



Linear Regression

Paul D. Allison, Ph.D.

Linear Regression Analysis

Paul D. Allison, Ph.D.



1

What is Linear Regression?

A statistical technique for studying the relationship between a single dependent variable y and one or more independent variables (the x 's).

Goal: A linear equation of the form

$$\hat{y} = b_0 + b_1x_1 + \dots + b_kx_k$$

where \hat{y} is the predicted value of y and the b 's are the estimated "regression coefficients".

Two main uses:

1. Generate predictions about y based on knowledge of x 's.
2. Estimate and test hypotheses about the "causal" effect of each x on y , "controlling" for the other x 's.

2

Prediction

How does Zillow produce an estimate of a home's value?

- Collect data on sale price of millions of homes.
- Also collect data on “location, lot size, square footage, number of bedrooms and bathrooms, actual property taxes paid, exceptions to tax, actual sale prices over time of the home itself and comparable recent sales of nearby homes.”
- Formula is “proprietary” but I’m pretty sure it’s based on some form of linear regression.
- Potential complications: Different information is available for different homes, how to deal with spatial information, change over time, different real estate markets, etc.

3

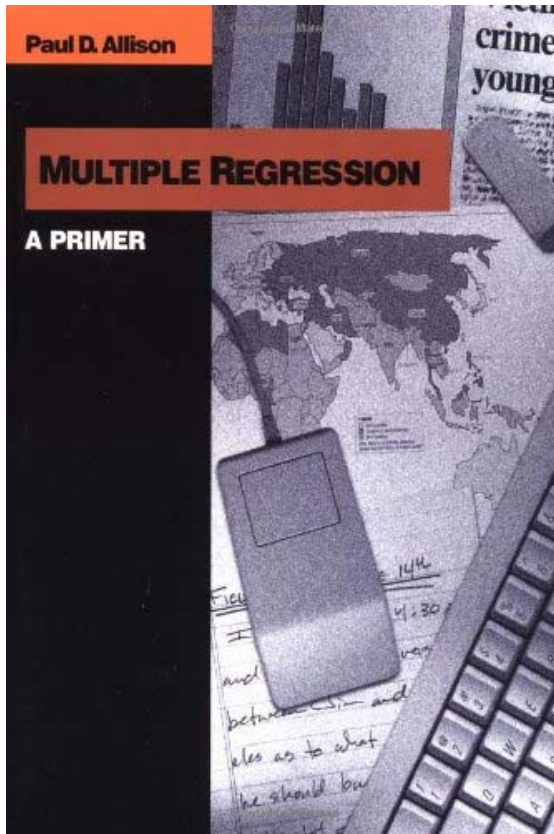
Causal Effects

Does divorce lead to bad outcomes for children?

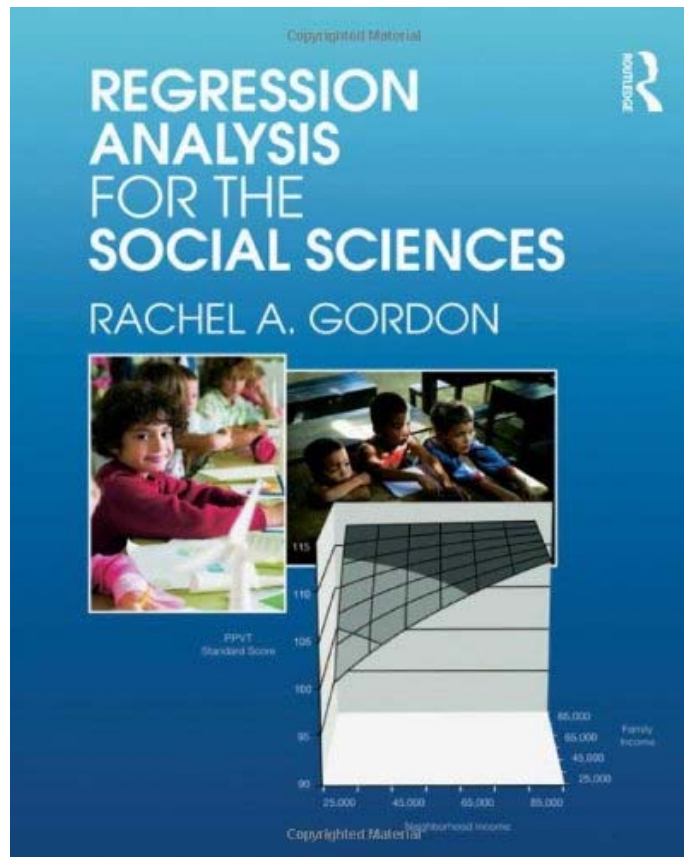
Allison, Paul D. and Frank F. Furstenberg, Jr. (1989) “How marital dissolution affects children: Variations by age and sex.” *Developmental Psychology* 25: 540-549.

