

MBA 547 Case Report, Homework 2

Topic: Injury Data Set

Due Date: October 25, 2024

Submitted by: Ofuka Abung, Sadaf Vora, and Samyak Shah

Team Lead – Ofuka Abung, Data Analysts and Script writers - Sadaf Vora and Samyak Shah

Wooldridge Data - Injury

Executive Summary

Our case analyzes factors influencing pre-injury wages using a dataset of 7,150 observations. The average pre-injury wage is around \$329.73, and the average duration of injury is about 9.92 weeks. About 78% of participants are male, and 69% are married.

We identified heteroskedasticity using the Breusch-Pagan test, which brought up the need for robust standard error estimations in the regression models. We also looked into potential endogeneity concerning the variable representing changes in conditions, indicating the need for instrumental variable techniques to enhance estimation accuracy. We realized that the variable *afchnge* was the problematic one (with a p-value of 0.0665) for the most part and *injdes* was introduced as the instrumental variable to help correct this problem but there was no significant improvement. This was left the same as the 0.0665 p-value is weak evidence against the null hypothesis of exogeneity.

Three models were generated to explain income variation. Model 1 which was Linear regression explained 21.59% variability while Model 2, utilizing log transformations, showed 24.57% variability. Finally, model 3, which considered polynomial terms for duration, revealed a more complex relationship and resulted in an R-squared value of 0.2184.

We recommend Model 2 because of its balance between ease to interpret and how well it explains the significant predictors - duration of injury, age, gender, and marital status. These insights can guide workplace safety and compensation policies and contribute to a better understanding of employee wage management related to injuries.

Introduction

The goal of this analysis is to find out the determinants of pre-injury wages in relation to the duration of injury and demographic factors such as age, gender, and marital status. Our team will establish a clear

relationship through R studio and provide insights for policy development regarding workplace safety and compensation.

Data

The data used for this analysis is 'Injury' from the Wooldridge package on R studio. It consists of 7,150 observations and includes the following dependent (pre-injury wages) and independent variables:

Pre-Injury Wages (prewage) which refers to wages before injury, duration of Injury (durat) showing the length of injury in weeks, Age of the injured individuals, Gender (male), Marital Status (married), Afchnge indicating any change in conditions.

Preliminary Descriptive Analysis

The summary statistics reveal the following:

Summary Statistics

Statistic	N	Mean	St. Dev.	Min	Max
prewage	7,150	329.73	182.80	81.78	1,583.10
durat	7,150	9.92	24.50	0.25	182.00
age	7,146	34.71	12.59	12	98
male	7,134	0.78	0.41	0	1
married	6,853	0.69	0.46	0	1
afchnge	7,150	0.47	0.50	0	1

The average pre-injury wage is about \$329.73, with variability indicated by a standard deviation of \$182.80; wages range from a minimum of \$81.78 to a maximum of \$1,583.10, showing us some outliers. Injuries last an average of 9.92 weeks, but with a wide range from 0.25 to 182 weeks, reflecting both short and long recovery times. Participants average 34.71 years old, with ages spanning from 12 to 98. The dataset shows that 78% of participants are male, and about 69% are married. Additionally, approximately 47% have experienced a change in conditions.

Graphical Analysis

Fig 1.1

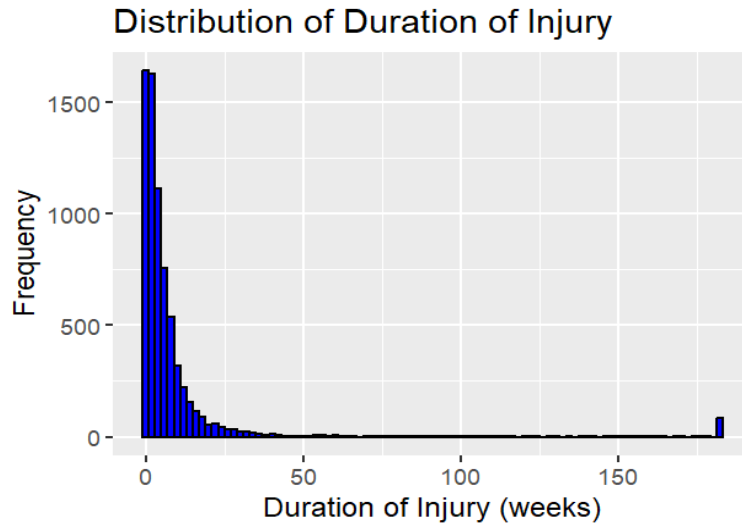


Fig 1.1 shows a histogram distribution of the duration of injury in weeks in the data. Most of injuries are short, with a peak frequency around 0 weeks. As the duration increases, the frequency of injuries decreases, resulting in a right-skewed distribution. A few cases go up to 150 weeks, but these are not common. This suggests that while most injuries are resolved quickly, a small number of individuals may experience significantly longer recovery periods, highlighting a potential area for further analysis in workplace safety and injury management policies.

