Exercise 1: Do problem 7.8. In part 2., the textbook contains a significant mistake. It should say, "WLS should be used with variance function $Var(Weight|Age) = SD^2\sigma^2/n$. " The point that the author is trying to make is that the applicable weights are n_i/SD^2. Skip parts 4 and 5.

7.8.1 Draw a scatter plot go Weight versus Age, and comment on the applicability of the usual assumptions of linear regression model. Also draw a scatterplot of SD versus Age, then summarize the information in this plot.

Solution:
Plotting both scatterplots in r, it does seem like the first one fits all the assumptions of constant variance, and linearity. However looking at the standard deviation scatterplot, it is clear that the variance is not constant across all data.

Figure 1: Weight vs Age

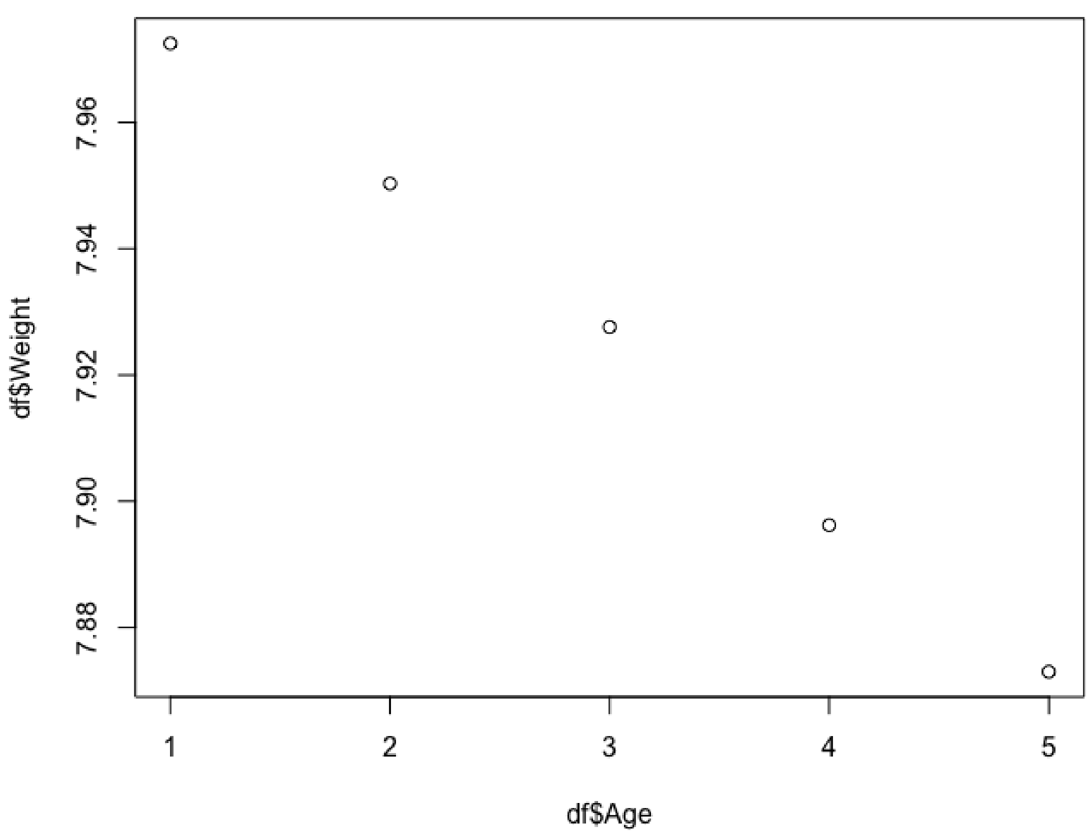
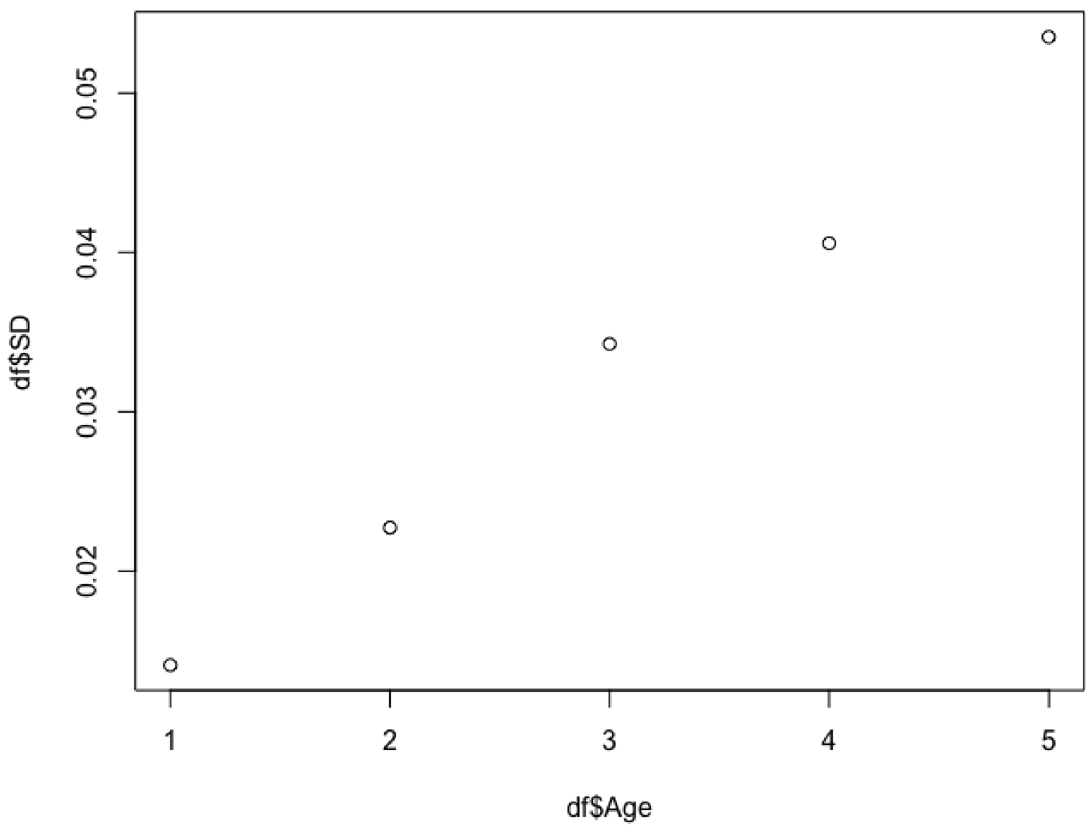


