

This work is licensed under a [Creative Commons Attribution-NonCommercial-ShareAlike License](https://creativecommons.org/licenses/by-nc-sa/4.0/). Your use of this material constitutes acceptance of that license and the conditions of use of materials on this site.



Copyright 2006, The Johns Hopkins University and John McGready. All rights reserved. Use of these materials permitted only in accordance with license rights granted. Materials provided "AS IS"; no representations or warranties provided. User assumes all responsibility for use, and all liability related thereto, and must independently review all materials for accuracy and efficacy. May contain materials owned by others. User is responsible for obtaining permissions for use from third parties as needed.



JOHNS HOPKINS
BLOOMBERG
SCHOOL *of* PUBLIC HEALTH

*Relating a Continuous Outcome
to More than One Predictor:
Multiple Linear Regression*

John McGready
Johns Hopkins University

Why multiple linear regression?

Interpreting coefficients from multiple linear regression

Statistical inference on multiple regression coefficients

ANOVA as a linear regression model

Statistical interaction and linear regression



JOHNS HOPKINS
BLOOMBERG
SCHOOL *of* PUBLIC HEALTH

Section A

Why (Multiple) Linear Regression?

Recall hemoglobin (Hb)/PCV example from the last section (are you sick of it yet?)

At the end of the analysis, we found a positive significant association between Hb and PCV

- ★ *We estimated the magnitude of this association (slope) and put 95% confidence limits on it*

We also estimated the R^2 for regression of 51%, indicating that PCV explains about half of the original variation in Hb

- ★ *This means there is still 49% of the original variability left to explain!*

Multiple Linear Regression

In the data set on 21 individuals for whom we have the hemoglobin and PCV data, we also have information on each subject's age

Here is a partial listing of the full data set

	Hb	PCV	age
1.	12	35	20
2.	10.7	39	22
3.	12.4	47	26
4.	14.2	53	28
5.	13.1	30	28

Continued

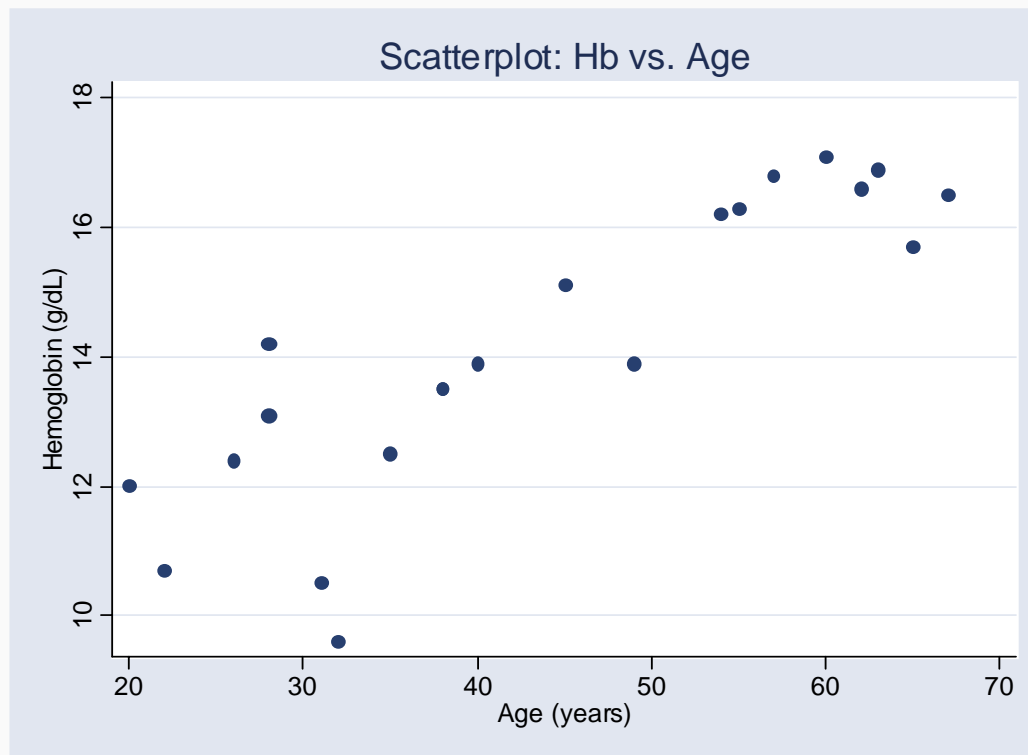
Question—can age tell us anything more about Hb above and beyond what PCV already tells us?

If age does add information about Hb, does it also change the estimated relationship between Hb and PCV?

★ *In other words, does age confound the Hb/PCV relationship?*

Multiple Linear Regression

Let's start investigating—Is there a relationship between Hb and age?



Continued

Multiple Linear Regression

Let's start investigating—Is there a relationship between Hb and age?

```
. regress Hb age
```

Source	SS	df	MS			
Model	78.9138751	1	78.9138751	Number of obs =	21	
Residual	26.4375502	19	1.39145001	F(1, 19) =	56.71	
				Prob > F =	0.0000	
				R-squared =	0.7491	
				Adj R-squared =	0.7358	
				Root MSE =	1.1796	
Total	105.351425	20	5.26757126			

Hb	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
age	.1295612	.0172041	7.53	0.000	.0935526	.1655698
_cons	8.497691	.7925723	10.72	0.000	6.838818	10.15656

Continued

Multiple Linear Regression

Let's start investigating—Is there a relationship between Hb and age?

```
. regress Hb age
```

Source	SS	df	MS				
Model	78.9138751	1	78.9138751	Number of obs =	21		
Residual	26.4375502	19	1.39145001	F(1, 19) =	56.71		
Total	105.351425	20	5.26757126	Prob > F =	0.0000		
				R-squared =	0.7491		
				Adj R-squared =	0.7358		
				Root MSE =	1.1796		

Hb	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
age	.1295612	.0172041	7.53	0.000	.0935526	.1655698
_cons	8.497691	.7925723	10.72	0.000	6.838818	10.15656

Continued

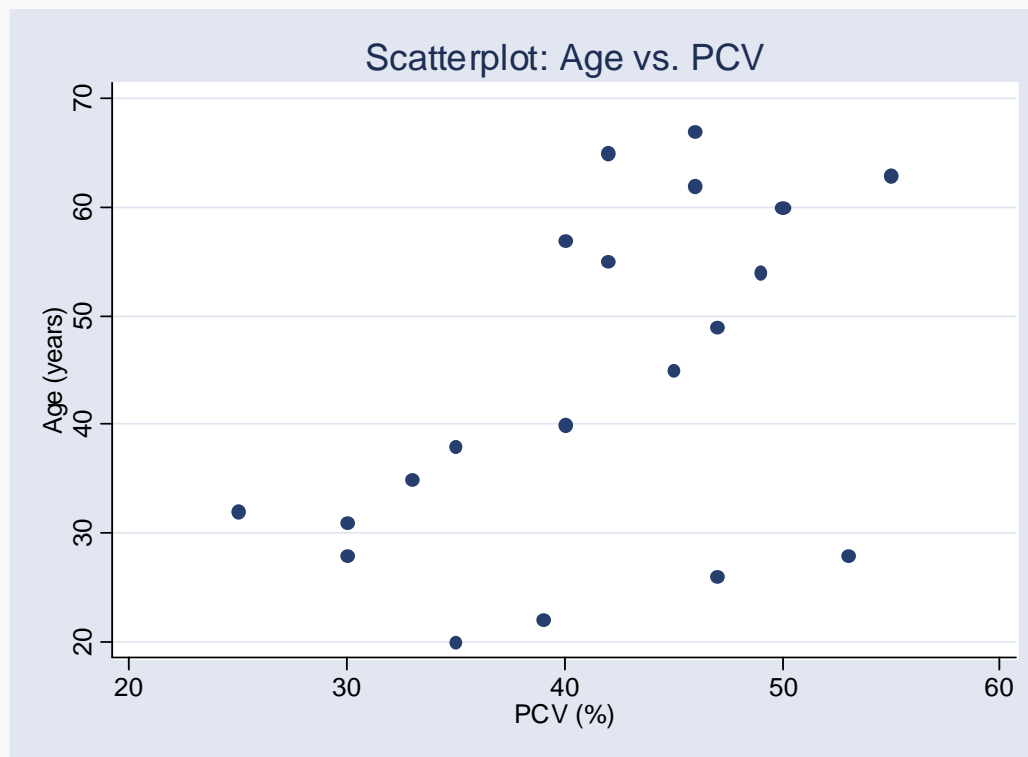
Let's start investigating—Is there a relationship between Hb and age?

- ★ *Evidence of linear relationship in scatterplot*
- ★ *Positive statistically significant slope of 0.12, 95% CI (0.09, 0.17)*
- ★ *R² of 75%—age alone explains more variability in Hb than PCV alone*
- ★ *Maybe the question we should ask is “can PCV tell us anything about Hb above and beyond age?”*

2nd part of investigation—If age does add information about Hb, does it also change the estimated relationship between Hb and PCV?

