# Interaction Effects and Polynomial Regression

2025-04-29

```r
library(ISLR)
library(ggplot2)
data(Wage)
head(Wage)
```

```
##        year age            maritl     race       education             region
## 231655 2006  18 1. Never Married 1. White    1. < HS Grad 2. Middle Atlantic
## 86582  2004  24 1. Never Married 1. White 4. College Grad 2. Middle Atlantic
## 161300 2003  45       2. Married 1. White 3. Some College 2. Middle Atlantic
## 155159 2003  43       2. Married 3. Asian 4. College Grad 2. Middle Atlantic
## 11443  2005  50      4. Divorced 1. White      2. HS Grad 2. Middle Atlantic
## 376662 2008  54       2. Married 1. White 4. College Grad 2. Middle Atlantic
##              jobclass         health health_ins  logwage      wage
## 231655  1. Industrial      1. <=Good      2. No 4.318063  75.04315
## 86582  2. Information 2. >=Very Good      2. No 4.255273  70.47602
## 161300  1. Industrial      1. <=Good     1. Yes 4.875061 130.98218
## 155159 2. Information 2. >=Very Good     1. Yes 5.041393 154.68529
## 11443  2. Information      1. <=Good     1. Yes 4.318063  75.04315
## 376662 2. Information 2. >=Very Good     1. Yes 4.845098 127.11574
```

```r
wage <- Wage
head(wage)
```

```
##        year age            maritl     race       education             region
## 231655 2006  18 1. Never Married 1. White    1. < HS Grad 2. Middle Atlantic
## 86582  2004  24 1. Never Married 1. White 4. College Grad 2. Middle Atlantic
## 161300 2003  45       2. Married 1. White 3. Some College 2. Middle Atlantic
## 155159 2003  43       2. Married 3. Asian 4. College Grad 2. Middle Atlantic
## 11443  2005  50      4. Divorced 1. White      2. HS Grad 2. Middle Atlantic
## 376662 2008  54       2. Married 1. White 4. College Grad 2. Middle Atlantic
##              jobclass         health health_ins  logwage      wage
## 231655  1. Industrial      1. <=Good      2. No 4.318063  75.04315
## 86582  2. Information 2. >=Very Good      2. No 4.255273  70.47602
## 161300  1. Industrial      1. <=Good     1. Yes 4.875061 130.98218
## 155159 2. Information 2. >=Very Good     1. Yes 5.041393 154.68529
## 11443  2. Information      1. <=Good     1. Yes 4.318063  75.04315
## 376662 2. Information 2. >=Very Good     1. Yes 4.845098 127.11574
```

```r
colnames(wage)
```

```
##  [1] "year"       "age"        "maritl"     "race"       "education"
##  [6] "region"     "jobclass"   "health"     "health_ins" "logwage"
## [11] "wage"
```

# Problem 0, Continued

1. Null Hypothesis: For the dummies of the "maritl" variable, beta_married = beta_widowed = b_divorced = b_separated = 0 Alternate Hypothesis : At least one of the dummies of the "maritl" variable != 0
2. Null Model without Marital Status wagehat_i = b0 + b1 * age_i + e_i

```
wage_null <- lm(wage ~ age, data = wage)
```

3. Formula for test statistic

Test statistic being used: F-statistic FS = (RSS_null - RSS_full)/(df_null - df_full) ALL DIVIDED BY (RSS_full/df_full)

