

Stata Walkthrough 4: Regression, Prediction, and Forecasting

Over drinks the other evening, my neighbor told me about his 25-year-old nephew, who is dating a 35-year-old woman. “God, I can’t see them getting married,” he said. I raised my eyebrow, because as an economic demographer, I know that spouses’ ages are very predictable: they tend to be similar, with the husband just a couple of years older than his wife. Strange cases do occur, but they tend to involve older men and younger women. The opposite is fairly unlikely. However, I didn’t know exactly what would constitute a range of “likely outcomes” for the age of the woman that a 25-year-old guy would marry.

In this Stata exercise, we’re going to do two things:

1. We will use data on married couples ages’ to estimate the relationship between spouses’ ages.
2. We will use this to predict a range of “likely outcomes” for the age of the wife of a twenty-five-year-old man.

You’ll need to load the database of “U.S. married couples, March 2005” (**marriedmar05.dta**) from my webpage. These data come from the Current Population Survey (done by the Bureau of Labor Statistics), and they are a random sample of all households in the U.S. I have restricted the sample to only couples that identify themselves as “married.” We have 34,674 of these couples.

In general, you should begin any project by exploring the data. In this case, that’s simple, since this database contains two variables: **hage** and **wage**, the husband’s age and the wife’s age. First, let’s look at some descriptive statistics: type **summarize**.

Variable	Obs	Mean	Std. Dev.	Min	Max
wage	34674	45.78664	13.26782	15	79
hage	34674	48.07369	13.62849	15	79

Values of each variable range from 15 to 79. The mean age of a married man is 48.1 years, and the mean age of a married woman is 45.8 years. We can see that there’s a strong correlation between the two variables by typing **corr hage wage**:

	wage	hage
wage	1.0000	
hage	0.9243	1.0000

In other words, older men tend to be married to older women—this should be no surprise. However, the sample means reveal that married men are older than married women on average, so we can also infer that typically each husband must be older than his wife.

We might be interested in the distribution of the age difference, so let's create a new variable for the difference in ages:

```
gen dage = hage - wage
label variable dage "Difference between spouses' ages"
```

