Teoria e Pratica dei Modelli Statistici

Homework 2

Edoardo Ricca

1. *Run a regression of average hourly earnings (AHE) on age (Age), gender (Female), and education (Bachelor). If Age increases from 25 to 26, how are earnings expected to change? If Age increases from 33 to 34, how are earnings expected to change?*

Running the regression:

$$ahe = \beta_0 + \beta_1 age + \beta_2 female + \beta_3 bachelor + \varepsilon_i$$

leads to the following estimates:

ahe	Coefficients	Std. error
Intercept	3.300273	0.872861
age	0.3144165	0.0289972
female	-2.493215	0.1656032
bachelor	5.335746	0.1643152

If *age* increases from 25 to 26, keeping every other variable constant, the expected change in earnings is going to be equal to the age coefficient β_1 which is approximately 0.31. This also applies to the increase from 33 to 34 since, being a multiple linear model, a unit increase in age will result in a change of 0.31$ in the average value of earnings per hour.

2. *Run a regression of the logarithm average hourly earnings, ln(AHE), on Age, Female, and Bachelor. If Age increases from 25 to 26, how are earnings expected to change? If Age increases from 33 to 34, how are earnings expected to change?*

Running the log-linear regression:

$$log(ahe) = \beta_0 + \beta_1 age + \beta_2 female + \beta_3 bachelor + \varepsilon_i$$

leads to the following estimates:

ln(ahe)	Coefficients	Std. error
Intercept	1.788448	0.0622994
age	0.0214183	0.0020696
female	-0.1800279	0.0118197
bachelor	0.3827709	0.0117278

In the log-linear model, the coefficient β_1 represents the relative change in the conditional mean of *ahe* given an increase of one unit in the variable *age*, keeping every other variable constant. Therefore, in both cases, earnings are expected to increase by approximately 2.14%.

3. *Run a regression of the logarithm average hourly earnings, ln(AHE), on ln(Age), Female, and Bachelor. If Age increases from 25 to 26, how are earnings expected to change? If Age increases from 33 to 34, how are earnings expected to change?*

Running the log-log regression:

$$log(ahe) = \beta_0 + \beta_1 log(age) + \beta_2 female + \beta_3 bachelor + \varepsilon_i$$

leads to the following estimates:

ln(ahe)	Coefficients	Std. error
Intercept	0.2675497	0.2065148
ln(age)	0.6369272	0.0608679
female	-0.1798992	0.0118176
bachelor	0.2675497	0.2065148

In the log-log model, the coefficient β_1 represents the elasticity of *ahe* with respect to *age*. This means that a 1% increase in *age* results in approximately β_1% increase in *ahe* (considering the passage from *age* to *ln(age)* then to *ln(ahe)* and finally *ahe*). An increase from 25 to 26 is a 4% increase in age, implying that this would result in approximately a 2.5478% increase in the conditional expectation of earnings. Instead, an increase from 33 to 34 is approximately a 3.03% increase, implying that the conditional expectation of earnings will increase by approximately 1.93%. In general, this model implies that the percentage increase in earnings declines as age increases since each unit increase in age implies a smaller percentage variation.

4. *Run a regression of the logarithm average hourly earnings, ln(AHE), on Age, Age^2, Female, and Bachelor. If Age increases from 25 to 26, how are earnings expected to change? If Age increases from 33 to 34, how are earnings expected to change?*

The polynomial model:

$$ln(ahe) = \beta_0 + \beta_1 age + \beta_2 age^2 + \beta_3 female + \beta_4 bachelor + \varepsilon_i$$

leads to the following estimates:

ln(ahe)	Coefficients	Std. error
Intercept	-0.2298404	0.7064571
age	0.1590073	0.0480171
age^2	-0.0023236	.0008101
female	-0.1793598	0.0118148
bachelor	0.3815594	0.0117282

