STAT 4620 Final Project Report

Daphne Kaur, Arjun Narang, Charlene Li, and Samhit Kasichainula

1. Introduction

Our goal is to model the relationship between workers' wages and a set of demographic and labor-market characteristics using the Wage dataset from the ISLR package. Understanding how factors such as age, education, job classification, year, and marital status contribute to wage outcomes is important for both labor economics research and practical workforce planning. The objective is to identify a model that is both interpretable and well-supported by the data, and to summarize insights about which variables most strongly influence wages and how these effects behave across demographic groups.

2. Data Description and Cleaning

2.1 Dataset Structure

The Wage dataset includes 3000 observations with 13 variables. Table 1 shows the description and measurement type for each variable.

Table 1: Variable Descriptions for Wage Dataset

Variable	Description	Type
X	Respondent ID (unique identifier)	Numerical: Discrete
year	Year that wage information was recorded	Numerical: Discrete
age	Age of worker	Numerical: Discrete
maritl	Marital status: 1. Never Married, 2. Married, 3. Widowed, 4. Divorced, 5. Separated	Categorical: Nominal
race	Race of worker: 1. White, 2. Black, 3. Asian, 4. Other	Categorical: Nominal
education	Highest education level: 1. < HS Grad, 2. HS Grad, 3. Some College, 4. College Grad, 5. Advanced Degree	Categorical: Ordinal
region	Region of country (mid-atlantic only)	Categorical: Nominal
jobclass	Job class: 1. Industrial, 2. Information	Categorical: Nominal
health	Self-reported health status: 1. <=Good, 2. >=Very Good	Categorical: Ordinal
health_ins	Health insurance status: 1. Yes, 2. No	Categorical: Binary
logwage	Log-transformed workers wage	Numerical: Continuous
wage	Workers raw wage	Numerical: Continuous
Resp	Mystery response variable	Numerical: Continuous

2.2 Data Cleaning & Preprocessing

All original variables are complete. The appended variable Resp contains missing values. To examine potential bias, we created an indicator variable marking whether Resp was missing. Comparing summary statistics of numerical predictors between missing and non-missing Resp cases showed similar mean values, suggesting that the missingness is likely missing at random (Table 2). Therefore, removing rows with missing Resp will not bias the analysis. We also checked for duplicate rows and found none. Thus, our final cleaned dataset excludes the identifier variable X and all rows with missing Resp.

Table 2: Mean of Numeric Variables by Missingness of Resp

missing	X	year	age	logwage	wage	Resp
FALSE	219006.4	2005.79	42.42	4.65	111.73	35.66
TRUE	212854.1	2005.62	42.15	4.67	110.65	NaN

3. Exploratory Data Analysis

3.1 Summary of Numeric Variables & Categorical Variables

The variable wage shows a wide range, with a median that is lower than the mean, indicating the presence of some high earners. The variable logwage is the logarithmic transformation of wage, which reduces skewness and makes the distribution more symmetric (Table 3). The categorical variables were summarized by their most frequent categories (Table 4).

Table 3: Summary Statistics – Numeric Variables

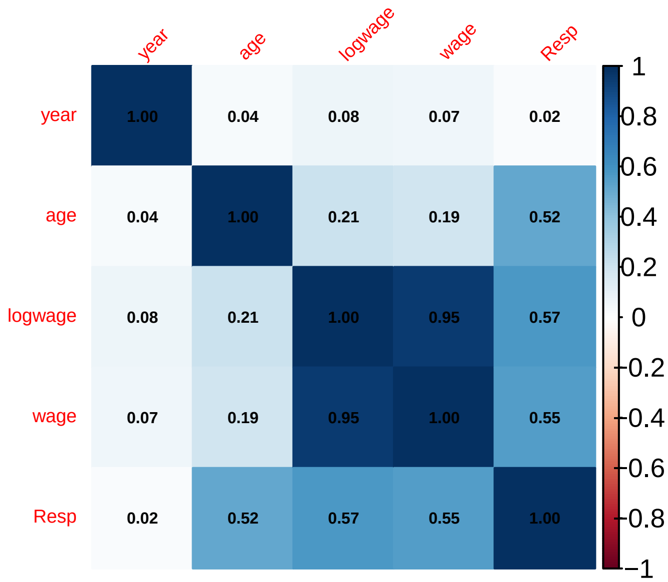
Variable	Min	Median	Mean	Max
year	2003.00	2006.00	2005.79	2009.00
age	18.00	42.00	42.42	80.00
logwage	3.00	4.65	4.65	5.76
wage	20.09	104.92	111.73	318.34
Resp	1.00	36.95	35.66	59.11

Table 4: Most Frequent Category – Categorical Variables

Variable	Most Frequent Category
maritl	2. Married
race	1. White
education	2. HS Grad
region	2. Middle Atlantic
jobclass	1. Industrial
health	2. >=Very Good
health_ins	1. Yes

3.2 Correlation Analysis

The variables wage and logwage are highly correlated, as expected due to the log transformation. The response variable Resp shows moderate positive correlations with wage, logwage, and age, suggesting that these variables may be influential in predicting Resp. The year variable shows a weak correlation with Resp, indicating limited time-based effects.

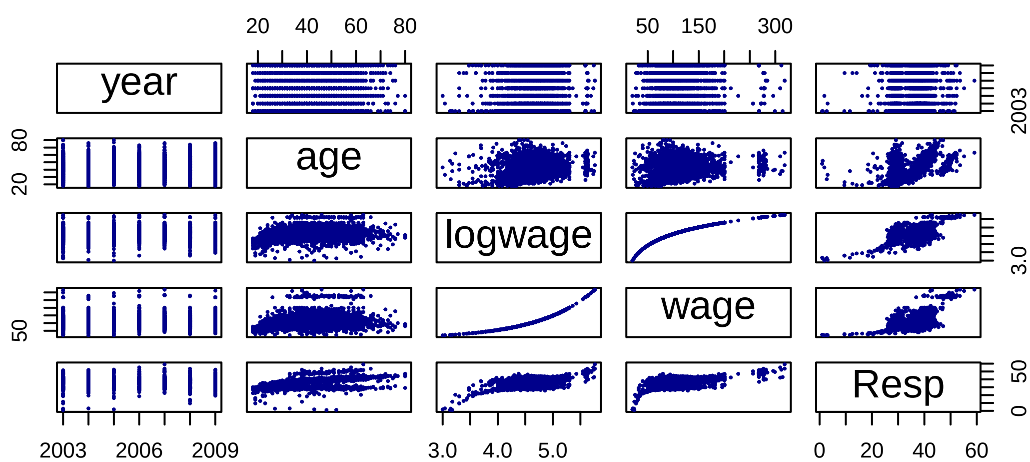


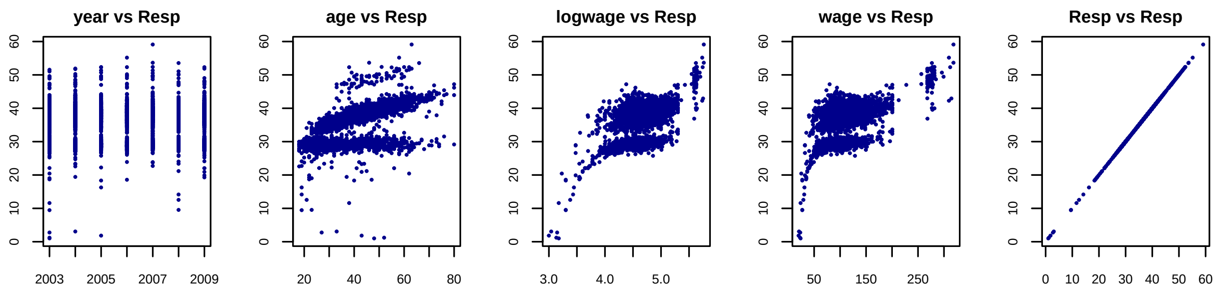
3.3 Numeric Variables vs Response

From scatterplot matrix, there is a positive, non-linear relationship. Wages tend to increase with age up to around 60, after which they seem to level off or slightly decrease. The variance also appears to increase with age. The plot also confirms that Resp is highly correlated with logwage and wage, and positively related to age.

When we specifically look at the relationship between Resp and the continuous predictors:

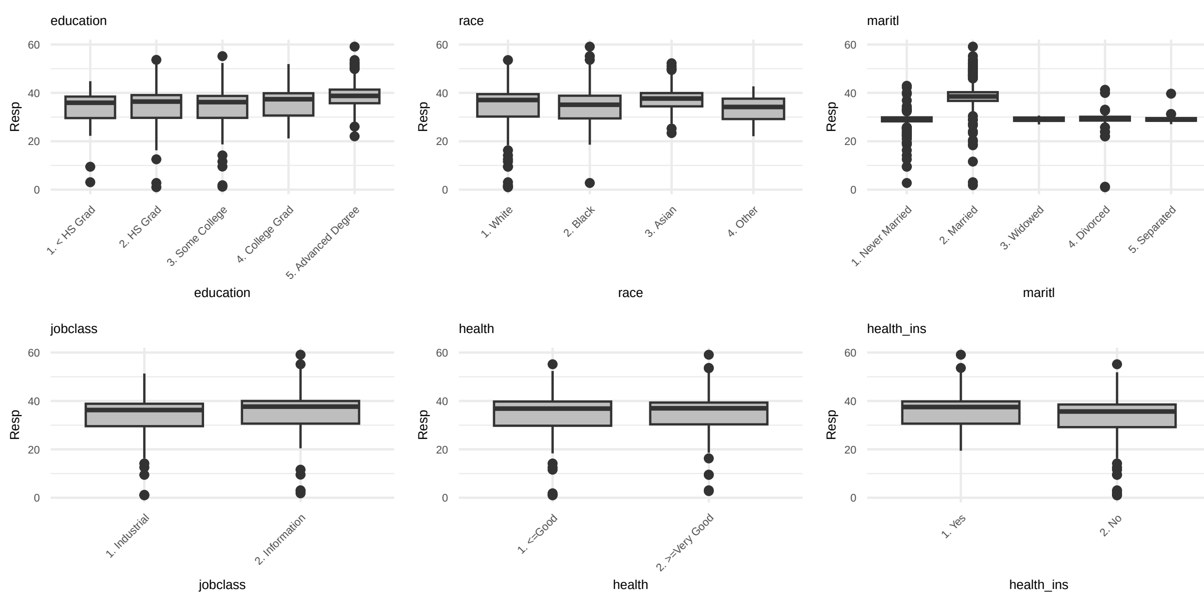
- age vs. Resp: The three apparent horizontal lines or clusters are most likely caused by the presence of a strong categorical predictor. A small number of highly influential categories in a predictor variable can create distinct, separated bands in a scatter plot when plotted against any continuous variable (like age).
- logwage/Wage vs. Resp: The plots are separated into distinct clusters primarily due to the influence of a strong categorical predictor.





3.4 Categorical Variables vs Response

- education: Shows a strong, positive relationship with the median of Resp. This indicates education is a vital predictor and should be included in your models. The differences between the groups are substantial.
- maritl (Marital Status): Reveals significant differences, especially between Married individuals (highest median/spread) and others. This also suggests it's a strong predictor.
- race, health, health_ins: Show weaker relationships with smaller differences in medians, but they still have some predictive value. They should be considered but might not be as impactful as education or marital status.
- jobclass: Shows almost no difference in the median Resp between 'Industrial' and 'Information' sectors. This suggests jobclass might be a weak predictor or could even be excluded from simpler models.
- Outlier Identification: The outliers show data points with unusually low or high wages for their respective categories, which may warrant further investigation for potential influential observations.



3.5 Variables interations with Age and Logwage

Based on what we observed, age and logwage seem to be good predictors, so let's try to break them further.

Significant interactions with age:

- age × maritl (VERY strong)
- age × health
- age × health_ins

Our interaction analysis shows that the effect of age on the response variable varies significantly by marital status, self-reported health, and health insurance status, with marital status showing the strongest interaction effect.

```
##
## Call:
## lm(formula = Resp ~ age * education, data = data_clean)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -36.464  -2.133   1.460   2.786  18.285
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   24.120395   1.006477  23.965  < 2e-16 ***
## age                            0.247916   0.023062  10.750  < 2e-16 ***
## education2. HS Grad            0.967445   1.152390   0.840  0.40125
## education3. Some College      -0.122795   1.220315  -0.101  0.91985
## education4. College Grad       1.348483   1.243922   1.084  0.27843
## education5. Advanced Degree    4.274717   1.445696   2.957  0.00313 **
## age:education2. HS Grad       -0.012724   0.026357  -0.483  0.62930
## age:education3. Some College   0.015126   0.028214   0.536  0.59190
## age:education4. College Grad   0.002141   0.028392   0.075  0.93989
## age:education5. Advanced Degree -0.025325   0.032253  -0.785  0.43241
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.714 on 2930 degrees of freedom
## Multiple R-squared:  0.2984, Adjusted R-squared:  0.2962
## F-statistic: 138.5 on 9 and 2930 DF,  p-value: < 2.2e-16
```

```
##
## Call:
## lm(formula = Resp ~ age * jobclass, data = data_clean)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -36.410  -2.284   1.420   2.679  17.964
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)              24.41599    0.45339  53.852   <2e-16 ***
## age                       0.25397    0.01054  24.097   <2e-16 ***
## jobclass2. Information    1.50252    0.67798   2.216   0.0268 *
## age:jobclass2. Information -0.01231    0.01539  -0.800   0.4238
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.786 on 2936 degrees of freedom
## Multiple R-squared:  0.2756, Adjusted R-squared:  0.2748
## F-statistic: 372.3 on 3 and 2936 DF,  p-value: < 2.2e-16
```

```
##
## Call:
## lm(formula = Resp ~ age * health, data = data_clean)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -35.719  -2.321   1.305   2.755  18.574
##
```

```
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)              25.06366    0.62412  40.159  < 2e-16 ***
## age                       0.22823    0.01338  17.051  < 2e-16 ***
## health2. >=Very Good     -0.66872    0.74268  -0.900  0.36798
## age:health2. >=Very Good  0.04718    0.01635   2.885  0.00394 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.763 on 2936 degrees of freedom
## Multiple R-squared:  0.2824, Adjusted R-squared:  0.2816
## F-statistic:   385 on 3 and 2936 DF,  p-value: < 2.2e-16
```

```
##
## Call:
## lm(formula = Resp ~ age * maritl, data = data_clean)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -36.432  -0.974  -0.085   0.824  16.549
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)            27.451294   0.368393  74.516  < 2e-16 ***
## age                     0.042513   0.010764   3.949 8.02e-05 ***
## maritl2. Married        1.579807   0.461416   3.424 0.000626 ***
## maritl3. Widowed        1.671434   3.059256   0.546 0.584865
## maritl4. Divorced       1.579128   1.060950   1.488 0.136751
## maritl5. Separated     -0.121284   1.766523  -0.069 0.945267
## age:maritl2. Married    0.172202   0.012341  13.954  < 2e-16 ***
## age:maritl3. Widowed   -0.043987   0.057946  -0.759 0.447853
## age:maritl4. Divorced  -0.042042   0.023153  -1.816 0.069493 .
## age:maritl5. Separated  0.001752   0.039946   0.044 0.965014
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.81 on 2930 degrees of freedom
## Multiple R-squared:  0.7508, Adjusted R-squared:  0.7501
## F-statistic: 980.9 on 9 and 2930 DF,  p-value: < 2.2e-16
```

```
##
## Call:
## lm(formula = Resp ~ age * health_ins, data = data_clean)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -36.222  -2.315   1.338   2.704  18.296
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        26.246081   0.433896  60.489  < 2e-16 ***
## age                 0.231193   0.009672  23.903  < 2e-16 ***
## health_ins2. No    -2.799121   0.687987  -4.069 4.85e-05 ***
## age:health_ins2. No  0.037786   0.016010   2.360   0.0183 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.773 on 2936 degrees of freedom
## Multiple R-squared:  0.2794, Adjusted R-squared:  0.2786
## F-statistic: 379.4 on 3 and 2936 DF,  p-value: < 2.2e-16
```

