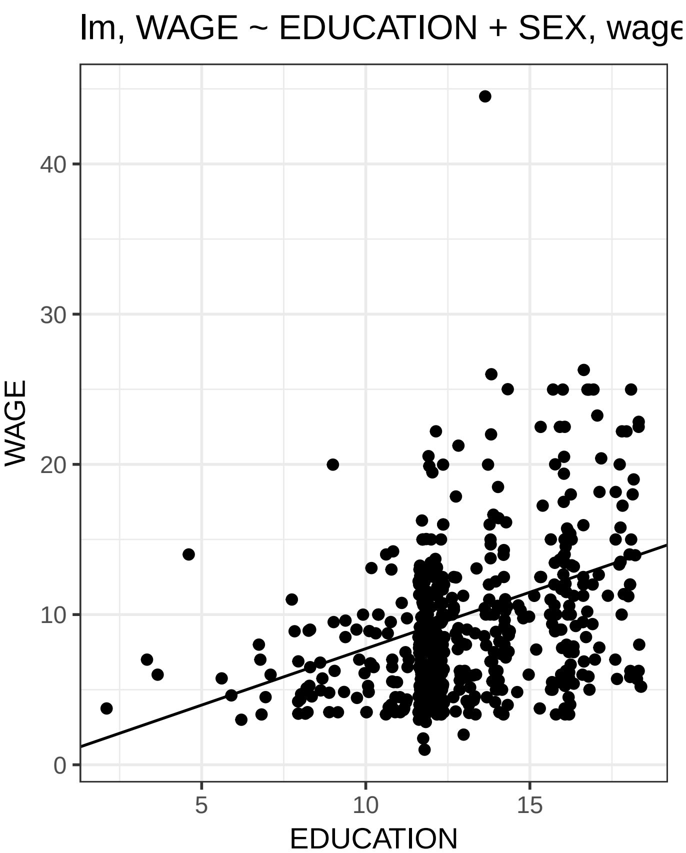
IE5

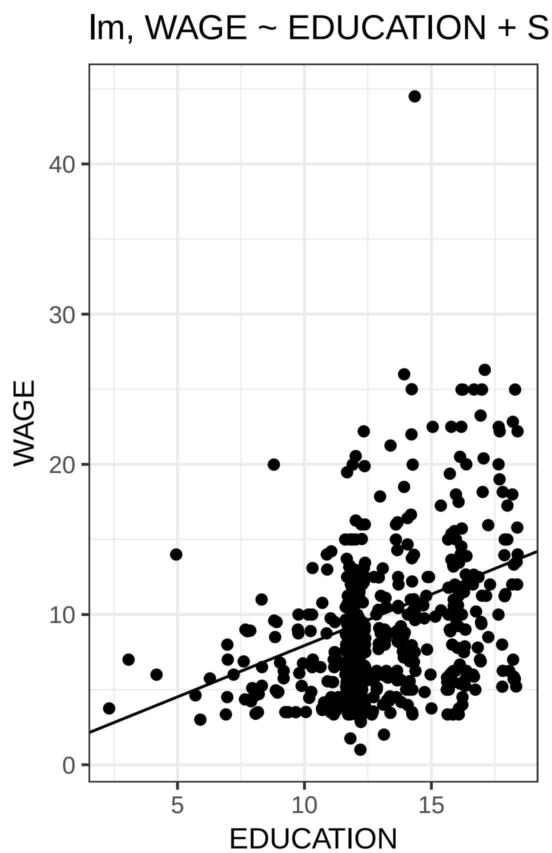
Ben Kaufman
Vic Chan
Ziwei Zhang

Models 1 and 2

$$M1 : E(Wage|Education, Sex) = \beta_{0,1} + \beta_{1,1}Education + \beta_{2,1}Sex \quad (1)$$
$$M2 : E(Wage|Education, Sex) = \beta_{0,2} + \beta_{1,2}Education + \beta_{2,2}Sex + \beta_{3,2}Sex : Education \quad (2)$$

