7 Regression with indicator variables

One of the most useful concepts in applied economics is the *indicator variable*, which signals the presence or absence of a characteristic. Indicator variables are also known as *binary* or *Boolean* variables and are well known to econometricians as *dummy variables* (although the meaning of that latter term is shrouded in the mists of time). Here we consider how to use indicator variables

- to evaluate the effects of qualitative factors;
- in models that mix quantitative and qualitative factors;
- in seasonal adjustment; and
- to evaluate structural stability and test for structural change.

7.1 Testing for significance of a qualitative factor

Economic data come in three varieties: quantitative (or cardinal), ordinal (or ordered), and qualitative.[1] In chapter 3, I described the first category as *continuous* data to stress that their values are quantities on the real line that may conceptually take on any value. We also may work with *ordinal* or ordered data. They are distinguished from cardinal measurements in that an ordinal measure can express only inequality of two items and not the magnitude of their difference; for example, a Likert scale of "How good a job has the president done? 5 = great, 4 = good, 3 = fair , 2 = poor, 1 = very poor" will generate ordered numeric responses. A response of 5 beats 4, which in turn beats 3 for voter satisfaction. But we cannot state that a respondent of 5 is five times more likely to support the president than a voter responding 1, nor 25% more likely than a respondent of 4, and so on. The numbers can be taken only as *ordered*. They could be any five ordered points on the real line (or the set of integers). The implication: if data are actually ordinal rather than cardinal, we should not mistake them for cardinal measures and should not use them as a response variable or as a regressor in a linear regression model.

In contrast, we often encounter economic data that are purely *qualitative*, lacking any obvious ordering. If these data are coded as string variables, such as `M` and `F` for survey respondents' genders, we are not likely to mistake them for quantitative values. We hope that few researchers would contemplate using five-digit ZIP codes (U.S. postal codes) in a quantitative setting. But where a quality may be coded numerically, there

1. I discuss censored data in chapter 10.

is the potential to misuse this qualitative factor as quantitative. This misuse of course is nonsensical: as described in section 2.2.4, we can `encode` a two-letter U.S. state code (AK, AL, AZ, ..., WY) into a set of integers 1, ..., 50 for ease of manipulation, but we should never take those numeric values as quantitative measures.

How should we evaluate the effects of purely qualitative measures? Since the answer to this question will apply largely to ordinal measures as well, it may be taken to cover all nonquantitative economic and financial data. To test the hypothesis that a qualitative factor has an effect on a response variable, we must convert the qualitative factor into a set of *indicator variables*, or dummy variables. Following the discussion in section 4.5.3, we then conduct a *joint test* on their coefficients. If the hypothesis to be tested includes one qualitative factor, the estimation problem may be described as a one-way ANOVA. Economic researchers consider that ANOVA models may be expressed as linear regressions on an appropriate set of indicator variables.[2]

The equivalence of one-way ANOVA and linear regression on a set of indicator variables that correspond to one qualitative factor generalizes to multiple qualitative factors. If two qualitative factors (e.g., race and sex) are hypothesized to affect income, an economic researcher would regress income on two appropriate sets of indicator variables, each representing one of the qualitative factors. If we include one or many qualitative factors in a model, we will estimate a linear regression on several indicator (dummy) variables.

7.1.1 Regression with one qualitative measure

Consider measures of the six New England states' per capita disposable personal income (`dpipc`) for 1981–2000 as presented in section 6.2.2. Does the state of residence explain a significant proportion of the variation in `dpipc` over these two decades? We calculate the average `dpipc` (in thousands of dollars) over the two decades by using `mean` (see [R] **mean**):

2. Stata's `anova` command has a `regress` option that presents the results of ANOVA models in a regression framework.

```
. use http://www.stata-press.com/data/imeus/NEdata, clear
. mean dpipc, over(state)
Mean estimation                   Number of obs    =       120
          CT: state = CT
          MA: state = MA
          ME: state = ME
          NH: state = NH
          RI: state = RI
          VT: state = VT
```

Over	Mean	Std. Err.	[95% Conf.	Interval]
dpipc				
CT	22.32587	1.413766	19.52647	25.12527
MA	19.77681	1.298507	17.20564	22.34798
ME	15.17391	.9571251	13.27871	17.06911
NH	18.66835	1.193137	16.30582	21.03088
RI	17.26529	1.045117	15.19586	19.33473
VT	15.73786	1.020159	13.71784	17.75788

States' average dpipc in 2000 varies considerably between Connecticut ($22,326) and Maine ($15,174). But are these differences statistically significant? Let us test this hypothesis with regress. We first must create the appropriate indicator variables. One way to do this (which I prefer to using xi) is, as described in section 2.2.4, to use tabulate and its generate() option to produce the desired variables. The following command generates six indicator variables, but we recognize that these six indicator variables must be *mutually exclusive and exhaustive* (MEE). Each observation must belong to one and only one state. Also the mean of an indicator variable is the fraction or proportion of the sample satisfying that characteristic. Those means must sum to 1.0 across any complete set of indicator variables.

If tabulate generates a set of indicator variables $\mathbf{D}_{N\times g}$, where there are G groups (here, six), then $\mathbf{D}\boldsymbol{\iota} = \boldsymbol{\iota}$, where $\boldsymbol{\iota}$ is the units vector. If we sum the indicator variables across the g categories, we must produce an N-vector of ones. For that reason, we must drop one of the indicator variables when running a regression to avoid perfect collinearity with the constant term. We fit the regression model, dropping the first indicator variable (that for CT):

```
. tabulate state, generate(NE)
```

state	Freq.	Percent	Cum.
CT	20	16.67	16.67
MA	20	16.67	33.33
ME	20	16.67	50.00
NH	20	16.67	66.67
RI	20	16.67	83.33
VT	20	16.67	100.00
Total	120	100.00	

```
. regress dpipc NE2-NE6
```

Source	SS	df	MS	Number of obs =	120
				F(5, 114) =	5.27
Model	716.218512	5	143.243702	Prob > F =	0.0002
Residual	3099.85511	114	27.1917115	R-squared =	0.1877
				Adj R-squared =	0.1521
Total	3816.07362	119	32.0678456	Root MSE =	5.2146

dpipc	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
NE2	-2.549057	1.648991	-1.55	0.125	-5.815695	.7175814
NE3	-7.151959	1.648991	-4.34	0.000	-10.4186	-3.88532
NE4	-3.65752	1.648991	-2.22	0.029	-6.924158	-.3908815
NE5	-5.060575	1.648991	-3.07	0.003	-8.327214	-1.793937
NE6	-6.588007	1.648991	-4.00	0.000	-9.854646	-3.321369
_cons	22.32587	1.166013	19.15	0.000	20.01601	24.63573

This regression produces estimates of a constant term and five coefficients. We have excluded the first state (CT), so the constant term is the mean of CT values over time, identical to the `means` output above. The coefficients reported by `regress` represent the differences between each state's mean `dpipc` and that of CT.[3] The state means shown in the `mean` output above are six points on the real line. Are their differences statistically significant? It does not matter how we measure those differences, whether from the VT mean value of 15.7 or from the CT mean value of 22.3. Although we must exclude one state's indicator variable from the regression, the choice of the excluded class is arbitrary and will not affect the statistical judgments.