Interaction tests show that logwage interacts strongly with health_insurance status, indicating that the wage–response relationship differs significantly for individuals who lack health insurance. Marital status shows a borderline-significant interaction, suggesting small differences among marital groups. Other variables (education, jobclass, health) do not show meaningful interactions with logwage.

Significant interactions with logwage:

- logwage × health_ins (strongest)
- logwage × maritl

```
##
## Call:
## lm(formula = Resp ~ logwage * education, data = data_clean)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -21.4318  -3.4202   0.4979   3.3329  14.5442
##
## Coefficients:
##                                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)                          -7.5057     4.5991  -1.632   0.1028
## logwage                               9.5466     1.0439   9.146   <2e-16 ***
## education2. HS Grad                   0.1344     5.1380   0.026   0.9791
## education3. Some College            -11.6923     5.3420  -2.189   0.0287 *
## education4. College Grad              5.1910     5.2524   0.988   0.3231
## education5. Advanced Degree          -6.7308     5.6107  -1.200   0.2304
## logwage:education2. HS Grad          -0.1630     1.1599  -0.141   0.8883
## logwage:education3. Some College      2.0911     1.1966   1.748   0.0807 .
## logwage:education4. College Grad     -1.4812     1.1709  -1.265   0.2060
## logwage:education5. Advanced Degree   1.0683     1.2280   0.870   0.3844
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.556 on 2930 degrees of freedom
## Multiple R-squared:  0.3448, Adjusted R-squared:  0.3428
## F-statistic: 171.4 on 9 and 2930 DF,  p-value: < 2.2e-16
```

```
##
## Call:
## lm(formula = Resp ~ logwage * jobclass, data = data_clean)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -21.5334  -3.4330   0.4929   3.4276  13.9825
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    -5.4762     1.6410  -3.337 0.000857 ***
## logwage                         8.8189     0.3572  24.689  < 2e-16 ***
## jobclass2. Information         -1.6739     2.3052  -0.726 0.467812
## logwage:jobclass2. Information   0.3954     0.4941   0.800 0.423594
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.618 on 2936 degrees of freedom
## Multiple R-squared:  0.3253, Adjusted R-squared:  0.3246
## F-statistic: 471.8 on 3 and 2936 DF,  p-value: < 2.2e-16
```

```
##
## Call:
## lm(formula = Resp ~ logwage * health, data = data_clean)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -21.2496  -3.3901   0.4582   3.3853  13.6121
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  -7.6387     2.1353  -3.577 0.000353 ***
## logwage                       9.4104     0.4663  20.180  < 2e-16 ***
## health2. >=Very Good          0.5992     2.5239   0.237 0.812358
## logwage:health2. >=Very Good  -0.2760     0.5471  -0.505 0.613928
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.61 on 2936 degrees of freedom
## Multiple R-squared:  0.3278, Adjusted R-squared:  0.3272
## F-statistic: 477.4 on 3 and 2936 DF,  p-value: < 2.2e-16
```

```
##
## Call:
## lm(formula = Resp ~ logwage * maritl, data = data_clean)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -26.1661  -1.5550  -0.0425   1.5210  13.9609
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  4.0400     1.4291   2.827  0.00473 **
## logwage                      5.5400     0.3184  17.400  < 2e-16 ***
## maritl2. Married             5.3401     1.6461   3.244  0.00119 **
## maritl3. Widowed            18.3466    13.3674   1.372  0.17002
## maritl4. Divorced           -4.8871     2.9775  -1.641  0.10084
## maritl5. Separated           8.1478     6.2599   1.302  0.19316
## logwage:maritl2. Married     0.6660     0.3622   1.839  0.06605 .
## logwage:maritl3. Widowed    -4.0931     2.9020  -1.410  0.15852
## logwage:maritl4. Divorced    0.9841     0.6516   1.510  0.13104
## logwage:maritl5. Separated  -1.8076     1.3670  -1.322  0.18616
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.67 on 2930 degrees of freedom
## Multiple R-squared:  0.775,  Adjusted R-squared:  0.7743
## F-statistic:  1121 on 9 and 2930 DF,  p-value: < 2.2e-16
```

```
##
## Call:
## lm(formula = Resp ~ logwage * health_ins, data = data_clean)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -19.4380  -3.4463   0.4241   3.3726  14.1796
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)              -3.6252     1.5696  -2.310    0.021 *
## logwage                   8.4247     0.3305  25.494  < 2e-16 ***
## health_ins2. No         -10.0078     2.4300  -4.118 3.92e-05 ***
## logwage:health_ins2. No   2.3026     0.5302   4.343 1.46e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.6 on 2936 degrees of freedom
## Multiple R-squared:  0.3308, Adjusted R-squared:  0.3301
## F-statistic: 483.7 on 3 and 2936 DF,  p-value: < 2.2e-16
```

