

End Term

Saket Choudhary

December 9, 2015

Problem 1

Define the likelihood function: $L(\theta|N, n, k)$ Then,

$$L(\theta|N, n, k) = \frac{\binom{\theta}{k} \binom{N-\theta}{n-k}}{\binom{N}{n}}$$

We find the MLE using first principle.

In order to ensure MLE, we need to ensure: $\frac{L(\theta|N, n, k)}{L(\theta-1|N, n, k)} > 1$

$$\begin{aligned} \frac{L(\theta|N, n, k)}{L(\theta-1|N, n, k)} &> 1 \\ \frac{\binom{\theta}{k} \binom{N-\theta}{n-k}}{\binom{N}{n}} &> 1 \\ \frac{\binom{\theta-1}{k} \binom{N-\theta+1}{n-k}}{\binom{N}{n}} &> 1 \\ \frac{\theta(N-\theta+1-(n-k))}{(\theta-k)(N-\theta+1)} &> 1 \\ \theta n &> -kN - k \\ \theta &< \frac{k(N+1)}{n} \end{aligned}$$

Similarly, $\frac{L(\theta|N, n, k)}{L(\theta+1|N, n, k)} > 1$

$$\begin{aligned} \frac{L(\theta|N, n, k)}{L(\theta+1|N, n, k)} &> 1 \\ \frac{(\theta-k+1)(N-\theta)}{(\theta+1)(N-\theta-(n-k))} &> 1 \\ \theta &> \frac{k(N+1)}{n} - 1 \end{aligned}$$

Thus, $\frac{k(N+1)}{n} - 1 < \theta < \frac{k(N+1)}{n}$

and hence a valid choice for MLE is $\hat{\theta} = \lfloor \frac{N(k+1)}{n} \rfloor$

Part (b)

N=19, n=4, k=3

```
N <- 10
n <- 4
k <- 3
theta <- ceiling(k*(N+1)/n)
```

$$\hat{\theta} = 9$$

To find one p-value for $H_\theta : \theta = 4$ we need to calculate: $\sum_{\theta=4}^9 Pr(X = k)$

```
#theta = seq(4,9,1);
#s <- sapply(theta, function(x) choose(x,3) * choose(10-x,1)/(choose(10,4)))
k <- seq(3,4,1);
s <- sapply(theta, function(x) choose(4,k) * choose(10-4,k)/(choose(10,4)) )
sum(s)
```

```
## [1] 0.452381
```

Thus, one sided p-value is **0.452381**

Problem 2

$Y \sim \text{Binomial}(n, \pi)$ and $\hat{\pi} = \frac{Y}{n}$ Define $g(\pi) = \log \frac{\pi}{1-\pi}$ Then by delta method $E[g(\pi)] = g(E[\hat{\pi}])$

$$\begin{aligned} E[g(\hat{\pi})] &= g(E[\hat{\pi}]) \\ &= \log \frac{\hat{\pi}}{1 - \hat{\pi}} \\ &= \log \frac{\frac{Y}{n}}{1 - \frac{Y}{n}} \\ &= \log \frac{Y}{n - Y} \end{aligned}$$

$$\text{Var}(g(\hat{\pi})) = g'(\pi)^2 \text{Var}(\hat{\pi}) = E[g(\hat{\pi})^2] - (E[g(\hat{\pi})])^2$$

Thus mean square error is given by

$$E[g(\hat{\pi})^2] = g'(\pi)^2 \text{Var}(\hat{\pi}) + (E[g(\hat{\pi})])^2$$

where $\text{Var}(\hat{\pi}) = \hat{\pi}(1 - \hat{\pi}) = \frac{Y}{n}(1 - \frac{Y}{n})$ and $g'(\hat{\pi}) = \frac{1}{\hat{\pi}} + \frac{1}{1-\hat{\pi}} = \frac{1}{\hat{\pi}(1-\hat{\pi})}$

Thus, $\text{Var}(\hat{\pi}) = \frac{1}{\hat{\pi}^2(1-\hat{\pi})^2} \times \hat{\pi}(1 - \hat{\pi}) = \frac{1}{\hat{\pi}(1-\hat{\pi})}$

Thus, $E[g(\hat{\pi})^2] = \frac{1}{\hat{\pi}(1-\hat{\pi})} + \log^2 \frac{\hat{\pi}}{1-\hat{\pi}}$

Problem 3

A possible model to describe the given two-way factorial experiment would be two-way factorial ANOVA. Since there are several treatment groups, we will test for within group versus between group variance for the two independent factors (and their interaction)

Problem 4

Part (a)

In this part, the variables X_1 and X_2 are correlated. The main difficulty that would potentially arise would be in interpreting the coefficients associated with these variables. Coefficients by definition imply the amount by which the mean response changes when all other covariates are held fixed. However in this case since X_1 and X_2 are highly correlated, changing one would also imply changing the other.