The test for relevance of the qualitative factor `state` is merely the ANOVA F statistic for this regression. The ANOVA F, as section 4.3.2 describes, tests the null hypothesis that all slope coefficients are jointly zero. In this context, that is equivalent to testing that all six state means of `dpipc` equal a common μ. The strong rejection of that hypothesis from the ANOVA F statistic implies that the New England states have significantly different levels of per capita disposable personal income.

Another transformation of indicator variables to produce *centered indicators* is often useful. If we create new indicators $d_i^* = d_i - d_g$, where d_g is the indicator for the excluded class, we can use the $(g-1)$ d_i^* variables in the model rather than the original d_i variables. As discussed above, the coefficients on the original d_i variables are contrasts with the excluded class. The d_i^* variables, which are trinary (taking on values of $-1, 0, 1$) will be contrasts with the grand mean. The constant term in the regression on d_i^* will be the grand mean, and the individual d_i^* coefficients are contrasts with that mean. To illustrate,

3. For instance, $22.32587 - 2.549057 = 19.77681$, the mean estimate for MA given above.

```
. forvalues i=1/5 {
  2.   generate NE_`i' = NE`i'-NE6
  3.   }
. regress dpipc NE_*
```

Source	SS	df	MS		Number of obs =	120
					F(5, 114) =	5.27
Model	716.218512	5	143.243702		Prob > F =	0.0002
Residual	3099.85511	114	27.1917115		R-squared =	0.1877
					Adj R-squared =	0.1521
Total	3816.07362	119	32.0678456		Root MSE =	5.2146

dpipc	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
NE_1	4.167853	1.064419	3.92	0.000	2.059247	6.276459
NE_2	1.618796	1.064419	1.52	0.131	-.48981	3.727402
NE_3	-2.984106	1.064419	-2.80	0.006	-5.092712	-.8754996
NE_4	.5103331	1.064419	0.48	0.633	-1.598273	2.618939
NE_5	-.8927223	1.064419	-0.84	0.403	-3.001328	1.215884
_cons	18.15802	.4760227	38.15	0.000	17.21502	19.10101

This algebraically equivalent model has the same explanatory power in terms of its ANOVA F statistic and R^2 as the model including five indicator variables. For example, $4.168 + 18.158 = 22.326$, the mean income in CT. Below we use `lincom` to compute the coefficient on the excluded class as minus the sum of the coefficients on the included classes.

```
. lincom -(NE_1+NE_2+NE_3+NE_4+NE_5)
 ( 1) - NE_1 - NE_2 - NE_3 - NE_4 - NE_5 = 0
```

dpipc	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
(1)	-2.420154	1.064419	-2.27	0.025	-4.52876	-.3115483

7.1.2 Regression with two qualitative measures

We can use two sets of indicator variables to evaluate the effects of two qualitative factors on a response variable. Take for example the Stata manual dataset `nlsw88`, an extract of the U.S. National Longitudinal Survey (NLSW) for employed women in 1988. We restrict the sample of 2,246 working women to a subsample for which data on hourly `wage`, `race`, and an indicator of `union` status are available. This step reduces the sample to 1,878 workers. We also have data on a measure of job `tenure` in years.

```
. use http://www.stata-press.com/data/imeus/nlsw88, clear
(NLSW, 1988 extract)
. keep if !missing(wage + race + union)
(368 observations deleted)
. generate lwage = log(wage)
```

```
. summarize wage race union tenure, sep(0)
```

Variable	Obs	Mean	Std. Dev.	Min	Max
wage	1878	7.565423	4.168369	1.151368	39.23074
race	1878	1.292332	.4822417	1	3
union	1878	.2454739	.4304825	0	1
tenure	1868	6.571065	5.640675	0	25.91667

We model `lwage`, the log of the reported wage, as the response variable. The variable `race` is coded 1, 2, or 3 for `white`, `black`, or `other`. We want to determine whether the variance in (log) wages is significantly related to the factors `race` and `union`. We cannot fit a regression model with two complete sets of dummies, so we will exclude one dummy from each group.[4] The regression estimates show the following:

```
. tabulate race, generate(R)
```

race	Freq.	Percent	Cum.
white	1,353	72.04	72.04
black	501	26.68	98.72
other	24	1.28	100.00
Total	1,878	100.00	

```
. regress lwage R1 R2 union
```

Source	SS	df	MS
Model	29.3349228	3	9.77830761
Residual	473.119209	1874	.252464893
Total	502.454132	1877	.267690001

```
Number of obs =    1878
F(  3,  1874) =   38.73
Prob > F      =  0.0000
R-squared     =  0.0584
Adj R-squared =  0.0569
Root MSE      =  .50246
```

lwage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
R1	-.0349326	.1035125	-0.34	0.736	-.2379444	.1680793
R2	-.2133924	.1049954	-2.03	0.042	-.4193126	-.0074721
union	.239083	.0270353	8.84	0.000	.1860606	.2921054
_cons	1.913178	.1029591	18.58	0.000	1.711252	2.115105

```
. test R1 R2  //          joint test for the effect of race
 ( 1)  R1 = 0
 ( 2)  R2 = 0
       F(  2,  1874) =   23.25
            Prob > F =    0.0000
```

A test for the significance of the qualitative factor `race` is the joint test for the coefficients of `R1`, `R2` equaling zero. When taking `other` as the excluded class for `race` we do not find that β_{R1} (the coefficient for `white`) differs from zero. But this coefficient is the contrast between the mean of `lwage` for `other` and the mean for `white`. The mean for `R2` (`black`), on the other hand, is distinguishable from that for `other`. These coefficients,

4. We could include one complete set of dummies in an equation without a constant term, but I do not recommend that approach. The absence of a constant term alters the meaning of many summary statistics.

taken together, reflect the effects of `race` on `lwage`. Those regressors should be kept or removed *as a group*. In particular, we should not use the t statistics for individual indicator variables to make inferences beyond noting, as above, the differences between group means. The magnitudes of those coefficients and their t statistics depend on the choice of excluded class, which is arbitrary.

The model of two qualitative factors illustrated here is a special case in that it assumes that the effects of the two qualitative factors are independent and strictly additive. That is, if you are black, your (log) wage is expected to be 0.213 lower than that of the `other` race category,[5] whereas if you are a union member, it is predicted to be 0.239 higher. What would this regression model predict that a black union member would earn, relative to the excluded class (a nonunion member of `other` race)? It would predict merely the sum of those two effects, or +0.026, since the union effect is slightly stronger than the black effect. We have a 3×2 two-way table of `race` and `union` categories. We can fill in the six cells of that table from the four coefficients estimated in the regression. For that approach to be feasible, we must assume independence of the qualitative effects so that the joint effect (reflected by a cell within the table) is the sum of the marginal effects. The effect of being black and a union member is taken to be the sum of the effects of being black, independent of union status, and that of being a union member, independent of `race`.