3.6 Compare Models With and Without Interaction

The interaction between age and marital status is highly significant ($p < 0.001$), indicating that the effect of age on Resp depends strongly on marital status. The interactions of age with health and health insurance are also significant ($p < 0.01$ and $p < 0.05$), suggesting that the influence of age on Resp varies slightly depending on a person's health and health insurance status.

```
## Analysis of Variance Table
##
```

```
## Model 1: Resp ~ age + maritl
## Model 2: Resp ~ age * maritl
##   Res.Df   RSS Df Sum of Sq      F    Pr(>F)
## 1   2934 25331
## 2   2930 23129  4      2202 69.738 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## Analysis of Variance Table
##
## Model 1: Resp ~ age + health
## Model 2: Resp ~ age * health
##   Res.Df   RSS Df Sum of Sq      F   Pr(>F)
## 1   2937 66801
## 2   2936 66612  1    188.85 8.3237 0.003942 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## Analysis of Variance Table
##
## Model 1: Resp ~ age + health_ins
## Model 2: Resp ~ age * health_ins
##   Res.Df   RSS Df Sum of Sq      F  Pr(>F)
## 1   2937 67014
## 2   2936 66887  1     126.9 5.5703 0.01833 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Table 5: ANOVA Interaction Summary

Interaction	F	p-value	Interpretation
age × maritl	69.738	$< 2.2\text{e-}16$	Highly significant → include interaction
age × health	8.324	0.00394	Significant → interaction matters
age × health_ins	5.570	0.0183	Significant → include interaction

The interaction between logwage and health insurance is highly significant ($p < 0.001$), meaning the effect of wages on Resp depends substantially on whether someone has health insurance. The interaction between logwage and marital status is barely significant ($p = 0.047$), suggesting that wages' effect on Resp may slightly differ across marital status groups.

```
## Analysis of Variance Table
##
## Model 1: Resp ~ logwage + health_ins
## Model 2: Resp ~ logwage * health_ins
##   Res.Df   RSS Df Sum of Sq      F    Pr(>F)
## 1   2937 62518
## 2   2936 62119  1    398.99 18.858 1.456e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## Analysis of Variance Table
```

```
##
## Model 1: Resp ~ logwage + maritl
## Model 2: Resp ~ logwage * maritl
##   Res.Df   RSS Df Sum of Sq      F  Pr(>F)
## 1   2934 20957
## 2   2930 20888  4    68.884 2.4157 0.04675 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Table 6: ANOVA Interaction Summary

Interaction	F	p-value	Interpretation
logwage × health_ins	18.858	1.456e-05	Interaction significantly improves the model, so you should include it.
logwage × maritl	2.416	0.04675	Interaction barely significant, so including it may slightly improve the model.