A Probabilistic View of Linear Regression

Brian Keng — 2016-05-14 20:43

One thing that I always disliked about introductory material to linear regression is how randomness is explained. The explanations always seemed unintuitive because, as I have frequently seen it, they appear as an after thought rather than the central focus of the model. In this post, I'm going to try to take another approach to building an ordinary linear regression model starting from a probabilistic point of view (which is pretty much just a Bayesian view). After the general idea is established, I'll modify the model a bit and end up with a Poisson regression using the exact same principles showing how generalized linear models aren't any more complicated. Hopefully, this will help explain the "randomness" in linear regression in a more intuitive way.

Background

The basic idea behind a regression is that you want to model the relationship between an outcome variable y (a.k.a dependent variable, endogenous variable, response variable), and a vector of explanatory variables $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ (a.k.a. independent variables, exogenous variables, covariates, features, or input variables). A linear regression relates y to a linear predictor function of $\mathbf{x}$ (how they relate is a bit further down). For a given data point i, the linear function is of the form:

$$f(i) = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} \tag{1}$$

Notice that the function is linear in the parameters $\beta = (\beta_0, \beta_1, \ldots, \beta_n)$, not necessarily in terms of the explanatory variables. It's possible to use a non-linear function of another explanatory variable as an explanatory variable itself, e.g. $f(i) = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3$ is a linear predictor function.

There are usually two main reasons to use a regression model:

- Predicting a future value of y given its corresponding explanatory variables. An example of this is predicting a student's test scores given attributes about the students.
- Quantifying the strength of the relationship of y in terms of its explanatory variables. An example of this is determining how strongly the unit sales of a product varies with its price (i.e. price elasticity).

The simplest form of linear regression model equates the outcome variable with the linear predictor function (ordinary linear regression), adding an error term (ε) to model the noise that appears when fitting the model. The error term is added because the y variable almost never can be exactly determined by $\mathbf{x}$, there is always some noise or uncertainty in the relationship which we want to model.

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \varepsilon_i \tag{2}$$

From this equation, most introductory courses will go into estimating the β parameters using an ordinary least squares approach given a set of $(y_i, \mathbf{x_i})$ pairs, which then can be used for either prediction or quantification of strength of the relationship. Instead of going the traditional route, let's start from the ground up by specifying the probability distribution of y and working our way back up.

Modeling the Outcome as a Normal Distribution

Instead of starting off with both y and $\mathbf{x}$ variables, we'll start by describing the probability distribution of *just* y and *then* introducing the relationship to the explanatory variables.

A Constant Mean Model

First, let's model y as a standard normal distribution with a zero (i.e. known) mean and unit variance. Note this does *not* depend any explanatory variables (no $\mathbf{x}$'s anywhere to be seen):

$$Y \sim N(0, 1) \tag{3}$$

In this model for y, we have nothing to estimate -- all the normal parameter distribution parameters are already set (mean $\mu = 0$, variance $\sigma^2 = 1$). In the language of linear regression, this model would be represented as $y = 0 + \varepsilon$ with no dependence on any $\mathbf{x}$ values and ε being a standard normal distribution. Please note that even though *on average* we expect $y = 0$, we still a expect certain amount of fluctuation or randomness about the 0.

Next, let's make it a little bit more interesting by assuming a fixed *unknown* mean and variance σ^2 corresponding to $y = \mu + \varepsilon$ regression model (here ε is a zero mean and σ^2 variance):

$$Y \sim N(\mu, \sigma^2) \tag{4}$$

We are still not modeling the relationship between y and $\mathbf{x}$ (bear with me here, we'll get there soon). In Equation 4, if we're given a set of $(y_i, \mathbf{x_i})$, we can get an unbiased estimate for μ by just using the mean of all the y_i's (we can also estimate σ^2 but let's keep it simple for now). A more round about (but more insightful) way to find this estimate is to maximize the likelihood function.

