MATH 392 Problem Set 9

Sam D. Olson

1. **GLM with Gaussian Response**

Consider the special case of the generalized (simple) linear model where we assume independent Gaussian errors and are linked to the linear predictor using the identify function (vanilla simple linear regression). We can summarize that model as follows (note that in the notation used in this problem, everything is a scalar):

i. If $X = x$, then $Y = \beta_0 + \beta_1 x + \epsilon$ where the β_j are (unknown) parameters and the ϵ is a random variable.
ii. $\epsilon \overset{iid}{\sim} N(0, \sigma^2)$ for some (unknown) parameter σ^2.

When this model was first derived, and only (i) was assumed, the parameters were estimated by minimizing the residual sum of squares (these are the least-squares estimates, $\hat{\beta}^{LS}$). This yielded

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$
$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$
$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^n (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x))^2$$

Now that we have added (ii), we have specified a full density function for the random variable Y, $f(y|X = x, \beta_0, \beta_1, \sigma^2)$, which enables a familiar route to estimation: maximum likelihood.

Find the maximum likelihood estimates of β_0, β_1, and σ^2.

a) Provide the derivation of closed-form solutions, if they exist. For reference, if $X \sim N(\mu, \sigma^2)$, $f(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{(x-\mu)^2}{2\sigma^2}}$.

(*):

Note: For X_i, $f(y_i|x_i, \beta_0, \beta_1, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i - (\beta_0 + \beta_1 x_i))^2}{2\sigma^2}}$

Thus, for $X_1, ..., X_n$, we may write:

$$f(y_1|x_1, \beta_0, \beta_1, \sigma^2)...f(y_n|x_n, \beta_0, \beta_1, \sigma^2) = \prod_{i=1}^n f(y_i|x_i, \beta_0, \beta_1, \sigma^2) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i - (\beta_0 + \beta_1 x_i))^2}{2\sigma^2}}$$

To ease our future calculations, we take the log-likelihood, turning the above relation into a summation instead of a product. To that end, let L denote the log, giving us:

$$L(\beta_0, \beta_1, \sigma^2) = L(f(y_1|x_1, \beta_0, \beta_1, \sigma^2)...f(y_n|x_n, \beta_0, \beta_1, \sigma^2)) = L(\prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i - (\beta_0 + \beta_1 x_i))^2}{2\sigma^2}}) = \sum_{i=1}^n log(\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i - (\beta_0 + \beta_1 x_i))^2}{2\sigma^2}})$$

We isolate each of the logs into the following relation:

(**): $L(\beta_0, \beta_1, \sigma^2) = nlog(\frac{1}{\sqrt{2\pi}}) + nlog(\frac{1}{\sqrt{\sigma^2}}) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2$

For the following derivations, we maximize the above relation with respect to our parameter of interest, which yields the maximum likelihood estimate. This involves deriving the equation (w.r.t. our parameter) and setting equal to zero, then solving for that parameter.

(1): β_0^{MLE}

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

Note, the expansion of the summation as follows:

(***): $(y_i - (\beta_0 + \beta_1 x_i))^2) = y_i^2 - 2y_i(\beta_0 + \beta_1 x_i) + (\beta_0 + \beta_1 x_i)^2 = y_i^2 - 2y_i\beta_0 - 2y_i\beta_1 x_i + \beta_0^2 + 2\beta_0\beta_1 x_i + \beta_1^2 x_i^2$

Using relation (***), we may simplify the derivation as:

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

Setting the above expression equal to zero, we may derive the MLE as:

$\beta_0^{MLE} = \bar{y}_n - \beta_1 \bar{x}_n$

(2): β_1^{MLE}

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

Using relation (***), we may simplify the derivation as:

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

Setting the above expression equal to zero, we may derive the MLE as:

$\beta_1^{MLE} = \frac{\sum_{i=1}^{n} (x_i - \bar{x}_n)(y_i - \bar{y}_n)}{\sum_{i=1}^{n} (x_i - \bar{x}_n)^2} = \frac{ss_{xy}}{ss_x}$

(3): σ^2_{MLE}

$\frac{\partial(L(\beta_0,\beta_1,\sigma^2))}{\partial\sigma^2} = \frac{\partial}{\partial\sigma^2}(-nlog(\sigma^2) - \frac{1}{\sigma^2} \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2)$

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

Setting the above equation equal to zero, we have:

$\sigma^2_{MLE} = \frac{1}{n} \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2)$

Commentary

The MLEs of β_0, β_1, and σ^2 are the same as the least squares estimates of β_0, β_1, and σ^2. We will take advantage of this when finding the MLEs using numerical optimization in part b).

b) Describe in pseudocode (or actual R code) how to find them using numerical optimization.

