# Exercises 4: More Regression with R

*Peter Tempfli*

*3/1/2019*

# 1, Consider the following model:

$Y = \beta 0 + \beta 1X1 + \beta 2X2 + X3\beta 3 + U$

## a, H0 : β0 = 0 & β1 = 0 & β2 = 0 & β3 = 0

```
R = diag(4)
r = cbind(c(0,0,0,0))
R
```

```
##      [,1] [,2] [,3] [,4]
## [1,]    1    0    0    0
## [2,]    0    1    0    0
## [3,]    0    0    1    0
## [4,]    0    0    0    1
```

```
r
```

```
##      [,1]
## [1,]    0
## [2,]    0
## [3,]    0
## [4,]    0
```

## b H0 : β0 = 0 & β1 = 0

```
R = rbind(c(1,0,0,0), c(0,1,0,0))
r = cbind(c(0,0))
R
```

```
##      [,1] [,2] [,3] [,4]
## [1,]    1    0    0    0
## [2,]    0    1    0    0
```

```
r
```

```
##      [,1]
## [1,]    0
## [2,]    0
```

## c H0 : β0 = 1 & β1 = 1

```
R = rbind(c(1,0,0,0), c(0,1,0,0))
r = cbind(c(1,1))
R
```

```
##      [,1] [,2] [,3] [,4]
## [1,]    1    0    0    0
## [2,]    0    1    0    0
```

```
r
```

```
##      [,1]
## [1,]    1
## [2,]    1
```

## d H0 : β1 = 0

```
R = rbind(c(0,1,0,0))
r = cbind(c(0))
R
```

```
##      [,1] [,2] [,3] [,4]
## [1,]    0    1    0    0
```

```
r
```

```
##      [,1]
## [1,]    0
```

## e H0 : β1 + β2 = 1

```
R = rbind(c(0,1,1,0))
r = cbind(c(1))
R
```

```
##      [,1] [,2] [,3] [,4]
## [1,]    0    1    1    0
```

```
r
```

```
##      [,1]
## [1,]    1
```

# 2 - models

$Y = \exp(X\beta + U)$ (1)

$Y = X\beta + U$ (2)

```
data("CPS1985")
CPS1985$experience_2 <- CPS1985$experience^2
CPS1985$experience_3 <- CPS1985$experience^3
CPS1985$experience_4 <- CPS1985$experience^4
CPS1985$experience_5 <- CPS1985$experience^5

mod1 <- lm(log(wage) ~ education + married + gender + experience + experience_2 + exp
erience_3, data = CPS1985)
summary(mod1)
```

```
##
## Call:
## lm(formula = log(wage) ~ education + married + gender + experience +
##     experience_2 + experience_3, data = CPS1985)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.23291 -0.27937  0.01609  0.27946  2.22333
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   4.643e-01  1.290e-01   3.599 0.000349 ***
## education     9.292e-02  7.997e-03  11.620  < 2e-16 ***
## marriedyes    3.871e-02  4.322e-02   0.896 0.370808
## genderfemale -2.522e-01  3.853e-02  -6.544 1.42e-10 ***
## experience    6.105e-02  1.200e-02   5.089 5.00e-07 ***
## experience_2 -1.990e-03  5.937e-04  -3.352 0.000861 ***
## experience_3  2.139e-05  8.394e-06   2.548 0.011103 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4418 on 527 degrees of freedom
## Multiple R-squared:  0.3072, Adjusted R-squared:  0.2993
## F-statistic: 38.94 on 6 and 527 DF,  p-value: < 2.2e-16
```

```
mod2 <- lm(wage ~ education + married + gender + experience + experience_2 + experien
ce_3, data = CPS1985)
summary(mod2)
```

```
##
## Call:
## lm(formula = wage ~ education + married + gender + experience +
##     experience_2 + experience_3, data = CPS1985)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -9.701 -2.558 -0.538  1.815 39.155
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -5.385e+00  1.290e+00  -4.174 3.51e-05 ***
## education     9.042e-01  7.999e-02  11.305  < 2e-16 ***
## marriedyes    2.082e-01  4.323e-01   0.482 0.630268
## genderfemale -2.319e+00  3.854e-01  -6.017 3.33e-09 ***
## experience    4.017e-01  1.200e-01   3.348 0.000872 ***
## experience_2 -1.128e-02  5.939e-03  -1.899 0.058130 .
## experience_3  1.131e-04  8.396e-05   1.347 0.178521
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.419 on 527 degrees of freedom
## Multiple R-squared:  0.269,  Adjusted R-squared:  0.2607
## F-statistic: 32.33 on 6 and 527 DF,  p-value: < 2.2e-16
```

