Linear Regression

Frank Zhong

August 2024

1 Introduction

In statistics, linear regression is a statistical model which estimates the linear relationship between a scalar response and one or more explanatory variables (also known as dependent and independent variables) [1]. The case of one explanatory variable is called simple linear regression; for more than one, the process is called multiple linear regression. This term is distinct from multivariate linear regression, where multiple correlated dependent variables are predicted, rather than a single scalar variable.

2 Construction of the Model

2.1 General Formula

The simple linear regression only has one variable, its formula is given by:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

Where the variable x_i and the output y_i follows an linear relationship with slope β_1 and y-intercept β_0. Since such a relationship may not hold exactly for the largely unobserved population of values of the independent (variables) and dependent (outputs) variables; we call the unobserved deviations from the above equation the errors. In this formula the error (residue) is denoted as ϵ_i. To build up the model and predict, our goal is to find the estimates for slope β_1 and y-intercept β_0. The concept of likelihood function therefore come in.

2.2 Assumptions

The linear regression model has the following basic assumptions [2]:

- **Linearity**: The relationship between x and the mean of y is linear.
- **Homoscedasticity**: The variance of residual is equal for any value of x.
- **Independence**: Observations are independent of each other.
- **Normality**: The residue is normally distributed.

2.3 Likelihood Function

A likelihood function (often simply called the likelihood) measures how well a statistical model explains observed data by calculating the probability of seeing that data under different parameter values of the model [3], written as:

$$\mathcal{L}(\theta|x)$$

Which represents the probability of obtaining the parameter θ given that we have the observed value x. Hence it's not hard to understand that we aim to maximise this function value, so that we have high probability of getting θ given x, that is to say, the parameter θ best fits the observations.

2.4 Maximum Likelihood Estimation

Suppose we denote the estimate for coefficients as slope $\hat{\beta_1}$ and y-intercept $\hat{\beta_0}$, then the output y_i follows the normal distribution:

$$y_i \sim N\left(\hat{\beta_0} + \hat{\beta_1} x_i \ , \ \sigma^2\right)$$

Therefore we have the probability density function of y_i:

$$\mathbb{P}\left(y_i\right) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}\left(y_i - \hat{\beta_0} - \hat{\beta_1} x_i\right)^2}$$

Since all y_is are independent, the likelihood function is the combined probability for all outputs, that is [4]:

$$L\left(\hat{\beta_0}, \hat{\beta_1}, \sigma^2\right) = \mathbb{P}\left(y_1, y_2, \cdots, y_n\right) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}\left(y_i - \hat{\beta_0} - \hat{\beta_1} x_i\right)^2}$$

Which can be simplified to:

$$L\left(\hat{\beta_0}, \hat{\beta_1}, \sigma^2\right) = \frac{1}{\sigma^n \left(2\pi\right)^{\frac{n}{2}}} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i - \hat{\beta_0} - \hat{\beta_1} x_i\right)^2}$$

To ease our calculation process, we take the log likelihood, which will pull down the exponent and gives the same result as maximising the likelihood function.

$$\mathcal{L} = \ln L\left(\hat{\beta_0}, \hat{\beta_1}, \sigma^2\right) = -n \ln\left(\sigma\sqrt{2\pi}\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i - \hat{\beta_0} - \hat{\beta_1} x_i\right)^2$$

Based on calculus, to maximise a function, we can take its derivative and set them to 0.

$$\frac{\partial L}{\partial \hat{\beta_0}} = \frac{1}{2\sigma^2}\sum_{i=1}^{n} 2\left(y_i - \hat{\beta_0} - \hat{\beta_1} x_i\right)$$

$$\frac{\partial L}{\partial \hat{\beta}_1} = \frac{1}{2\sigma^2} \sum_{i=1}^{n} 2 \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \right) x_i$$

By setting the two derivatives to 0, we obtain:

$$\sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \right) = 0 \tag{1.1}$$

$$\sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \right) x_i = 0 \tag{1.2}$$