If so, age would also have to be related to PCV

Is age related to PCV?



So, lets recap:

- ★ *Hb is related to age*
- ★ *Age explains more variability in Hb than PCV*
- ★ *Visual evidence of age/PCV relationship*

Given the facts above, it's pretty clear that some of the original relationship we estimated between Hb and PCV was because of Hb/PCV/age relationship

How could we estimate the age-adjusted relationship between Hb and PCV?

- ★ *We want to see how different it is from unadjusted estimated of the slope, the 0.20 we saw in the regression of Hb on PCV*

Well, we could stratify the sample by age

Multiple Linear Regression

Sort individuals into age groups (21–25, 26–30, etc . . .)

For each age group, perform a simple linear regression of Hb (y) on PCV (x)

Get the slope estimate, \hat{b}_1 , for each age group

Calculate a weighted average of slopes based on number of observations in each age group

The resulting weighted average would be our age-adjusted slope estimate for the Hb/PCV relationship

- ★ *We would also need to figure out how to get a confidence interval for this*
- ★ **WHAT A PAIN!**

Further, we would have to take the same approach to adjust Hb/age relationship for PCV

Then, suppose both adjusted estimates were “significant”

- ★ *How could we figure out how variability in Hb was explained by BOTH age and PCV?*
- ★ *How could we predict Hb from subjects' age and PCV?*

There's got to be a better, easier way to do this!

And there is—multiple linear regression

Multiple Linear Regression

Multiple linear regression (MLR) is a method to estimate an equation to relate the mean of a continuous measure **y** to multiple independent variables (the **x**'s) in one equation!

MLR allows the estimation of the adjusted relationships and evaluates the overall explanatory power of multiple predictors of an outcome

Multiple Regression Equation

A multiple linear regression equation is just an extension of a simple linear regression equation—just add more **x**'s!

$$E[y] = \hat{b}_0 + \hat{b}_1x_1 + \hat{b}_2x_2 + \hat{b}_3x_3\dots\dots$$

Could also be expressed as:

$$\hat{y} = \hat{b}_0 + \hat{b}_1x_1 + \hat{b}_2x_2 + \hat{b}_3x_3\dots\dots$$

Multiple Regression Equation

So in the hemoglobin example, we can fit an MLR of hemoglobin on PCV and age in one equation

$$\hat{y} = \hat{b}_0 + \hat{b}_1 x_1 + \hat{b}_2 x_2$$

- ★ Here :
- ★ \hat{y} = *estimated mean hemoglobin*
- ★ x_1 = *PCV*
- ★ x_2 = *age*

Multiple Regression with Stata

How can we estimate these regression coefficients for this equation from our data?

★ *Stata will do it*

Recall data set up in Stata (here are first five observations)

	Hb	PCV	age
1.	12	35	20
2.	10.7	39	22
3.	12.4	47	26
4.	14.2	53	28
5.	13.1	30	28

Just like before, we use the regress command

- ★ *General syntax*
- ★ *regress y X₁ X₂ X₃ ...*

So for the hemoglobin/PCV example:

- ★ *regress Hb PCV age*
(y)(X₁) (X₂)

Hemoglobin and PCV

```
. regress Hb PCV age
```

Source	SS	df	MS			
Model	88.7923124	2	44.3961562	Number of obs = 21		
Residual	16.5591129	18	.919950714	F(2, 18) = 48.26		
Total	105.351425	20	5.26757126	Prob > F = 0.0000		
				R-squared = 0.8428		
				Adj R-squared = 0.8254		
				Root MSE = .95914		

Hb	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
PCV	.1023427	.0312317	3.277	0.004	.0367274	.167958
age	.1013414	.0164271	6.169	0.000	.0668293	.1358534
_cons	5.516596	1.114866	4.948	0.000	3.174349	7.858842

Continued

Hemoglobin and PCV

```
. regress Hb PCV age
```

Source	SS	df	MS			
Model	88.7923124	2	44.3961562	Number of obs = 21		
Residual	16.5591129	18	.919950714	F(2, 18) = 48.26		
Total	105.351425	20	5.26757126	Prob > F = 0.0000		
				R-squared = 0.8428		
				Adj R-squared = 0.8254		
				Root MSE = .95914		

Hb	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
PCV	.1023427	.0312317	3.277	0.004	.0367274	.167958
age	.1013414	.0164271	6.169	0.000	.0668293	.1358534
_cons	5.516596	1.114866	4.948	0.000	3.174349	7.858842

Continued

Hemoglobin and PCV

```
. regress Hb PCV age
```

Source	SS	df	MS			
Model	88.7923124	2	44.3961562	Number of obs = 21		
Residual	16.5591129	18	.919950714	F(2, 18) = 48.26		
Total	105.351425	20	5.26757126	Prob > F = 0.0000		
				R-squared = 0.8428		
				Adj R-squared = 0.8254		
				Root MSE = .95914		

Hb	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
PCV	.1023427	.0312317	3.277	0.004	.0367274	.167958
age	.1013414	.0164271	6.169	0.000	.0668293	.1358534
_cons	5.516596	1.114866	4.948	0.000	3.174349	7.858842

Continued

Hemoglobin and PCV

```
. regress Hb PCV age
```

Source	SS	df	MS	Number of obs = 21	
Model	88.7923124	2	44.3961562	F(2, 18) =	48.26
Residual	16.5591129	18	.919950714	Prob > F =	0.0000
Total	105.351425	20	5.26757126	R-squared =	0.8428
				Adj R-squared =	0.8254
				Root MSE =	.95914

Hb	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
PCV	.1023427	.0312317	3.277	0.004	.0367274	.167958
age	.1013414	.0164271	6.169	0.000	.0668293	.1358534
_cons	5.516596	1.114866	4.948	0.000	3.174349	7.858842

Continued

The resulting equation:

- ★ $\hat{y} = 5.5 + .102*PCV + .101*Age$
- ★ *(It is a coincidence that coefficients for age and PCV are nearly equal)*
- ★ *What is the interpretation of the intercept?*

Interpretation

- ★ *For a given age, hemoglobin increases by an estimated .10 gm/dl for every one percent increase in PCV*
- ★ *We estimate that two groups of subjects of the same age who differ by one percentage in PCV levels will have hemoglobin levels that differ on average by .10 gm/dL*

Interpretation

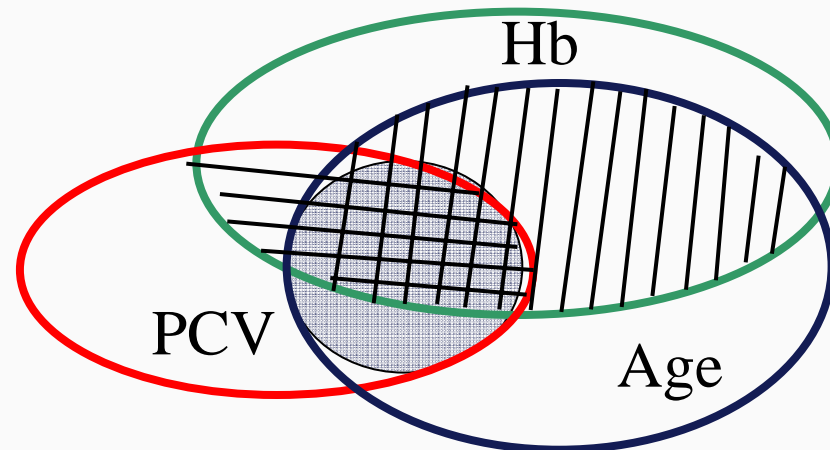
- ★ *For a given PCV level, hemoglobin increases an estimated .10 gm/dl for every one year increase in age*
- ★ *We estimate that two groups of subjects of the same PCV level who differ by one year of age will have hemoglobin levels that differ on average by .10 gm/dL*

Comparison of SLR and MLR results

Model x's	PCV Coef	Age Coef	R^2
PCV	.20	–	51%
age	–	.13	75%
PCV + age	.10	.10	84%

Comparison of SLR and MLR results

- ★ *Coefficients of both PCV and age changed value when both were in model*
- ★ *Together, age and PCV explain more variability in Hb than either alone*



General interpretation of regression coefficients from equation

$$\star \hat{y} = \hat{b}_0 + \hat{b}_1 x_1 + \hat{b}_2 x_2$$

Consider two observations with the same value of x_2 , but observation one with x_1 one unit higher than observation two