# Model diagnostics

I chose the *First (exponential)* model, because:

1. parameters are more significant (so parameters have a bigger influance on Y)
2. R-squered is slightly better (so the model fits better the actual databoints). Both models have the same number of parameters, so Adjusted R-squared is not really important here
3. Both models have similar p-values (2.2e-16) – this is a very low value, so the probability that the model's result is only by chanse are slow.

# c heteroskedasticity- robust standar errors and P-values

(the first model (exponential) was choosen, called `mod1`

```
summary(mod1, robust=T)
```

```
##
## Call:
## lm(formula = log(wage) ~ education + married + gender + experience +
##     experience_2 + experience_3, data = CPS1985)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.23291 -0.27937  0.01609  0.27946  2.22333
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   4.643e-01  1.290e-01   3.599 0.000349 ***
## education     9.292e-02  7.997e-03  11.620  < 2e-16 ***
## marriedyes    3.871e-02  4.322e-02   0.896 0.370808
## genderfemale -2.522e-01  3.853e-02  -6.544 1.42e-10 ***
## experience    6.105e-02  1.200e-02   5.089 5.00e-07 ***
## experience_2 -1.990e-03  5.937e-04  -3.352 0.000861 ***
## experience_3  2.139e-05  8.394e-06   2.548 0.011103 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4418 on 527 degrees of freedom
## Multiple R-squared:  0.3072, Adjusted R-squared:  0.2993
## F-statistic: 38.94 on 6 and 527 DF,  p-value: < 2.2e-16
```

