4319
1.5	.4332	.4345	.4357	.4370	.4382	.4394	.4406	.4418	.4429	.4441

4 Testing Hypotheses about a Single Linear Combination of the Parameter

Consider

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

where jc = number of years attending a two-year college

$univ$ = number of years at a four-year college

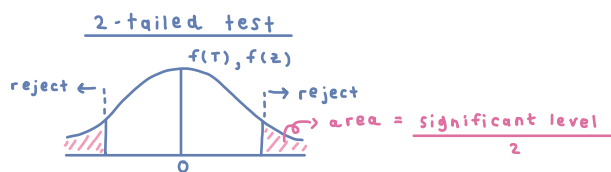
$\text{exp } er$ = months in the workforce.

We want to test whether $\beta_1 = \beta_2$. $\rightarrow$ 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

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



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

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

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

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

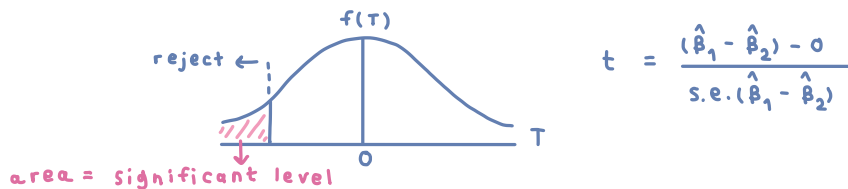
Not a very straight forward to calculate
 $\Rightarrow$ we use a variable transformation trick!
 $\Rightarrow$ see notes!

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)

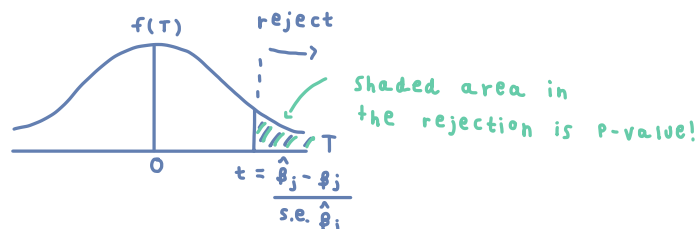


Then, go to the extra note **

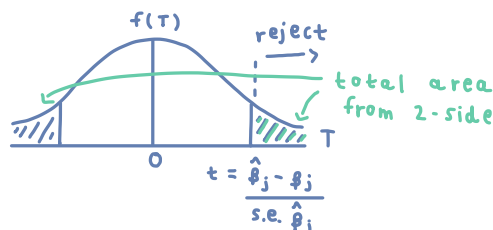
5 Computing p-Values for t-Tests

- 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 1 term) than our t in the H_0 test.

Extra note

* In class exercise

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

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

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

$$t = \frac{\hat{\theta}_1 - 1}{\text{s.e.}(\hat{\theta}_1)} \gg \text{we need our regression to have } \theta_1 \text{ in it.}$$

So, STATA or OLS estimation will automatically give $\hat{\theta}_1$ & $\text{s.e.}\hat{\theta}_1$

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

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

sub in the main regression and get

$$\begin{aligned} Y &= \beta_0 + (\theta_1 + 3\beta_2)X_1 + \beta_2 X_2 + \beta_3 X_3 + u \\ &= \beta_0 + \theta_1 X_1 + 3\beta_2 X_1 + \beta_2 X_2 + \beta_3 X_3 + u \\ &= \beta_0 + \theta_1 X_1 + \beta_2 (X_2 + 3X_1) + \beta_3 X_3 + u \end{aligned}$$

>> Now, the explanatory variables are going to be $X_1, X_2 + 3X_1$, and X_3

$$\gg \text{ We can calculate } t = \frac{\hat{\theta}_1 - 1}{\text{s.e.}\hat{\theta}_1}$$

Example 1: $H_0 : \beta_j \geq 0$, $H_a : \beta_j < 0$, 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%?

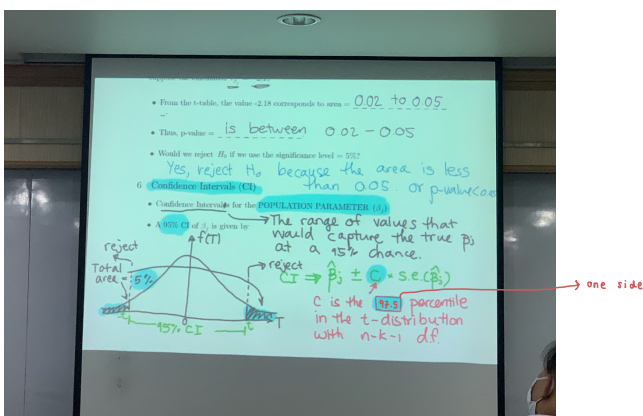
Example 2: $H_0 : \beta_j = a_j$, $H_a : \beta_j \neq a_j$, d.f. = 18.

