Dis 3: Simple (Univariate) Linear Regression

1 Overview

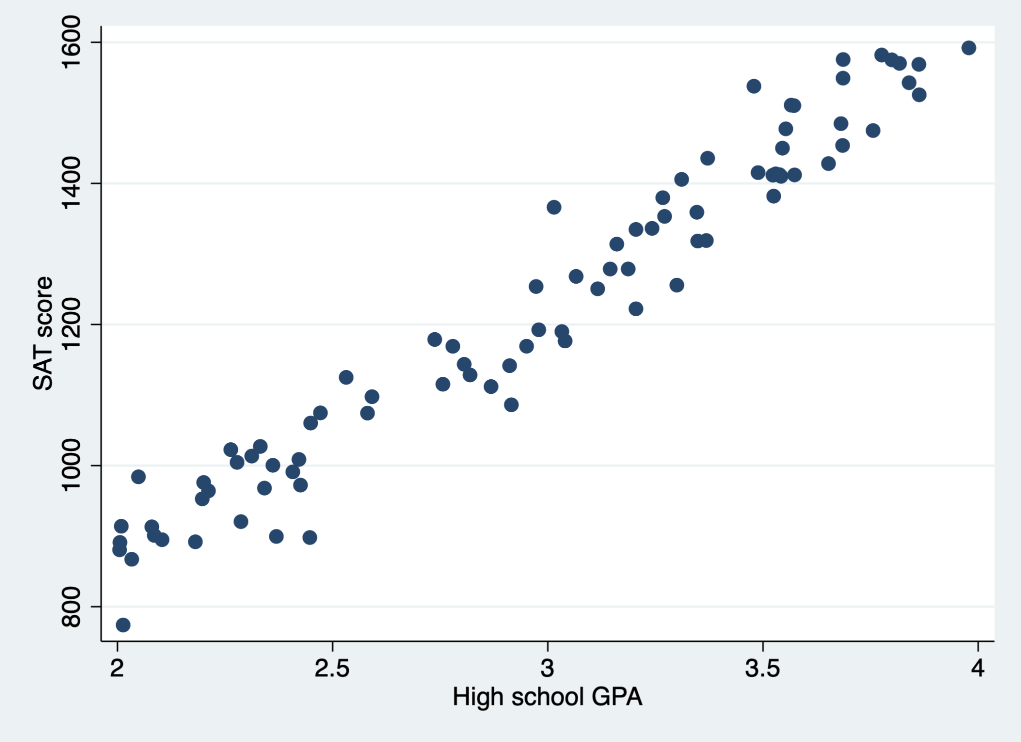
1.1 Why do we need regression? (What are we studying in econometrics?)

- "Econometrics is the application of **statistical methods** to economic data in order to give empirical content to **economic relationships**." – Wikipedia
- Some statistical methods used to describe relationship between variables:
  - **Covariance**
    → Positive vs. negative relationship, but the scale of relationship is ambiguous.
  - **Coefficient of correlation**
    → Positive vs. negative relationship, and the scale is normalized between -1 and 1 (inclusive). But unclear on how change in one variable quantifies to change in the other.
  - **Simple (Univariate) linear regression**
    → Positive vs. negative relatioship. Measurable scale (through statistical significance). Tells us rate of change.

1.2 Simple (Univariate) linear regression

Suppose that we collect a simple random sample of size n: $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, where X is high school GPA, and Y is student's correpsonding SAT test score.

- Looking at the scatter plot between the two variables, they seem to be positively correlated, and they seem to follow a linear relationship:



- Then how do we project a line onto this data?
  - **Model**: $y_i = \beta_0 + \beta_1 x_i + u_i$

Names for y	**Names for x**
Dependent variable	Independent variable
Response variable	Explanatory variable
Regressand	Regressor
Predicted variable	Predictor variable