Which are identical to the least square estimation [4]. Therefore we can use least squares estimates to maximise the likelihood function value in linear regressions. *Proof.* The mean squared error is given by:

$$MSE\left(\beta_0, \beta_1\right) = \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \beta_0 - \beta_1 x_i \right)^2$$

Similarly, we take the derivatives to find the maximum values:

$$\frac{\partial MSE}{\partial \beta_0} = -\frac{2}{n} \sum_{i=1}^{n} \left(y_i - \beta_0 - \beta_1 x_i \right)$$

$$\frac{\partial MSE}{\partial \beta_1} = -\frac{2}{n} \sum_{i=1}^{n} \left(y_i - \beta_0 - \beta_1 x_i \right) x_i$$

Solving for the two derivatives equal to 0 gives:

$$\frac{1}{n} \sum_{i=1}^{n} \left(y_i - \beta_0 - \beta_1 x_i \right) = 0 \tag{2.1}$$

$$\frac{1}{n} \sum_{i=1}^{n} \left(y_i - \beta_0 - \beta_1 x_i \right) x_i = 0 \tag{2.2}$$

Which, up to a factor of $\frac{1}{n}$, shares the same result as the maximum likelihood estimation (Eq. 1.1 and Eq. 1.2).
Following Eq. 2.1 and Eq. 2.2, we can obtain the estimates for the regression coefficients:

$$\bar{y} - \hat{\beta}_0 - \hat{\beta}_1 \bar{x} = 0 \tag{3.1}$$

$$\overline{xy} - \hat{\beta}_0 \bar{x} - \hat{\beta}_1 \bar{x^2} = 0 \tag{3.2}$$

Substituting Eq. 3.1 into Eq. 3.2, we obtain:

$$\overline{xy} - \bar{x}\bar{y} + \hat{\beta_1}\bar{x}^2 - \hat{\beta_1}\bar{x^2} = 0$$

From which we can derive for β_1:

$$\hat{\beta_1} = \frac{\overline{xy} - \bar{x}\bar{y}}{\bar{x^2} - \bar{x}^2} = \frac{\mathbb{E}\left(xy\right) - \mathbb{E}\left(x\right)\mathbb{E}\left(y\right)}{\mathbb{E}\left(x^2\right)\mathbb{E}\left(x\right)^2} = \frac{c_{xy}}{s_x^2}$$

Where c_{xy} denotes the covariance of x and y, and s_x^2 denotes the variance of x. In addition, we see from Eq. 3.1 that:

$$\bar{y} = \hat{\beta_0} + \hat{\beta_1}\bar{x}$$

Which represents that the regression line always passes the point $(\bar{x}, \bar{y})$.

3 Evaluation of the Model

3.1 Coefficient of Determination

The coefficient of determination, or denoted as R^2, is the proportion of the variation in the dependent variable that is predictable from the independent variable(s) [5]. The R^2 is a value lying between $[0, 1]$ and the larger R^2 value shows that the model has a better goodness of fit, therefore the model will be more trustworthy.

Suppose we define the residual sum of squares SS_{res} and total sum of squares SS_{tot} to be:

$$SS_{res} = \sum_{i=1}^{n} \left(y_i - \hat{y_i}\right)^2 = \sum_{i=1}^{n} e_i^2$$

$$SS_{tot} = \sum_{i=1}^{n} \left(y_i - \bar{y}\right)^2 \propto Var\left(y\right)$$

Then we will have:

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

Hence if the model exactly fits the data, we will have $SS_{res} = 0$ and $R^2 = 1$, and on the other hand, if the model always estimates output to be $\bar{y}$ then $SS_{res} = SS_{tot}$ and $R^2 = 0$.

