Regression Analysis, Model Selection and Bayesian Inference

Julia Navarro

April 2021

1 Multiple Linear Regression

In the following pages we will infer how explanatory variables affect the ferritin levels in the bloodstream. We will be analyzing the given dataset from 202 Australian athletes, considering fitting a model, eliminating insignificant variables and undertaking model diagnostics, and ensuring the assumptions of a linear model are met. We will conclude with comparing the predictions to the actual values and proving/disproving if the model is an accurate one.

1(a)

	Lean Body Mass (LBM)	Red Cell Count (RCC)	White Cell Count (WCC)	Hema-tocrit (Hc)	Haemo-globin (Hg)	BMI	Sum of Skin Folds (SSF)	% Body Fat (X.Bfat)	Ferritin (Ferr)
Minimum	34.36	3.800	3.300	35.90	11.60	16.75	28.00	5.630	8.00
1st Quartile	54.67	4.372	5.900	40.60	13.50	21.08	43.85	8.545	41.25
Median	63.03	4.755	6.850	43.50	14.70	22.72	58.60	11.65	65.50
Mean	64.87	4.719	7.109	43.09	14.57	22.96	69.02	13.507	76.88
3rd Quartile	74.75	5.030	8.275	45.58	15.57	24.46	90.35	18.08	97.00
Maximum	106.00	6.720	14.300	59.70	19.20	34.42	200.80	35.520	234.00

Interpreting the above table, we see that there is noticable spread for the following variables: LBM, SSF and Ferr. Note that the ferritin varies from 8 to 234 μmol/L. Next we shall visually investigate the relationships between these variables, using the code in R (line 10).

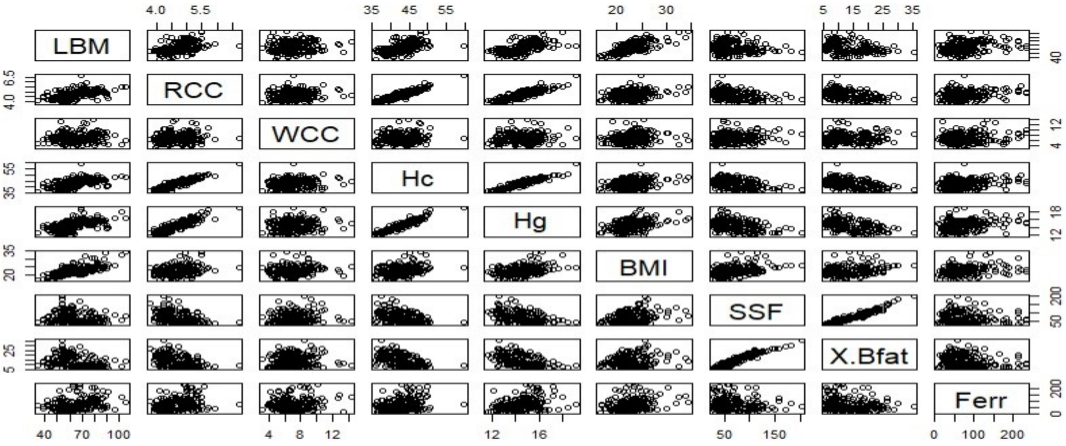


Figure 1: Plots of Pairs

We see that RCC positively correlates with Hc and Hg, whereas the correlation of LBM and Ferr is vague; there neither positive nor negative correlation to be clearly found. A strong positive relation exists between SSF and X.Bfat. Some variables show signs of some relation, but these are weak.

1(b)

Consider the two plots below:

