Stat 200: Introduction to Statistical Inference **Winter 2023**

Lecture 13: Simple linear regression

Lecturer: Dominik Rothenhäusler *March 1*

Note: *These lecture notes were written by Art Owen. If you like the material, he gets the credit! These notes have not been subjected to the usual scrutiny reserved for formal publications. They are meant as a memory aid for students who took stat 200 at Stanford University. They may be distributed outside this class only with the permission of Art Owen. Also, Stanford University holds the copyright.*

Abstract

These notes are mnemonics about what was covered in class. They don't replace being present or reading the book. Reading ahead in the book is very effective.

13.1 Context

The setup for linear regression is that we have variables X and Y. We want to study how Y is related to X. In simple linear regression X is a real valued variable and so is Y. This lecture is just an introduction to regression. There are several whole courses devoted to it on our campus. They can involve more than one X variable, polynomials in X, or many more elaborate machine learning type problems, models and algorithms. For this course we just look at Y and one X. Those later courses involve more linear algebra and computation than this course assumes. As will be explained later, we will consider the Y values to be random and the X values to be non-random (fixed), so we have x and Y.

13.2 One line

We could predict Y by a line like $\beta_0 + \beta_1 x$. If we have (x_i, Y_i) data then a good line would minimize

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 x_i)^2.$$

One could reasonably choose instead to minimize

$$\frac{1}{n}\sum_{i=1}^{n}|Y_i - \beta_0 - \beta_1 x_i|,$$

but this choice is more difficult computationally and theoretically, so we focus on the first one, called ***least squares***. For future data values where we get some x and want to predict Y we might want to minimize a population quantity like

$$\mathbb{E}((Y - \beta_0 - \beta_1 x)^2).$$

Notice that we are minimizing the vertical separation between a Y_i and it's corresponding $\beta_0 + \beta_1 x_i$ when Y is plotted versus x. This makes sense if we think of predicting an unknown Y from a known x.

© Stanford University 2023

13.3 Least squares estimators

For a given (β_0, β_1) the sum of squared errors is

$$S = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_i)^2.$$

We will assume that $\sum_i (x_i - \bar{x})^2 > 0$. If that would not be true then all the x_i would be equal and there would be no way to get a good Y versus x slope. Now

$$\frac{\partial}{\partial \beta_0} S = -2 \sum_i (Y_i - \beta_0 - \beta_1 x_i), \quad \text{and}$$
$$\frac{\partial}{\partial \beta_1} S = -2 \sum_i x_i (Y_i - \beta_0 - \beta_1 x_i).$$

We set these both to zero to get estimates

$$0 = -2 \sum_i (Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i), \quad \text{and}$$
$$0 = -2 \sum_i x_i (Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i).$$

Rearranging the first one and dividing by n yields

$$\bar{Y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}.$$

This means that the point $(\bar{x}, \bar{Y})$ lies on the least squares line. We can replace $\hat{\beta}_0$ by $\bar{Y} - \hat{\beta}_1 \bar{x}$ in the second equation, and after some algebra we get

$$\hat{\beta}_1 = \frac{\overline{xY} - \bar{x}\bar{Y}}{\overline{x^2} - \bar{x}^2}$$

for

$$\overline{xY} = \frac{1}{n} \sum_i x_i Y_i \quad \text{and} \quad \overline{x^2} = \frac{1}{n} \sum_i x_i^2.$$

It is handy to introduce variables

$$s_{xy} = \frac{1}{n} \sum_i (x_i - \bar{x})(Y_i - \bar{Y}), \quad s_{xx} = \frac{1}{n} \sum_i (x_i - \bar{x})^2, \quad \text{and} \quad s_{yy} = \frac{1}{n} \sum_i (Y_i - \bar{Y})^2.$$

Now

$$\hat{\beta}_1 = \frac{s_{xy}}{s_{xx}}$$

is a sample covariance between x_i and Y_i. Notice that the denominator was assumed to be positive, so $\hat{\beta}_1$ is well defined.

Because $\sum_i (x_i - \bar{x}) = \sum_i (Y_i - \bar{Y}) = 0$ we can write

$$s_{xy} = \frac{1}{n} \sum_i (x_i - \bar{x}) Y_i = \frac{1}{n} \sum_i x_i (Y_i - \bar{Y}), \quad \text{and} \quad s_{xx} = \frac{1}{n} \sum_i (x_i - \bar{x}) x_i.$$

Having these choices lets us get some simpler derivations. For instance

$$\hat{\beta}_1 = \frac{1}{n} \sum_{i=1}^{n} \frac{(x_i - \bar{x}) Y_i}{s_{xx}}$$

and $\hat{\beta}_0$ has a similar formula.