In this polynomial model, interpreting the coefficients of the single polynomial terms of age becomes uninformative since the terms influence each other. Instead, we can easily verify that:

$$\mathbb{E}[log(ahe)|age+1,(age+1)^2,female,bachelor] - \mathbb{E}[log(ahe)|age,age^2,female,bachelor]$$

is equal to

$$\begin{aligned}&= \beta_0 + \beta_1(age+1) + \beta_2(age+1)^2 + \beta_3 female + \beta_4 bachelor - (\beta_0 + \beta_1 age + \beta_2 age^2 + \beta_3 female + \beta_4 bachelor)\\ &= \beta_1 + \beta_2[(age+1)^2 - age^2]\end{aligned}$$

therefore, an increase from 25 to 26 will lead to a percentage increase in the average earnings of approximately 4.05% and an increase from 33 to 34 will lead to an approximately 0.33% increase. This model implies, like the previous, that the older the worker, the less a unit increase will impact their salary.

5. *Compare the models specified in (2), (3) and (4). Which model do you prefer?*

All the models have similar fit since:

Variable	log-linear model	log-log model	polynomial model
age	0.02141825		0.15900727
female	-0.18002792	-0.17989917	-0.17935976
bachelor	0.38277091	0.38259616	0.38155945
ln(age)		0.63692721	
age^2			-0.00232357
intercept	1.7884485	0.26754968	-0.22984037
r^2	0.18280581	0.18313176	0.18394241

but as we know, we cannot compare models with transformed variables and a different number of variables/parameters using the R^2 coefficient. Since the log-log model and the polynomial model have the same number of parameters and the same response variable, we can conclude that the polynomial model is a slightly better fit. If instead we use some fit indexes, in particular the Akaike Information Criterion and the Bayesian Information criterion we obtain the following:

	log-linear model	log-log model	polynomial model
AIC	7226.16	7223.801	7219.933
BIC	7252.898	7250.54	7253.355

As we can see, we have close results within models with a better fit for the polynomial model according to the AIC and a better fit for the log-log model according to the BIC (which punishes the number of parameters in the polynomial model). In my opinion, both the log-log and the polynomial model capture the slightly non-linear relationship between age and earnings, which seems reasonable since we could expect a unit increase in age to impact more in term of earnings while being at the beginning of one's career compared to later in life. In both cases, it is not sensible to look at the coefficients for age since elasticity is not very useful for age and the polynomial structure renders the single coefficients not interpretable. Still, due to the AIC and the R^2 coefficient I would probably choose the polynomial model, which also minimize the squared sum of residuals (1171.97 versus approximately 1173 for the others).

6. *Plot the regression relation between Age and ln(AHE) from (2), (3), and (4) for males with a high school diploma. Describe the similarities and differences between the estimated regression functions. Would your answer change if you plotted the regression function for females with college degrees?*

