52571: Regression and Statistical Models

Spring 2024

Asaf Weinstein

1 What is regression?

In Statistics, *regression* refers to the modeling and analysis of the relationship between a response variable Y and a set of explanatory variables $X_1, \ldots, X_p$. To use the most basic terms, in a regression problem we are trying to learn how Y changes as $X_1, \ldots, X_p$ change. In that sense, regression analysis studies the *dependence* of the response on the explanatory variables. Here are some real-life examples:

Example 1. $Y =$ weight of a person. $X =$ height of a person

Example 2. $Y =$ percent body weight from fat. $X =$ circumference of abdomen.

Example 3. $Y =$ typical (median) house value in a neighborhood of the city of Boston. $X_1 =$ average distance from employment centers, $X_2 =$ number of rooms per house, $X_3 =$ level of air pollution.

In the elementary description above we have intentionally avoided 'causal' terminology: we did not state our aim as studying what change in Y is *caused* by a change in the values of $X_1, \ldots, X_p$, because the tools covered in this course are generally designed to learn only about *association* between $X_1, \ldots, X_p$ and Y, not causation. This caveat should be kept in mind throughout the course.

In the first two examples $p = 1$, as we have a single explanatory variable. In the third example $p = 3$ as we have three explanatory variables. Now let's think about the general description of a regression problem in the context of the first example above, and imagine that X and Y above were to be measured on each male person in the entire population in the United States. We all know that taller people tend to have higher weight. At the same time, if I consider men with height just about $176\,\mathrm{cm}$, e.g., then of course we do not expect all of them to have exactly the same weight. In other words, weight is not fully *determined* by height, and it is certainly possible to find, for example, a $172\,\mathrm{cm}$ tall male whose weight is greater than that of a $176\,\mathrm{cm}$ tall male. What exactly could we be referring to, then, when we say that taller people "tend" to have higher weight? We can imagine that, if for some fixed value of x we considered only men of height $X \in (x - \Delta, x + \Delta)$, for small Δ, then for every x the corresponding weights, the corresponding values of the response,

$$\mathscr{Y}_x := \{Y : X \in (x - \Delta, x + \Delta)\},$$

have some *distribution*. This can be regarded as approximating the *conditional distribution* of Y on $X = x$. When we say that taller people *tend* to have higher weight, it is reasonable to expect that the *average* of values in $\mathscr{Y}_x$ increase with x. We can imagine a function that maps each x to the average of the values in $\mathscr{Y}_x$—the *conditional mean* of Y for $X = x$. This function is called the *general regression function*, and is standardly the main object of interest in regression analysis. As illustrated in Figure 9, this line can be thought of as being formed, approximately, by taking small bins on the X axis, and drawing a horizontal line, that spans the bin width, at the average value of Y in every bin separately. The function $x \mapsto \mathrm{mean}\,(\mathscr{Y}_x)$ describes the

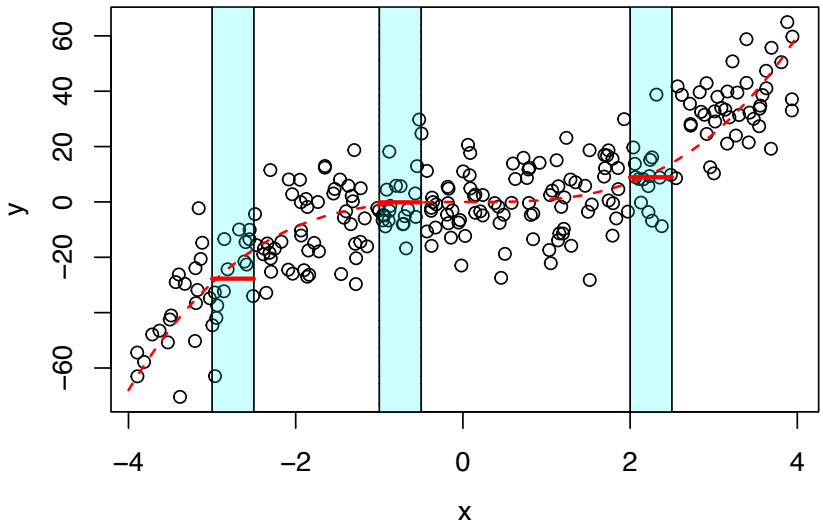

Figure 1: An illustration for general regression. The observed datapoints are represented by the black circles. Light blue rectangles are bins of small width 2Δ on the X-axis, and horizontal red lines mark the average of points (X, Y) with X values in the corresponding bin, i.e. the average of $\mathscr{Y}_x$, for three different values of x. The dashed red line represents a "true" regression line which can be imagined to approximately form if we binned tightly on the entire X range, and consider the (step) function obtained by the horizontal red lines.