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

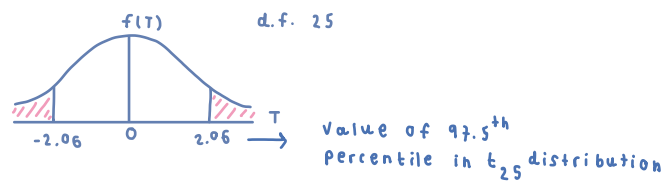
- From the t-table, the value -2.18 corresponds to area = _____.
- Thus, p-value = _____.
- Would we reject H_0 if we use the significance level = 5%?

6 Confidence Intervals (CI)

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

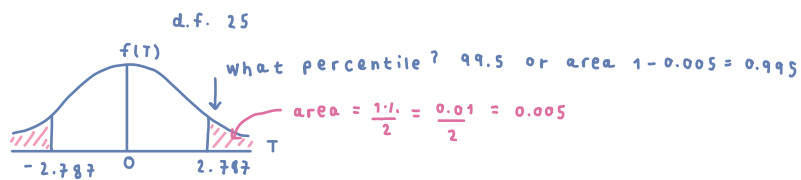


Example 1: **95% CI**



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

Example 2: **99% CI**



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

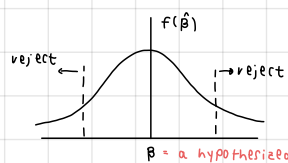
Inference → Hypothesis testing about " β " the true parameter

$$\text{wage} = \beta_0 + \beta_1 \text{educ} + \beta_2 \text{exp} + \dots + u$$

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

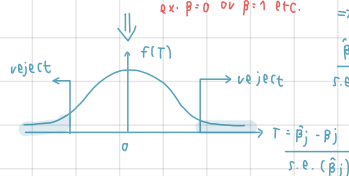
e.g. $\beta_j = 0 \rightarrow x_j$ has no impact on y

$\beta_j = 1 \rightarrow 1$ unit $\uparrow$ in x_j correspond to 1 unit $\uparrow$ in y

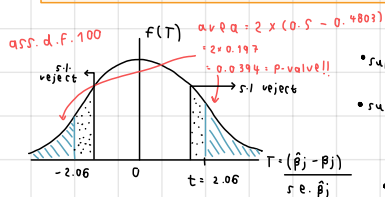
β = a hypothesized value
ex. $\beta = 0$ or $\beta = 1$ etc.

⇒ t-test "How??"

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



Significant level = total area in the rejection region



* Suppose, we calculate a t-statistic = $\frac{\hat{\beta}_j - \beta_j}{\text{s.e.}(\hat{\beta}_j)} = 5.78$

* Suppose, we are testing $H_0: \beta_j = 0, H_1: \beta_j \neq 0$

↳ 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 $p\text{-value} < \text{significance level} \Rightarrow \text{reject } H_0$!!

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:

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

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.

<i>salary</i>	= season salary
<i>years</i>	= years in major leagues
<i>gamesyr</i>	= games per year in the league
<i>bavg</i>	= career batting average
<i>hrunsyr</i>	= homeruns per year
<i>rbisyr</i>	= 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		
Model	308.989208	5	61.7978416	F(5, 347) = 117.06		
Residual	183.186327	347	.527914487	Prob > F = 0.0000		
				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		
Model	293.864058	2	146.932029	F(2, 350) = 259.32		
Residual	198.311477	350	.566604221	Prob > F = 0.0000		
				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: $\log(\text{salary})$			
Independent Variables	(1)	(2)	(3)
$\log(\text{sales})$	.224 (.027)	.158 (.040)	.188 (.040)
$\log(\text{mktval})$	—	.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	177	177	177
R-squared	.281	.304	.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 change 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	Number of obs = 506		
Model	51.4933152	5	10.298663	F(5, 500) = 155.62		
Residual	33.0889098	500	.06617782	Prob > F = 0.0000		
Total	84.582225	505	.167489554	R-squared = 0.6088		
				Adj R-squared = 0.6049		
				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 would be the % change in price when the number of room increases from 5 to 6?

3 Models with Interaction Terms

Consider

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

where

price = housing price

sqrf t = 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

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

Consider Model 2

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

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 arbitrarily assign a dummy variable to describe them.

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

TABLE 7.1

A Partial Listing of the Data in WAGE1.RAW

<i>person</i>	<i>wage</i>	<i>educ</i>	<i>exper</i>	<i>female</i>	<i>married</i>
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

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

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*.

$$\log(wage) = \beta_0 + \delta_0 female + \delta_1 married + \beta_1 educ + \beta_2 exper + \beta_3 exper^2 + \beta_4 tenure + \beta_5 tenure^2 + u.$$

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

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

$$\log(wage) = \beta_0 + \delta_0 marrmale + \delta_1 marrfem + \delta_2 singfem + \beta_1 educ + \beta_2 exper + \beta_3 exper^2 + \beta_4 tenure + \beta_5 tenure^2 + u. \quad (8.1)$$

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

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

lwage	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
δ_0 marrmale	.2126757	.0553572	3.84	0.000	.103923	.3214284
δ_1 marrfem	-.1982676	.0578355	-3.43	0.001	-.311889	-.0846462
δ_2 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: