Topic #3: Linear Models & Linear Regression

Objectives

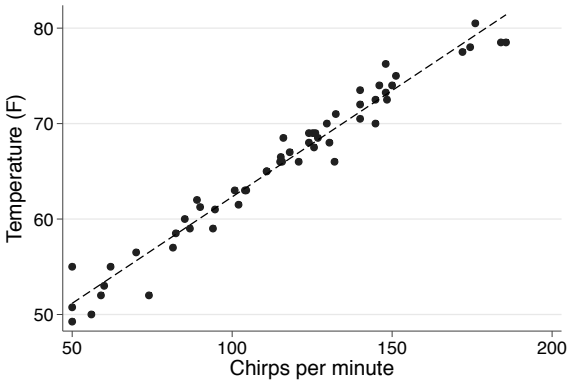
- Create scatterplots to display the relationship between two variables
- Derive the least squares criterion
- Interpret the correlation between two variables
- Using technology, find the least squares regression line to model the relationship between two variables
- Interpret the coefficient of determination and use it evaluate the appropriateness of a linear model
- Interpret the slope and y-intercept of a least-squares regression line
- Interpolate and extrapolate from the regression line
- Given output from a computer program, interpret coefficients in a multiple regression analysis

Outside of a math classroom, it's rare for two variables to have a perfect linear relationship. Consider that *snowy tree cricket* example from the last activity. Do you *really* believe the chirp rate for those crickets increases exactly 12 chirps-per-minute each time the temperature increases by 3 degrees?

The scatterplot on the right shows data collected from crickets in Boulder, Colorado from August to September in 2007.

Each dot on the scatterplot displays the number of times a cricket chirped in a minute along with the temperature during that minute. A total of 55 observations were plotted on the graph.

As you can see, the chirp rates do not all fall exactly on a straight line. Even so, it looks like a linear function does a pretty good job of describing the relationship between temperature and chirp rates.



http://blog.globe.gov/sciblog/2007/10/

Even if we expect two variables to have a linear relationship, our measurement will lead to the data roughly following a linear function. Our goal in this activity is to learn how we can find the line that best fits a dataset (if, in fact, we think those variables should have a linear relationship).

1) We're going to fit linear functions to actual datasets (not scenarios created for math textbooks), so we're going to have to deal with error. Even though our linear functions won't fit our data *perfectly*, they might fit good enough. At some point, we're going to have to decide how to determine if the fit is *good enough*. For now, let's investigate how we might quantify error.

   To do this, we're going to have a contest (which I will explain in class). In this contest, you'll have to fill-in the first two columns of this table:

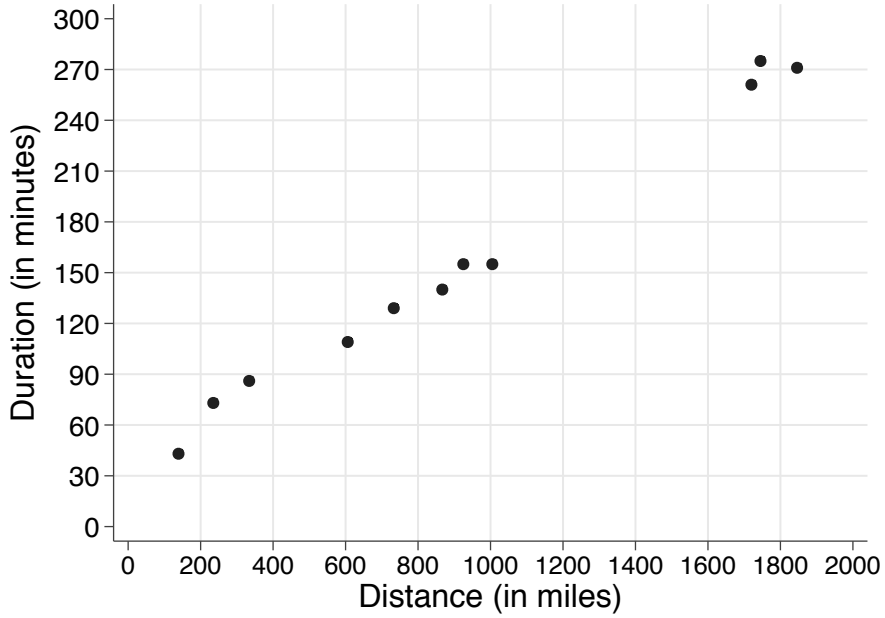
	Individual	Guess	Actual	Error
1				
2				
3				
4				
5				
6				
7				
8				
9				
10				
			Total:	