- For 1197 children, we compared those from intact families with those whose parents divorced or separated (328).
- Children of divorce did worse on measures of delinquency, hyperactivity, academic difficulty, distress, etc.
- Problem: Couples who divorced differed in many other respects from those who did not.
- Solution: Estimate linear regressions that controlled for child's age, race, sex, birth order, region of residence, mother's education, religious preference, age at birth of the child, age at birth of first child, and foreign or U.S. birth.
- Result: Most “effects” of divorce remained strong after controls.

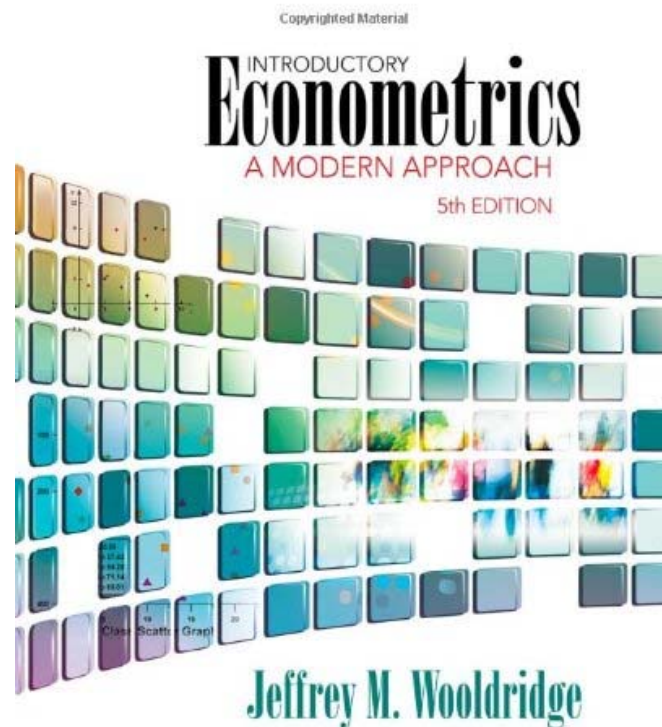
4



5



6



7

Beauty Data Set

Source: Hamermesh, D.S. and J.E. Biddle (1994) "Beauty and the Labor Market." *American Economic Review* 84: 1174-1194.

1260 employed adults in 1977 in the US

wage	hourly wage (THIS WILL BE OUR DEPENDENT VARIABLE)
exper	years of workforce experience
looks	from 1 to 5 (as rated by interviewers)
union	=1 if union member , else 0
goodhlth	=1 if good health, else 0
black	=1 if black, else 0
female	=1 if female , else 0
married	=1 if married, else 0
south	=1 if live in south, else 0
bigcity	=1 if live in big city, else 0
smllcity	=1 if live in small city, else 0
service	=1 if service industry, else 0
educ	years of schooling

8

Questions

Can we construct a “good” model to predict hourly wages?

Does more education lead to higher wages?

