# STAT120C Homework 6

**Assigned Thursday May 16th, 2019**

**Due Thursday May 23th, 2019 by 5pm in the Dropbox in DBH**

1. Consider the linear regression model

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i,$$

with $\varepsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$, $i = 1, \cdots, n$. Let $\hat{Y}_h = \hat{\beta}_0 + \hat{\beta}_1 X_h$ be the MLE of the mean at covariate value $X_h$.

   (a) Show $\hat{Y}_h$ is unbiased for $\mathbb{E}[Y|X = X_h]$.

   (b) Write the formulas for $\text{Var}(\hat{\beta}_0), \text{Var}(\hat{\beta}_1)$, and $\text{Cov}(\bar{Y}, \hat{\beta}_1)$.

   (c) Use the formulas from (b) to calculate $\text{Var}(\hat{Y}_h)$.

   (d) What is the distribution of $\hat{Y}_h$?

   (e) How does $\text{Var}(\hat{Y}_h)$ change as the distance between $X_h$ and $\bar{X}$ increases?

   (f) Suppose we estimate $\sigma^2$ by $s^2 = SSE/(n-2)$. Derive the distribution for

$$\frac{\hat{Y}_h - \mathbb{E}[Y|X = X_h]}{\sqrt{s^2[\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum_{i=1}^n (X_i - \bar{X})^2}]}}.$$

   You can use the fact that $SSE/\sigma^2 \sim \chi^2_{n-2}$ without proof.

   (g) What is a $(1-\alpha)100\%$ confidence interval for $\mu(X_h) = E(Y_h|X_h)$?

   (h) Suppose we observe a new observation $Y_{new}$ at covariate value $X = X_{new}$. What is a $(1-\alpha)100\%$ **prediction interval** for $Y_{new}$?

   (i) Give an intuitive explanation for why the prediction interval from (h) is different than the confidence interval from (g).

2. Suppose $Y \sim \text{Exp}(\mu)$, which has pdf $f(y) = \frac{1}{\mu}\exp(-y/\mu)$.

   (a) Use the following `R` code to generate data from the model $Y_i \sim \text{Exp}(0.05/X_i)$, and provide the scatterplot of $Y$ against $X$.

```
set.seed(123)
n <- 500

X <- rnorm(n, 3, 1)
Y <- rexp(n, X)
```

   (b) Fit the model $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$ using the `lm` function in `R` and provide a plot of the best fit line on the scatterplot of $Y$ vs $X$, and the residual vs fitted plot. What assumptions of the linear regression model appear to be violated by the data?

   (c) Derive the variance stabilizing transformation for $Y \sim \text{Exp}(\mu)$.

   (d) Apply the variance stabilizing transformation to the data $Y_i$ and fit the linear regression model to the transformed data. Provide the new residuals by fitted plot and comment on whether the linear regression model assumptions are satisfied by the transformed data.

3. Consider the multiple linear regression model

$$Y_i = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_{p-1,i} + \varepsilon_i,$$

where $X_{1i}, \ldots, X_{p-1,i}$ are observed covariate values for observation $i$, and $\varepsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$.

(a) What is the interpretation of $\beta_1$ in this model?

(b) Write the matrix form of the model. Label the response vector, design matrix, coefficient vector, and error vector, and specify the dimensions and elements for each.

(c) Write the likelihood, log-likelihood, and $\frac{\partial \ell}{\partial \beta}$ in matrix form.

(d) Solve $\frac{\partial \ell}{\partial \beta} = 0$ for $\hat{\beta}$, the MLE of the coefficient vector.

(e) Calculate $\text{Var}(\hat{\beta})$.

(f) Show that when $p = 2$, the estimates for $\beta_0$, $\beta_1$ from (d) are the same as the formulas derived in the simple linear regression case.