2) The following table displays information about United Airlines flights out of Chicago (on May 1, 2013):

Fly from Chicago to...	Miles	Duration (in min)	Cost ($)	Prediction	Squared Error
Moline	139	43	513		
Detroit	235	73	100		
Minneapolis	334	86	87		
Atlanta	606	109	152		
New York (LaGuardia)	733	129	195		
Boston	867	140	93		
Houston	925	155	522		
Orlando	1005	155	119		
Seattle	1720	261	201		
Los Angeles	1745	275	214		
San Francisco	1846	271	266		

Note: Information accessed from United Airlines website on 11/17/2012. All flights were one-way flights on 5/1/13.

Below, I've created a scatterplot showing the relationship between distance and the duration of each flight.



Is the duration of a flight a function of the distance?

What assumptions are you making?

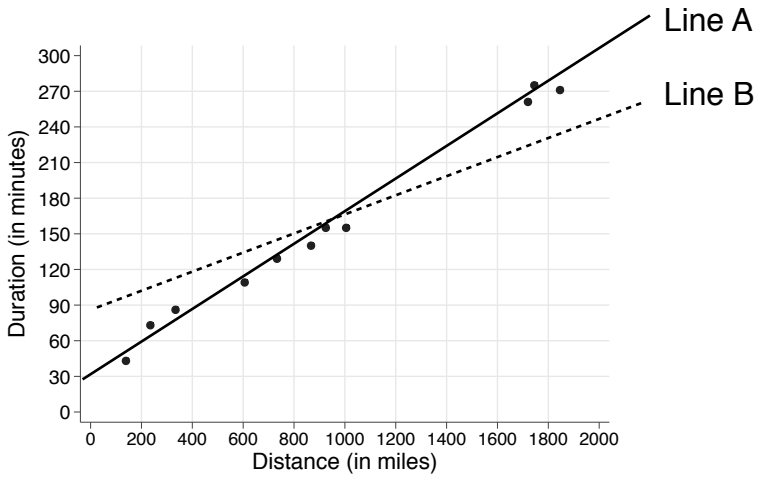
Would you say a linear function appropriately models the relationship between distance and duration?

Sketch the linear function you think best fits this data, estimate its slope and y-intercept, and write out the formula.

3) According to your formula, how long should it take to fly 139 miles (from Chicago to Moline)?

How about 1745 miles (from Chicago to Los Angeles)? How accurate is your formula?

4) We'll learn how to use our calculators to find the best fitting line. Before that, let's see if we can agree on what it means for a line to fit the data "best." Let's start with a simple example. Suppose two students came up with the following lines:

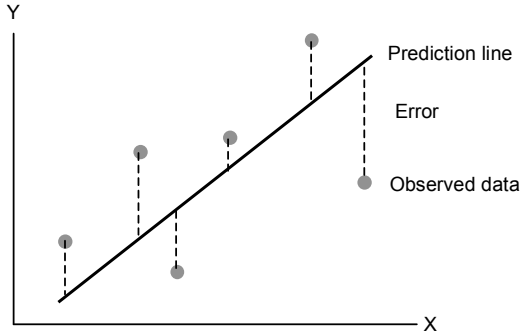


Obviously, it looks like Line A fits the data better. How do we know this? How can we quantify which line fits best?