- ★ *Obs 1:* $x_1 = k + 1, x_2 = c$
- ★ *Obs 2:* $x_1 = k, x_2 = c$

Let's write out the predicted y for each observation

★ *Obs 1:*

$$\hat{y} = \hat{b}_0 + \hat{b}_1(k + 1) + \hat{b}_2c$$

★ *Obs 2:*

$$\hat{y} = \hat{b}_0 + \hat{b}_1(k) + \hat{b}_2c$$

Let's write out the predicted y for each observation

★ *Obs 1:*

$$\hat{y} = \hat{b}_0 + \hat{b}_1(k) + \hat{b}_1 + \hat{b}_2c$$

★ *Obs 2:*

$$\hat{y} = \hat{b}_0 + \hat{b}_1(k) + \hat{b}_2c$$

Subtracting yields

★ *Obs 1:*

$$\hat{y} = \hat{b}_0 + \hat{b}_1(k) + \hat{b}_1 + \hat{b}_2 c$$

★ *Obs 2:*

$$\hat{y} = \hat{b}_0 + \hat{b}_1(k) + \hat{b}_2 c$$

Subtracting yields

★ *Obs 1:*

$$\hat{y} = \hat{b}_0 + \hat{b}_1(k) + \hat{b}_1 + \hat{b}_2 c$$

★ *Obs 2:*

$$\hat{y} = \hat{b}_0 + \hat{b}_1(k) + \hat{b}_2 c$$

★ *All that remains is \hat{b}_1*

\hat{b}_1 is the expected (mean) change in y per unit change in x_1 if x_2 is held constant

\hat{b}_1 estimates relationship between y and x_1 adjusted (controlling) for x_2

\hat{b}_2 is the expected (mean) change in y per unit change in x_2 if x_1 is held constant

\hat{b}_2 estimates relationship between y and x_2 adjusted (controlling) for x_1



JOHNS HOPKINS
BLOOMBERG
SCHOOL *of* PUBLIC HEALTH

Section A

Practice Problems

1. Consider the multiple linear regression equation described by the following equation:

$$\hat{y} = \hat{b}_0 + \hat{b}_1x_1 + \hat{b}_2x_2 + \hat{b}_3x_3$$

1. What is the most general interpretation of the slope estimate \hat{b}_1 in this equation?

2. Consider the following results from a MLR regressing weight (lbs) on height (in) and age (years) for a sample of 12 nutritionally deficient children. The results, on the next slide, are missing some key information.

Stata results

```
. regress wt ht age
```

Source	SS	df	MS	
Model	692.822607	2	346.411303	Number of obs = 10
Residual	195.427393	9	21.7141548	F(2, 9) = 15.95
Total	888.25	11	80.75	Prob > F = 0.001
				R-squared = 0.780
				Adj R-squared = 0.7311
				Root MSE = 4.659

	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
wt					
ht	.722038	.2608051			
age	2.050126	.9372256			
_cons	6.553048	10.94483			

2. Write out the regression equation described by these results.

What is the interpretation of the coefficient for height?



JOHNS HOPKINS
BLOOMBERG
SCHOOL *of* PUBLIC HEALTH

Section A

Practice Problem Solutions

1. Consider the multiple linear regression equation described by the following equation:

$$\hat{y} = \hat{b}_0 + \hat{b}_1x_1 + \hat{b}_2x_2 + \hat{b}_3x_3$$

1. What is the most general interpretation of the slope estimate \hat{b}_1 in this equation?

- ★ *The expected change in y for a one-unit increase in x_1 , after adjustment for x_2 and x_3 (holding x_2 and x_3 constant). In other words, if we were to compare the mean of y for two groups who differ by one unit in x_1 and have the same values of x_2 and x_3 , the estimated mean difference would be \hat{b}_1 .*

2. Consider the following results from a MLR regressing weight (lbs) on height (in) and age (years) for a sample of 12 nutritionally deficient children. The results, on the next slide, are missing some key information.

Stata results

```
. regress wt ht age
```

Source	SS	df	MS	
Model	692.822607	2	346.411303	Number of obs = 1
Residual	195.427393	9	21.7141548	F(2, 9) = 15.95
Total	888.25	11	80.75	Prob > F = 0.001
				R-squared = 0.780
				Adj R-squared = 0.7311
				Root MSE = 4.659

	Coef.	Std. Err.	t	P> t	[95% Conf. Interval
ht	.722038	.2608051			
age	2.050126	.9372256			
_cons	6.553048	10.94483			

2a. Write out the regression equation described by these results.

$$\hat{y} = 6.55 + 2.1x_1 + 0.72x_2$$

Where \hat{y} = mean weight

x_1 = age

x_2 = height

2b. What is the interpretation of the coefficient for height?

- ★ *After adjusting for age, it is expected that children will differ on average by 0.72 lbs for every one inch difference in height (taller children will weigh more as the coefficient is positive).*
- ★ *In other words, if we were to compare average weights for two groups of children of the same age who differ by one inch in height, the expected difference in mean weights would be 0.72 lbs.*



JOHNS HOPKINS
BLOOMBERG
SCHOOL *of* PUBLIC HEALTH

Section B

Statistical Inference on Multiple Regression
Coefficients

Statistical Inference for Multiple Linear Regression Coefficients

As before, we are estimating regression parameters (coefficients) from a sample

Each slope is an estimate of the true value in the population

Statistical Inference for Multiple Linear Regression Coefficients

Each estimated regression coefficient
($\hat{b}_1, \hat{b}_2, \hat{b}_3, \dots, \hat{b}_p$) has an associated standard error (SE)

Statisticians have developed formulas for standard errors of multiple linear regression coefficients

Statistical Inference for Multiple Linear Regression Coefficients

Generally, the standard errors for each \hat{b}_j gets smaller with larger sample size

Statistical Inference for Multiple Linear Regression Coefficients

Standard errors for \hat{b}_j also depend on ...

- ★ *Distance of points from “line”*
- ★ *Variability in x_j*
- ★ *Number of x variables*
- ★ *Correlation between x variables*

Statistical Inference for Multiple Linear Regression Coefficients

95% confidence interval for regression coefficient \hat{b}_j

$$\hat{b}_j \pm t_{n-(p+1)} SE(\hat{b}_j)$$

Where $t_{n-(p+1)}$ is the t-value with **n-(p+1)** degrees of freedom, where **n** is the sample size and **p** is total number of predictors

(in large samples, just $\hat{b}_j \pm 2SE(\hat{b}_j)$)

95% CI for coefficient of PCV in regression model of Hb on PCV and age

```
. regress Hb PCV age
```

Source	SS	df	MS			
Model	88.7923124	2	44.3961562	Number of obs =	21	
Residual	16.5591129	18	.919950714	F(2, 18) =	48.26	
Total	105.351425	20	5.26757126	Prob > F =	0.0000	
				R-squared =	0.8428	
				Adj R-squared =	0.8254	
				Root MSE =	.95914	

Hb	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
PCV	.1023427	.0312317	3.277	0.004	.0367274	.167958
age	.1013414	.0164271	6.169	0.000	.0668293	.1358534
_cons	5.516596	1.114866	4.948	0.000	3.174349	7.858842

95% confidence interval for regression coefficient for PCV (b_1)

$$\star \hat{b}_1 \pm t_{18} SE(\hat{b}_1)$$

95% confidence interval for regression coefficient for PCV (b_1)

$$\star .102 \pm 2.1*(.0312)$$

Where 2.1 is the t-value from a t-distribution with $21 - (2+1) = 18$ d.f.

95% confidence interval for regression coefficient for PCV

★ $.102 \pm .0655$

★ $(0.037, 0.1675)$

After accounting for age, a positive association was estimated between hemoglobin and PCV

Amongst subjects of the same age, a one percent increase in PCV percent is associated with a .10 g/dL increase in hemoglobin on average

Accounting for sampling variability, this increase could be as small as .04 g/dL and as large as .17 g/dL.

95% CI for coefficient of age in regression model of Hb on PCV and age

```
. regress Hb PCV age
```

Source	SS	df	MS			
Model	88.7923124	2	44.3961562	Number of obs =	21	
Residual	16.5591129	18	.919950714	F(2, 18) =	48.26	
Total	105.351425	20	5.26757126	Prob > F =	0.0000	
				R-squared =	0.8428	
				Adj R-squared =	0.8254	
				Root MSE =	.95914	

Hb	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
PCV	.1023427	.0312317	3.277	0.004	.0367274	.167958
age	.1013414	.0164271	6.169	0.000	.0668293	.1358534
_cons	5.516596	1.114866	4.948	0.000	3.174349	7.858842

95% confidence interval for regression coefficient for PCV

★ $(0.067, 0.136)$

After accounting for PCV, a positive association was estimated between hemoglobin and age

Amongst subjects of the same PCV, a one year increase in age is associated with a .10 g/dL in hemoglobin on average

Accounting for sampling variability, this increase could be as small as .07 g/dL and as large as .14 g/dL

Hypothesis testing for multiple regression coefficients

★ $H_o: b_j = 0$

★ $H_a: b_j \neq 0$

Hypothesis test

★ $H_o: b_j = 0$

★ $H_a: b_j \neq 0$

Is there a statistically significant relationship between y and X_j after adjusting for all other factors in the equation?

