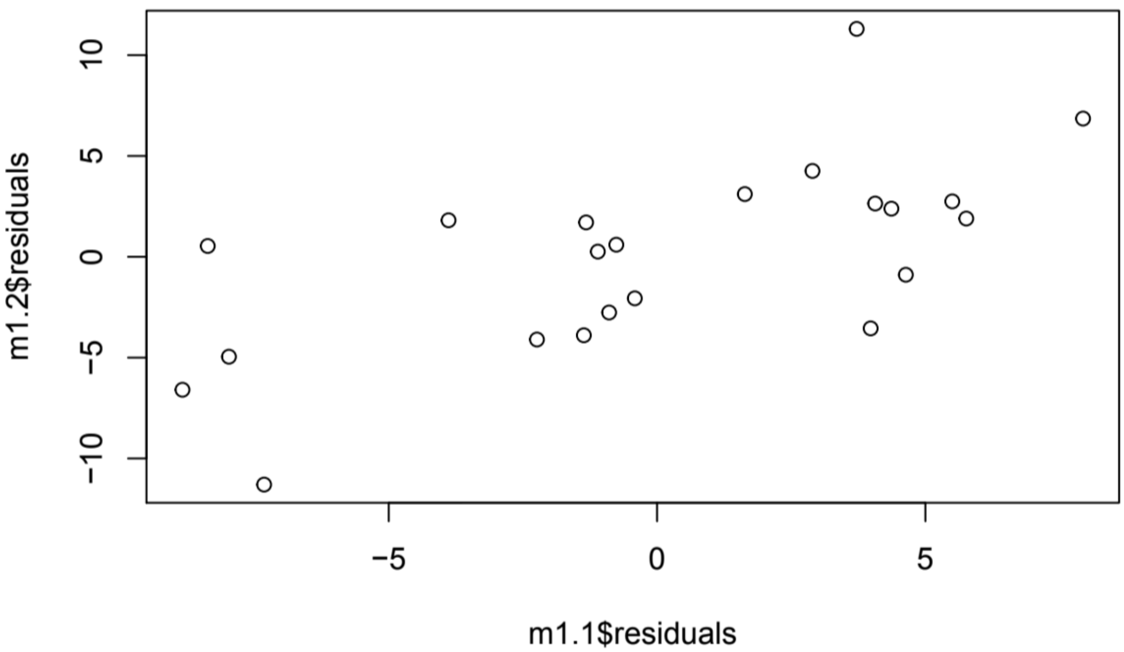
Math 564: HW#4

Aleksei Sorokin | A20394300 | asorokin@hawk.iit.edu

Problem 1: Ex 4.8.b

1. Regress F on P_1
2. Regress P_2 on P_1
3. Plot the residuals from step 1 vs the residuals from step 2

```
df1 = read.csv(url("http://www1.aucegypt.edu/faculty/hadi/RABE5/Data5/P083.txt"),sep='\t')
m1.1 = lm(F~P1,data=df1)
m1.2 = lm(P2~P1,data=df1)
plot(m1.1$residuals,m1.2$residuals)
```



Because the residuals vs residuals plot shows a linear relationship, we conclude that both predictors should be included. Therefore, the best model is

$$F = \beta_0 + \beta_1 P_1 + \beta_2 P_2 + \epsilon.$$

Problem 2

```
df2 = read.csv('table6.2.txt',sep='\t')
names(df2) = c('t','Nt')
```