One way to determine which line best fits best is to use the *least squares criterion*.

The best fitting line **minimizes the squared vertical distances from each observation to the line**.

If you take a statistics class, you'll learn more about this concept.



Sketch the squared errors for Line A and Line B on the scatterplot for question #3.

5) Let's go ahead and use our calculator to find the best-fitting line for our data and quantify the error.

Before we begin, you'll want to enable a special feature on your calculator (if you have a TI-83 or TI-84):

```
Enter the CATALOG menu (located above the zero button)
Move the cursor down to DIAGNOSTICON (or enter "D" to move to it more quickly)
Hit ENTER twice (you should see DONE displayed on the screen)
```

Now, let's enter our data into our calculators.

```
Enter the STAT menu
We want to enter data, so select EDIT...
Enter the data into the lists L1 and L2, pressing ENTER after each entry
QUIT to leave the data entry screen
```

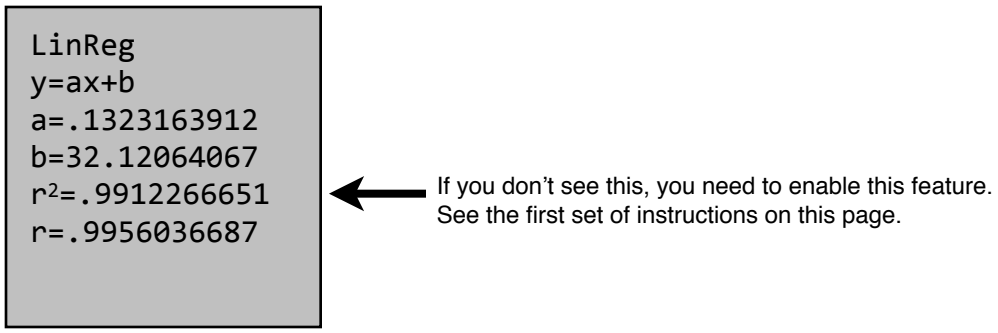
Before we calculate the best-fitting line, we can display a scatterplot.

```
Press [2ND] [Y=] to access the STAT PLOT editor
Press [ENTER] to edit Plot1
Press [ENTER] to turn ON Plot1
Scroll down & highlight the scatterplot graph type (first option in the first row)
Press [ENTER] to select the scatterplot
Make sure XList is set to L1 and Ylist is set to L2.
Press [ZOOM][9] to perform a ZoomStat and display your scatterplot
```

To find the line of best fit, you'll need to...

```
Press STAT to enter the statistics menu
Move right to highlight the CALC menu
Select LINREG(ax+b) to calculate a linear regression
(by default, this treats L1 as the independent var. and L2 as the dependent var.)
Press ENTER twice to calculate the least-squares regression line
```

You should see this screen...



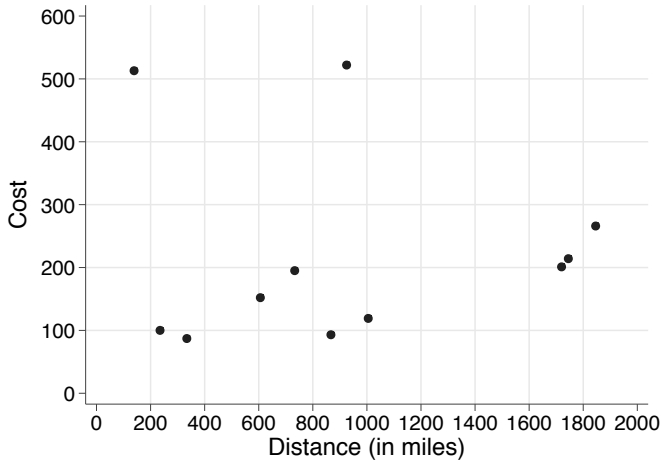
6. Write out the linear function that best fits our distance and duration data. What do the slope and y-intercept represent? What are the domain and range of this function?