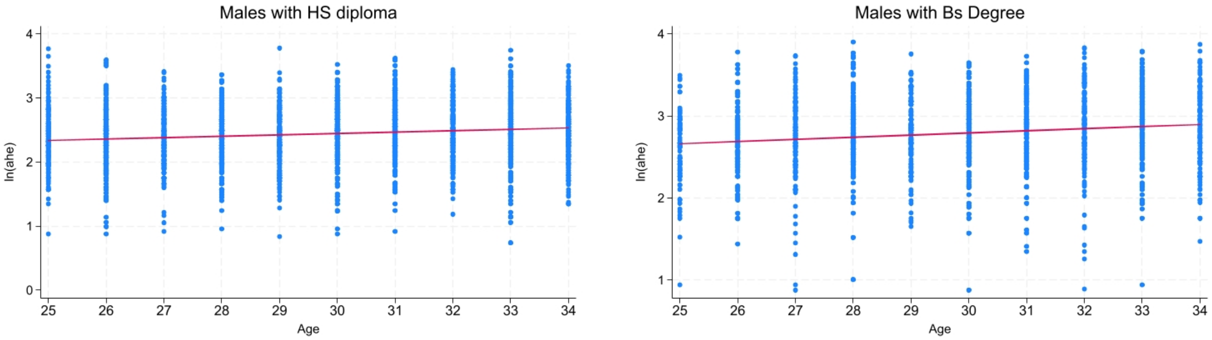

Figure 1: Log-Linear model

Starting with the log-linear model, we immediately notice that males with a bachelor's degree have a higher intercept, indicating a higher starting salary compared to males with a high school diploma. It also seems that the regression line is steeper for males with a bachelor's degree indicating that age (and possibly career progression) leads to a higher percentage increase each year compared to males with a high-school diploma.

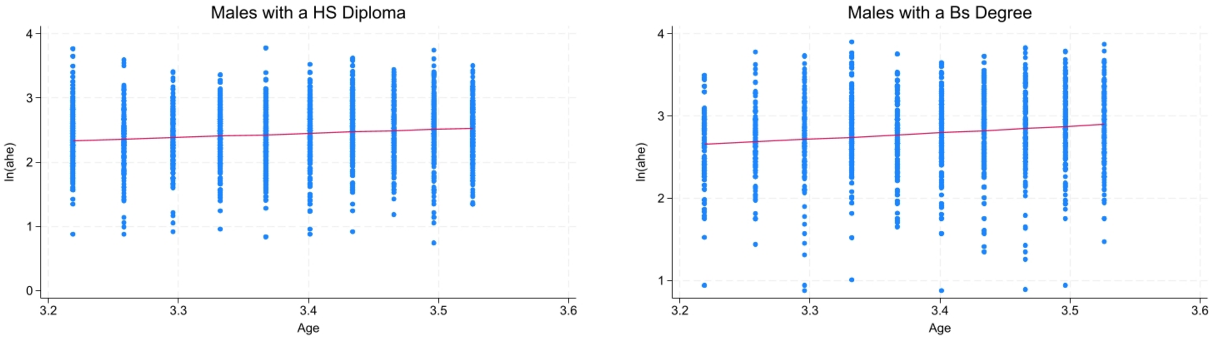

Figure 2: Log-Log model

In the Log-Log model, we also observe a higher intercept and steeper curve indicating the same results as before but percentage-wise (in terms of elasticity).

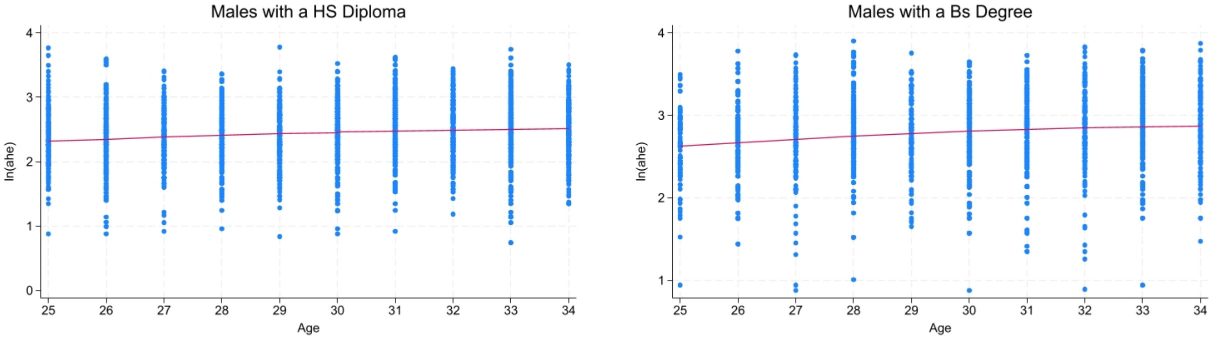

Figure 3: Polynomial model

Lastly, in the polynomial model, we also observe a higher intercept and a steeper curve. In addition, we observe a more convex curve in males with a bachelor's degree indicating that males with a bachelor's degree, while having a higher increase in earnings (in relative change terms), experience a decreasing increase (second derivative) more rapidly compared to males with a high school diploma.