Now let's look at the frequency distribution of this variable by typing

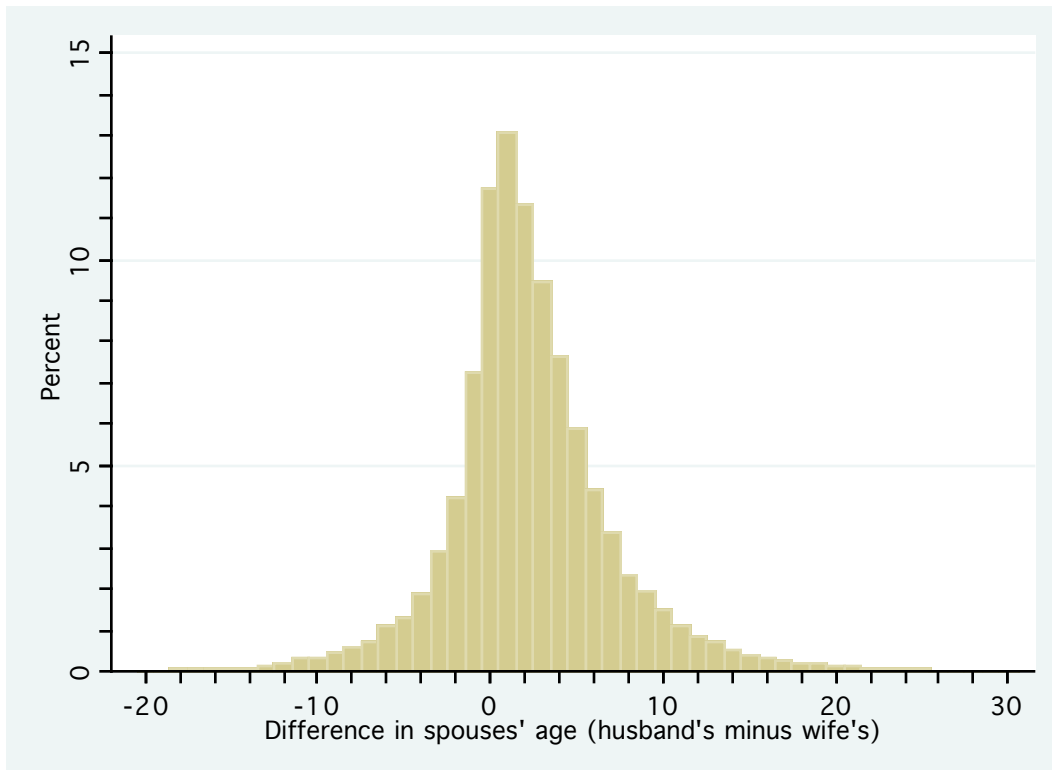
```
histogram dage
```

The image is a bit ugly. First, the graph goes from -50 to +50, even though there's basically nobody out in those tails. (There are some people, but they're not significant enough to register in the graph.) For all practical purposes, the range goes from -20 to +30, so we'll restrict it to that. Second, Stata is using a default bin-width of 2.44 units. There's a natural width in this case, I'd think: one, a single year. Let's tell Stata to make each bin of width one. Third, there aren't nearly enough tick marks or labels on the axes to give us a clear indication of the values, so we'll add those. Finally, I think it's easier to interpret percentages than frequencies, so we'll specify that option. Here's the complete command and the output:

```
histogram dage if dage >= -20 & dage <= 30, width(1)
ylabel(0 5 10 15) ytick(1(1)15) xlabel(-20 (10) 30)
xtick(-20(2) 30) percent
```

There's still one small problem: the bins start just to the right of the corresponding number. We can correct this by specifying **start(-20.5)** in our list of options; that way, Stata would place the bin for -20.5 to -19.49 (including the values of -20) right over the number -20 in the graph, and so forth. (To do this correction, click on the last command that you typed, in the "Review" window; it should appear in your

command line, and you can tack on the extra option.) The final output should look this:



You can see the asymmetry in the distribution: the frequency of “-1” (men who are one year younger than their wives) is roughly the same as the frequency of “+3” (men who are three years older than their wives). The frequency of “-2” is about the same as the frequency of “+6”. Finally, only a small number (0.33%) of men are ten years younger than their wives (the match that my neighbor worries about). I knew that it was unlikely.

This isn’t a perfect answer to the question, however—it tells us the chance for the overall population. We have information about the specific age of this person (he is twenty-five), and we should use that knowledge to improve our predictions. The relationship might not be the same for all ages. For example, it could be the case that men typically marry women almost exactly their own age initially; then, when they get to a midlife crisis (or a seven-year-itch, whichever comes first), they trade in their spouse for a younger model. That could generate something like the distribution that we observe; however, it would be inaccurate to claim that the distribution of wives’ ages is the same for men at all ages. After all, our data says that

7.0% of men are ten years older than their wives—but surely this isn’t true for seventeen-year-old grooms?

We want to predict the wife’s age as a function of the husband’s age. We will assume that the relationship can be described as:

$$wage_i = \beta_0 + \beta_1 \cdot hage_i + e_i$$

One note: when we’re doing prediction, we don’t necessarily have to believe that the husband’s age *causes* the wife’s age (even though we model it that way).

From the descriptive statistics, we have enough information to calculate our best prediction of $\hat{\beta}_1$ and $\hat{\beta}_0$:

$$\hat{\beta}_1 = r_{xy} \frac{s_y}{s_x} = (0.9243) \left(\frac{13.26782}{13.62849} \right) = 0.8998$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} = 45.78664 - (0.8998)48.07369 = 2.5300$$

However, it’s generally easier to let Stata do the work. Type **reg wage hage**, and you should get the output:

Source	SS	df	MS	Number of obs = 34674		
Model	5214158.71	1	5214158.71	F(1, 34672) =	.	
Residual	889502.864	34672	25.6547896	Prob > F =	0.0000	
				R-squared =	0.8543	
				Adj R-squared =	0.8543	
Total	6103661.57	34673	176.035	Root MSE =	5.0651	

wage	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
hage	.8998053	.0019959	450.82	0.000	.8958933	.9037174
_cons	2.529682	.0997317	25.36	0.000	2.334205	2.72516

Let’s look first at the bottom portion of this table. The estimated coefficients ($\hat{\beta}_1$ and $\hat{\beta}_0$) are exactly what we found using the formulas. In addition, Stata calculates the “standard error” in each estimate ($s_{\hat{\beta}_1}$ and $s_{\hat{\beta}_0}$, in the notation from class). At the right-hand side, it gives us a 95% confidence interval for each parameter. We can very quickly test a (two-tailed) hypothesis using these. In the top-right of the

table, we have the R^2 value of the regression: 0.8543. This means that husbands' ages can predict 85.43% of the variation in wives' ages. . (Remember that the R^2 is also equal to the model sum of squares divided by the total sum of squares: $5214158.71 / 6103661.57 = 0.8543$. We can also calculate the R^2 value as simply the correlation coefficient, squared? Well, $(0.9243)^2 = 0.8543$.) Finally, in the top-left, we have the model (explained), residual, and total sum of squares.

