Dis 3: Simple (Univariate) Linear Regression

1 Overview

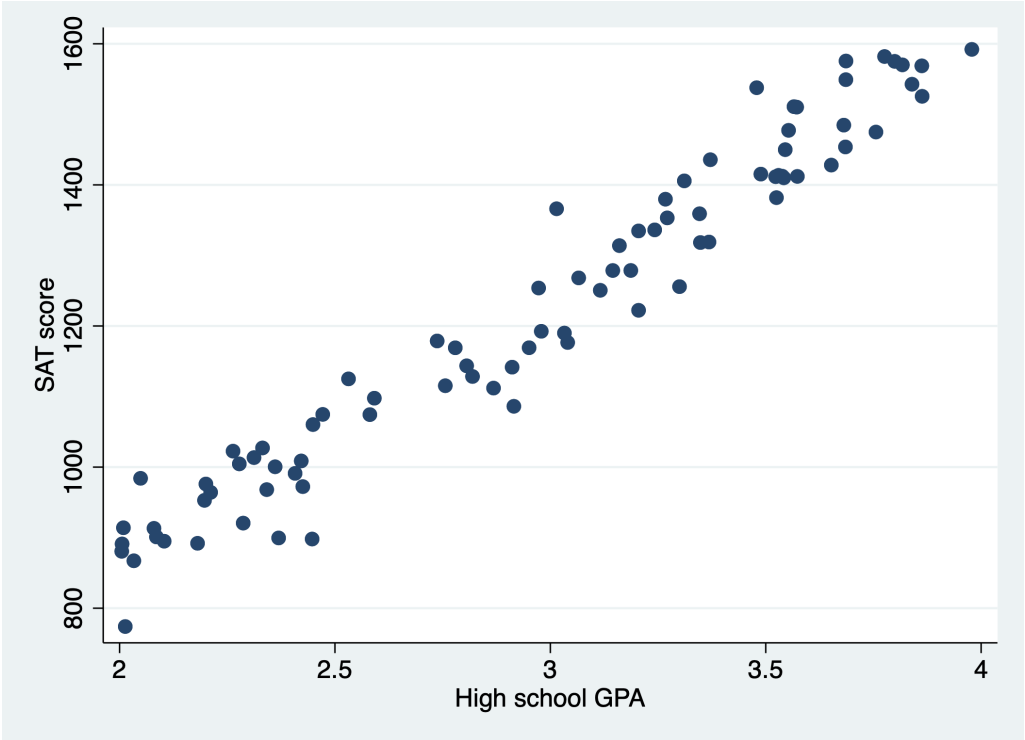
1.1 Why do we need regression? (What are we studying in econometrics?)

- "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**
    $\rightarrow$ Positive vs. negative relationship, but the scale of relationship is ambiguous.
  - **Coefficient of correlation**
    $\rightarrow$ 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**
    $\rightarrow$ 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:

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

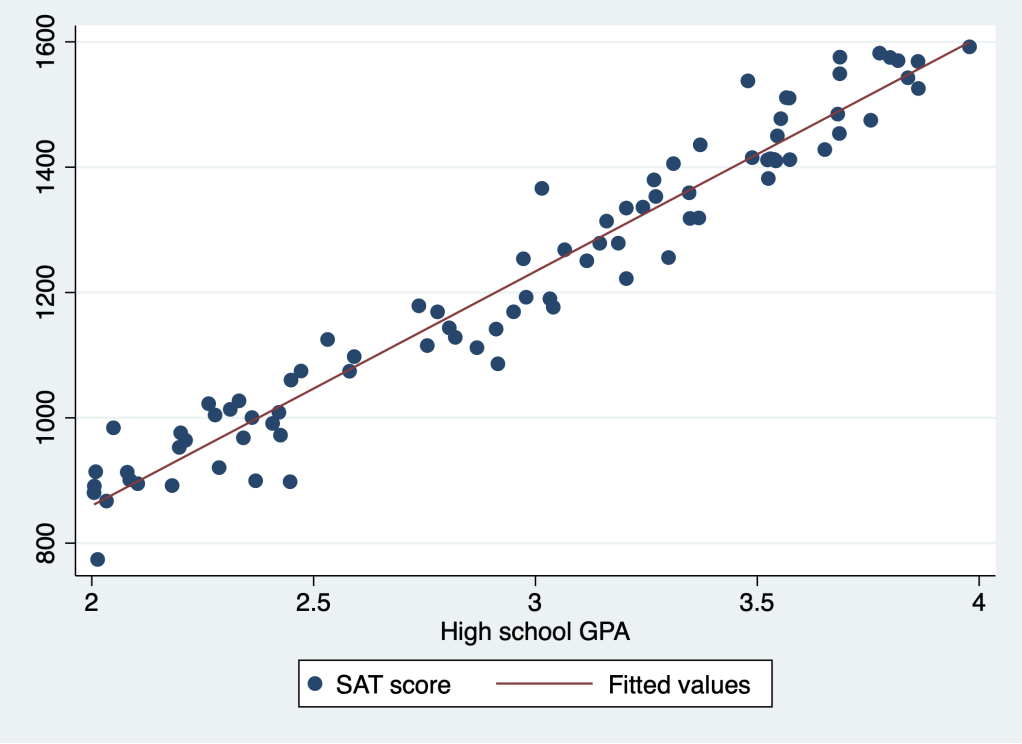
    - The way they are defined tells us that:
      - **Errors are unobservable** (since we don't assume that we know the true parameters β_0 and β_1 – otherwise, what's the point of estimating the model?).
      - **Residuals are observable**, and can be used to tell us something about how well the projected 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)**.