Is that true even if we control for other characteristics?

What’s the magnitude of the effect?

Do women make less than men?

Do blacks make less than non-blacks?

Do more attractive people get higher wages?

Do union workers make more than non-union workers?

9

Preliminary Steps

Get familiar with the data

- Check the distribution of each variable.
- Compute means, minima, maxima, frequency tables, histograms, etc.
- Look for outliers and coding errors.
- Check for missing data.

10

Using Stata

```
. use "C:\data\beauty.dta"
. summarize
```

Variable	Obs	Mean	Std. Dev.	Min	Max
wage	1260	6.30669	4.660639	1.02	77.72
lwage	1260	1.6588	.5945075	.0198026	4.353113
exper	1260	18.20635	11.96349	0	48
looks	1260	3.185714	.6848774	1	5
union	1260	.2722222	.4452804	0	1
goodhlth	1260	.9333333	.2495429	0	1
black	1260	.0738095	.2615645	0	1
female	1260	.3460317	.4758923	0	1
married	1260	.6912698	.462153	0	1
south	1260	.1746032	.3797781	0	1
bigcity	1260	.2190476	.4137652	0	1
smlcity	1260	.4666667	.4990857	0	1
service	1260	.2738095	.4460895	0	1
educ	1260	12.56349	2.624489	5	17

11

Using SAS

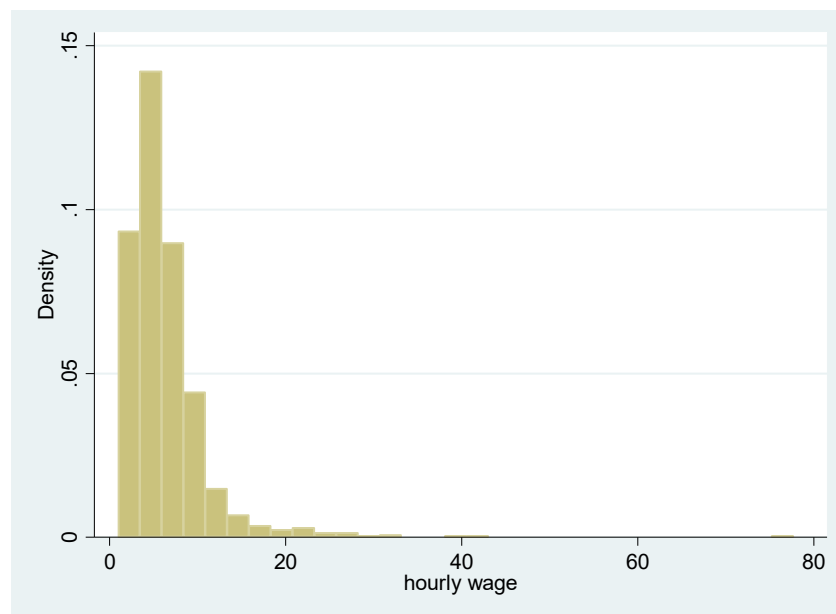
```
PROC MEANS DATA=my.beauty; RUN;
```

Variable	N	Mean	Std Dev	Minimum	Maximum
wage	1260	6.3066905	4.6606390	1.0200000	77.7200000
lwage	1260	1.6587996	0.5945075	0.0198026	4.3531130
exper	1260	18.2063492	11.9634853	0	48.0000000
looks	1260	3.1857143	0.6848774	1.0000000	5.0000000
union	1260	0.2722222	0.4452804	0	1.0000000
goodhlth	1260	0.9333333	0.2495429	0	1.0000000
black	1260	0.0738095	0.2615645	0	1.0000000
female	1260	0.3460317	0.4758923	0	1.0000000
married	1260	0.6912698	0.4621530	0	1.0000000
south	1260	0.1746032	0.3797781	0	1.0000000
bigcity	1260	0.2190476	0.4137652	0	1.0000000
smlcity	1260	0.4666667	0.4990857	0	1.0000000
service	1260	0.2738095	0.4460895	0	1.0000000
educ	1260	12.5634921	2.6244892	5.0000000	17.0000000