7. Use your linear function to "predict" the duration of flights in our table (on the second page of this activity). Calculate the squared error of each of your predictions. What is the total sum of this squared error?

8. Calculate the best-fitting line to model the cost of a flight as a function of its distance. Record the formula along with the values of r and R^2 estimated by your calculator. Explain what the slope and y-intercept represent.

9. Use your linear function to "predict' how much it should cost for a 1005 mile flight to Orlando. How much error is in this prediction?

10. Sketch that best-fitting line on the following scatterplot and comment on how well the line represents the data. Would you say the relationship between the distance of a flight and its cost is linear?



11. As we just demonstrated, just because a line is the best-fitting doesn't mean it actually fits (or describes) the data. Our calculator will always find the best-fitting line, so it's up to us to judge whether we should try to fit a line to a set of data.

    Thankfully, your calculator gives you some additional information we can use to determine how well the best-fitting line actually fits our data.

    The r value reported by your calculator is a correlation coefficient. It gives an index of how close the data fall on a straight line. Your textbook (in section 2.2) gives a good overview of what a correlation coefficient represents.

    In this class, we'll be more interested in R2, the coefficient of determination. We can interpret the coefficient of determination as: the amount of variation in (independent variable) that's explained by (dependent variable).

    So, $R^2 = 0.99$ (that we found in question 4) can be interpreted as: 99% of the variation in flight duration is explained by distance. High values of R2 give us confidence that our model fits the data well.

12. Interpret the R^2 you found when you modeled cost as a function of distance. What other factors could explain the variation in cost?

13. At the very beginning of this activity, we looked at the relationship between temperature and cricket chirp rates. I typed this data into a computer program called *Stata* and had it run a linear regression analysis to find the linear function that models temperature as a function of chirp rates. Here's the output it gave me:

```
      Source |       SS       df       MS              Number of obs =      55
-------------+------------------------------           F(  1,    53) = 1302.52
       Model |  3375.21407     1  3375.21407           Prob > F      =  0.0000
    Residual |  137.338201    53  2.59128681           R-squared     =  0.9609
-------------+------------------------------           Adj R-squared =  0.9602
       Total |  3512.55227    54  65.0472643           Root MSE      =  1.6097

------------------------------------------------------------------------------
       tempf |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      chirps |   .2229494   .0061775    36.09   0.000     .2105589    .2353399
       _cons |   40.02525   .7441376    53.79   0.000      38.5327     41.5178
------------------------------------------------------------------------------
```

Write out the formula for the best-fitting linear function and determine if a linear function is an appropriate model for this data. Interpret the slope and y-intercept.

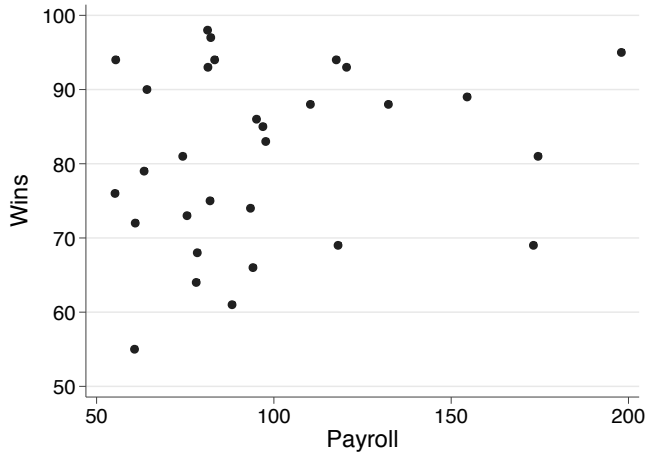
14. The U.S. Census Bureau reports the total amount of money American citizens spend on reading each year. Type this data into your calculator, generate a scatterplot, sketch that scatterplot below, find the best-fitting line to predict reading expenditures as a function of the year, and comment on how well that line describes the relationship between these variables: Source: http://www.census.gov/compendia/statab/2012/tables/12s1232.pdf

