STA 9797 – Project 1

Erik Carrion

4 November 2021

Introduction

The data under consideration is the 1985 Consumer Supplementary Survey which includes data on wages, occupation, marital status, geographic location, and others. The purpose of our analysis is to isolate those variables which have the most pronounced effect on wage and then, dependent upon our analysis, construct a 2-way model which sufficiently accounts for the variation in wage data.

In order to facilitate our analysis, we have converted the AGE, EDUC and EXPERIENCE variables into categorical variables:

1. For Education we created three categories to reflect whether the individual had less than a high school education, up to high school education, and more than a high school education. We coded these as (1, 2, 3).
2. For Experience, the subjects exhibited a range from 1 to 55 years of experience. We coded experience into the following categories labeled as (1, 2, 3, 4, or 5):
    a. 0 to 10 years
    b. 11 to 20 years
    c. 21 to 30 years
    d. 31 to 40 years
    e. More than 40 years
3. For Age we constructed 5 ranges coded (1, 2, 3, 4, 5):
    a. Less than 20 years old
    b. Between 20 and 30 years old
    c. Between 30 and 40 years old
    d. Between 40 and 50 years old
    e. Greater than 50 years old

When considering wages there are a number of factors which play a role in how much one earns. Of those factors we postulate that age, experience, & occupation will have the largest effect on the wage outcome reported.

In this analysis we ask and answer the following three questions:

1. Are our factors of interest individually significant?
2. Are there differences in the levels of the factors which are significant?
3. Does your Occupation and level of Education lead to higher wages?

Univariate Analysis

Prior to our analysis we hypothesized that Age, Education, Experience, and Occupation play a significant role in explaining the variability in reported Wages. The overall results of our 1-Way models are below.

Education

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	1725.36929	862.68465	37.09	<.0001
Error	531	12351.32939	23.26051		
Corrected Total	533	14076.69868			

Experience

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	582.80995	145.70249	5.71	0.0002
Error	529	13493.88873	25.50830		
Corrected Total	533	14076.69868			

Age

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	834.67184	208.66796	8.34	<.0001
Error	529	13242.02684	25.03219		
Corrected Total	533	14076.69868			

Occupation

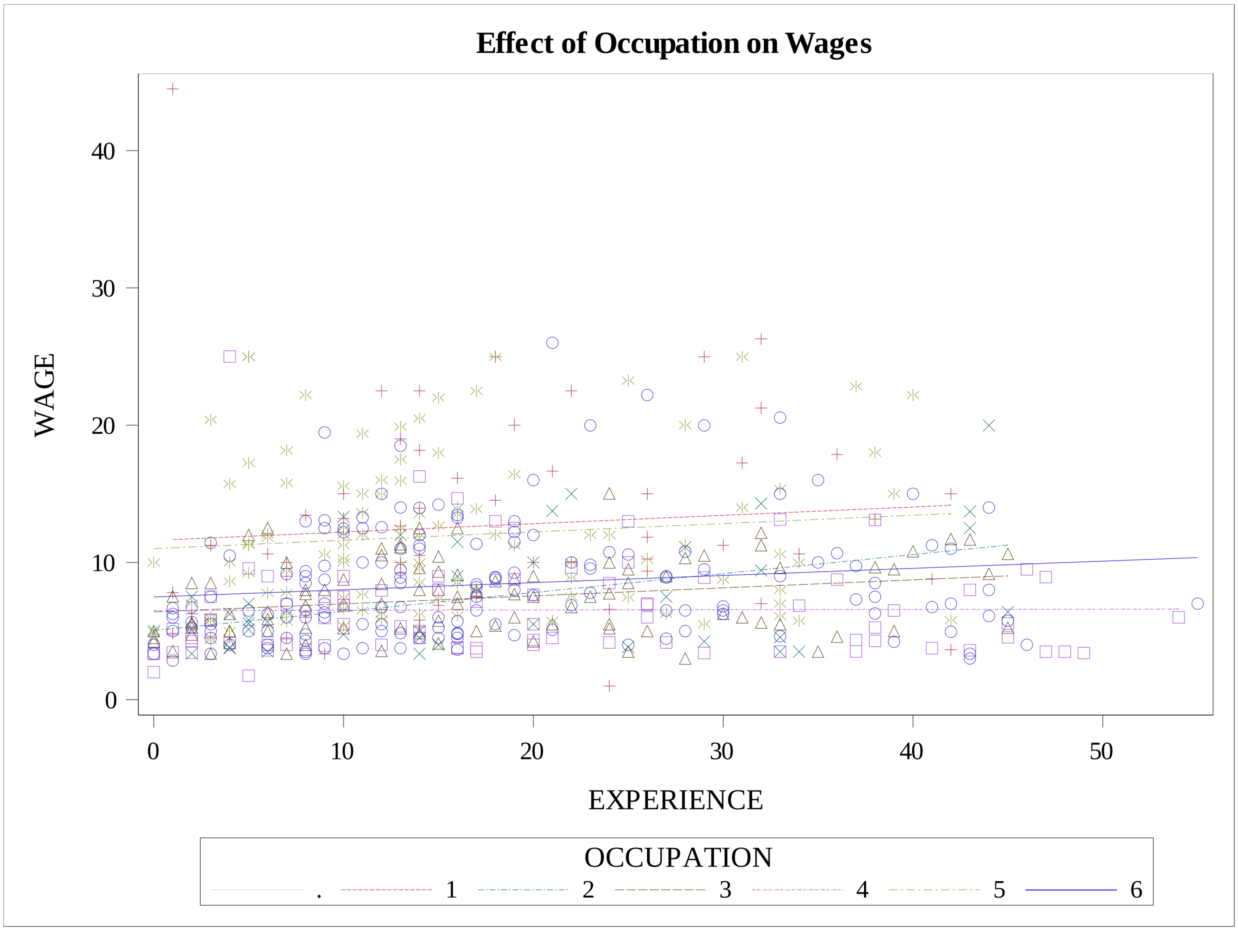
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	2537.69723	507.53945	23.22	<.0001
Error	528	11539.00145	21.85417		
Corrected Total	533	14076.69868			

