Additional task

1. Estimate the following wage equation with least squares and heteroskedasticity-robust standard errors, and report the results.

$$\ln(WAGE) = \beta_1 + \beta_2 EDUC + \beta_3 EXPER + \beta_4 EXPER^2 + \beta_5 (EXPER * EDUC) + e$$

```
data_1 <- read.csv(file='D:\\BIRMINGHAM\\STUDIES\\SEMESTER2\\STATMETHODS\\ADDITIONALTSK\\cps4_small.csv')

model_ols <- lm(log(wage) ~ educ + exper + I(exper^2) + I(exper*educ), data = data_1)
summary(model_ols)
cov1 <- hccm(model_ols, type="hc1")
model_heterorobust <- coeftest(model_ols, vcov.=cov1)
model_heterorobust
```

```
> summary(model_ols)

Call:
lm(formula = log(wage) ~ educ + exper + I(exper^2) + I(exper *
    educ), data = data_1)

Residuals:
     Min       1Q   Median       3Q      Max
-2.28227 -0.32856 -0.02725  0.33751  1.47088

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     5.297e-01  2.267e-01   2.336  0.01969 *
educ            1.272e-01  1.472e-02   8.642  < 2e-16 ***
exper           6.298e-02  9.536e-03   6.604 6.48e-11 ***
I(exper^2)     -7.139e-04  8.804e-05  -8.109 1.49e-15 ***
I(exper * educ) -1.322e-03  4.949e-04  -2.672  0.00766 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5057 on 995 degrees of freedom
Multiple R-squared:  0.2445,	Adjusted R-squared:  0.2415
F-statistic: 80.52 on 4 and 995 DF,  p-value: < 2.2e-16
```

```
> model_heterorobust

t test of coefficients:

                   Estimate  Std. Error t value  Pr(>|t|)
(Intercept)      5.2968e-01  2.5283e-01  2.0950   0.03642 *
educ             1.2720e-01  1.6960e-02  7.4999 1.413e-13 ***
exper            6.2981e-02  1.1378e-02  5.5355 3.969e-08 ***
I(exper^2)      -7.1394e-04  9.2013e-05 -7.7591 2.114e-14 ***
I(exper * educ) -1.3224e-03  6.3679e-04 -2.0766   0.03809 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

All coefficients in the fitted model are significant with 5% significance level, according to their t-values. According to the model, an increase in 1 year of education, while all other variables are constant, will lead to an increase of 12.72% - 0.132 %***exper** in earnings per hour, so with increasing experience the rate of increasing wage of education become less. Also, the model captures the diminishing returns in post-education years experience; the change in wage to a change in experience is = 6.298% - 0.071%***exper** – 0.132%***educ**. So, generally speaking, with **exper** increasing, **wage** increase. However, the rate of increase becomes less with **exper** and **educ** increasing. The model explains 24.45% of the variation in wage. Relying on F-test, the overall model is significant.

After getting heteroskedasticity-robust se, we see that the se of **exper** term was overrated in one order, and other se of terms were slightly overrated as well.

2. Add MARRIED to the equation and re-estimate. Holding education and experience constant, do married workers get higher wages? Using a 1% significance level, test a null hypothesis that wages of married workers are less than or equal to those of unmarried workers against the alternative that wages of married workers are higher.

```
model_ols2 <- lm(log(wage) ~ educ + exper + I(exper^2) + I(exper*educ) + married, data = data_1)
summary(model_ols2)
cov2 <- hccm(model_ols2, type="hc1")
model_heterorobust_2 <- coeftest(model_ols2, vcov.=cov2)
model_heterorobust_2
```

```
> summary(model_ols2)

Call:
lm(formula = log(wage) ~ educ + exper + I(exper^2) + I(exper *
    educ) + married, data = data_1)

Residuals:
     Min       1Q   Median       3Q      Max
-2.29834 -0.32252 -0.02409  0.33333  1.45621