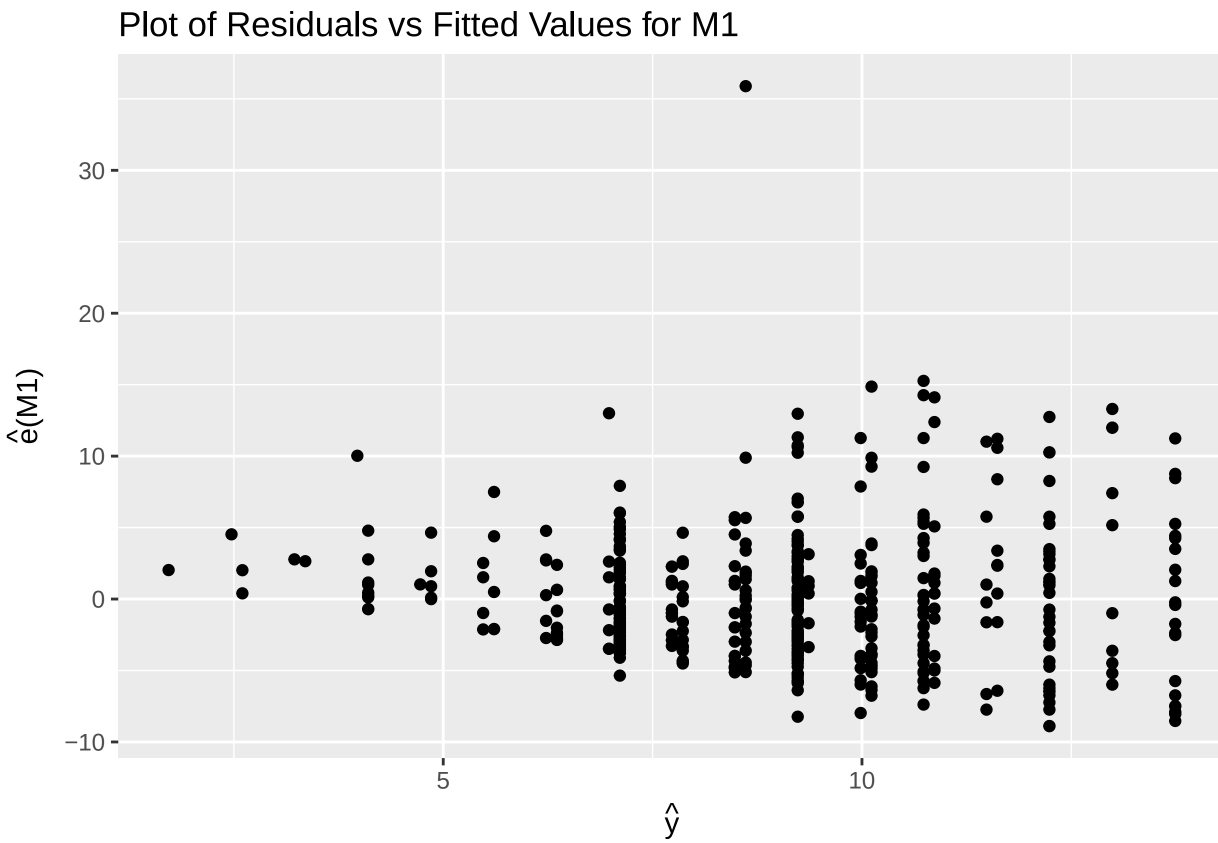
Throw out M2 because interaction term is not significant and hardly changes R^2