While all four models are significant, an inspection of the F Statistic reveals that Occupation and Education were significantly larger than Age and Experience which is reflected in their respective R^2's:

Variable	R^2
Age	0.059295
Education	0.122569
Occupation	0.180276
Experience	0.0414

Age did not play a large part, explaining only 5.93% of the variability in the data. Interestingly, Experience performed worse than Age alone. Given one's occupation, we expect wages to grow significantly as you become more experienced, knowledgeable, and skilled in your field. The data, however, do not bear that out.

In the graph below we see that given your occupation, wages don't grow substantially as you gain more experience, regardless of your occupation. The primary driver in income differences is Occupation with those working in Management or Professional fields earning significantly more than those working in Sales, Clerical, or Other occupations.



Checking Model Assumptions

While the models are all significant, we need to check our assumptions regarding normality and homogeneity of variance for each of the 4 factors. First, we check the results of Levene's test for the homogeneity of variance and conclude that we can reject the null hypothesis of equal variances for Education and Experience.

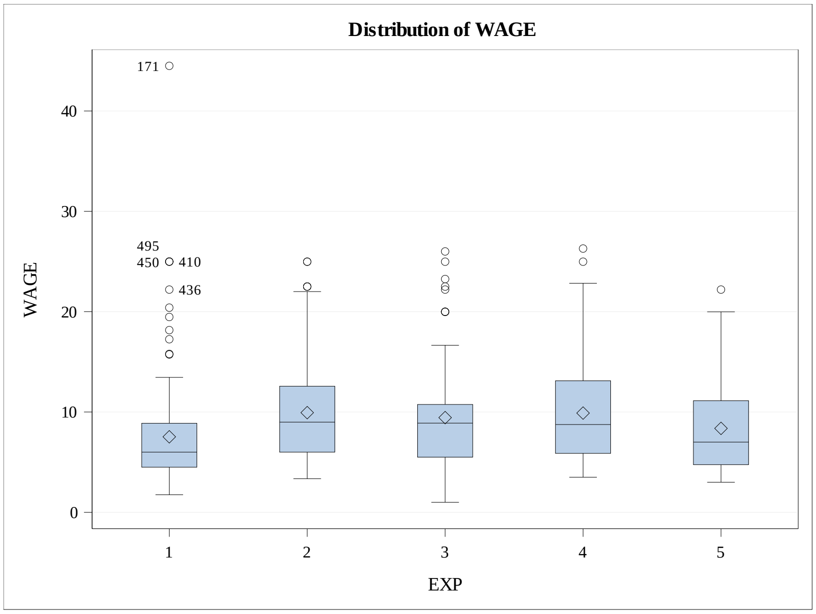
Levene's Test		
	F Statistic	**P > F**
Education	10.68	<.001
Occupation	0.23	0.9187
Experience	6.97	<.001
Age	0.74	0.563