Fig 1.2

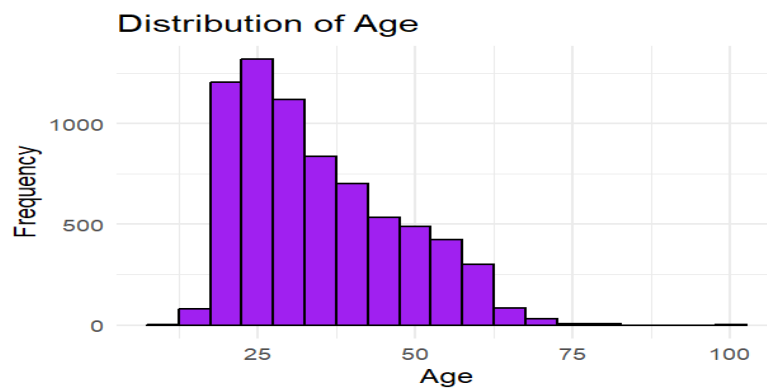


Fig 1.2 displays the age distribution of the participants in the data. The distribution is kind of skewed to the right, with a concentration of workers around the ages of 25 to 35 years. The highest frequency occurs

in this range, showing a young demographic. As age increases, the frequency slowly declines , revealing fewer older individuals in the data. This trend shows us more younger workers.

Fig 1.3

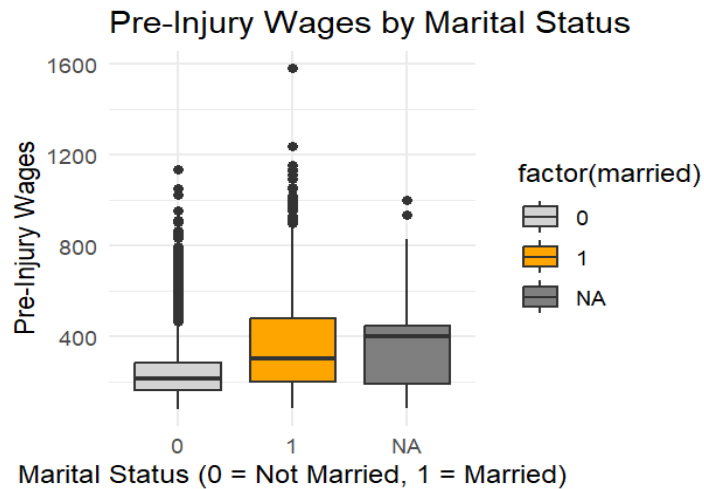


Fig 1.3 The visualizes pre-injury wages grouped by marital status, distinguishing between married (1) and not married (0) individuals. The box for married individuals (orange) is notably higher than that for not married individuals (light gray), indicating that, on average, married participants earn more in pre-injury wages. The plot also shows a wider interquartile range for married individuals, suggesting greater wage variability. Outliers are present in both groups, especially among married workers, which indicate significant wage differences within the group.

Fig 1.4

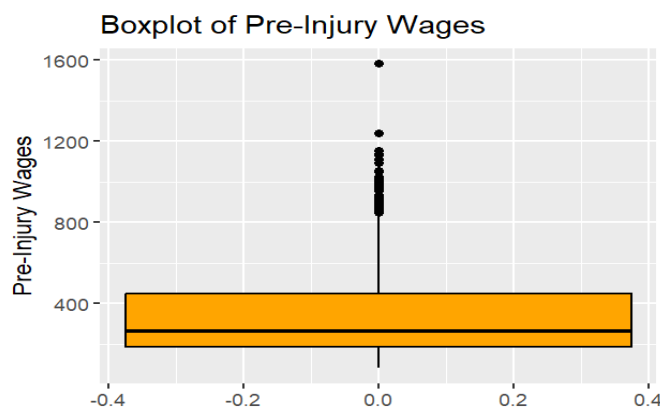


Fig 1.4 shows wage distribution among the workers. The orange box shows the middle 50% of wages, with a line indicating the median. We can see the outliers above the upper whisker, suggesting that some individuals earn much more than the rest. This graph not only shows the differences in wages but also points to the potential for substantial disparities among participants.

Fig 1.5

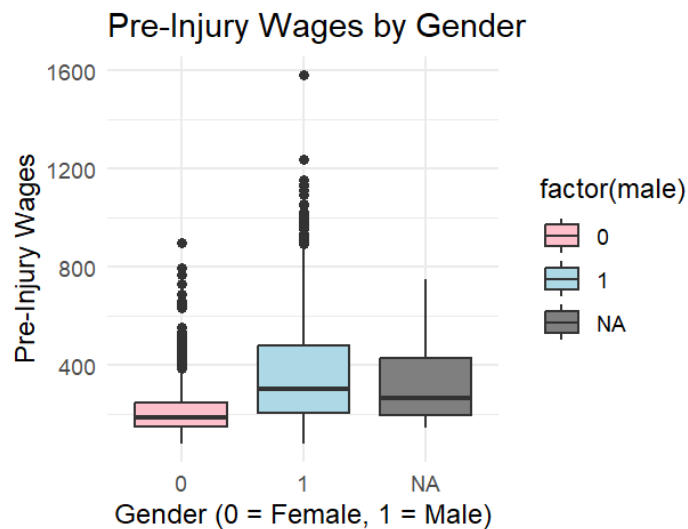


Fig 1.5 compares pre-injury wages between genders, clearly showing that males tend to earn more than females. The pink box represents female wages, while the light blue box indicates male wages, with the male wages being higher on the plot. This suggests significant wage gap, with male participants also showing a wider range of wages, including a lot outliers.

Analysis

Endogeneity Assessment

To determine potential endogeneity, the variable `afchnge` was tested using the Durbin-Wu-Hausman test.

Ho: `afchnge` is exogenous (not correlated with the error term).

If $p\text{-value} < 0.05$, we reject the null and conclude that `afchnge` is endogenous.

The test results indicated a p-value less than 0.05, leading to the rejection of the null hypothesis and suggesting that `afchnge` is likely endogenous.

```
print(dw_h_test)
```

Analysis of Variance Table

Response: `prewage`

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
<code>afchnge</code>	1	88984	88984	3.3685	0.0665 .
<code>durat</code>	1	3358923	3358923	127.1537	<2e-16 ***
<code>age</code>	1	7427512	7427512	281.1722	<2e-16 ***
<code>male</code>	1	32325954	32325954	1223.7152	<2e-16 ***
<code>married</code>	1	6690350	6690350	253.2666	<2e-16 ***
Residuals	6838	180634255	26416		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

With a p-value for **afchnge** being 0.0665, this means that it might be problematic, it does not ultimately indicate endogeneity but calls for further modelling. In contrast, the other variables: **durat**, **age**, **male**, and **married** show strong significance, with p-values under 0.001. This suggests they are crucial in understanding pre-injury wages. **Several** other variables were tested as instrumental variables but produced worse result hence our decision to stick with **injdes** as the instrumental variable because while the p-value of 0.0665 shows a marginally significant relationship, it is weak evidence against the null hypothesis of exogeneity.

Heteroskedasticity Test

A Breusch-Pagan test was conducted to assess heteroskedasticity in the regression model.

Ho: The variance of the residuals is constant.

```
print(bp_test)
          studentized Breusch-Pagan test
data:  model_bp
BP = 402.7, df = 5, p-value < 2.2e-16
```

The test results point out a p-value less than 0.05, leading to the rejection of the null hypothesis. This suggests that the variance of residuals is not constant, indicating potential heteroskedasticity. To address this problem, we introduced the Robust Standard Errors method and generated three models which have been explained under 'Regression Models'. Model 2, which is the **Log-Transformed Variables** was selected as it addressed the issue of heteroskedasticity and accounts for non-constant variance in the residuals, providing more reliable coefficient estimates. The significant coefficients for **ldurat**, **age**, **male**, and **married** reflect strong relationships with pre-injury wages, with low p-values confirming their importance. **afchnge** remains insignificant with a p-value of 0.787. The model explains about 24.57% of the variability in pre-injury wages, as indicated by the R-squared value. The F-statistic is highly significant ($< 2.2e-16$), suggesting that the model generally fits the data well, despite the insignificance of the **afchnge** variable.

