Dis 13: Simple Linear Regression

Related textbook chapter: 16
Ch 16 handout and solution offered by Dr. Pac can be accessed here: Handout Solution
This handout incorporates reviews with all exercises from the handout given by Dr. Pac.

1 Motivation

- Congratulation! You've made it to the last chapter of Econ 310!
- Throughout this semester, we have learned many useful statistical tools. But the most important tool is **statistical inference** – using statistics and their relevant distributions to draw inference about population parameters.
- Statistical inference is at the heart of your future econometrics class (either Econ 400 or 410).
- You might ask: what are we going to study in econometrics?
  - "Econometrics is the application of **statistical methods** to economic data in order to give empirical content to **economic relationships**." – Wikipedia

    (Put in other words, econometrics asks you to recover true economic relationships from sample data by using statistical methods.)
  - 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 linear regression**

      → Positive vs. negative relatioship. Measurable scale (through statistical inference on slope coefficient). Tells us rate of change from x to y (via the slope coefficient β_1).

2 Simple Linear Regression

- Suppose that high school GPA (x) and student SAT test score (y) follow a linear relationship, which is expressed as the following:

$$y = \beta_0 + \beta_1 x + \varepsilon$$

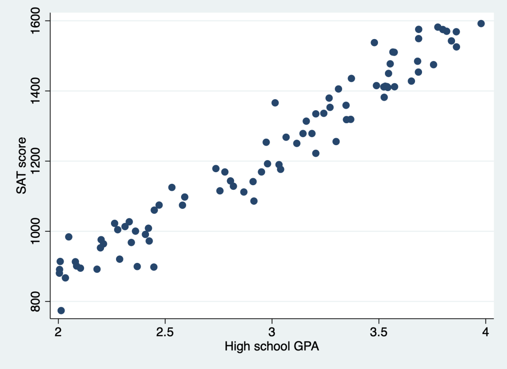
  Notice that the above describes the **true** relationship between y and x. Some terminologies:
  - y: dependent variable (in this case, the student SAT score)
  - x: independent variable (in this case, the student high school GPA)
  - β_0: true intercept
  - β_1: true slope
  - ε: error term

- With a true linear relationship established, how do we recover the true parameters β_0 and β_1?

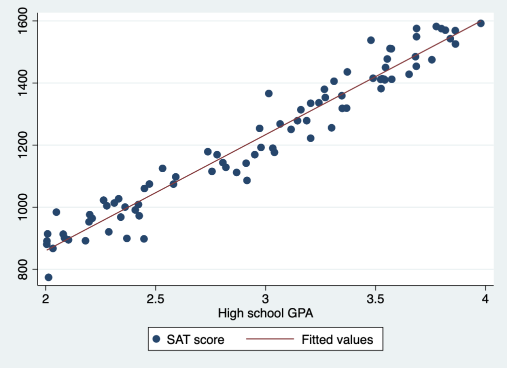
  $\Rightarrow$ use sample data to estimate them!

2.1 How to estimate the linear line?

- Let's say that we collect a simple random sample of size n: $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$. Again, x is high school GPA, and y is the student's correpsonding SAT test score.
- The scatter plot created for the sample data looks like the following:



- How to fit a linear line in our sample data in order to estimate β_0 and β_1?



  Solution: minimize the **sum of squared distance** between each (x_i, y_i) data point and the fitted line.

- Denote the estimated fitted line as $\hat{y}$, where

$$\hat{y} = b_0 + b_1 x$$

  - $\hat{y}$: predicted y / fitted y
  - b_0: estimate of intercept
  - b_1: estimate of slope

  The estimates of b_0 and b_1 are obtained by

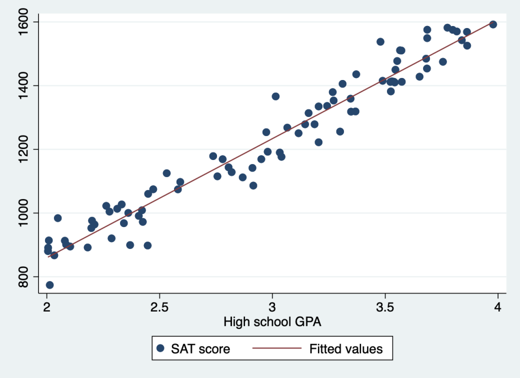
$$\min_{b_0, b_1} \sum_i (y_i - \hat{y}_i)^2 \quad \Leftrightarrow \quad \min_{b_0, b_1} \sum_i (y_i - b_0 - b_1 x_i)^2$$