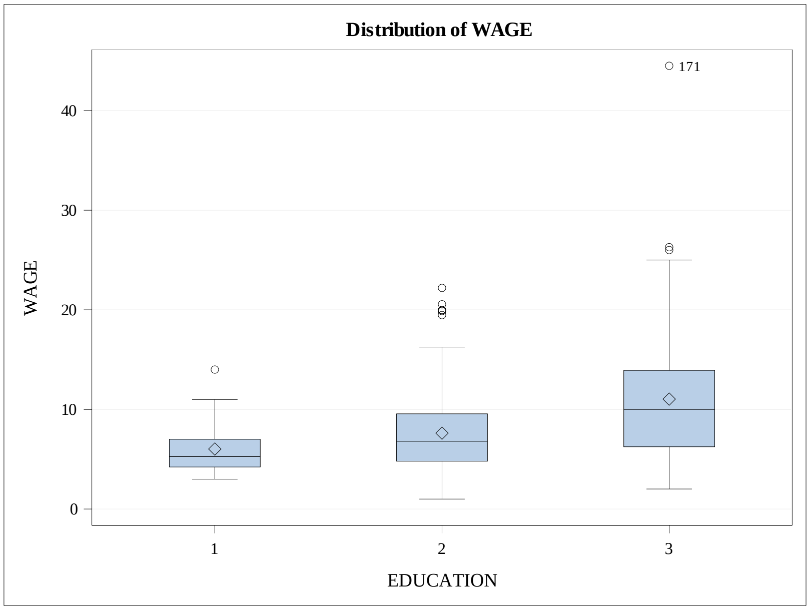
A rule of thumb is that ANOVA will be robust so long as the largest variance is not 4 times greater than smallest variance. If we look at this ratio for our 4 factors, we have the following:

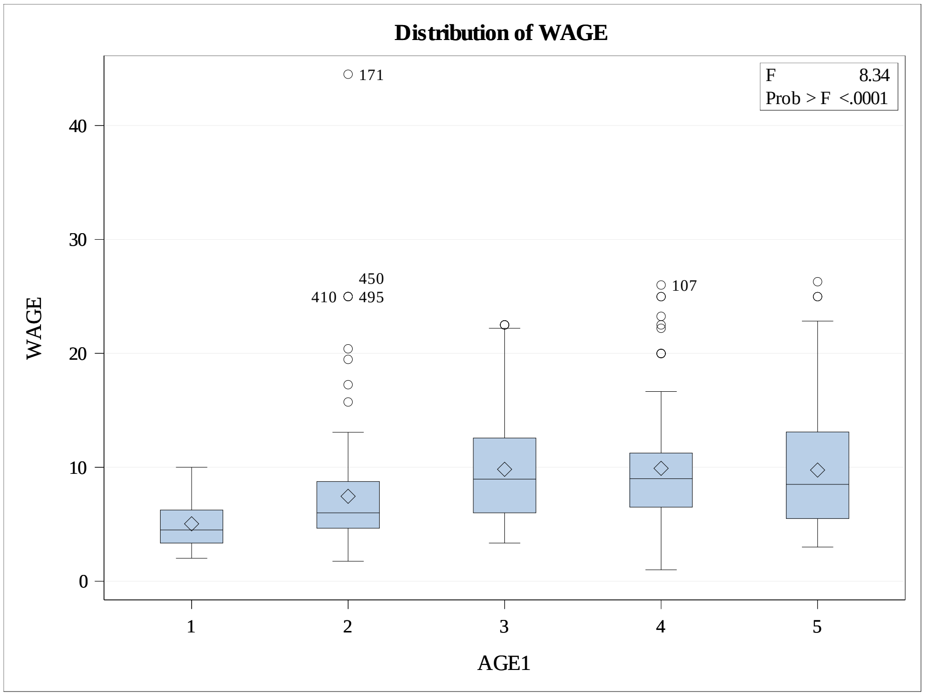
Variance Ratio (Biggest/Smallest)	
Education	5.35
Occupation	7.87
Experience	1.43
Age	7.01

For the variables which Levene's concluded had different variances, Education has a ratio greater than 4 while Experience does not which is evident in the boxplots. Interestingly, the other two factors which Levene's test concluded had equal variances, exhibit a wide gap in the range of the variances.

