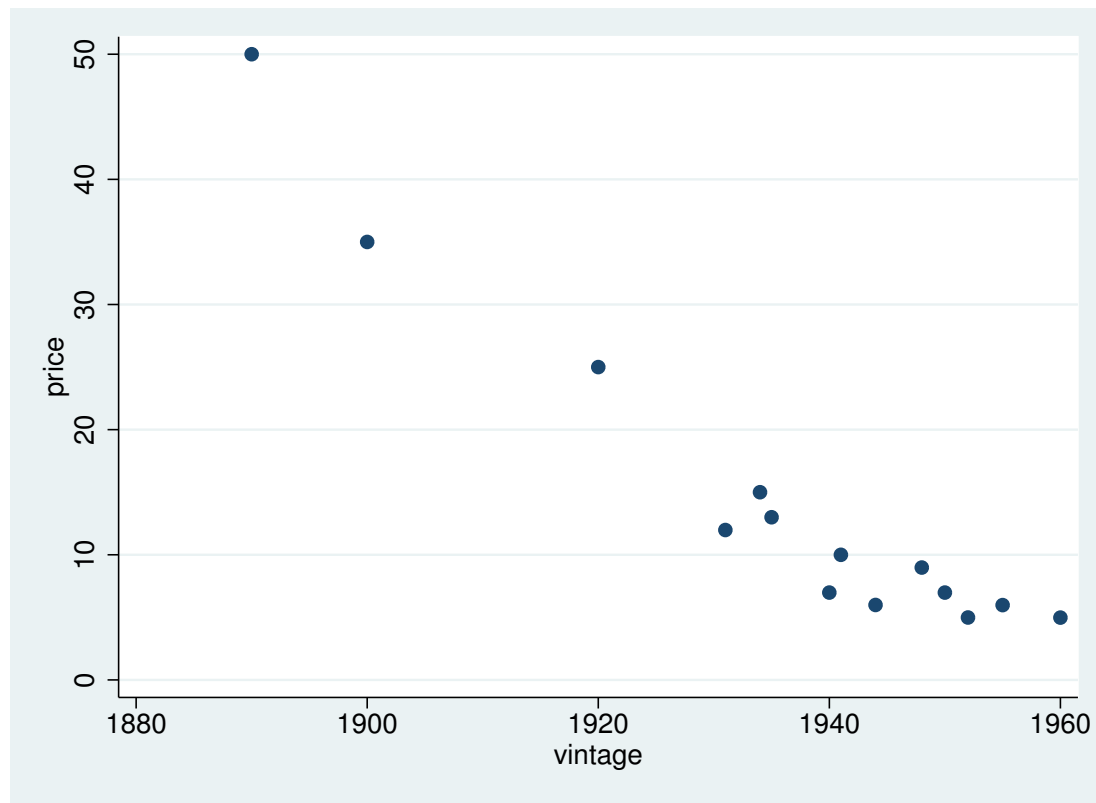


Common Transformations - an example

Example: Wine price and age

A listing of port wine vintage year and price (source: Chicago Maroon, ad for the Party Mart)

	vintage yr	price (\$)
1.	1890	50
2.	1900	35
3.	1920	25
4.	1931	11.98
5.	1934	15
6.	1935	13
7.	1940	6.98
8.	1941	10
9.	1944	5.99
10.	1948	8.98
11.	1950	6.98
12.	1952	4.99
13.	1955	5.98
14.	1960	4.98