Fitted vs Residuals for M1

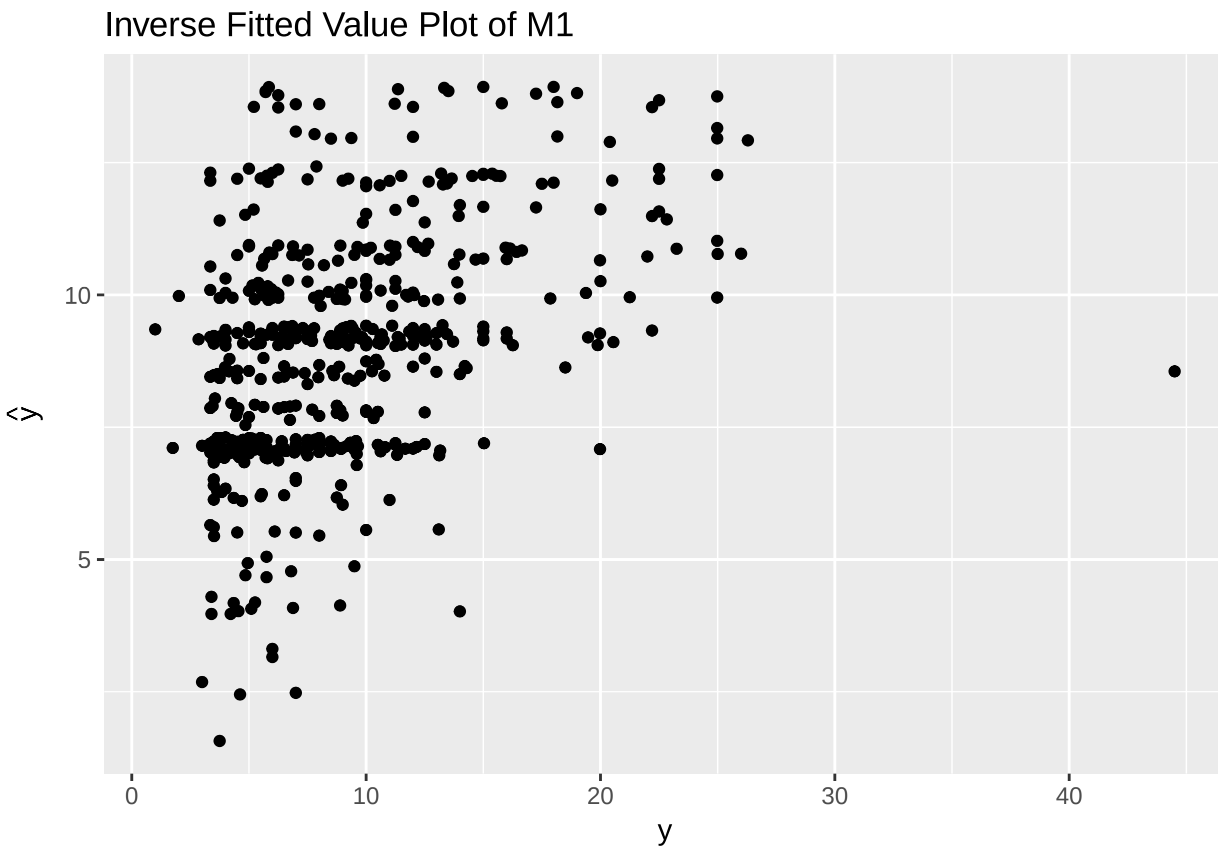
THIS PLOT SUGGESTS NON-CONSTANT VARIANCE, WITHHOLDING TRANSFORMATION UNTIL MORE VARIABLES ADDED TO THE MODEL



Inverse fitted values plot for M1

THIS SUGGESTS A LOG TRANSFORM

```
fitted_M1 <- fitted(M1)
inv_fitt_df <- data.frame(fitted_M1, wages$WAGE)
inv_fitt_value_M1 <- ggplot( inv_fitt_df, aes ( x =wages.WAGE , y = fitted_M1) )
inv_fitt_value_M1 +
  geom_jitter(height=0.2,width=0) +
  xlab("y") + ylab(expression(hat(y))) + ggtitle("Inverse Fitted Value Plot of M1")
```