By plotting the regression for females with a bachelor's degree, it would be sensible to compare them to males with a bachelor's degree.

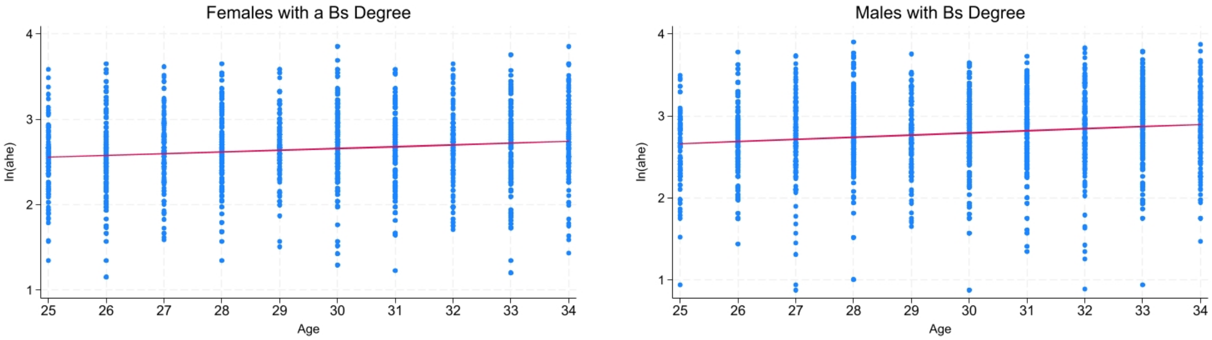

Figure 4: Log-Linear model

Starting with the log-linear model, we notice that males with a bachelor's degree have a higher intercept than females, indicating a higher starting salary compared to females with a bachelor's degree. It also seems that the regression line is steeper for males indicating that age (and, again, possibly career progression) leads to a higher percentage increase in salary each year for males compared to females. Comparing the plot of the estimated regression of males with a high school diploma to the one of females with a bachelor's degree, they seem to be almost similar indicating that females with a bachelor's degree have almost the same effect of age on salary than males with a high school diploma.

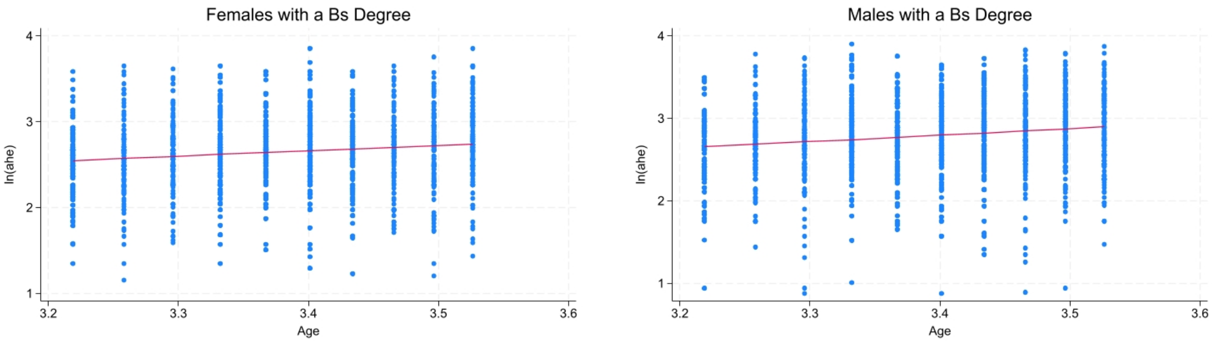

Figure 5: Log-Log model

In the Log-Log model, again, we observe the same results as before but percentage-wise (in terms of elasticity).