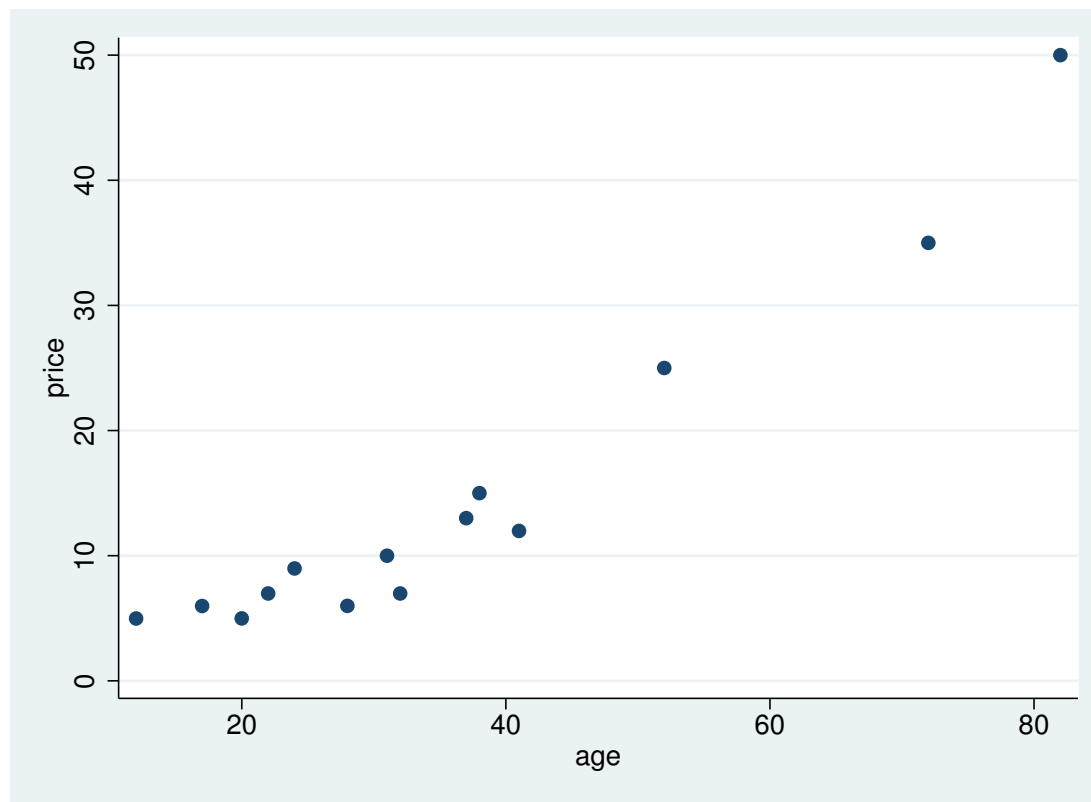
We want to predict price with vintage year or some function of it.

Common Transformations - an example

First, convert vintage year to a more useful quantity. Instead of vintage year, let's define $\text{age} = (1972 - \text{vintage})$.

```
. gen age = 1972-vintage  
. scatter price age
```

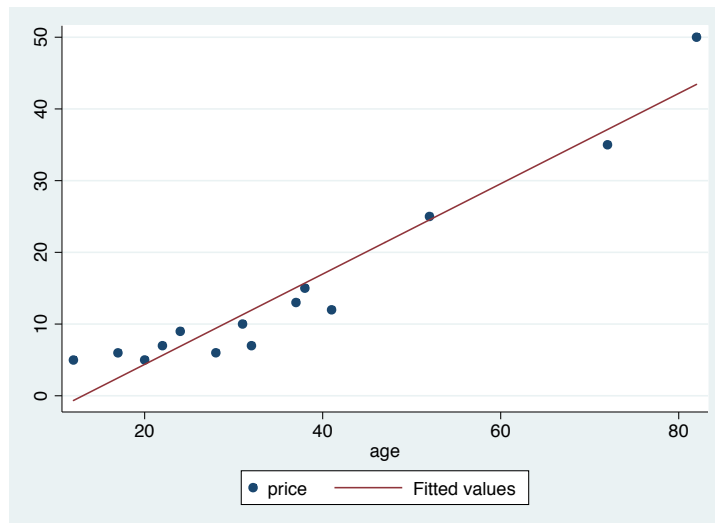
Why 1972?



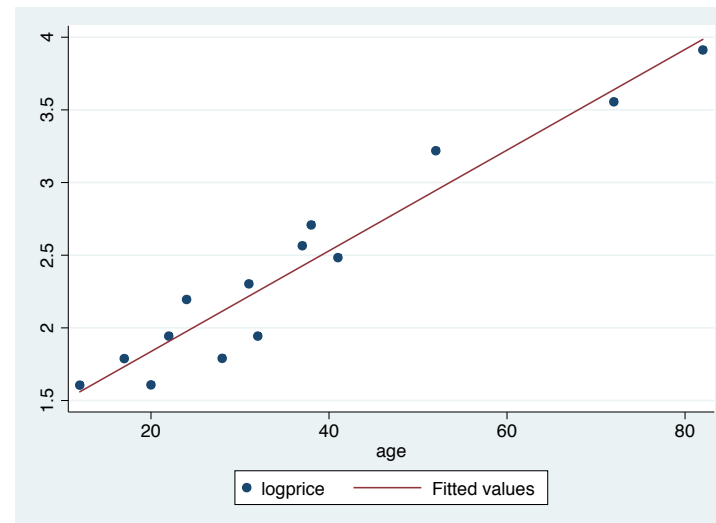
Do we have the right scale for Y (price)? There is some curvature upwards, and we may try the log transform on Y . Note: default base in Stata is e , or natural log

```
. gen logprice=log(price)
. twoway (scatter price age) (lfit price age)
. twoway (scatter logprice age) (lfit logprice age)
```