Next, we want to predict the expected age of each person's spouse. There are three things that we might want to know: the best single guess of the wife's average or actual age (these are the same), a range of estimates for the *average* age, or a range of estimates for the *actual* age.

We can obtain a point estimate of the wife's age for each person by using the formula:

$$wagehat_i = \hat{\beta}_0 + \hat{\beta}_1 \cdot hage_i$$

To do this in Stata, we could type:

```
gen wagehat = 2.529682 + .8998053 * hage
```

This is a bit time-consuming and error-prone, so Stata has a pre-programmed shortcut for many post-estimation predictions. You can read **help regress postestimation** for a complete description, but let me give you a partial list:

Command	What it does
predict varname1, xb	Calculates \hat{y}_i for each observation, and saves it to a new variable called varname1
predict varname2, r	Calculates the residual $\hat{e}_i = y_i - \hat{y}_i$ for each observation, and saves it to a new variable called varname2
predict varname3, stdp	Calculates the standard deviation in prediction $s_{\hat{y}}$ for each observation, and saves it to a new variable called varname3
predict varname4, stdf	Calculates the standard deviation in forecast $s_{y-\hat{y}}$ for each observation, and saves it to a new variable called

	varname4
--	-----------------

The names of the variables can be whatever you like, but the options (**xb**, **r**, **stdp**, and **stdf**) need to be stated just like this. In this case, we would type:

```
predict wagehat2, xb
```

Then Stata will create a new variable called **wagehat2** with the predicted value of each wife's age. By typing **corr wagehat wagehat2**, you can check that these values are the same as the ones obtained by using the formula $wagehat_i = 2.529682 + 0.8998053 \cdot hage_i$:

	wagehat	wagehat2
wagehat	1.0000	
wagehat2	1.0000	1.0000

A correlation of 1.0000 means that the variables are exactly the same.

With this variable, we will graph the predicted age of wives. Let's type:

```
graph twoway (scatter wage hage) (line wagehat2 hage)
```

If you want to know the predicted value for a man of a specific age (like my neighbor's nephew), you can type:

```
summ wagehat2 if hage==25
```

This gives us the following information:

Variable	Obs	Mean	Std. Dev.	Min	Max
wagehat2	306	25.02481	0	25.02481	25.02481

This tells us that, on average, a 25-year-old man is married to a 25.02-year-old woman. Even though men are *typically* married to women who are 2.42 years younger, 25-year-old men are not—they are usually married to women of their same age (more or less). We also see that the standard deviation in the variable is zero. We can infer from this that the variable's value is the same for all people in this sample (that is, for all men aged twenty-five).

However, we might worry that our prediction is inaccurate, so we might want to form a confidence interval for this estimated average age. We know that the formula for this confidence interval is:

$$\hat{Y}_i \pm z_{\alpha/2}(s_{\hat{y}})$$

where $s_{\hat{y}}$ is the “standard deviation of the prediction” (given by some formula in the book). We can ask Stata to calculate the standard error of the prediction for us:

```
predict syhat, stdp
```

The option “**, stdp**” tells Stata that we want the standard deviation of the prediction; it creates a variable named **syhat** with this value (again, this name can be whatever you choose). A 95% confidence interval would be:

```
gen lowerpred = wagehat2 - 1.96 * syhat  
gen upperpred = wagehat2 + 1.96 * syhat
```

We can see these ranges if we create the graph:

```
graph twoway (scatter wage hage) (line wagehat2 hage) (line  
lowerpred hage) (line upperpred hage)
```

You might notice that the lower-bound and upper-bound lines almost coincide exactly with the prediction line. This is because we have estimated the model very precisely: we have a lot of data, we have a lot of variation in **hage**, and the **hage** is a good predictor of **wage**. As a consequence, we are fairly confident that we have estimated the prediction line accurately. (In practice, you will almost never find such close bounds for your predictions—but spouse’s age is one of the most predictable variables there is.)