Maximizing Likelihood

Consider that we have n points, each of which is drawn in an independent and identically distributed (i.i.d.) way from the normal distribution in Equation 4. For a given, μ, σ^2, the probability of those n points being drawn define the likelihood function, which are just the multiplication of n normal probability density functions (PDF) (because they are independent).

$$\mathcal{L}(\mu|y) = \prod_{i=1}^{n} P_Y(y_i|\mu, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(y_i-\mu)^2}{2\sigma^2}} \tag{5}$$

Once we have a likelihood function, a good estimate of the parameters (i.e. μ, σ^2) is to just find the combination of parameters that maximizes this function for the given data points. In this scenario, the data points are fixed (we have observed n of them with known values) and we are trying to estimate the unknown values for μ (or σ^2). Here we derive the maximum likelihood estimate for μ:

$$\begin{aligned}
\hat{\mu} = \arg\max_{\mu} \mathcal{L}(\mu|y) &= \arg\max_{\mu} \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(y_i-\mu)^2}{2\sigma^2}} \\
&= \arg\max_{\mu} \log\Big(\prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(y_i-\mu)^2}{2\sigma^2}}\Big) \\
&= \arg\max_{\mu} \sum_{i=1}^{n} \log(\frac{1}{\sigma\sqrt{2\pi}}) + \log(e^{-\frac{(y_i-\mu)^2}{2\sigma^2}}) \\
&= \arg\max_{\mu} \sum_{i=1}^{n} \log(e^{-\frac{(y_i-\mu)^2}{2\sigma^2}}) \\
&= \arg\max_{\mu} \sum_{i=1}^{n} -\frac{(y_i-\mu)^2}{2\sigma^2} \\
&= \arg\min_{\mu} \sum_{i=1}^{n} (y_i-\mu)^2
\end{aligned} \tag{6}$$

We use a couple of tricks here. It turns out maximizing the likelihood is the same as maximizing the log-likelihood [1] and it makes the manipulation much easier. Also, we can remove any additive or multiplicative constants where appropriate because they do not affect the maximum likelihood value.

To find the actual value of the optimum point, we can take the partial derivative of Equation 6 with respect to μ and set it to zero:

$$\begin{aligned}
\frac{\partial}{\partial\mu} \log \mathcal{L}(\mu|y) &= 0 \\
\frac{\partial}{\partial\mu} \sum_{i=1}^{n} (y_i-\mu)^2 &= 0 \\
\sum_{i=1}^{n} -2(y_i-\mu) &= 0 \\
n\mu &= \sum_{i=1}^{n} y_i \\
\mu &= \frac{1}{n} \sum_{i=1}^{n} y_i
\end{aligned} \tag{7}$$

Which is precisely the mean of the y values as expected. Even though we knew the answer ahead of time, this work will be useful once we complicate the situation by introducing the explanatory variables.

Finally, the expected value of y is just the expected value of a normal distribution, which is just equal its mean:

$$E(y) = \mu \tag{8}$$

A Couple of Important Ideas

So far we haven't done anything too interesting. We've simply looked at how to estimate a "regression" model $y = \mu + \varepsilon$, which simply relates the outcome variable y to a constant μ. Another way to write this in terms of Equation 2 would be $y = \beta_0 + \varepsilon$, where we just relabel $\mu = \beta_0$.

Before we move on, there are two points that I want to stress that might be easier to appreciate with this extremely simple "regression". First, y is a random variable. Assuming our model represents the data correctly, when we plot a histogram it should bell shaped and centered at μ. This is important to understand because a common misconception with regressions is that y is a deterministic function of the $\mathbf{x}$ (or in this case constant) values. This confusion probably comes about because the error term ε error term is tacked on at the end of Equation 2 reducing its importance. In our constant modeling of y, it would be silly to think of y to be exactly equal to μ -- it's not. Rather, the values of y are normally distributed around μ with μ just being the expected value.