Statistical Inference for Multiple Linear Regression Coefficients

Hypothesis test

★ Calculate test statistic:

$$t = \frac{\hat{b}_j}{SE(\hat{b}_j)}$$

Hypothesis test for coefficient of PCV in regression of Hb on PCV and age

★ *Calculate test statistic:*

$$t = \frac{.102}{.0312} = 3.27$$

Hypothesis test for coefficient of PCV in regression of Hb on PCV and age

- ★ *Get a p -value from a t -table with 18 d.f., or Stata (better choice!)*
- ★ *$p = .004$*

After accounting for age, there is a statistically significant relationship between hemoglobin and PCV

P-value for testing coefficient of PCV in regression model of Hb on PCV and age

```
. regress Hb PCV age
```

Source	SS	df	MS		Number of obs =	21
Model	88.7923124	2	44.3961562		F(2, 18) =	48.26
Residual	16.5591129	18	.919950714		Prob > F =	0.0000
Total	105.351425	20	5.26757126		R-squared =	0.8428
					Adj R-squared =	0.8254
					Root MSE =	.95914

Hb	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
PCV	.1023427	.0312317	3.277	0.004	.0367274	.167958
age	.1013414	.0164271	6.169	0.000	.0668293	.1358534
_cons	5.516596	1.114866	4.948	0.000	3.174349	7.858842

Hypothesis test for coefficient of age in regression of Hb on PCV and age

- ★ *Get a p -value from a t -table with 18 d.f., or Stata (better choice!)*
- ★ *$p < .001$*

After accounting for PCV, there is a statistically significant relationship between hemoglobin and age

P-value for testing coefficient of age in regression model of Hb on PCV and age

```
. regress Hb PCV age
```

Source	SS	df	MS			
Model	88.7923124	2	44.3961562	Number of obs =	21	
Residual	16.5591129	18	.919950714	F(2, 18) =	48.26	
Total	105.351425	20	5.26757126	Prob > F =	0.0000	
				R-squared =	0.8428	
				Adj R-squared =	0.8254	
				Root MSE =	.95914	

Hb	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
PCV	.1023427	.0312317	3.277	0.004	.0367274	.167958
age	.1013414	.0164271	6.169	0.000	.0668293	.1358534
_cons	5.516596	1.114866	4.948	0.000	3.174349	7.858842

Notes on T-Test for Regression Coefficients

Evaluates the relationship of a predictor with the outcome after accounting for other predictor variables in the model

Does the predictor add additional information about the mean of y in addition to the information in the other predictors (x s)?

There is a more general test associated with MLR called the “overall F-test”

Answers the general question “are any of the predictors in the model statistically important?”

Formulation

- ★ $H_0: b_1, b_2, b_3 \dots b_p = 0$
- ★ $H_a: \text{At least one } b \neq 0$
- ★ *P-value allows evaluation of whether regression model is “better than nothing”*
- ★ *If null is rejected, individual t-tests for predictors can be performed to see which coefficients are statistically important*

The Overall F-Test in Hemoglobin, PCV, Age Example

P-value for overall F-test

```
. regress Hb PCV age
```

Source	SS	df	MS
Model	88.7923124	2	44.3961562
Residual	16.5591129	18	.919950714
Total	105.351425	20	5.26757126

```
Number of obs = 21  
F( 2, 18) = 48.26  
Prob > F = 0.0000  
R-squared = 0.8428  
Adj R-squared = 0.8254  
Root MSE = .95914
```

Hb	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
PCV	.1023427	.0312317	3.277	0.004	.0367274	.167958
age	.1013414	.0164271	6.169	0.000	.0668293	.1358534
_cons	5.516596	1.114866	4.948	0.000	3.174349	7.858842

Continued

The Overall F-Test in Hemoglobin, PCV, Age Example

P-value of overall F-test is testing:

- ★ $H_0: b_1 = b_2 = 0$
- ★ $H_a: \text{at least one of } b_1, b_2 \neq 0$



JOHNS HOPKINS
BLOOMBERG
SCHOOL *of* PUBLIC HEALTH

Section B

Practice Problems

1. Consider the following results from a MLR regressing weight (lbs) on height (in) and age (years) for a sample of 12 nutritionally deficient children. The results, on the next slide, are missing some key information.

Here is the regression output from Stata

```
. regress wt ht age
```

Source	SS	df	MS			
Model	692.822607	2	346.411303	Number of obs =	12	
Residual	195.427393	9	21.7141548	F(2, 9) =	15.95	
Total	888.25	11	80.75	Prob > F =	0.0011	
				R-squared =	0.7800	
				Adj R-squared =	0.7311	
				Root MSE =	4.6598	

	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
ht	.722038	.2608051			
age	2.050126	.9372256			
_cons	6.553048	10.94483			

Continued

2a. If we call the coefficient of height in this model b_1 (estimated by \hat{b}_1), what substantive question is being addressed by the statistical hypothesis test:

$$H_o: b_1 = 0$$

$$H_a: b_1 \neq 0$$

2b. Compute a 95% CI for b_1 (the appropriate t-value that cuts off 95% for a t_9 distribution is 2.26). Comment on both the statistical and scientific interpretations of the results.

2c. Based on the results given is it possible to ascertain whether child's age confounds the relationship between weight and height?



JOHNS HOPKINS
BLOOMBERG
SCHOOL *of* PUBLIC HEALTH

Section B

Practice Problem Solutions

1. Consider the following results from a MLR regressing weight (lbs) on height (in) and age (years) for a sample of 12 nutritionally deficient children. The results, on the next slide, are missing some key information.

Here is the regression output from Stata

```
. regress wt ht age
```

Source	SS	df	MS	
Model	692.822607	2	346.411303	
Residual	195.427393	9	21.7141548	
Total	888.25	11	80.75	

				Number of obs =	
				F(2, 9) =	15.95
				Prob > F =	0.001
				R-squared =	0.780
				Adj R-squared =	0.731
				Root MSE =	4.659

	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
ht	.722038	.2608051			
age	2.050126	.9372256			
_cons	6.553048	10.94483			

Continued

2a. If we call the coefficient of height in this model b_1 (estimated by \hat{b}_1), what substantive question is being addressed by the statistical hypothesis test:

$$H_o: b_1 = 0$$

$$H_a: b_1 \neq 0$$

2a. Substantively speaking, the question being asked is “after adjusting for the association with a child’s age, is a child’s height statistically significantly associated with his/her weight?”

2b. Compute a 95% CI for b_1 (the appropriate t-value that cuts off 95% for a t_9 distribution is 2.26). Comment on both the statistical and scientific interpretations of the results.

2b. Because we are estimating an intercept, and two slopes in this model we lose three degrees of freedom. Hence, we want to get the number of standard errors we need to add and subtract to our estimate to get 95% confidence from a t-table with nine df—but I already told you that this value is 2.26.

To construct a 95% CI:

$$\hat{b}_1 \pm t_{9,.95} se(\hat{b}_1)$$

$$0.72 \pm 2.26 * 0.26$$

$$.72 \pm .59$$

$$(.13, 1.31)$$

Substantive Interpretation

A multiple linear regression was used to estimate the age adjusted association between child weight and height in a sample of 12 malnourished children. In this sample, two groups of children of the same age who differ by one inch in height will differ by 0.72 pounds on average (95% CI 0.13–1.31 pounds), the taller children weighing more.

Based on the results given is it possible to ascertain whether child's age confounds the relationship between weight and height?

Without seeing the results of a simple linear regression of weight on height alone, it is not possible to comment based solely on the results given. Just FYI, the results of this regression are on the following slide.

Regression of weight on height

```
. regress wt ht
```

Source	SS	df	MS			
Model	588.922523	1	588.922523	Number of obs =	12	
Residual	299.327477	10	29.9327477	F(1, 10) =	19.67	
Total	888.25	11	80.75	Prob > F =	0.0013	
				R-squared =	0.6630	
				Adj R-squared =	0.6293	
				Root MSE =	5.4711	

wt	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
ht	1.07223	.241731	4.44	0.001	.5336202	1.610841
_cons	6.189849	12.84875	0.48	0.640	-22.43894	34.81864



JOHNS HOPKINS
BLOOMBERG
SCHOOL *of* PUBLIC HEALTH

Section C

Handling Multiple Categorical Predictors in
Multiple Linear Regression: ANOVA as a Regression
Model

Sometimes, regression scenarios include predictors which are not continuous, not binary, but multi-categorical

Examples

- ★ *Subject's race (White, African-American, Hispanic, Asian, Other)*
- ★ *City of residence (Baltimore, Chicago, Tokyo, Madrid)*

How can this type of situation be handled in a regression framework?