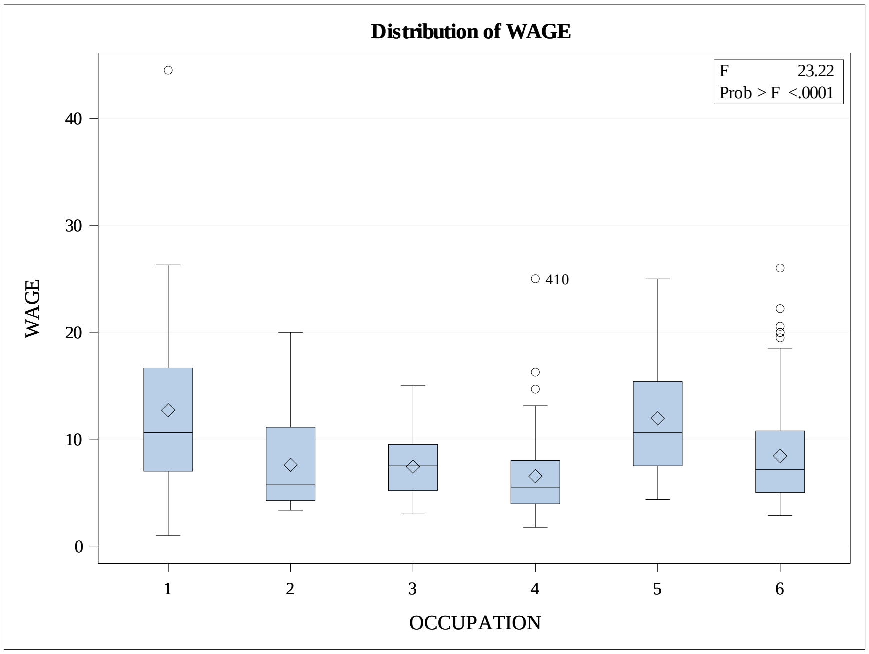
Looking at the boxplot for Experience, we can see the variation about their means is in roughly the same range for each level. However, that is not the case for Education. As Education increases, it's accompanied by a widening of the interquartile range.


Distribution of WAGE
WAGE
0
10
20
30
40
171
495
450
410
436
1
2
3
4
5
EXP


Distribution of WAGE
WAGE
0
10
20
30
40
171
1
2
3
EDUCATION

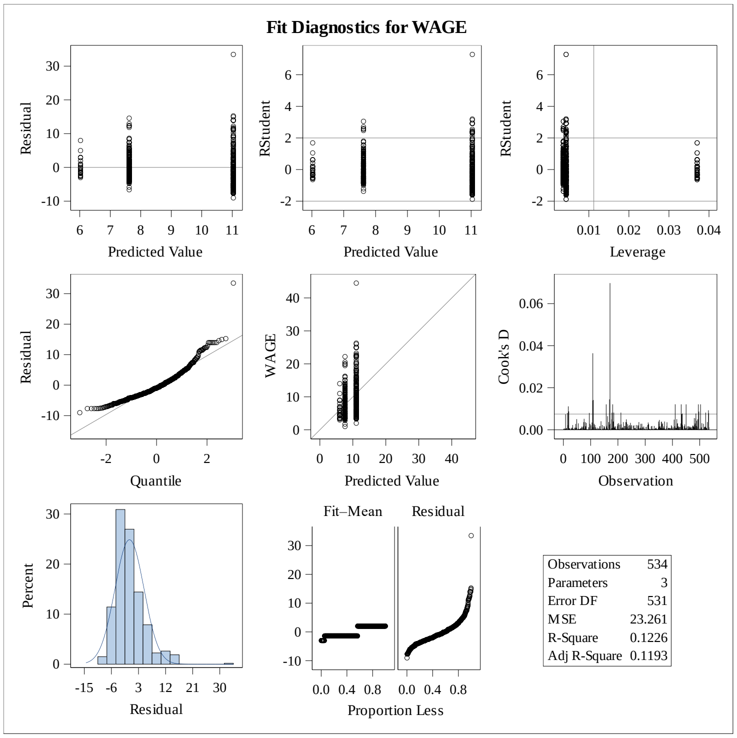


As expected, wages grow in line with a subject's Age. Dispersion seems relatively symmetric, but the data is affected by a number of outliers