Pseudocode - MLE

There are a number of options available for estimating the MLE, ranging from general optimzation using 'optim', generating a generalized linear model, to dedicated maximum likelihood package such as 'maxLik'.

Code Example - MLE

```
# specify likelihood function, using gaussian assumptions
l_gauss <- function(n, sigma, B, X, Y) {
  (-(n / 2) * log(2 * pi)) - (n * log( sigma)) - ((1 / (2 * (sigma^2))) * (sum((Y - (X %*% B))^2)))
}
# Set Parameters
n <- 500
p <- 1
sigma <- 2
# Generate X
X <- cbind(1, rmvnorm(n, mean = rep(0, p), sigma = diag(p)/2))
# Set Beta values
B <- c(2, 4)
# Generate Y
Y <- rnorm(n, mean = 4, sd = sqrt(sigma))
# Optim example
optim(par = c(0, 0), fn = l_gauss, X = X, Y = Y, n = n, sigma = sigma)
```

```
## $par
## [1] -3.183805e+55  4.996335e+54
##
## $value
## [1] -6.351068e+112
##
## $counts
## function gradient
##      501       NA
##
## $convergence
## [1] 1
##
## $message
## NULL
```

```
# maxLik example
ml <- maxLik(l_gauss, start = c(0, 0), X = X, Y = Y, n = n, sigma = sigma)
ml$estimate
```

```
## [1] 4.090525475 0.006275676
```

```
# glm example
df <- data.frame(Y = Y, x1 = X[, 2])
coef(glm(Y ~ x1, data = df, family = "gaussian"))
```

```
## (Intercept)          x1
## 4.090525474 0.006275676
```

Commentary

In the above code, the estimates of the coefficients are quite off. However, the maxLik estimates and the glm estimates are similar.

---

2. **Logistic Regression MLEs: Bias, Variance, and Shrinkage**:

For this problem you'll be working in a setting when the design matrix including the intercept is an $n \times 2$ matrix X and the response is Bernoulli with an inverse logit link function to the linear predictor (logistic regression). Since there is no closed form of the MLE, you'll be using simulation, meaning you'll need to specify values for all of the parameters needed to generate data from the Logistic Regression model.

a) *How does a single estimate compare with the true mean function?* Simulate one data set and fit one model using the MLEs. Construct a scatterplot with the simulated data, the estimated mean function ($\hat{E}(Y|X = x)$) and the true mean function ($\hat{E}(Y|X = x)$).

b) *Is the MLE Biased?* Simulate many data sets and fit many models using MLE. Create a plot similar to the previous, but with *all* of the fitted models' mean functions plotted. To make this more complex plot intelligible, I recommend the gghighlight package.

c) *How does the bias of an estimate change with sample size for a particular value of the parameter?* For a single fixed value of β_1, construct a plot that shows the relationship between n and the bias of the corresponding MLE.

d) *How does the bias of an estimate change with sample size for multiple values of the parameter?* Extend the idea of the previous plot by expressing the relationship between the value of β_1 and the corresponding element of $\hat{\beta}^{MLE}$ for various fixed values of β_1. Examine this relationship at a handful of sample sizes n.

e) *Can I perform shrinkage on logistic coefficients?* The original motivation for ridge regression was to make the $X'X$ matrix in OLS invertible. Statisticians have since realized the practical value of its variance-reducing characteristics when shrinking $\hat{\beta}$ towards zero. Traditional ridge regression is performed by adding a penalty term, $\lambda \sum_{j=1}^{p} \beta_j^2$, to the RSS. In logistic regression, we instead of finding our estimates by minimizing RSS, we choose to maximize the likehood. Although the original motivation of matrix inversion is lost, it can still be perfectly valid and valuable to shink the logistic regression estimates by penalizing the likelihood.

   For the same data set that you used to create the plot in part a, find three more estimated mean functions, each one corresponding to a different value of λ, and add them to the plot. Admittedly, this will demonstrate the *downside* of penalized regression: that we have actually *increased* the bias. The plot should have five overlaid sigmoid curves. Play around with the values of λ so that you can the shape of all five on the same plot.

f) To bring in a sense of both bias and variance, select one of your values of λ and use it to replicate the plot from part b, but now with two clumps of sigmoids: one corresponding to the MLE, the other to the ridge estimates. Use color to differentiate between the two clups. Describe what the plot demonstrates about the bias and variance for the MLE and the ridge estimates in this setting.