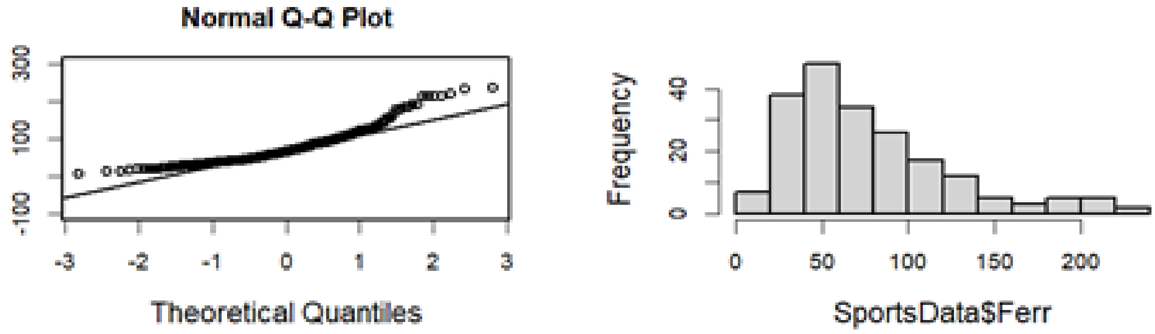


Figure 2

Here we can see that the QQ-plot suggests that the distribution of the errors in the variable Ferr are not distributed normally. Additionally, the histogram is skewed positively, not fitting a bell shape, which indicates Ferr itself is not normally distributed. We can therefore try and transform the variable Ferr. Consider possibly taking the natural logarithm of this variable.

2(a)

A regression model with Ferr as the response and other ten variables as predictors would take the form:

$$\hat{T}F=\hat{\beta}_0+\hat{\beta}_1 Sport+\hat{\beta}_2 Sex+\hat{\beta}_3 LBM+\hat{\beta}_4 RCC+\hat{\beta}_5 WCC+\hat{\beta}_6 Hc+\hat{\beta}_7 Hg+\hat{\beta}_8 BMI+\hat{\beta}_9 SSF+\hat{\beta}_{10} Bfat$$

Where $\hat{T}F$ represents ferritin in the *Training* dataset. Note that in the above equation Sport and Sex are categorical variables, and that our response variable and beta values are all approximations.

2(b)

Let us consider two models: *TrainingModel1* (has all 10 variables) and *TrainingModel2* (variables with a low p-value were removed leaving; only Sport, Sex, RCC and BMI remain). Let us compare these two models using the ANOVA test in R (line 71).

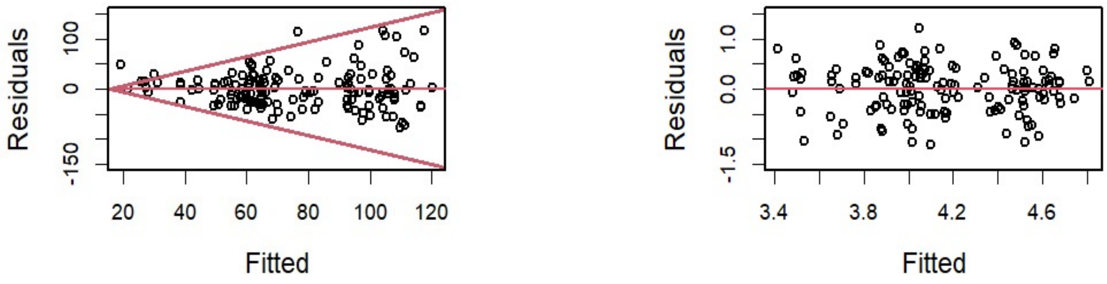
This gives us the values F=1.7495 and p=0.1152 > 0.05. Since the p value is larger than 0.05 there is no evidence to reject H_0. This means that both models perform similar, however, we keep the one with less variables for less complexity. Alternatively, we can use the AIC and BIC tests: The lower the value the better the model (lines 81-84). We obtain the smallest values for *TrainingModel2* (AIC=1444.099 and BIC= 1485.381, compared with AIC=1444.46 and BIC=1503.436 for *TrainingModel1*).

This supports the claim that *TrainingModel2* would be a better model. The ANOVA comparison above in fact said both models were similar, but by using the AIC and BIC we can confirm that the model with less explanatory variables is better. Hence, we will perform model diagnostics on *TrainingModel2*.

2(c(i))

We will now check for constant variance, independence, and normality assumptions of the errors in *TrainingModel2*.

To check if the constant variance assumption holds, we interpret the output in R (line 115-177). This plots our fitted model against its residuals. We observe a sort of funneling pattern (see plot (a) of the figure below) and hence we require a transformation of the response variable.



(a) $Ferr$ (b) $\log Ferr$

Figure 3

To resolve this non-constant variance, we will take the natural logarithm of the response variable, Ferr. Redoing the analysis (line 122, 125-126) yields a graph with constant variance, see plot (b) above. Some small clustering is still visible, but note the different scales on the two plots, as this truly highlights how much better our transformed response variable is in terms of constant variance.