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      0.5410611  0.2268944   2.385  0.01728 *
educ             0.1261199  0.0147433   8.554  < 2e-16 ***
exper            0.0613731  0.0096289   6.374 2.82e-10 ***
I(exper^2)      -0.0006934  0.0000897  -7.729 2.64e-14 ***
I(exper * educ) -0.0013091  0.0004949  -2.645  0.00829 **
married          0.0402895  0.0337911   1.192  0.23342
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5056 on 994 degrees of freedom
Multiple R-squared:  0.2456,	Adjusted R-squared:  0.2418
F-statistic: 64.73 on 5 and 994 DF,  p-value: < 2.2e-16
```

```
> model_heterorobust_2

t test of coefficients:

                   Estimate  Std. Error t value  Pr(>|t|)
(Intercept)      5.4106e-01  2.5421e-01  2.1284   0.03355 *
educ             1.2612e-01  1.7056e-02  7.3943 3.015e-13 ***
exper            6.1373e-02  1.1588e-02  5.2964 1.454e-07 ***
I(exper^2)      -6.9335e-04  9.5567e-05 -7.2551 8.074e-13 ***
I(exper * educ) -1.3091e-03  6.3842e-04 -2.0506   0.04057 *
married          4.0289e-02  3.3923e-02  1.1877   0.23525
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

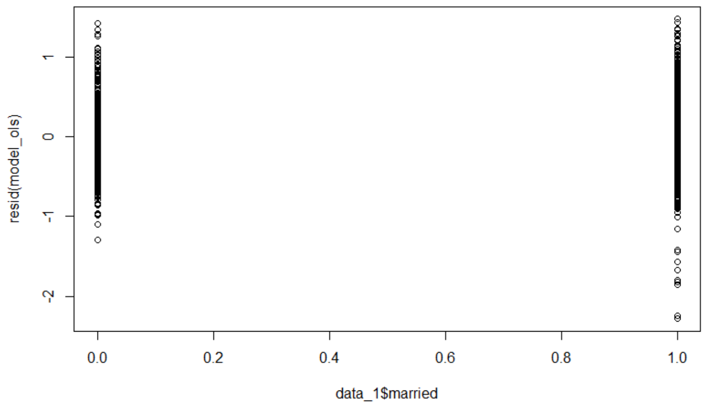
The married term is not significantly important in the equation according to the t-test (p-value = 0.235).

```
> model_heterorobust_2[6]/model_heterorobust_2[12] < 2.326
[1] TRUE
```

According to the test with 1% significance level, married workers get higher wages by 4%.

3. Plot the residuals from part (1) against MARRIED. Is there evidence of heteroskedasticity?

```
model_1_res = resid(model_ols)
plot(data_1$married, resid(model_ols))
```



It looks like, yes, for married workers, the interval of residuals is bigger.

4. Estimate the model in part (1) twice---once using observations on only married workers and once using observations on only unmarried workers. Use the Goldfeld-Quandt test and a 1% significance level to test whether the error variances for married and unmarried workers are different.

```
li <- data_1[which(data_1$married == 1),]
hi <- data_1[which(data_1$married == 0),]
eqli <- lm(log(wage) ~ educ + exper + I(exper^2) + I(exper*educ), data=li)
eqhi <- lm(log(wage) ~ educ + exper + I(exper^2) + I(exper*educ), data=hi)

dfli <- eqli$df.residual
dfhi <- eqhi$df.residual

sigsqli <- glance(eqli)$sigma^2
sigsqhi <- glance(eqhi)$sigma^2

fstat <- sigsqli/sigsqhi
Fc <- qf(0.99, dfhi, dfli)
fstat < Fc
```