Second, $\mu = \frac{1}{n} \sum_{i=1}^{n} y_i$ (from Equation 7) is a *point estimate*. We don't know its exact value, whatever we estimate will probably not be equal to its "true" value (if such a thing exists). Had we sampled our data points slightly differently, we would get a slightly different estimate of μ. *This all points to the fact that μ is a random variable* [2]. I won't talk too much more about this point since it's a bit outside scope for this post but perhaps I'll discuss it in the future.

Modeling Explanatory Variables

Now that we have an understanding that y is a random variable, let's add in some explanatory variables. We can model the expected value of y as a linear function of p explanatory variables [3] similar to Equation 2:

$$E(y|\mathbf{x}) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p \tag{9}$$

Combining this Equation 8, the mean of y is now just this linear function. Thus, y is a normal variable with mean as a linear function of $\mathbf{x}$ and a fixed standard deviation:

$$y \sim N(\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p, \sigma^2) \tag{10}$$

This notation makes it clear that y is still a random normal variable with an expected value corresponding to the linear function of $\mathbf{x}$. The problem now is trying to find estimates for the p β_i parameters instead of just a single μ value.

Maximizing Likelihood

To get point estimates for the β_i parameters, we can again use a maximum likelihood estimate. Thankfully, the work we did above did not go to waste as the steps are the same up to Equation 6. From there, we can substitute the linear equation from Equation 9 in for μ and try to find the maximum values for the vector of β values:

$$\begin{aligned}\beta &= \arg\min_{\beta} \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p))^2 \\ &= \arg\min_{\beta} \sum_{i=1}^{n} (y_i - \hat{y_i})^2 \end{aligned} \tag{11}$$

We use the notation $\hat{y_i} = E(y|\mathbf{x}, \beta)$ to denote the predicted value (or expected value) of y of our model. Notice that the estimate for the β values in Equation 11 is precisely the equation for ordinary least squares estimates.

I won't go into detail of how to solve Equation 11 but any of the standard ideas will work such as a gradient descent or taking partial derivatives with respect to all the parameters, set them to zero and solve the system of equations. There are a huge variety of ways to solve this equation that have been studied quite extensively.

Prediction

Once we have the coefficients for our linear regression from Equation 11, we can now predict new values. Given a vector of explanatory variables $\mathbf{x}$, predicting y is a simple computation:

$$\hat{y_i} = E(y_i|\mathbf{x_i}) = \beta_{i0} + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} \tag{12}$$

I included the expectation here to emphasize that we're generating a point estimate for y. The expectation is the most likely value for y (according to our model) but our model is really predicting that y is most likely a band of values within a few σ of this expectation. To actually find this range, we would need to estimate σ but it's a bit outside the scope of this post.

Many times though, a point estimate is good enough and we can use it directly as a new prediction point. With classical statistics, you can also derive a confidence interval or a prediction interval around this point estimate to gain some insight into the uncertainty of it. A full Bayesian approach is probably better though since you'll explicitly state your assumptions (e.g. priors).

Generalized Linear Models (GLM)

Changing up some of the modeling decision we made above, we get a different type of regression model that is not any more complicated. Generalized linear models are a generalization of the ordinary linear regression model we just looked at above except that it makes different choices. Namely, the choice of probability distribution and choice of how the mean of the outcome variable relate to the explanatory variables (i.e. "link function"). The above methodology for deriving ordinary linear regression can be equally applied to any of the generalized linear models. We'll take a look at a Poisson Regression as an example.

Poisson Regression

The first big difference between ordinary and Poisson regression is the distribution of the outcome variable y. A Poisson regression uses a Poisson distribution (duh!) instead of a normal distribution:

$$\begin{aligned} Y &\sim Poisson(\lambda) \\ E(Y) &= Var(Y) = \lambda \end{aligned} \tag{13}$$

The Poisson distribution is a discrete probability distribution with a single parameter λ. Since the Poisson regression is discrete, so is our outcome variable. Typically, a Poisson regression is used to represent count data such as the number of letters of mail (or email) in a day, or perhaps the number of customers walking into a store.