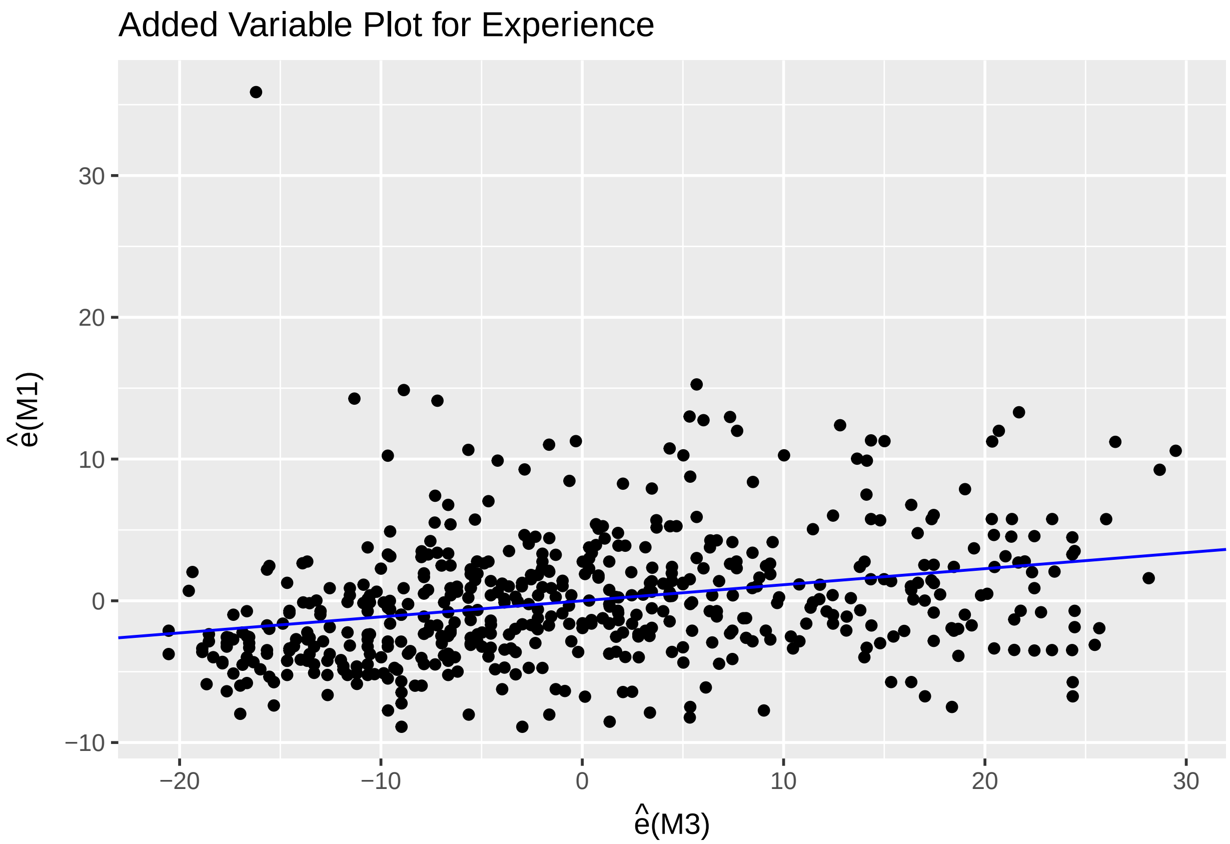


Adding Experience to Our Model

$$M3: E(Wage|Education, Sex, Experience) = \beta_{0,3} + \beta_{1,3}Education + \beta_{2,3}Sex + \beta_{3,3}Experience \quad (3)$$

ADDED VARIABLE PLOT

LOOKS PRETTY LINEAR, R^2 SUGGESTS 8% OF REMAINING VARIANCE EXPLAINED




```
## R squared value of regression on AVP for experience: 0.0798

##
## Call:
## lm(formula = resid_M1 ~ resid_M3, data = avp_df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -9.571 -2.746 -0.653  1.893 37.724
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 3.739e-17  1.924e-01   0.000        1
## resid_M3    1.133e-01  1.668e-02   6.794 2.93e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.445 on 532 degrees of freedom
## Multiple R-squared:  0.07983,    Adjusted R-squared:  0.07811
## F-statistic: 46.16 on 1 and 532 DF,  p-value: 2.931e-11
```