3.2 Adjusted R^2

R^2 is not a good measure of the predictive ability of a model. It measures how well the model fits the historical data, but not how well the model will forecast

future data. In addition, adding any variable tends to increase the value of R^2, even if that variable is irrelevant. For these reasons, forecasters should not use R^2 to determine whether a model will give good predictions, as it will lead to over-fitting.
Hence we turn to the adjusted R^2:

$$\bar{R}^2 = 1 - \left(1 - R^2\right) \frac{T-1}{T-k-1}$$

where T is the number of observations and k is the number of predictors. This is an improvement on R^2, as it will no longer increase with each added predictor. Using this measure, the best model will be the one with the largest value of $\bar{R}^2$.

3.3 Residues

The difference between the observed y values and fitted $\hat{y}$ values are defined as residues:

$$e_t = y_t - \hat{y_t} = y_t - \hat{\beta}_0 - \hat{\beta}_1 x_{1,t} - \hat{\beta}_2 x_{2,t} - \cdots - \hat{\beta}_k x_{k,t}$$

Each residual is the unpredictable component of the associated observation. The residuals have some useful properties including the following two:

$$\sum_{t=1}^{T} e_t = 0$$

$$\sum_{t=1}^{T} x_{k,t} e_t = 0$$

As a result of these properties, it is clear that the average of the residuals is zero, and that the correlation between the residuals and the observations for the predictor variable is also zero.
After selecting the regression variables and fitting a regression model, it is necessary to plot the residuals to check that the underlying assumptions of the model have been satisfied.

3.4 Plotting the Residues

- **Autocorrelation Function (ACF) Plot**: The ACF plot is a graphical representation of the correlation of a time series with itself at different lags. The correlation coefficient is a measure of how closely two variables are related.

- **Histogram**: Histogram gives a rough sense of the underlying distribution of the data. We can use it to check if our residues are approximately normally distributed.

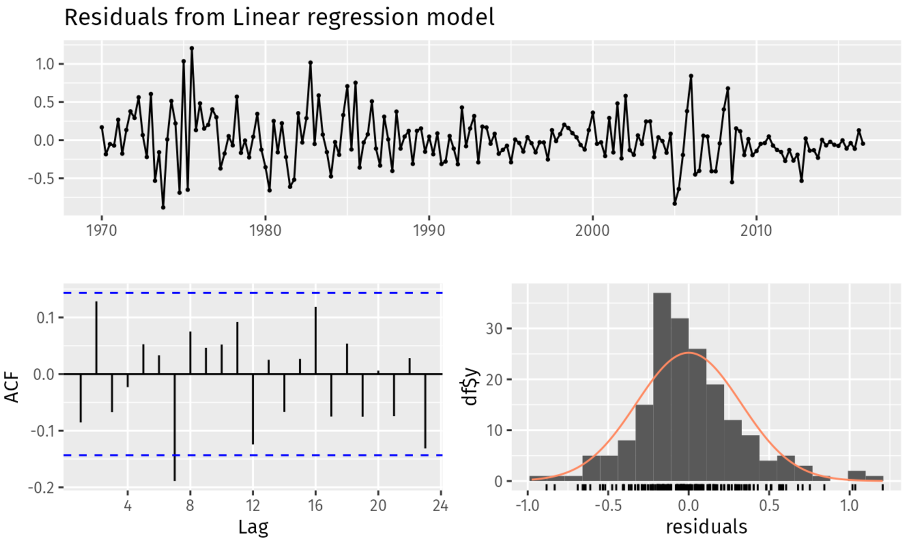

Figure 1: Example of a residue plot for linear regression.

Figure 1 shows a time plot, the ACF and the histogram of the residuals from a multiple regression model.

The time plot shows some changing variation over time, but is otherwise relatively unremarkable. This heteroscedasticity will potentially make the prediction interval coverage inaccurate.

The histogram shows that the residuals seem to be slightly skewed, which may also affect the coverage probability of the prediction intervals.

