Week 3: Simple Linear Regression

Marcelo Coca Perraillon

University of Colorado
Anschutz Medical Campus

Health Services Research Methods I
HSMP 7607
2019

Outline

- Putting a structure into exploratory data analysis
- Covariance and correlation
- Simple linear regression
- Parameter estimation
- Regression towards the mean

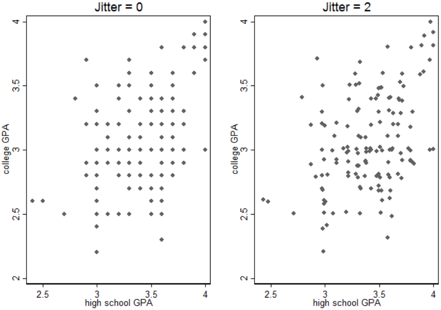
Big picture

- We did exploratory data analysis of two variables X, Y in the first homework
- Now we are going to provide **structure** to the analysis
- We will assume a relationship (i.e. a **functional form**) and estimate parameters that **summarize** that relationship
- We will then **test** hypotheses about the relationship
- For the time being, we will focus on **two continuous** variables

Example data

- We will use data from Wooldridge on grades for a sample 141 college students (see today's do file)

```
use GPA1.DTA, clear
sum colgpa hsgpa
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
      colgpa |        141    3.056738    .3723103        2.2          4
       hsgpa |        141    3.402128    .3199259        2.4          4
```

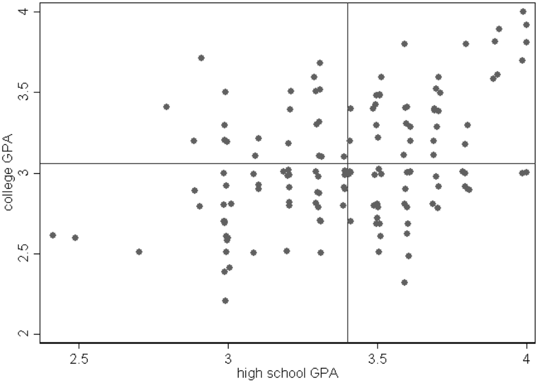


Covariance and correlation

- A simple summary of the relationship between two variables is the **covariance**:
- $COV(X,Y) = E[(X - E(X)(Y - E(Y))] = E(XY) - E(X)E(Y)$
- $COV(X,Y) = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})$
- For each pair x_i, y_i we calculate the product of the deviations of each variable from its mean
- The covariance will be closer to zero if observations are closer to their mean (for one or both variables); it can be positive or negative
- The scale is the product of the scales of $X$and Y (e.g. age* age, grades*age, etc)

```
. corr colgpa hsgpa, c
(obs=141)
             |   colgpa    hsgpa
-------------+------------------
      colgpa |  .138615
       hsgpa |  .049378  .102353
```

Graphical intuition?



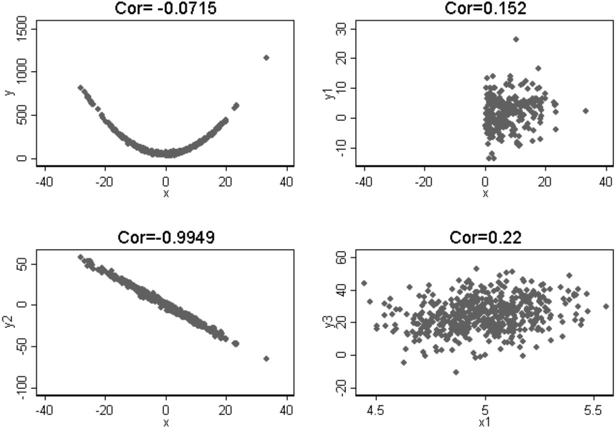
- If $COV(X, Y) > 0$ then positive relationship between x and y
- If $COV(X, Y) < 0$ then negative relationship between x and y