YEAR	READING
1985	141
1990	153
1994	165
1995	163
1996	159
1997	164
1998	161
1999	159
2000	146
2001	141
2002	139
2003	127
2004	130
2005	126
2006	117
2007	118
2008	116
2009	110

15. The following table displays the payroll for each Major League Baseball team in 2012 along with the number of games each team won during the regular season. Find the formula for the best-fitting line to model wins as a function of payroll. Interpret the slope and y-intercept. Does a linear function adequately model this data?

Team	Payroll	Wins
New York Yankees	$198.0	95
Philadelphia Phillies	$174.5	81
Boston Red Sox	$173.2	69
Los Angeles Angels	$154.5	89
Detroit Tigers	$132.3	88
Texas Rangers	$120.5	93
Miami Marlins	$118.1	69
San Francisco Giants	$117.6	94
St. Louis Cardinals	$110.3	88
Milwaukee Brewers	$97.7	83
Chicago White Sox	$96.9	85
Los Angeles Dodgers	$95.1	86
Minnesota Twins	$94.1	66
New York Mets	$93.4	74
Chicago Cubs	$88.2	61

Team	Payroll	Wins
Atlanta Braves	$83.3	94
Cincinnati Reds	$82.2	97
Seattle Mariners	$82.0	75
Baltimore Orioles	$81.4	93
Washington Nationals	$81.3	98
Cleveland Indians	$78.4	68
Colorado Rockies	$78.1	64
Toronto Blue Jays	$75.5	73
Arizona Diamondbacks	$74.3	81
Tampa Bay Rays	$64.2	90
Pittsburgh Pirates	$63.4	79
Kansas City Royals	$60.9	72
Houston Astros	$60.7	55
Oakland Athletics	$55.4	94
San Diego Padres	$55.2	76




```
      Source |       SS       df       MS              Number of obs =      30
-------------+------------------------------           F(  1,    28) =    1.12
       Model |  158.558318     1  158.558318           Prob > F      =  0.2994
    Residual |  3971.44168    28  141.837203           R-squared     =  0.0384
-------------+------------------------------           Adj R-squared =  0.0040
       Total |        4130    29  142.413793           Root MSE      =   11.91

------------------------------------------------------------------------------
        wins |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     payroll |   .0634843   .0600436     1.06   0.299    -.0595095     .186478
       _cons |   74.77706   6.274477    11.92   0.000     61.92438    87.62974
------------------------------------------------------------------------------
```

16. Every situation we've looked at so far has modeled a dependent variable as a linear function of a single independent variable. How can we handle situations with two or more independent variables?

- The wind chill is a function of air temperature and air speed: Wind Chill = f(air temperature, air speed)
- The monthly payment on a loan is a function of the amount borrowed, the interest rate, and the length of a loan: Payment = f(amount, interest rate, length)
- We may be able to predict your grade in this course as a function of many independent variables: Grade = f(last math class taken, grade in that class, high school GPA, ACT Math score, age, gender)

We can conduct a multiple regression analysis to determine the function that best fits a data set with multiple independent variables. Given the following scenarios and computer output, write out the best-fitting function, interpret the coefficients, and comment on how well the function models the data.

Scenario A: Some occupations are considered to be more prestigious than others. For example, many would agree that a heart surgeon has a more prestigious occupation than a waitress. We're going to model the prestige of an occupation as a function of several characteristics.

Source: Canada (1971). Census of Canada. Vol. 3, Part 6. Statistics Canada, 19-21.