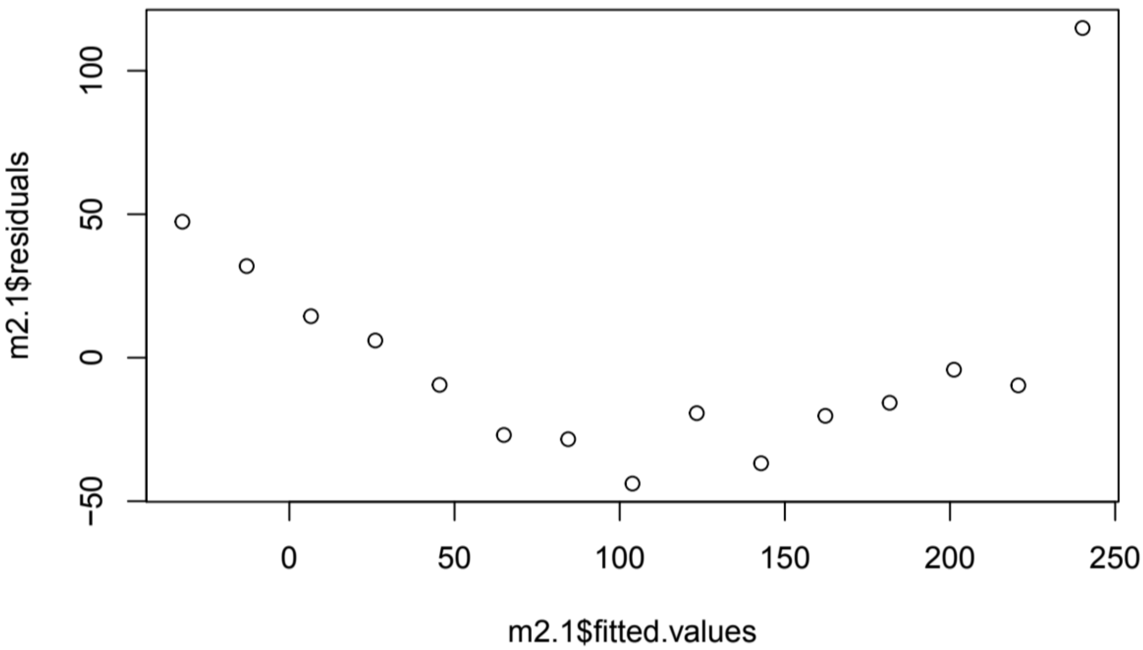
Part 1

```
m2.1 = lm(Nt~t,data=df2)
summary(m2.1)
```

```
##
```

```
## Call:
## lm(formula = Nt ~ t, data = df2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -43.867 -23.599  -9.652  10.223 114.883
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   259.58      22.73  11.420 3.78e-08 ***
## t             -19.46       2.50  -7.786 3.01e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 41.83 on 13 degrees of freedom
## Multiple R-squared:  0.8234, Adjusted R-squared:  0.8098
## F-statistic: 60.62 on 1 and 13 DF,  p-value: 3.006e-06
```

```
plot(m2.1$fitted.values,m2.1$residuals)
```



The residuals follow a clear pattern, so the linear assumption is violated.