13.4 Regression model

We now have least squares estimates $\hat{\beta}_0$ and $\hat{\beta}_1$. We could ask if they're good. For that we would have to define good. Let's define good as coming out close to true values β_0 and β_1. Close can mean small error but now we have to define what true values would mean here. Consider the model

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

with $\mathbb{E}(\varepsilon_i) = 0$ and $\mathrm{Var}(\varepsilon_i) = \sigma^2$. Later we will also consider $\varepsilon_i \sim N(0, \sigma^2)$ (independently).

Our (x, Y) pairs have Y values scattered above and below the line $\beta_0 + \beta_1 x$, which we take to be the true line. We use a notion of the true line because we want to consider future as yet unobserved data. If all we want is a short summary of the line fitting our past data then $\hat{\beta}_0 + \hat{\beta}_1 x$ is it.

13.5 Fixed or random x_i

A very slippery point is whether to treat the x values as fixed numbers or as random variables. Sometimes the x_i are fixed on purpose. We set the oven to $x_i = 400$ degrees and see how much the cake rises. Other times x and Y are jointly observed and the x is every bit as random as Y. For instance if we relate systolic blood pressure to diastolic blood pressure they're both equally random. Tomorrow we might want to fit DBP as a function of SBP.

Regression is much easier to study for fixed x so we usually assume that. Even when they were random. Here is one reason. Suppose that $X \sim f_X(x; \theta)$ for some parameter θ. Now suppose that given $X = x$ the distribution of Y is $f_{Y|X}(y \mid x; \beta_0, \beta_1, \sigma)$ (such as $N(\beta_0 + \beta_1 x, \sigma^2)$. The likelihood function for all of our data is

$$\prod_{i=1}^{n} f_X(x_i; \theta) \times \prod_{i=1}^{n} f_{Y|X}(y_i \mid x_i; \beta_0, \beta_1, \sigma).$$

The first factor has nothing to do with β_0 and β_1 (or σ). If we maximize the likelihood, the first factor doesn't come into it. If we take likelihood ratios, the first factor cancels. We might as well work with

$$\prod_{i=1}^{n} f_{Y|X}(y_i \mid x_i; \beta_0, \beta_1, \sigma).$$

This is the likelihood from $Y_i \mid X_i = x_i$. We work in a setting where the random X_i have come out equal to x_i. Our probabilities are conditional on $X_i = x_i$. Conditionally on $X_i = x_i$, the predictors are not random any more.

You can imagine that (X_i, Y_i) pairs are generated somewhere and on one day you're given $X_1, \ldots, X_n$ independent draws from f_X. At that point the data have not told you anything about β_0 or β_1. The next day $Y_i \mid X_i = x_i$ are revealed to you. That is informative about β_0 and β_1 using $f_{Y|X}(Y_i \mid X_i = x_i; \beta_0, \beta_1, \sigma)$.

The easier analysis is with X fixed, so that is the one we'll do.

13.6 Expected value

The slope coefficient β_1 is ordinarily much more interesting and imporant than the intercept. If $\beta_1 > 0$ then it means that $\mathbb{E}(Y_i)$ increases with x_i if $\beta_1 < 0$ then $\mathbb{E}(Y_i)$ decreases with x_i and $\beta_1 = 0$ means that $\mathbb{E}(Y_i)$ is the same for all x_i.

We saw in class that

$$\begin{aligned}
\mathbb{E}(\hat{\beta}_1) &= \mathbb{E}\Big(\frac{1}{n}\sum_{i=1}^{n}\frac{X_i-\bar{X}}{s_{xx}}Y_i\Big), \quad \text{easier than using } Y_i-\bar{Y}\\
&= \sum_{i=1}^{n}\frac{x_i-\bar{x}}{ns_{xx}}\mathbb{E}(Y_i), \quad \text{only } Y_i \text{ is random}\\
&= \sum_{i=1}^{n}\frac{x_i-\bar{x}}{ns_{xx}}(\beta_0+\beta_1 x_i), \quad \text{also } \mathbb{E}(\varepsilon_i)=0\\
&= \sum_{i=1}^{n}\frac{x_i-\bar{x}}{ns_{xx}}\beta_1 x_i, \quad \text{because } \sum_i (x_i-\bar{x})=0\\
&= \beta_1\sum_{i=1}^{n}\frac{x_i-\bar{x}}{ns_{xx}}(x_i-\bar{x}), \quad \text{because } \sum_i (x_i-\bar{x}) \text{ is still } 0\\
&= \beta_1\frac{s_{xx}}{s_{xx}}, \quad \text{because } s_{xx}=\sum_i (x_i-\bar{x})^2/n\\
&= \beta_1.
\end{aligned}$$

Similarly you can find that $\mathbb{E}(\hat{\beta}_0)=\beta_0$. Also for any particular value of x, call it x_*, we get $\mathbb{E}(\hat{\beta}_0+\hat{\beta}_1 x_*) = \beta_0+\beta_1 x_*$.