This looks improved with respect to linearity



(a) Original scale for price



(b) Log scale scale for price

Common Transformations - an example

- But how should we interpret this model? And what is the model now anyway? It is specified as:

$$E(\log \text{Price}) = \beta_0 + \beta_1 \text{Age}$$

This means that the mean of $\log(\text{Price})$ increases by a fixed amount for each increment in Age. What does this mean on the scale of Price (i.e. what does it mean in dollars)?

$$\exp(E(\log \text{Price})) = \exp(\beta_0) \exp(\beta_1 \text{Age}).$$

More simply written:

$$E(\text{Price}) \stackrel{>}{=} k e^{(\beta_1 \text{Age})} \quad \text{Not true for the mean, but true for median.}$$

where $k = \exp(\beta_0)$. So, there is a a nonlinear effect of age on price

Common Transformations - an example

If the age of Port increases by one year, from T to $(T + 1)$,

$$\exp(\log \widehat{\text{Price}}_{old}) = \exp(\hat{\beta}_0) \exp(\hat{\beta}_1 T).$$

$$\exp(\log \widehat{\text{Price}}_{new}) = \exp(\hat{\beta}_0) \exp(\hat{\beta}_1 (T + 1)).$$

Then

$$\widehat{\text{Price}}_{new} = \widehat{\text{Price}}_{old} \times \exp(\hat{\beta}_1).$$

- The prices (in \$) would be expected to (approximately) increase by a fixed multiple $\exp(\beta_1)$ per year. One could convert to % increase by $(\exp(\beta_1) - 1) \times 100\%$.
- Note that if one works on the $\log(\text{price})$ scale, $(\hat{\beta}_0 + \hat{\beta}_1(T + 1)) - (\hat{\beta}_0 + \hat{\beta}_1(T)) = \beta_1$ or the increment in ~~price~~ **average log(price)** ~~on the log scale~~

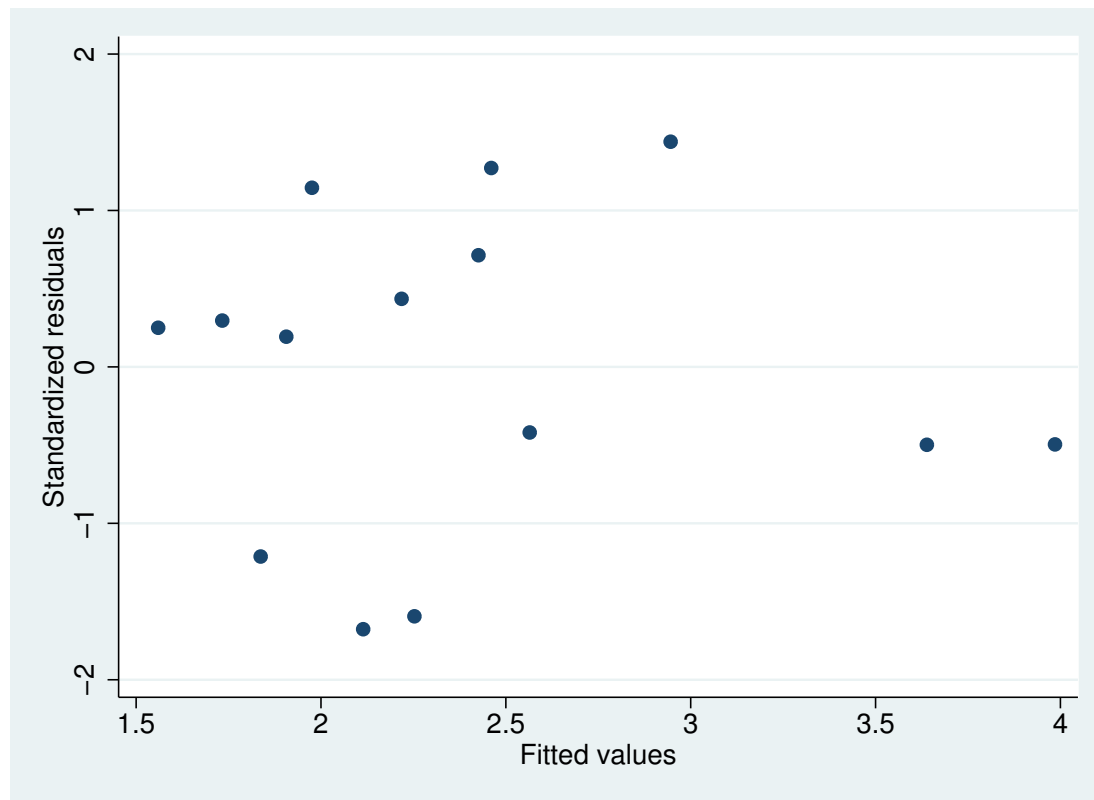
Common Transformations - an example

The results for this regression:

```
. reg logprice age
```

Source	SS	df	MS			
-----+-----				Number of obs =	14	
Model	6.40578846	1	6.40578846	F(1, 12) =	157.05	
Residual	.489463007	12	.040788584	Prob > F =	0.0000	
-----+-----				R-squared =	0.9290	
Total	6.89525147	13	.530403959	Adj R-squared =	0.9231	
-----				Root MSE =	.20196	
logprice	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
-----+-----						
age	.0346517	.0027651	12.53	0.000	.0286271	.0406763
_cons	1.143891	.1139306	10.04	0.000	.8956572	1.392124

