7.4

Intro

In this exercise, we will extend our MLE analysis of the simple linear regression to take into account the variance (σ^2) of our prediction. More specifically, we will prove that the empirical variance is also the MLE. This is makes a lot of sense if you stop to think about it: the variance is a quadratic measure of the dispersion of the data around the mean. Meanwhile, our cost function is also a quadratic measure of the dispersion of the data around our prediction. Since we are predicting the mean of $y|x$, our cost function has a similar role to the variance.

Now let's talk briefly about the math. Due to the nature of the log likelihood function for linear regression, the addition of the variance as a parameter to be estimated will not change any of the weights w that we find by least squares. In fact, we will derive the expression for σ^2 based solely on the new equation $\frac{\partial l}{\partial \sigma} = 0$ that the MLE optimization problem introduces. Now, let's go to the solution.

Solution

The full log likelihood for the least squares (taking into account the σ^2) is expressed as:

$$l(w, \sigma^2) = -\frac{1}{2\sigma^2} RSS(w) - \frac{N}{2} log(2\pi\sigma^2) \tag{1}$$

Now, we need to derive the log likelihood with respect to the new variable and equal it to 0:

$$\begin{aligned} &\frac{\partial l(w, \sigma^2)}{\partial \sigma} = \frac{1}{\sigma^3} RSS(w) - \frac{N}{2} \frac{1}{2\pi\sigma^2} 2\pi\sigma = \\ &\frac{RSS(w)}{\sigma^3} - \frac{N}{\sigma} = 0 \\ &\hat{\sigma}^2 = \frac{RSS(w)}{N} \end{aligned} \tag{2}$$

Since the solution for $\nabla_w l(w, \sigma^2)$ remains the same ($\hat{w} = argmin RSS(w)$), our MLE for the variance becomes: $\hat{\sigma}^2 = \frac{RSS(\hat{w})}{N}$.

Conclusion

In this exercise we went further in our analysis of linear regression and proved that the expression for the MLE of the variance is the empirical variance. We already did a minor discussion of why this result makes sense in the intro. As a final commentary, notice that this empirical variance is with respect to the variable $y|x$ not y. Based on this, the goal of regression could be though as: find a set of weights w such that the predictions $y|x$ will have a smaller variance than if we just predicted $\hat{y}$. This line of reasoning is explicited encoded in the coefficient of determination R^2. Since $R^2 = 1 - \frac{\sum e_i^2}{\sum (y_i - \hat{y})^2} = 1 - \frac{\hat{\sigma}^2_{y|x}}{\hat{\sigma}^2_y}$, we see that when our regression model makes $\hat{\sigma}^2_{y|x}$ small compared to $\hat{\sigma}^2_y$, we have a good score.