ECON2300 - Introductory Econometrics

Tutorial 6: Nonlinear Regression Functions

Tutor: Francisco Tavares Garcia

Quiz 5 is available!

Quiz 5 Now Available

Posted on: Friday, 4 April 2025 08:00:00 o'clock AEST

The following test has been made available in Quizzes: Problem Solving, Data Analysis and Short Report: ECON2300 Quiz 5 (Semester 1, 2025).

The due date is 4pm, Thursday, 10 April, 2025.

Course Link/Assessment/Quizzes: Problem Solving, Data Analysis and Short Report/ECON2300 Quiz 5 (Semester 1, 2025)

> Posted by: April Deng
> Posted to: [ECON2300] Introductory Econometrics (St Lucia). Semester 1, 2025 ECON2300_7520_20889

Quiz Opening Day Change

Posted on: Thursday, 3 April 2025 14:49:45 o'clock AEST

Dear Student,

Beginning with Quiz 5, quizzes will open on Friday at 8:00 AM and continue to be due on Thursday at 4PM. Thus, they will be open for 7 days. The change takes effect from 4th April, 2025.

The ECON2300 Team

THE UNIVERSITY OF QUEENSLAND AUSTRALIA

Project 1 is available!

Presentation of Modelling Results and Submission of Project Report

Please read this carefully

- For plots, your axes should be appropriately labelled and the plot should be titled.
- Please present estimated models in a table format, following Lecture 5, slide 30 as a template. Please use the following convention to denote statistical significance of coefficients: significant at the *5% level, or **1% level.
- Include the R code and output as an *appendix section* to the project report. The section should be labelled "Appendix" and appear at the end of the project report.
- Please submit your project report via the submission link provided in the course's Blackboard site. The submission must be a single "pdf" file. Projects submitted in any other format will receive a deduction of 5%.

- Download the files for tutorial 06 from Blackboard,
- save them into a folder for this tutorial.

Tutorial 6 [Week 7] Nonlinear Regression Functions

Attached Files:
- Lead_Mortality_Description.pdf (76.299 KB)
- lead_mortality.csv (19.463 KB)
- cps12.csv (155.882 KB)
- CPS12_Description.pdf (64.449 KB)
- tutorial6.pdf (71.147 KB)

In Week 7, tutors will aim to cover E8.1a)-b) and E8.2a)-g)

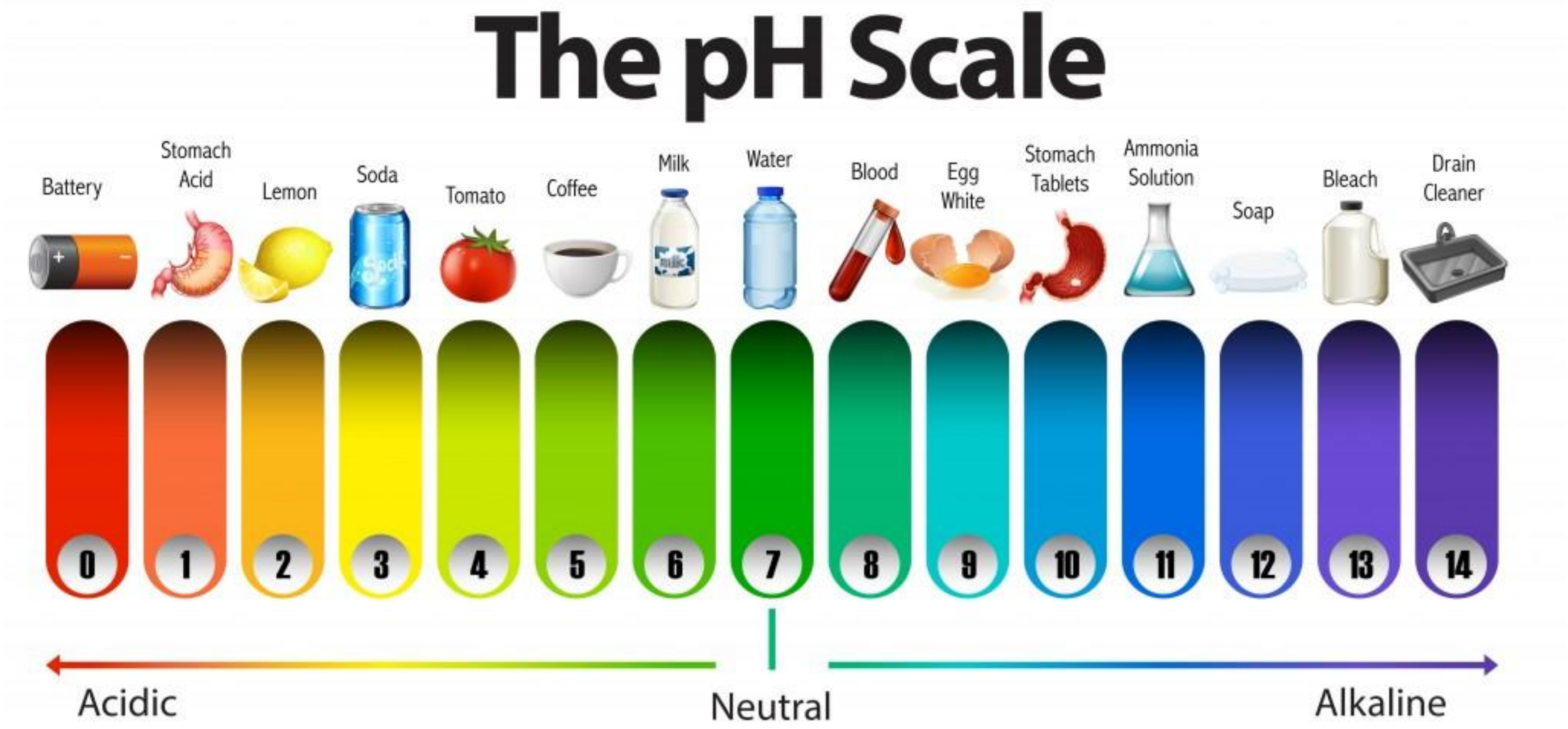
Now, let’s download the script for the tutorial.