# d Removing outliers

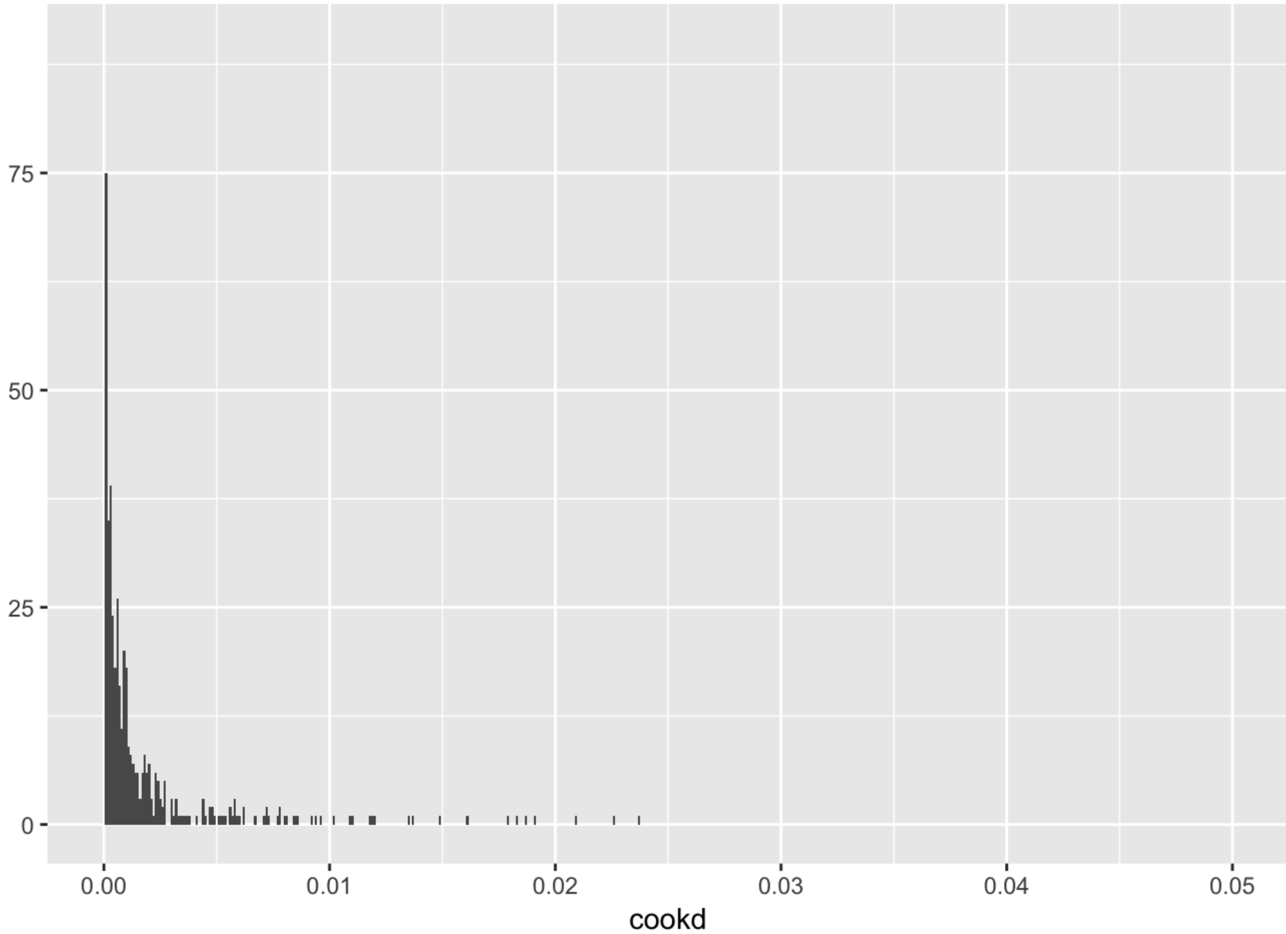
1. I'm going to visualize the Cook distances for every datapoint

```
cookd <- cooks.distance(mod2)
summary(cookd)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## 0.0000000 0.0001042 0.0004558 0.0019695 0.0014177 0.2101321
```

```
qplot(cookd, binwidth = 0.0001, xlim=c(0, 0.05))
```

```
## Warning: Removed 1 rows containing non-finite values (stat_bin).
```




I want to cut the outliers at 3rd quanticile, so i create the boolean-selector:

```
cutter_at_3rd_qu <- cookd < 0.0014177
```