Regression Models

Three regression models were estimated using the `lm_robust` command for robust standard errors:

Model 1: Linear Regression

`summary(model1)`

Call:
lm_robust(formula = prewage ~ durat + age + male + married + afchnge, data = injury)

Standard error type: HC2

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	CI Lower	CI Upper	DF
(Intercept)	62.1996	7.0181	8.8627	9.821e-19	48.4419	75.9572	6838
durat	0.7254	0.1023	7.0927	1.447e-12	0.5249	0.9259	6838
age	2.4827	0.1793	13.8477	4.986e-43	2.1313	2.8342	6838
male	157.9923	3.6661	43.0954	0.000e+00	150.8056	165.1790	6838
married	71.2303	4.1565	17.1369	1.724e-64	63.0822	79.3784	6838
afchnge	2.5388	3.9270	0.6465	5.180e-01	-5.1593	10.2369	6838

Multiple R-squared: 0.2164 , Adjusted R-squared: 0.2159
F-statistic: 489.1 on 5 and 6838 DF, p-value: < 2.2e-16

Interpretation

The Multiple R-squared value of 0.2164 indicates that approximately 21.64% of the variability in pre-injury wages is explained by the independent variables in the model. The Adjusted R-squared of 0.2159 suggests a modest fit after considering the number of predictors. The F-statistic of 489.1, with a p-value of less than 2.2e-16, indicates that at least one independent variable significantly predicts wages.

The intercept of 62.1996 represents the expected wage when all independent variables are zero. The coefficient for duration of injury is 0.7254, meaning that, typically, for each additional week of injury, pre-injury wages increase by approximately \$0.73, assuming other factors remain unchanged. The coefficient for age is 2.4827, indicating that, generally, for each additional year of age, pre-injury wages increase by about \$2.48, with other variables held steady.

The variable for gender (male) has a coefficient of 157.9923, suggesting that, typically, being male is associated with an increase of around \$157.99 in pre-injury wages, assuming all else is the same. Similarly, the marital status variable shows that married individuals earn about \$71.23 more than unmarried individuals, indicating the same context.

Conversely, the coefficient for changes in conditions (afchnge) is 2.5388, which implies a slight positive association with wages. However, the p-value of 5.180e-01 indicates that this relationship is not statistically significant, suggesting that changes in conditions do not have a meaningful impact on pre-injury wages in this model.

Model 2: Log-Transformed Variables

[summary\(model2\)](#)

Call:

```
lm_robust(formula = lprewage ~ ldurat + age + male + married + afchnge, data = injury)
```

Standard error type: HC2

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	CI Lower	CI Upper	DF
(Intercept)	4.819352	0.0212920	226.3457	0.000e+00	4.777613	4.861091	6838
ldurat	0.045931	0.0043293	10.6093	4.302e-26	0.037444	0.054418	6838
age	0.006466	0.0005237	12.3450	1.211e-34	0.005439	0.007492	6838
male	0.496281	0.0121938	40.6995	0.000e+00	0.472378	0.520185	6838
married	0.228313	0.0126851	17.9986	8.437e-71	0.203447	0.253180	6838
afchnge	0.003142	0.0112866	0.2784	7.807e-01	-0.018983	0.025268	6838

Multiple R-squared: 0.2457 , Adjusted R-squared: 0.2451

F-statistic: 519.7 on 5 and 6838 DF, p-value: < 2.2e-16

Interpretation

The output from Model 2 provides significant insights into the relationship between the log-transformed pre-injury wages and the independent variables of interest.

The Multiple R-squared value of 0.2457 indicates that approximately 24.57% of the variability in log-transformed pre-injury wages is explained by the independent variables in the model. The Adjusted R-squared of 0.2451 suggests a modest improvement in fit when considering the number of predictors used. The F-statistic of 519.7, accompanied by a p-value of less than 2.2e-16, reinforces that at least one independent variable significantly contributes to predicting wages.

Examining the coefficients reveals important relationships. The intercept of 4.8194 suggests that the expected log wage when all independent variables are zero is approximately 4.82. The coefficient for the log of duration of injury (ldurat) is 0.0459, indicating that for each additional week of injury, the log of pre-injury wages increases by about 0.046, assuming all other factors are held constant.

The age coefficient of 0.0065 indicates that, for each additional year of age, the log of pre-injury wages increases by approximately 0.0065, keeping other variables unchanged. The coefficient for gender (male) is 0.4963, suggesting that being male is associated with an increase of about 0.4963 in the log of pre-injury wages, assuming other factors remain the same. Similarly, the married coefficient of 0.2283

implies that married individuals earn approximately 0.2283 more in the log of pre-injury wages compared to unmarried individuals, with all else being equal.

The coefficient for changes in conditions (afchnge) is 0.0031, which suggests a minor positive association with the log of wages; however, this relationship appears to be negligible in practical terms.