The ACF plot shows a significant spike at lag 7, but it is not quite enough for the Breusch-Godfrey test to be significant at the 5% level. In any case, the autocorrelation is not particularly large, and at lag 7 it is unlikely to have any noticeable impact on the forecasts or the prediction intervals.

- **Residues vs. predictor variables plot**: We would expect the residuals to be randomly scattered without showing any systematic patterns. A simple and quick way to check this is to examine scatter plots of the residuals against each of the predictor variables. If these scatterplots show a pattern, then the relationship may be nonlinear and the model will need to be modified accordingly.

It is also necessary to plot the residuals against any predictors that are not in the model. If any of these show a pattern, then the corresponding predictor may need to be added to the model.

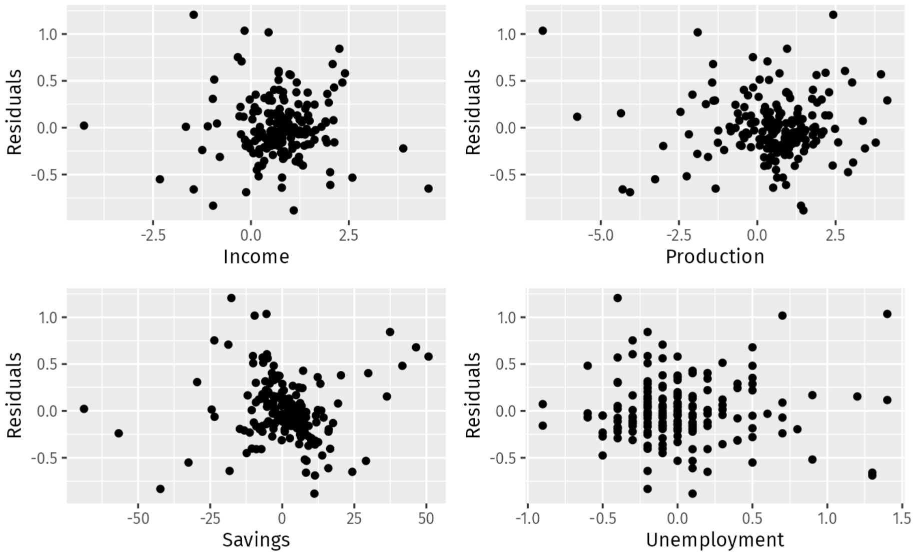

Figure 2: Example of a residue vs. predictor variable plot.

Figure 2 shows a residue vs. predictor variable plot. Each of the four residue seem to be randomly scattered, therefore no further modification is required.

- **Residues vs. fitted values plot**: A plot of the residuals against the fitted values should also show no pattern, so that the distribution of residues satisfies the Guaasian-Markov Condition. If a pattern is observed, there may be "heteroscedasticity" in the errors which means that the variance of the residuals may not be constant. If this problem occurs, a transformation of the forecast variable such as a logarithm or square root may be required.

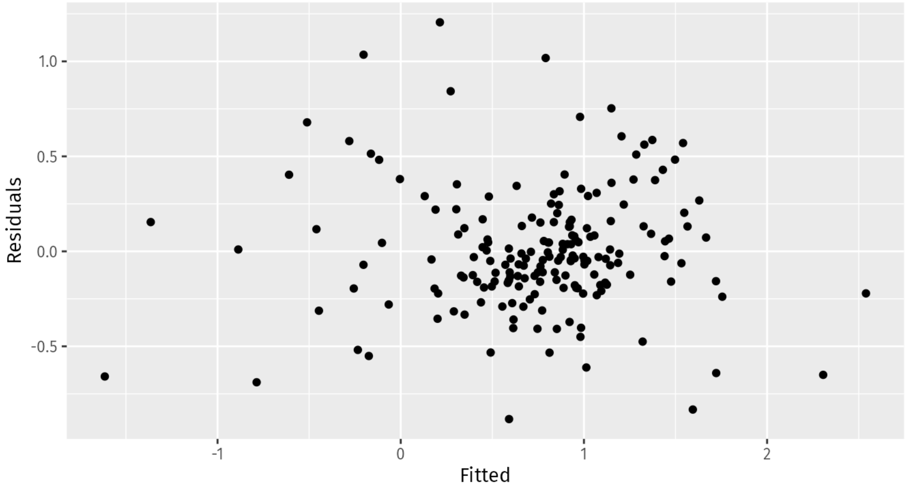

Figure 3: Example of a residue vs. fitted value plot.