population of all males in the USA. In other words, to know this function *exactly* would require access to the weights and heights of all US males. To have access to the records of all subjects in the population is usually impractical. This is where Statistics enters: we will take on the task of *estimating* this function when we only have access to a *sample* from the population.

To estimate a general regression function based on a sample only, could be quite challenging, especially when the number of explanatory variables p is large. In this course we will focus on the case where the population regression line is assumed to be a *linear* function of the explanatory variables, which in the single explanatory case illustrated in Figure 1 the dashed red line would simply be a straight line. Our goal will be to estimate this *linear regression function* and provide statistical inference for it. (Remark: linear regression analysis in fact has a meaning also when the population regression function is not linear, but it is simpler to describe the target in linear regression analysis as we did above, and it is also consistent with what will follow in subsequent chapters).

2 Simple regression and the Least-Squares method

We start with the situation of a single explanatory variable, $p = 1$, in which case we say that we're dealing with *simple* regression. Suppose that we have a dataset of n sample points,

$$(x_i, y_i), \qquad i = 1, ..., n,$$

where x_i records a measurement on some explanatory variable $X \in \mathbb{R}$, and y_i records a measurement of some response variable $Y \in \mathbb{R}$. For illustration, we will use a dataset with $n = 50$ measurements of X =car

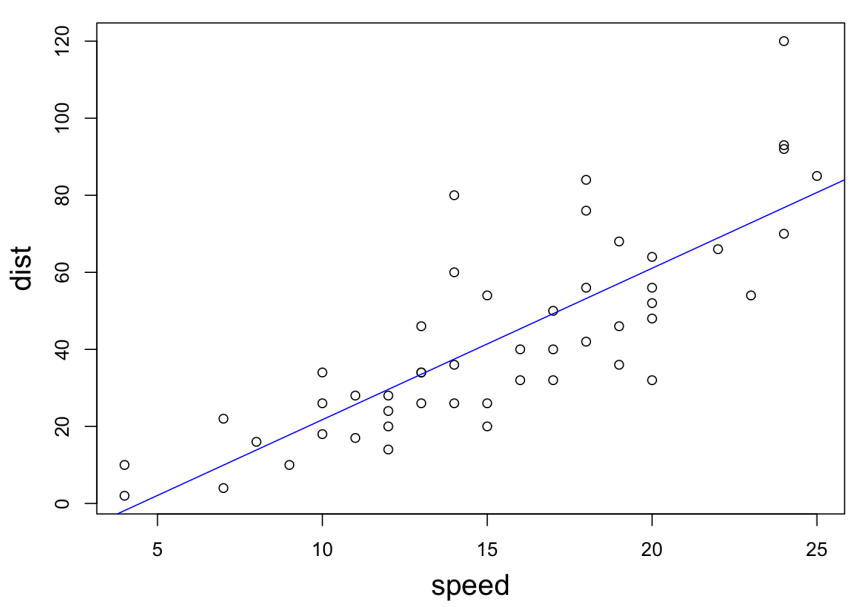

Figure 2: Car data: stopping distance (Y) vs. car speed (X). Blue line is the least squares line.

speed, Y =stopping distance. Figure 2 shows a *scatter plot* of the data, a graph of y_i versus x_i for each of the data points. We want to use this data to learn how X *linearly* explains Y[1]. This means we want to *fit* to (estimate from) the scatter of points a straight line,

$$y = \hat{\beta}_0 + \hat{\beta}_1 x, \tag{1}$$

where the intercept $\hat{\beta}_0$ and slope $\hat{\beta}_1$ are functions of the data $(x_i, y_i),\ i = 1, ..., n$ (hence the "hats" in the notation). For any fitted line (1), we define the *predicted values* (also called *fitted values*) to be

