Northeastern University

ALY-6010
PROBABILITY THEORY AND INTRODUCTORY STATISTICS

BY INSTRUCTOR
ROY WADA

MODULE-6 PROJECT REPORT
SUBMITTED BY

ITI ROHILLA

On Dec-17-2023

SUMMARY

The full-sample regression reveals significant relationships between lnWage and several predictors, including experience, education, gender, and marital status.

Positive coefficients for experience, education, gender (male), and being married suggest that these factors contribute to higher lnWages on average. The model, with an R-squared value of approximately 35.73%, explains a substantial portion of the variance in lnWage.

Further exploration through occupation-specific regressions uncovers nuanced patterns within different job sectors. The analysis by occupation levels (white and blue) indicates distinct impacts of experience, education, gender, and marital status on lnWage within each category.

Understanding these variations is crucial for tailoring policies and decisions to the unique dynamics of each occupation, providing a more accurate depiction of wage determinants.

The introduction of a categorical variable, "occupation," with dummy variables reveals that being in the "white" occupation category is associated with a significantly higher lnWage.

The positive coefficient for "occupation-white" is both substantial and statistically significant, emphasizing the impact of occupation on wage differentials.

Part-1

The regression results provide valuable insights into the relationship between the natural logarithm of wages (lnWage) and several independent variables, including experience, education, gender, and marital status. Let's interpret the key results.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	6.44419291	0.09041222	71.2756818	4.26000584329609e-292
experience	0.0062895	0.00139513	4.50819214	7.8884866217137e-06
education	0.07676437	0.00532855	14.4062421	1.55234485940021e-40
gendermale	0.34065754	0.06502853	5.23858599	2.25515213652282e-07
marriedyes	0.12592024	0.05237396	2.40425266	0.016512804

Coefficients:

1. **Intercept (6.444193):**
    - The intercept represents the estimated lnWage when all other variables are zero. In this context, it may not have a direct, meaningful interpretation, as having zero experience, education, and other factors is not realistic.
2. **Experience (0.006289):**
    - The coefficient for experience is 0.006289, indicating that for each additional experience unit, the expected lnWage increases by approximately 0.63%. This suggests a positive association between experience and wages.
3. **Education (0.076764):**
    - The coefficient for education is 0.076764, suggesting that a one-unit increase in education is associated with an approximate 7.68% increase in lnWage. This indicates a positive relationship between education and wages.
4. **Gender (male, 0.340658):**
    - The coefficient for the 'gendermale' variable is 0.340658. Since it is a binary variable (1 for male, 0 for female), this implies that being male is associated with an estimated 34.07% higher lnWage compared to being female.
5. **Marital Status (married, 0.125920):**
    - The coefficient for 'marriedyes' is 0.125920, suggesting that being married is associated with an estimated 12.59% higher lnWage compared to being unmarried.

Statistical Significance:

- All coefficients are statistically significant at conventional levels (indicated by the asterisks). The p-values for each coefficient are well below 0.05, suggesting that the relationships between lnWage and experience, education, gender, and marital status are unlikely to be due to random chance.

Model Fit:

- The R-squared value (0.3573) indicates that the model explains approximately 35.73% of the variance in lnWage. This suggests that the selected independent variables collectively contribute to explaining the variation in wages, though a substantial portion remains unexplained.

Conclusion:

In summary, the regression analysis highlights significant relationships between lnWage and experience, education, gender, and marital status. The model suggests that, on average, individuals with more experience, higher education levels, male gender, and married status tend to have higher lnWages. However, it's important to note that the model's explanatory power is not exhaustive, and there may be other factors influencing wage variation not captured in the model.

Part-2

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)...1	6.60282182910031	0.161225602876584	40.9539286024856	2.02614093982667E-121
experience...2	0.006827593756214	0.00211851442140073	3.22282146736565	0.00141667541251907
education...3	0.0708446344910968	0.00985663183564695	7.18750945276094	5.80387953535108E-12
gendermale...4	0.350111535924302	0.0941437695976899	3.71890288035474	0.000240918832563838
marriedyes...5	0.0586933493712288	0.0790168326012325	0.742795521397719	0.458217291151194
(Intercept)...6	6.6457614700492	0.139957801349777	47.4840373738107	1.43314974055762E-141
experience...7	0.00507432033479737	0.00182776462393741	2.7762438709784	0.00584480729265743
education...8	0.0513100517554049	0.00965321978387815	5.31533031508284	2.08162987952516E-07

gendermale...9	0.361059664645658	0.0892050181673569	4.0475263843148	6.59246944744323E-05
marriedyes...10	0.193439020993486	0.0681860715567611	2.83692866559205	0.00486512178318644

Occupation Level: "White"

For individuals in the "white" occupation category, the regression analysis reveals interesting insights into the factors influencing lnWage. The intercept of 6.60 suggests a high baseline lnWage for this category when all other predictors are zero.