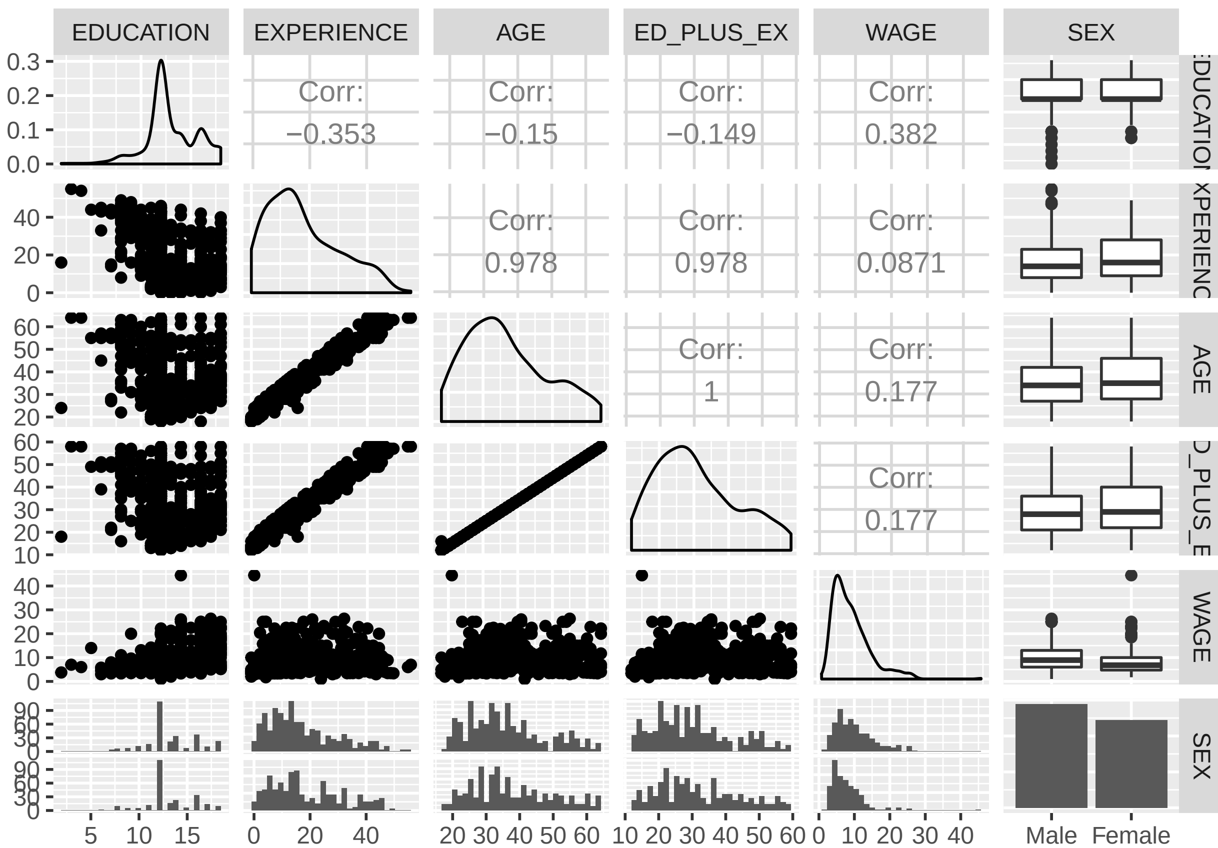
CREATING MODEL 4

$$M4 : E(Wage|Education, Sex, Experience) = \beta_{0,4} + \beta_{1,4}Education + \beta_{2,4}Sex + \beta_{3,4}Experience \tag{4}$$
$$\tag{5}$$

Age and Collinearity

LOOK AT CORRELATION BETWEEN EXPERIENCE AND AGE, EXP+EDU AND AGE, explain why education alone