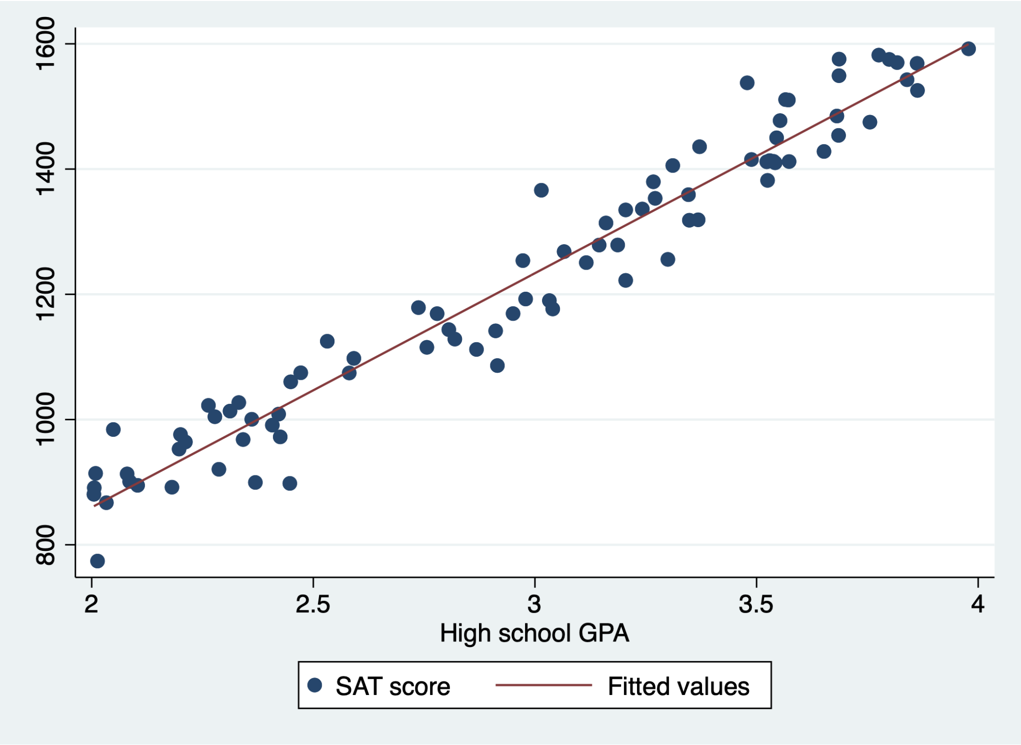
  - **Actual estimates**: $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$
  - **Error vs. Residual**:
    - Error term u_i comes from the model:

$$u_i = y_i - (\beta_0 + \beta_1 x_i)$$

    - Residual $\hat{u}_i$ comes from the estimates:

$$\begin{aligned} \hat{u}_i &= y_i - \hat{y}_i \\ &= y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i) \end{aligned}$$