We can also obtain the upper and lower bounds of the 95% confidence interval for *expected age of wife* at any specific value of **hage**. To do this, we would type:

```
summ lowerpred upperpred if hage==25
```

The output looks like this:

Variable	Obs	Mean	Std. Dev.	Min	Max
lowerpred	306	24.91998	0	24.91998	24.91998
upperpred	306	25.12965	0	25.12965	25.12965

Because the standard deviation is zero, this is simply a constant value for everyone in this group. For all men aged 25, the 95% confidence interval for average wife's age is from 24.92 to 25.13. (As a note, we would not be able to reject the hypothesis that a 25-year-old man married a woman of the same age, 25, on average.)

Finally, we want to come up with a range of “likely outcomes” for the actual value—our forecast interval. We know the formula for this:

$$\hat{Y}_i \pm z_{\alpha/2}(s_{y-\hat{y}})$$

where $s_{y-\hat{y}}$ is the “standard deviation of forecast”. We have a formula for this, but Stata can also generate it for us:

```
predict syminusyhat, stdf
```

Now it is simple to construct a 95% confidence interval for the actual age of the spouse:

```
gen lowerfore= wagehat2 - 1.96 * syminusyhat  
gen upperfore = wagehat2 + 1.96 * syminusyhat
```

Let's generate a graph that shows these bounds for all ages:

```
graph twoway (scatter wage hage) (line wagehat2 hage) (line  
lowerfore hage) (line upperfore hage)
```

There are a few things that bother me about this graph. Given how many observations we have, the dots are so big that they become a meaningless jumble. We can specify the option `msize(.2)` to shrink the dots to 20% of their default size. Also, it annoys me a big that the forecast lines extend into implausible ranges (it suggests that a 16-year-old guy has a 7-year-old wife!). We don't observe this in the data, so I'm going to cut off that portion of the line—and similarly, where it goes over the values we observe in the data, at the top right. Finally, we should place a title on the vertical axis.


```
graph twoway (scatter wage hage, msize(.2) ytitle("Wife's Age")) (line wagehat2 hage) (line lowerfore hage if lowerfore>15) (line upperfore hage if upperfore<80)
```

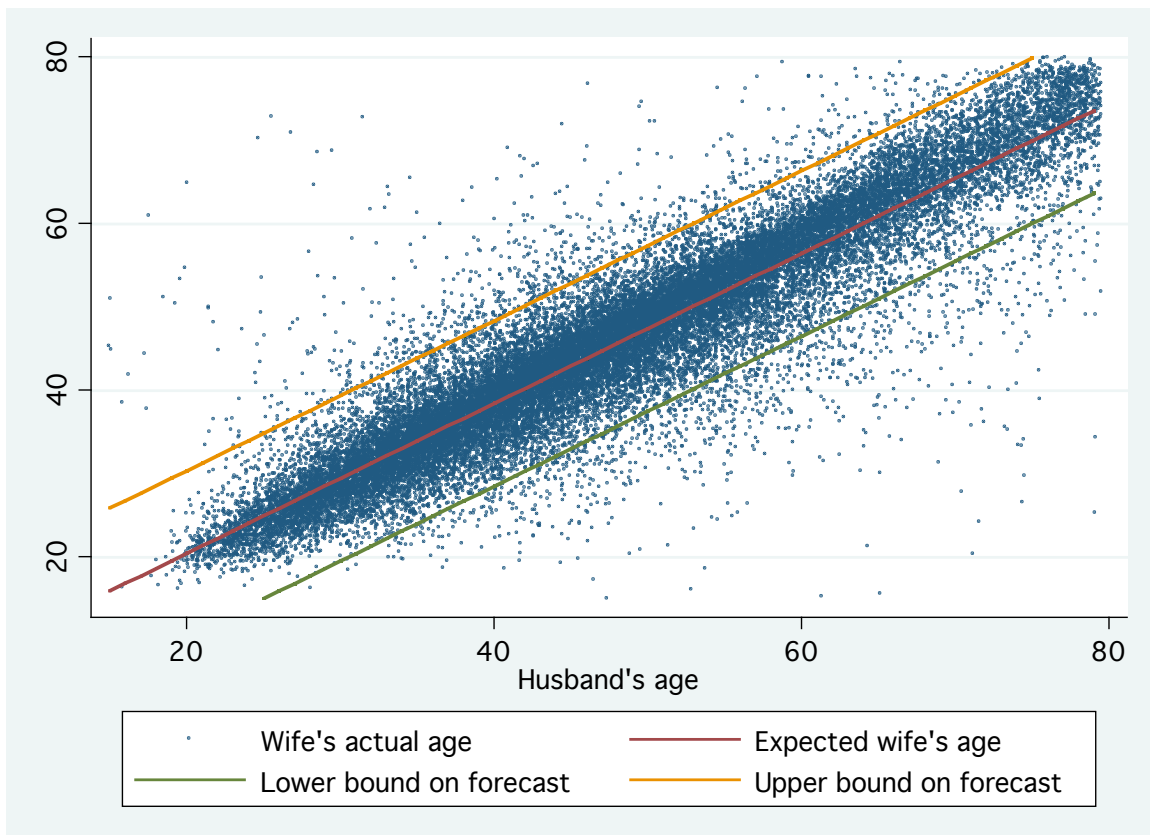
This looks a bit better, but the data is still too dense to give us a clear idea of how many people are in the center of the distribution. I'm going to add one more option: `jitter(2)`, which will place each plot at a random location *near* its actual value.

```
graph twoway (scatter wage hage, msize(.2) jitter(2)) (line wagehat2 hage) (line lowerfore hage if lowerfore>15) (line upperfore hage if upperfore<80)
```