Interaction effects

Although sometimes this independence of qualitative factors is plausible, often it is not an appropriate assumption. Consider variations of the unemployment rate across age and race. Teenagers have a hard time landing a job because they lack labor market experience, so teenage unemployment rates are high relative to those of prime-aged workers. Likewise, minority participants generally have higher unemployment rates, whether due to discrimination or other factors such as the quality of their education. These two effects may not be merely additive. Perhaps being a minority teenager involves two strikes against you when seeking employment. If so, the effects of being both minority and a teenager are greater than the sum of their individual contributions. This reasoning implies that we should allow for *interaction effects* in evaluating these qualitative factors, which will allow their effects to be correlated, and requires that we estimate all six elements in the 3×2 table from the last regression example.

In regression, interactions involve products of indicator variables. Dummy variables may be treated as algebraic or Boolean. Adding indicator variables is equivalent to the Boolean "or" operator (|), denoting the union of two sets, whereas multiplying two indicator variables is equivalent to the Boolean "and" operator (&), denoting the intersection of sets. We may use either syntax in Stata's `generate` statements, remembering that we need to handle missing values properly.

5. This prediction translates into roughly 21%, using the rough approximation that $\log(1 + x) \simeq x$, although this approximation should really be used only for single-digit x.

How can we include a `race*union` interaction in the last regression? Since we need two `race` dummies to represent the three classes and one `union` dummy to reflect that factor, we need two interaction terms in the model: the interaction of each included `race` dummy with the `union` dummy. In the model

$$\texttt{lwage}_i = \beta_1 + \beta_2 \texttt{R1}_i + \beta_3 \texttt{R2}_i + \beta_4 \texttt{union}_i + \beta_5(\texttt{R1}_i \times \texttt{union}_i) + \beta_6(\texttt{R2}_i \times \texttt{union}_i) + u_i$$

the mean log wage for those in race `R1` (`white`) is $\beta_1 + \beta_2$ for nonunion members, but $\beta_1 + \beta_2 + \beta_4 + \beta_5$ for union members. Fitting this model yields the following:

```
. generate R1u = R1*union
. generate R2u = R2*union
. regress lwage R1 R2 union R1u R2u
```

Source	SS	df	MS	Number of obs =	1878
Model	33.3636017	5	6.67272035	F(5, 1872) =	26.63
Residual	469.09053	1872	.250582548	Prob > F =	0.0000
				R-squared =	0.0664
				Adj R-squared =	0.0639
Total	502.454132	1877	.267690001	Root MSE =	.50058

lwage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
R1	-.1818955	.1260945	-1.44	0.149	-.4291962	.0654051
R2	-.4152863	.1279741	-3.25	0.001	-.6662731	-.1642995
union	-.2375316	.2167585	-1.10	0.273	-.6626452	.187582
R1u	.4232627	.2192086	1.93	0.054	-.0066561	.8531816
R2u	.6193578	.2221704	2.79	0.005	.1836302	1.055085
_cons	2.07205	.1251456	16.56	0.000	1.82661	2.317489

```
. test R1u R2u    //    joint test for the interaction effect of race*union
 ( 1)  R1u = 0
 ( 2)  R2u = 0
       F(  2,  1872) =    8.04
            Prob > F =    0.0003
```

The joint test of the two interaction coefficients `R1u` and `R2u` rejects the null hypothesis of independence of the qualitative factors `race` and `union` at all conventional levels. Because the interaction terms are jointly significant, it would be a misspecification to fit the earlier regression rather than this expanded form. In regression, we can easily consider the model with and without interactions by merely fitting the model with interactions and performing the joint test that all interaction coefficients are equal to zero.

7.2 Regression with qualitative and quantitative factors

Earlier, we fitted several regression models in which all the regressors are indicator variables. In economic research, we often want to combine quantitative and qualitative information in a regression model by including both continuous and indicator regressors.

Returning to the nlsw88 dataset, we might model the log(wage) for qualitative factors race and union, as well as a quantitative factor tenure, the number of years worked in the current job. Estimation of that regression yields

```
. regress lwage R1 R2 union tenure
      Source |       SS       df       MS              Number of obs =    1868
-------------+------------------------------           F(  4,  1863) =   85.88
       Model |  77.1526731     4  19.2881683            Prob > F      =  0.0000
    Residual |  418.434693  1863  .224602626           R-squared     =  0.1557
-------------+------------------------------           Adj R-squared =  0.1539
       Total |  495.587366  1867  .265445831           Root MSE      =  .47392

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          R1 |   -.070349   .0976711    -0.72   0.471    -.2619053    .1212073
          R2 |  -.2612185   .0991154    -2.64   0.008    -.4556074   -.0668297
       union |   .1871116   .0257654     7.26   0.000     .1365794    .2376438
      tenure |   .0289352   .0019646    14.73   0.000     .0250823    .0327882
       _cons |   1.777386   .0975549    18.22   0.000     1.586058    1.968715
------------------------------------------------------------------------------

. test R1 R2  // joint test for the effect of race
 ( 1)  R1 = 0
 ( 2)  R2 = 0
       F(  2,  1863) =   29.98
            Prob > F =    0.0000
```

These results illustrate that this analysis-of-covariance model accounts for considerably more of the variation in lwage than does its counterpart based on only qualitative factors.[6] How might we interpret $\widehat{\beta}_{\texttt{tenure}}$? Using the standard approximation that $\log(1+x) \simeq x$,[7] we see that a given worker with 1 more year on her current job can expect to earn about 2.89% more (roughly, the semielasticity of wage with respect to tenure). How do we interpret the constant term? It is the mean log wage for a nonunion worker of other race with zero years of job tenure. Here that is a plausible category, since you might have less than 1 year's tenure in your current job. In other cases—for instance, where age is used as a regressor in a labor market study—the constant term may not correspond to any observable cohort.

The predictions of this model generate a series of parallel lines in {log(wage), tenure} space: a total of six lines, corresponding to the six possible combinations of race and union, with their intercepts computed from their coefficients and the constant term. We can separately test that those lines are distinct with respect to a qualitative factor: for instance, following the regression above, we jointly tested R1 and R2 for significance. If that test could not reject its null that each of those coefficients is zero, we would conclude that the {log(wage), tenure} profiles do not differ according to the qualitative factor race, and the six profiles would collapse to two.

6. I earlier noted that the form of this model with interaction terms was to be preferred; for pedagogical reasons, we return to the simpler form of the model.