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



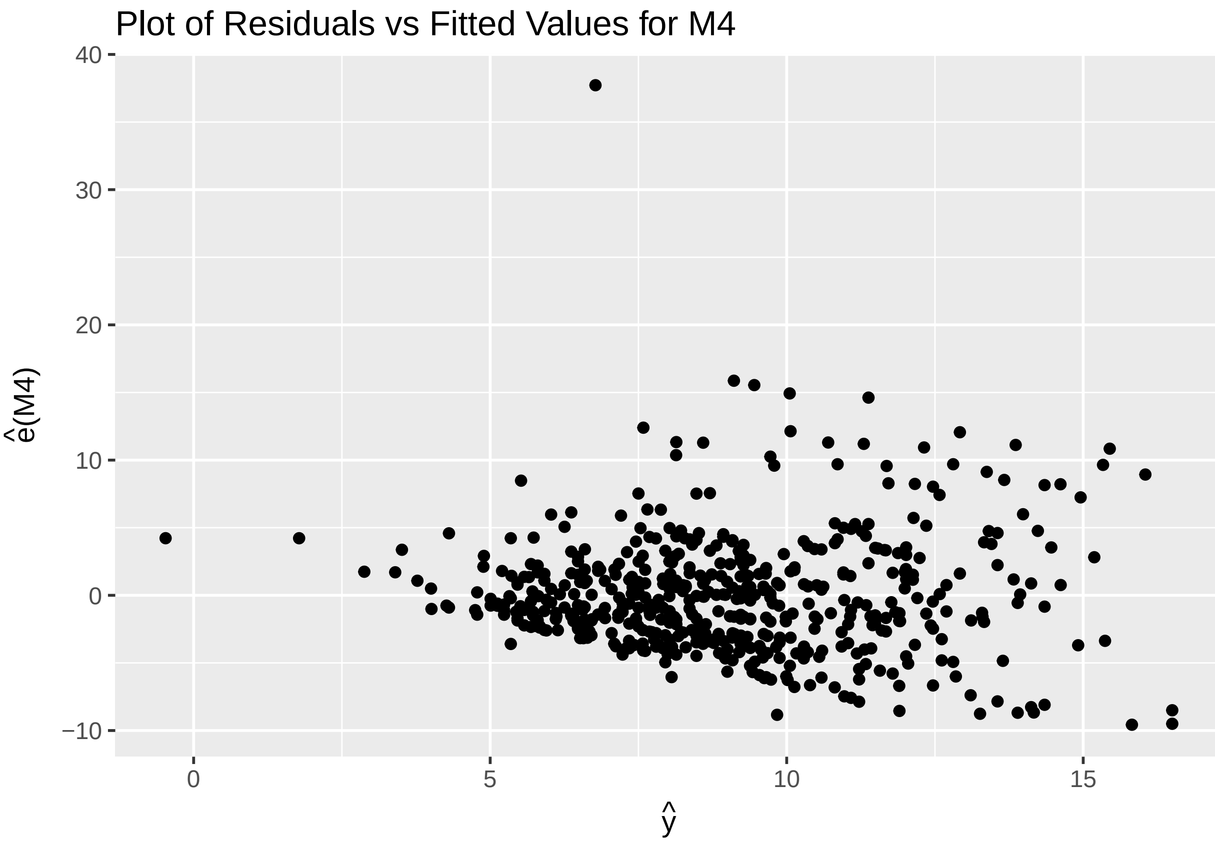
FROM THIS WE SEE THAT AGE = 6 + (EDUCATION + EXPERIENCE) THIS MAKES SENSE BECAUSE ALMOST EVERYONE STARTS FIRST GRADE AT AGE 6 IN THE US

```
##
## Call:
## lm(formula = AGE ~ ED_PLUS_EX, data = wages)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.9805  0.0017  0.0090  0.0147  0.0228
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 5.9674844  0.0210985   282.8   <2e-16 ***
## ED_PLUS_EX  1.0008114  0.0006396  1564.8   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 0.173 on 532 degrees of freedom
## Multiple R-squared:  0.9998, Adjusted R-squared:  0.9998
## F-statistic: 2.448e+06 on 1 and 532 DF,  p-value: < 2.2e-16
```

Variable Transformation

STILL HAVE CONE SHAPE WHICH IMPLIES NONCONSTANT VARIANCE



THIS SUGGESTS A LOG TRANSFORM

USING MODEL 4 IT IS SUGGESTED BY THE BOX COX METHOD THAT WE TRANSFORM OUR RESPONSE VARIABLE USING LOG

```
summary(powerTransform(M4,family="bcPower"))
```

```
## bcPower Transformation to Normality
##    Est Power Rounded Pwr Wald Lwr bnd Wald Upr Bnd
## Y1    -0.008           0      -0.1324       0.1165
##
## Likelihood ratio tests about transformation parameters
##                               LRT df     pval
## LR test, lambda = (0)   0.01574383  1 0.900148
## LR test, lambda = (1) 243.50191305  1 0.000000
```

NEW LOG MODEL

Now that we have properly justified our usage of modeling log(WAGE) rather than wage let's now define a new model:

$$M5 : E(log(Wage)|Education, Sex, Experience) = \beta_{0,5} + \beta_{1,5}Education + \beta_{2,5}Sex + \beta_{3,5}Experience \quad (6)$$

Adding Occupation M6

$$\begin{aligned} M6 : E[log(Wage)|...Occupation] = \beta_{0,6} + \beta_{1,6}Education + \beta_{2,6}Sex + \beta_{3,6}Experience + \\ \beta_{4,6}Management + \beta_{5,6}Sales + \beta_{6,6}Clerical + \beta_{7,6}Service + \beta_{8,6}Professional \end{aligned}$$