The positive coefficient for experience (0.0068) indicates that as experience increases, lnWage tends to rise. Similarly, the positive coefficient for education (0.0708) suggests that higher levels of education are associated with higher lnWage.

The gender coefficient (0.3501) implies a positive association between being male and lnWage, while the relatively small and statistically insignificant coefficient for being married (0.0587) suggests that marital status may not strongly influence lnWage in this "white" occupation.

Occupation Level: "Blue"

Moving to individuals in the "blue" occupation category, similar patterns emerge. The intercept of 6.65 indicates a high baseline lnWage for this category. Positive coefficients for experience (0.0051), education (0.0513), and being male (0.3611) suggest that, like the "white" category, more experience, higher education, and being male are associated with higher lnWage.

Notably, the coefficient for being married is larger in this "blue" occupation (0.1934) compared to the "white" occupation, implying that marital status may have a more pronounced effect on lnWage in this category.

In summary, the regression analysis by occupation levels reveals distinct patterns in how experience, education, gender, and marital status influence lnWage in the "white" and "blue" occupational categories. These findings provide a more nuanced understanding of the wage determinants within specific professional contexts, emphasizing the importance of considering occupational heterogeneity in wage analysis.

Subset Regressions v/s Full-Sample Regression

1. **Occupational-Specific Effects:**
    - **Subset Regressions:** By running regressions separately for each occupational category, you capture variations in the relationship between predictors (experience, education, gender, marital status) and wages that are specific to each job sector.
    - **Full-Sample Regression:** The full-sample regression estimates a single set of coefficients for the entire dataset, assuming that the relationships are consistent across all occupational categories.
2. **Heterogeneity in Coefficients:**
    - **Subset Regressions:** Coefficients for each predictor variable may vary across different occupational categories. For example, the impact of education on wages may differ between occupations, reflecting the diverse nature of work and salary structures.
    - **Full-Sample Regression:** Assumes a uniform relationship between predictors and wages for the entire dataset, potentially oversimplifying the complex dynamics present in different job sectors.
3. **Identification of Significant Predictors:**
    - **Subset Regressions:** Allow for the identification of significant predictors within each occupational category. It helps discern which variables are particularly influential in explaining wage variations within specific occupations.

- **Full-Sample Regression:** Provides an overall assessment of the significance of predictors across the entire dataset but might obscure specific effects within subsets.

4. **Precision of Estimates:**
    - **Subset Regressions:** Estimates for coefficients in subset regressions may have higher precision because they focus on a more homogeneous group of observations.
    - **Full-Sample Regression:** Estimates may have lower precision if the dataset contains diverse occupational categories with different underlying relationships.
5. **Generalization vs. Specificity:**
    - **Subset Regressions:** Provide more specific insights into the relationships within each occupational category, allowing for a more tailored understanding of wage determinants.
    - **Full-Sample Regression:** Offers a more generalized view applicable to the entire dataset but might overlook specific nuances present in individual job sectors.

In summary, the subset regressions offer a more detailed and context-specific analysis, revealing how the relationships between predictors and wages vary across different occupational categories. This approach is valuable for understanding the heterogeneity in the labor market and tailoring insights to specific job sectors.

Understanding data by two different methods:

Analyzing the data through subset regressions for various occupational categories offers a more nuanced perspective on wage determinants. This method recognizes and embraces the inherent diversity within the labor market, allowing for the identification of occupation-specific patterns and trends.

By conducting regressions on subsets based on occupational categories, the analysis provides context-specific insights into the impact of

predictors like experience, education, gender, and marital status, revealing how these factors uniquely influence wages across different job sectors.

The granularity of subset regressions not only aids in tailoring policies and decision-making to the specific needs of each occupation but also enhances model fit and predictive accuracy within homogeneous groups. This approach enables a more accurate depiction of wage differentials, identifies outliers or anomalies, and contributes to an overall improved understanding of how socioeconomic factors shape earnings within specific occupational contexts.