Figure 3 shows a residue vs. fitted value plot. The random scatter distribution suggests the errors are homoscedastic.

- **Normal quantile-quantile (Q-Q) plot**: Q-Q plots allow us to compare the quantiles of two sets of numbers and hence the distributions of them. This kind of comparison is much more detailed than a simple comparison of means or medians, however, we should be careful that more observations are required than for simple comparisons.

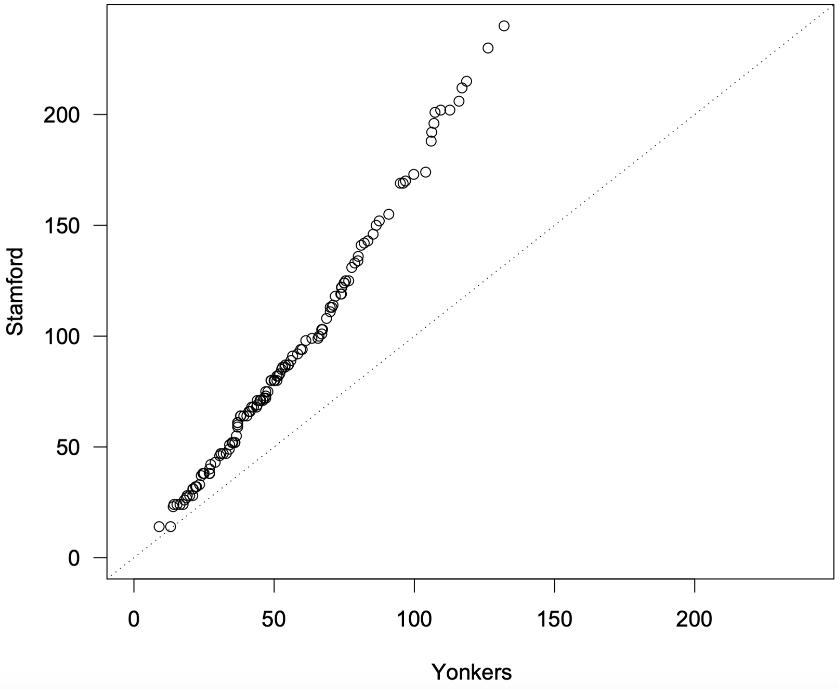

Figure 4: Example of a Q-Q plot.

We can see from Figure 4 that the distribution of the variable Stamford is not the same as that of Yonkers, since the scattered points are not lying on the line $y = x$ (the dashed line in the plot). That is to say, the quantiles from both data frames are not similar.

- **Scale-location plot**: Similar to the residue vs. fitted value plot, except that the y axis is changed from residue to standardized residual. This makes it easier to notice the range of residues' distribution. An example is shown in Figure 5.