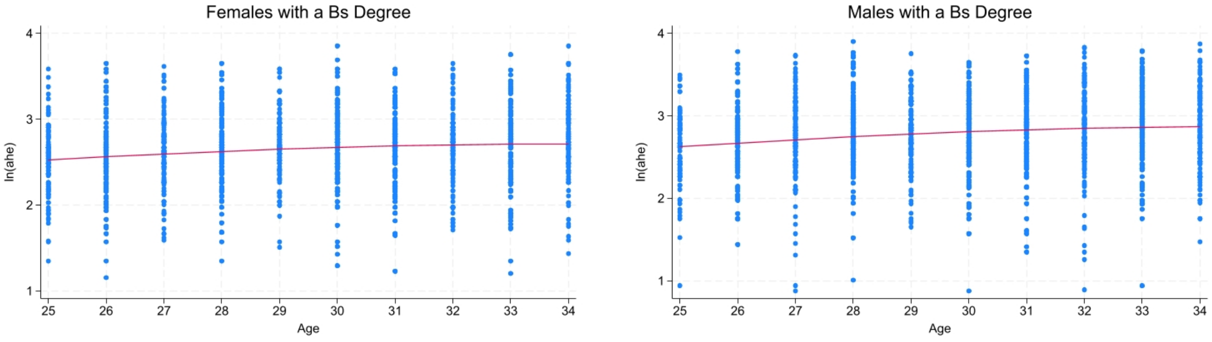

Figure 6: Polynomial model

Lastly, in the polynomial model, we also observe a higher intercept and a steeper curve. Females with a bachelor's degree seem to have almost a flat line, indicating a more steady growth of salary compared to the males.

In summary, females with a bachelor's degree exhibit lower starting salaries and slower salary growth than males with a bachelor's degree. Their salary seems to behave more similarly to males with a high school diploma. It should be considered though that bachelor is only a binary variable and it doesn't allow for a partition into the respective field of studies. This model then would not capture the distribution of males and females into the highest and lowest-paid majors. Of course, regardless of the type of major, the similarity between females with a bachelor's degree and males with a high school diploma indicates a form of possible discrimination. In addition, even if we had the distribution into the highest-paid major by sex, we would still be unable to understand if the lower presence of females in high-paid discipline is the result of a personal choice or the effect of cultural pressure (or others).

7. *Run a regression of ln(AHE), on Age, Age2, Female, Bachelor, and the interaction term FemalexBachelor. What does the coefficient on the interaction term measure? Alexis is a 30- year-old female with a bachelor's degree. What does the regression predict for her value of ln(AHE)? Jane is a 30-year-old female with a high school degree. What does the regression predict for her value of ln(AHE)? What is the predicted difference between Alexis's and Jane's earnings? Bob is a 30-year-old male with a bachelor's degree. What does the regression predict for his value of ln(AHE)? Jim is a 30-year-old male with a high school degree. What does the regression predict for his value of ln(AHE)? What is the predicted difference between Bob's and Jim's earnings?*

Running the regression:

$$log(ahe) = \beta_0 + \beta_1 age + \beta_2 age^2 + \beta_4 female + \beta_4 bachelor + \beta_5 female \cdot bachelor + \varepsilon$$

leads to the following estimates: These results tell us that being female reduces the average relative change in earnings by approximately 21.8% (keeping all the other variables constant) and that having a bachelor's degree increases it by approximately 34.6%. However, thanks to the interaction term we can use its coefficient β_5 as a measure of the additional effect of being a female and simultaneously having a bachelor's degree. The positive coefficient tells us that, in this combination, the negative effect of being a female is surpassed by the positive effect of having a bachelor's degree resulting in an additional effect of approximately 8.5% which shows us how being a female drastically reduces the additional advantages of having a bachelor's degree (while remaining positive).

log_ahe	Coefficient	Std. err.
age	0.1615669	0.0479743
age^2	-0.0023639	0.0008094
female	-0.2180115	0.0159837
bachelor	0.3455236	0.0154355
interaction	0.0849926	0.0237006
intercept	-0.2554836	0.7057851