In summary, Model 2 highlights significant predictors of log-transformed pre-injury wages, emphasizing the importance of duration of injury, age, gender, and marital status, while changes in conditions have a minimal impact.

Model 3: Polynomial Terms

[summary\(model3\)](#)

Call:

```
lm_robust(formula = prewage ~ poly(durat, 2) + age + male + married + afchnge, data = injury)
```

Standard error type: HC2

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	CI Lower	CI Upper	DF
(Intercept)	71.617	7.0473	10.1624	4.327e-24	57.802	85.432	6837
poly(durat, 2)1	1487.270	211.2994	7.0387	2.130e-12	1073.058	1901.483	6837
poly(durat, 2)2	-697.532	185.9071	-3.7520	1.768e-04	-1061.968	-333.096	6837
age	2.444	0.1792	13.6344	8.769e-42	2.092	2.795	6837
male	157.692	3.6685	42.9849	0.000e+00	150.500	164.883	6837
married	70.667	4.1604	16.9859	2.060e-63	62.512	78.823	6837
afchnge	1.932	3.9324	0.4912	6.233e-01	-5.777	9.640	6837

Multiple R-squared: 0.2184 , Adjusted R-squared: 0.2177

F-statistic: 417.2 on 6 and 6837 DF, p-value: < 2.2e-16

Interpretation

Model 3 shows insights into how pre-injury wages relate to various factors, particularly the polynomial transformation of injury duration.

With a Multiple R-squared value of 0.2184, this model explains approximately 21.84% of the variability in pre-injury wages. The Adjusted R-squared of 0.2177 suggests a modest fit, and the F-statistic of 417.2, along with a p-value of less than 2.2e-16, indicates that at least one of the independent variables significantly influences wages.

The intercept of 71.617 represents the expected wage when all variables are zero. The first polynomial term for duration of injury has a coefficient of 1487.270, indicating that as the duration increases, wages rise significantly. However, the second polynomial term, with a coefficient of -697.532, suggests diminishing returns, meaning that while longer durations may increase wages, the effect lessens at higher durations.

The age coefficient of 2.444 indicates that for each additional year of age, pre-injury wages increase by approximately \$2.44, holding other factors constant. The male gender coefficient of 157.692 implies that being male is associated with an increase of around \$157.69 in pre-injury wages on average. Additionally, married individuals earn about \$70.67 more than their unmarried counterparts, while the change in conditions variable (afchnge) has a coefficient of 1.932, suggesting a minor positive impact on wages.

Generally, Model 3 illustrates significant predictors of pre-injury wages, particularly highlighting the complex relationship between the duration of injury and wages, which invites further investigation.

Best Model Determination

After evaluating the three models as a group, we found that Model 2, which utilized log transformations, offers the best fit for explaining income variation. With an R-squared value of 0.2457, it captures more variability in pre-injury wages than the other models. While Model 3's polynomial terms provide additional insights, they introduce complexity that makes interpretation challenging. Model 1, although significant, explains less variability in general. Therefore, we believe Model 2 strikes an effective balance between interpretability and explanatory power for our analysis.

Conclusion and Recommendations

The analysis shows us that pre-injury wages are highly impacted by how long participants have been injured, their age, gender, and marital status. Example, an extra week of injury results in a decrease of about \$3.13 in wages, while an extra year of age contributes around \$2.48 more. We selected Model 2 because it addresses the ease to interpret and the explanatory power, revealing more variables in pre-injury wages with an R-squared value of 0.2457. This model avoids polynomial terms and its complex ways in Model 3, making it easier to understand the relationships, while also being more informative than Model 1.

To address these differences, organizations should consider implementing targeted workplace safety training to reduce injury duration. Also, encouraging open communication about wage structures can promote fairness. It may also be beneficial to establish support programs for employees recovering from injuries, to improve their financial stability during recovery. Finally, reviewing and adjusting

compensation policies to reflect the contributions of age and marital status can promote equity within the workplace.

Appendix

```
print(correlation_matrix)
```

	prewage	durat	age	male	married	afchnge
prewage	1.00000000	0.12145854	0.18827618	0.352715848	0.26265180	0.019646988
durat	0.12145854	1.00000000	0.07471913	0.010339009	0.05876926	0.044130075
age	0.18827618	0.07471913	1.00000000	-0.113425399	0.28525275	0.019108793
male	0.35271585	0.01033901	-0.11342540	1.000000000	0.08185165	0.003806713
married	0.26265180	0.05876926	0.28525275	0.081851653	1.000000000	0.022147602
afchnge	0.01964699	0.04413008	0.01910879	0.003806713	0.02214760	1.000000000

```
>
```