All parameter estimates are unbiased.

13.7 Variance

The variance of $\hat{\beta}_1$ is quite a bit easier to work out when only Y_i are random. For instance, the Y values are all in the numerator. We get (in very slow motion)

$$\begin{aligned}
\mathrm{Var}(\hat{\beta}_1) &= \mathrm{Var}\Big(\frac{1}{n}\frac{\sum_i (x_i-\bar{x})(Y_i-\bar{Y})}{s_{xx}}\Big)\\
&= \mathrm{Var}\Big(\frac{1}{n}\frac{\sum_i (x_i-\bar{x})Y_i}{s_{xx}}\Big), \quad \text{something is } \textbf{\textit{still}}\ 0\\
&= \frac{1}{n^2}\sum_i\Big(\frac{x_i-\bar{x}}{s_{xx}}\Big)^2\mathrm{Var}(Y_i), \quad Y_i \text{ are independent}\\
&= \frac{1}{n^2}\sum_i\Big(\frac{x_i-\bar{x}}{s_{xx}}\Big)^2\mathrm{Var}(\beta_0+\beta_1 x_i+\varepsilon_i), \quad \text{plug in}\\
&= \frac{1}{n^2}\sum_i\Big(\frac{x_i-\bar{x}}{s_{xx}}\Big)^2\mathrm{Var}(\varepsilon_i), \quad x_i \text{ not random}\\
&= \frac{\sigma^2}{n^2}\sum_i\Big(\frac{x_i-\bar{x}}{s_{xx}}\Big)^2, \quad \mathrm{Var}(\varepsilon_i)=\sigma^2\\
&= \frac{\sigma^2}{n^2}\frac{ns_{xx}}{s_{xx}^2}\\
&= \frac{\sigma^2}{ns_{xx}}.
\end{aligned}$$

	What you assume	What you get
1)	$\mathbb{E}(\varepsilon_i) = 0$	$\mathbb{E}(\hat{\beta}_j) = \beta_j$, $j = 0, 1$
2)	Independent ε_i with $\mathrm{Var}(\varepsilon_i) = \sigma^2$	$\mathrm{Var}(\hat{\beta}_1) = \sigma^2/(ns_{xx})$
3)	1), 2) and $\varepsilon_i \sim N(0, \sigma^2)$ and $n \geqslant 3$	$(\hat{\beta}_1 - \beta_1)/(s/\sqrt{ns_{xx}}) \sim t_{(n-2)}$

Table 13.1: Three regression conclusions and the assumptions behind them.

So bigger n is better. Get a bigger sample size if you can. Smaller σ is better. Sometimes you cannot do anything about that but sometimes you can get smaller σ from more expensive equipment or some other improvement.

The most interesting one is that bigger s_{xx} is better. The more spread out the x_i are the better we can estimate the slope β_1. When you're picking the x_i, if you can spread them out more, then it is more informative. There will always be tradeoffs. You cannot simply send $x_1 \to -\infty$ and $x_2 \to \infty$ in real settings.

By the same token smaller s_{xx} is worse. This can mean that in a situation where Y really does vary with x you could get a very bad estimate of β_1 maybe even the wrong sign if the x_i you had did not vary much. Ironically taking care to keep an important x variable in a tight range can remove your ability to measure the effect it has on Y. In class we considered examples like entrance tests to an academic program.

13.8 Normal distribution

Now suppose that $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ where $\varepsilon_i \sim N(0, \sigma^2)$. For known σ we can make tests and confidence intervals using

$$\frac{\hat{\beta}_1 - \beta_1}{\sigma/\sqrt{ns_{xx}}} \sim N(0, 1).$$

We reject $H_0 : \beta_1 = \beta_1^*$ if $|\hat{\beta}_1 - \beta_1^*| > \sigma Z^{1-\alpha/2}/\sqrt{ns_{xx}}$ with the most popular hypothesized value being $\beta_1^* = 0$ (i.e., no effect of x on $\mathbb{E}(Y)$).

