Problem Sheet 3

Theory of Inference

Dom Hutchinson

Question 1

Part a)

$$
\begin{array}{rrcl}
 & \|\tilde{\mathbf{y}} - \tilde{\mathbf{X}}\beta\|^2 & = & \sum_i (\tilde{y}_i - \sum_j \tilde{X}_{ij}\beta_j)^2 \\
 & & = & \sum_i (W_{ii}y_i - \sum_j W_{ii}X_{ij}\beta_j)^2 \\
 & & = & \sum_i W_{ii}^2 (y_i - \sum_j X_{ij}\beta_j)^2 \\
\text{Define} & W_{ij} & := & \begin{cases} \frac{1}{\sqrt{n_{se(i)}}} & i \equiv j \\ 0 & \text{otherwise} \end{cases} \\
\implies & \|\tilde{\mathbf{y}} - \tilde{\mathbf{X}}\beta\|^2 & = & \sum_i \frac{1}{\sqrt{n_{se(i)}}} (y_i - \sum_j X_{ij}\beta_j)^2
\end{array}
$$

Part b)

$$
\begin{array}{rrcl}
\text{We have} & \|\tilde{\mathbf{y}} - \tilde{X}\beta\|^2 & = & (\tilde{\mathbf{y}} - \tilde{X}\beta)^T(\tilde{\mathbf{y}} - \tilde{X}\beta) \\
 & & = & (y - X\beta)^T W^2 (y - X\beta) \\
 & & = & y^T W^2 y - y^T W^2 X\beta - \beta^T X^T W^2 y + \beta^T X^T W^2 X\beta \\
\implies & \frac{\partial}{\partial\beta}\|\tilde{\mathbf{y}} - \tilde{X}\beta\| & = & -X^T W^2 y - X^T W^2 y + 2X^T W^2 X\beta \\
\text{Setting} & 0 & = & \frac{\partial}{\partial\beta}\|\tilde{\mathbf{y}} - \tilde{X}\beta\| \\
\implies & 0 & = & 2(X^T W^2 X\hat{\beta}_{\text{WLS}} + X^T W^2 X\hat{\beta}_{\text{WLS}}) \\
\implies & \hat{\beta}_{\text{WLS}} & = & (X^T W^2 X)^{-1} X^T W^2 y \\
\implies & \mathbb{E}[\hat{\beta}_{\text{WLS}}] & = & \mathbb{E}((X^T W^2 X)^{-1} X^T W^2 y) \\
 & & = & (X^T W^2 X)^{-1} X^T W^2 \mathbb{E}(y) \\
 & & = & (X^T W^2 X)^{-1} X^T W^2 X\beta \\
 & & = & \beta
\end{array}
$$

Since $\mathbb{E}(\hat{\beta}_{\text{WLS}}) = \beta$ then $\hat{\beta}_{\text{WLS}}$ is an unbiased estimator of β.

Part c)

Since this estimator is unbiased and fulfils all the residual assumptions I would recommend the weighted least squares estimator here as it takes into account more information than a standard least squares estimator.

Question 2

Model doesn't account for the fact that the markets close at night. There are likely to be sizable changes in the price of a share between closing and opening of the markets. This could be seen by regular spikes in model residual values at the start of each day.

Question 3

Part a)

If we know that an observation is not in group one or two, then it most be in group three. Thus the groups are not independent & the model matrix is rank deficient. This means we cannot invert the matrix.

Part b)

```r
m1<-lm(weight ~ group,data=PlantGrowth)
```

In this interpretation value which is the expected value for observations in `ctrl` group. There are then two binary variables which state whether the observation was made from `grouptrt1` or `grouptrt2` (these are not independent). The parameters associated with each of these variables is the expected difference in `weight` for observations in that group, compared to `ctrl`.

Part c)

$H_0 : \boldsymbol{\beta} = \begin{pmatrix} \mu \\ 0 \\ 0 \end{pmatrix}$ against $H_1 : \boldsymbol{\beta} \neq \begin{pmatrix} \mu \\ 0 \\ 0 \end{pmatrix}$.

```r
anova(m1)
```

```
## Analysis of Variance Table
##
## Response: weight
##           Df  Sum Sq Mean Sq F value  Pr(>F)
## group      2  3.7663  1.8832  4.8461 0.01591 *
## Residuals 27 10.4921  0.3886
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We reject null hypothesis at the 5% level. Thus concludinf that `group` is needed in the model.
It uses the results in second 3.4.4 of the notes, in particular

$$F = \frac{(RSS_0 - RSS_1)/q}{RSS_1/(n-p)}$$

where $n = 30,\ p = 3,\ q = 3$. $TODO check this

Part d)

```r
m0<-lm(weight~1,data=PlantGrowth)
anova(m0,m1)
```

```
## Analysis of Variance Table
##
## Model 1: weight ~ 1
## Model 2: weight ~ group
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1     29 14.258
## 2     27 10.492  2    3.7663 4.8461 0.01591 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Same result as in part c).

Part e)

```r
summary(m1)
```

```
##
## Call:
```

```
## lm(formula = weight ~ group, data = PlantGrowth)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.0710 -0.4180 -0.0060  0.2627  1.3690
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   5.0320     0.1971  25.527   <2e-16 ***
## grouptrt1    -0.3710     0.2788  -1.331   0.1944
## grouptrt2     0.4940     0.2788   1.772   0.0877 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6234 on 27 degrees of freedom
## Multiple R-squared:  0.2641, Adjusted R-squared:  0.2096
## F-statistic: 4.846 on 2 and 27 DF,  p-value: 0.01591
```

$H_0 : \boldsymbol{\beta} = \mathbf{0}$ against $H_1 : \boldsymbol{\beta} \neq \mathbf{0}$. The p-Value is unchanged.

Part f)

We have that

$$\frac{\hat{\beta}_2 - \beta_i}{\hat{\sigma}_{\hat{\beta}_2}} \sim t_{27}$$

Thus

$$\begin{aligned}1 - 2\alpha &= \mathbb{P}\left(-t_{27}(\alpha) < \frac{\hat{\beta}_2 - \beta_i}{\hat{\sigma}_{\hat{\beta}_2}} < t_{27}(\alpha)\right) \\ &= \mathbb{P}(\hat{\beta}_2 - t_{27}(\alpha)\hat{\sigma}_{\hat{\beta}_2} < \beta_2 < \hat{\beta}_2 + t_{27}(\alpha)\hat{\sigma}_{\hat{\beta}_2})\end{aligned}$$

In this scenario we have that

$$\begin{aligned}\hat{\beta}_2 &= \bar{x}_{\text{trt2}} - \bar{x}_{\text{ctrl}} \\ &= 5.525 - 5.032 \\ &= 0.4940 \\ \hat{\sigma}^2_{\hat{\beta}_2} &= \tfrac{1}{10-1}\textstyle\sum_{i=1}^{10}(y_i - \hat{y}_i)^2 \\ &= \tfrac{1}{9}1.76284 \\ &= 0.1959 \\ \implies \hat{\sigma}_{\hat{\beta}_2} &= 0.4425733\end{aligned}$$

Giving us the following 90% confidence interval for difference between treatment 2 and the control

$$\begin{aligned}&[0.4940 - 1.703 \cdot 0.4425733, 0.4940 + 1.703 \cdot 0.4425733] \\ &= [-0.272, 4.342]\end{aligned}$$