Homework 05

Joseph Blubaugh
jblubau1@tamu.edu

STAT 608-720

1. a)

$$
\begin{aligned}
RSS(\vec{\alpha}) &= \sum e_i^2 \\
&= (\vec{y} - x\vec{\alpha})(\vec{y} - x\vec{\alpha}) \\
&= \vec{y}'\vec{y} - \vec{y}x\alpha - \vec{\alpha}'x'y + \vec{\alpha}x'x\vec{\alpha} \\
\\
&= \vec{y}'\vec{y} - 2\vec{y}'x\vec{\alpha} + \vec{\alpha}'x'x\vec{\alpha} \\
\frac{dRSS(\vec{\alpha})}{d\vec{\alpha}} &= -2x'\vec{y} + 2x'x\hat{\vec{\alpha}} = 0 \\
\alpha &= (x'x)^{-1}x'y
\end{aligned}
$$

b) Where $Var(e_i) = \frac{\sigma^2}{w}$ and $w = 1$, for 1…n and $w = 2$, for all n+1

$$
\begin{aligned}
RSS(\vec{\alpha}) &= \sum e_i^2 \\
&= (\vec{y} - xw\vec{\alpha})(\vec{y} - xw\vec{\alpha}) \\
&= \vec{y}'\vec{y} - \vec{y}xw\alpha - \vec{\alpha}'x'wy + \vec{\alpha}x'wx\vec{\alpha} \\
\\
&= \vec{y}'\vec{y} - 2\vec{y}'xw\vec{\alpha} + \vec{\alpha}'x'wx\vec{\alpha} \\
\frac{dRSS(\vec{\alpha})}{d\vec{\alpha}} &= -2x'w\vec{y} + 2x'wx\hat{\vec{\alpha}} = 0 \\
\alpha &= (x'wx)^{-1}x'wy
\end{aligned}
$$

c) For $w = \frac{1}{2}$, Since the variances are not equal, the weighted regression is a better estimate than the simple regression.

$$
\begin{aligned}
\alpha &= (x'x)^{-1}x'y \\
E(\alpha|x) &= \alpha \\
Var(\alpha|x) &= \sigma^2(x'x)^{-1}
\end{aligned}
$$

$$
\begin{aligned}
\alpha &= (x'wx)^{-1}x'wy \\
E(\alpha|x) &= \frac{\alpha}{2} \\
Var(\alpha|x) &= \frac{\sigma^2(x'x)^{-1}}{2}
\end{aligned}
$$

2.

$$
\begin{aligned}
x^2 &= (x'x) \\
Var(e_i|x_i) &= \sigma^2(x'x) \\
w &= \frac{1}{\sqrt{x}} \\
Var(e_i|x_i) &= \sigma^2(x'x)^{-1} \\
w\hat{y_i} &= \hat{\beta}x_i w + \hat{e}_i w
\end{aligned}
$$

3. a) A weighted model is neccesary because there will be differing N for each subdivision. Subdivisions with higher N should have less variance so it would make sense to weight these observations higher.

   b) The model is not valid because there appears to be no straight light relationships between the percent measurements and price per square foot.

   c) In order to obtain a valid regression model I would try logging both the explanatory and response variables. This makes sense because it may be easier to interpret percent change effects on the response variable.

4. a) $\mathbf{w} = \begin{bmatrix} \sqrt{1} & \sqrt{1} & \sqrt{4} & \sqrt{4} \end{bmatrix}'$

   b)

$$
\mathbf{x} = \begin{bmatrix} \sqrt{1}1 & \sqrt{1}0 \\ \sqrt{1}0 & \sqrt{1}1 \\ \sqrt{4}1 & \sqrt{4}1 \\ \sqrt{4}1 & \sqrt{4}1 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 4 & 4 \\ 4 & 4 \end{bmatrix}
$$

   c) Since we have a larger sample in the measurements with 4 coins we are more likely to have less variance, so it makes sense for the weights to be heavier for those samples. The estimates are unbiased because they are constant.

5. a) This is not BLUE because there is no specification of the distribution. If its not symmetrical then this would not work.

   b) This is not BLUE because the expected value of $E(3y_11 - y_12 - y_13 - y_14 - y_15) \neq \alpha_1$

   c) i. Yes because $E(\hat{\beta}) = E(y_4 + 2y_3 - 2y_2 - y_1) = y_i$
      ii. $Var(\hat{\beta} = \frac{1}{25}Var(e_4 + 2e_3 - 2e_2 - e_1)$
      iii. $E(\hat{\beta}) = \begin{bmatrix} \frac{1}{30} & \frac{1}{15} & \frac{1}{10} & \frac{2}{15} \end{bmatrix} y_i$ and $Var(\hat{\beta}) = \frac{1}{30}\sigma^2$
      iv. The sampling variance of the least squares estimator is 1/5 smaller than the sample variance

6. a) In this case and interaction term is neccesary because we want to have a separate slope and intercept for boys and girls feet.

   b) Where, y = width, x1 = length, x2 is a dummy variable with 0 = boy, 1 = girl.

$$
y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{1i} x_{2i} + e_i
$$

   c) 1st column = intercept, 2nd column = width in inches, 3rd column = dummy variable {boy = 0, girl = 1}, 4th column = interaction term between width and sex

$$
\mathbf{X} = \begin{bmatrix} 1 & 7 & 0 & 0 \\ 1 & 7 & 1 & 7 \\ 1 & 8 & 0 & 0 \\ 1 & 6 & 0 & 6 \\ 1 & 9 & 1 & 9 \end{bmatrix}
$$

d) H_o : mean width between girls and boys feet are equal, H_1 : mean width between girls and boys feet are not equal