The R^2 is high. The standardized residuals seem not too bad:



The model also suggests that ^{median} port prices increase by about 3.5% per year:

$$\exp(.0346517) = 1.0353$$

Note that for small x , $\exp(x)$ is approximately $1 + x$.

How do we do predictions with this model?

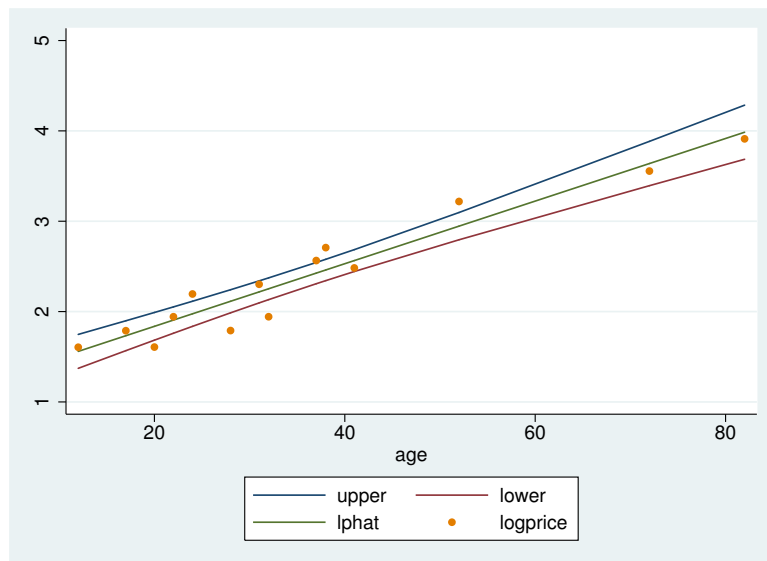
```
. predict sepred, stdp
. predict lphat
.
. sort age
. gen upper=lphat + invttail(12, 0.025) * sepred
. gen lower=lphat - invttail(12, 0.025) * sepred
. scatter upper lower lphat logprice age, c(1 1 1 .) s(i i i o)
```

Converting back to the original model:

```
. gen dollarp=exp(lphat)
. gen dollaru=exp(upper)
. gen dollarl=exp(lower)
. scatter price dollarp dollaru dollarl age, s(o i i i) c(. 1 1 1)
  xlabel(0(10)90) ylabel(0(20)100)
```

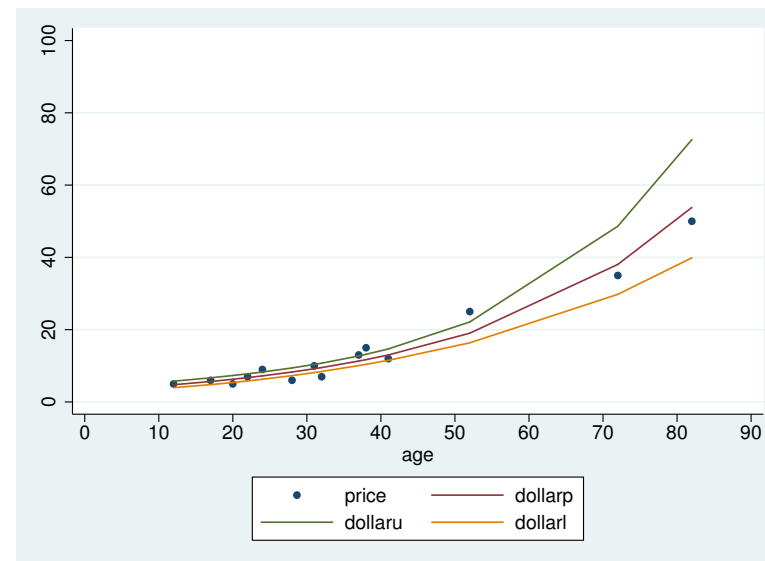
Note that once you convert back to the original scale, the confidence bands are not symmetric any longer. They are also increasing in width.

Confidence bands for average log(price)



(c) Log scale (in log-dollars)

Confidence bands for median price



(d) Original scale (in dollars)