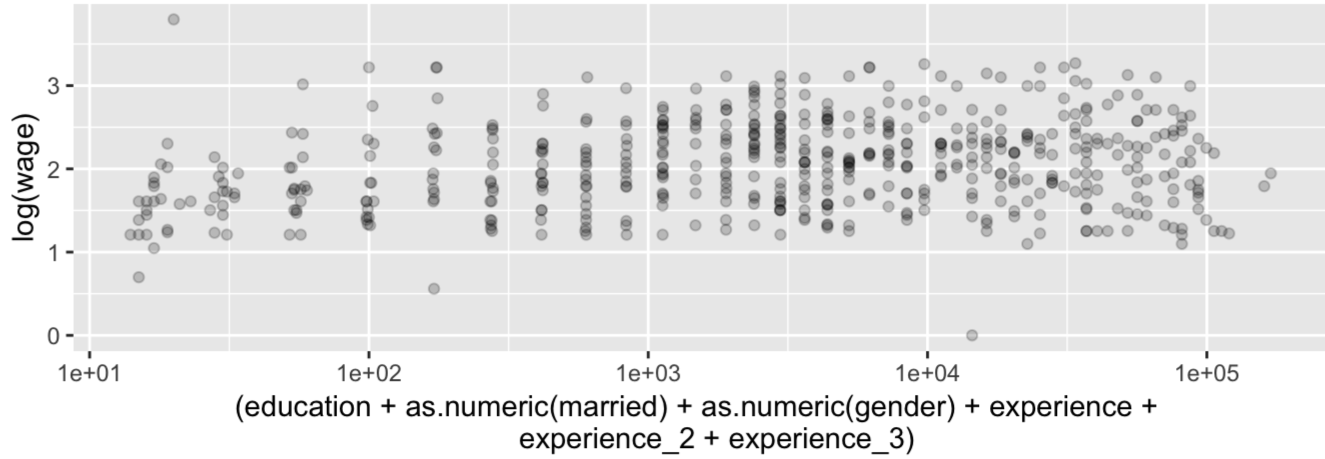
Visaulize the original set and the cutted set:

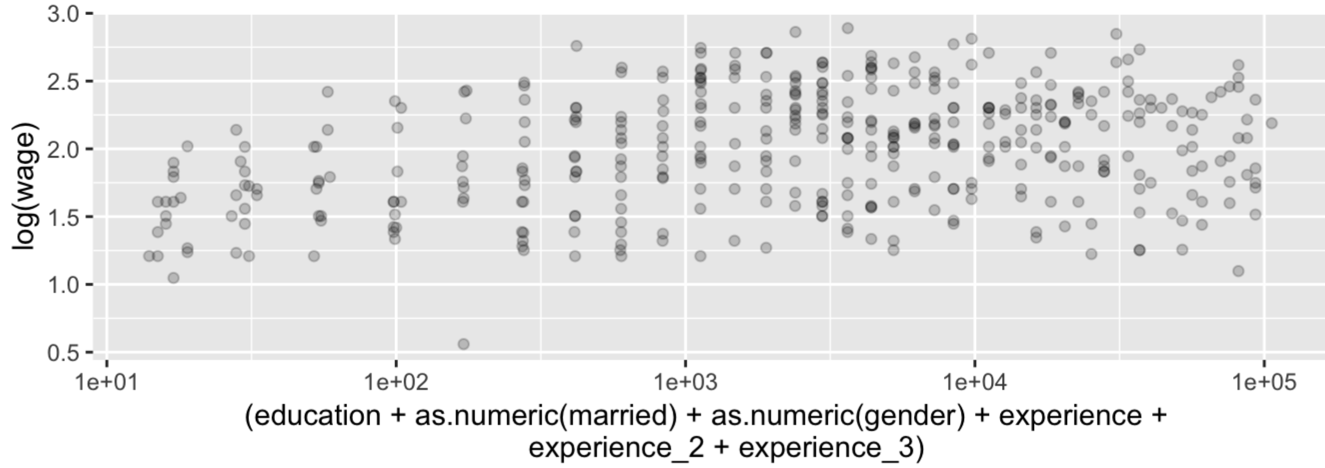
```
cutted_data <- CPS1985[cutter_at_3rd_qu,]

p1 <- ggplot(data=CPS1985, aes(x= (education + as.numeric(married) + as.numeric(gende
r) + experience + experience_2 + experience_3), y=log(wage))) +
  geom_jitter(alpha=0.2) +
  scale_x_log10()

p2 <- ggplot(data=cutted_data, aes(x= (education + as.numeric(married) + as.numeric(g
ender) + experience + experience_2 + experience_3), y=log(wage))) +
  geom_jitter(alpha=0.2) +
  scale_x_log10()

grid.arrange(p1, p2)
```