We'll explore with an example using a data set containing information about average SAT scores in 51 U.S. states (treating D.C. as a state)—the averages were based on random samples of students taken within each of the 51 states

The SAT (Scholastic Aptitude Test) is taken by many U.S. high school students to fulfill requirements for entry into most colleges or universities

The test is made up of two components: verbal and quantitative (math)

This analysis will use the quantitative score, which ranges from 200-800 (we will refer to these simply as SAT scores for simplicity)

This data comes from the book *Statistics with Stata 8*, by Lawrence Hamilton

Data consists of 51 observations: the cumulative average SAT quantitative section scores for the 51 U.S. states for students taking the test in 1990

Additional information on each observation includes geographical region of the state (West, Northeast, South, Midwest) and per-pupil education expenditures in each state in 1990

A key question

- ★ *Do average SAT scores differ across the four regions of the country and, if so, what is the magnitude of these differences?*

Snippet of the Data

Here is a snippet of the data in Stata (*msat* is mean SAT score for the state, *region* is a “labeled” numerical variable)

```
+-----+
|  msat   region |
+-----+
1.   515     South |
2.   481     West  |
3.   490     West  |
4.   523     South |
5.   482     West  |
+-----+
6.   506     West  |
7.   468     N. East |
```

Analysis of variance testing for differences between the four states yields a p-value of less than 0.001

So there are at least some statistical differences in SAT quantitative scores across the four regions—but in order to find out which regions are statistically different and to figure out by how large (and in what direction) these differences occur would require a lot of t-tests

ANOVA as a Regression Model

Could this analysis be done by a linear regression relating SAT scores to region?

How can we handle a predictor that takes on four categories?

ANOVA as a Regression Model

Arbitrarily give each region a numerical value ($X_1 = 1$ for Western region states, 2 for Northeastern states, 3 for Southern states, and 4 for mid-Western states for example), and fit SLR of

$$\hat{y} = \hat{b}_0 + \hat{b}_1 x_1$$

Where \hat{y} is estimate mean SAT score, and X_1 is region as defined above

Potential Problem

- ★ Coding is arbitrary, could have assigned $X_1 = 1$ for Midwest, etc. . . .
- ★ Estimated coefficient of region will depend on arbitrary coding

Potential Problem

Coding “assumes” mean SAT score differences between regions “incremental”

- ★ *Example—difference in average SAT scores between Southern states ($X_1 = 3$) and Western States ($X_1 = 1$) is twice the difference between Northeastern States ($X_1 = 2$) and Western States ($X_1 = 1$)*

This is not a good idea!!!

ANOVA as a Regression Model

Alternative approach—designate one region as “reference” region, say Western region, and make binary indicators for each of the three other regions

- ★ $X_1 = 1$ if Northeastern state, 0 otherwise
- ★ $X_2 = 1$ if Southern state, 0 otherwise
- ★ $X_3 = 1$ if mid-Western state, 0 otherwise

ANOVA as a Regression Model

Here is a table showing the x values for each region

Region	X_1	X_2	X_3
West	0	0	0
Northeast	1	0	0
South	0	1	0
MidWest	0	0	1

Fit the regression model

$$\hat{y} = \hat{b}_0 + \hat{b}_1x_1 + \hat{b}_2x_2 + \hat{b}_3x_3$$

Here, each coefficient estimates mean SAT score difference between region that has a corresponding x value of 1 and the reference region (Western states)

Notice, the intercept has meaning here—it's the estimated mean when all x 's are 0, the estimated mean SAT score for Western states!

Example

- ★ For Northeastern states ($X_1=1, X_2=0, X_3=0$) model predicts

$$\hat{y} = \hat{b}_0 + \hat{b}_1 * 1 + \hat{b}_2 * 0 + \hat{b}_3 * 0$$

- ★ For Western states ($X_1=0, X_2=0, X_3=0$) model predicts

$$\hat{y} = \hat{b}_0 + \hat{b}_1 * 0 + \hat{b}_2 * 0 + \hat{b}_3 * 0$$

Example

- ★ For Northeastern states ($\mathbf{X}_1 = 1, \mathbf{X}_2 = 0, \mathbf{X}_3 = 0$) model predicts

$$\hat{y} = \hat{b}_0 + \hat{b}_1$$

- ★ For Western states ($\mathbf{X}_1 = 0, \mathbf{X}_2 = 0, \mathbf{X}_3 = 0$) model predicts

$$\hat{y} = \hat{b}_0$$

Example

$$\star \text{ So } \hat{y}_{NE} - \hat{y}_W = \hat{b}_0 + \hat{b}_1 - \hat{b}_0 = \hat{b}_1$$

Similar results can be shown for other coefficients

ANOVA as a Regression Model

Stata results

Notice, data in the following format . . .

```
+-----+
      msat  region
-----
1.    515      3
2.    481      1
3.    490      1
4.    523      3
5.    482      1
-----
6.    506      1
7.    468      2
8.    464      3
```

“xi” option before regression command will automatically create binary indicators for a multi-categorical variable

Syntax

★ *xi: regress msat i.region*

ANOVA as a Regression Model

Stata results

```
. xi: regress msat i.region
i.region          _Iregion_1-4          (naturally coded; _Iregion_1 omitted)
```

Source	SS	df	MS	Number of obs =	50
Model	26433.6728	3	8811.22428	F(3, 46) =	12.26
Residual	33050.8072	46	718.495808	Prob > F =	0.0000
Total	59484.48	49	1213.96898	R-squared =	0.4444
				Adj R-squared =	0.4081
				Root MSE =	26.805

msat	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
_Iregion_2	-32.01709	11.62333	-2.75	0.008	-55.41364 -8.620547
_Iregion_3	-11.64904	10.00874	-1.16	0.250	-31.79559 8.497512
_Iregion_4	35.45513	10.7305	3.30	0.002	13.85576 57.0545
_cons	498.4615	7.434306	67.05	0.000	483.4971 513.426

Resulting regression equation

$$\hat{y} = \hat{b}_0 + \hat{b}_1x_1 + \hat{b}_2x_2 + \hat{b}_3x_3$$

$$\hat{y} = 498.5 - 32.0x_1 - 11.6x_2 + 35.5x_3$$

ANOVA as a Regression Model

Overall F-test

```
. xi: regress msat i.region
i.region      _Iregion_1-4      (naturally coded; _Iregion_1 omitted)
```

Source	SS	df	MS	Number of obs =	50
Model	26433.6728	3	8811.22428	F(3, 46) =	12.26
Residual	33050.8072	46	718.495808	Prob > F =	0.0000
Total	59484.48	49	1213.96898	R-squared =	0.4444
				Adj R-squared =	0.4081
				Root MSE =	26.805

msat	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
_Iregion_2	-32.01709	11.62333	-2.75	0.008	-55.41364 -8.620547
_Iregion_3	-11.64904	10.00874	-1.16	0.250	-31.79559 8.497512
_Iregion_4	35.45513	10.7305	3.30	0.002	13.85576 57.0545
_cons	498.4615	7.434306	67.05	0.000	483.4971 513.426

Continued

This is the overall test for . . .

- ★ H_0 —*no differences in mean SAT scores across the four regions*
- ★ H_a —*at least one region has different mean SAT scores than the other*
- ★ *This is the same exact test that we did with the traditional ANOVA approach*

ANOVA as a Regression Model

Some of the estimated regional differences

```
. xi: regress msat i.region  
i.region          _Iregion_1-4          (naturally coded; _Iregion_1 omitted)
```

Source	SS	df	MS	Number of obs =	50
Model	26433.6728	3	8811.22428	F(3, 46) =	12.26
Residual	33050.8072	46	718.495808	Prob > F =	0.0000
Total	59484.48	49	1213.96898	R-squared =	0.4444
				Adj R-squared =	0.4081
				Root MSE =	26.805

msat	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
_Iregion_2	-32.01709	11.62333	-2.75	0.008	-55.41364 -8.620547
_Iregion_3	-11.64904	10.00874	-1.16	0.250	-31.79559 8.497512
_Iregion_4	35.45513	10.7305	3.30	0.002	13.85576 57.0545
_cons	498.4615	7.434306	67.05	0.000	483.4971 513.426

A statistically significant relationship was found between mean SAT scores and student's region of the country ($p < .0001$ by F-test)

Students from Northeastern states had SAT scores of 32 points lower on average than students from Western states (95% CI 8.6 to 55.4 points lower)

Students from Southern states had SAT scores of 11.6 points lower on average than students from Western states (95% CI 31.6 points lower to 8.5 points higher)

Students from mid-Western states had SAT scores of 35.5 points higher on average than students from Western states (95% CI 13.9 points to 57.0 points higher)

Regional differences account for 44% of the variation in SAT scores

What about other comparisons—for example, SAT scores for Northeastern states to mid-Western states?