In the more realistic setting of unknown σ, so long as $n \geqslant 3$ we can estimate σ^2 by

$$s^2 = \frac{1}{n-2}\sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2.$$

This has $(n-2)s^2 \sim \chi^2_{(n-2)}\sigma^2$ independent of the pair $(\hat{\beta}_0, \hat{\beta}_1)$. As a result

$$\frac{\hat{\beta}_1 - \beta_1}{s/\sqrt{ns_{xx}}} \sim t_{(n-2)}$$

and the confidence interval for β_1 is

$$\hat{\beta}_1 \pm t^{1-\alpha/2}_{(n-2)}\frac{s}{\sqrt{ns_{xx}}}.$$

It is $n-2$ degrees of freedom because we have fit 2 parameters to the n data points.

13.9 Causality

Suppose that $\beta_1 > 0$. That doesn't mean that increasing x_i is going to increase Y_i. The relationship might or might not be causal. This is true even if you have big data. Even if you really want a causal relationship.

Even if you used much more sophisticated methods than linear regression. Even if your boss or advisor or others want a causal relationship. Even if you ***don't*** want a causal relationship. Even if you "controled for" some other variables. To judge whether it is causal or not involves reasoning outside of whatever is present in the set of (x_i, Y_i) values. You can't establish causality from the scatterplot alone or a regression output alone. There is every possibility that you will work with people who want a causal claim from regression output.

Experimentation on x (setting it by a randomization) can establish causal relationships. George Box said "to find out what happens to a system when you interfere with it, you have to inferfere with it". Outside of that, one has to think hard about the logic of the situation, bring in other scientific knowlege and maybe make untestable assumptions. Such 'causal inference' from observational data is a very active area of research. Roughly speaking you'll need an assumption about causality in order to get a conclusion about causality. [I remember Judea Pearl saying something like that but I cannot track down the precise quote and it might have been in a seminar.]

It might be useful to visualize the relationship ***as if*** it were causal. That might help to understand it. It might even give you some ideas to test in future experimental data. But a causal interpretation could well be wrong.

Under our model $\mathbb{E}(Y_i) = \beta_0 + \beta_1 x_i$. Suppose we have a different point with a different value of x, say x_j for $j \neq i$. The model says that

$$\mathbb{E}(Y_j) - \mathbb{E}(Y_i) = (\beta_0 + \beta_1 x_j) - (\beta_0 + \beta_1 x_i) = \beta_1(x_j - x_i).$$

This doesn't mean that if you change x_i to x_j for observation i that Y_i will consequently go up by $\beta_1(x_j - x_i)$. It means that if you happen upon an observation with $x = x_i$ and then later happen upon a different observation from this system with $x = x_j$ then you expect the second one to have a larger value of Y by the amount $\beta_1(x_j - x_i)$.

13.10 Regression and correlation

The estimated slope is $\hat{\beta}_1 = s_{xy}/s_{xx}$. If we formed a sample correlation of these variables it would be $r = s_{xy}/\sqrt{s_{xx}s_{yy}}$. So we see $\hat{\beta}_1 = r\sqrt{s_{yy}}/\sqrt{s_{xx}}$.

Let some particular value of x be x_*. If we got that value we might predict Y by $y_* = \hat{\beta}_0 + \hat{\beta}_1 x_*$. Some algebra in Rice shows that

$$\frac{y_* - \bar{Y}}{\sqrt{s_{yy}}} = r\frac{x_* - \bar{x}}{\sqrt{s_{xx}}}.$$

The LHS is how many standard deviations y_* is from the average $\bar{Y}$. The RHS is r times how many standard deviations x_* is from the average $\bar{x}$. Unless the line is a perfect fit $|r| < 1$. Galton compared heights of sons to their fathers. He got $0 < r < 1$. That means that tall fathers had taller than average sons, but the sons were fewer standard deviations above average than the fathers. Heights in those families had ***regressed towards the mean***. This is where the term ***regression*** came from. Similarly, short fathers had shorter than average sons, but the sons were fewer standard deviations below average than the fathers.

13.11 More generally

This example gives you a taste of dependence modeling. Instead of a model $Y_i \sim f_Y(y;\theta)$ we have $Y_i \sim f_Y(y;\theta_i)$ so every obsevation could have its own parameter. Furthermore $\theta_i = g(x_i)$ so that parameter

depends on other variables.

These notes give a frequentist approach in which β_j and σ and x_i are fixed and then ε_i and Y_i and $\hat{\beta}_j$ and s are random.

In a Bayesian approach, β_j and σ get to be random too with prior and posterior distributions. You could run Bayesian analysis with random x_i or with fixed x_i. I don't see Bayesian regression in Rice Ch 14.