$\beta_{2,6}$ is the wage gap between men and women in occupation other

OCCUPATION INTERACTION

$$\begin{aligned} M7 : E[log(Wage)|...Occupation] = {} & \beta_{0,7} + \beta_{1,7}Education + \beta_{2,7}Sex + \beta_{3,7}Experience + \\ & \beta_{4,7}Management + \beta_{5,7}Sales + \beta_{6,7}Clerical + \beta_{7,7}Service + \\ & \beta_{8,7}Professional + \beta_{9,7}Sex : Management + \beta_{10,7}Sex : Sales \\ & \beta_{11,7}Sex : Clerical + \beta_{12,7}Sex : Service + \beta_{13,7}Sex : Professional \end{aligned}$$

$\beta_{2,7}$ is the average difference in wage in occupation other $\beta_{4:8,7}$ are the differences in wages for each occupation for men $\beta_{9:13,7}$ are the differences in M/F wages compared to difference in occupation other

~~~~Ignore Below This Point~~~~

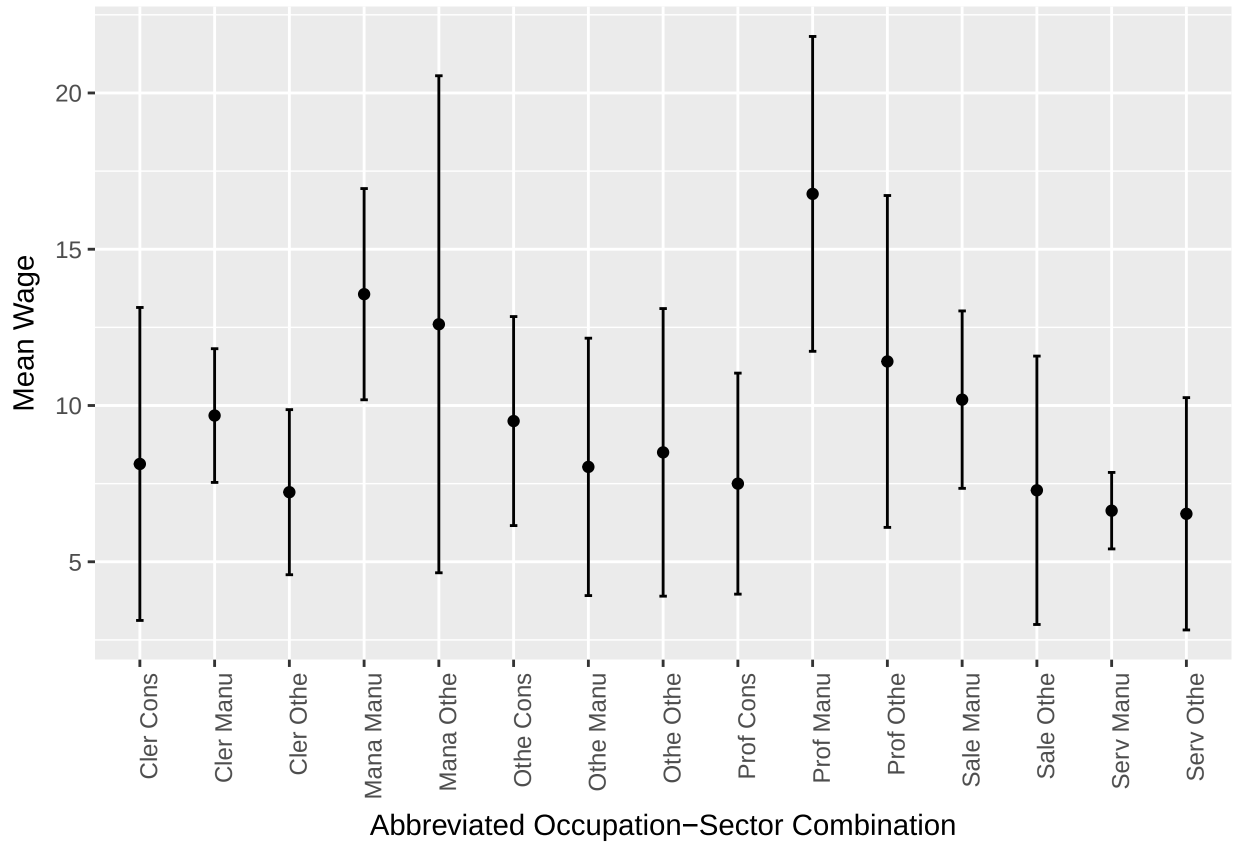
Exploring Occupation

Now that we are accounting for a person's education and experience we can now try to explore the effect of a person's occupation on wage. Intuitively you can think to yourself, does a hamburger flipper make as much as a CEO of a major company; the answer is definitely not. We can apply at least some level of reasoning to this claim. Considering our current model, which accounts for education and experience, it is insufficient to claim that the CEO makes more than the hamburger flipper because the CEO has a PHD and the hamburger flipper didn't graduate high school nor is it sufficient to say that the CEO has had more experience nor is any combination of the prior two arguments. At this point we claim that the occupation you are in can determine your wage independently (enough) of education and experience. Some reasons may include the level of physical risk involved in job, for example, if there are two jobs that require the same level of education and experience and one job has a higher level of risk than the other, say a tree logger (one of the most dangerous jobs) and a cashier, we would expect the tree logger to earn more because his wage will compensate him for the additional risk he takes on.

OCCUPATION	SECTOR	Total	Male	Female	MeanWage	se
Other	Other	68	62	6	8.5006	4.601049
Other	Manufacturing	68	44	24	8.0360	4.117607
Other	Construction	20	20	0	9.5020	3.343877
Management	Other	49	31	18	12.5990	7.950997

OCCUPATION	SECTOR	Total	Male	Female	MeanWage	se
Management	Manufacturing	6	3	3	13.5617	3.378878
Sales	Other	34	18	16	7.2874	4.294022
Sales	Manufacturing	4	3	1	10.1875	2.838537
Clerical	Other	88	21	67	7.2270	2.641008
Clerical	Manufacturing	7	0	7	9.6786	2.139217
Clerical	Construction	2	0	2	8.1300	5.006316
Service	Other	81	33	48	6.5351	3.716362
Service	Manufacturing	2	1	1	6.6350	1.223295
Professional	Other	91	42	49	11.4091	5.308922
Professional	Manufacturing	12	9	3	16.7708	5.037080
Professional	Construction	2	2	0	7.5000	3.535534