If you've done this right, you should have an image that looks like a photograph of a galaxy, with a very high concentration of dots in a narrow band. Finally, let's fix the labels on the variables:

```
label variable wagehat2 "Expected wife's age"
label variable upperfore "Upper bound on forecast"
label variable lowerfore "Lower bound on forecast"
label variable hage "Husband's age"
```

Now generate the graph again. We should get:



The final question we have is “what is a likely range of age for the wife of a 25-year-old (and does it include 35)?” We would type:

```
summ upperfore lowerfore if hage==25
```

This gives us the following values:

Variable	Obs	Mean	Std. Dev.	Min	Max
upper	306	34.95288	0	34.95288	34.95288
lower	306	15.09675	0	15.09675	15.09675

In other words, from 15.10 years to 34.95. The value of 35 is just outside this 95% confidence interval—it could happen, but only about 5% of men would be in such a relationship. My initial intuition was correct: this would be an unlikely match.

That wraps up this little exercise. I’d like to add a few final notes about whether this analysis was entirely valid.

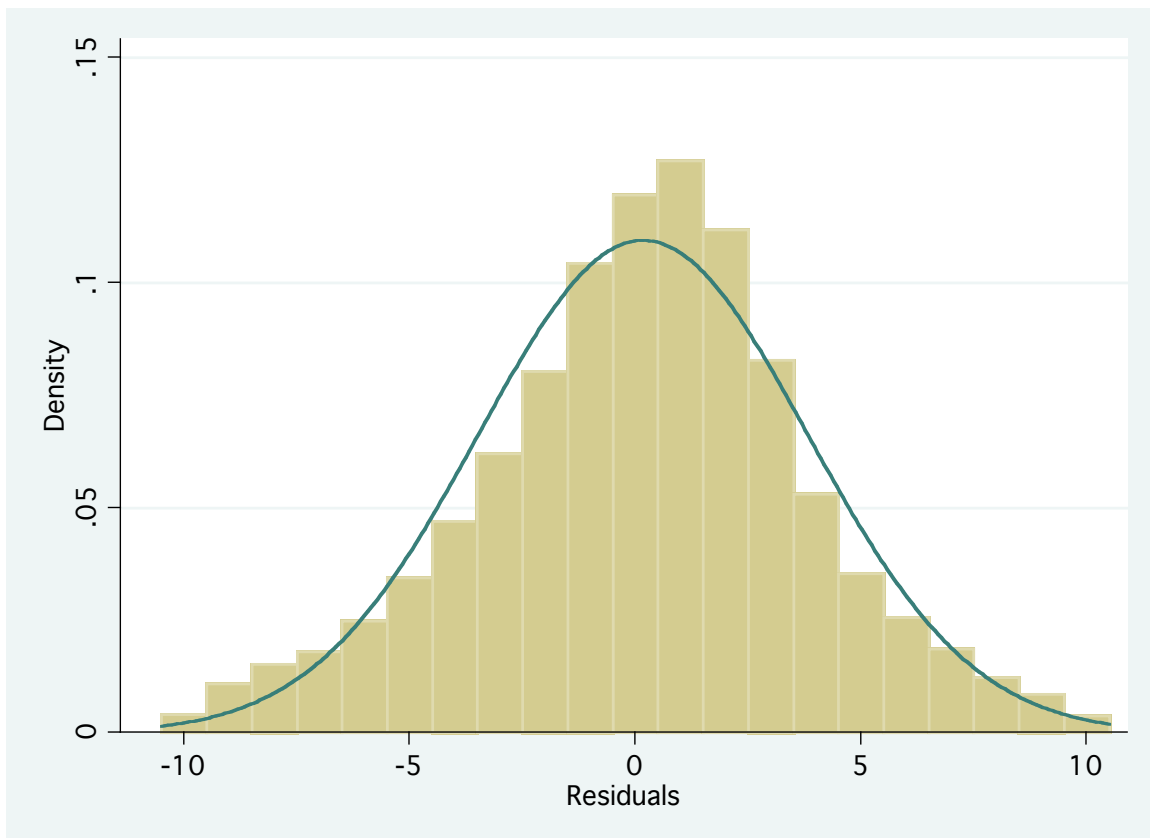
The “ordinary least squares regression” model makes several assumptions. The first is that we have some variation in the explanatory variable: the husband’s age, in this case. Clearly, that is true; we would not be able to compute $\hat{\beta}_1$ otherwise.