4. ANOVA

```
wage_full <- lm(wage ~ age + maritl, data = wage)

print(summary(wage_null))
```

```
##
## Call:
## lm(formula = wage ~ age, data = wage)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -100.265  -25.115   -6.063   16.601  205.748
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 81.70474    2.84624   28.71   <2e-16 ***
## age          0.70728    0.06475   10.92   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 40.93 on 2998 degrees of freedom
## Multiple R-squared:  0.03827,    Adjusted R-squared:  0.03795
## F-statistic: 119.3 on 1 and 2998 DF,  p-value: < 2.2e-16
```

```
print(summary(wage_full))
```

```
##
## Call:
## lm(formula = wage ~ age + maritl, data = wage)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -100.97  -24.41   -5.56   15.65  219.25
##
## Coefficients:
```

```
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)         78.65466    2.79883  28.103  < 2e-16 ***
## age                  0.43212    0.07105   6.082 1.34e-09 ***
## maritl2. Married    20.81989    2.00197  10.400  < 2e-16 ***
## maritl3. Widowed    -1.06328    9.40852  -0.113    0.910
## maritl4. Divorced    3.93218    3.38700   1.161    0.246
## maritl5. Separated  3.62631    5.67963   0.638    0.523
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 40.04 on 2994 degrees of freedom
## Multiple R-squared:  0.0809, Adjusted R-squared:  0.07936
## F-statistic:  52.7 on 5 and 2994 DF,  p-value: < 2.2e-16
```

```r
anova(wage_null, wage_full)
```

```
## Analysis of Variance Table
##
## Model 1: wage ~ age
## Model 2: wage ~ age + maritl
##   Res.Df     RSS Df Sum of Sq     F    Pr(>F)
## 1   2998 5022216
## 2   2994 4799644  4    222572 34.71 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

4. ANOVA Continued

RSS_null = 5022216 RSS_full = 4799644 df_null = 2998 df_full = 2994 numerator_df = 4 f-stat = 34.71 p-value = < 2.2e-16

5. Conclusion of Test

From our ANOVA test between the null and full models, we obtained an F-statistic of 34.71, and a corresponding p-value of < 2.2e-16. This p-value is way below 0.05, giving us enough evidence to reject the null hypothesis and say that at least one of the dummy marital variables is statistically significant. This means that adding marital status to our model improves it's predictions significantly.

#PROBLEM ONE

## Provide data visualization that will help determine if there's a strong interaction between respective variables in explaining the response.

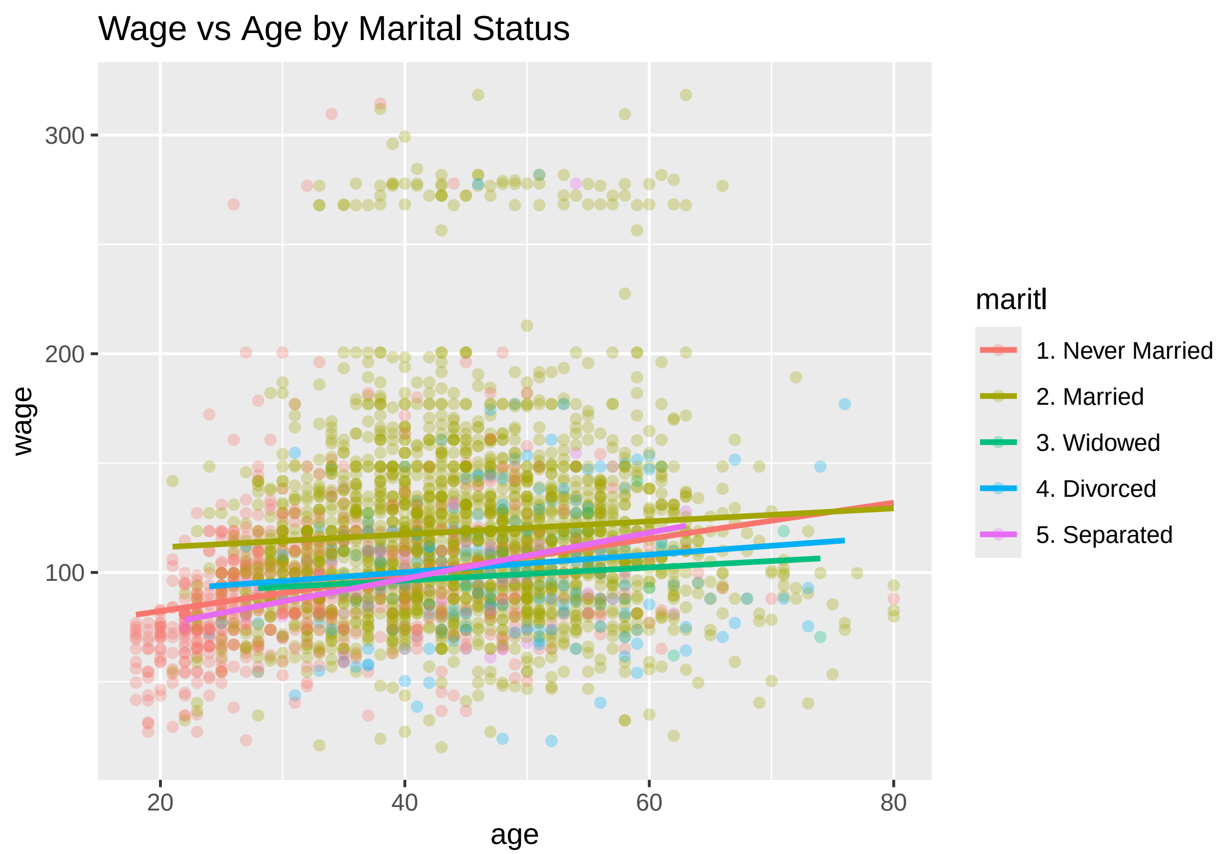
```r
ggplot(Wage, aes(x = age, y = wage, color = jobclass)) +
  geom_point(alpha = 0.3) +
  geom_smooth(method = "lm", se = FALSE) +
  labs(title = "Wage vs Age by Job Class")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```