```
## # A tibble: 15 x 8
## # Groups:   OCCUPATION [?]
##    OCCUPATION   SECTOR        Total  Male Female MeanWage    se OCC_SEC
##    <fct>        <fct>         <int> <int>  <int>    <dbl> <dbl> <chr>
##  1 Other        Other            68    62      6     8.50  4.60 Othe Othe
##  2 Other        Manufacturing    68    44     24     8.04  4.12 Othe Manu
##  3 Other        Construction     20    20      0     9.50  3.34 Othe Cons
##  4 Management   Other            49    31     18    12.6   7.95 Mana Othe
##  5 Management   Manufacturing     6     3      3    13.6   3.38 Mana Manu
##  6 Sales        Other            34    18     16     7.29  4.29 Sale Othe
##  7 Sales        Manufacturing     4     3      1    10.2   2.84 Sale Manu
##  8 Clerical     Other            88    21     67     7.23  2.64 Cler Othe
##  9 Clerical     Manufacturing     7     0      7     9.68  2.14 Cler Manu
## 10 Clerical     Construction      2     0      2     8.13  5.01 Cler Cons
## 11 Service      Other            81    33     48     6.54  3.72 Serv Othe
## 12 Service      Manufacturing     2     1      1     6.64  1.22 Serv Manu
## 13 Professional Other            91    42     49    11.4   5.31 Prof Othe
## 14 Professional Manufacturing    12     9      3    16.8   5.04 Prof Manu
## 15 Professional Construction      2     2      0     7.50  3.54 Prof Cons
```



$$E(log(Wage)|\mathbf{X}) = \beta_0 + \beta_1 Education + \beta_2 Age + \beta_3 Sex + \beta_4 Management + \beta_5 Sales + \beta_6 Clerical + \beta_7 Service + \beta_8 Profes \tag{7}$$