Figure 2: SD of Weight vs Age

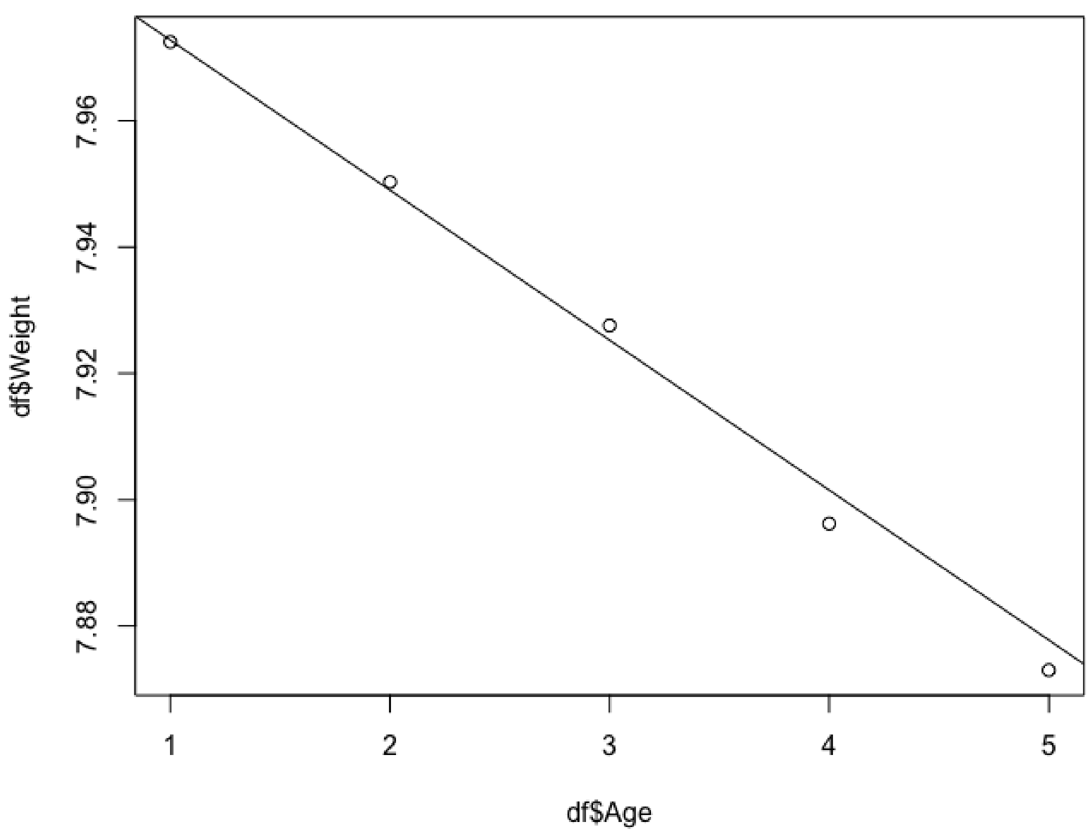


7.8.2 Fit a WLS regression with Weight as the response, using $Var(Weight|Age) = SD^2\sigma^2/n$ as the variance function.

Solution:
Fitting the model in r we get the following,
Code:

```
> WLS_Model <- lm(Weight~Age, weights = n/SD^2, data = df)
> summary(WLS_Model)

Call:
lm(formula = Weight ~ Age, data = df, weights = n/SD^2)

Weighted Residuals:
      1       2       3       4       5
-0.2091  0.5017  0.3875 -0.5383 -0.4339

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.9965218  0.0013220    6049 9.96e-12 ***
Age         -0.0237562  0.0008797     -27 0.000111 ***
---
Residual standard error: 0.5549 on 3 degrees of freedom
Multiple R-squared:  0.9959,    Adjusted R-squared:  0.9945
F-statistic: 729.2 on 1 and 3 DF,  p-value: 0.0001114


> plot(df$Age, df$Weight)
> abline(WLS_Model)
```

Figure 3: WLS model for Weight vs Age



7.8.3 Is the fitted regression consistent with the known standard weight for a new coin.

Solution:
The problem statement gives 7.9876g as the standard weight of a gold sovereign. We can see if our WLS regression is consistent by computing the confidence interval on the intercept. Doing so we get, a 95 percent confidence interval of (7.99231466, 8.00072893) which excludes the standard weight. I would see about experimenting with other weights especially since when we square the RSE we get a value around $1/4$ which goes against our assumption that $\sigma^2 = 1$.
Code:

```
> confint(WLS_Model)
                  2.5 %      97.5 %
(Intercept)  7.99231466  8.00072893
Age         -0.02655593 -0.02095642
```

Exercise 2: Use salarygov data. Although the response (MaxSalary) is a maximum of , rather than a mean of, sub-observations, fit the WLS model with weight that represent rows' differing sample sizes. Your model should include the predictor Female_dominated, the spline bases for Score, and the interaction terms between these. A description of how to create Female_dominated is given in 5.9.3. For the splines, use B-splines with 3 degrees of freedom. Once the model is fitted do the following:

a. Report the fitted model.

Solution:

Code:

```
> df <- salarygov
> df$Female_dominated = df$NW/df$NE
> for(i in 1:length(df$Female_dominated)){
+ df$Female_dominated[i] = ifelse(df$Female_dominated[i] >= .70, 1, 0)
+ }
> df$Female_dominated<-factor(df$Female_dominated)
-----------------------------------------------------------------------------
> ScoreSpline <- bs(df$Score, df = 3, degree = 2)
> df$S1 = ScoreSpline[,1]
> df$S2 = ScoreSpline[,2]
> df$S3 = ScoreSpline[,3]
-----------------------------------------------------------------------------

> WLS_Model <- lm(MaxSalary ~ Female_dominated + S1 + S2 + S3 +
                  Female_dominated:S1 + Female_dominated:S2 + Female_domina
                  weights = NE, data = df)

> summary(WLS_Model)
Call:
lm(formula = MaxSalary ~ Female_dominated + S1 + S2 + S3 + Female_dominated
    Female_dominated:S2 + Female_dominated:S3, data = df, weights = NE)

Weighted Residuals:
    Min      1Q  Median      3Q     Max
-7763.4  -527.4   -84.7   353.5  8717.7

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)           1425.33      98.29  14.501   <2e-16 ***
Female_dominated1     -318.88     130.02  -2.453   0.0145 *
S1                     305.70     142.76   2.141   0.0327 *
S2                    3347.49     119.05  28.117   <2e-16 ***
S3                    5262.94     337.46  15.596   <2e-16 ***
Female_dominated1:S1   232.81     212.25   1.097   0.2733
```

```
Female_dominated1:S2    -438.77      225.48   -1.946    0.0522 .
Female_dominated1:S3  -2442.93     1444.81   -1.691    0.0915 .
---
Residual standard error: 1113 on 487 degrees of freedom
Multiple R-squared:  0.8761,    Adjusted R-squared:  0.8743
F-statistic:   492 on 7 and 487 DF,  p-value: < 2.2e-16
```

b Perform a partial F-test on the interaction terms to determine if female-dominated occupations require different spline coefficints than other occupations.

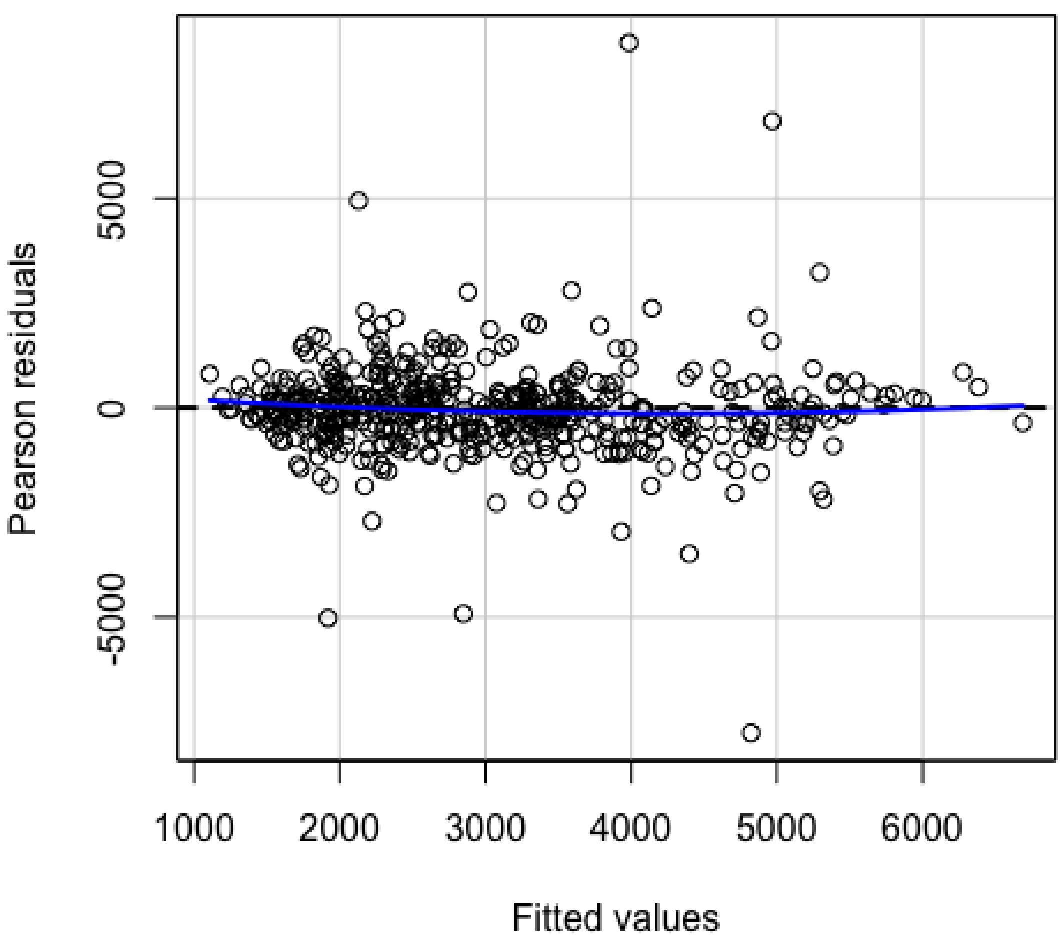
Solution:
With a p-value of .00436 the following F-test tells us that female-dominated occupations require different spline coefficients.
Code:

```
> WLS_Model_NOInt <- lm(MaxSalary ~ Female_dominated + S1 + S2 + S3,
                        weights = NE, data = df)

> anova(WLS_Model_NOInt, WLS_Model)
Analysis of Variance Table

Model 1: MaxSalary ~ Female_dominated + S1 + S2 + S3
Model 2: MaxSalary ~ Female_dominated + S1 + S2 + S3 + Female_dominated:S1 +
    Female_dominated:S2 + Female_dominated:S3
  Res.Df       RSS Df Sum of Sq      F   Pr(>F)
1    490 626409067
2    487 603678615  3  22730452 6.1124 0.000436 ***
---
```