The normality of errors can be verified using a QQ-plot:

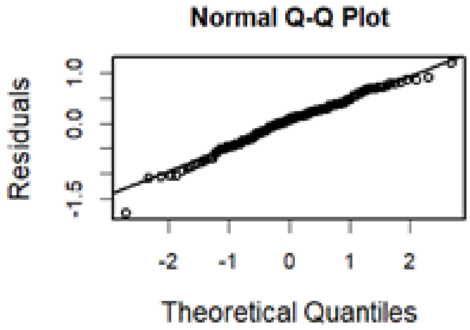


Figure 4

The above figure shows us that the errors follow a normal distribution, because the points follow the line so closely. We can confirm this using the Shapiro-Wilk test in R (line 142)(the null hypothesis being that residuals are distributed normally). The p-value is 0.1089 (at the significance level of 0.05, p-value ¿ 0.05), we do not reject this hypothesis – so the residuals (errors) are normally distributed.

Lastly, we check independence assumption by plotting each residual next to each other in order(line 159-161).

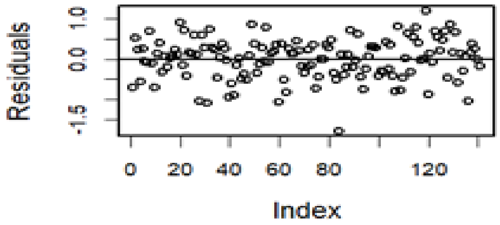


Figure 5

The figure shows no signs of dependence, and so the errors are independent, as required.

We have now checked that the assumptions hold and continue onto searching for outliers, large leverage and influential points.

2(c(ii))

The figure below will aid in identifying unusual observations (lines 178-188):

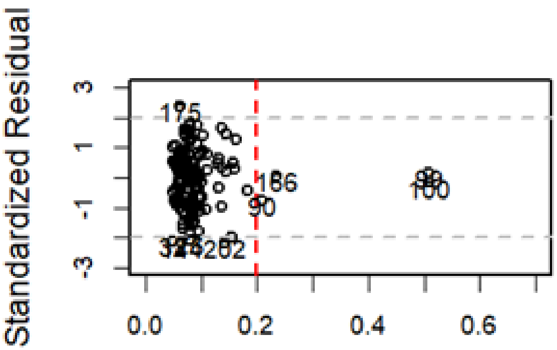


Figure 6

Outliers here can be found by searching for points above or below the grey dotted lines, either above the line y=2 or below y= $-$ 2. Points 90, 99, 100 and 166 show signs of leverage and hence could be inspected more closely, as could the outliers. Typically ,though, large leverages do not effect the model. For example points 99 and 100 both concern Females in the Gym sport category. Taking a closer look, we see that both have a low LBM (34.36 and 39.03 respectively). Recalling from 1a) that the lower bound for LBM was around 34.36, we will assume that this is what may have flagged these points as having large leverage. A possible way to deal with these would be to check why the values are so disparate. For example, there may have been an error in recording the data. If there is a justifiable reason, these can be removed, however, this could result in more outliers appearing.

No influential observations can be observed in our model; no data points both have large leverage *and* lie outside our limits (grey lines).

2(d)

Using the code in R (line 200) to output a summary table, we can observe any significant predictors. This shows that, when taking Basketball as the reference among the sport types, the following sports show significance: Rowing, Swimming, Tennis and Sprint, whilst the other Sports have an overall small positive effect on Ferritin. The largest influence is tennis, giving an estimate of 0.71634.

The sex variable has two levels: SexMale and SexFemale. Comparing Male to Female shows that the level SexMale has roughly a 0.56517 additional effect on Ferr than SexFemale has, meaning that the SexMale is better at predicting the Ferritin levels in the blood than the female gender is. Amongst the other variables, BMI acts as a low predictor, having a positive effect of Ferritin compared to the small negative effect of the variable RCC. The effect of both on the overall model are fairly negligible though. The Adjusted R-squared gives 0.259, and hence we conclude that it is not a strong model. We could increase the fit by adding new significant variables.

3

We now test the model using the testing set.