Correlation

- The sign of the covariance is useful but the magnitude is not because it **depends on the unit of measurement**
- The **correlation** (ρ) scales the covariance by the standard deviation of each variable:
- $Cor(X, Y) = \frac{1}{n-1} \sum_{i=1}^{n} \frac{(y_i - \bar{y})(x_i - \bar{x})}{S_y S_x}$
- $Cor(X, Y) = \frac{COV(X,Y)}{S_y S_x}$
- $-1 \leq Cor(X, Y) \geq 1$
- Closer to 1 or -1, stronger relationship
- Grades data:

```
  corr colgpa hsgpa
             |   colgpa    hsgpa
-------------+------------------
      colgpa |   1.0000
       hsgpa |   0.4146   1.0000
```

Examples



- ρ close to 0 **does NOT imply** X and Y are not related
- ρ measures the **linear** relationship between two variables

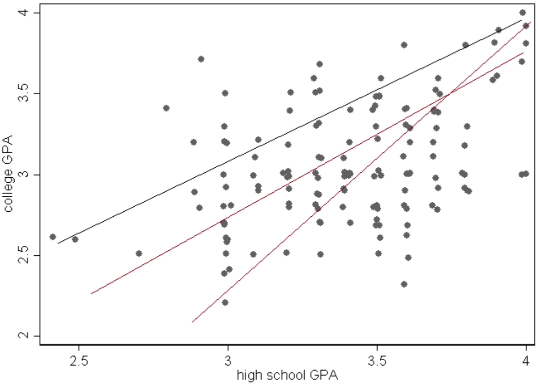
Going beyond the correlation coefficient

- We need more flexibility to understand the relationship between X and Y; the correlation is useful but it is limited to a linear relationship and we can't study changes in Y for changes in X using ρ
- A useful place to start is assuming a more specific functional form:
- $Y = \beta_0 + \beta_1 X + \epsilon$
- The above model is the an example of **simple linear regression** (SLR)
- **Confusion alert**: it's **linear on the parameters** β_i; $Y = \beta_0 + \beta_1 X^2 + \epsilon$ is also a SLR model
- In the college grades example, we have $n = 141$ observations. We could write the model as
- $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, where $i = 1, .., n$. College grades is y and high school grades is x

The role of ϵ

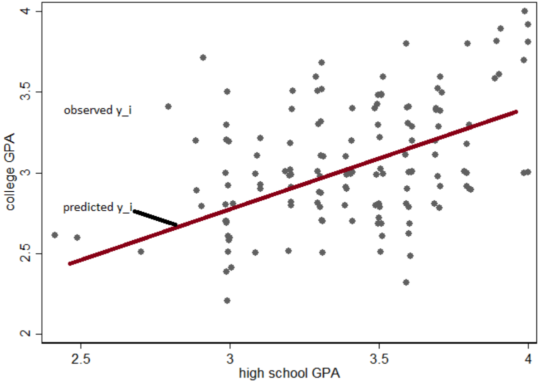
- $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, where $i = 1, .., n$
- We are assuming that X and Y are related as described by the above equation plus an error term ϵ
- In general, we want the error, or the **unexplained part of the model**, to be as small as possible
- How do we find the optimal β_j? One way is to find the values of β_0 and β_1 that are **as close as possible** to **all** the points x_i, y_i
- These values are $\hat{\beta}_0$ and $\hat{\beta}_1$ the prediction is $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$
- This is equivalent to say that we want to make the ϵ as small as possible
- Obviously, the relationship is not going to be perfect so $\epsilon_i \neq 0$ for most observations

Some possible lines (guesses)



- I used a graphic editor to draw some possible lines; I wanted to draw the lines as close as possible to most of the points
- The line is affected by the mass of points and extreme values

A more systematic way



- The error will be the difference $\epsilon = (y_i - \hat{y}_i)$ for each point; we don't want a positive error to cancel out a negative one so we take the square: $\epsilon_i^2 = (y_i - \hat{y}_i)^2$

The method of ordinary least squares (OLS)