Re-building the model with the filtered dataset

```
mod1_filtered <- lm(log(wage) ~ education + married + gender + experience + experienc
e_2 + experience_3, data = cutted_data)
```

Testing original `mod1` against the same model run on filtered data `mod1_filtered` :

```
summary(mod1)
```

```
##
## Call:
## lm(formula = log(wage) ~ education + married + gender + experience +
##     experience_2 + experience_3, data = CPS1985)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.23291 -0.27937  0.01609  0.27946  2.22333
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   4.643e-01  1.290e-01   3.599 0.000349 ***
## education     9.292e-02  7.997e-03  11.620  < 2e-16 ***
## marriedyes    3.871e-02  4.322e-02   0.896 0.370808
## genderfemale -2.522e-01  3.853e-02  -6.544 1.42e-10 ***
## experience    6.105e-02  1.200e-02   5.089 5.00e-07 ***
## experience_2 -1.990e-03  5.937e-04  -3.352 0.000861 ***
## experience_3  2.139e-05  8.394e-06   2.548 0.011103 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4418 on 527 degrees of freedom
## Multiple R-squared:  0.3072, Adjusted R-squared:  0.2993
## F-statistic: 38.94 on 6 and 527 DF,  p-value: < 2.2e-16
```

```
summary(mod1_filtered)
```

```
##
## Call:
## lm(formula = log(wage) ~ education + married + gender + experience +
##     experience_2 + experience_3, data = cutted_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.96535 -0.19780  0.00798  0.22482  0.61574
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.120e-01  1.023e-01   1.095 0.274022
## education     1.158e-01  6.908e-03  16.762  < 2e-16 ***
## marriedyes    7.158e-03  3.269e-02   0.219 0.826812
## genderfemale -2.921e-01  2.885e-02 -10.123  < 2e-16 ***
## experience    7.360e-02  1.022e-02   7.204 3.01e-12 ***
## experience_2 -2.536e-03  5.522e-04  -4.592 5.91e-06 ***
## experience_3  3.003e-05  8.353e-06   3.595 0.000366 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2851 on 393 degrees of freedom
## Multiple R-squared:  0.5519, Adjusted R-squared:  0.5451
## F-statistic: 80.69 on 6 and 393 DF,  p-value: < 2.2e-16
```

# Conclusion

1. Filtered data has slightnly better significance values, but not a big difference. This is the right behaviour, because we are using the same dataset.
2. Filered data has significantly better R-squared value (0.5 against 0.3)
3. The conclusion is that filtering outliers makes the model fit better. However, if filtering too much outliers, it can result in a model which doesn't fit to the real world data.

# 3, Polynomial model of degree five

```
mod5 <- lm(log(wage) ~ education + married + gender + experience + experience_2 + exp
erience_3 + experience_4 + experience_5, data = CPS1985)

mod_lin <- lm(log(wage) ~ education + married + gender + experience, data = CPS1985)

summary(mod5)
```

```
##
## Call:
## lm(formula = log(wage) ~ education + married + gender + experience +
##     experience_2 + experience_3 + experience_4 + experience_5,
##     data = CPS1985)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.21983 -0.27947  0.01676  0.28669  2.25429
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   4.162e-01  1.392e-01   2.990  0.00292 **
## education     9.217e-02  8.075e-03  11.414  < 2e-16 ***
## marriedyes    3.519e-02  4.343e-02   0.810  0.41812
## genderfemale -2.525e-01  3.861e-02  -6.541 1.46e-10 ***
## experience    9.297e-02  3.653e-02   2.545  0.01120 *
## experience_2 -5.978e-03  4.497e-03  -1.329  0.18429
## experience_3  2.116e-04  2.250e-04   0.941  0.34736
## experience_4 -3.817e-06  4.847e-06  -0.787  0.43140
## experience_5  2.710e-08  3.741e-08   0.724  0.46909
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4422 on 525 degrees of freedom
## Multiple R-squared:  0.3083, Adjusted R-squared:  0.2978
## F-statistic: 29.25 on 8 and 525 DF,  p-value: < 2.2e-16
```

```
summary(mod_lin)
```

```
##
## Call:
## lm(formula = log(wage) ~ education + married + gender + experience,
##     data = CPS1985)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.18058 -0.30892  0.01226  0.30221  2.03436
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.648276   0.120773   5.368 1.20e-07 ***
## education     0.096871   0.008001  12.108  < 2e-16 ***
## marriedyes    0.090522   0.042744   2.118   0.0347 *
## genderfemale -0.254981   0.039283  -6.491 1.97e-10 ***
## experience    0.011639   0.001760   6.614 9.18e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4508 on 529 degrees of freedom
## Multiple R-squared:  0.2758, Adjusted R-squared:  0.2703
## F-statistic: 50.36 on 4 and 529 DF,  p-value: < 2.2e-16
```