Using R (lines 214-215) we test *TrainingModel2* given the testing set. This gives the predicted ferritin levels and a 95% confidence range (this means we have an upper and lower limit on the values). This range is quite large, suggesting that the model is not confident since the variance is substantial.

We can also graphically demonstrate this by plotting the predicted values from a new model, the *TestingModel*, against the actual ferritin values in this set (see lines 218-227). The figure below shows that most points do not lie on the line, again suggesting that the model is not very accurate in its predictions. (using code from Statology[1])

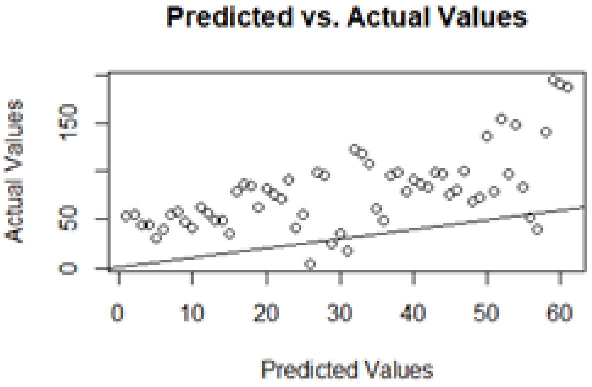


Figure 7

Next, we can observe the difference of the two by calculating the errors (lines 231-235). We create a 61 by 2 matrix containing the actual ferritin values and the predicted ones. We then subtract one from the other, take the absolute value, to obtain the errors. Most show deviations from the actual values.

Now to briefly summarize everything discussed throughout the last few pages. We initially fitted a model using 9 explanatory variables and reduced it down to 4. Continuing, we checked that this new model satisfied the required conditions, and discussed influential observations. Finally, we tested whether or not the predicted values for ferritin resembled the actual values within the testing data set, which told us the model is not very accurate.

2 Bayesian Inference

a) Find $\theta|x$, given $x|\theta \sim N(\theta, 4)$ and $\theta \sim N(12, 9)$. The Likelihood is given as follows, recalling and applying the formula for the normal distribution:

$$\begin{aligned} f(x|\theta) &= \frac{1}{2\sqrt{2\pi}} exp(-\frac{1}{8}(x^2 - 2x\theta + \theta^2)) \\ x &= 13.25 \text{ gives :} \\ &= \frac{1}{2\sqrt{2\pi}} exp(-\frac{1}{8}(13.25^2 - 2 \cdot 13.25\theta + \theta^2)) \\ &= \frac{1}{2\sqrt{2\pi}} exp(-\frac{1}{2 \cdot 4}(\theta - 13.25)^2) \end{aligned}$$

Ignoring constants we have $f(x|\theta) \propto exp(-\frac{1}{8}(\theta^2 - 26.5\theta))$ and $f(x|\theta) \sim N(13.25, 4)$. Next we find the Prior distribution:

$$\begin{aligned} f(\theta) =& \frac{1}{3\sqrt{2\pi}} exp(-\frac{1}{18}(\theta - \mu)^2) \\ \mu =& 12 \text{ gives :} \\ =& \frac{1}{3\sqrt{2\pi}} exp(-\frac{1}{18}(\theta^2 - 2 \cdot 12\theta + 12^2) \end{aligned}$$

Ignoring constants we have $f(\theta) \propto exp(-\frac{1}{18}(\theta^2 - 24\theta))$. Now combining yields:

$$\begin{aligned} f(\theta|x) &\propto f(x|\theta) \cdot f(\theta) \\ &\propto exp(-\frac{1}{8}(\theta^2 - 26.5\theta))exp(-\frac{1}{18}(\theta^2 - 24\theta)) \\ &= exp(-\frac{(\theta^2 - 26.5\theta)}{8} - \frac{(\theta^2 - 24\theta)}{18}) \\ &= exp(\frac{-18(\theta^2 - 26.5\theta) - 8(\theta^2 - 24\theta)}{144}) \\ &= exp(\frac{-26\theta^2 + 669\theta}{2 \cdot 72}) = exp(-\frac{1 \cdot 26}{2 \cdot 72}(\theta^2 - \frac{669}{26}\theta)) \\ &= exp(-\frac{1}{2} \cdot 26 \frac{(\theta^2 - 2 \cdot \frac{669}{52}\theta + (\frac{669}{52})^2)}{72}) \\ &= exp(-\frac{1}{2}\frac{(\theta - \frac{669}{52})^2}{\frac{72}{26}}), \end{aligned}$$

and this gives the posterior distribution:

$$\boxed{\theta|x \sim N(\frac{669}{52}, \frac{72}{26})}.$$

b)