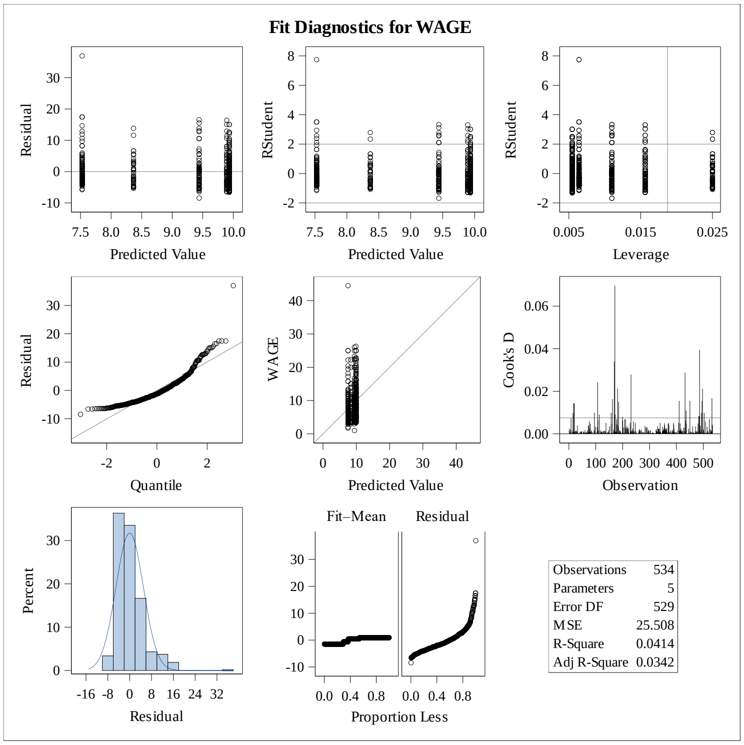


When it comes to Occupation, we see that salaries are comparatively highly variable for those who listed their Occupation as Management or Professional. Given that Occupation and Education display violations of our equal variance assumption, we could transform the data to ensure homogeneity of variance. For our analysis, we will assume that the ANOVA is robust to this violation.

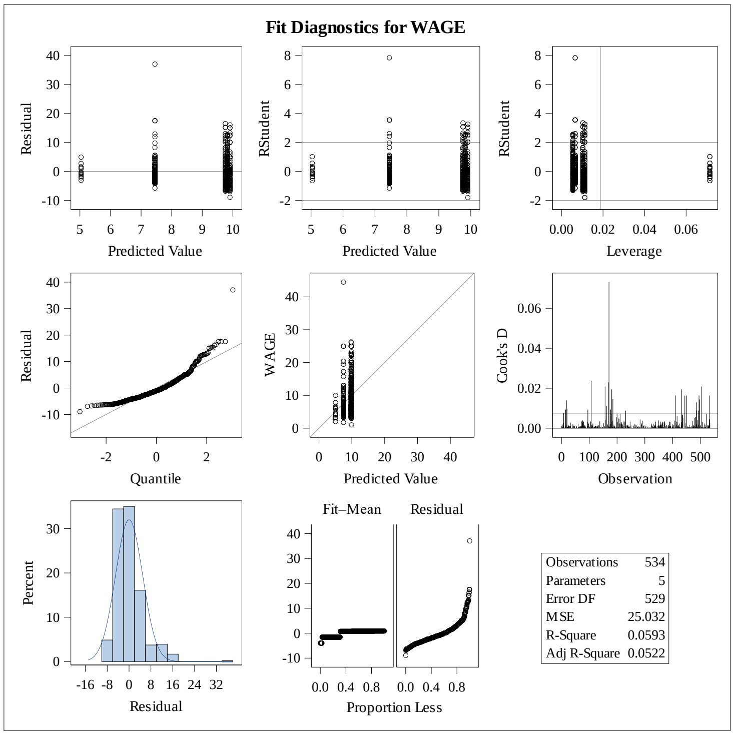
Checking on the normality of our variables we look to the diagnostic plots and see that each of our 4 factors do not violate our assumption of normality.



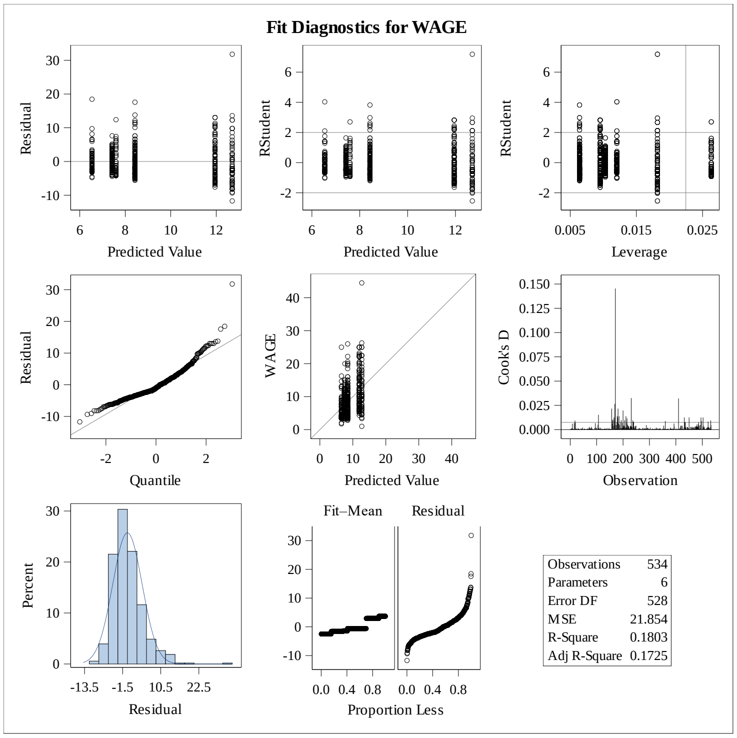
Education



Experience3



Age



Observations	534
Parameters	6
Error DF	528
MSE	21.854
R-Square	0.1803
Adj R-Square	0.1725

