STAT 151A: Lab 7
Midterm Review

Stephanie DeGraaf

October 5, 2018

1. **Normal Regression Theory**
   Suppose $X \in \mathbb{R}^{n\times p}$ and $X^T X$ is invertible. Consider the normal linear model $Y = X\beta + e$, where $\beta \in \mathbb{R}^p$ and $e \sim N(0, \sigma^2 I_n)$. Let $\hat{\beta}$ be the least squares solution.

   (a) Give the explicit form of $\hat{\beta}$. Show that $\hat{\beta}$ is a multivariate normal with $\mathbb{E}(\hat{\beta}) = \beta$ and $cov(\hat{\beta}) = \sigma^2(X^T X)^{-1}$.
   The least squares solution is $\hat{\beta} = (X'X)^{-1}X'Y$. Since Y is a linear function of the multivariate normal ϵ, Y has the multivariate normal distribution $N(X\beta, \sigma^2 I_n)$. Since $\hat{\beta}$ is a linear transformation of Y, it is also a multivariate normal. The parameters of the distribution can be calculated as

   $$\mathbb{E}\hat{\beta} = \mathbb{E}((X'X)^{-1}X'Y) = \mathbb{E}((X'X)^{-1}X'X\beta) = \mathbb{E}\beta = \beta$$

   and

   $$\begin{aligned} var(\hat{\beta}) = var((X'X)^{-1}X'Y) &= (X'X)^{-1}X'var(Y)((X'X)^{-1}X')' \\ &= \sigma^2(X'X)^{-1}X'X(X'X)^{-1\prime} \\ &= \sigma^2(X'X)^{-1}. \end{aligned}$$

   (b) Let $\hat{Y} = X\hat{\beta}$ and $\hat{e} = Y - \hat{Y}$. Show that $\hat{Y} \sim N(X\beta, \sigma^2 H)$ and $\hat{e} \sim N(0, \sigma^2(I - H))$, where $H = X(X^T X)^{-1}X^T$, the hat matrix.
   From part (a), $\hat{\beta}$ is a multivariate normal, and since $\hat{Y}$ is a linear transformation of $\hat{\beta}$, it is also a multivariate normal. Similarly, since Y and $\hat{Y}$ are multivariate normals, $\hat{e}$ is also multivariate normal. The parameters for $\hat{Y}$ can be derived by

   $$\mathbb{E}\hat{Y} = \mathbb{E}(X\hat{\beta}) = X\mathbb{E}\hat{\beta} = X\beta,$$
   $$var(\hat{Y}) = var(X\hat{\beta}) = Xvar(\hat{\beta})X' = X\sigma^2(X'X)^{-1}X' = \sigma^2 H.$$

   The parameters for $\hat{e}$ can be derived by noting that $\hat{e} = (I - H)Y$, so

   $$\mathbb{E}\hat{e} = \mathbb{E}(Y - \hat{Y}) = \mathbb{E}Y - \mathbb{E}\hat{Y} = X\beta - X\beta = 0,$$

and

$$\begin{aligned} var(\hat{e}) = var((I-H)Y) &= (I-H)var(Y)(I-H)' \\ &= \sigma^2(I-H)(I-H)' \\ &= \sigma^2(I-H)^2 \\ &= \sigma^2(I-H), \end{aligned}$$

from the fact that $(I-H)$ is a projection matrix (as shown in the next parts).

(c) Show that H and $I-H$ are idempotent.
By definition,

$$H^2 = X(X^TX)^{-1}X^TX(X^TX)^{-1}X^T = XI_n(X^TX)^{-1}X^T = X(X^TX)^{-1}X^T = H,$$

and therefore,

$$(I-H)^2 = (I-H)(I-H) = I - 2H + H^2 = I - 2H + H = I - H.$$

(d) Show that $\hat{Y}$ and $\hat{e}$ are independent.
$\hat{Y}$ and $\hat{e}$ are jointly normal because they can be expressed in the form of $\hat{Y}$ and Y, which are themselves independent normal random variables (because X and Y are independent). Thus, $\hat{Y}$ and $\hat{e}$ are independent if their covariance is equal to zero.

$$\begin{aligned} cov(\hat{Y}, \hat{e}) &= cov(HY, (I-H)Y) \\ &= HCov(Y, Y)(I-H) \\ &= \sigma^2 H(I-H) \\ &= \sigma^2(H - H^2) = \sigma^2(H-H) = 0. \end{aligned}$$

Therefore, $\hat{Y}$ and $\hat{e}$ are independent.

2. **Simple Linear Regression**
Suppose we want to predict the longevity of some insects based on their thorax size. We observe 10 insects, measure their thorax sizes, and record their life span. Then we fit a simple linear regression of their longevities $y = (y_1, ..., y_{10})$ on their thorax sizes $x = (x_1, ..., x_{10})$. We computed the following information:

$$\begin{aligned} \bar{x} = 0.752 \qquad s_x = 0.064 \\ \bar{y} = 49.2 \qquad s_y = 9.4. \\ r = 0.7067. \end{aligned}$$

(a) Report the least squares estimates of the intercept and slope.

$$\hat{\beta}_1 = r(s_y/s_x) = 0.7067 * (9.4/0.064) = 103.7966$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} = -28.85504.$$

(b) Report the total sum of squares, TSS.

$$TSS = \sum (y_i - \bar{y})^2 = (n-1)s_y^2 = 9(9.4)^2 = 795.$$

(c) Report the regression sum of squares, RegSS.

$$\begin{aligned} R^2 = (0.7067)^2 = 0.4994249 &= RegSS/TSS = RegSS/795 \\ \Rightarrow RegSS &= 795 * 0.49942494 = 397.0428. \end{aligned}$$

(d) Report the residual standard error.

$$RSE = \sqrt{\frac{RSS}{(n-p-1)}} = \sqrt{\frac{TSS - RegSS}{(n-p-1)}} = \sqrt{\frac{795 - 397.0428}{8}} = 7.052989$$

(e) Explain the assumption of homoskedastic errors and how you would assess the validity of this assumption.

We assume that the variability in y is constant across x values. We could assess this with a plot of the residuals by x values.

(f) Predict the average life span of these insects with thorax size of 0.85mm.

$$\hat{y} = -28.85505 + 103.7966(0.85) = 59.37206.$$

3. **Estimating $\hat{\beta}$**

(a) State the Gauss-Markov Theorem, including all required assumptions.

Assume the errors have expectation zero, are uncorrelated and have equal variances. Then the OLS estimator is the best linear unbiased estimator (BLUE) of the coefficients. Here "best" means giving the lowest variance of the estimate, as compared to other unbiased, linear estimators. The errors do not need to be normal, nor do they need to be independent and identically distributed (only uncorrelated with mean zero and homoscedastic with finite variance).

(b) Write the definition of estimability.

$\lambda^T \beta$ is estimable if $\lambda \in C(X^T)$.

(c) Suppose $\lambda^T \beta$ is estimable.

i. If $(X^T X)$ is not invertible, what is a solution for β? Is this solution unique?
A solution $\hat{\beta}$ is any $\hat{\beta}$ satisfying the normal equations $X^T X b = X^T y$. There are infinitely many solutions.

ii. If $(X^T X)$ is invertible, what is a solution for β? Is this solution unique?
The least squares solution $\hat{\beta} = (X^T X)^{-1} X^T y$ is the unique solution.

iii. What is an estimate for $\lambda^T \beta$? Is this estimate unique?
When $\lambda^T \beta$ is estimable, use the least squares estimate $\lambda^T \hat{\beta}$. This estimate is the same for every least squares solution $\hat{\beta}$.