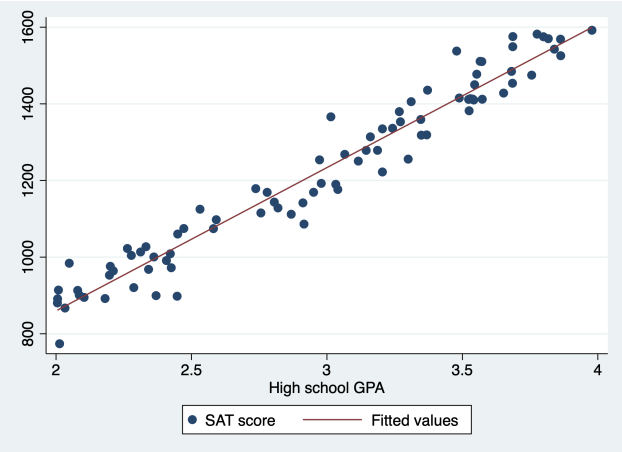
    - **Additional assumptions required for performing the estimation**:
      1. **Zero conditional mean**: $E(u|x) = 0$.
         $\rightarrow$ error term, conditional on information from x, does NOT further explain y.
      2. **I.I.D. Data**: Data (x_i, y_i) are i.i.d. (independent and identically distributed).
      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.
         $\rightarrow$ size of the error, conditional on information from x, does NOT vary greatly throughout the data.

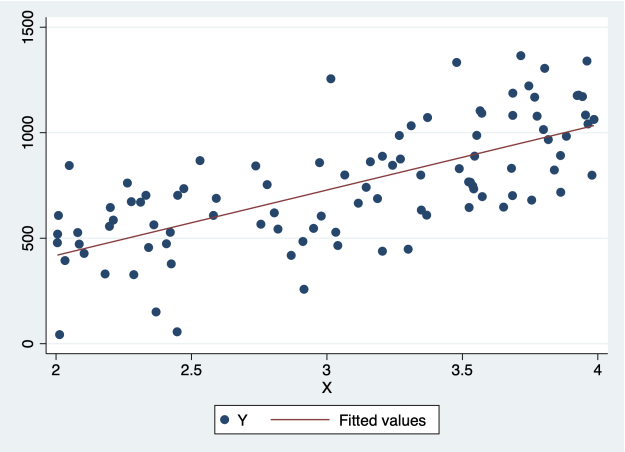


Estimates achieved are

* $\hat{\beta}_1 = \frac{\widehat{Cov}(X,Y)}{\widehat{Var}(X)} \quad \rightarrow$ similar to how coefficient of correlation looks like!
* $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \quad \rightarrow$ 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?

$\Rightarrow$ **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.

   First load the dataset into Stata by running

   ```
   use "http://fmwww.bc.edu/ec-p/data/wooldridge/sleep75.dta", clear
   ```

   To regress sleep on totwrk, run

   ```
   reg sleep totwrk
   ```

   The regression result output is the following:

Source	SS	df	MS	Number of obs	=	706
				F(1, 704)	=	81.09
Model	14381717.2	1	14381717.2	Prob > F	=	0.0000
Residual	124858119	704	177355.282	R-squared	=	0.1033
				Adj R-squared	=	0.1020
Total	139239836	705	197503.313	Root MSE	=	421.14

sleep	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
totwrk	-.1507458	.0167403	-9.00	0.000	-.1836126	-.117879
_cons	3586.377	38.91243	92.17	0.000	3509.979	3662.775

   In the "Coef." column at the bottom table, the second row "_cons" records $\hat{\beta}_0$, and the first row "totwrk" records coefficient on `totwrk`, which is $\hat{\beta}_1$. Thus,

   - β_0 estimate, $\hat{\beta}_0$, is 3586.377
   - β_1 estimate, $\hat{\beta}_1$, is -0.1507458
   - The predicted line is thus

   $$\begin{aligned}\widehat{\text{sleep}}_i &= \hat{\beta}_0 + \hat{\beta}_1 \text{totwrk}_i \\ &= 3586.377 - 0.1507458 \times \text{totwrk}_i\end{aligned}$$

   (b) Interpret the estimated slope.

