Regression Modeling with Actuarial and Financial Applications

by Edward Frees

Ch. 1 notes

- Regression is a method that is widely used by many disciplines and in many situations.
- Regression allows us to make statements about variables after having controlled for values of known explanatory variables.

- Explanatory variables – x values, these are used to help explain or predict another variable.
- Response variable – y value, this is the outcome that you are trying to predict, using the information that you have from the x values.

Normal distribution

- This distribution is very common and is assumed to be true in many formulas and proofs, we will learn how to often ‘fix’ the data when it is not Normal.
- The formula looks like this:

$$f(y) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma^2}(y-\mu)^2\right)$$

- Fortunately, we don’t have to work this, they have made z tables so that you can easily look up probabilities for z scores.
- In the formula, μ is the location parameter, which describes the average of the data and σ is the scale parameter, which describes how spread out the data is.
- Also known as the Gaussian distribution.

Basic summary statistics

Consider a random sample observed $Y_1, Y_2, \ldots, Y_n$. Data sorted in ascending order: $Y_{(1)}, Y_{(2)}, \ldots, Y_{(n)}$. A statistic is a numerical summary measure of the sample.

- sample mean - average value of the data: $\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$.
- sample median M: the midpoint of the data, which can be computed as $M = \begin{cases} Y_{(n+1)/2}, & \text{if } n \text{ is odd,} \\ \frac{1}{2}\left[Y_{(n/2)} + Y_{(n/2+1)}\right], & \text{if } n \text{ is even.} \end{cases}$
- standard deviation - a measure of the spread of the data:
  $$s_Y = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2} = \sqrt{\frac{1}{n-1}\left(\sum_{i=1}^{n} Y_i^2 - n\overline{Y}^2\right)}.$$
- percentile - a value that a specified fraction of the data is below it, e.g. 25th percentile, 95th percentile. The 50th percentile is also the median.
- minimum - smallest value of the data: $Y_{(1)}$.
- maximum - largest value of the data: $Y_{(n)}$.

Graphs

- Histogram – Provides a visual of how data is spread out, but cannot provide information on individual observations.

- Box Plot – Helps identify shape of distribution of data (is it Normal or not? – it is if each side of the box and each of the whiskers are roughly equal sizes to each other). Also identifies outliers as data points outside 1.5*IQR.

- QQ plot – Helps assess whether the data is normally distributed. If the points lie along the straight 45 degree line, the data is normally distributed. Large deviations from a straight line is a problem.

Transformations

- If the data is found to be not normally distributed, we often will transform the data to ‘fix’ it and make it be normally distributed so that certain statistical methods will work.

- Log = Log(y) = usually the best and most common (this one or ln) , it shrinks large values a lot and changes small values very little, this helps suck the data in so it is not affected by a lot of outlier type points.
- Ln = ln(y) = the natural log, same wonderful effects as log. As a warning - in a number of statistical textbooks and references, they will talk about taking the log of something when they are actually referring to the ln.
- Square root = √(y)
- Reciprocal = (1/y)
- Negative Reciprocal = -(1/y)

- There is no good way to just look at your data and figure out which transformation is best, you often just start with ln or log and then move down the list from there. Try different ones and see what makes your data look best.

Population distributions

- Population: entire collection of objects of interest.
- Sample: (random) subset of population.
- Statistical thinking: draw inferences about population by using sample data.
- Model: mathematical abstraction of the real world used to make statistical inferences.
- Assumptions:
  - model provides a reasonable fit to sample data, and
  - sample is representative of entire population.
- Normal distribution: simple, effective model (“bell-curve”).
  - continuous, symmetric curve, and
  - has lots of desirable properties.

Random sampling

- Population parameters: numerical summary measures of the population, such as mean, μ and standard deviation σ.
- Sample statistics: analogous sample measures such as sample mean $\overline{Y}$ and sample standard deviation s_Y.
- Statistical inference: use sample statistics to infer about (likely values of) population parameters.
- Example: you may use sample mean to estimate population mean.
- Next question is: how far off might this estimate be?
    - Could be long way off if Y is very variable and/or sample size is small.
- Can quantify uncertainty using sampling distributions.

Central Limit Theorem - normal version

- Suppose $Y_1, Y_2, \ldots, Y_n$ is a random sample from a population with mean μ and variance σ^2.
- Central Limit Theorem:
  - The sample mean $\overline{Y}$ has a Normal distribution: $\overline{Y} \sim \mathrm{N}(\mu, \sigma^2/n)$.
  - Hence: $Z = \dfrac{\overline{Y} - \mu}{\sigma/\sqrt{n}} \sim \mathrm{N}(0, 1)$.

Student t-distribution

- Main drawback to CLT: must know the population variance, σ^2.
- But we rarely know the variance, hence what would be a good estimate of variance? The sample variance: s_Y^2.
- Replacing σ^2 with s_Y^2 requires use of a t-distribution rather than the normal:
    - The t-distribution is just like the normal, but more spread out (fatter in tails) to reflect the additional uncertainty.
    - This additional uncertainty is because of replacing the population variance with sample variance.
    - s_Y^2 provides a good estimate of σ^2 when n is large.
    - t-distribution accounts for this using degrees of freedom (df $= n - 1$ in this case).
    - As df increase, t-distribution approaches the normal.