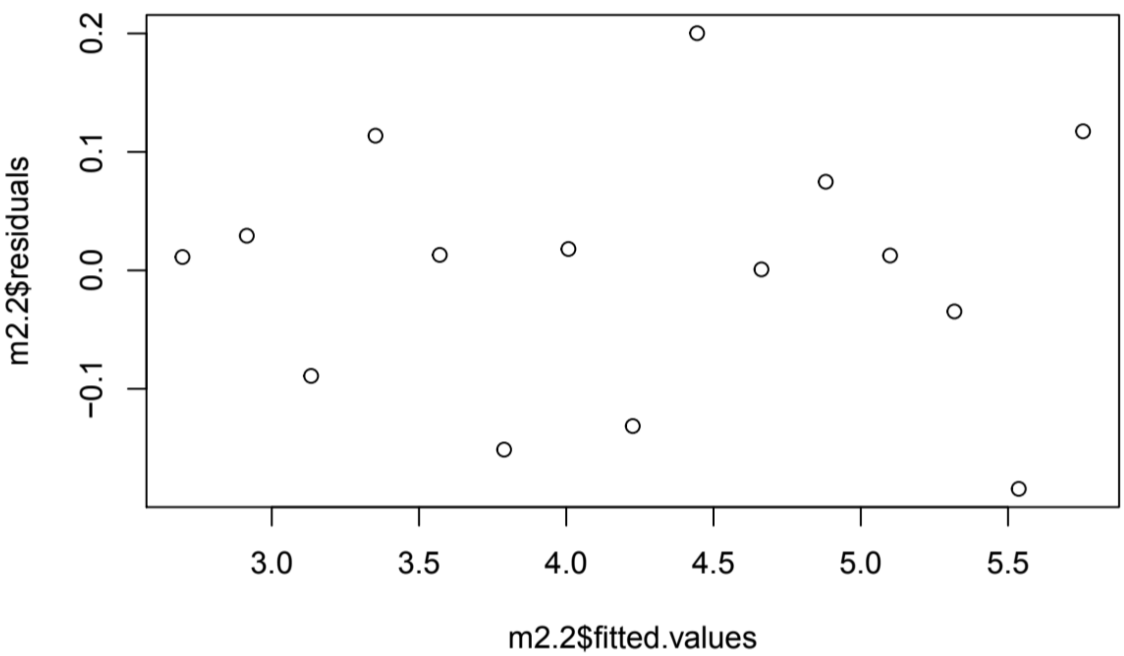
Part 2

```
df2$logNt = log(df2$Nt)
m2.2 = lm(logNt~t,data=df2)
summary(m2.2)
```

```
##
## Call:
## lm(formula = logNt ~ t, data = df2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.18445 -0.06189  0.01253  0.05201  0.20021
##
```

```
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.973160   0.059778   99.92  < 2e-16 ***
## t           -0.218425   0.006575  -33.22 5.86e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.11 on 13 degrees of freedom
## Multiple R-squared:  0.9884, Adjusted R-squared:  0.9875
## F-statistic:  1104 on 1 and 13 DF,  p-value: 5.86e-14
```

```
plot(m2.2$fitted.values,m2.2$residuals)
```



The linear assumption is no longer violated as shown by the lack of a trend in the above plot. The transformed model is

$$\widehat{\log(n_t)} = 5.97 - 0.218\,t.$$

Problem 3

```
df3 = read.csv('table6.6.txt',sep='\t')
```

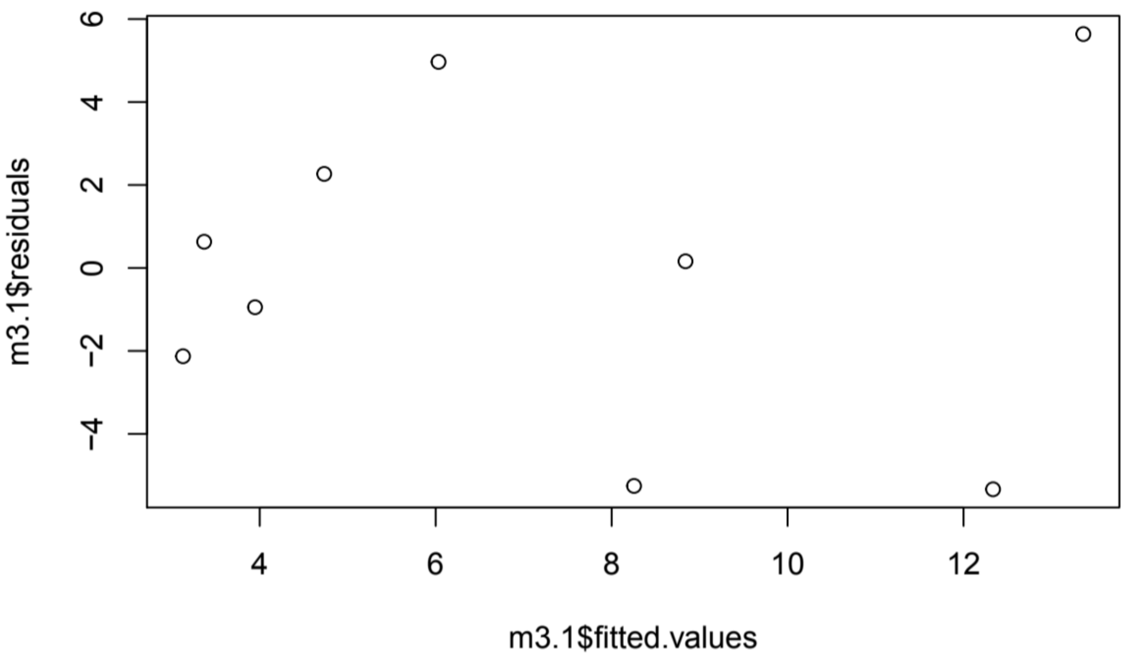
Part 1

```
m3.1 = lm(Y~N, data=df3)
summary(m3.1)
```

```
##
## Call:
## lm(formula = Y ~ N, data = df3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.3351 -2.1281  0.1605  2.2670  5.6382
##
## Coefficients:
```

```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.1402     3.1412  -0.045   0.9657
## N            64.9755    25.1959   2.579   0.0365 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.201 on 7 degrees of freedom
## Multiple R-squared:  0.4872, Adjusted R-squared:  0.4139
## F-statistic:  6.65 on 1 and 7 DF,  p-value: 0.03654
```

```
plot(m3.1$fitted.values,m3.1$residuals)
```



