

Assignment 5a

Satvik Saha

2024-10-01

Answer 1

Consider simple linear regression with 3 data points, where $\{x_i\}$ are equally spaced. By shifting and scaling, we may consider the case $x_1 = -1, x_2 = 0, x_3 = 1$. Then, the slope is given by

$$\hat{b} = \frac{n \sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{n \sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2} = \frac{y_3 - y_1}{2}.$$

This is independent of y_2 ! Similarly, the intercept is given by

$$\hat{a} = \frac{\sum_{i=1}^n y_i \sum_{i=1}^n x_i^2 - \sum_{i=1}^n x_i \sum_{i=1}^n x_i y_i}{n \sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2} = \frac{y_1 + y_2 + y_3}{6}.$$

- (a) The slope $b = (2 - 0)/2 = 1$ (this is unaffected by the shift in x).
- (b) The intercept can be alternatively calculated as the mean of $y_i - \hat{b}x_i$ (recall the discussion on regression on a constant), hence we calculate $a = (0 + 1 + 0)/3 = 1/3$.
- (c) Since the x_i 's are spaced by 50 years, we scale our slope down, giving $b = (6 - 1.5) \times 10^9 / (2 \times 50) = 4.5 \times 10^7$ people per year. For the intercept, we calculate the mean $a = (1.5 + 2.5 + 6) \times 10^9 / 3 - b \times 1950 = (10/3 - 87.75) \times 10^9 \approx 84.42 \times 10^9$ people.

Answer 2

- (a) We prepare the data as follows.

```
df <- read.table("hibbs.dat", header = TRUE)
df$more <- as.numeric(df$growth > 2.0)
head(df)
```

##	year	growth	vote	inc_party_candidate	other_candidate	more
## 1	1952	2.40	44.60	Stevenson	Eisenhower	1
## 2	1956	2.89	57.76	Eisenhower	Stevenson	1
## 3	1960	0.85	49.91	Nixon	Kennedy	0
## 4	1964	4.21	61.34	Johnson	Goldwater	1
## 5	1968	3.02	49.60	Humphrey	Nixon	1
## 6	1972	3.62	61.79	Nixon	McGovern	1

The average difference in vote share has a standard error which can be estimated by computing the variances in each group, then combining them.

```
df.less <- df[df$more == 0, ]
df.more <- df[df$more == 1, ]

diff.hat <- mean(df.more$vote) - mean(df.less$vote)
diff.se <- sqrt(sd(df.more$vote)^2/nrow(df.more) + sd(df.less$vote)^2/nrow(df.less))
```

```
diff.hat
```

```
## [1] 5.5075
```

```
diff.se
```

```
## [1] 2.502052
```

(b) We perform the linear fit using `lm` to eliminate the effect of priors, regularization, etc.

```
diff.fit <- lm(vote ~ more, data = df)
summary(diff.fit)
```

```
##
```

```
## Call:
```

```
## lm(formula = vote ~ more, data = df)
```

```
##
```

```
## Residuals:
```

```
##      Min       1Q   Median       3Q      Max
## -10.2088  -3.3706   0.1287   3.3038   6.9812
```

```
##
```

```
## Coefficients:
```

```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   49.301      1.769   27.866 1.15e-13 ***
## more          5.508      2.502    2.201  0.045 *
```

```
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
```

```
## Residual standard error: 5.004 on 14 degrees of freedom
```

```
## Multiple R-squared:  0.2571, Adjusted R-squared:  0.204
```

```
## F-statistic: 4.845 on 1 and 14 DF,  p-value: 0.045
```

The estimate for the slope and its standard error agrees with our calculation precisely.

Research homework assignment

Consider the model $y_i \sim \theta + \epsilon_i$, with $\epsilon_i \stackrel{iid}{\sim} N(0, \sigma)$. With just this setup, we may use the OLS estimate $\hat{\theta} = \bar{y} = \frac{1}{n} \mathbf{1}^\top \mathbf{y}$, hence our predictions are $\hat{\mathbf{y}} = \frac{1}{n} \mathbf{1} \mathbf{1}^\top \mathbf{y} = H \mathbf{y}$. The effective number of parameters here becomes $\text{trace}(H) = 1$. Note that σ here is a nuisance parameter, which we are not interested in. In the MLE setup, where both θ and σ play a role, it would be natural to consider 2 parameters.

Now, add the restriction that $\theta \in (0, 1)$; perhaps this can be achieved by imposing a prior on θ supported on $[0, 1]$, say a uniform one. With a prior $\pi(\sigma^2) \propto 1/\sigma^2$, we may compute the Bayes estimate for θ , which under the squared error loss is the posterior mean of θ ,

$$\hat{\theta}(\mathbf{y}) = \frac{\int_0^1 \int_0^\infty \theta \cdot (2\pi\sigma^2)^{-n/2} e^{-\sum_i (y_i - \theta)^2 / 2\sigma^2} \cdot (\sigma^2)^{-1} d\sigma^2 d\theta}{\int_0^1 \int_0^\infty (2\pi\sigma^2)^{-n/2} e^{-\sum_i (y_i - \theta)^2 / 2\sigma^2} \cdot (\sigma^2)^{-1} d\sigma^2 d\theta} = \frac{\int_0^1 \theta \cdot \left(1 + \frac{(\bar{y} - \theta)^2}{\frac{1}{n} \sum_i (y_i - \bar{y})^2}\right)^{-n/2} d\theta}{\int_0^1 \left(1 + \frac{(\bar{y} - \theta)^2}{\frac{1}{n} \sum_i (y_i - \bar{y})^2}\right)^{-n/2} d\theta}.$$

Thus, $\hat{\mathbf{y}} = \mathbf{1} \hat{\theta}(\mathbf{y})$. To obtain the analogous version of effective number of parameters, we would need to compute

$$n_{\text{eff}} = \sum_i \frac{\partial}{\partial y_i} \hat{\theta}(\mathbf{y}).$$

This setup does not admit nice computations!

Another estimator involves simply clamping \bar{y} within $(0, 1)$, via

$$\hat{\theta}(\mathbf{y}) = \begin{cases} 0, & \text{if } \bar{y} < 0, \\ \bar{y}, & \text{if } \bar{y} \in [0, 1], \\ 1, & \text{if } \bar{y} > 1. \end{cases}$$

With this interpretation, we have the awkward

$$n_{\text{eff}} = \sum_i \frac{\partial}{\partial y_i} \hat{\theta}(\mathbf{y}) = \hat{\theta}(\mathbf{y}) = \begin{cases} 1, & \text{if } \bar{y} \in [0, 1], \\ 0, & \text{otherwise.} \end{cases}$$

Of course, this kind of estimator that clamps \bar{y} is very different from a regularized estimator, where n_{eff} would be some positive number between 0 and 1.

Intuitively, it ought to make sense that restricting θ to an interval, say $(-\epsilon, \epsilon)$, should reduce the number of effective parameters with decreasing ϵ . If this restriction is modeled using a uniform prior on $(-\epsilon, \epsilon)$, this would degenerate to a δ_0 prior, which completely eliminates the role of the parameter θ . However, without a broader context with which to judge the ‘amount of freedom a parameter enjoys’, it becomes difficult to quantify this with actual numbers. We may be able to make comparisons between models with restrictions of this kind, but this reasoning does not hold for more general priors.