The regression predicts the following values for each young worker:

	Prediction
Alexis	2.6765199
Jane	2.2460036
Δ_{AJ}	5.08 $
Bob	2.8095388
Jim	2.2460036
Δ_{BJ}	7.15 $

8. *Is the effect of Age on earnings different for males than females? Specify and estimate a regression that you can use to answer this question.*

If we want to observe a possible difference in the effect of age on earnings between males and females we could use the model in the following form:

$$log(ahe) = \beta_0 + \beta_1 age + \beta_2 age^2 + \beta_3 female + \beta_4 female \cdot age + \beta_5 bachelor + \varepsilon_i$$

The model leads to the following computations:

ln(ahe)	Coefficient	Std. err.
age	0.1614155	0.0480829
age^2	-0.0023356	0.0008103
female	-0.0599406	0.124949
fem_age	-0.0040212	0.0041885
bachelor	0.3812919	0.0117316
intercept	-0.290701	0.7093004

Which leads to the following estimates for the effect of of age:

$$\hat{f}_m = \beta_1 age + \beta_2 age^2$$
$$\hat{f}_f = \beta_1 age + \beta_2 age^2 + \beta_3 + \beta_4 age$$

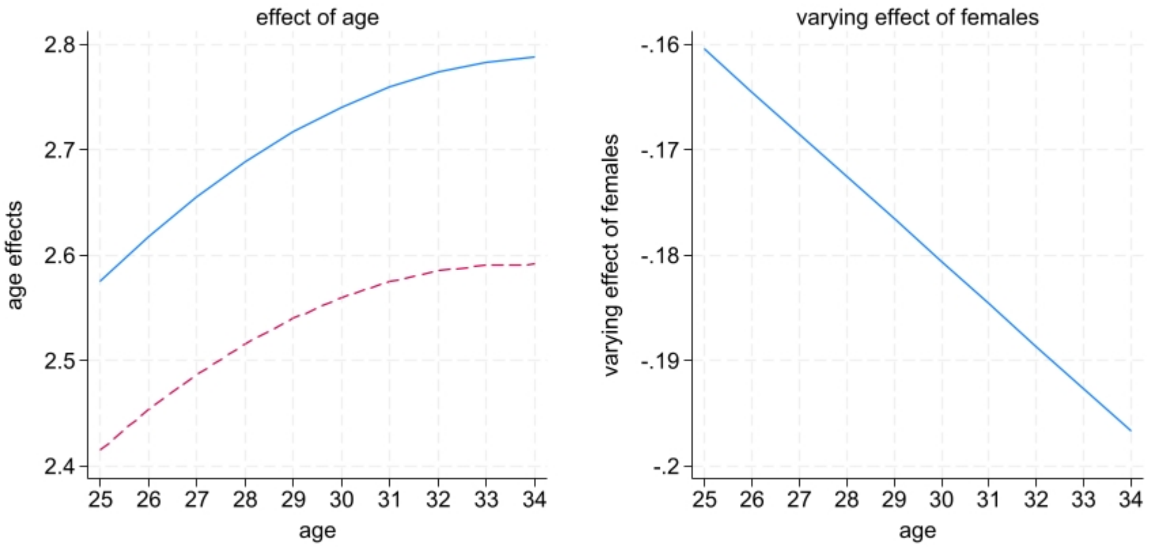

Figure 7: (left) Age effect for males (blu, solid line) and females (red, dashed line)

These plots shows different effects of age on the conditional expected value of earnings for males and females. The plot on the shows not only a significant starting difference but also a continuously increasing one. It seems that age has a higher effect on average relative change of earnings in males than in females. This difference seems also to become worse with age, according to the plot on the right.

