Project Step 2

Alec Wang, Gary Han

2023-11-04

Recap

This dataset includes 14 parameters from 30162 adults collected during the 1994 census as predictors for whether or not income exceeds $50k/yr. We have sampled 500 entries from the dataset for analysis. Note that we have used the same seed for the sample in this step so results are consistent.
Below is a table detailing the 14 parameters and the response, that were collected in the census. The link to the source is here (https://archive.ics.uci.edu/dataset/2/adult)

Field	Description
age	Age in years of individual (Integer)
workclass	Class of work of individual (7 categories)
fnlwgt	Number of people the entry represents (Integer)
education	Highest level of education of individual (16 categories)
education-num	Maps each category in education to a number (Integer)
marital-status	Marital status of individual (7 categories)
occupation	Description of occupation (14 categories)
relationship	Relationship of individual relative to others (6 categories)
race	Category of race of individual (5 categories)
sex	Biological sex of individual (2 categories)
capital-gain	Capital gain of individual (Integer)
capital-loss	Capital loss of individual (Integer)
hours-per-week	Hours worked per week by individual (Integer)
income	Whether or not income is above $50k/yr (2 categories)

First few lines of the dataset read:

	age	workclass	fnlwgt	education	education-num	marital-status	occupation	relationship	race	sex	capital-gain	capital-loss	hours-per-week	native-country	i
15632	42	State-gov	83411	HS-grad	9	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	40	United-States	>
26579	21	Private	169699	HS-grad	9	Never-married	Adm-clerical	Not-in-family	White	Female	0	0	40	United-States	<
18454	51	Federal-gov	163671	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	1902	40	United-States	<
27304	31	Private	58582	Some-college	10	Never-married	Craft-repair	Not-in-family	White	Male	0	0	46	United-States	<
32293	41	Private	318046	Some-college	10	Married-civ-spouse	Transport-moving	Husband	White	Male	0	0	48	United-States	>
11802	27	Private	158647	Some-college	10	Never-married	Adm-clerical	Not-in-family	White	Female	0	0	40	United-States	<
22992	36	Private	198841	Masters	14	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	40	United-States	>
656	29	State-gov	71592	Some-college	10	Never-married	Adm-clerical	Unmarried	Asian-Pac-Islander	Female	0	0	40	Philippines	<

Hypothesis

The hypothesis that we will be testing is:

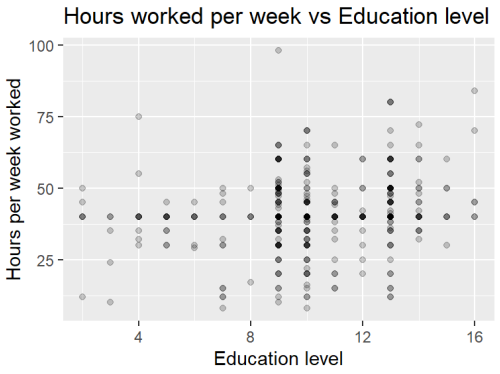
$$H_0 : \text{Hours per week worked is not linearly related to education level, } \beta_1 = 0$$
$$H_a : \text{Hours per week worked is linearly related to education level, } \beta_1 \neq 0$$

Linear regression assumptions

Before we proceed, we would like to check these assumptions:

- The variance between observations is constant (homoscedasticity)
- The model is correctly specified (linearity): ($E[Y] = \beta_0 + \beta_1 x$)

To do so, we construct a scatter plot of the variables:

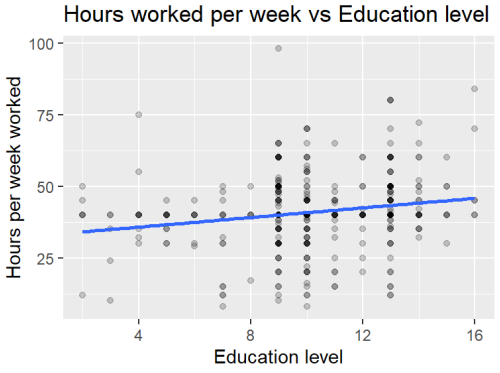


Scatter plot of our variables of interest

From inspection we see that the there does seem to be a slight positive correlation between our predictor and response variable. The variance does not seem to be constant. An example of this is that the spread of the response variable for $x = 15$ is wider than the spread for $x = 6$. However, as prof. Mouti said, this is to be expected for real life data, so we are just going to move forward and fit a linear model to the data anyways.