Model of married workers

```
> summary(eqli)

Call:
lm(formula = log(wage) ~ educ + exper + I(exper^2) + I(exper *
    educ), data = li)

Residuals:
     Min       1Q   Median       3Q      Max
-2.37423 -0.34481  0.00957  0.34195  1.44652

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      0.9196961  0.3557963   2.585 0.009986 **
educ             0.1008275  0.0221957   4.543 6.77e-06 ***
exper            0.0506938  0.0149271   3.396 0.000731 ***
I(exper^2)      -0.0007088  0.0001379  -5.141 3.75e-07 ***
I(exper * educ) -0.0004620  0.0007478  -0.618 0.536990
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5353 on 576 degrees of freedom
Multiple R-squared:  0.2109,	Adjusted R-squared:  0.2054
F-statistic: 38.48 on 4 and 576 DF,  p-value: < 2.2e-16
```

Model of unmarried workers

```
> summary(eqhi)

Call:
lm(formula = log(wage) ~ educ + exper + I(exper^2) + I(exper *
    educ), data = hi)

Residuals:
     Min       1Q   Median       3Q      Max
-1.26823 -0.30372 -0.06065  0.29208  1.44465

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      0.1974877  0.2944715   0.671  0.50282
educ             0.1512920  0.0194232   7.789 5.48e-14 ***
exper            0.0728360  0.0127057   5.733 1.91e-08 ***
I(exper^2)      -0.0007014  0.0001193  -5.880 8.45e-09 ***
I(exper * educ) -0.0021448  0.0006538  -3.280  0.00112 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4614 on 414 degrees of freedom
Multiple R-squared:  0.2753,	Adjusted R-squared:  0.2683
F-statistic: 39.32 on 4 and 414 DF,  p-value: < 2.2e-16
```

```
> fstat < Fc
[1] FALSE
```

Since the f-stat is more than the f-critical value, we reject the null hypothesis and conclude that the <u>variances are different</u> for married and unmarried workers.

5. Find generalized least squares of the model in part (1). Compare the estimates and standard errors with those obtained in part (1) **using traditional OLS with the White's correction**. You can also apply GLS to the model in part (2) which includes MARRIED, is MARRIED significant? Can you exclude it from part (2) and focus on the model in part (1)?

```
w <- 1/lm(abs(model_ols$residuals) ~ model_ols$fitted.values)$fitted.values^2

model_fgls_known <- lm(log(wage) ~ educ + exper + I(exper^2) + I(exper*educ), weights=w, data=data_1)
summary(model_fgls_known)
model_heterorobust
```

```
> summary(model_fgls_known)

Call:
lm(formula = log(wage) ~ educ + exper + I(exper^2) + I(exper *
    educ), data = data_1, weights = w)

Weighted Residuals:
    Min      1Q  Median      3Q     Max
-5.6624 -0.8245 -0.0686  0.8489  3.7166

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      5.056e-01  2.233e-01   2.264  0.02376 *
educ             1.290e-01  1.457e-02   8.849  < 2e-16 ***
exper            6.447e-02  9.379e-03   6.874  1.1e-11 ***
I(exper^2)      -7.149e-04  8.626e-05  -8.288  3.7e-16 ***
I(exper * educ) -1.430e-03  4.880e-04  -2.931  0.00346 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.273 on 995 degrees of freedom
Multiple R-squared:  0.243,	Adjusted R-squared:  0.2399
F-statistic: 79.83 on 4 and 995 DF,  p-value: < 2.2e-16
```

```
> model_heterorobust

t test of coefficients:

                   Estimate  Std. Error t value  Pr(>|t|)
(Intercept)      5.2968e-01  2.5283e-01  2.0950   0.03642 *
educ             1.2720e-01  1.6960e-02  7.4999 1.413e-13 ***
exper            6.2981e-02  1.1378e-02  5.5355 3.969e-08 ***
I(exper^2)      -7.1394e-04  9.2013e-05 -7.7591 2.114e-14 ***
I(exper * educ) -1.3224e-03  6.3679e-04 -2.0766   0.03809 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We see that estimates in the GLS model are a little bit bigger in absolute values, signs are the same, and se are less. So we can conclude that the GLS model is better than the OLS model.