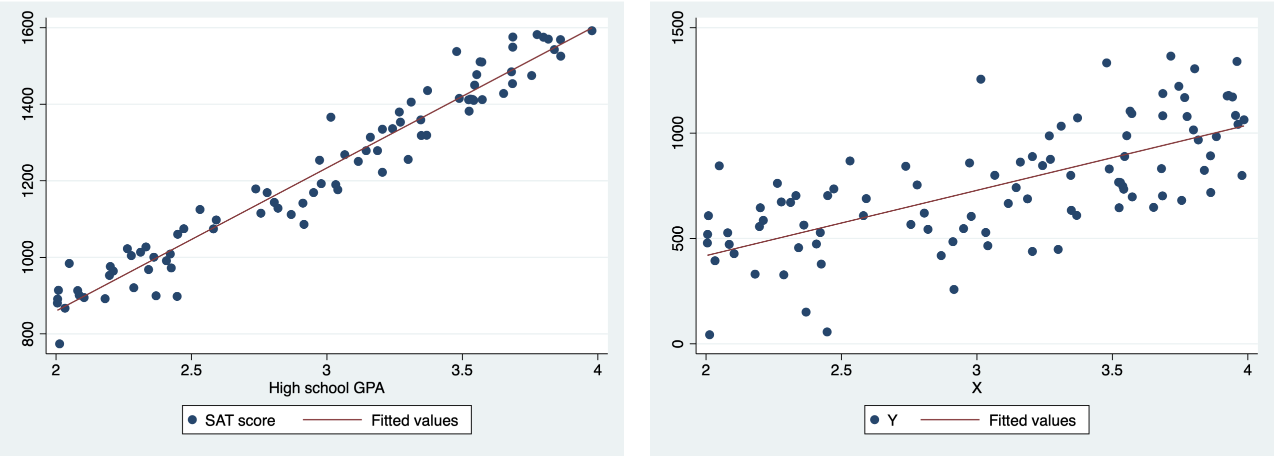
    - The way they are defined tells us that:
      - **Errors are unobservable** (since we don't assume that we know the true parameters β_0 and β_1 – otherwise, what's the point of estimating the model?).
      - **Residuals are observable**, and can be used to tell us something about how well the projected line fits our data.
      - Restrictions on errors are related to how well the model is constructed / how well we can **interpret** the model.
      - Restrictions on residuals are related to how well we can **estimate parameters to best fit** the model.
  - **How do we achieve the estimates?**:
    Choose $\hat{\beta}_0$ and $\hat{\beta}_1$ such that they minimize squared distance from data to the projected line.
    $\Leftrightarrow$ Choose $\hat{\beta}_0$ and $\hat{\beta}_1$ such that they minimize squared residuals.
    Hence why such simple linear regression is also known as **Ordinary Least Squares (OLS)**.
    - **Additional assumptions required for performing the estimation**:
      1. **Zero conditional mean**: $E(u|x) = 0$.
         → error term, conditional on information from x, does NOT further explain y.
      2. **I.I.D. Data**: Data (x_i, y_i) are i.i.d. (independent and identically distributed).
      3. **Large outliers are unlikely**: There doesn't exist some pair of (x_i, y_i) that live in a dramatically different region.
      4. **Homoskedasticity**: $Var(u|x) = \sigma^2$ is a constant.
         → size of the error, conditional on information from x, does NOT vary greatly throughout the data.



Estimates achieved are

- $\hat{\beta}_1 = \frac{\widehat{Cov}(X,Y)}{\widehat{Var}(X)}$ → similar to how coefficient of correlation looks like!
- $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$ → the point $(\bar{x}, \bar{y})$ passes through the fitted line!

– **How to quantify the fitness of the projected line?**



The line on the left certainly seems to fit the data much better than the one on the right. How can we measure this?

⇒ **Use** R^2:

$$R^2 = \frac{SSE}{SST} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

  where SSE stands for **explained sum of squares**, SST stands for **total sum of squares**.

  Some properties of R^2:

  - $0 \leq R^2 \leq 1$
  - The better fit the projected line, the bigger the R^2.
  - **Interpretation of R^2 in words**:
    Say that $R^2 = 0.35$. This implies that about 35% of the sample variation in y is explained by variation in x.

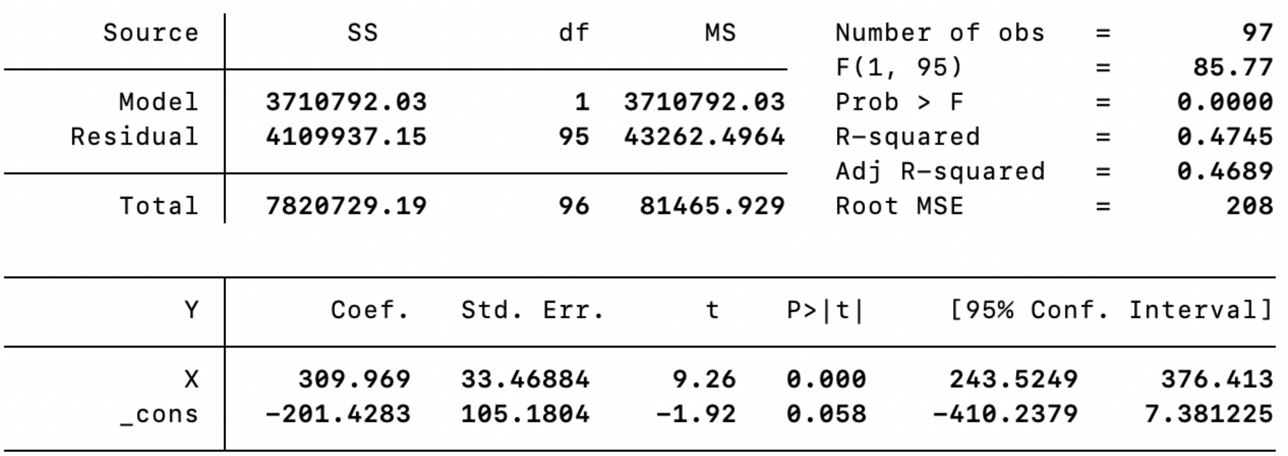
- **How to interpret the fitted line generated by OLS?**
  - General rule of thumb: **correlation instead of causation**
  - Say our fitted line is of the form