7. See section 4.3.4.

Testing for slope differences

The model we have fitted is parsimonious and successful, given that it considers one quantitative factor. But are the true {log(wage), tenure} profiles parallel? Say that the unionized sector achieves larger annual wage increments by using its organized bargaining power. Might we expect two otherwise identical workers—one union, one nonunion—to have different profiles, with the unionized worker's profile steeper? To test that hypothesis, I return to the notion of an *interaction effect*, but here we interact a continuous measure (`tenure`) with the indicator variable `union`:

```
. quietly generate uTen = union*tenure
. regress lwage R1 R2 union tenure uTen
```

Source	SS	df	MS	Number of obs =	1868
				F(5, 1862) =	69.27
Model	77.726069	5	15.5452138	Prob > F =	0.0000
Residual	417.861297	1862	.224415304	R-squared =	0.1568
				Adj R-squared =	0.1546
Total	495.587366	1867	.265445831	Root MSE =	.47372

lwage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
R1	-.0715443	.0976332	-0.73	0.464	-.2630264	.1199377
R2	-.2638742	.0990879	-2.66	0.008	-.4582093	-.0695391
union	.2380442	.0409706	5.81	0.000	.157691	.3183975
tenure	.0309616	.0023374	13.25	0.000	.0263774	.0355458
uTen	-.0068913	.0043112	-1.60	0.110	-.0153467	.001564
_cons	1.766484	.0977525	18.07	0.000	1.574768	1.9582

The tenure effect is now measured as $\partial \texttt{lwage}/\partial \texttt{tenure} = \widehat{\beta}_{\texttt{tenure}}$ for nonunion members, but $(\widehat{\beta}_{\texttt{tenure}} + \widehat{\beta}_{\texttt{uTen}})$ for union members. The difference between those values is the estimated coefficient $\widehat{\beta}_{\texttt{uTen}}$, which is not significantly different from zero at the 10% level, but negative. Counter to our intuition, the data cannot reject the hypothesis that the slopes of the union and nonunion profiles are equal.

But what about the profiles for `race`? It is often claimed that minority hires are not treated equally over time, for instance, that promotions and larger increments go to whites rather than to blacks or Hispanics. We interact the race categories with `tenure`, in effect allowing the slopes of the {log(wage), tenure} profiles to differ by race:

```
. quietly generate R1ten = R1*tenure
. quietly generate R2ten = R2*tenure
. regress lwage R1 R2 union tenure R1ten R2ten
```

Source	SS	df	MS	Number of obs =	1868
				F(6, 1861) =	57.26
Model	77.2369283	6	12.8728214	Prob > F =	0.0000
Residual	418.350438	1861	.224798731	R-squared =	0.1558
				Adj R-squared =	0.1531
Total	495.587366	1867	.265445831	Root MSE =	.47413

lwage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
R1	-.082753	.1395	-0.59	0.553	-.3563459	.1908398
R2	-.291495	.1422361	-2.05	0.041	-.570454	-.012536
union	.1876079	.0257915	7.27	0.000	.1370246	.2381912
tenure	.0257611	.0186309	1.38	0.167	-.0107785	.0623007
R1ten	.0024973	.0187646	0.13	0.894	-.0343045	.0392991
R2ten	.0050825	.018999	0.27	0.789	-.032179	.0423441
_cons	1.794018	.1382089	12.98	0.000	1.522957	2.065078

```
. test R1ten R2ten
 ( 1)  R1ten = 0
 ( 2)  R2ten = 0
       F(  2,  1861) =    0.19
            Prob > F =    0.8291
```

We cannot reject the null hypothesis that both interaction coefficients are zero, implying that we do not have evidence against the hypothesis that one slope over categories of `race` suffices to express the effect of `tenure` on the wage. There does not seem to be evidence of *statistical discrimination* in wage increments, in the sense that the growth rates of female workers' wages do not appear to be race related.[8]

This last regression estimates five {log(wage), tenure} profiles, where the profiles for union members and nonunion members have equal slopes for a given race (with intercepts 0.188 higher for union members). We could fully interact `tenure` with both qualitative factors and estimate six {log(wage), tenure} profiles with different slopes:

```
. regress lwage R1 R2 union tenure uTen R1ten R2ten
```

Source	SS	df	MS		
Model	77.8008722	7	11.1144103	Number of obs	= 1868
Residual	417.786494	1860	.224616394	F(7, 1860)	= 49.48
				Prob > F	= 0.0000
				R-squared	= 0.1570
				Adj R-squared	= 0.1538
Total	495.587366	1867	.265445831	Root MSE	= .47394

lwage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
R1	-.0697096	.1396861	-0.50	0.618	-.3436676	.2042485
R2	-.2795277	.1423788	-1.96	0.050	-.5587668	-.0002886
union	.238244	.0410597	5.80	0.000	.1577161	.3187718
tenure	.0304528	.0188572	1.61	0.106	-.0065308	.0674364
uTen	-.0068628	.0043311	-1.58	0.113	-.0153572	.0016316
R1ten	-.0001912	.0188335	-0.01	0.992	-.0371283	.0367459
R2ten	.0023429	.0190698	0.12	0.902	-.0350576	.0397433
_cons	1.76904	.1390492	12.72	0.000	1.496331	2.041749

8. We could certainly use these findings to argue that black women with a given job tenure earn lower wages than do white women or those of other races, but that outcome could be related to other factors: the workers' ages, levels of education, employment location, and so forth.

```
. test uTen R1ten R2ten
 ( 1)  uTen = 0
 ( 2)  R1ten = 0
 ( 3)  R2ten = 0
       F(  3,  1860) =    0.96
            Prob > F =    0.4098
```

The joint test conducted here considers the null of one slope for all six categories versus six separate slopes. That null is not rejected by the data, so one slope will suffice.

Before leaving this topic, consider a simpler model in which we consider only the single indicator variable `union` and one quantitative measure, `tenure`. Compare the equation

$$\texttt{lwage}_i = \beta_1 + \beta_2 \texttt{union}_i + \beta_3 \texttt{tenure}_i + \beta_4(\texttt{union}_i \times \texttt{tenure}_i) + u_i \qquad (7.1)$$

with the equations

$$\begin{aligned} \texttt{lwage}_i &= \gamma_1 + \gamma_2 \texttt{tenure}_i + v_i, \ i \neq \texttt{union} \\ \texttt{lwage}_i &= \delta_1 + \delta_2 \texttt{tenure}_i + \omega_i, \ i = \texttt{union} \end{aligned} \qquad (7.2)$$

That is, we estimate separate equations from the nonunion and union cohorts. The point estimates of β from (7.2) are identical to those that may be computed from (7.1), but their standard errors will differ since the former are computed from smaller samples. Furthermore, when the two equations are estimated separately, each has its own σ^2 estimate. In estimating (7.1), we assume that u is homoskedastic over union and nonunion workers, but that may not be an appropriate assumption. From a behavioral standpoint, collective bargaining may reduce the *volatility* of wages (e.g., by ruling out merit increments in favor of across-the-board raises), regardless of the effects of collective bargaining on the *level* of wages. Estimating these equations for the `nlsw88` data illustrates these points. First, I present the regression over the full sample:

```
. regress lwage union tenure uTen
```

Source	SS	df	MS	Number of obs =	1868
Model	64.0664855	3	21.3554952	F(3, 1864) =	92.25
Residual	431.52088	1864	.231502618	Prob > F =	0.0000
				R-squared =	0.1293
Total	495.587366	1867	.265445831	Adj R-squared =	0.1279
				Root MSE =	.48115

lwage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
union	.2144586	.0414898	5.17	0.000	.1330872	.29583
tenure	.0298926	.0023694	12.62	0.000	.0252456	.0345395
uTen	-.0056219	.0043756	-1.28	0.199	-.0142035	.0029597
_cons	1.655054	.0193938	85.34	0.000	1.617018	1.69309