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i. \tag{2}$$

Intuitively, a "good" fit would make the predicted values $\hat{y}_i$ close to the observed values y_i. For any candidate straight line $y = b_0 + b_1 x$, different metrics can be used to measure how well the predicted values agree with the observations. The most standard one is to measure this by the sum of the squared errors between $\hat{y}_i$ and y_i,

$$Q(b_0, b_1) = \sum_{i=1}^{n} \big(y_i - (b_0 + b_1 x_i)\big)^2,$$

and notice that the objective function $Q(b_0, b_1)$ depends on the observations (x_i, y_i). This suggests to obtain the *estimates* $\hat{\beta}_0$ and $\hat{\beta}_1$ by minimizing the objective over (b_0, b_1),

$$(\hat{\beta}_0, \hat{\beta}_1) \ = \ \underset{(b_0, b_1)}{\arg\min}\, Q(b_0, b_1). \tag{3}$$

The criterion entailing the estimation of the slope and intercept by minimizing the sum of squared errors, is called the *least squares* (LS) criterion. The values $(\hat{\beta}_0, \hat{\beta}_1)$ in (3) are the *least squares estimates* of the

[1] for this particular dataset, a straight line indeed seems appropriate for describing the "trend", but what follows does not assume this.

coefficients. The line obtained by substituting $(\hat{\beta}_0, \hat{\beta}_1)$ into (1), appearing in blue in the figure, is called the *least squares line*, or the estimated (linear) *regression line*. From now on, whenever we use the symbols $\hat{\beta}_0, \hat{\beta}_1$ and unless indicated otherwise, we will refer to the least squares coefficients, although we should bear in mind that many other methods (criteria) can be used to fit a line to the data (for example, minimizing the sum of absolute errors instead of the squared errors).

An advantage of using *squared* errors in the objective function, is that it makes $Q(b_0, b_1)$ differentiable, and so for finding a minimum we just look for a point where the partial derivatives vanish:

$$\frac{\partial Q}{\partial b_0} = -2\sum_{i=1}^{n}(y_i - [b_0 + b_1 x_i])\left(\frac{\partial}{\partial b_0}[b_0 + b_1 s_i]\right) = -2\sum_{i=1}^{n}(y_i - [b_0 + b_1 x_i])$$
$$\frac{\partial Q}{\partial b_1} = -2\sum_{i=1}^{n}(y_i - [b_0 + b_1 x_i])\left(\frac{\partial}{\partial b_1}[b_0 + b_1 x_i]\right) = -2\sum_{i=1}^{n}(y_i - [b_0 + b_1 x_i])\, x_i$$

Then, we get

$$\sum_{i=1}^{n}(y_i - [b_0 + b_1 x_i]) = 0$$
$$\sum_{i=1}^{n}(y_i - [b_0 + b_1 x_i])\, x_i = 0$$

which can be rewritten as

$$b_0 + \bar{x} b_1 = \bar{y}$$
$$\bar{x} b_0 + \left(n^{-1}\sum_{i=1}^{n} x_i^2\right) b_1 = n^{-1}\sum_{i=1}^{n} x_i y_i$$

where $\bar{x} := \frac{1}{n}\sum_{i=1}^{n} x_i, \bar{y} := \frac{1}{n}\sum_{i=1}^{n} y_i$. From the first equation we have $b_0 = \bar{y} - \bar{x} b_1$. Substituting this into the second equation and solving for b_1 gives

$$b_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}$$

which can also be written as

$$b_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

Exercise. Prove the equivalence between (1.3) and (1.4). Use the fact that for any vector $z = (\boldsymbol{z}_1, \ldots, z_n)$, we have

$$\begin{aligned}\sum_{i=1}^{n}(z_i - \bar{z})^2 &= \sum_{i=1}^{n}\left(z_i^2 - 2\bar{z} z_i + \bar{z}^2\right) \\ &= \sum_{i=1}^{n} z_i^2 - 2\bar{z}\sum_{i=1}^{n} z_i + \sum_{i=1}^{n}\bar{z}^2 \\ &= \sum_{i=1}^{n} z_i^2 - 2n\bar{z}^2 + n\bar{z}^2 \\ &= \sum_{i=1}^{n} z_i^2 - n\bar{z}^2\end{aligned}$$

(Remark: if we define a random variable Z by $P(Z = z_i) = \frac{1}{n}$, then this is just the familiar relationship

$$\underbrace{V(Z)}_{n^{-1}\sum(z_i-\bar{z})^2} = \underbrace{EZ^2}_{n^{-1}\sum z_i^2} - \underbrace{(EZ)^2}_{\bar{z}^2}).$$

To summarize, the LS solution is given by

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \tag{4}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \tag{5}$$

Thus, the least squares estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ have closed-form expressions as functions of the data. In fact, the LS coefficients are linear in the y_i 's (formally meaning that (1.7) and (1.8), viewed as functions of $\boldsymbol{y} = (y_1, \ldots, y_n)$ only, are affine).