12

Histogram of wage

In Stata: **hist wage**



13

Distribution of looks

```
. tab looks
```

from 1 to 5	Freq.	Percent	Cum.
1	13	1.03	1.03
2	142	11.27	12.30
3	722	57.30	69.60
4	364	28.89	98.49
5	19	1.51	100.00
Total	1,260	100.00	

Note:

1=homely, 2=quite plain, 3=average, 4=good looking, 5=strikingly beautiful

14

Stata Regression Predicting Wages

```
regress wage exper looks union goodhlth black female married
south bigcity smllcity service educ
```

Source	SS	df	MS	Number of obs =	1260
Model	6011.43607	12	500.953006	F(12, 1247) =	29.28
Residual	21336.0031	1247	17.1098662	Prob > F =	0.0000
				R-squared =	0.2198
				Adj R-squared =	0.2123
Total	27347.4392	1259	21.7215561	Root MSE =	4.1364

wage	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
exper	.078945	.0106672	7.40	0.000	.0580174 .0998725
looks	.4138739	.1742518	2.38	0.018	.0720148 .755733
union	.6114653	.2670238	2.29	0.022	.0875999 1.135331
goodhlth	-.0524494	.4755118	-0.11	0.912	-.9853408 .8804419
black	-.1106673	.4612814	-0.24	0.810	-1.015641 .7943061
female	-2.127858	.2763716	-7.70	0.000	-2.670063 -1.585654
married	.8213589	.2744173	2.99	0.003	.2829883 1.359729
south	.3572703	.3116461	1.15	0.252	-.2541384 .9686789
bigcity	1.720111	.3363336	5.11	0.000	1.060269 2.379954
smllcity	.5875548	.2736183	2.15	0.032	.0507518 1.124358
service	-.478788	.2882098	-1.66	0.097	-1.044218 .0866417
educ	.4246026	.0500969	8.48	0.000	.3263191 .5228861
_cons	-2.306654	.9797571	-2.35	0.019	-4.228808 -.3844996

15

Stata Regression Predicting Wages

```
regress wage exper looks union goodhlth black female married
south bigcity smllcity service educ
```

Source	SS	df	MS	Number of obs =	1260
Model	6011.43607	12	500.953006	F(12, 1247) =	29.28
Residual	21336.0031	1247	17.1098662	Prob > F =	0.0000
				R-squared =	0.2198
				Adj R-squared =	0.2123
Total	27347.4392	1259	21.7215561	Root MSE =	4.1364

wage	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
exper	.078945	.0106672	7.40	0.000	.0580174 .0998725
looks	.4138739	.1742518	2.38	0.018	.0720148 .755733
union	.6114653	.2670238	2.29	0.022	.0875999 1.135331
goodhlth	-.0524494	.4755118	-0.11	0.912	-.9853408 .8804419
black	-.1106673	.4612814	-0.24	0.810	-1.015641 .7943061
female	-2.127858	.2763716	-7.70	0.000	-2.670063 -1.585654
married	.8213589	.2744173	2.99	0.003	.2829883 1.359729
south	.3572703	.3116461	1.15	0.252	-.2541384 .9686789
bigcity	1.720111	.3363336	5.11	0.000	1.060269 2.379954
smllcity	.5875548	.2736183	2.15	0.032	.0507518 1.124358
service	-.478788	.2882098	-1.66	0.097	-1.044218 .0866417
educ	.4246026	.0500969	8.48	0.000	.3263191 .5228861
_cons	-2.306654	.9797571	-2.35	0.019	-4.228808 -.3844996

This is an F test of the null hypothesis that all 12 coefficients are 0. The p-value is extremely low, so we can reject that hypothesis and conclude that at least one coefficient is not 0.

16