Occupation

Given that Age and Experience were the lease significant variables of interest we will not include them for further investigation, and we will proceed with an analysis of Occupation and Education. Now we will consider the difference among the group means for Occupation and Education.

Difference in Mean Levels of Occupation and Education

To assess differences while maintaining an overall significance level of .05 we employ Tukey's Test. For Education we have the following results:

Comparisons significant at the 0.05 level are indicated by ***.				
EDUCATION Comparison	Difference Between Means	Simultaneous 95% Confidence Limits		
3 - 2	3.4172	2.4067	4.4277	***
3 - 1	5.0133	2.7083	7.3182	***
2 - 3	-3.4172	-4.4277	-2.4067	***
2 - 1	1.5961	-0.6900	3.8821	
1 - 3	-5.0133	-7.3182	-2.7083	***
1 - 2	-1.5961	-3.8821	0.6900	

The results show that groups 3 (College or Above) & 2 (High School Education) and 3 & 1 (Less Than Highschool Education) are significantly different with group 3 outpacing their less educated counterparts by $3.41/hour and $5.01/hour respectively.

Comparisons significant at the 0.05 level are indicated by ***.				
OCCUPATION Comparison	Difference Between Means	Simultaneous 95% Confidence Limits		
1 - 5	0.7566	-1.4690	2.9822	
1 - 6	4.2775	2.1807	6.3744	***
1 - 2	5.1114	2.2908	7.9319	***
1 - 3	5.2814	3.0245	7.5384	***
1 - 4	6.1665	3.8417	8.4913	***
5 - 1	-0.7566	-2.9822	1.4690	
5 - 6	3.5210	1.8331	5.2088	***
5 - 2	4.3548	1.8235	6.8861	***
5 - 3	4.5249	2.6418	6.4079	***
5 - 4	5.4100	3.4461	7.3738	***
6 - 1	-4.2775	-6.3744	-2.1807	***
6 - 5	-3.5210	-5.2088	-1.8331	***
6 - 2	0.8338	-1.5850	3.2527	
6 - 3	1.0039	-0.7250	2.7328	

Comparisons significant at the 0.05 level are indicated by ***.				
OCCUPATION Comparison	Difference Between Means	Simultaneous 95% Confidence Limits		
6 - 4	1.8890	0.0724	3.7056	***
2 - 1	-5.1114	-7.9319	-2.2908	***
2 - 5	-4.3548	-6.8861	-1.8235	***
2 - 6	-0.8338	-3.2527	1.5850	
2 - 3	0.1701	-2.3889	2.7290	
2 - 4	1.0552	-1.5638	3.6741	
3 - 1	-5.2814	-7.5384	-3.0245	***
3 - 5	-4.5249	-6.4079	-2.6418	***
3 - 6	-1.0039	-2.7328	0.7250	
3 - 2	-0.1701	-2.7290	2.3889	
3 - 4	0.8851	-1.1142	2.8844	
4 - 1	-6.1665	-8.4913	-3.8417	***
4 - 5	-5.4100	-7.3738	-3.4461	***
4 - 6	-1.8890	-3.7056	-0.0724	***
4 - 2	-1.0552	-3.6741	1.5638	
4 - 3	-0.8851	-2.8844	1.1142	

When it comes to occupation, those who are Professionals or in Management outpace every other group. Among the other groups while there are some statistically significant differences, note is the magnitude of these differences are relatively small compared to the differences between Managers and any other group, for example.