When looking at the graph of the slopes of job class v. age v. wage, the lines appear slightly different than each other. However, as we learned in class, although they appear to be different slopes, given our confidence level, etc, they may still be parallel within our confidence bands.Although the line for information job class is slightly above industrial, the lines are roughly parallel, suggesting jobclass doesnt not interact with age much to affect wage.

```
ggplot(wage, aes(x = age, y = wage, color = maritl)) +
  geom_point(alpha = 0.3) +
  geom_smooth(method = "lm", se = FALSE) +
  labs(title = "Wage vs Age by Marital Status")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```




When looking at the graph of the slopes of marital status v age v wage, the lines are definitely varied across marital status. However, most of the differences are not drastic enough to raise attention. The most noticable difference is that the "never married" and maybe "separated" slopes appears a bit steeper than the others, and intersects across other lines. Although this line may still fall within the acceptable range of the null hypothesis, it is hard to tell and it may be worth examining further.

### Confirm your hunch from part (a) by actually writing out the appropriate model, fitting it, and conducting the respective statistical test for significance of the overall interaction.

I think the hunch refers to the never married dummy variable's slope appearing steeper than the others.

Null Hypothesis: b3 = 0 (beta3 being the slope of the interaction term between marital and age, aka slope of interaction term = 0) Alternate Hypothesis:b3 != 0

```r
interaction_full  <- lm(wage ~ age * maritl, data = wage)

interaction_null <- lm(wage ~ age + maritl, data = wage)
```

```r
print(interaction_full)
```

```
##
## Call:
## lm(formula = wage ~ age * maritl, data = wage)
##
```

```
## Coefficients:
##           (Intercept)                    age        maritl2. Married
##               65.8098                 0.8263                 39.7248
##       maritl3. Widowed      maritl4. Divorced      maritl5. Separated
##               18.6807                18.1354                -10.6207
##   age:maritl2. Married   age:maritl3. Widowed   age:maritl4. Divorced
##               -0.5293                -0.5301                 -0.4227
## age:maritl5. Separated
##                0.2241
```

```r
summary(interaction_full)
```

```
##
## Call:
## lm(formula = wage ~ age * maritl, data = wage)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -98.660 -24.678  -5.099  15.705 217.119
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)             65.8098     5.1857  12.691  < 2e-16 ***
## age                      0.8263     0.1517   5.448 5.50e-08 ***
## maritl2. Married        39.7248     6.5024   6.109 1.13e-09 ***
## maritl3. Widowed        18.6807    38.6555   0.483  0.62895
## maritl4. Divorced       18.1354    14.8593   1.220  0.22238
## maritl5. Separated     -10.6207    25.1344  -0.423  0.67265
## age:maritl2. Married    -0.5293     0.1740  -3.042  0.00237 **
## age:maritl3. Widowed    -0.5301     0.7478  -0.709  0.47849
## age:maritl4. Divorced   -0.4227     0.3242  -1.304  0.19233
## age:maritl5. Separated   0.2241     0.5682   0.394  0.69337
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 39.99 on 2990 degrees of freedom
## Multiple R-squared:  0.08414,    Adjusted R-squared:  0.08138
## F-statistic: 30.52 on 9 and 2990 DF,  p-value: < 2.2e-16
```