The second difference between ordinary and Poisson regressions is how we relate the linear function of explanatory variables to the mean of the outcome variable. The Poisson regression assumes that the logarithm of the expected value of the outcome is equal to the linear function of the explanatory variables:

$$\log E(Y) = \log \lambda = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} \tag{14}$$

Now with these two equations, we can again derive the log-likelihood function in order to derive an expression to estimate the β parameters (i.e. the maximum likelihood estimate). Using the same scenario as Equation 6, namely $n\left(y_i, \mathbf{x_i}\right)$ i.i.d. points, we can derive a log likelihood function (refer to the Wikipedia link for a reference of the probability mass function of a Poisson distribution):

$$\begin{aligned}
\arg\max_{\beta} \mathcal{L}(\beta|y_i) &= \prod_{i=1}^{n} \frac{\lambda^{y_i} e^{-\lambda}}{y_i!} \\
\arg\max_{\beta} \log \mathcal{L}(\beta|y) &= \sum_{i=1}^{n} \left(y_i \log \lambda - \lambda - \log y_i!\right) \\
&= \sum_{i=1}^{n} \left(y_i \log \lambda - \lambda\right) \\
&= \sum_{i=1}^{n} \left(y_i (\beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}) - e^{(\beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip})}\right)
\end{aligned} \tag{15}$$

You can arrive at the last line by substituting Equation 14 in. Unlike ordinary linear regression, Equation 15 doesn't have a closed form for its solution. However, it is a convex function meaning that we can use a numerical technique such as gradient descent to find the unique optimal values of β that maximize the likelihood function.

Prediction of Poisson Regression

Once we have a point estimate for β, we can define the distribution for our outcome variable:

$$Y \sim Poisson(exp\{\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p\}) \tag{16}$$

and correspondingly our point prediction of $\hat{y}_i$ given its explanatory variables:

$$\hat{y}_i = E(y_i) = exp\{\beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}\} \tag{17}$$

Other GLM Models

There are a variety of choices for the distribution of Y and the link functions. This table from Wikipedia has a really good overview from which you can derive the other common types of GLMs.

The logistic regression is actually a type of GLM with outcome variable modeled as a Bernoulli distribution and link function as the logit function (inverse of the logistic function, hence the name). In the same way as we did for the ordinary and Poisson regression, you can derive a maximum likelihood expression and numerically solve for the required coefficients (there is no closed form solution similar to the Poisson regression).

Conclusion

Linear regression is such a fundamental tool in statistics that sometimes it is not explained in enough detail (or as clearly as it should be). Building up a regression model from the bottom up is much more interesting than the traditional method of presenting the end result and scarcely relating it back to its probabilistic roots. In my opinion, there's a lot of beauty in statistics but only because it has its roots in probability. I hope this post helped you see some of the beauty of this fundamental topic too.

References and Further Reading

- Wikipedia: Linear Regression, Ordinary Least Squares, Generalized linear models, Poisson Regression

[1] Since logarithm is monotonically increasing, it achieves the same maximum as the logarithm of a function at the same point. It's also much more convenient to work with because many probability distributions have an exponents or are multiplicative. The logarithm brings down the exponents and changes the multiplications to additions.

[2] This is true at least in a Bayesian interpretation. In a frequentist interpretation, there is a fixed true value of μ, and what is random is the confidence interval we can find that "traps" it. I've written a bit about it here.

[3] We explicitly use the conditional notation here because the value of y depends on $\mathbf{x}$.

Bayesian logistic Poisson probability regression

Previous post

Next post

Hi, I'm Brian Keng. This is the place where I write about all things technical.

Twitter: @bjlkeng

Signup for Email Blog Posts
Email Address

Subscribe