The second assumption is that “unobservables” e_i are uncorrelated with the explanatory variable x_i . In this case, you can think of the “unobservable” as “marrying someone younger or older than we would predict.” We are assuming that this does not tend to vary with the husband’s age. That can’t really be true for younger men: if someone is seventeen years old, then we predict that he should marry a woman who is 17.82 years old. He could marry someone older (even substantially older) than this with no problem, but he cannot easily marry someone younger (especially substantially younger) because of legal restrictions. This means that young men will tend to marry only women who are older than we predict. I might argue that the same is true at the opposite end of the spectrum: we could imagine a 60-year-old man married to someone thirty years younger than we would expect, but it would be hard to imagine that he would be equally likely to be married to someone thirty years *younger* than we would predict. Because of these tendencies, one of our OLS assumptions is violated. Unfortunately, this will happen often when we’re doing research. In these cases, we have to ask ourselves: *how bad is the violation?*

I'd argue that it's probably slight, in this case. These violations apply to the extremes of the distribution; they don't apply to the majority of our observations. I wouldn't worry about it. (I'll also add: when I used more sophisticated econometric tools to correct for this problem, I found virtually the same estimates of $\hat{\beta}_0$ and $\hat{\beta}_1$.)

Finally, there is the assumption that the variance in unobservables is the same for all observations (and that these unobservables are distributed normally). In other words, that there is the same variation in the ages of young men's wives as there is in middle-aged and older men's wives. I know that this isn't true: there's much more variation for middle-aged men than there is for anyone else. Also, what about the assumption that the variance is distributed normally? The histogram of **dage** revealed an asymmetry in the distribution. Let's look at the residuals that we obtained in this case:

```
predict resid, r
histogram r if r>-10 & r<10, width(1) normal start(-10.5)
```



The blue curve represents a normal distribution with the same mean and variance as **resid**. We see that this doesn't quite match up: if a wife is older than we predict, she is usually older by only one or two years. These categories are overrepresented in our histogram (relative to the normal distribution). In such a large sample, there's almost no room for leeway: if the histogram doesn't match up perfect with a normal curve, the variable is almost certainly *not* distributed normally.

What effect does this violation have? It means that our calculations of the confidence intervals are wrong. Remember that 95% of observations should be within our confidence intervals; 2.5% should be above, and 2.5% should be below. Let's see how many observations are actually above or below our forecasts. First we'll create dummy variables for "wife's actual age exceeds the upper bound" and "wife's actual age is below the lower bound":

```
gen above = (wage > upperfore)
gen below = (wage < lowerfore)
summ above below
```

This should give us the following results:

Variable	Obs	Mean	Std. Dev.	Min	Max
above	34674	.0226971	.1489382	0	1
below	34674	.0325316	.1774096	0	1

You can interpret the means of the dummy variables as the fraction of the sample that falls into that category (this is always the interpretation of the mean). We have 2.26% of observations above the upper bound, and 3.25% below the lower bound. Altogether, our "95% confidence interval" contains 94.48% of observations. This is a bit off (again, in a sample this size, we should get almost exactly 95% if the normality assumption is correct)—but it's not bad. The more troubling feature is the asymmetry: we should have 2.5% above, but we have slightly less; and we should have 2.5% below, but we have substantially more. This happens because the distribution (of unobservables) is not symmetric.

If we wanted to do this perfectly, we should model the unobservables with some other probability distribution—one that is not symmetric. However, that's best left for a later day.