```r
anova(interaction_full, interaction_null)
```

```
## Analysis of Variance Table
##
## Model 1: wage ~ age * maritl
## Model 2: wage ~ age + maritl
##   Res.Df     RSS Df Sum of Sq      F  Pr(>F)
## 1   2990 4782710
## 2   2994 4799644 -4    -16934 2.6467 0.03182 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We use an incremental F-test to determine whether adding a set/subset of predictors improves the model's performance. In this case, we are testing the interaction term between age and marital status. The ANOVA function llooks at the null model without the interaction term and compares it to the full model with the interaction term. Because the corresponding p-value, 0.0318, is less than 0.05, we reject the null hypothesis and conclude that at least one of the dummy variables has an interaction with age in explaining wage.

#c If you found an interaction…

```r
print((summary(interaction_full)))
```

```
##
## Call:
## lm(formula = wage ~ age * maritl, data = wage)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -98.660 -24.678  -5.099  15.705 217.119
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)              65.8098     5.1857  12.691  < 2e-16 ***
## age                       0.8263     0.1517   5.448 5.50e-08 ***
## maritl2. Married         39.7248     6.5024   6.109 1.13e-09 ***
## maritl3. Widowed         18.6807    38.6555   0.483  0.62895
## maritl4. Divorced        18.1354    14.8593   1.220  0.22238
## maritl5. Separated      -10.6207    25.1344  -0.423  0.67265
## age:maritl2. Married     -0.5293     0.1740  -3.042  0.00237 **
## age:maritl3. Widowed     -0.5301     0.7478  -0.709  0.47849
## age:maritl4. Divorced    -0.4227     0.3242  -1.304  0.19233
## age:maritl5. Separated    0.2241     0.5682   0.394  0.69337
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 39.99 on 2990 degrees of freedom
## Multiple R-squared:  0.08414,    Adjusted R-squared:  0.08138
## F-statistic: 30.52 on 9 and 2990 DF,  p-value: < 2.2e-16
```

When looking at the summary for the full model and specifically the p-values and t-values of all the interaction terms, one comes up as statistically significant. The interaction term for marital status "Married" has a p-value $< 0.05$, which is why we can reject the null hypothesis for this term. It appears that age and dummy variable Married work together to impact wage.