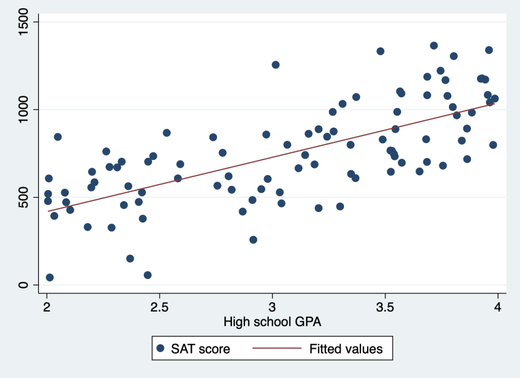
Taking first order conditions to solve the minimization problem yields

$$b_1 = \frac{s_{xy}}{s_x^2} \qquad\qquad b_0 = \bar{y} - b_1\bar{x}$$

2.2 How to evaluate level of fitness for the estimated linear line?

- In the scatter plot that we just looked at, the estimated linear line seems to fit the data pretty well.
- However, say that you've gathered another set of data on high school GPA (x) and SAT score (y), but this new dataset is a bit more noisy. The scatter plot of the new dataset looks like the one on the right:





- How can we tell the above two cases apart?
  - The estimated line fits the data on the left better than the data on the right.
  - Use some goodness of fit measure to differentiate the level of fitness for the estimated line!
- Goodness of fit measure: **coefficient of determination**, or R^2

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

where

  - SSE = sum of squares for error $= \sum_{i=1}^n (y_i - \hat{y}_i)^2$
  - SST = total sum of squares $= \sum_{i=1}^n (y_i - \bar{y})^2$
- 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:

    Say that $R^2 = 0.35$, then about 35% of the sample variation in y is explained by variation in x.

2.3 How to interpret the linear line?

- Continue with the example of student high school GPA (x) and SAT score (y). Say that the fitted line is estimated to be $\hat{y} = 120 + 400x$

- How do we interpret $b_1 = 400$?
  - b_1 gives us an estimate on the rate of change from x onto y.
  - **For this example**: For every 1.0 point increase in high school GPA, their predicted SAT score increases by 400.
- How do we interpret $b_0 = 120$?
  - b_0 tells us what predicted y would be if $x = 0$.
  - **For this example**: If a student has high school GPA = 0, then their predicted SAT score is 120.
  - Keep in mind, if your dataset doesn't contain actual data points near where $x = 0$, then it might not make sense for you to interpret b_0.

    For example, it is virtually impossible for us to observe a student in our data where their GPA is 0.0. So saying that someone with 0.0 GPA has predicted SAT score equals to 120 (a) doesn't make sense, and (b) doesn't really offer any meaningful information.

2.4 How to perform hypothesis testing on the slope coefficient?

- Since the estimate of β_1 (i.e. b_1) gives us an estimate on the rate of change from x onto y, it is often informative for one to perform hypothesis testing on the true slope coefficient (often to check whether β_1 is positive / negative / equal to 0).
- If one goes down the test statistic & rejection region route:

$$\text{test statistic} = \frac{b_1 - \beta_{1,H_0}}{s_\varepsilon / \sqrt{(n-1)s_x^2}} \sim t_{n-2}$$

  where s_ε is the standard error of the regression:

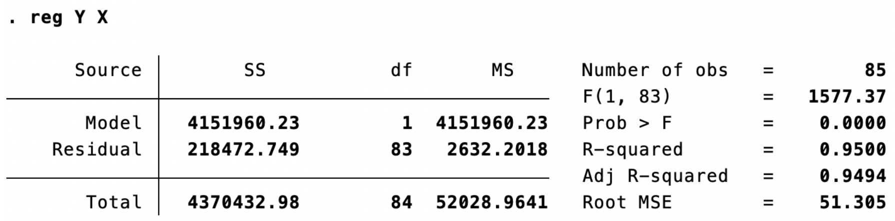
$$s_\varepsilon = \text{Root MSE (Root Mean Square Error)} = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2}}$$