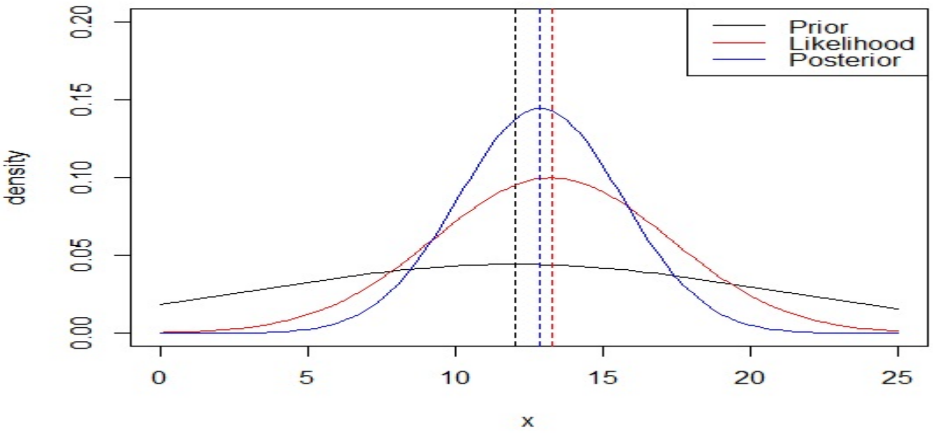


Figure 8: Prior, Likelihood and Posterior distribution plotted in R

The graph indicates that after using the prior and likelihood functions together, the spread, or variation, of the posterior distribution has been reduced, so that the certainty of our belief has increased.

c) We next obtain a formula for the posterior mean and variance of the mean parameter. We have:

$$\begin{aligned} \text{Prior}:\ f(\mu) &\propto exp(-\frac{1}{2}(\frac{\mu-12}{3})^2), \\ \text{Likelihood}:\ f(\bar{x}|\mu) &\propto exp(-\frac{1}{2}(\frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}})^2). \end{aligned}$$

Now we find $f(\mu|\bar{x}) \propto\ f(\mu)f(\bar{x}|\mu)$.

$$\begin{aligned} f(\mu|\bar{x}) &\propto exp(-\frac{1}{2}(\frac{\mu-12}{3})^2)exp(-\frac{1}{2}(\frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}})^2) \\ &= exp(-\frac{1}{2}(\frac{\mu^2-24\mu+12^2}{9}+\frac{\bar{x}^2-2\bar{x}\mu+\mu^2}{\frac{\sigma^2}{n}})). \end{aligned}$$

We again discard the constants, to obtain:

$$\begin{aligned} &\propto exp(-\frac{1}{2}(\frac{\mu^2-24\mu}{9}+\frac{n\cdot(-2\bar{x}\mu+\mu^2)}{\sigma^2}) \\ &= exp(-\frac{1}{2}(\frac{\sigma^2(\mu^2-24\mu)+9n(\mu^2-2\bar{x}\mu)}{9\sigma^2})) \\ &= exp(-\frac{1}{2}(\frac{(\sigma^2+9n)\mu^2+(-24\sigma^2-18n\bar{x})\mu}{9\sigma^2})) \\ &= exp(-\frac{1}{2}(\sigma^2+9n)(\frac{\mu^2-\frac{(24\sigma^2+18n\bar{x})\mu}{\sigma^2+9n}}{9\sigma^2})) \\ &= exp(-\frac{1}{2}(\sigma^2+9n)(\frac{\mu^2-\frac{2\mu(24\sigma^2+18n\bar{x})}{2(\sigma^2+9n)}}{9\sigma^2})) \\ &= exp(-\frac{1}{2}(\sigma^2+9n)\frac{\mu-(\frac{12\sigma^2+9n\bar{x}}{\sigma^2+9n})^2}{9\sigma^2}) \\ &= exp(-\frac{1}{2}\frac{\mu-(\frac{12\sigma^2+9n\bar{x}}{\sigma^2+9n})^2}{\frac{9\sigma^2}{\sigma^2+9n}}) \end{aligned}$$