# c continued

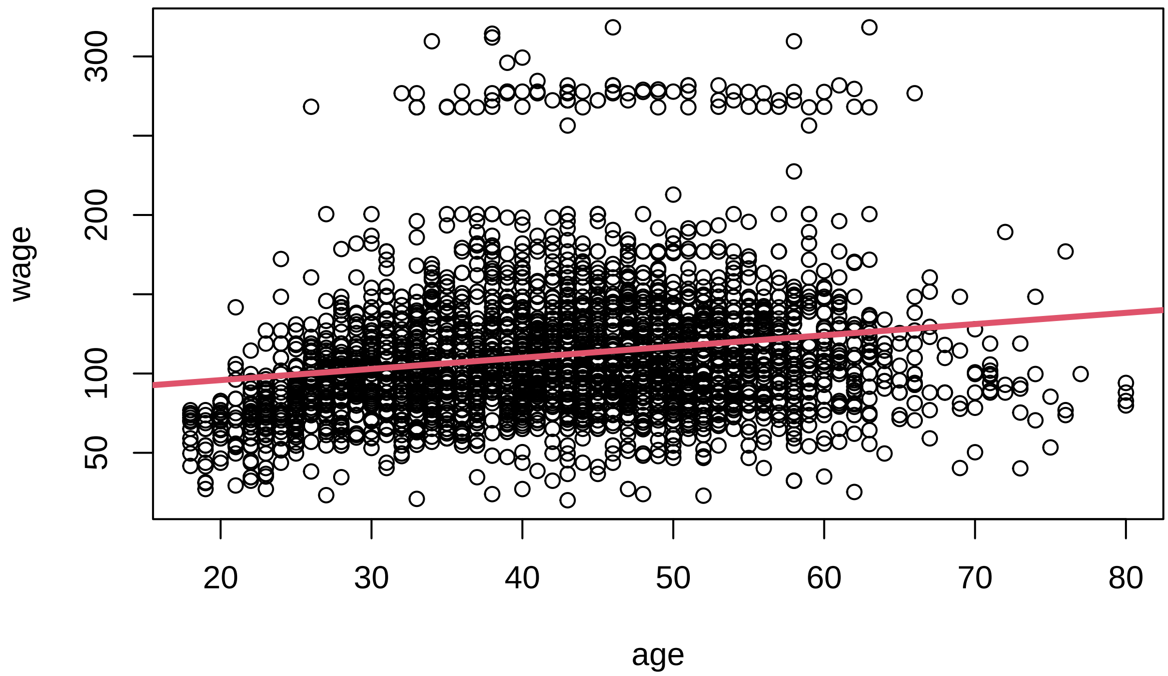
The most significant interaction term was age:maritl2.Married. The two main variables involved in this term are age and maritl2.Married. Main effect of age: 0.826 Among all individuals who have mever married, each 1 year increase in age results in a ~0.83 unit increase in wage, on average. Main effect for maritl2.Married: 39.72 Among all individuals at baseline age zero, married individuals wages are expected to be 39.72 units higher than the baseline never married.

#Problem Two

Simple Regression While the simple linear regression of wage ~ age is not horrible, there is a trend amongst the observations and it is not capturing it well. The plot has noticable curvature, and the fitted line is just running through the middle of the cluster. There appears to be a noticable pattern of clustering at the top as well, and the bottom portion appears curved downwards.

```r
simple_wage <- lm(wage ~ age, data=Wage)

plot(wage ~ age, data=Wage)
abline(simple_wage, lwd=3, col=2)
```




```r
summary(simple_wage)
```

```
##
## Call:
## lm(formula = wage ~ age, data = Wage)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -100.265  -25.115   -6.063   16.601  205.748
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 81.70474    2.84624   28.71   <2e-16 ***
## age          0.70728    0.06475   10.92   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 40.93 on 2998 degrees of freedom
## Multiple R-squared:  0.03827,    Adjusted R-squared:  0.03795
## F-statistic: 119.3 on 1 and 2998 DF,  p-value: < 2.2e-16
```

Test Quadratic Regression

```r
poly_wage <- lm(wage ~ age + I(age^2), data=wage)
summary(poly_wage)
```

```
##
## Call:
## lm(formula = wage ~ age + I(age^2), data = wage)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -99.126 -24.309  -5.017  15.494 205.621
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -10.425224   8.189780  -1.273    0.203
## age           5.294030   0.388689  13.620   <2e-16 ***
## I(age^2)     -0.053005   0.004432 -11.960   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 39.99 on 2997 degrees of freedom
## Multiple R-squared:  0.08209,    Adjusted R-squared:  0.08147
## F-statistic:   134 on 2 and 2997 DF,  p-value: < 2.2e-16
```

Null Hypothesis: B2 = 0, B2 being the coefficient for the polynomial term Alternate Hypothesis: B2!= 0