In order to overcome this, we need to choose either of X_1 or X_2 as a covariate in the linear regression model (essentially discarding the other) based on which one of these predictors best captures the ‘reality’ of the independent variable.

Part (b)

There exist two outliers one on the top right and one on the bottom left. They are outliers considering (X_1, X_2) together but not individually. Since we are minimizing the squared error to determine the coefficients, the presence of such outlier points will affect the coefficients (and they will turn out to be smaller in magnitude). One way to overcome such outliers is to either use regularization (L1, L2) or to completely neglect such outliers.

Problem 5

Given $\log \frac{p}{1-p} = 3.2 - 0.078 \times \text{age}$ for men and $\log \frac{p}{1-p} = 1.6 - 0.078 \times \text{age}$ for women where p denotes the probability of survival.

Consider

$$\begin{aligned}\log \frac{p}{1-p} &= \beta_0 + \beta_1 x \\ \frac{1-p}{p} &= \exp(-(\beta_0 + \beta_1 x)) \\ p &= \frac{1}{1 + \exp(-(\beta_0 + \beta_1 x))}\end{aligned}$$

```
pman <- 1/(1+exp(-(3.2-0.078*25)))
pwoman <- 1/(1+exp(-(1.6-0.078*50)))
```

Thus, Estimated probability of survival of man of age 25 = 0.7772999 and of woman aged 50 = 0.091123
Age at which probability of survival is 0.5:

$$\begin{aligned}\beta_0 + \beta_1 x &= \log(0.5/0.5) \\ \hat{x} &= \frac{\beta_0}{\beta_1}\end{aligned}$$

Thus, man's age at which probability of surviving is 0.5: 41.025641 and for woman: 20.5128205

Problem 6

```
library(car)
problem6 <- read.csv('problem6.csv')
model <- lm(y~x1+x2, problem6)
vif(model)
```

```
##           x1           x2
## 1.666667 1.666667
```

We first consider *VIF1* $X_1 = \beta_0 + \beta_1 X_2$ $\bar{X}_1 = 2.5$ and $\bar{X}_2 = 0$ Also, $\sum_i (X_{2i} - \bar{X}_2)^2 = (-1 - 0)^2 + (0 - 0)^2 + (1 - 0)^2 + (0 - 0)^2 = 2$

$$\beta_1 = \frac{\sum_i (X_{1i} - \bar{X}_1)(X_{2i} - \bar{X}_2)}{\sum_i (X_{2i} - \bar{X}_2)^2} = \frac{(1-2.5)(-1-0) + (2-2.5)(0) + (3-2.5)(1-0) + (4-2.5)(0)}{2} = 1$$

$$\beta_0 = \bar{X}_1 - \beta_1 \bar{X}_2 = 2.5 - 0 = 2.5$$

$$X_2 = 2.5 + X_1$$

$$SS_{Total} = \sum_i (X_{1i} - \bar{X}_1)^2 = (1 - 2.5)^2 + (2 - 2.5)^2 + (3 - 2.5)^2 + (4 - 2.5)^2 = 5$$

$$SS_{Res} = (1 - 2.5)^2 + (2 - 2.5)^2 + (3 - 2.5)^2 + (4 - 2.5)^2 = 3$$

$$\text{Thus } R_1^2 = 1 - SS_{Res}/SS_{Tot} = 1 - 3/5 = 0.4$$

$$\text{And } VIF1 = 1/(1 - R_1^2) = 1.6667$$

Similarly for *VIF2* we consider $X_2 = \beta_0 + \beta_1 X_1$

$$\sum_i (X_{1i} - \bar{X}_1)^2 = 5 \text{ and hence } \beta_1 = \frac{2}{5} \text{ Thus, } \beta_0 = \bar{X}_2 - \beta_1 \bar{X}_1 = 0 - 2.5 * 2/5 = -1$$

$$\text{Thus, } X_2 = -1 + 0.4X_1$$

$$SS_{Tot} = \sum_i (X_{2i} - \bar{X}_2)^2 = 2$$

$$\text{And } SS_{Res} = (-1 + 0.6)^2 + (0 + 0.2)^2 + (1 - 0.2)^2 + (0 - 0.6)^2 = 1.2$$

$$\text{and hence } R_2^2 = 1 - 1.2/2 = 0.4 \text{ and } VIF2 = 1/(1 - R_2^2) = 1.667$$

$$\begin{aligned} \frac{SE(\beta_2)}{SE(\beta_1)} &= \frac{\sigma \sqrt{\frac{1}{(n-1)s_{X_2}^2}}}{\sigma \sqrt{\frac{1}{(n-1)s_{X_1}^2}}} \\ &= \frac{s_{X_1}}{s_{X_2}} \\ &= \sqrt{\frac{2}{5}} \\ &= 0.632455 \end{aligned}$$