- If one goes down the confidence interval route, the confidence interval of $(1-\alpha)$ confidence level is

$$\left[b_1 - t_{\alpha/2} \times \frac{s_\varepsilon}{\sqrt{(n-1)s_x^2}} \quad , \quad b_1 + t_{\alpha/2} \times \frac{s_\varepsilon}{\sqrt{(n-1)s_x^2}} \right]$$

Side note on s_ε (Root MSE):

- s_ε measures the standard deviation of regression:
  - If the linear line (i.e. the linear model) fits the data well, then s_ε should be relatively small.
  - If the linear line fits the data poorly, then s_ε should be relatively large.
- s_ε can be found from Stata regression output:



```
. reg Y X

      Source |       SS           df       MS      Number of obs   =        85
-------------+----------------------------------   F(1, 83)        =   1577.37
       Model |  4151960.23         1  4151960.23   Prob > F        =    0.0000
    Residual |  218472.749        83   2632.2018   R-squared       =    0.9500
-------------+----------------------------------   Adj R-squared   =    0.9494
       Total |  4370432.98        84  52028.9641   Root MSE        =    51.305
```

In the upper left corner, the *SS* column, *Residual* row records the value of SSE, which in this specific example is 218472.749.

The *Number of obs = 85* on the right hand side tells us that $n = 85$. This allows us to calculate s_ε by hand:

$$s_\varepsilon = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2}} = \sqrt{\frac{218472.749}{85-2}} = 51.305$$

Alternatively, looking at the right side, Root MSE $= 51.305$ directly tells us the value of s_ε.

3 Exercises

1. Attempting to analyze the relationship between sales (y) and advertising (x), the owner of a furniture store recorded monthly sales (in hundreds of thousands of dollars) and his monthly advertising budget (in thousands of dollars) and for a random sample of 12 days. You do not have the raw data, but you do have the following statistics: $\bar{x} = 5$, $\bar{y} = 10.8$, $s_x = 5.6$, $s_y = 20.2$, $s_{xy} = 56.7$, $s_\varepsilon = 18.3$.

   (a) Calculate and interpret the least squares regression coefficients.

   The OLS estimate of the slope is:

   $$b_1 = \frac{s_{xy}}{s_x^2} = \frac{56.7}{(5.6)^2} = 1.808$$

   Based on the estimated slope, we can obtain the estimated intercept:

   $$b_0 = \bar{y} - b_1\bar{x} = 10.8 - 1.808 \times 5 = 1.76$$

   Thus, the estimated linear regression line is:

   $$\begin{aligned}\hat{y} &= b_0 + b_1 x \\ &= 1.76 + 1.808x\end{aligned}$$

   Interpretation of b_0: If no money is spent on advertising, then predicted monthly sales are 1.76 hundred thousand dollars.

   Interpretation of b_1: For every 1 thousand dollars spent on advertising, predicted monthly sales increase by 1.808 hundred thousand dollars.

   (b) Using a 5% significance level, estimate a confidence interval for β_1.

The test statistic related to β_1 is:

$$\text{test statistic} = \frac{b_1 - \beta_1}{s_\varepsilon / \sqrt{(n-1)s_x^2}} \sim t_{n-2}$$

Given that we have 5% significance level, the confidence level is 95%. Using a t-distribution with degree of freedom $= n - 2 = 12 - 2 = 10$, we can find the critical value $t_{\alpha/2} = 2.228$. This means that $P(\text{test statistic} > 2.228) = \alpha/2 = 0.05/2 = 0.025$. The confidence interval bounds are:

$$b_1 \mp t_{\alpha/2} \times \frac{s_\varepsilon}{\sqrt{(n-1)s_x^2}}$$

Plug in $b_1 = 1.808$, $t_{\alpha/2} = 2.228$, $s_\varepsilon = 18.3$, $n = 12$, and $s_x^2 = (5.6)^2$, we find the estimated 95% confidence interval to be $[-0.387, 4.003]$.

(c) Using a 5% significance level, test whether the slope coefficient is greater than zero.

i. Hypotheses:

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 > 0$$

ii. Find test statistic and its distribution:

$$\text{test statistic} = \frac{b_1 - \beta_{1,H_0}}{s_\varepsilon / \sqrt{(n-1)s_x^2}} \sim t_{n-2}$$

Since $n = 12$, t-distribution's degree of freedom is $n - 2 = 10$.

iii. Construct rejection region:

Since our alternative hypothesis is $H_1 : \beta_1 > 0$, one would reject the null and accept the alternative hypothesis only when test statistic is in the far right side of the distribution.