# 3 a,

On the linear model the only one linear parameter of `experience` is much more significant that the same paramater on 1 to 5 degree; meanwhile the adjusted R-squared is similar; so we should choose the simpler model.

# 3 b,

Joint test 5-degree X Linear:

```
R = rbind(c(0,1,0,0,0,0,0,0,0),c(0,0,1,0,0,0,0,0,0),c(0,0,0,1,0,0,0,0,0),c(0,0,0,0,1,
0,0,0,0))
r = cbind(c(0,0,0,0))
R
```

```
##      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9]
## [1,]    0    1    0    0    0    0    0    0    0
## [2,]    0    0    1    0    0    0    0    0    0
## [3,]    0    0    0    1    0    0    0    0    0
## [4,]    0    0    0    0    1    0    0    0    0
```

```
r
```

```
##      [,1]
## [1,]    0
## [2,]    0
## [3,]    0
## [4,]    0
```

```
lht(mod5, R, rhs=r)
```

```
## Linear hypothesis test
##
## Hypothesis:
## education = 0
## marriedyes = 0
## genderfemale = 0
## experience = 0
##
## Model 1: restricted model
## Model 2: log(wage) ~ education + married + gender + experience + experience_2 +
##     experience_3 + experience_4 + experience_5
##
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    529 138.69
## 2    525 102.68  4    36.006 46.025 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

# 3 c,

Joint test 5-degree X 3-degree:

```
R = rbind(c(0,1,0,0,0,0,0,0,0),c(0,0,1,0,0,0,0,0,0),c(0,0,0,1,0,0,0,0,0),c(0,0,0,0,1,
0,0,0,0),c(0,0,0,0,0,1,0,0,0),c(0,0,0,0,0,0,1,0,0))
r = cbind(c(0,0,0,0,0,0))
R
```

```
##      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9]
## [1,]    0    1    0    0    0    0    0    0    0
## [2,]    0    0    1    0    0    0    0    0    0
## [3,]    0    0    0    1    0    0    0    0    0
## [4,]    0    0    0    0    1    0    0    0    0
## [5,]    0    0    0    0    0    1    0    0    0
## [6,]    0    0    0    0    0    0    1    0    0
```

```
r
```

```
##      [,1]
## [1,]    0
## [2,]    0
## [3,]    0
## [4,]    0
## [5,]    0
## [6,]    0
```

```
lht(mod5, R, rhs=r)
```

```
## Linear hypothesis test
##
## Hypothesis:
## education = 0
## marriedyes = 0
## genderfemale = 0
## experience = 0
## experience_2 = 0
## experience_3 = 0
##
## Model 1: restricted model
## Model 2: log(wage) ~ education + married + gender + experience + experience_2 +
##     experience_3 + experience_4 + experience_5
##
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    531 146.56
## 2    525 102.68  6    43.882 37.395 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

## d, Conclusion

Adding the same parameter on higher power does not increades the performance – even the `experience_2` and `experience_3` (which were relevant) got non-relevant. In my option the effect is because the 4th and the 5th degrees makes these parameters too dominant, so they outperform the effect of the other parameters

However it is interesting that we also can add `experience_3 + experience_4 + experience_5` BUT NOT linear and `experience_2` , and get this parameters linear.