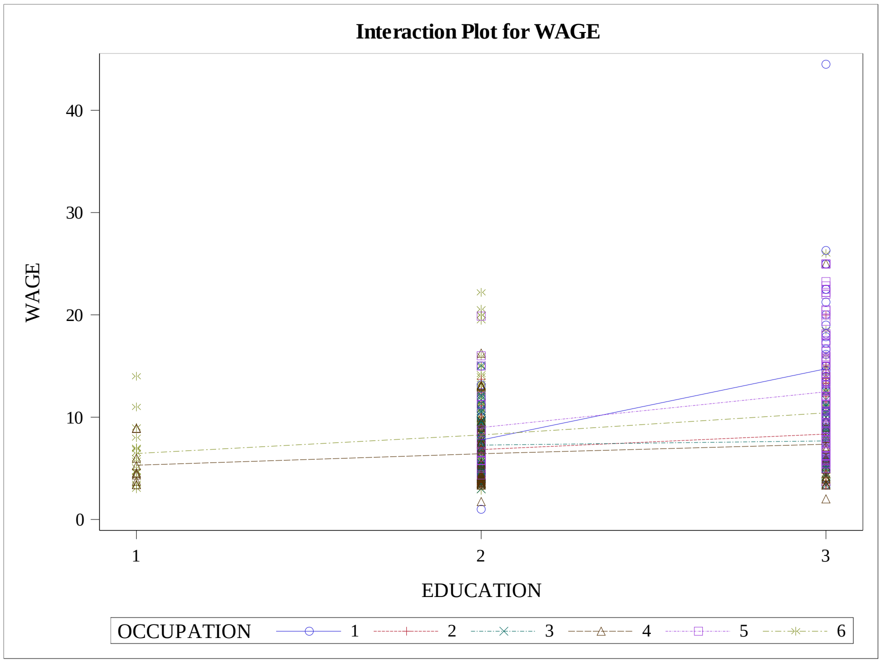
While our assumption regarding constant variance was violated, thus raising the possibility of a transformation on the data, it is outside the scope of this analysis. We will proceed to test a 2-factor model including Occupation and Education as our primary factors of interest.

2 Factor Analysis

To determine if we will employ a saturated or reduced model, we first test the significance of the interaction between Occupation and Education. Our analysis yields the following result:

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	13	3482.60264	267.89251	13.15	<.0001
Error	520	10594.09604	20.37326		
Corrected Total	533	14076.69868			

R-Square	Coeff Var	Root MSE	WAGE Mean
0.247402	50.01821	4.513675	9.024064



Knowing that the interaction is significant we now develop a full model including the interaction term. We will be treating this as a mixed model with Education treated as our fixed factor and Occupation as our random factor. The levels of Education are exhaustive and thus represent the population of all possible levels of Education. We treat Occupation as a random factor as they are not exhaustive of all possible levels of the population. Running the analysis yields the following:

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	13	3482.60264	267.89251	13.15	<.0001
Error	520	10594.09604	20.37326		
Corrected Total	533	14076.69868			

R-Square	Coeff Var	Root MSE	WAGE Mean
0.247402	50.01821	4.513675	9.024064

Source	DF	Type III SS	Mean Square	F Value	Pr > F
EDUCATION	2	630.2431009	315.1215505	15.47	<.0001
OCCUPATION	5	872.4630050	174.4926010	8.56	<.0001
EDUCATION*OCCUPATION	6	388.8774721	64.8129120	3.18	0.0045

Not only is the overall model significant but the Type III sum of squares shows that both main effects and the interaction are significant. Further, the overall model has an R^2 of .2474 which is exactly the same as the model which includes only the interaction term. While the main effects are both significant if we look to the results of the random effects analysis, we see that education is barely significant while Occupation is not at all significant. However, because the interaction is significant, Occupation remains in the final model.

Source	DF	Type III SS	Mean Square	F Value	Pr > F
EDUCATION	2	630.243101	315.121550	5.70	0.0335

Error	7.0786	391.075120	55.247835		
Error: 0.7848*MS(EDUCATION*OCCUPATION) + 0.2152*MS(Error)					

Source	DF	Type III SS	Mean Square	F Value	Pr > F
OCCUPATION	5	872.463005	174.492601	2.74	0.1240

Error	6.1027	388.374968	63.639869		
Error: 0.9736*MS(EDUCATION*OCCUPATION) + 0.0264*MS(Error)					

Source	DF	Type III SS	Mean Square	F Value	Pr > F
EDUCATION*OCCUPATION	6	388.877472	64.812912	3.18	0.0045

Error: MS(Error)	520	10594	20.373262		

While the R^2 is reasonable, there are a number of factors that determine average wages. For example, state of residence, the unemployment or inflation rate. Because there are other factors, not considered, a large portion explainable variability is retained within the residuals.

Conclusions & Discussion

Based on the results of our analysis we can conclude that hourly wages are determined in part by the interaction between Education and Occupation. If you want to earn a higher hourly wage you should be a Professional or work in Management but in order to do that you need to get a better than high school education.