Data: Title: Name of occupation
Education: Average years of education for occupational incumbents (in 1971)
Income: Average income, in dollars, of incombents (in 1971)
%women: Percentage of incumbents who are women (in 1971)
Prestige: Pineo-Porter Prestige score (from a survey conducted in the mid-1960s)

#	Title	Education	Income	%women	Prestige
1	Physicians	15.96	25308	10.56	87.2
2	University Professors	15.97	12480	19.59	84.6
3	Lawyers	15.77	19263	5.13	82.3
…	…	…	…	…	…
18	Medical Technicians	12.79	5180	76.04	67.5
19	Secondary Teachers	15.08	8034	46.8	66.1
…	…	…	…	…	…
26	Elementary Teachers	13.62	5648	83.78	59.6
…	…	…	…	…	…
99	Bartenders	8.5	3930	15.51	20.2
100	Elevator Operators	7.58	3582	30.08	20.1
101	Janitors	7.11	3472	33.57	17.3
102	Newsboys	9.62	918	7	14.8

```
      Source |       SS       df       MS              Number of obs =     102
-------------+------------------------------           F(  3,    98) =  129.19
       Model |  23861.8558     3  7953.95195           Prob > F      =  0.0000
    Residual |  6033.57026    98  61.5670435           R-squared     =  0.7982
-------------+------------------------------           Adj R-squared =  0.7920
       Total |  29895.4261   101  295.994318           Root MSE      =  7.8465

------------------------------------------------------------------------------
    prestige |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   education |   4.186637   .3887013    10.77   0.000     3.415272    4.958002
      income |   .0013136   .0002778     4.73   0.000     .0007623    .0018648
    percwomn |  -.0089052   .0304071    -0.29   0.770     -.069247    .0514367
       _cons |  -6.794334   3.239089    -2.10   0.039     -13.2222   -.3664676
------------------------------------------------------------------------------
```

Scenario B: If a pregnant woman smokes, does it affect the health of her child? To study this, a researcher collected the following data for 3,978 firstborn children.

Source: Abrevaya, J. (2006). Estimating the effect of smoking on birth outcomes using a matched panel data approach. Journal of Applied Econometrics 21: 489-519.

Data: Birth weight: Weight of child at birth, in grams
Mother Age: Age of mother at time of birth
Child Male?: Was the child male? (1 = yes, 0 = no)
Mother married?: Was the mother married? (1 = yes, 0 = no)
Mother high school?: Did the mother complete high school? (1 = yes, 0 = no)
Mother college?: Did the mother complete college? (1 = yes, 0 = no)
Black: Was the mother African-American? (1 = yes, 0 = no)

Birth Weight (grams)	Mother Age	Mother smoke?	Child male?	Mother married?	Mother High School?	Mother college grad?	Black
2790	16	0	0	0	0	0	1
2693	17	0	0	0	0	0	1
3600	20	0	0	0	0	0	1
2807	22	1	0	1	1	0	0
2948	23	1	0	1	1	0	0
...	...	...	...	...	...	...	...

```
      Source |       SS       df       MS              Number of obs =    3978
-------------+------------------------------          F(  7,  3970) =   47.16
       Model |  78376210.7     7  11196601.5           Prob > F      =  0.0000
    Residual |   942457129  3970  237394.743           R-squared     =  0.0768
-------------+------------------------------          Adj R-squared =  0.0751
       Total |  1.0208e+09  3977   256684.27           Root MSE      =  487.23

------------------------------------------------------------------------------
       birwt |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        mage |   5.870645   1.654518     3.55   0.000     2.626861    9.114429
       smoke |  -241.7262   23.63754   -10.23   0.000    -288.0691   -195.3834
        male |   102.1545    15.4649     6.61   0.000     71.83462    132.4744
     married |   80.82202   28.16412     2.87   0.004     25.60453    136.0395
      hsgrad |   17.57377   19.43135     0.90   0.366    -20.52259    55.67014
    collgrad |   21.36287    20.3324     1.05   0.293    -18.50006    61.22581
       black |  -213.2207   32.46166    -6.57   0.000    -276.8638   -149.5776
       _cons |   3189.924   48.31306    66.03   0.000     3095.203    3284.645
------------------------------------------------------------------------------
```