The size of residuals appears to grow with the fitted Y values, so the heteroscedastic assumption is violated.

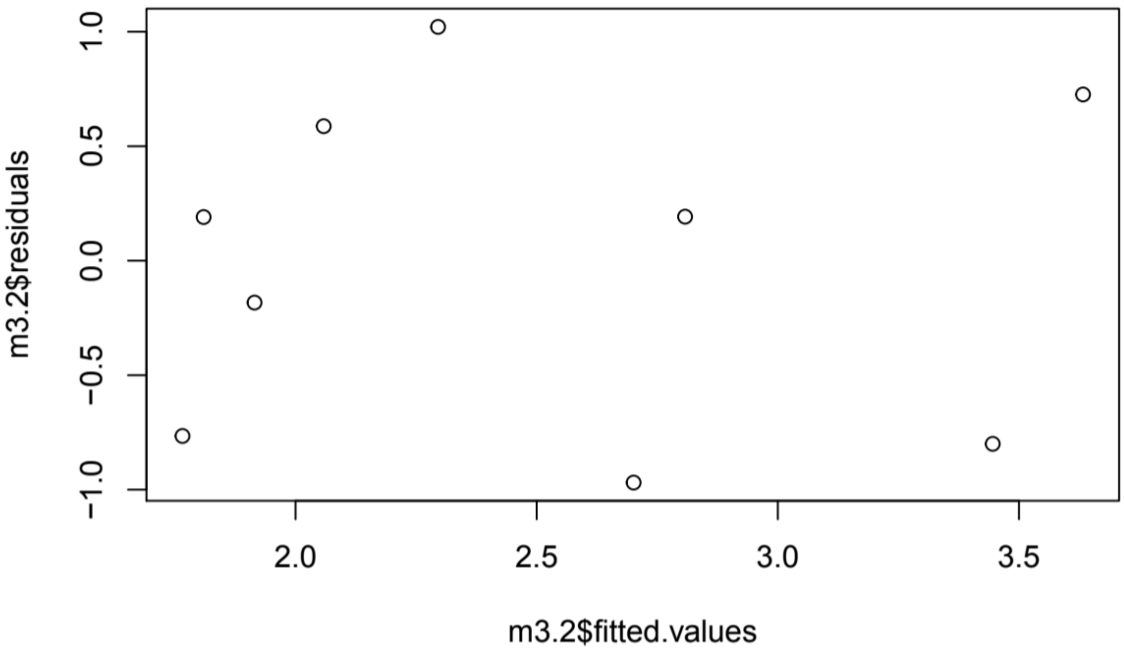
Part 2

```
df3$sqrtY = sqrt(df3$Y)
m3.2 = lm(sqrtY~N,data=df3)
summary(m3.2)
```

```
##
## Call:
## lm(formula = sqrtY ~ N, data = df3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.9690 -0.7655  0.1906  0.5874  1.0211
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.1692     0.5783   2.022   0.0829 .
## N            11.8564     4.6382   2.556   0.0378 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7733 on 7 degrees of freedom
```

```
## Multiple R-squared:  0.4828, Adjusted R-squared:  0.4089
## F-statistic: 6.535 on 1 and 7 DF,  p-value: 0.03776
```

```
plot(m3.2$fitted.values,m3.2$residuals)
```