And so, we obtain the following formulae for the mean and variance:

$$\mu_{\text{post}} = \frac{12\sigma^2+9n\bar{x}}{\sigma^2+9n},$$

and

$$\sigma^2_{\text{post}} = \frac{9\sigma^2}{\sigma^2+9n}.$$

d)We now substitute the following values into the above two equations to obtain the posterior mean and variance of the mean parameter. Let $n = 20, \bar{x} = 11.85, \sigma^2 = 4$.

$$\begin{aligned} \mu_{\text{post}} &= \frac{12\cdot 4+9\cdot 20\cdot 11.85}{4+9\cdot 20} \approx 11.85326087, \\ \sigma^2_{\text{post}} &= \frac{9\cdot 4}{4+9\cdot 20} \approx 0.195652173. \end{aligned}$$

We now plot the following three distributions in R:

$$\begin{aligned} \text{Prior}: f(\mu) &\sim N(12,9) \\ \text{Likelihood}: f(x|\mu) &\sim N(11.85,4) \\ \text{Posterior}: f(\mu|x) &\sim N(\mu_{\text{post}},\sigma^2_{\text{post}}) \end{aligned}$$

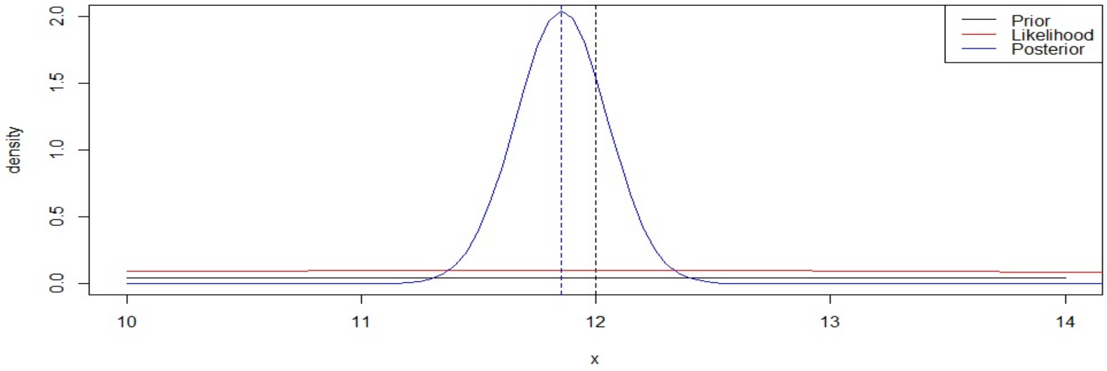


Figure 9: Prior, Likelihood and Posterior distribution plotted again in R

Observing, we notice the peak of the posterior distribution dominates the graph as the other two curves appear essentially flat. This implies a very strong certainty about our updated belief.

We can see that the Likelihood has a very similar mean to the posterior distribution. The given prior reinforces the belief that θ lies near the mean of the likelihood distribution.

e) By sending the same signal multiple times we are essentially taking the limit as follows:

$$\begin{aligned}&\text{As } n \to \infty \text{, by Rule 1 we have that :}\\ &\lim_{n\to\infty} \mu_{\text{post}} = \bar{x} = 11.85\\ &\lim_{n\to\infty} \sigma^2_{\text{post}} = 0.\end{aligned}$$

Therefore as the number of signals (n) that are sent tends to infinity, we have that $\mu|x \sim N(11.85, 0)$. Moreover, we can calculate roughly how many observations are equivalent to the prior given, by using:

$$n_0 = \frac{\sigma^2}{\sigma_0^2},$$

where $\sigma^2 = 9$ and $\sigma_0^2 = 4$ to obtain $n_0 \approx 2.25$. This essentially tells us that using the prior with $\sigma^2 = 9$ has the same effect as adding about 3 observations.

References

[1] Zach, *LaTeX: How to plot predicted values in R (with examples)*. Statology, [online] Available at https://www.statology.org/plot-predicted-values-in-r/.