Part-3

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	6.57094378142419	0.096971089356639	67.7618847536883	2.85124748980595E-280
experience	0.00596919421644382	0.00138582813582024	4.30731204119389	1.93596640359174E-05
education	0.0617129713950761	0.00687530268830479	8.97603701143991	3.74521359048361E-18
gendermale	0.351845856364718	0.0645304059542373	5.4524041986383	7.31944885709012E-08
marriedyes	0.127651727774393	0.0519083712629225	2.45917420771731	0.0142121254954976
occupationwhite	0.128412841850652	0.037560051948171	3.418867525206	0.000672272351832111

Impact of a categorical variable

The impact of categorical variables on regression can be understood by examining the coefficients associated with them. In your example, the

categorical variable is "occupation," which has been encoded as dummy variable. The dummy variables represent different levels or categories of the "occupation" variable.

This coefficient represents the change in the predicted value of the dependent variable (lnWage) when the "occupation" is coded as "white" compared to the reference category (which is not explicitly shown in your output but is the category excluded in the dummy coding). Here are some key interpretations:

1. **The magnitude of the Coefficient:**
    - A positive coefficient (0.128413) suggests that, compared to the reference category, individuals with the "white" occupation tend to have a higher predicted lnWage.
2. **Significance Level:**
    - The p-value associated with the coefficient (0.000672) is less than the conventional significance level of 0.05, indicating that the "occupationwhite" variable is statistically significant.
3. **Interpretation:**
    - Holding other variables constant, individuals with the "white" occupation have a predicted lnWage that is 0.128413 units higher than those in the reference category.
    -

In summary, the coefficient for the "occupation-white" variable provides insights into the impact of the "white" category of the occupation variable on the dependent variable (lnWage). It suggests a positive impact, but the magnitude and significance should be considered when interpreting the results. It's important to note that the interpretation of coefficients for dummy variables is always in comparison to the reference category.

Understanding intercepts and lines

In the provided linear regression model, there is one intercept and multiple coefficients representing the slopes of the lines associated with each predictor variable. The intercept is the value of the dependent variable when all predictor variables are zero.

Intercept:

- The intercept is represented by the term **(Intercept)** in the summary output.
- In your example, the intercept is given by:

(Intercept): 6.570944

- This means that when all predictor variables (experience, education, gender, married, occupation) are zero, the predicted value of **lnWage** is 6.570944.

Lines (Coefficients):

- The lines are represented by the coefficients associated with each predictor variable.
- For example:

experience: 0.005969
education: 0.061713
gender male: 0.351846
married-yes: 0.127652.
occupation-white: 0.128413

Each of these coefficients represents the change in the predicted value of **lnWage** for a one-unit change in the corresponding predictor variable, holding other variables constant.

So, there is one intercept and five lines (associated with the five predictor variables) in this multiple linear regression model.

CONCLUSION

In conclusion, the regression analysis underscores the multifaceted nature of wage determination. While certain predictors exhibit consistent impacts across the entire dataset, the occupation-specific analysis highlights the importance of considering occupational heterogeneity in wage studies. Tailoring insights to specific job sectors are crucial for a more accurate understanding of wage dynamics.

The positive association between lnWage and being in the "white" occupation category suggests that occupation plays a substantial role in determining wages, beyond the effects of individual characteristics. Policymakers and stakeholders should consider these findings to formulate targeted interventions and strategies that address the unique dynamics of different occupations, ultimately contributing to a more equitable and informed approach to wage-related decisions.

Additionally, the robustness of the subset regressions in capturing occupation-specific effects enhances the precision of estimates and provides a more tailored understanding of wage determinants. This approach allows for a nuanced analysis that goes beyond the limitations of a one-size-fits-all model, offering a valuable framework for policymakers and researchers seeking to unravel the intricacies of wage differentials across diverse occupational contexts.

CITATIONS

- Chi-Square Test in R | Explore the Examples and Essential concepts!

  https://data-flair.training/blogs/chi-square-test-in-r/

- ANOVA in R | A Complete Step-by-Step Guide with Examples

  https://www.scribbr.com/statistics/anova-in-r/

- How to Draw a Horizontal Barplot in R
  https://www.geeksforgeeks.org/how-to-draw-a-horizontal-barplot-in-r/

- How to Create Summary Tables in R?

  https://www.geeksforgeeks.org/how-to-create-summary-tables-in-r/