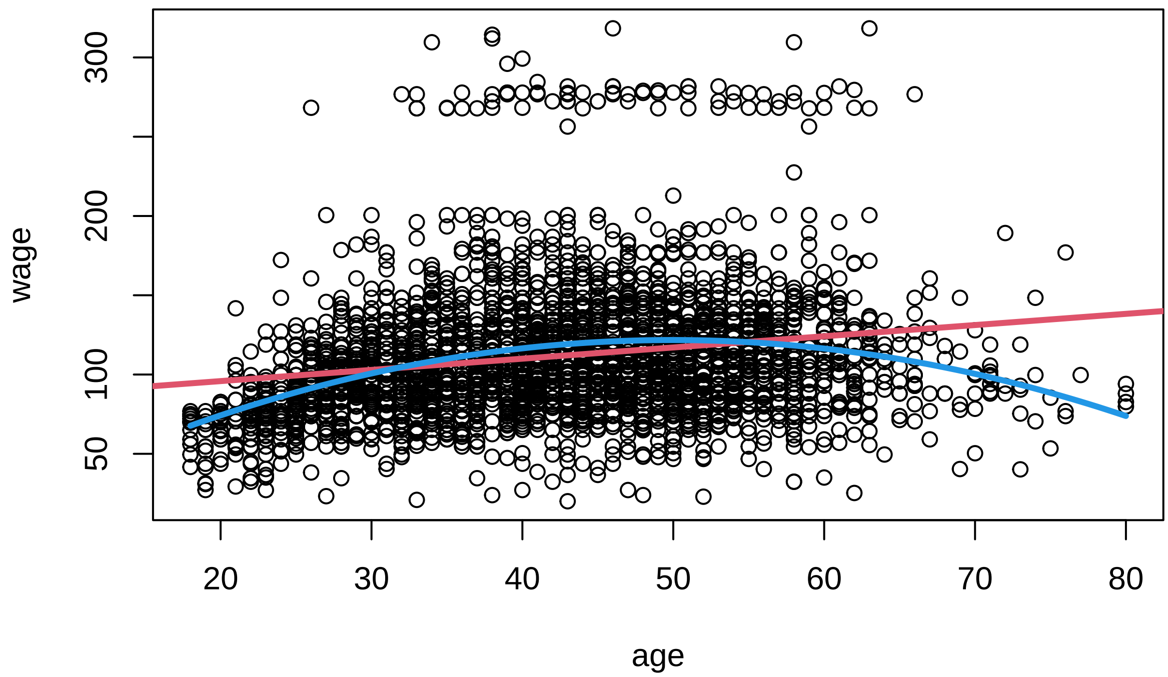
wage_i = b0 +( b1 * age_i) + (b2 * age^2_i) + error_i, where e is N ~(0,sigma^2)

When we add the polynomial variable, age^2, the corresponding p-value returns far below 0.05, with a test statistic of -11.96. There is enough evidence to reject the null hypothesis. Additionally, the RSS of the model went down a little bit when we included the polynomial term, and the R-squared value more than doubled.Overall, there is strong evidence of a quadratic relationship between age and wage.

Quadratic Line

```r
#Need to run all together to show in knit


plot(wage ~ age, data=Wage)
abline(simple_wage, lwd=3, col=2)
x.grid <- seq(from = min(Wage$age), to = max(Wage$age), length = 100)
lines(x = x.grid,
      y = predict(poly_wage, newdata = data.frame(age = x.grid)),
      col = 4, lwd = 3)
```




While the new curved line is not perfect, it is significantly more representative of what is actually going on. Before, our fitted line was just trying to cut through the center of the majority of the points. Now, the new curve moderately captures the overall curve of the observations. The lower majority (lower as in physically not numerically) of observations are curved downwards, and our curve reflects that. There is still a little bit of uncertainty regarding the cluster of "outliers" on the top.

#Summary

In this homework/set of lectures, we wrapped up linear regression with ways to combat their limitations. We started with using interaction terms to tackle the restriction of additivity for linear models. By incorporating an interaction term, we learned in lecture that we can measure how the two terms work together to explain the response variable. In this homework, we solidified this by adding an interaction term to the wages dataset, and performing an incremental f-test to determine if the term adds anything to our model, which it did. If it did not, we would just drop the variable and focus on the main effects, as we discussed. After we finished addressing additivity, we addressed the restriction of linearity. Before, one of our key assumptions was that the true relationship between the response and predictors in the real world was linear, which is hardly ever the case for the questions we ask. By introducing synthetic polynomial terms into our regression, we are able to try and model non-linear relationships and make predictions. As we've been doing most of the semester, we performed the hypothesis testing process we learned in class to evaluate if a quadratic term is appropriate.