# Qualifying Exam: CAS MA575, Linear Models

Boston University, Fall 2016

1. Consider the linear regression model

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e},$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top \in \mathbb{R}^p$ is the regression coefficient to be estimated. Instead of using ordinary least squares, an alternative is to consider the ridge estimator

$$\tilde{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I}_{p\times p})^{-1} \boldsymbol{X}^\top \boldsymbol{Y},$$

where $\lambda \geq 0$ is a tuning parameter prespecified by the user and $\boldsymbol{I}_{p\times p}$ is the identity matrix in $\mathbb{R}^{p\times p}$. Throughout this problem, assume that the design matrix $\boldsymbol{X}$ is deterministic and that the errors are independent normal random variables with mean zero and variance $\sigma^2 > 0$.

   (a) Find $E(\tilde{\boldsymbol{\beta}})$. When is $\tilde{\boldsymbol{\beta}}$ an unbiased estimator of $\boldsymbol{\beta}$?

   (b) Given a data, explain the behavior of $\tilde{\boldsymbol{\beta}}$ when $\lambda \to \infty$, namely when an extremely large tuning parameter is used.

   (c) Find the covariance matrix of $\tilde{\boldsymbol{\beta}}$.

   (d) Suppose $\sigma^2 = 1$ is known and one is interested in testing if $\beta_1 = 0$. Is it possible to achieve this by using the ridge estimator $\tilde{\boldsymbol{\beta}}$? Explain your strategy in detail (including the form of your test statistic and the distribution that you will use to obtain the cut-off value).

   (e) Suppose $\sigma^2 = 1$ is known and one is interested in testing if $\beta_1 = \beta_2 = \cdots = \beta_p = 0$. Is it possible to achieve this by using the ridge estimator $\tilde{\boldsymbol{\beta}}$? Explain your strategy in detail (including the form of your test statistic and the distribution that you will use to obtain the cut-off value).

2. Consider the following data

| *Degree* | *Response* |
|---|---|
| Undergraduate | 37.7, 32.5, 34.1 |
| Master | 27.8, 22.7, 31.6, 36.5, 41.3 |
| Doctoral | 38.2, 44.6, 35.4, 33.7, 40.2 |

The associated `R` output of regressing the response on the degree variable is given below.

```
Call:
lm(formula = Y ~ Degree)

Residuals:
   Min     1Q Median     3Q    Max 
-9.280 -3.020 -0.380  2.933  9.320 

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)    
(Intercept)              38.420      2.441  15.737  2.2e-08 ***
DegreeMaster             -6.440      3.453  -1.865   0.0917 .  
DegreeUndergraduate      -3.653      3.987  -0.916   0.3810    
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 5.459 on 10 degrees of freedom
Multiple R-squared:  0.2589,    Adjusted R-squared:  0.1107 
F-statistic: 1.747 on 2 and 10 DF,  p-value: 0.2235
```

Let

$$\begin{aligned}\beta_1 &= E(\text{Response} \mid \text{Degree} = \text{Undergraduate})\\ \beta_2 &= E(\text{Response} \mid \text{Degree} = \text{Master})\\ \beta_3 &= E(\text{Response} \mid \text{Degree} = \text{Doctoral})\end{aligned}$$

Throughout this problem, assume that the errors are independent normal random variables with mean zero and variance $\sigma^2 > 0$. Also assume that the errors are independent of the variable `Degree`.

(a) Find the least squares estimate $\hat{\beta}_1$ for $\beta_1$.

(b) Find the least squares estimate $\hat{\xi}$ for $\xi = \beta_2 - \beta_1$. Is it unbiased? Prove.

(c) Suppose one is interested in the squared difference $\zeta = (\beta_2 - \beta_1)^2$. Let $\hat{\xi}$ be as in part (b), is $\hat{\xi}^2$ an unbiased estimate of $\zeta$? Prove.

(d) If $\hat{\xi}^2$ is not an unbiased estimate of $\zeta$, provide a way to correct for the bias and give an unbiased estimate of $\zeta$.

(e) Is it possible to provide an unbiased estimate of $\gamma = |\beta_2 - \beta_1|$? If yes, describe your strategy in achieving this. Note that $\gamma$ can be interpreted as the absolute difference between `Undergraduate` and `Master`.