Part (b)

```
c <- summary(model)
s <- c$coefficients
SE1 <- s[2,2]
SE2 <- s[3,3]
b1 <- s[2,1]
b2 <- s[3,1]
```

Problem 7

$$\begin{aligned}\mu_{11} &= \mu + \alpha_1 + \gamma_1 + \theta_{11} \\ \mu_{12} &= \mu + \alpha_1 + \gamma_2 + \theta_{12} \\ \mu_{21} &= \mu + \alpha_2 + \gamma_1 + \theta_{21} \\ \mu_{22} &= \mu + \alpha_2 + \gamma_2 + \theta_{22} \\ \mu_{31} &= \mu + \alpha_3 + \gamma_1 \\ \mu_{32} &= \mu + \alpha_3 + \gamma_2\end{aligned}$$

Applying constraints $\sum_i \alpha_i = 0$ we get :

$$\begin{aligned}\alpha_1 + \alpha_2 + \alpha_3 &= 0 \\ 2 - 3 + \alpha_3 &= 0 \\ \alpha_3 &= 1\end{aligned}$$

Similarly $\sum_i \gamma_i = 0$

$$\begin{aligned}\gamma_1 + \gamma_2 &= 0 \\ 5 + \gamma_2 &= 0 \\ \gamma_2 &= -5\end{aligned}$$

Now, $\sum_j \theta_{ij} = 0 \forall i$

$$\begin{aligned}\theta_{11} + \theta_{12} &= 0 \\ \theta_{12} &= -4\end{aligned}$$

Similarly, $\sum_i \theta_{ij} = 0 \forall j$

$$\begin{aligned}\theta_{11} + \theta_{21} &= 0 \\ \theta_{21} &= -4\end{aligned}$$

Similarly $\theta_{22} = 4$

Thus,

$$\begin{aligned}\mu_{11} &= \mu + \alpha_1 + \gamma_1 + \theta_{11} = 12 \\ \mu_{12} &= \mu + \alpha_1 + \gamma_2 + \theta_{12} = -6 \\ \mu_{21} &= \mu + \alpha_2 + \gamma_1 + \theta_{21} = -1 \\ \mu_{22} &= \mu + \alpha_2 + \gamma_2 + \theta_{22} = -3 \\ \mu_{31} &= \mu + \alpha_3 + \gamma_1 = 7 \\ \mu_{32} &= \mu + \alpha_3 + \gamma_2 = 3\end{aligned}$$

Problem 8

Given model: $T = A(t - t_0)^p z^q$; $T(t, z)$

Part (a)

To estimate A, p and q, we consider the log transformed model: $\log(Z) = \log(A) + p \log(t - t_0) + q \log(z)$, this is essentially a linear regression model of the following form : $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$ where:

$$\begin{aligned} Y &= \log(T) \\ \beta_0 &= \log(A) \\ \beta_1 &= p \\ X_1 &= \log(t - t_0) \\ \beta_2 &= q \\ X_2 &= \log(z) \end{aligned}$$

and the coefficients $(\beta_0, \beta_1, \beta_2)^T = (X^T X)^{-1} X^T Y$

Part (b)

To test the hypothesis that temperature does not change in time for each depth, we need to check for (in the above restated log model) $H_0 : \beta_1 = 0$; $H_1 : \beta_1 \neq 0$ or alternatively $H_0 : p = 0$; $H_1 : p \neq 0$

Part (c)

To test the hypothesis that temperature does not depend on both time and depth we need to do the following test in lieu of the above model:

$H_0 : \beta_1 = 0, \beta_2 = 0$ and $H_1 : \beta_1 \neq 0, \beta_2 \neq 0$ alternatively $H_0 : p = 0, q = 0$ and $H_1 : p \neq 0; q \neq 0$

Part (d)

She can use the model for prediction of Temperature conditions in a given region for 1 year, since there is autocorrelation and hence the next year's prediction will depend on the current year's parameters (which are known).

Part (e)

Prediction for 20 years will not be accurate since the data for next year depends on knowing the actual parameters for the current year (since the temperature readings are correlated)

Part (f)

This model cannot be used to make predictions in other parts of North America accurately since the model used to infer the coefficients used z_i from a specific area. The model was trained using data from a specific area and hence cannot necessarily be generalised.