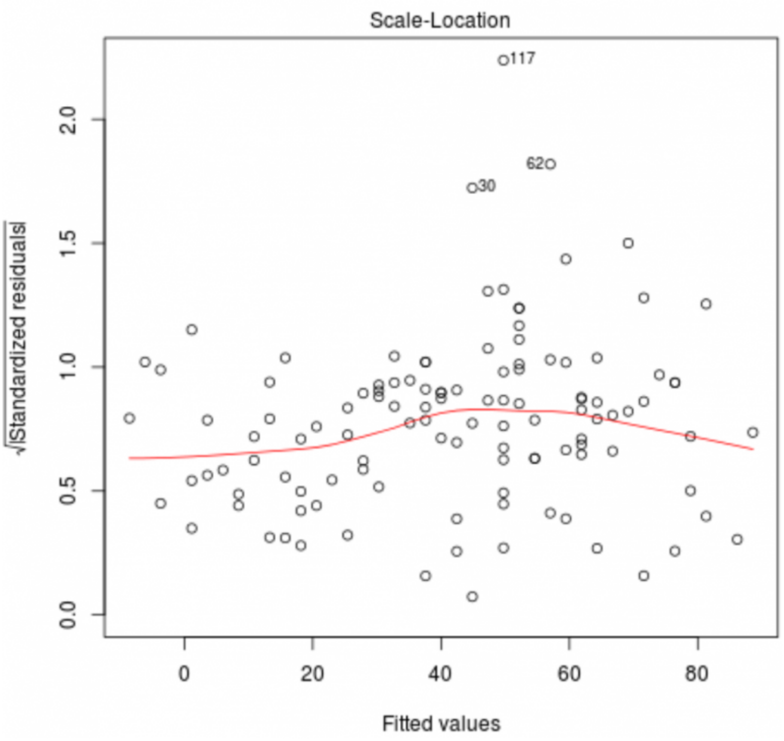

Figure 5: Example of a scale-location plot.

- **Residue vs. leverage plot**: We can find out the outliers, leverages and influential points from a residue vs. leverage plot.

Outliers: Points far from the regression line, cannot be explained well by the model.
In the residue vs. leverage plot: Check the residue values. For small sample sizes, points with standardised residues bigger than 2 are considered as outliers and bigger than 4 for large sample sizes.

Leverage points: Points with extreme x values. Some of them fits the model well and hence are named good leverage points, which improve the universality of the model. However, good leverage points will over-raise the R^2 value and our confidence.
In the residue vs. leverage plot: Check the leverage values (Cook's distance), points with Cook's distance larger than $\frac{n}{4}$ are considered as leverage points.

Influential points: Points that are both outliers and bad leverage points. They lay great impact on the stability of the model and pull the regression line towards themselves.
In the residue vs. leverage plot: The points that are both outliers and leverage points.

The calculation for Cook's distance of the i^{th} observation is given by:

$$D_i = \frac{\sum_{j=1}^{n} \left(\hat{y_j} - \hat{y_{j(i)}}\right)^2}{ps^2}$$

where p is the number of independent variables and $\hat{y_{j(i)}}$ is the fitted response value obtained when excluding observation i, and s^2 is the mean squared error of the regression model:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y_i})^2$$