Remark. That the LS coefficients are linear in y is unrelated to the fact that we are attempting to fit a linear function (a straight line) to the data: e.g., if we used absolute deviations then the fitted line is still a straight line, but the coefficients will not be linear functions in $\boldsymbol{y}$.

Recall the definition of the predicted values, $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$. We define the i th residual to be

$$e_i := y_i - \hat{y}_i.$$

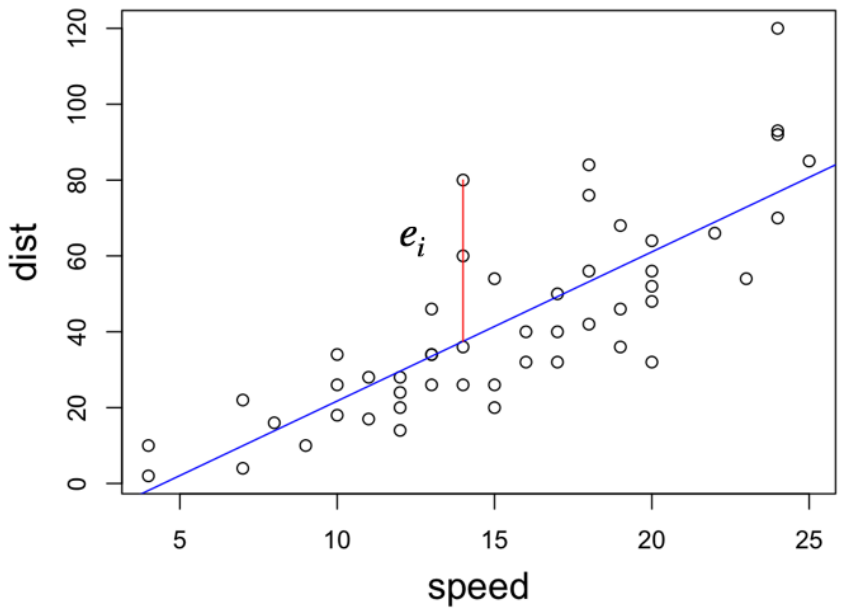

Figure 3: Car data: residuals.

Proposition 1. *For the least squares fit we have the following properties:*

1. $\sum_{i=1}^{n} e_i = 0$
2. $\sum_{i=1}^{n} x_i e_i = 0$
3. $\sum_{i=1}^{n} \hat{y}_i e_i = 0$
4. $n^{-1} \sum_{i=1}^{n} \hat{y}_i = \bar{y}$

Proof. Properties 1 and 2 are true because the LS solution $(\hat{\beta}_0, \hat{\beta}_1)$ satisfies equations (1.3) and (1.4), respectively. Property 3 follows from properties 1 and 2 . For Property 4 , using 1 and the definition of the residuals, we have

$$\frac{1}{n} \sum \hat{y}_i = \frac{1}{n} \sum (y_i - e_i) = \frac{1}{n} \sum y_i = \bar{y}.$$

□

<u>Sums of squares decomposition</u>

$$SST := \sum_{i=1}^{n} (y_i - \bar{y})^2 \qquad \text{Sum of Squares Total}$$

$$SSR := \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \qquad \text{Sum of Squares Regression}$$

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 := \sum_{i=1}^{n} e_i^2 \qquad \text{Sum of Squares Error}$$

SST measures the total ("marginal") variation in y_i; SSR measures the variation explained by the linear regression, i.e., the variation in the y_i which can be accounted for (linearly) by the x_i; and SSE is the leftover variation, i.e, the variation in the y_i which cannot be explained by the x_i.