The t test for `uTen` indicates that the effects of `tenure` do not differ significantly across the classifications. We now fit the model over the union and nonunion subsamples:

```
. regress lwage tenure if !union
```

Source	SS	df	MS	Number of obs =	1408
				F(1, 1406) =	148.43
Model	36.8472972	1	36.8472972	Prob > F =	0.0000
Residual	349.032053	1406	.248244703	R-squared =	0.0955
				Adj R-squared =	0.0948
Total	385.87935	1407	.274256823	Root MSE =	.49824

lwage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
tenure	.0298926	.0024536	12.18	0.000	.0250795	.0347056
_cons	1.655054	.0200828	82.41	0.000	1.615659	1.69445

```
. predict double unw if e(sample), res
(470 missing values generated)
. regress lwage tenure if union
```

Source	SS	df	MS	Number of obs =	460
				F(1, 458) =	55.95
Model	10.0775663	1	10.0775663	Prob > F =	0.0000
Residual	82.4888278	458	.180106611	R-squared =	0.1089
				Adj R-squared =	0.1069
Total	92.5663941	459	.201669704	Root MSE =	.42439

lwage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
tenure	.0242707	.0032447	7.48	0.000	.0178944	.0306469
_cons	1.869513	.0323515	57.79	0.000	1.805937	1.933088

```
. predict double nunw if e(sample), res
(1418 missing values generated)
```

The `Root MSE` values are different for the two subsamples and could be tested for equality as described in section 6.2.2's treatment of groupwise heteroskedasticity:[9]

```
. generate double allres = nunw
(1418 missing values generated)
. replace allres = unw if unw<.
(1408 real changes made)
. sdtest allres, by(union)
Variance ratio test
```

Group	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf.	Interval]
nonunion	1408	5.19e-17	.0132735	.4980645	-.0260379	.0260379
union	460	6.47e-17	.0197657	.4239271	-.0388425	.0388425
combined	1868	5.50e-17	.0111235	.4807605	-.0218157	.0218157

```
    ratio = sd(nonunion) / sd(union)                 f =   1.3803
Ho: ratio = 1                     degrees of freedom = 1407, 459
    Ha: ratio < 1               Ha: ratio != 1                 Ha: ratio > 1
  Pr(F < f) = 1.0000         2*Pr(F > f) = 0.0000           Pr(F > f) = 0.0000
```

9. We could instead use `egen double allres = rowtotal(nunw unw)`, but we would then have to use `replace allres=. if nunw ==. & unw==.` to deal with observations missing from both subsamples. Those observations would otherwise be coded as zeros.

We conclude that contrary to our prior results, nonunion workers have a significantly smaller variance of their disturbance process than union members. We should either correct for the heteroskedasticity across this classification or use robust standard errors to make inferences from a model containing both union and nonunion workers. To illustrate the latter point:

```
. regress lwage union tenure uTen, robust
Linear regression                               Number of obs =    1868
                                                F(  3,  1864) =  109.84
                                                Prob > F      =  0.0000
                                                R-squared     =  0.1293
                                                Root MSE      =  .48115
------------------------------------------------------------------------------
             |               Robust
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       union |   .2144586   .0407254     5.27   0.000     .1345864    .2943308
      tenure |   .0298926   .0023964    12.47   0.000     .0251928    .0345924
        uTen |  -.0056219   .0038631    -1.46   0.146    -.0131984    .0019546
       _cons |   1.655054   .0210893    78.48   0.000     1.613693    1.696415
------------------------------------------------------------------------------
```

Although robust standard errors increase the t statistic for uTen, the coefficient is not significantly different from zero at any conventional level of significance. We conclude that an interaction of tenure and union is not required for proper specification of the model.

7.3 Seasonal adjustment with indicator variables

Economic data with a time-series dimension often must be *seasonally adjusted*. For instance, monthly sales data for a set of retail firms will have significant variations around the holidays, and quarterly tax collections for municipalities located in a tourist area will fluctuate widely between the tourist season and off-season. A common method of seasonal adjustment involves modeling the seasonal factor in the time series as being either additive or multiplicative. An *additive* seasonal factor increases (decreases) the variable by the same *dollar amount* every January (or first quarter), with the amount denominated in units of the variable. In contrast, a *multiplicative* seasonal factor increases (decreases) the variable by the same *percentage* every January (or first quarter).

The primary concern here is that some economic data are made available in seasonally adjusted (SA) form. For flow series such as personal income, this concept is often indicated as seasonally adjusted at an annual rate (SAAR). Other economic data that may be used in a model of household or firm behavior are denoted as not seasonally adjusted (NSA). The two types of data should not be mixed in the same model: for instance, an NSA response variable versus a set of regressors, each of which is SA. Such a regression will contain seasonality in its residuals and will fail any test for independence of the errors that considers AR(4) models (for quarterly data) or AR(12) models (for monthly data). If we recognize that there are seasonal components in one or more data

series, we should use some method of seasonal adjustment unless all series in the model are NSA.

Deseasonalization with either the additive or multiplicative form of the seasonal model requires that a set of *seasonal dummies* be created by defining the elements of the set with statements like

```
. generate mseas1 = (month(dofm(datevar)) == 1)
. generate qseas1 = (quarter(dofq(datevar)) == 1)
```

for data that have been identified as monthly or quarterly data to Stata, respectively, by `tsset datevar`. The variable `mseas1` will be 1 in January and 0 in other months; `qseas1` will be 1 in the first quarter of each year and 0 otherwise. The `month()` and `quarter()` functions, as well as the more arcane `dofm()` and `dofq`, are described in [D] **functions** under the headings *Date functions* and *Time-series functions*. The set of seasonal dummies is easily constructed with a `forvalues` loop, as shown in the example below.

To remove an additive seasonal factor from the data, we regress the series on a constant term and all but one of the seasonal dummies

```
. regress sales mseas*
. regress taxrev qseas*
```

for monthly or quarterly data, respectively. After the regression, we use `predict` with the `residuals` option to produce the deseasonalized series. Naturally, this series will have a mean of zero, since it comes from a regression with a constant term; usually it is "rebenched" to the original series' mean, as I illustrate below. We use the `turksales` dataset, which contains quarterly turkey sales data for 1990q1–1994q4, as described by `summarize`:

```
. use http://www.stata-press.com/data/imeus/turksales, clear
. summarize sales
```

Variable	Obs	Mean	Std. Dev.	Min	Max
sales	40	105.6178	4.056961	97.84603	112.9617

We first find the mean of the quarterly `sales` series and generate three quarterly dummy variables:

```
. summarize sales, meanonly
. local mu = r(mean)
. forvalues i=1/3 {
  2.         generate qseas`i' = (quarter(dofq(t)) == `i')
  3. }
```

We then run the regression to evaluate the importance of seasonal factors:

```
. regress sales qseas*
```