Again, the size of residuals appears to grow with the fitted Y values, so the heteroscedastic assumption is violated. The transformed model is

$$\widehat{\sqrt{Y}} = 1.17 + 11.9\,N.$$

Problem 4

```
df4 = read.csv('table6.9.txt',sep='\t')
```

WLS

```
m4.wls = lm(Y~X,data=df4,weight=1/(df4$X^2))
summary(m4.wls)
```

```
##
## Call:
## lm(formula = Y ~ X, data = df4, weights = 1/(df4$X^2))
##
## Weighted Residuals:
##       Min        1Q    Median        3Q       Max
## -0.041477 -0.013852 -0.004998  0.024671  0.035427
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 3.803296   4.569745   0.832    0.413
## X           0.120990   0.008999  13.445 6.04e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.02266 on 25 degrees of freedom
## Multiple R-squared:  0.8785, Adjusted R-squared:  0.8737
```

```
## F-statistic: 180.8 on 1 and 25 DF,  p-value: 6.044e-13
```

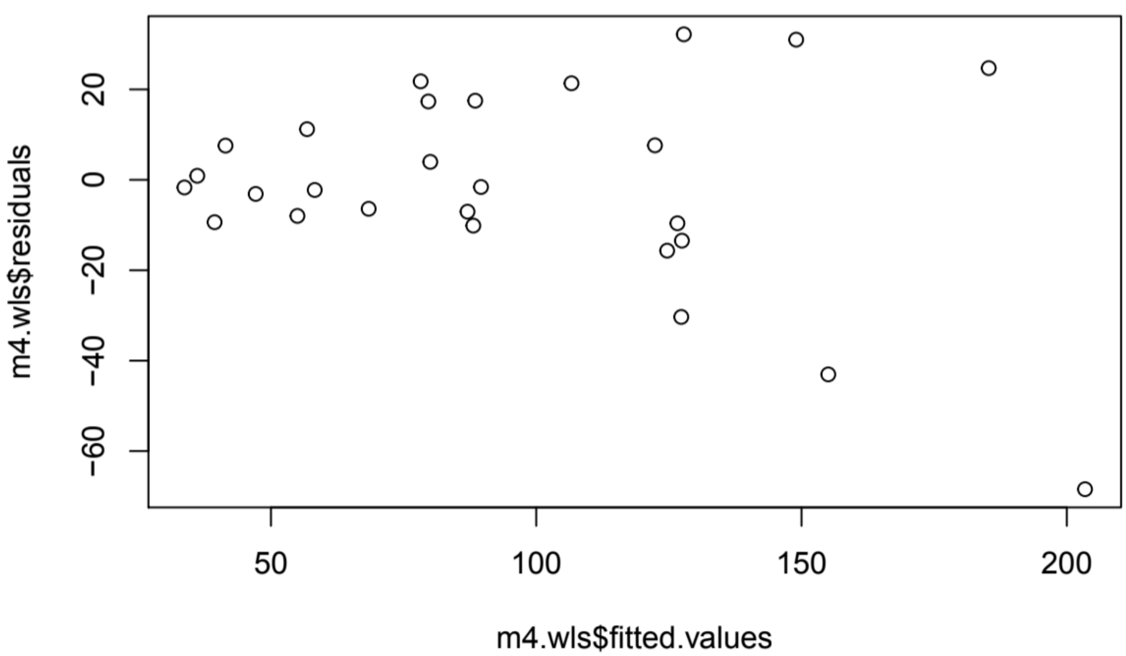
Y'=Y/X, X'=1/X

```
df4$Yp = df4$Y/df4$X
df4$Xp = 1/df4$X
m4.tf = lm(Yp~Xp,data=df4)
summary(m4.tf)
```