9. *Is the effect of Age on earnings different for high school than college graduates? Specify and estimate a regression that you can use to answer this question.*

Similarly, we can estimate the effect using the following model:

$$log(ahe) = \beta_0 + \beta_1 age + \beta_2 age^2 + \beta_3 female + \beta_4 bachelor + \beta_5 bachelor \cdot age + \varepsilon_i$$

The model leads to the following computations:

ln(ahe)	Coefficient	Std. err.
age	0.1572434	0.0480846
age^2	-0.0023149	0.0008103
female	-0.1791579	0.0118188
bachelor	0.2940732	0.1248572
edu_age	0.0029452	0.0041847
intercept	-0.1851663	0.7093331

Which leads to the following estimates for the effect of of age:

$$\hat{f}_m = \beta_1 age + \beta_2 age^2$$
$$\hat{f}_f = \beta_1 age + \beta_2 age^2 + \beta_4 + \beta_5 age$$

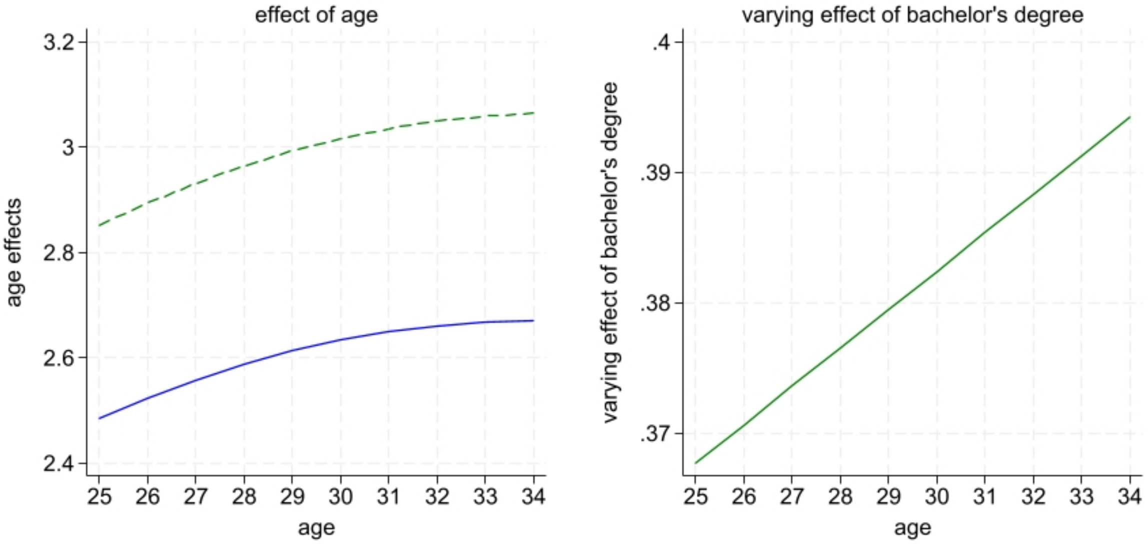

Figure 8: (left) Age effect for highschool diploma (blu, solid line) and bachelor's degree (green, dashed line)

These plots shows different effects of age on the conditional expected value of earnings depending on whether people have a high school diploma or a bachelor's degree. The plot on the left shows a significant starting difference that is not reduced with a variation in age. Instead, the plot on the left suggests that the difference becomes larger with age in favour of having a bachelor's degree.

10. . *After running all of these regressions (and any others that you want to run), summarize the effect of age on earning for young workers.*

According to our models, it seems that age positively influences the average earnings of young workers. This is in line with the expectation of career progression as employees become older. The average relative increase in salary seem also to follow a non-linear growth function with convex curve, indicating that the increases tend to become lower as age increases. The regressions also indicate a significant difference of the effect of age on females and males. Females tend to have a lower starting salary which then proceed to grow less rapidly then the males counterpart. The same difference is present when comparing people with bachelor's degree to people with a high school diploma. When looking at females with bachelor's degree, while their average salary growth is higher compared to males with a high school diploma, it is still more similar than compared to males with a bachelor's degree.