c. Give the residuals-vs-fitted values plot from the model that include the interaction terms and interpret the plot in terms of model assumptions.

Solution:
The residual plot seems to show some level of non-constant variance still. it does seem to be opening outwards (like this"¡") with a significant clustering towards the initial values and around zero. Definitely does not look like random scatter.

Figure 4: Residual vs Fitted Plot



Code:

```
> residualPlot(WLS_Model)
> residualPlots(WLS_Model)
                 Test stat Pr(>|Test stat|)
Female_dominated
S1                 -4.5584        6.531e-06 ***
S2                 -9.1824        < 2.2e-16 ***
S3                  0.5970           0.5508
Tukey test         -9.6050        < 2.2e-16 ***
```

Exercise 3.: The Blackmore data set in alr4 provides the number of hours of exercise performed each week by 236 teenage girls at five different ages. It also provides a categorical indicator of whether the subject was hospitalized for an eating disorder. Do the following,

a. Fit a mixed model that controls for age and group as fixed effects and has a random intercept for subject. Give the estimated variance component for subject and interpret it.

Solution:
Fitting the mixed model with the lmer function we get a estimated variance component for subject of 3.898, a non-insignificant in the mean exercise score and subjects.
Code:

```
> df <- Blackmore
> MixedModel <- lmer(exercise ~ age + group + (1 | subject), data = d

> summary(MixedModel)
 Linear mixed model fit by REML ['lmerMod']
 Formula: exercise ~ age + group + (1 | subject)
    Data: df

REML criterion at convergence: 4704.5

 Scaled residuals:
     Min      1Q  Median      3Q     Max
 -2.5411 -0.5217 -0.0894  0.3182  7.5223

 Random effects:
  Groups   Name        Variance Std.Dev.
  subject  (Intercept) 3.898    1.974
  Residual             6.217    2.493
 Number of obs: 945, groups:  subject, 231

 Fixed effects:
              Estimate Std. Error t value
 (Intercept)   -3.3988     0.4144  -8.202
 age            0.4500     0.0300  14.998
 grouppatient   1.2993     0.3147   4.129

 Correlation of Fixed Effects:
             (Intr) age
 age         -0.807
 grouppatint -0.433 -0.031
```

b. Test the variance component for subject is equal to 0 using a likelihood ratio test. Report a test statistic, p-value, and your conclusion.

Solution:
Fitting the regular fixed model, which excludes the random intercept subject parameter, we get a chi squared test statistic of 184.26 with a p-value on the order of 10^{-16} which means that the random intercept subject predictor is significant.
Code:

```
> FixedModel <- lm(exercise ~ age + group, data = df)

> anova(MixedModel, FixedModel)
refitting model(s) with ML (instead of REML)
Data: df
Models:
FixedModel: exercise ~ age + group
MixedModel: exercise ~ age + group + (1 | subject)
           npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
FixedModel    4 4889.2 4908.6 -2440.6   4881.2
MixedModel    5 4706.9 4731.2 -2348.5   4696.9 184.26  1  < 2.2e-16 ***
```

c. Produce a normal probability plot of the predicted random effects for subject. Interpret the plot and what it says about your model.

Solution:
Producing the normal probability plot in r we can see that the random effects do not look to be normally distributed. We should not trust this model, and probably experiment with other mixed models.

Figure 5: normal probability plot for predicted random effects.