$$\hat{y}_i = 10 + 3x_i$$

    One way to correctly interpret this result: One unit increase in x is **associated** with 3 units of increase in y.

- How to perform simple linear regression (OLS) in Stata?

```
reg y x
```



```
      Source |       SS           df       MS      Number of obs   =        97
-------------+----------------------------------   F(1, 95)        =     85.77
       Model |  3710792.03         1  3710792.03   Prob > F        =    0.0000
    Residual |  4109937.15        95  43262.4964   R-squared       =    0.4745
-------------+----------------------------------   Adj R-squared   =    0.4689
       Total |  7820729.19        96   81465.929   Root MSE        =       208

------------------------------------------------------------------------------
           Y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           X |    309.969   33.46884     9.26   0.000     243.5249     376.413
       _cons |  -201.4283   105.1804    -1.92   0.058    -410.2379    7.381225
------------------------------------------------------------------------------
```

- How to produce the scatter plot with the OLS fitted line in Stata?

```
twoway scatter y x || lfit y x
```

2 Problems

1. Load the following dataset from `http://fmwww.bc.edu/ec-p/data/wooldridge/sleep75.dta` into Stata (don't forget to first change your working directory).

   Dataset codebook is available at `http://fmwww.bc.edu/ec-p/data/wooldridge/sleep75.des`

   (a) You are interested in studying the tradeoff between time spent sleeping and working in a sample of individuals aged 23 - 65. Suppose that the true model is:

   $$\text{sleep}_i = \beta_0 + \beta_1 \text{totwrk}_i + u_i$$

   Use Stata to estimate β_0 and β_1 and write down the predicted line.

   (b) Interpret the estimated slope.

   (c) Interpret the estimated constant (or called "intercept").

   (d) Obtain the residuals from the linear regression using the `predict` command, and save it in a variable named `u_hat`

   (e) Now that we have an estimated linear model, if you have a new set of data, then Stata can predict the value of sleep given a new set of totwrk data. To do so,

      i. Load this week's discussion dataset.

      ii. This dataset records some new levels of `totwrk`, but in order for the predicted model to work, we need to first rename the variable so that it's named as `totwrk` (this way, Stata can recognize that this is the independent variable we were regressing onto earlier).

iii. Use the `predict` command to predict the new set of sleep value. Name the variable recording the predicted values as `sleep_new_hat`

iv. Let's verify by hand that `predict` is doing its job. The first new observation of `totwrk` is 2418. Calculate by hand what the predicted `sleep` value would be using our linear model, and then verify that the value generated by the `predict` command is correct.

(f) Does the model fit the data well? Give quantitative evidence.

(g) Visualize how well the model fits the data by plotting the fitted line with the scatter plot of data points.

(h) Is there any source other than `totwrk` that can affect the value of `sleep`?

(i) Given your answer to part (h), do you think the assumption that $E(u|x) = 0$ is valid?

2. Load the `http://fmwww.bc.edu/ec-p/data/wooldridge/sleep75.dta` dataset back into Stata.

   (a) What's the mean level of sleep? And what's the number of observations for sleep?

   (b) What's the standard deviation of sleep?

   (c) Assume that sleep is normally distributed. Construct the 95% confidence interval for the mean level of sleep by hand.

   (d) Continue to assume that sleep is normally distributed. Construct the 95% confidence interval for the mean level of sleep using Stata's `ci means` command.