Source	SS	df	MS	Number of obs =	40
				F(3, 36) =	4.03
Model	161.370376	3	53.7901254	Prob > F =	0.0143
Residual	480.52796	36	13.3479989	R-squared =	0.2514
				Adj R-squared =	0.1890
Total	641.898336	39	16.4589317	Root MSE =	3.6535

sales	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
qseas1	-5.232047	1.633891	-3.20	0.003	-8.545731	-1.918362
qseas2	-2.842753	1.633891	-1.74	0.090	-6.156437	.4709317
qseas3	-.8969368	1.633891	-0.55	0.586	-4.210621	2.416748
_cons	107.8608	1.155335	93.36	0.000	105.5177	110.2039

The ANOVA F statistic from the regression indicates that seasonal factors explain much of the variation in `sales`. To generate the deseasonalized series, we use `predict` to recover the residuals and add the original mean of the series to them:

```
. predict double salesSA, residual
. replace salesSA = salesSA + `mu'
(40 real changes made)
```

We can now compare the two series:

```
. summarize sales salesSA
```

Variable	Obs	Mean	Std. Dev.	Min	Max
sales	40	105.6178	4.056961	97.84603	112.9617
salesSA	40	105.6178	3.510161	97.49429	111.9563

```
. label var salesSA "sales, seasonally adjusted"
. tsline sales salesSA, lpattern(solid dash)
```

The deseasonalized series has a smaller standard deviation than the original as the fraction of the variation because the seasonality has been removed. This effect is apparent in the graph of the original series and the smoother deseasonalized series in figure 7.1.

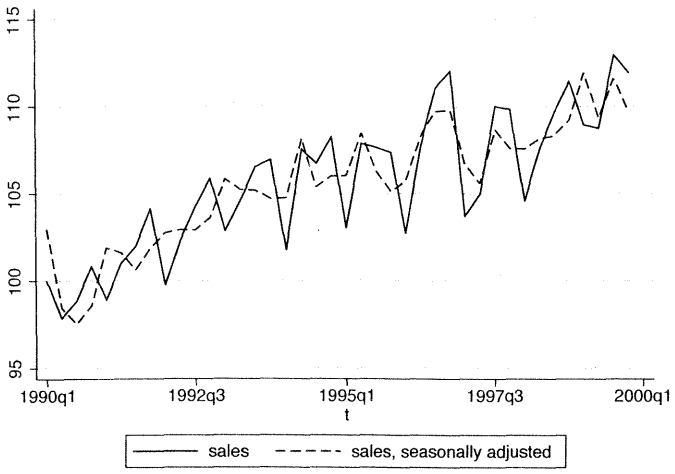

Figure 7.1: Seasonal adjustment of time series

We may also want to remove the trend component from a series. To remove a linear trend, we merely regress the series on a time trend. For a multiplicative (geometric, or constant growth rate) trend, we regress the logarithm of the series on the time trend. In either case, the residuals from that regression represent the detrended series.[10] We may remove both the trend and seasonal components from the series in the same regression, as illustrated here:

```
. regress sales qseas* t
```

Source	SS	df	MS
Model	552.710487	4	138.177622
Residual	89.1878487	35	2.54822425
Total	641.898336	39	16.4589317

Number of obs	= 40
F(4, 35)	= 54.23
Prob > F	= 0.0000
R-squared	= 0.8611
Adj R-squared	= 0.8452
Root MSE	= 1.5963

sales	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
qseas1	-4.415311	.7169299	-6.16	0.000	-5.870756	-2.959866
qseas2	-2.298262	.7152449	-3.21	0.003	-3.750287	-.846238
qseas3	-.6246916	.7142321	-0.87	0.388	-2.07466	.8252766
t	.2722452	.0219686	12.39	0.000	.2276466	.3168438
_cons	69.47421	3.138432	22.14	0.000	63.10285	75.84556

```
. test qseas1 qseas2 qseas3
 ( 1)  qseas1 = 0
 ( 2)  qseas2 = 0
 ( 3)  qseas3 = 0
       F(  3,    35) =   15.17
            Prob > F =    0.0000
```

10. For more detail, see Davidson and MacKinnon (2004, 72–73).

```
. predict double salesSADT, residual
. replace salesSADT = salesSADT + 'mu'
(40 real changes made)
. label var salesSADT "sales, detrended and SA"
. tsline sales salesSADT, lpattern(solid dash) yline('mu')
```

The trend `t` is highly significant in these data. A joint F test for the seasonal factors shows that they are also significant beyond a trend term. The detrended and deseasonalized series, rebenched to the mean of the original series (shown by the horizontal line), is displayed in figure 7.2.

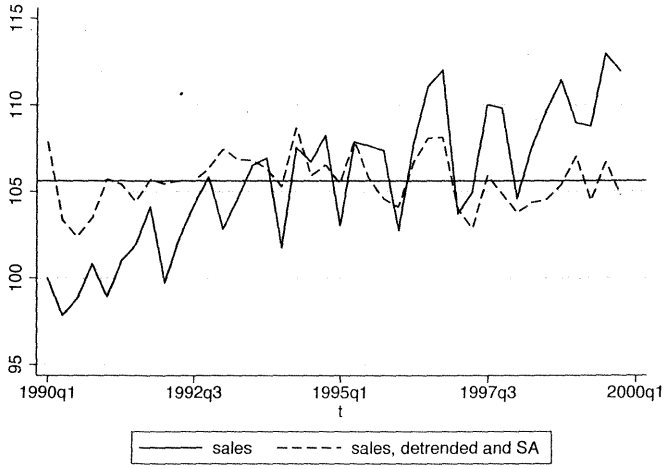

Figure 7.2: Seasonal adjustment and detrending of time series

Several other methods of seasonal adjustment and detrending for time-series data are implemented in Stata under the heading `tssmooth`; see in particular [TS] **tssmooth shwinters**. As Davidson and MacKinnon (2004, 584–585) point out, the seasonal adjustment methods used by government statistics bureaus can be approximated by a *linear filter*, or τ-term moving average. In this context, `tssmooth ma` or the `egen` function `filter()` available in the `egenmore` package from `ssc` may be helpful.

If you are interested in filtering time-series data to identify business cycles, see the author's `bking` (Baxter–King bandpass filter) and `hprescott` (Hodrick–Prescott filter) routines, both available from the SSC archive (see [R] **ssc**).

7.4 Testing for structural stability and structural change

Indicator variables are used to test for structural stability in a regression function in which we specify a priori the location of the possible structural breakpoints. In (7.1) and (7.2), we found that the intercept of the regression differed significantly between union and nonunion cohorts but that one slope parameter for `tenure` was adequate. In further testing, we found that the σ_u^2 differed significantly between these two cohorts in the sample. If we doubt structural stability—for instance, an industry-level regression over a set of natural resource–intensive and manufacturing industries—we may use indicator variables to identify groups within the sample and test whether the intercept and slope parameters are stable over these groups. In household data, a function predicting food expenditures might not be stable over families with different numbers of children. Merely including the number of children as a regressor might not be adequate if this relationship is nonlinear in the number of mouths to feed.

Structural instability over cohorts of the sample need not be confined to shifts in the intercept of the relationship. A structural shift may not be present in the intercept, but it may be an important factor for one or more slope parameters. If we question structural stability, we should formulate a general model in which all regressors (including the constant term) are interacted with cohort indicators and test down where coefficients appear to be stable across cohorts.