- ★ *One approach—recode indicators for region making “Mid-West” the reference group—more work!*
- ★ *Another option—use existing coefficients*

Recall $\hat{b}_1 = \hat{y}_{NE} - \hat{y}_W = -32.0$ estimates the average difference in SAT scores for Northeastern States minus (compared to) Western States

Recall $\hat{b}_3 = \hat{y}_{MW} - \hat{y}_W = 35.5$ estimates the average difference in SAT scores for Midwestern States minus (compared to) Western States

So :

$$\begin{aligned}\hat{b}_1 - \hat{b}_3 &= \hat{y}_{NE} - \hat{y}_W - (\hat{y}_{MW} - \hat{y}_W) \\ &= \hat{y}_{NE} - \hat{y}_W - \hat{y}_{MW} + \hat{y}_W \\ &= \hat{y}_{NE} - \hat{y}_{MW}\end{aligned}$$

So the estimated mean difference in SAT scores between Northeastern States and mid-Western states is given by $(-32.0 - 35.5) = -67.5$ points

We can employ Stata to do this and get a 95% CI (just FYI)

The “lincom” command can be run after any regression to give estimates for differences in coefficients

ANOVA as a Regression Model

We need to use names Stata gives to coefficients in commands

```
. xi: regress msat i.region
i.region      _Iregion_1-4      (naturally coded; _Iregion_1 omitted)
```

Source	SS	df	MS	Number of obs =	50
Model	26433.6728	3	8811.22428	F(3, 46) =	12.26
Residual	33050.8072	46	718.495808	Prob > F =	0.0000
Total	59484.48	49	1213.96898	R-squared =	0.4444
				Adj R-squared =	0.4081
				Root MSE =	26.805

msat	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
<u>_Iregion_2</u>	-32.01709	11.62333	-2.75	0.008	-55.41364 -8.620547
<u>_Iregion_3</u>	-11.64904	10.00874	-1.16	0.250	-31.79559 8.497512
<u>_Iregion_4</u>	35.45513	10.7305	3.30	0.002	13.85576 57.0545
_cons	498.4615	7.434306	67.05	0.000	483.4971 513.426

Continued

Syntax

★ *lincom _Iregion_2 - _Iregion_4*

```
. lincom _Iregion_2- _Iregion_4  
( 1)  _Iregion_2 - _Iregion_4 = 0
```

msat	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
(1)	-67.47222	11.81979	-5.71	0.000	-91.26423 -43.68021

ANOVA is just a specific form of linear regression

In general, if we have a categorical predictor with k categories, we designate one category as the reference group, and create $k-1$ binary indicators X_1, X_2, X_{k-1} for all other levels of the predictor

Coefficients are interpretable as mean difference in the outcome between each of the $k-1$ categories and the reference group

Not only do we get an overall test for any mean outcome differences between the groups being compared, we also get estimates and 95% CIs for some of the differences

This approach also gives an R^2 value

We can also expand regression model to include more predictors (example—SAT scores predicted by both region and per-pupil state expenditures)



JOHNS HOPKINS
BLOOMBERG
SCHOOL *of* PUBLIC HEALTH

Section D

Statistical Interaction and Linear Regression

Data on elevation and percentage of dead or badly damaged trees, from 64 Appalachian sites (reported by *Committee on Monitoring and Assessment of Trends in Acid Deposition*, 1986)

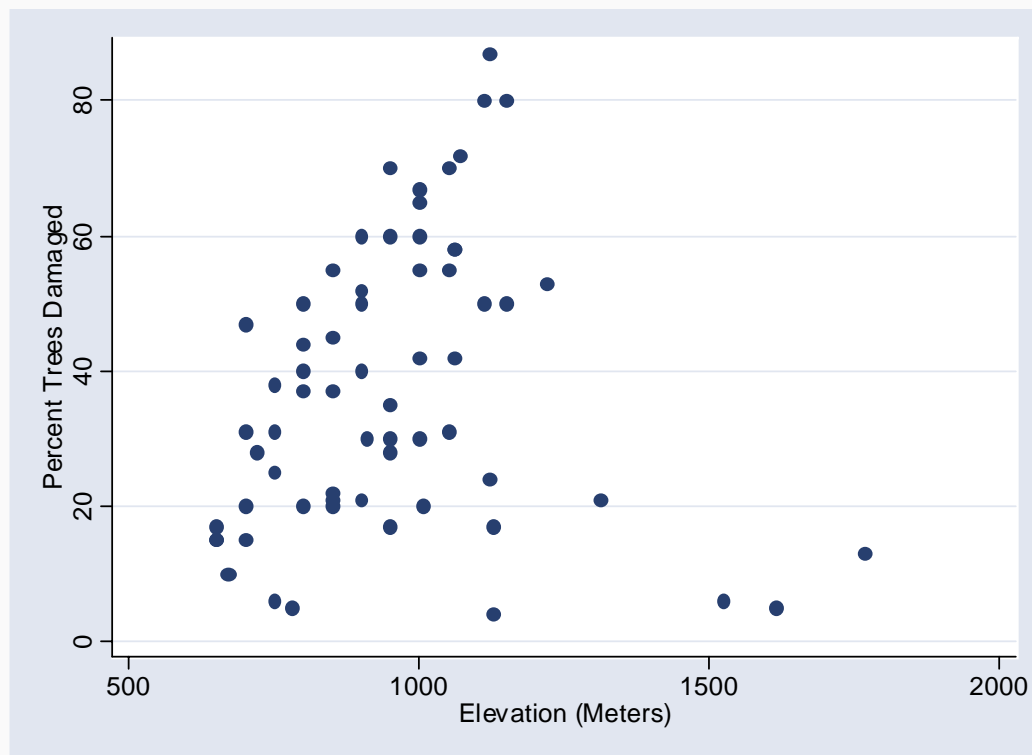
Eight of the 64 sites are in Southern states

Data, as it appears in Stata

```
+-----+
| region   elev   damage |
+-----+-----+
1.  | South   1615     5   |
2.  | South   1768    13   |
3.  | South   1524     6   |
4.  | South   1311    21   |
5.  | South   1128     4   |
+-----+-----+
6.  | South   1005    20   |
7.  | South   1128    17   |
8.  | South   1052    31   |
9.  | North    670    10   |
10. | North    720    28   |
+-----+-----+
```

	region	elev	damage
1.	South	1615	5
2.	South	1768	13
3.	South	1524	6
4.	South	1311	21
5.	South	1128	4
6.	South	1005	20
7.	South	1128	17
8.	South	1052	31
9.	North	670	10
10.	North	720	28

Scatterplot of percent of damaged trees on elevation



SLR of percent of damaged trees on elevation

```
. regress damage elev
```

Source	SS	df	MS			
Model	238.14136	1	238.14136	Number of obs =	64	
Residual	27807.7961	62	448.512841	F(1, 62) =	0.53	
				Prob > F =	0.4689	
				R-squared =	0.0085	
				Adj R-squared =	-0.0075	
Total	28045.9375	63	445.173611	Root MSE =	21.178	

damage	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
elev	.0088774	.0121831	0.73	0.469	-.0154762	.033231
_cons	29.07121	11.90825	2.44	0.018	5.266964	52.87546

Scatterplot of percent of damaged trees on region



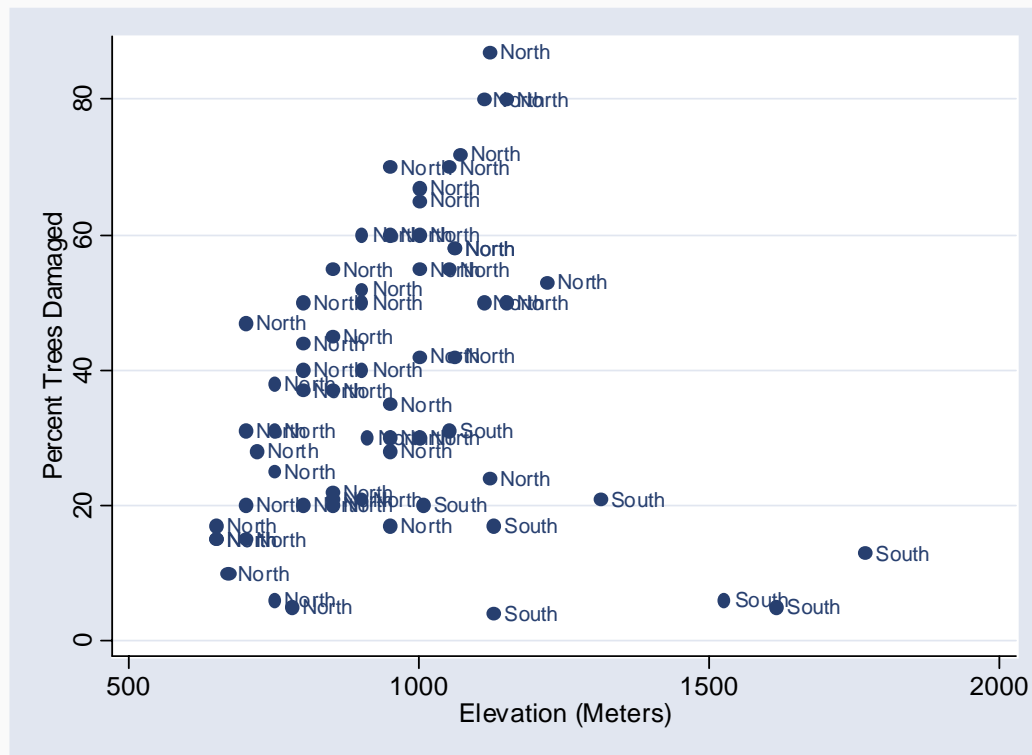
SLR of percent of damaged trees on region

```
. regress damage region
```