Joint hypothesis shows that removing paramers creates worse F-values.

# 4

```
CPS1985$marr_wo <-  ifelse(CPS1985$gender == 'female' & CPS1985$married == 'yes', 1,
0)

mod5_2 <- lm(log(wage) ~ marr_wo + education + married + gender + experience + experi
ence_2 + experience_3 + experience_4 + experience_5, data = CPS1985)
summary(mod5_2)
```

```
##
## Call:
## lm(formula = log(wage) ~ marr_wo + education + married + gender +
##     experience + experience_2 + experience_3 + experience_4 +
##     experience_5, data = CPS1985)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.2494 -0.2846  0.0096  0.2760  2.1819
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.917e-01  1.391e-01   2.817  0.00504 **
## marr_wo      -1.848e-01  8.169e-02  -2.263  0.02406 *
## education     9.029e-02  8.086e-03  11.165  < 2e-16 ***
## marriedyes    1.228e-01  5.806e-02   2.115  0.03488 *
## genderfemale -1.305e-01  6.624e-02  -1.971  0.04930 *
## experience    9.451e-02  3.639e-02   2.597  0.00966 **
## experience_2 -6.348e-03  4.482e-03  -1.416  0.15731
## experience_3  2.327e-04  2.243e-04   1.038  0.29998
## experience_4 -4.279e-06  4.833e-06  -0.885  0.37630
## experience_5  3.048e-08  3.729e-08   0.817  0.41414
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4405 on 524 degrees of freedom
## Multiple R-squared:  0.315,  Adjusted R-squared:  0.3032
## F-statistic: 26.77 on 9 and 524 DF,  p-value: < 2.2e-16
```

# b, Do men have a wage premium from marriage?

- genderfemale beta = -1.305e-0
- marr_wo beta = -1.848e-01

From this model we don’t know: we see that being non-married women has a negative effect; however it doesn’t indicates that being married male has a positive effect. However, we can see this from the following model:

```
CPS1985$marr_male <-  ifelse(CPS1985$gender == 'male' & CPS1985$married == 'yes', 1,
0)
mod5_3 <- lm(log(wage) ~ marr_male + education + married + gender + experience + expe
rience_2 + experience_3 + experience_4 + experience_5, data = CPS1985)
summary(mod5_3)
```

```
##
## Call:
## lm(formula = log(wage) ~ marr_male + education + married + gender +
##     experience + experience_2 + experience_3 + experience_4 +
##     experience_5, data = CPS1985)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.2494 -0.2846  0.0096  0.2760  2.1819
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.917e-01  1.391e-01   2.817  0.00504 **
## marr_male     1.848e-01  8.169e-02   2.263  0.02406 *
## education     9.029e-02  8.086e-03  11.165  < 2e-16 ***
## marriedyes   -6.203e-02  6.097e-02  -1.017  0.30943
## genderfemale -1.305e-01  6.624e-02  -1.971  0.04930 *
## experience    9.451e-02  3.639e-02   2.597  0.00966 **
## experience_2 -6.348e-03  4.482e-03  -1.416  0.15731
## experience_3  2.327e-04  2.243e-04   1.038  0.29998
## experience_4 -4.279e-06  4.833e-06  -0.885  0.37630
## experience_5  3.048e-08  3.729e-08   0.817  0.41414
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4405 on 524 degrees of freedom
## Multiple R-squared:  0.315,  Adjusted R-squared:  0.3032
## F-statistic: 26.77 on 9 and 524 DF,  p-value: < 2.2e-16
```

marr_male beta = 1.848e; so it has a positive effect.