In regard to our initial questions:

1. Are our factors of interest individually significant? We tested them individually to determine their inclusion in our 2-Way model, hypothesizing that of the observed variables, these would be significantly related to hourly wage. The results of our ANOVA indicate they are all positively related to hourly wage.

   We proceeded with a 1-Way ANOVA because it allows us to test whether the levels of the factor under consideration are in fact significant because it tests the following two models:

   Reduced Model: $Y_{ij} = \mu + \epsilon_{ij}$ versus the

   Full model of $Y_{ij} = \mu_j + \epsilon_{ij}$ which implies a treatment level effect.

   Because of this, the ANOVA procedure was the appropriate model to employ for this analysis because it will tell us whether the factors of interest have a "true" effect or if they don't.

2. Are there differences in the levels of the factors which are significant? Knowing that the factors are significant is the first step. We wanted to investigate the differences between the levels to identify those levels which had the most pronounced effect on hourly wage.

   The results of our analysis showed that those with a college degree or better earn more than their less educated counterparts. Further, if you are a professional or in management you can expect to earn more than if you were in Sales or another type of profession.

   Because we are making pairwise comparisons, we employed Tukey's Studentized Range (HSD) Test which allows us to maintain a significance level of .05.

3. Does your Occupation and level of Education lead to higher wages? We employed a mixed effects model to estimate the relationship between Hourly Wage and Occupation, Education, and the Interaction. Our results from the analysis, in combination with the prior analysis, to conclude that if you want to earn a

higher hourly wage you should work in Management or as a Professional but in order to do that you have to get more than a high school education.

The random effects model considers the nature of the random factor and allows for a more powerful test.

Code

```
proc import file = "CSS85.xlsx" dbms = xlsx out = css replace replace;
run;

/* One Way ANOVA for Factors of Interest */

ods rtf file = "One Way ANOVAs .rtf";

/* Does Education Affect Wage? */
proc glm data = css PLOTS = DIAGNOSTICS;
        class EDUCATION;
        model WAGE = EDUCATION;
        means EDUCATION /HOVTEST = LEVENE;
        title "Effect of Eduction on Wages";
run;

/* Does Experience affect Wage? */
proc glm data = css plots=diagnostics;
        CLASS EXP;
        model WAGE = EXP;
        means EXP / TUKEY hovetest = levene;
        title "Effect of Experience on Wages";
run;

/* Does Age affect Wage? */

proc glm data = css plots = diagnostics;
        class AGE1;
        model WAGE = AGE1;
        means AGE1 / tukey hovtest = levene;
        title "Effect of Age on Wages";
        run;


/* Does Occupation affect Wage? - yes and it will be includedin the model */
proc glm data = css plots = diagnostics;
        CLASS OCCUPATION;
        model WAGE = OCCUPATION;
        means OCCUPATION / tukey hovtest=levene;
        title "Effect of Occupation on Wages";
run;
ods rtf close;

/* Means */
ods rtf file = "Univariate Means.rtf";
proc means data = css mean var median q1 q3;
        class EXP;
        var WAGE;
        title "Means Experience";
RUN;
proc means data = css mean var median q1 q3;
        class AGE1;
        var WAGE;
        title "Means Age";
RUN;
proc means data = css mean var median q1 q3;
        class EDUCATION;
        var WAGE;
        title "Means Education";
RUN;
proc means data = css mean var median q1 q3;
        class OCCUPATION;
        var WAGE;
        title "Means Occupation";
RUN;
ods rtf close;

ods rtf file = "Plot Experience x Wage by Occupation.rtf";
proc sgplot data = css;
        reg x = EXPERIENCE y = WAGE / group = OCCUPATION;
        run;
ods rtf close;

/* 2 way ANOVA for Education & Occupation*/
/* TEST INTERACTION FIRST */

ODS RTF FILE = "INTERACTION ANALYSIS.RTF";
proc glm data = css;
        class EDUCATION OCCUPATION;
        model WAGE = EDUCATION*OCCUPATION / SOLUTION;
        TITLE "Test of Interaction of Occupation and Education on Wage";
RUN;
ODS RTF CLOSE;

/* the interaction is significant so it stays in the model */
```

```
ODS RTF FILE = "2 Factor Analysis Education and Occupation";
proc glm data = css MANOVA;
        class EDUCATION OCCUPATION;
        MODEL WAGE = EDUCATION OCCUPATION EDUCATION*OCCUPATION;
        RANDOM OCCUPATION EDUCATION*OCCUPATION/ test;
        MEANS EDUCATION OCCUPATION / SCHEFFE;
        TITLE "Effect of Education and Occupation on Wages";
RUN;
ods rtf close;
```