- We want to find $\hat{\beta}_i$ that minimizes the sum of all errors:
  $S(\beta_0, \beta_1) = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$
- The solution is fairly simply with calculus. We solve the system of equations:
  $\frac{\partial S(\beta_0,\beta_1)}{\partial \beta_0} = 0$
  $\frac{\partial S(\beta_0,\beta_1)}{\partial \beta_1} = 0$
- The solution is $\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$ and $\beta_0 = \bar{y} - \hat{\beta}_1 \bar{x}$
- To get predicted values, we use $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$
- $S(\beta_0, \beta_1)$ is also denoted by SSE, *sum of squares for error*

Deriving the formulas

- We start with:
  $\frac{\partial SSE}{\partial \beta_0} = \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2 = -2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i) = 0$
- We can them multiply by $-\frac{1}{2}$ and distribute the summation:
  $\sum_{i=1}^{n} y_i - n\beta_0 - \beta_1 \sum_{i=1}^{n} x_i = 0$
- And almost there. Divide by n and solve for β_0: $\hat{\beta_0} = \bar{y} - \beta_1 \bar{x}$
- For β_1, more steps but start with the other first order condition and plug in $\hat{\beta_0}$
  $\frac{\partial SSE}{\partial \beta_1} = -2\sum_{i=1}^{n} x_i(y_i - \beta_0 - \beta_1 x_i) = 0$

Interpreting the formulas

- Does the formula for $\hat{\beta}_1$ look familiar?
  $\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$
- We can multiply by $\frac{1/(n-1)}{1/(n-1)}$ and we get the formulas for the covariance and variance:
  $\hat{\beta}_1 = \frac{COV(Y,X)}{Var(X)}$
- Since $Var(X) > 0$, the sign of $\hat{\beta}_1$ depends on $COV(X, Y)$
- If the X and Y are not correlated, then $\hat{\beta}_1 = 0$
- So you can use a test for $\hat{\beta}_1 = 0$ as a **test for correlation**. But now you have more flexibility and are not constrained to a linear relationship correlation
- For example, you could test if $\gamma_1 = 0$ in $Y = \gamma_0 + \gamma_1 X^2$

Digression

- Not the only target function to minimize. We could also work with the absolute value, as in $|y_i - \hat{y}_i|$. This is called the least absolute errors regression; more robust to extreme values
- **Jargon alert**: *robust* means a lot of things in statistics. Whenever you hear that XYZ method is more robust, ask the following question: robust to what? It could be missing values, correlated errors, functional form...
- A very fashionable type of model in prediction and machine learning is the **ridge regression** (Lasso method, too)
- It minimizes the sum of errors $(y_i - \hat{y}_i)^2$ plus the sum of square betas $\lambda \sum_{i=1}^{j} \beta_j^2$
- It may look odd but we want to also make the betas as small as possible as a way to **select variables** in the model

Grades example

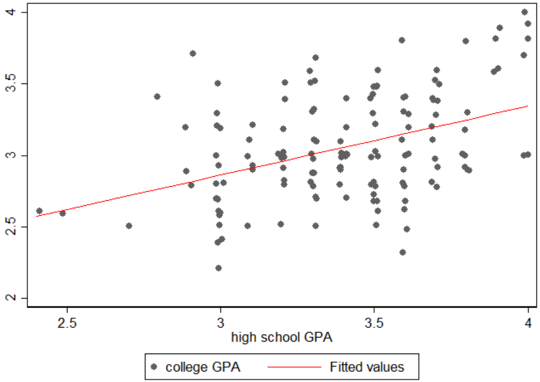
- In Stata, we use the reg command:

```
. reg colgpa hsgpa

      Source |       SS           df       MS      Number of obs   =       141
-------------+----------------------------------   F(1, 139)       =     28.85
       Model |  3.33506006         1  3.33506006   Prob > F        =    0.0000
    Residual |  16.0710394       139  .115618988   R-squared       =    0.1719
-------------+----------------------------------   Adj R-squared   =    0.1659
       Total |  19.4060994       140  .138614996   Root MSE        =    .34003

------------------------------------------------------------------------------
      colgpa |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       hsgpa |   .4824346   .0898258     5.37   0.000      .304833    .6600362
       _cons |   1.415434   .3069376     4.61   0.000     .8085635    2.022304
------------------------------------------------------------------------------
```