   The slope estimate ($\hat{\beta}_1$) implies that, every minute worked in a given week is associated with 0.1507458 minute less sleep at night per week.

   Notice that this interpretation is about correlation instead of causation: the linear regression model only tells us how the two variables are correlated. To say something like "every minute

worked in a given week *causes* 0.1507458 minute less sleep at night per week", we have to argue that the amount of sleep you get solely depends on the amount of work you do in a given week, and that's a different set of exercise.

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

The constant / intercept estimate ($\hat{\beta}_0$) implies that, for someone who works 0 minute in a given week, it is predicted that this person will get 3586.377 minutes of sleep at night per week.

Again, notice that we used language like "predicted" instead of "determined". This is because "predicted" implies correlation, whereas "determined" implies causation.

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

We want to use the option `residuals` in `predict` to obtain these residuals:

```
predict u_hat, residuals
```

(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.
After changing your working directory to where the dataset is located at, run the following command in Stata:

```
use "400_sp21_dis-3_dataset.dta", clear
```

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).
To rename variable `total_work` as `totwrk`, use the `rename` command in Stata:

```
rename total_work totwrk
```

iii. Use the `predict` command to predict the new set of sleep value. Name the variable recording the predicted values as `sleep_new_hat`
`predict` uses the previous linear model run in Stata to generate the predicted $\widehat{\text{sleep}}_i$ values. The command to run in Stata is

```
predict 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.
Our linear model predicts $\widehat{\text{sleep}}_i$ as $\widehat{\text{sleep}}_i = 3586.377 - 0.1507458 \times \text{totwrk}_i$. Now, if our new `totwrk`'s value is 2418, simply plug it into the estimated linear equation will give us predicted `sleep` value:

$$\widehat{\text{sleep}}_i = 3586.377 - 0.1507458 \times 2418 \approx 3221.8736556$$

which is basically what Stata's `predict` command generated for such level of `totwrk`.

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

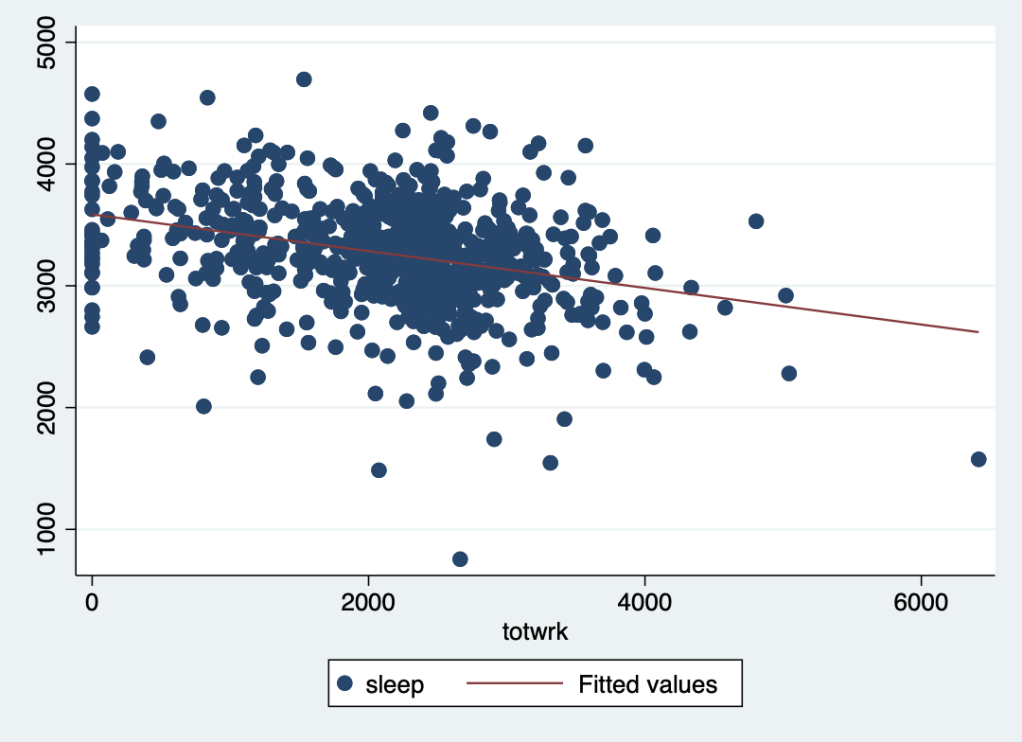
No, the linear model seems to have a poor fit on the data. This is because the R^2 is only 0.1033 for the model proposed, which can be interpreted as only 10.33% of the sample variation in sleep is explained by totwrk.

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

Running the following command in Stata:

```
twoway scatter sleep totwrk || lfit sleep totwrk
```

The graph generated should look like this:



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

Absolutely. I would imagine things like mood of the day, noise level at night, stress level, time of first meeting tomorrow all could potentially affect the amount of sleep a person gets.

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

Most likely not. Conditional on information from X, there likely are other things that can affect level of sleep, which means that the conditional expectation for our error term is not 0 – there are things in the uncaptured error term that could help us explain the level of sleep.

This gives rise to future topics that we'll explore (adding more variables to the regression for control, or use some special variable called instruments to address the problem).

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?
   Running the following command in Stata should give us answer to both questions:

   ```
   mean sleep
   ```

   Stata's result output looks like the following:

   ```
   Mean estimation                   Number of obs    =       706
   ```

	Mean	Std. Err.	[95% Conf.	Interval]
sleep	3266.356	16.72572	3233.517	3299.194

   Here, "Number of obs" at the upper right corner tells us that there are 706 observations of sleep in the dataset, and the "Mean" column tells us that the mean of `sleep` is 3266.356.

   (b) What's the standard deviation of sleep?
   Running the following command in Stata gives us that the standard deviation of sleep variable is 444.4134:

   ```
   tabstat sleep, statistics(sd)
   ```

   (c) Assume that sleep is normally distributed. Construct the 95% confidence interval for the mean level of sleep by hand.
   With sleep being normally distributed, the formula to construct the 95% confidence interval for the mean level of sleep is the following:

$$[\overline{\text{sleep}} - 1.96 * se(\overline{\text{sleep}}), \quad \overline{\text{sleep}} + 1.96 * se(\overline{\text{sleep}})]$$

   Here, $\overline{\text{sleep}} = 3266.356$ from our answer in part (a). The only unknown variable is $se(\overline{\text{sleep}})$, standard error of sleep at its mean. To calculate $se(\overline{\text{sleep}})$:

$$se(\overline{\text{sleep}}) = \sqrt{Var(\overline{\text{sleep}})} = \sqrt{Var\left(\frac{1}{n}\sum_{i=1}^{n}\text{sleep}_i\right)}$$

$$= \sqrt{\frac{1}{n^2}Var\left(\sum_{i=1}^{n}\text{sleep}_i\right)}$$

$$= \sqrt{\frac{1}{n^2}\sum_{i=1}^{n}Var\left(\text{sleep}_i\right)} = \sqrt{\frac{1}{n^2}nVar\left(\text{sleep}_i\right)}$$

(Sleep is randomly sampled, so it's i.i.d. (independent and identically distributed))

$$= \sqrt{\frac{Var\left(\text{sleep}_i\right)}{n}} = \sqrt{\frac{[sd(\text{sleep}_i)]^2}{n}}$$

Hence, given that $n = 706$ from part (a), and $sd(\text{sleep}_i) = 444.4134$ from part (b), we can calculate $se(\overline{\text{sleep}})$:

$$se(\overline{\text{sleep}}) = \sqrt{\frac{[sd(\text{sleep}_i)]^2}{n}} = \sqrt{\frac{[444.4134]^2}{706}} = 16.7257$$

Thus, the 95% confidence interval is constructed as

$$\begin{aligned}&[3266.356 - 1.96 \times 16.7257, \quad 3266.356 + 1.96 \times 16.7257] \\ &= [3233.5736, 3299.1383]\end{aligned}$$

Correct way to interpret the confidence interval: The 95% confidence interval estimator contains the true expectation of sleep 95% of the time; using this sample, our estimate of this interval is $[3233.5736, 3299.1383]$.

(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.

Running the following command in Stata:

```
ci means sleep, level(95)
```

Stata's result output looks like the following:

Variable	Obs	Mean	Std. Err.	[95% Conf.	Interval]
sleep	706	3266.356	16.72572	3233.517	3299.194

Here, Stata's confidence interval estimate is $[3233.517, 3299.194]$, which is very close to what we have (they are not exactly the same due to rounding errors).