Section 6.2.2 considers the possibility of heteroskedasticity over groups or cohorts in the data that may have been pooled. Beyond the possibility that σ_u^2 may differ across groups, we should be concerned with the stability of the regression function's coefficients over the groups. Whereas groupwise heteroskedasticity may be readily diagnosed and corrected, improperly specifying the regression function to be constant over groups of the sample will be far more damaging, rendering regression estimates biased and inconsistent. For instance, if those firms who are subject to liquidity constraints (because of poor credit history or inadequate collateral) behave differently from firms that have ready access to financial markets, combining both sets of firms in the same regression will yield a regression function that is a mix of the two groups' dissimilar behavior. Such a regression is unlikely to provide reasonable predictions for firms in *either* group. Placing the two groups in the same regression, with indicator variables used to allow for potential differences in structure between their coefficient vectors, is more sensible. That approach will allow those differences to be estimated and tested for significance.

7.4.1 Constraints of continuity and differentiability

It is easy to determine that the regression function should be allowed to exhibit various structural breaks. Tests may show that a representative worker's earnings–tenure profile should be allowed to have different slopes over different ranges of job tenure. You could accomplish this configuration by using a polynomial in tenure, but doing so may introduce unacceptable behavior (for instance, with `tenure` and `tenure`^{2}, there must be some tenure at which the profile turns downward, predicting that wages will fall with each additional year on the job). If we use the interaction terms with no further

constraints on the regression function, that piecewise linear function exhibits discontinuities over the groups identified by the interaction terms (e.g., the age categories in the sample). I illustrate, returning to the NLSW dataset and defining four job tenure categories: fewer than 2 years, 2–7 years, 7–12 years, and more than 12 years:

```
. use http://www.stata-press.com/data/imeus/nlsw88, clear
(NLSW, 1988 extract)
. generate lwage = log(wage)
. generate Ten2 = tenure<=2
. generate Ten7 = !Ten2 & tenure<=7
. generate Ten12 = !Ten2 & !Ten7 & tenure<=12
. generate Ten25 = !Ten2 & !Ten7 & !Ten12 & tenure<.
```

We now generate interactions of `tenure` with each of the tenure categories, run the regression on the categories and interaction terms,[11] and generate predicted values:

```
. generate tTen2 = tenure*Ten2
(15 missing values generated)
. generate tTen7 = tenure*Ten7
(15 missing values generated)
. generate tTen12 = tenure*Ten12
(15 missing values generated)
. generate tTen25 = tenure*Ten25
(15 missing values generated)
. regress lwage Ten* tTen*, nocons hascons
```

Source	SS	df	MS		
Model	76.6387069	7	10.9483867	Number of obs	= 2231
Residual	655.578361	2223	.294907045	F(7, 2223)	= 37.12
				Prob > F	= 0.0000
				R-squared	= 0.1047
				Adj R-squared	= 0.1018
Total	732.217068	2230	.328348461	Root MSE	= .54305

lwage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
Ten2	1.55662	.0383259	40.62	0.000	1.481462	1.631778
Ten7	1.708728	.060084	28.44	0.000	1.590901	1.826554
Ten12	1.870808	.1877798	9.96	0.000	1.502566	2.23905
Ten25	1.751961	.1691799	10.36	0.000	1.420194	2.083728
tTen2	.0897426	.0331563	2.71	0.007	.0247221	.1547631
tTen7	.0434089	.0140739	3.08	0.002	.0158095	.0710083
tTen12	.0154208	.019786	0.78	0.436	-.0233801	.0542218
tTen25	.0238014	.0102917	2.31	0.021	.0036191	.0439837

```
. predict double lwagehat
(option xb assumed; fitted values)
(15 missing values generated)
. label var lwagehat "Predicted log(wage)"
. sort tenure
```

11. We exclude the constant term so that all four tenure dummies can be included. The option `hascons` indicates to Stata that we have the equivalent of a constant term in the four tenure dummies `Ten2`-`Ten25`.

The predicted values for each segment of the wage–tenure profile can now be graphed:

```
. twoway (line lwagehat tenure if tenure<=2)
> (line lwagehat tenure if tenure>2 & tenure<=7)
> (line lwagehat tenure if tenure>7 & tenure<=12)
> (line lwagehat tenure if tenure>12 & tenure<.), legend(off)
```

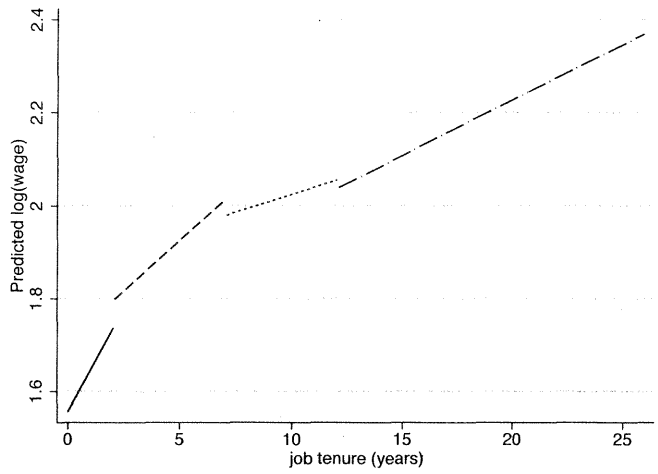

Figure 7.3: Piecewise wage–tenure profile

As we see in figure 7.3, this piecewise function allows for a different slope and intercept for each of the four ranges of job tenure, but it is not continuous. For instance, the estimates predict that at the point of 2 years' tenure, the average worker's log wage will abruptly jump from 1.73 per hour to 1.80 per hour and then decline from 2.01 per hour to 1.98 per hour at the point of 7 years' tenure.

We may want to allow such a profile to be flexible over different ranges of job tenure but force the resulting function to be *piecewise continuous* by using a *linear spline*: a mathematical function that enforces continuity between the adjacent segments. Spline functions are characterized by their *degree*. A linear spline is degree 1, a quadratic spline is degree 2, and so on. A linear spline will be continuous but not differentiable at the *knot points*: those points on the profile that define segments of the function. A quadratic spline is continuous and once differentiable. Since the function has constant first derivatives on both sides of the knot, there will be no kinks in the curve. Likewise, a cubic spline will be continuous and twice differentiable, and so on.

I illustrate using a linear spline to generate a piecewise continuous earnings–tenure profile. Stata's `mkspline` (see [R] **mkspline**) command automates this process for linear splines. Higher-order splines must be defined algebraically or by using a user-written

routine. We can use the `mkspline` command to generate a spline with knots placed at specified points or a spline with equally spaced knots.[12] Here we use the former syntax:

`mkspline` *newvar*$_1$ *#1* [*newvar*$_2$ *#2* [...]] *newvar*$_k$ = *oldvar* [*if*] [*in*]

where k *newvar*s are specified to define a linear spline of *varname* with $(k-1)$ knots, placed at the values *#1*, *#2*, ..., $\#(k-1)$ of the splined variable. The resulting set of *newvarname* variables may then be used as regressors.