To determine how big the right-side tail needs to be, notice that the significance level is set to be 5%, so we need the critical value from a t-distribution with 10 degrees of freedom, where the right-tail size is 5%. Looking up the t-distribution table, the associated critical value is 1.812. This yields the rejection region:

$$\text{Rejection region: test statistic} > 1.812$$

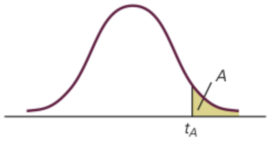
Calculating the value of test statistic using the given sample:

$$\text{test statistic} = \frac{b_1 - \beta_{1,H_0}}{s_\varepsilon / \sqrt{(n-1)s_x^2}} = \frac{1.808 - 0}{18.3 / \sqrt{(12-1) \times 5.6^2}} = 1.835$$

Since test statistic $= 1.835 > 1.812$, the test statistic is in the rejection region. Thus, we can reject the null at 5% significance level, and conclude that predicted monthly sales increase as the monthly advertising budget increases.

Probability table for a t-distribution

TABLE 4
Critical Values of the Student *t* Distribution



Degrees of Freedom	$t_{.100}$	$t_{.050}$	$t_{.025}$	$t_{.010}$	$t_{.005}$
1	3.078	6.314	12.706	31.821	63.657
2	1.886	2.920	4.303	6.965	9.925
3	1.638	2.353	3.182	4.541	5.841
4	1.533	2.132	2.776	3.747	4.604
5	1.476	2.015	2.571	3.365	4.032
6	1.440	1.943	2.447	3.143	3.707
7	1.415	1.895	2.365	2.998	3.499
8	1.397	1.860	2.306	2.896	3.355
9	1.383	1.833	2.262	2.821	3.250
10	1.372	1.812	2.228	2.764	3.169
11	1.363	1.796	2.201	2.718	3.106
12	1.356	1.782	2.179	2.681	3.055
13	1.350	1.771	2.160	2.650	3.012
14	1.345	1.761	2.145	2.624	2.977
15	1.341	1.753	2.131	2.602	2.947
16	1.337	1.746	2.120	2.583	2.921
17	1.333	1.740	2.110	2.567	2.898
18	1.330	1.734	2.101	2.552	2.878
19	1.328	1.729	2.093	2.539	2.861
20	1.325	1.725	2.086	2.528	2.845
21	1.323	1.721	2.080	2.518	2.831
22	1.321	1.717	2.074	2.508	2.819
23	1.319	1.714	2.069	2.500	2.807
24	1.318	1.711	2.064	2.492	2.797
25	1.316	1.708	2.060	2.485	2.787
26	1.315	1.706	2.056	2.479	2.779
27	1.314	1.703	2.052	2.473	2.771
28	1.313	1.701	2.048	2.467	2.763
29	1.311	1.699	2.045	2.462	2.756
30	1.310	1.697	2.042	2.457	2.750
35	1.306	1.690	2.030	2.438	2.724
40	1.303	1.684	2.021	2.423	2.704
45	1.301	1.679	2.014	2.412	2.690
50	1.299	1.676	2.009	2.403	2.678
55	1.297	1.673	2.004	2.396	2.668
60	1.296	1.671	2.000	2.390	2.660
65	1.295	1.669	1.997	2.385	2.654
70	1.294	1.667	1.994	2.381	2.648
75	1.293	1.665	1.992	2.377	2.643
80	1.292	1.664	1.990	2.374	2.639
85	1.292	1.663	1.988	2.371	2.635
90	1.291	1.662	1.987	2.368	2.632
95	1.291	1.661	1.985	2.366	2.629
100	1.290	1.660	1.984	2.364	2.626
110	1.289	1.659	1.982	2.361	2.621
120	1.289	1.658	1.980	2.358	2.617
130	1.288	1.657	1.978	2.355	2.614
140	1.288	1.656	1.977	2.353	2.611
150	1.287	1.655	1.976	2.351	2.609
160	1.287	1.654	1.975	2.350	2.607
170	1.287	1.654	1.974	2.348	2.605
180	1.286	1.653	1.973	2.347	2.603
190	1.286	1.653	1.973	2.346	2.602
200	1.286	1.653	1.972	2.345	2.601
∞	1.282	1.645	1.960	2.326	2.576