Note that we have

$$SSR := \sum_i (\hat{y}_i - \bar{y})^2 = \sum_i (\hat{\beta}_0 + \hat{\beta}_1 x_i - \bar{y})^2 = \sum_i [(\hat{\beta}_1 (x_i - \bar{x})]^2 = \hat{\beta}_1^2 \sum_i (x_i - \bar{x})^2 \tag{6}$$

where we used the fact that $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$.

Proposition 2. *SST decomposes as*

$$SST = SSR + SSE.$$

Proof. We have

$$\sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} [(y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})]^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 + 2 \sum_{i=1}^{n} (y_i - \hat{y}_i)(\hat{y}_i - \bar{y})$$

Now,

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) = \sum_{i=1}^{n} e_i (\hat{y}_i - \bar{y}) = \sum_{i=1}^{n} e_i \hat{y}_i - \bar{y} \sum_{i=1}^{n} e_i = 0 - 0,$$

using Properties $1 + 3$ of the LS estimates. Together, we get

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} [(y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})]^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 = SSE + SSR$$

□

Definition 1. For simple linear regression, the coefficient of determination ("R squared") is

$$R^2 := \frac{SSR}{SST}$$

The R^2 value measures the proportion of the total variance of the y_i's explained (linearly!) by x_i's. By Proposition 2, we have

$$R^2 = \frac{SST - SSE}{SST} = 1 - \frac{SSE}{SST}.$$

The R^2 value is connected to Pearson's correlation coefficient. Recall that for two vectors $\boldsymbol{u} = (u_1, \ldots, u_n)$, $\boldsymbol{v} = (v_1, \ldots, v_n)$, the (Pearson) correlation coefficient between $\boldsymbol{u}$ and $\boldsymbol{v}$ is defined as

$$r_{u,v} := \frac{\sum_i (u_i - \bar{u})(v_i - \bar{v})}{\sqrt{\sum_i (u_i - \bar{u})^2}\sqrt{\sum_i (v_i - \bar{v})^2}} \quad \frac{\langle u-\bar{u}, v-\bar{v}\rangle}{|u-\bar{u}||v-\bar{v}|} = \cos\theta$$

Proposition 3. *We have*

1. $\hat{\beta}_1 = \frac{s_y}{s_x} r_{x,y}$, *where* $s_x := \sqrt{\frac{1}{n-1}\sum_i (x_i - \bar{x})}$ *and* $s_y := \sqrt{\frac{1}{n-1}\sum_i (y_i - \bar{y})}$.
2. $r_{x,y} = \operatorname{sign}\left(\hat{\beta}_1\right)\sqrt{R^2}$
3. $r_{y,\hat{y}} = \sqrt{R^2}$

Note: this means that $|r_{x,y}| = |r_{y,\hat{y}}|$, and that $r_{y,\hat{y}} \geq 0$. This relation gives an interpretation for the correlation coefficient between $\boldsymbol{x}$ and $\boldsymbol{y}$: Since $r_{x,y}^2 = R^2$, the square of the Pearson correlation coefficient is the fraction of variation in $\boldsymbol{y}$ that can be linearly explained by $\boldsymbol{x}$ (i.e., by a least-squares regression of $\boldsymbol{y}$ on $\boldsymbol{x}$.

Proof. Using the definition of $r_{x,y}$,

$$\frac{s_y}{s_x} r_{x,y} = \frac{\sqrt{\frac{1}{n-1}\sum_i (y_i - \bar{y})}}{\sqrt{\frac{1}{n-1}\sum_i (x_i - \bar{x})}} \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2}\sqrt{\sum_i (y_i - \bar{y})^2}} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2} = \hat{\beta}_1. \tag{7}$$

Also, using Equation (6), we get

$$R^2 := \frac{SSR}{SST} = \frac{\hat{\beta}_1^2 \sum_i (x_i - \bar{x})^2}{\sum_i (y_i - \bar{y})^2} = \hat{\beta}_1^2 \frac{s_x^2}{s_y^2}.$$

Using the two displays above,

$$r_{x,y} = \hat{\beta}_1 \frac{s_x}{s_y} = \sqrt{\left(\hat{\beta}_1 \frac{s_x}{s_y}\right)^2} \operatorname{sign}\left(\hat{\beta}_1 \frac{s_x}{s_y}\right) = \sqrt{R^2} \cdot \operatorname{sign}(\hat{\beta}_1)$$

The last part #3 will be left as a homework exercise.