In the piecewise regression above, we estimated four slopes and four intercepts for a total of eight regression parameters. Fitting this model as a linear spline places constraints on the parameters. At each of the three knot points (2, 7, and 12 years) along the tenure axis, $\gamma+\delta$ `tenure` must be equal from the left and right. Simple algebra shows that each of the three knot points imposes one constraint on the parameter vector. The piecewise linear regression using a linear spline will have five parameters rather than eight:

```
. mkspline sTen2 2 sTen7 7 sTen12 12 sTen25 = tenure
. regress lwage sTen*
```

Source	SS	df	MS	Number of obs =	2231
				F(4, 2226) =	64.55
Model	76.1035947	4	19.0258987	Prob > F =	0.0000
Residual	656.113473	2226	.294749988	R-squared =	0.1039
				Adj R-squared =	0.1023
Total	732.217068	2230	.328348461	Root MSE =	.54291

lwage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
sTen2	.1173168	.0248619	4.72	0.000	.0685619	.1660716
sTen7	.0471177	.009448	4.99	0.000	.02859	.0656455
sTen12	.0055041	.0111226	0.49	0.621	-.0163076	.0273158
sTen25	.0237767	.0083618	2.84	0.005	.007379	.0401744
_cons	1.539985	.0359605	42.82	0.000	1.469465	1.610505

```
. predict double lwageSpline
(option xb assumed; fitted values)
(15 missing values generated)
. label var lwageSpline "Predicted log(wage), splined"
. twoway line lwageSpline tenure
```

The result of the piecewise linear estimation, displayed in figure 7.4, is a continuous earnings–tenure profile with kinks at the three knot points. From an economic standpoint, the continuity is highly desirable. The model's earnings predictions for tenures of 1.9, 2.0, and 2.1 years will now be smooth, without implausible jumps at the knot points.

12. The alternative syntax can also place knots at equally spaced percentiles of the variable with the `pctile` option.

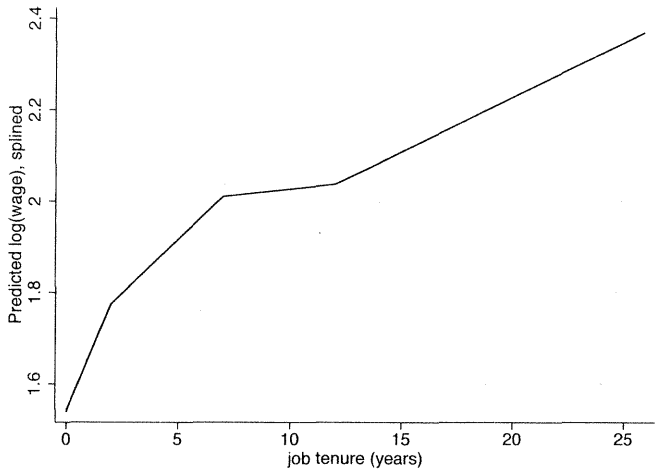

Figure 7.4: Piecewise linear wage–tenure profile

7.4.2 Structural change in a time-series model

With time-series data, a concern for structural stability is usually termed a test for *structural change*. We can allow for different slopes or intercepts for different periods in a time-series regression (e.g., allowing for a household consumption function to shift downward during wartime). Just as in a cross-sectional context, we should consider that both intercept and slope parameters may differ over various periods. Older econometrics texts often discuss this difference in terms of a *Chow test* and provide formulas that manipulate error sums of squares from regressions run over different periods to generate a test statistic. This step is not necessary since the Chow test is nothing more than the F test that all *regime dummy* coefficients are jointly zero. For example,

$$y_t = \beta_1 + \beta_2 x_{2t} + \beta_3 x_{3t} + \beta_4 gw_t + \beta_5 (x_{2t} \times gw_t) + \beta_6 (x_{3t} \times gw_t) + u_t$$

where $gw_t = 1$ during calendar quarters of the Gulf War. The joint test $\beta_4 = \beta_5 = \beta_6 = 0$ would test that this regression function is stable during the two regimes. We may also consider the intermediate cases: for instance, the coefficient on x_2 may be stable over peacetime and wartime, but the coefficient on x_3 (or the intercept) may not. We can easily handle more than two regimes by merely adding regime dummies for each regime and their interactions with the other regressors. We should also be concerned about the realistic possibility that the σ_u^2 has changed over regimes. We may deal with this possibility by computing robust standard errors for the regression with regime dummies, but we might want to estimate the differing variances for each regime, as this is a sort of groupwise heteroskedasticity where the groups are time-series regimes.

Sometimes a regime may be too short to set up the fully interacted model since it requires that the regression model be fitted over the observations of that regime. Since the model above contains three parameters per regime, it cannot be estimated over a regime with 4 or fewer observations. This problem often arises at the end of a time series. We may want to test the hypothesis that the last T_2 observations were generated by the same regime as the previous T_1 observations. Then we construct an F test by estimating the regression over all $T = T_1 + T_2$ observations and then estimating it again over the first T_1 observations. The sum of squared residuals ($\Sigma\widehat{u}_t^2$) for the full sample will exceed that from the first T_1 observations unless the regression fits perfectly over the additional T_2 data points. If the fit is very poor over the additional T_2 data points, we can reject the null of model stability over $[T_1, T_2]$. This Chow predictive F test has T_2 degrees of freedom in the numerator:

$$F(T_2, T_1 - k) = \frac{(\widehat{\mathbf{u}}_T'\widehat{\mathbf{u}}_T - \widehat{\mathbf{u}}_{T1}'\widehat{\mathbf{u}}_{T1})/T_2}{(\widehat{\mathbf{u}}_{T1}'\widehat{\mathbf{u}}_{T1})/(T_1 - k)}$$

where $\widehat{\mathbf{u}}_T$ is the residual vector from the full sample. Following a regression, the error sum of squares may be accessed as `e(rss)` (see [P] **ereturn**).

These dummy variable methods are useful when the timing of one or more structural breaks is known a priori from the economic history of the period. However, we often are not sure whether (and if so, when) a relationship may have undergone a structural shift. This uncertainty is particularly problematic when a change may be a gradual process rather than an abrupt and discernible break. Several tests have been devised to evaluate the likelihood that a change has taken place, and if so, when that break may have occurred. Those techniques are beyond the scope of this text. See Bai and Perron (2003).

Exercises

1. Using the dataset of section 7.1.2, test that `race` explains much of the variation in `lwage`.
2. Consider the model used in section 7.2 to search for evidence of statistical discrimination. Test a model that includes interactions of the factors `race` and `tenure`.
3. Consider the model used in section 7.3 to seasonally adjust turkey sales data. Fit a multiplicative seasonal model to these data. Is an additive seasonal factor or a multiplicative seasonal factor preferred?
4. Consider the model used in section 7.3 to seasonally adjust turkey sales data. Apply Holt–Winters seasonal smoothing (`tssmooth shwinters`), and compare the resulting series to that produced by seasonal adjustment with indicator variables.
5. Consider the model used in section 7.4.1. Use the alternate syntax of `mkspline` to generate three equally placed knots and estimate the equation. Repeat the exercise, using the `pctile` option. How sensitive are the results to the choice of linear spline technique?