Fitting

```
##
## Call:
## lm(formula = `hours-per-week` ~ `education-num`, data = adult)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -32.778  -3.291   0.060   4.248  58.060
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      32.4013     2.0425  15.864  < 2e-16 ***
## `education-num`   0.8376     0.1950   4.296 2.09e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.38 on 498 degrees of freedom
## Multiple R-squared:  0.03574,    Adjusted R-squared:  0.03381
## F-statistic: 18.46 on 1 and 498 DF,  p-value: 2.088e-05
```



Data plotted with the line of best fit (blue)

Hypothesis test

From the fit, our linear model is:

$$E[y] = \beta_0 + \beta_1 x$$
$$\beta_0 = 32.4013, \quad \beta_1 = 0.8376$$

Recall that our hypothesis test is:

$$H_0 : \beta_1 = 0$$
$$H_a : \beta_1 \neq 0$$

This is a global significance test (since we only have one parameter). Reading off the F-statistic portion of the printout, we get a p-value of

2.088×10^{-5}. Thus at a level of $\alpha = 0.05$, the data suggests that there is indeed a linear relationship between education level and hours per week worked.

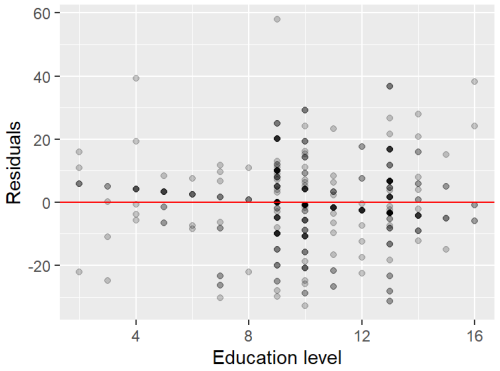
Confidence interval on β_1:

Using the function `confint` on our model, we can read of the 95 percent confidence interval for β_1:

$$\beta_1 \in (0.455, 1.221)$$

Thus we are 95% confidence the true slope, β_1 lies in the interval $(0.455, 1.221)$, under the assumption that the observations are normally distributed about the line of best fit. Note that $\beta_1 = 0$ is not in this interval which is what we expect from the theorem about test-interval duality.

Residual plot



Plot of the residuals. The y=0 was plotted in red for reference.

The data seems to be randomly clustered about the line $x = 0$, and there does not seem to be any clear patterns in the residuals. Thus a linear model would be appropriate for this data. From before, our R^2 value is 0.03574 which tells us that this regression model explains about 3.574% of the variance seen in the original data. This makes sense because looking back at our residual plot, the spread looks approximately the same (both in shape and magnitude) as the original scatter plot (so $SS_{res} \approx SS_T$ which gives a small R^2).

Individual Response

An interesting value to examine is the mean response of the number of hours worked for a bachelor's degree (education number of 13) since we are about to graduate soon.

Using the `predict` function, the 95% confidence interval for the mean number of hours worked for an individual with a bachelor's is $(41.81, 44.77)$, and the 95% prediction interval for the number of hours worked for an individual with a bachelors is $(20.89, 65.70)$. The prediction interval is bigger than the confidence interval since the confidence interval makes a statement regrading the mean while the prediction interval makes a statement regarding an individual, which means it has to take into account the random errors ε_i.

Conclusion

We found that the mean number of hours worked per week, $E[y]$, can be modeled by the education level, x, with the following model:

$$E[y] = \beta_0 + \beta_1 x$$
$$\beta_0 = 32.4013, \quad \beta_1 = 0.8376$$

Where the 95% confidence interval on β_1 was found to be $\beta_1 \in (0.455, 1.221)$. One thing that was interesting was we actually got that $\beta_1 \neq 0$ in our global significance test, as we did not think that the slight positive correlation observed in the data was significant enough to be characterized by a linear model. The results also surprised us because we thought that a lower education level correlates with more hours worked which is in direct conflict with what we found. Given the time, we would like to repeat this process with a bigger/different sample and see if we get the same results.