```
w2 <- 1/lm(abs(model_ols2$residuals) ~ model_ols2$fitted.values)$fitted.values^2

model_fgls_known2 <- lm(log(wage) ~ educ + exper + I(exper^2) + I(exper*educ) + married, weights=w2, data=data_1)
summary(model_fgls_known2)
```

```
> summary(model_fgls_known2)

Call:
lm(formula = log(wage) ~ educ + exper + I(exper^2) + I(exper *
    educ) + married, data = data_1, weights = w2)

Weighted Residuals:
    Min      1Q  Median      3Q     Max
-5.6968 -0.8155 -0.0619  0.8393  3.6875

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      5.114e-01  2.226e-01   2.297  0.02181 *
educ             1.283e-01  1.456e-02   8.812  < 2e-16 ***
exper            6.312e-02  9.434e-03   6.691 3.70e-11 ***
I(exper^2)      -6.938e-04  8.756e-05  -7.924 6.15e-15 ***
I(exper * educ) -1.440e-03  4.863e-04  -2.962  0.00313 **
married          4.197e-02  3.368e-02   1.246  0.21302
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.277 on 994 degrees of freedom
Multiple R-squared:  0.2439,	Adjusted R-squared:  0.2401
F-statistic: 64.14 on 5 and 994 DF,  p-value: < 2.2e-16
```

The term married is insignificant in the GLS model, so we can exclude it.

6. Find two 95% interval estimates for the marginal effect $\partial E(\ln(WAGE))/\partial EDUC$ for a worker with 12 years of education and 25 years of experience. Use the results from part (1) **with the White's correction** for one interval and the results from part (5) **GLS results** for the other interval. Comment on any differences.

```
lambda = as.numeric(model_heterorobust[2]) + as.numeric(model_heterorobust[5])*25
lambda
se = sqrt(cov1[7] + 625*cov1[25]+2*25*cov1[10])
t_crit = 1.645
interval_1 = c(lambda-se*t_crit, lambda+se*t_crit)
interval_1
```

```
lambda_2 = as.numeric(model_fgls_known$coefficients[2]) + as.numeric(model_fgls_known$coefficients[5])*25
lambda_2
se_2 = sqrt(vcov(model_fgls_known)[7] + 625*vcov(model_fgls_known)[25]+2*25*vcov(model_fgls_known)[10])
t_crit = 1.645
interval_2 = c(lambda_2-se_2*t_crit, lambda_2+se_2*t_crit)
interval_2
```

The interval of White's correction

```
> interval_1
[1] 0.08349394 0.10477736
```

The interval of GLS results

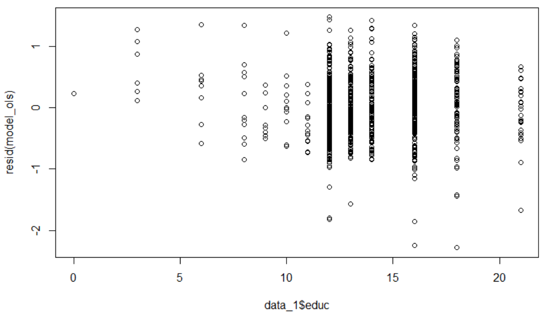
```
> interval_2
[1] 0.08325823 0.10316683
```

The interval estimates are pretty similar. Probably, the estimates from the GLS model are slightly less than those from White's correction.

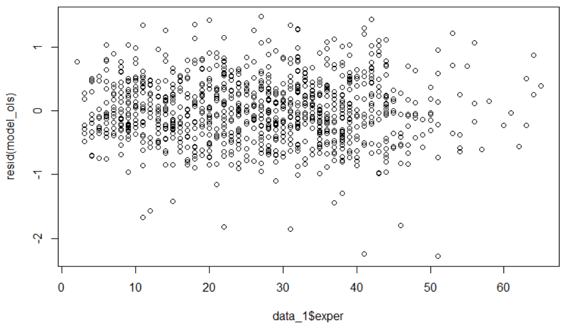
7. Using the model in part (1), plot the least squares residuals against EDUC and against EXPER. What do they suggest?