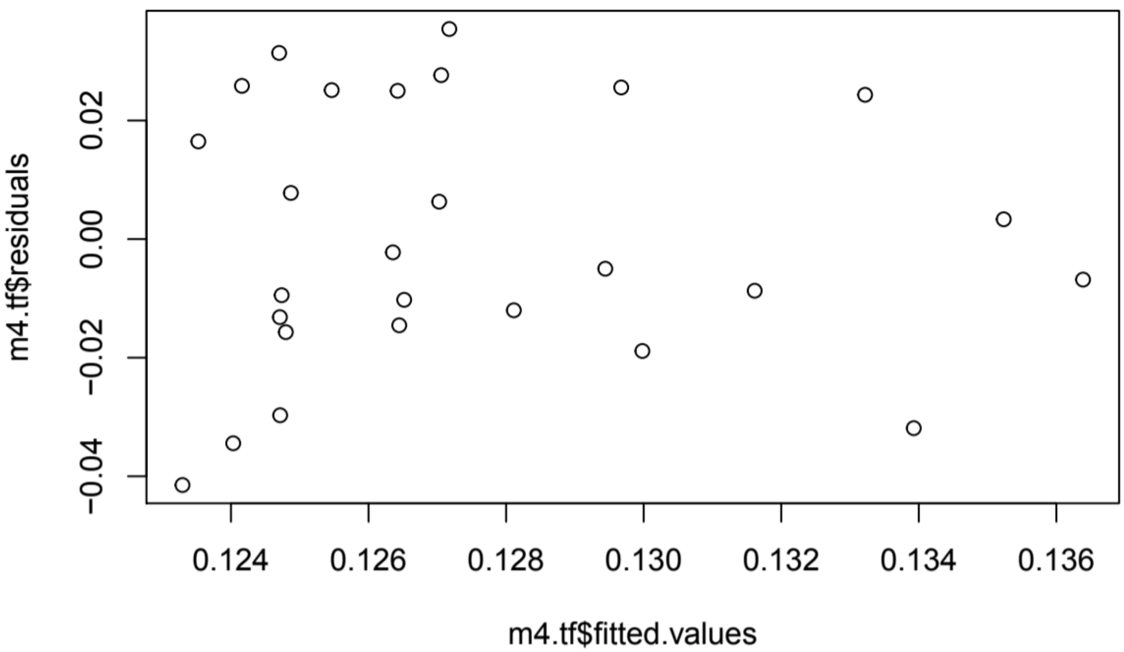
```
##
## Call:
## lm(formula = Yp ~ Xp, data = df4)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.041477 -0.013852 -0.004998  0.024671  0.035427
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.120990   0.008999  13.445 6.04e-13 ***
## Xp          3.803296   4.569745   0.832    0.413
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.02266 on 25 degrees of freedom
## Multiple R-squared:  0.02696,    Adjusted R-squared:  -0.01196
## F-statistic: 0.6927 on 1 and 25 DF,  p-value: 0.4131
```

Comparison

```
plot(m4.wls$fitted.values,m4.wls$residuals)
```




```
plot(m4.tf$fitted.values,m4.tf$residuals)
```



Weighted least squares (WLS) model:

$$\hat{Y} = 3.8 + 0.121\,X.$$

Transformed model

$$\widehat{Y'} = 0.121 + 3.8\,X'.$$

Both models do a good job removing heteroscedasticity. In fact, the models are equivalent since for $Y' = Y/X$ and $X' = 1/X$, the transformed model multiplied by X recovers the WLS model

$$Y = X(Y') = X(\beta_0 + \beta_1 X' + \epsilon) = \beta_1 + \beta_0 X + X\epsilon.$$

Problem 5

```
m4.ols = lm(Y~X,data=df4)
diag(vcov(m4.ols))
```

```
##  (Intercept)            X
## 9.143207e+01 1.282703e-04
```

```
diag(vcov(m4.wls))
```

```
##  (Intercept)            X
## 2.088257e+01 8.097546e-05
```

OLS gives $\text{var}(\beta_0) = 9.14 * 10$ and $\text{var}(\beta_1) = 1.28 * 10^{-4}$.

WLS gives $\text{var}(\beta_0) = 2.09 * 10$ and $\text{var}(\beta_1) = 8.10 * 10^{-5}$.

Therefore, WLS produced smaller coefficient variances.