BST 140.751
Problem Set 1

1 Regression and background

1. If your normalize (mean 0 and variance 1) your Y and X vectors, argue that the regression slope estimate is the correlation.

2. Let Y and X be one dimensional vectors of length n. Give the relationship between the slope from regressing Y on X and X on Y.

3. Consider the residuals after mean only regression. Argue that they sum to 0.

4. Consider the residuals after regression through the origin. Argue that they are orthogonal to the regressor.

5. Consider the residuals from ordinary linear regression. Argue that the residuals are orthogonal to both J_n and X.

2 Least squares

1. Show that $I - H$ is an idempotent matrix where H is idempotent.

2. Let $X = [X_1 X_2]$ be an $n \times 2$ design matrix and consider

$$||Y - X\beta||^2$$

   where $\beta = (\beta_1 \beta 2)'$. Show that $\hat{\beta}_2$ can be obtained by taking the residuals after regressing X_1 out of Y and X_2 then doing regression through the origin on the residuals.

3. Argue that X, X', $X'X$ and XX' all have the same matrix rank.

4. Suppose that X is such that $X'X = I$. Find the associated least squares estimate of β.

5. Suppose that the design matrix is of the form $J_A \otimes I$ where J_A is a vector of length A and I is a $B \times B$ identity matrix. Let Y be a $AB \times 1$ length vector. Find the least squares estimates associated with this design matrix.

3 Linear models

1. Let $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$ where $\epsilon_i \sim N(0, \sigma^2)$ for $i = 1, \ldots, n$.

   A. Derive the MLEs for β_0, β_1 and σ^2.

   B. Relate β_1 to the correlation between Y_i and X_i.

C. Suppose that you standardize (i.e. take $(Y_i - \bar{Y})/S_y$ and $(X_i - \bar{X})/S_X$) X_i and Y_i. Derive the estimates of β_0 and β_1.

2. Let $Y_{ij} = \alpha_0 + \beta_j + \epsilon_{ij}$ for $i = 1, \ldots, I$ and $j = 1, \ldots, J$.

   A. Write out the design matrix for the associated linear model.

   B. Show what the estimates are under the following constraints:

      i. $\alpha_0 = 0$
      ii. $\beta_1 = 0$
      iii. $\beta_J = 0$
      iv. $\sum_{j=1}^{J} \beta_j = 0$

3. Let Σ be a known matrix. Consider the model $Y = X\beta + \epsilon$ where $\epsilon \sim N(0, \Sigma)$. Derive the ML estimate of β.

4. Let P be a rotation matrix and consider the model $Y = X\beta + \epsilon$ where $\epsilon \sim N(0, \sigma^2 I)$. Suppose someone gave you the ML estimates for $\tilde{\beta}$ and $\tilde{\sigma}^2$ from fitting the model $\tilde{Y} = \tilde{X}\tilde{\beta} + \tilde{\epsilon}$ where $\tilde{Y} = PY$ and $\tilde{X} = PX$ and $\tilde{\epsilon} \sim N(0, \tilde{\sigma}^2)$. Relate these estimates to the ML estimates of β and σ^2.

5. Let $Y \mid \beta \sim N(X\beta, \sigma^2 I)$ and $\beta \sim N(\beta 0, \tau^2 I)$. What is the posterior distribution of β?

6. Consider the model $Y = X\beta + \epsilon$. Let F be an invertible $p \times p$ matrix and $\tilde{X} = XF$.

   A. Consider another model $Y = \tilde{X}\tilde{\beta} + \epsilon$. Argue that the models are equivalent.

   B. Show that the least squares estimate of $\tilde{\beta}$ from the second model is $F^{-1}\hat{\beta}$ where $\hat{\beta}$ is the least squares estimate from the first model.

   C. Suppose that you have a linear regression equation where one of the regressors is temperature. Use the results above to relate the beta coefficients if the regressor is input as Celsius or Fahrenheit.

7. Consider a linear model with iid errors $N(0, \sigma^2)$ errors. Show that $\frac{1}{n-p} e'e$, where e is the vector of residuals, is the ML estimate of σ^2. Further show that this estimate is unbiased.