```
plot(data_1$educ, resid(model_ols))
plot(data_1$exper, resid(model_ols))
```

Plot residuals against EDUC



Plot residuals against EXPER



It seems like variance changes with changes in educ and exper terms. We can guess that there are heteroskedasticity.

8. Using the model in part (1), test for heteroskedasticity using a Breusch-Pagan test. Now the variance depends on EDUC, EXPER and MARRIED. What do you conclude at a 5% significance level?

```
ressq <- resid(model_ols)^2
modres <- lm(ressq ~ educ + exper + married, data=data_1)
N <- nobs(modres)
gmodres <- glance(modres)
S <- gmodres$df
Rsqres <- gmodres$r.squared
chisq <- N*Rsqres
pval <- 1-pchisq(chisq,S)
pval
```

```
> pval
[1] 0.002146327
```

P-value (0.002) is less than 0.05, which means that heteroskedasticity exists.

9. Use the model in part (1) to extract residuals. Now estimate a variance function with unknown function form that includes EDUC, EXPER, and MARRIED and use it to estimate the standard deviation for each observation and list the first ten estimates. Hint: Don't take log of EDUC, EXPER, and MARRIED.

```
ehatsq <- resid(model_ols)^2

sighatsq.ols<- lm(log(ehatsq)~educ + exper + married, data=data_1)
summary(sighatsq.ols)
vari <- exp(fitted(sighatsq.ols))
sd <- sqrt(vari)
sd[1:10]
```

```
> summary(sighatsq.ols)

Call:
lm(formula = log(ehatsq) ~ educ + exper + married, data = data_1)

Residuals:
     Min       1Q   Median       3Q      Max 
-15.7817  -0.9769   0.4807   1.4909   4.1646 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -3.025504   0.413800  -7.312 5.42e-13 ***
educ         0.013910   0.026384   0.527    0.598    
exper        0.005160   0.005635   0.916    0.360    
married      0.045473   0.146057   0.311    0.756    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.229 on 996 degrees of freedom
Multiple R-squared:  0.001283,	Adjusted R-squared:  -0.001725 
F-statistic: 0.4264 on 3 and 996 DF,  p-value: 0.7341
```

First ten estimates of the standard deviation:

```
> sd[1:10]
        1         2         3         4         5         6         7         8         9        10 
0.2785646 0.2495720 0.2604898 0.2498239 0.2794415 0.2647028 0.2721703 0.2674490 0.2728735 0.2612313 
```

10. Use the model in part (1). Find generalized least squares estimates of the wage equation **based on findings in (9)**. Compare the GLS estimates and standard errors with those obtained from OLS estimation with heteroskedasticity-robust standard errors (White's correction).

```
model_fgls <- lm(log(wage) ~ educ + exper + I(exper^2) + I(exper*educ), weights=1/sd, data=data_1)
summary(model_fgls)
stargazer(model_ols, model_heterorobust, model_fgls,
          header=FALSE,
          title="Comparing various 'wage' models",
          type="text",
          keep.stat="n",
          omit.table.layout="n",
          star.cutoffs=NA,
          digits=3,
          intercept.bottom=FALSE,
          column.labels=c("OLS","WHITE","FGLS"),
          dep.var.labels.include = FALSE,
          model.numbers = FALSE,
          dep.var.caption="Dependent variable: 'wage'",
          model.names=FALSE,
          star.char=NULL)
```

```
> summary(model_fgls)

Call:
lm(formula = log(wage) ~ educ + exper + I(exper^2) + I(exper * 
    educ), data = data_1, weights = 1/sd)

Weighted Residuals:
    Min      1Q  Median      3Q     Max 
-4.2408 -0.6435 -0.0510  0.6610  2.8687 