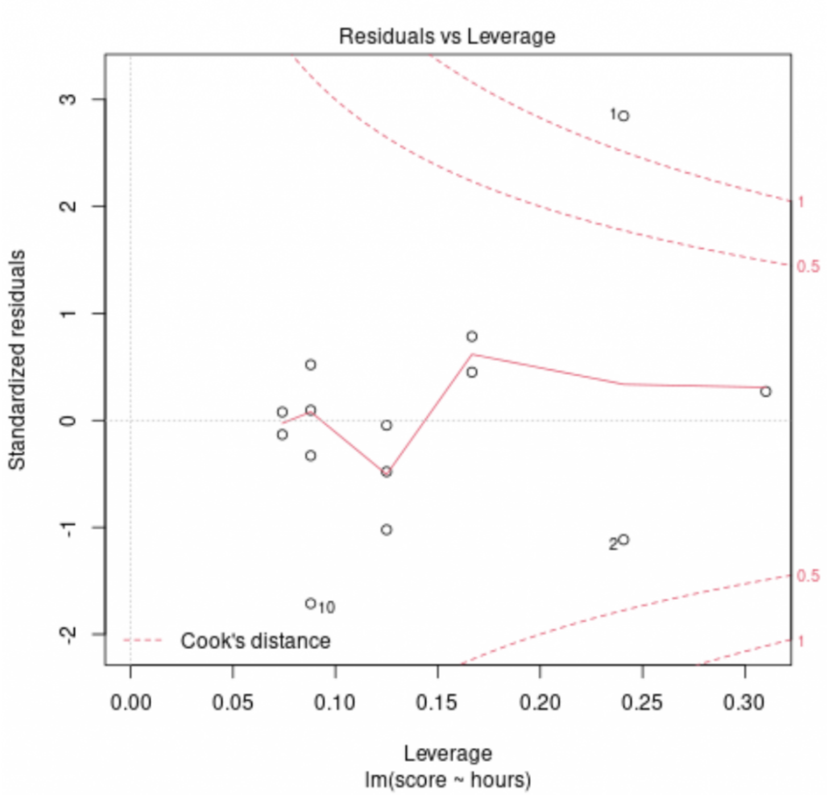

Figure 6: Example of a residue vs. leverage plot.

We observe in Figure 6 that point 1 is an influential point.

4 Visualisation of the Model

For linear regression, we often use the nomogram to visualise the regression results.

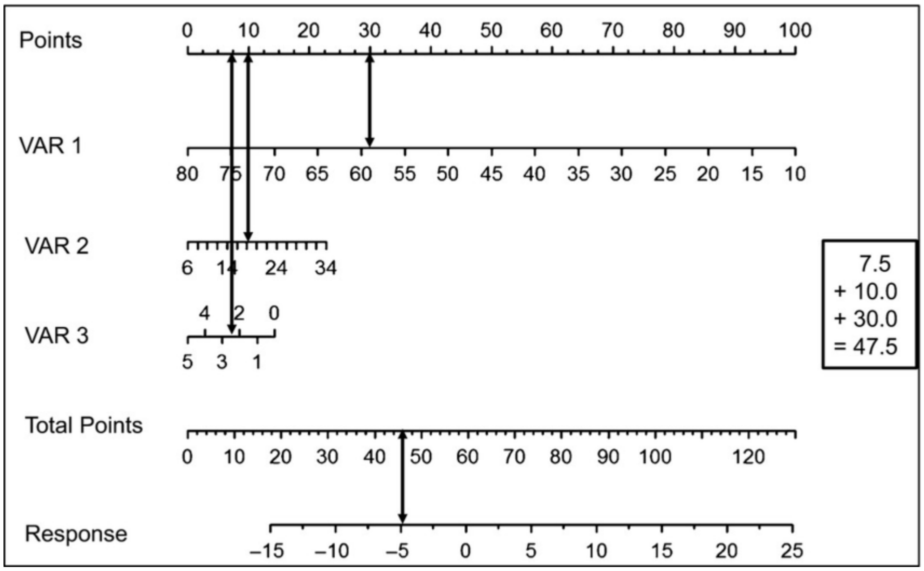

Figure 7: Example of a nomogram.

The example shows a linear regression with three variables (VAR 1, 2, 3). The length and density of the scale lines represents the influencing power of each variable. In Figure 7, obviously variable VAR 1 has the largest (negative) influencing power to the output.

References

[1] Wikipedia. *Linear regression.* URL: `https://en.wikipedia.org/wiki/Linear_regression` (visited on 08/07/2024).

[2] *Correlation and Regression with R.* URL: `https://sphweb.bumc.bu.edu/otlt/MPH-Modules/BS/R/R5_Correlation-Regression/R5_Correlation-Regression7.html` (visited on 08/08/2024).

[3] Wikipedia. *Likelihood function.* URL: `https://en.wikipedia.org/wiki/Likelihood_function` (visited on 08/08/2024).

[4] Cosma Shalizi. *36-401, Modern Regression, Section B.* URL: `https://www.stat.cmu.edu/~cshalizi/mreg/15/` (visited on 08/08/2024).

[5] Wikipedia. *Coefficient of determination.* URL: `https://en.wikipedia.org/wiki/Coefficient_of_determination` (visited on 08/09/2024).