   A. Argue that $\frac{1}{\sigma^2}(y - X\beta)'(y - X\beta)$ is χ^2_n

   B. Argue that $\frac{1}{\sigma^2} e'e$ is χ^2_{n-p}.

   C. Argue that $\frac{1}{\sigma^2}(y - X\beta)'X(X'X)^{-1}X'(y - X\beta)$ is χ^2_p.

   D. In each of the above cases, use the expected value calculation for quadratic forms to verify that the expected values equals the Chi squared df.

4 Multivariate means, variances and normals

1. Let X be a multivariate vector with mean μ. Show that $E[AX + b] = A\mu + b$.

2. Consider the previous problem; assume that Var$(X) = \Sigma$. Show that Var$(AX + b) = A\Sigma A'$.

3. Show that $E[(X - \mu)(X - \mu)'] = E[XX'] - \mu\mu'$.

4. Argue that Var(X) is non-negative definite.

5. Let $C(X, Y)$ be the multivariate covariance function, $E[(X - \mu_x)(Y - \mu_y)']$. Show that $C(X, Y) = E[XY'] - \mu_X \mu_y'$.

6. Show that $C(X_1 + X_2, Y) = C(X_1, Y) + C(X_2, Y)$.

7. Argue that $C(X, Y) = C(Y, X)'$.

8. Argue that $Var(X + Y) = Var(X) + C(X, Y) + C(Y, X) + V(Y)$.

9. Argue that $C(AX, BY) = AC(X, Y)B'$.

10. Let $X \sim N(0, I)$. Argue that $aX/\sqrt{a'a} \sim N(0, 1)$ for any non-zero vector a.

11. Let $X \sim N(0, I)$. Argue that if $AA' = I$ then $AX \sim N(0, I)$. Argue geometrically why this occurs.

12. Let X_i for $i = 1, \ldots, I$ be iid k dimensional vectors from a distribution with mean μ and variance Σ. What is the mean and variance of the multivariate pointwise sample average of the vectors?

13. Let X_i be iid k dimensional vectors from a distribution with mean μ and variance Σ. Give an unbiased estimate of Σ when μ is known.

14. Consider a covariance matrix that is of the form

$$\sigma^2 \mathbf{I} + \theta \mathbf{1}\mathbf{1}'$$

    where σ^2 and θ are positive constants and $\mathbf{1}$ is a vector of ones. Argue that this matrix describes random vectors where every pair of elements of the vector are equally correlated and every elemant has the same variance. Give this correlation and variance.

15. Let $X = (X_1' \; X_2')' \sim N(\mu, \Sigma)$.

    A. Derive the marginal distribution of X_1

    B. Derive the conditional distribution $X_1 \mid X_2$.

16. Let $X \mid \mu \sim N(\mu, \Sigma)$ and $\mu \sim N(\alpha, \tau I)$. Derive the distribution of $\mu \mid X$.

17. Argue that if $Y \sim N(\mu, \Sigma)$, the quadratic form $(Y - \mu)'\Sigma^{-1}(Y - \mu)$ is χ^2_p.

5 Computing and analysis

1. Write an R function that takes a Y vector and X matrix and obtains the least squares fit for the associated linear model.

2. Write and R function that takes an $n \times p$ data matrix, X, and "whitens" it via subtracting out a mean and multiplying by a matrix so that the resulting matrix has (p) sample column means of 0 and $p \times p$ sample covariance matrix of I.

3. You collect 100 blood pressure measurements from a population. Twenty one exhibit high blood pressure. Using a beta prior, plot the posteriors assuming: Beta(1, 1), Beta(2, 2), Beta(.5, .5). Calculate 95% equi-tail and HPD credible intervals.

4. You now collect 100 measurements from an otherwise similar population with different diets; 15 have high blood pressure. Plot the posteriors for the relative risk, risk difference and risk ratio. Assume Beta(1,1) priors for both populations.

5. One nuclear reactor test failed 51 times after having been monitored for $1,000$ days. Assuming a Gamma(.1,.1) prior and a Poisson model, plot the posterior for the failure rate. Give a credible interval (HPD and equi-tail).

6. A second reactor failed 25 times for 600 monitoring days. Plot the posterior and give credible intervals (HPD and equi-tail) for the relative rate with the other reactor.