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)     5.283e-01  2.235e-01   2.364  0.01825 *  
educ            1.273e-01  1.455e-02   8.751  < 2e-16 ***
exper           6.329e-02  9.483e-03   6.674 4.13e-11 ***
I(exper^2)     -7.143e-04  8.836e-05  -8.085 1.80e-15 ***
I(exper * educ) -1.345e-03  4.936e-04  -2.724  0.00655 ** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9824 on 995 degrees of freedom
Multiple R-squared:  0.2449,	Adjusted R-squared:  0.2419 
F-statistic: 80.69 on 4 and 995 DF,  p-value: < 2.2e-16
```

```
Comparing various 'wage' models
=============================================
                     Dependent variable: 'wage'
                    ---------------------------
                      OLS      WHITE     FGLS
---------------------------------------------
Constant             0.530     0.530     0.528
                    (0.227)   (0.253)   (0.223)

educ                 0.127     0.127     0.127
                    (0.015)   (0.017)   (0.015)

exper                0.063     0.063     0.063
                    (0.010)   (0.011)   (0.009)

I(exper2)           -0.001    -0.001    -0.001
                   (0.0001)  (0.0001)  (0.0001)

I(exper * educ)     -0.001    -0.001    -0.001
                   (0.0005)   (0.001)  (0.0005)

---------------------------------------------
Observations         1,000               1,000
=============================================
```

GLS estimates did not change much. However, standard errors are less.

11. Use the model in part (1). Find two 95% interval estimates for the marginal effect $\partial E(\ln(\text{WAGE}))/\partial EXPER$ for a worker with 16 years of education and 20 years of experience. Use least squares with heteroskedasticity-robust standard errors for one interval and the results from part (10) for the other. Comment on any difference.

```
lambda_exper1 = (as.numeric(model_heterorobust[3])
                +20*as.numeric(model_heterorobust[4])
                +16*as.numeric(model_heterorobust[5]))
se_exper1 = sqrt(cov1[13] + 400*cov1[19] + 256*cov1[25]
                + 2*20*cov1[18] + 2*16*cov1[23]+2*16*20*cov1[20])
t_crit = 1.645
interval_exper1 = c(lambda_exper1-se_exper1*t_crit, lambda_exper1+se_exper1*t_crit)
interval_exper1
```

```
lambda_exper2 = (as.numeric(model_fgls$coefficients[3])
                + 20*as.numeric(model_fgls$coefficients[4])
                + 16*as.numeric(model_fgls$coefficients[5]))
se_exper2 = sqrt(vcov(model_fgls)[13] + 400*vcov(model_fgls)[19]
                + 256*vcov(model_fgls)[25] + 2*20*vcov(model_fgls)[18]
                + 2*16*vcov(model_fgls)[23]+2*16*20*vcov(model_fgls)[20])
t_crit = 1.645
interval_exper2 = c(lambda_exper2-se_exper2*t_crit, lambda_exper2+se_exper2*t_crit)
interval_exper2
```

The interval from White's correction model

```
> interval_exper1
[1] 0.02264277 0.03244467
```

The interval from GLS model (unknown form)

```
> interval_exper2
[1] 0.02230892 0.03266676
```

Intervals are pretty similar.

12. Use the model in part (2). Forecast the wage of a married worker with 18 years of education and 16 years of experience. Use both the natural predictor and the corrected predictor.

```
pred_wage <- as.numeric(exp(model_ols2$coefficients[1]
                        + model_ols2$coefficients[2]*18 + model_ols2$coefficients[3]*16
                        + model_ols2$coefficients[4]*16*16 + 16*18*model_ols2$coefficients[5]
                        + model_ols2$coefficients[6]))
pred_wage
corrected_pred <- pred_wage * exp(summary(model_ols2)$sigma^2/2)
corrected_pred
```

Natural predictor:

```
> pred_wage
[1] 26.548
```

Corrected predictor:

```
> corrected_pred
[1] 30.16705
```

13. Are you happy about the above model? Do you have any other ideas to improve the model?

The model is pretty good. Adding new variables and changing the functional form of the model do not improve the model significantly.