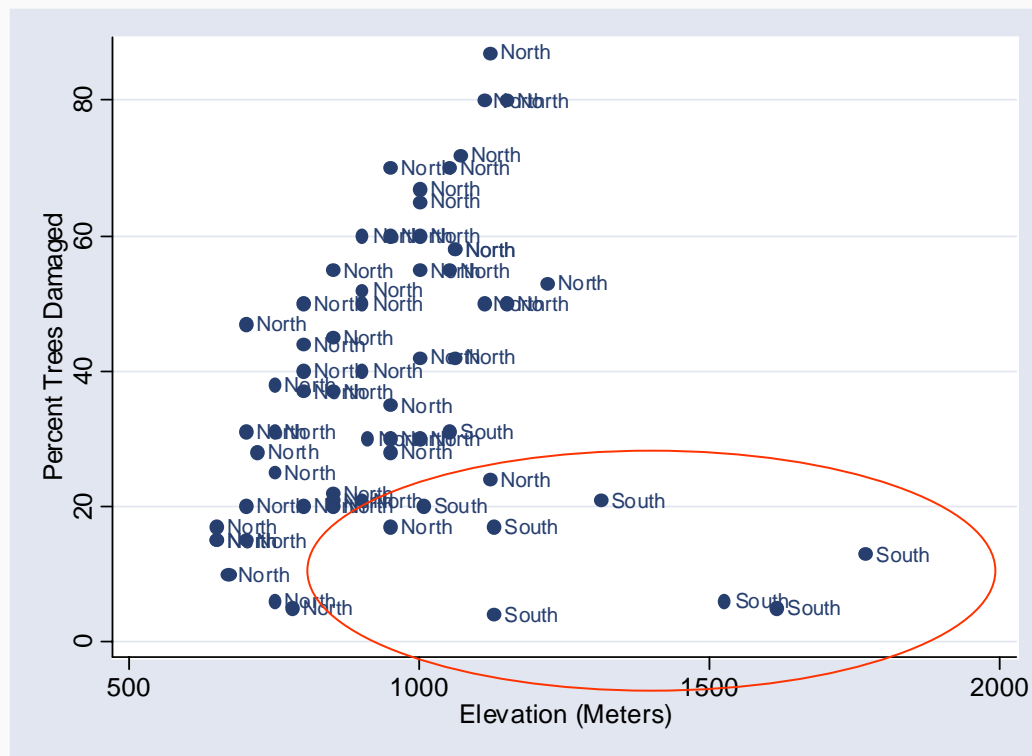
Source	SS	df	MS	
Model	4797.22321	1	4797.22321	Number of obs = 64
Residual	23248.7143	62	374.979263	F(1, 62) = 12.79
				Prob > F = 0.0007
				R-squared = 0.1710
				Adj R-squared = 0.1577
Total	28045.9375	63	445.173611	Root MSE = 19.364

damage	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
region	26.17857	7.319048	3.58	0.001	11.548 40.80914
_cons	14.625	6.846343	2.14	0.037	.9393562 28.31064

Scatterplot of damage by elevation with points labeled by region



Scatterplot of damage by elevation with points labeled by region

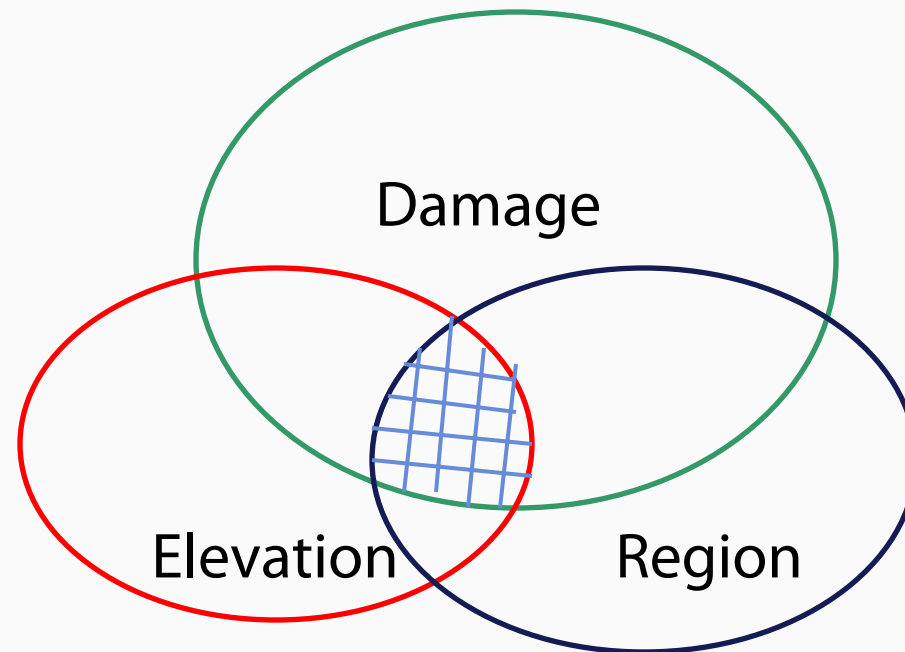


Southern region trees are at higher elevation

Southern region trees have less damage

- ★ *Damage related to elevation*
- ★ *Damage related to region*
- ★ *Elevation related to region*

Possible diagram of scenario



Let's examine damage/elevation relationship in Northern states

-> region = North

Source	SS	df	MS			
Model	10260.5277	1	10260.5277	Number of obs =	56	
Residual	12362.3116	54	228.931695	F(1, 54) =	44.82	
Total	22622.8393	55	411.324351	Prob > F =	0.0000	
				R-squared =	0.4535	
				Adj R-squared =	0.4434	
				Root MSE =	15.13	

damage	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
elev	.0909526	.0135857	6.69	0.000	.0637149	.1181904
_cons	-41.15124	12.40757	-3.32	0.002	-66.02692	-16.27555

Let's examine damage/elevation relationship in Southern states

```
-> region = South
```

Source	SS	df	MS	Number of obs = 8		
Model	170.09548	1	170.09548	F(1, 6) =	2.24	
Residual	455.77952	6	75.9632534	Prob > F =	0.1852	
Total	625.875	7	89.4107143	R-squared =	0.2718	
				Adj R-squared =	0.1504	
				Root MSE =	8.7157	

damage	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
elev	-.0172129	.0115029	-1.50	0.185	-.0453596	.0109338
_cons	37.2836	15.45254	2.41	0.052	-.5274096	75.09461

Overall slope of damage on elevation

★ 0.009

Slope of damage on elevation by region

★ *South: -0.017*

★ *North: 0.09*

Confounding?

- ★ *Southern trees—less damage, higher elevation*
- ★ *Overall slope showing very small relationship between damage and elevation is in part because of confounding between elevation and region as both relate to damage*
- ★ *How do you get “adjusted” relationship between damage and elevation after removing portion attributable to region/elevation relationship?*

Regression of damage on both elevation and region

```
. regress damage elev region
```

Source	SS	df	MS			
Model	10636.2722	2	5318.13608	Number of obs =	64	
Residual	17409.6653	61	285.40435	F(2, 61) =	18.63	
Total	28045.9375	63	445.173611	Prob > F =	0.0000	
				R-squared =	0.3792	
				Adj R-squared =	0.3589	
				Root MSE =	16.894	

damage	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
elev	.0567284	.0125418	4.52	0.000	.0316495	.0818073
region	49.73808	8.240274	6.04	0.000	33.26063	66.21552
_cons	-60.05084	17.55694	-3.42	0.001	-95.15811	-24.94358

After adjusting for region, a one meter increase in elevation increased the percentage of damaged trees by .05 on average (95% CI .03% to .08%)

After adjusting for elevation, trees in Northern states suffered tree damage on average 50% higher than Southern states (95% CI 32% to 62 %)