- So $\hat{\beta}_0 = 1.415434$ and $\hat{\beta}_1 = .4824346$. A predicted value is $\hat{y}_i = 1.415434 + .4824346(x_i = a)$

Grades example II

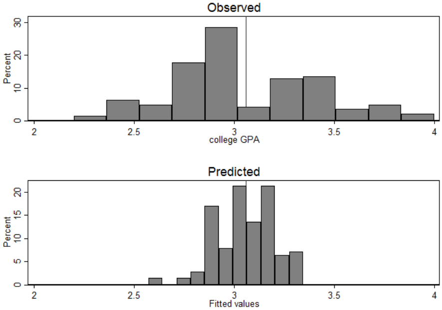
```
gen gpahat =  1.415434 + .4824346*hsgpa
gen gpahat0 = _b[_cons] + _b[hsgpa]*hsgpa
predict gpahat1
* ereturn list
* help reg
scatter colgpa hsgpa, jitter(2) || line gpahat1 hsgpa, color(red) sort ///
   saving(reg1.gph, replace)
```



How do observed and predicted values compare?

```
sum colgpa gpahat1
hist colgpa, percent title("Observed") saving(hisob.gph, replace) xline(3.06)
hist gpahat1, percent title("Predicted") saving(hispred.gph, replace) xline(3.06)
graph combine hisob.gph hispred.gph, col(1) xcommon

   Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
      colgpa |        141    3.056738    .3723103        2.2          4
     gpahat1 |        141    3.056738    .1543433   2.573277   3.345172
```

- Predictions **"regress" towards the mean**:



Regression towards the mean

- Regression towards the mean is an often-misunderstood concept
- In this example, our model is telling us that a student with a high high-school GPA is going to be more like an average college student (i.e. she will regress towards the mean)
- Why is that happening? Look at the data. Is that true in our sample?
- It happens because our prediction is using the information of **everybody** in the sample to make predictions for those with high high-school GPA
- It may also be because it's a *property* of the particular dataset or problem, like in the homework example

Confusion alert and iterated expectations

- From OLS, it is not clear that we are modeling the conditional expectation of Y given X: $E[Y|X]$ but **WE ARE (!!)**
- We are modeling how the mean of Y changes for different values of X
- The **mean of the predictions** from our model **will match the observed mean of Y**
- We can use the law of iterated expectations to go from the conditional to unconditional mean of Y:

```
qui sum hsgpa
di 1.415434 + .4824346*r(mean)
3.0567381
.sum colgpa
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
      colgpa |        141    3.056738    .3723103        2.2          4
```

Another way of writing the model

- When we cover Maximum Likelihood Estimation (MLE), it's going to become super clear that we are indeed modeling a **conditional expectation**
- For the rest of the semester and your career, it would be useful to write the estimated model as $E(\hat{y}_i|x) = \hat{\beta}_0 + \hat{\beta}_1 x$ or $E(\hat{y}_i) = \hat{\beta}_0 + \hat{\beta}_1 x$
- Next class we are going to start interpreting parameters. We will see that $\hat{\beta}_1$ tells you how the expected value/average y changes when x changes
- This is **subtle** but super important. It's not the change in y, it's the change in the average y
- Seeing it this way will make it easier later (trust me)
- To make it a bit more confusing: of course, we can use the model to make a prediction for one person. Say, a college student with a hs gpa of xx will have a college gpa of yy. But that prediction is based on the average of others

Big picture

- We started with a graphical approach to study the relationship of two continuous variables
- We then used the correlation coefficient to measure the **magnitude** and **direction** of the **linear** relationship
- We then considered a more **flexible** approach by assuming a more specific functional form and used the method of least squares to find the best parameters
- We now have a way of **summarizing** the relationship between X, Y
- We **didn't make any assumptions about the distribution** of Y (or X)
- Don't ever forget that we are modeling the conditional expectation (!!)
- Next class we will see other ways of thinking about SLR and causal inference