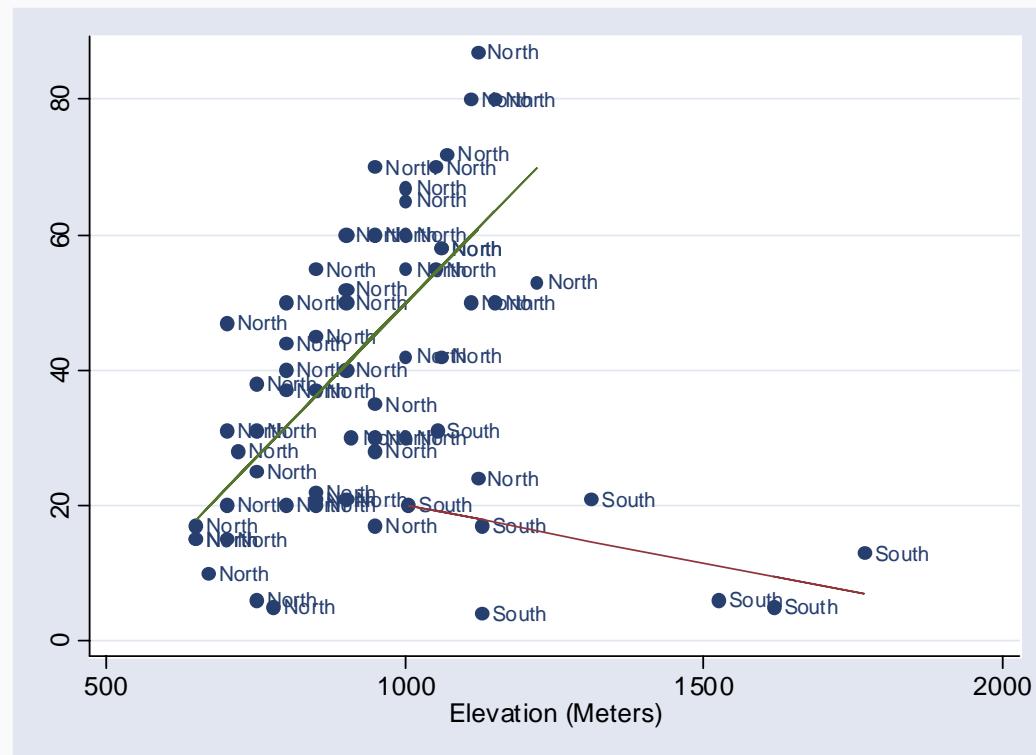
Together, elevation and region explain about 38% of the variability in percentage of damaged trees

This model “assumes” same damage/elevation for each region

Recall, when we regressed damage on elevation separately for each region, we got different estimates of the damage/elevation association

- ★ *South: -0.017*
- ★ *North: 0.09*
- ★ *Do we really think the underlying relationship between damage and elevation is the same for both regions?*

Scatterplot with separate regression lines for each region



Suppose we wanted allow for different estimates of the relationship between damage and elevation depending on region

We could run two separate regressions, damage on elevation for each region, and report the results separately

Does this Suggest Effect Modification?

Coefficient estimates for elevation “look different” for each region but are they really different?

We would need to do a formal hypothesis test/estimation of the difference comparing elevation coefficient across models! (a pain!)

Is there any way this could be done without running separate regressions?

More Elegant Solution!

Add an interaction term to the model which already include elevation (x_1) and region (x_2) as predictors

Add $x_3 = x_1 * x_2$

New model

$$\hat{y} = \hat{b}_0 + \hat{b}_1 x_1 + \hat{b}_2 x_2 + \hat{b}_3 x_3$$

What is value of x_3 for Northern States? Southern States?
(recall $x_2 = 1$ if North, 0 if South)

Model for Southern states ($x_2 = 0$)

$$\hat{y} = \hat{b}_0 + \hat{b}_1 x_1$$

So \hat{b}_1 estimates the magnitude of the relationship between damage and elevation in Southern states—it is the slope of elevation (x_1) for Southern states

Model for Northern states ($x_2=1$)

$$\hat{y} = \hat{b}_0 + \hat{b}_1 x_1 + \hat{b}_2 x_2 + \hat{b}_3 x_1$$

$$\hat{y} = \hat{b}_0 + \hat{b}_2 x_2 + (\hat{b}_1 + \hat{b}_3) x_1$$

So $(\hat{b}_1 + \hat{b}_3)$ estimates the magnitude of the relationship between damage and elevation in Northern states—it is the slope of elevation (x_1) for Northern states

So \hat{b}_3 is an estimate of the difference in the relationship between damage and elevation for the two different region

It allows for the estimated “effect” of elevation on damage to be different depending on the region

If there is no difference in the damage/elevation relationship, then the true value of b_3 would be 0

Testing the following hypothesis:

- ★ $H_0: b_3 = 0$
- ★ $H_a: b_3 \neq 0$
- ★ *Is sometimes called a "test of interaction"*

To create interaction term, use generate command

★ *generate interact = region*elev*

Results from Stata

```
. gen interact = region* elev
```

```
. regress damage region elev interact
```

Source	SS	df	MS	Number of obs =	64
Model	15227.8464	3	5075.94881	F(3, 60) =	23.76
Residual	12818.0911	60	213.634851	Prob > F =	0.0000
				R-squared =	0.5430
				Adj R-squared =	0.5201
Total	28045.9375	63	445.173611	Root MSE =	14.616

damage	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
region	-78.43484	28.55164	-2.75	0.008	-135.5466	-21.32306
elev	-.0172129	.0192905	-0.89	0.376	-.0557996	.0213738
interact	.1081655	.0233316	4.64	0.000	.0614954	.1548356
_cons	37.2836	25.91399	1.44	0.155	-14.55209	89.11929

Results from Stata

```
. gen interact = region* elev
. regress damage region elev interact
```

Source	SS	df	MS	Number of obs =	64
Model	15227.8464	3	5075.94881	F(3, 60) =	23.76
Residual	12818.0911	60	213.634851	Prob > F =	0.0000
				R-squared =	0.5430
				Adj R-squared =	0.5201
Total	28045.9375	63	445.173611	Root MSE =	14.616

damage	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
region	-78.43484	28.55164	-2.75	0.008	-135.5466 -21.32306
elev	-.0172129	.0192905	-0.89	0.376	-.0557996 .0213738
interact	.1081655	.0233316	4.64	0.000	.0614954 .1548356
_cons	37.2836	25.91399	1.44	0.155	-14.55209 89.11929

\hat{b}_2

\hat{b}_1

\hat{b}_3

\hat{b}_0

Model for Southern states ($x_2 = 0$)

$$\hat{y} = \hat{b}_0 + -.017x_1$$

The estimated relationship between damage and elevation for Southern states is negative

On average a one meter increase in elevation is associated with a .017 percent decrease in the percentage of damaged trees but these results are not conclusive of a negative relationship (95% CI: -.05% decrease to 0.02 % increase)

Model for Northern states ($x_2 = 1$)

$$\hat{y} = \hat{b}_0 + \hat{b}_2 x_2 + (\hat{b}_1 + \hat{b}_3) x_1$$

$$\hat{y} = \hat{b}_0 + \hat{b}_2 x_2 + (-.017 + .11) x_1$$

$$\hat{y} = \hat{b}_0 + \hat{b}_2 x_2 + (.09) x_1$$

The estimated relationship between damage and elevation for Northern states is positive

On average a one meter increase in elevation is associated with a .09 percent increase in the percentage of damaged trees

Results from Stata

```
. gen interact = region* elev
. regress damage region elev interact
```

Source	SS	df	MS	Number of obs =	64
Model	15227.8464	3	5075.94881	F(3, 60) =	23.76
Residual	12818.0911	60	213.634851	Prob > F =	0.0000
				R-squared =	0.5430
				Adj R-squared =	0.5201
Total	28045.9375	63	445.173611	Root MSE =	14.616

	damage	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
\hat{b}_2	region	-78.43484	28.55164	-2.75	0.008	-135.5466 -21.32306
\hat{b}_1	elev	-.0172129	.0192905	-0.89	0.376	-.0557996 .0213738
\hat{b}_3	interact	.1081655	.0233316	4.64	0.000	.0614954 .1548356
\hat{b}_0	_cons	37.2836	25.91399	1.44	0.155	-14.55209 89.11929

Confidence interval is trickier—must appeal to Stata and lincom command

```
. lincom elev+interact
```

```
( 1) elev + interact = 0
```

damage	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
(1)	.0909526	.013124	6.93	0.000	.0647007 .1172045

Statistical interaction (effect modification) occurs when the relationship between an outcome and predictor one is different depending on the level of predictor two

- ★ *BP treatment good for men, bad for women*
- ★ *Relationship between cholesterol level and red meat consumption different for smokers and non-smokers*

Interactions are usually investigated because of apriori assumptions/hypotheses on the part of the researchers

Linear regression allows for the inclusion of interactions (and formal tests of interaction) within a single regression model