- Copy the code from Github,
  - https://github.com/tavaresgarcia/teaching
- Save the scripts in the same folder as the data.

E8.1 Lead is toxic, particularly for young children, and for this reason government regulations severely restrict the amount of lead in our environment. But this was not always the case. In the early part of the 20th century, the underground water pipes in many U.S. cities contained lead, and lead from these pipes leached into drinking water. In this exercise you will investigate the effect of these lead water pipes on infant mortality using the dataset, `lead_mortality.csv`, which contains data on infant mortality, type of water pipes (lead or non-lead), water acidity (pH), and several demographic variables for 172 U.S cities in 1900; see also `Lead_Mortality_Description.pdf`.

Variable Name	Description
infrate	Infant mortality rate (deaths per 100 in population)
lead	Indicator =1 if city had lead pipes. (These are "lead-only" or "mixed-lead" cities.)
ph	Water pH
hardness	Water hardness index
population	City population (in 100s)
typhoid_rate	Typhoid death rate
np_tub_rate	Non-pulmonary tuberculosis death rate
mom_rate	Fraction of population who are women of child-bearing age
age	Average Age
foreign_share	Fraction of population who are foreign born
precipitation	Average precipitation in state
temperature	Average temperature in state
city	City
state	State
year	Year

E8.1 Lead is toxic, particularly for young children, and for this reason government regulations severely restrict the amount of lead in our environment. But this was not always the case. In the early part of the 20th century, the underground water pipes in many U.S. cities contained lead, and lead from these pipes leached into drinking water. In this exercise you will investigate the effect of these lead water pipes on infant mortality using the dataset, `lead_mortality.csv`, which contains data on infant mortality, type of water pipes (lead or non-lead), water acidity (pH), and several demographic variables for 172 U.S cities in 1900; see also `Lead_Mortality_Description.pdf`.



Source: ScienceABC

(a) Compute the average infant mortality rate, **infrate**, for cities with lead pipes and for cities with non-lead pipes. Is there a statistically significant difference in the average?

```
rm(list = ls())
setwd("/Users/uqdkim7/Dropbox/Teaching/R tutorials/Data")
LM <- read_csv("lead_mortality.csv") %>%
  mutate(lead_ph = lead*ph, lead_ph_65 = lead*(ph - 6.5))
attach(LM)

reg1 = lm_robust(infrate ~ lead, data = LM, se_type = "stata")
reg2 = lm_robust(infrate ~ lead + ph + lead_ph, data = LM, se_type = "stata")
reg3 = lm_robust(infrate ~ lead + ph + lead_ph_65, data = LM, se_type = "stata")

texreg(list(reg1, reg2, reg3), include.ci = F, caption.above = T,
          digits = 3, caption = "Lead and Infant Mortality",
       custom.model.names = c("(1)", "(2)", "(3)"))
```

THE UNIVERSITY OF QUEENSLAND AUSTRALIA

(a) Compute the average infant mortality rate, **infrate**, for cities with lead pipes and for cities with non-lead pipes. Is there a statistically significant difference in the average?

Table 1: Lead and Infant Mortality

	(1)	(2)	(3)
(Intercept)	0.381***	0.919***	0.919***
	(0.020)	(0.150)	(0.150)
lead	0.022	0.462*	0.092**
	(0.024)	(0.208)	(0.033)
ph		−0.075***	−0.075***
		(0.021)	(0.021)
lead_ph		−0.057*	
		(0.028)	
lead_ph_65			−0.057*
			(0.028)
R^2	0.005	0.272	0.272
Adj. R^2	-0.001	0.259	0.259
Num. obs.	172	172	172
RMSE	0.151	0.130	0.130

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Column (1) of Table 1 shows that the sample mean of **infrate** for cities with non-lead pipes and cities with lead pipes are 0.381 and 0.403, respectively. The difference in the sample means is 0.022 with a standard error of 0.024. The estimate implies that cities with lead pipes have a higher infant mortality rate (by 0.022 deaths per 100 people in the population), but the standard error is comparatively large (0.024) and so the difference is not statistically significant (t-statistic= 0.91).

THE UNIVERSITY OF QUEENSLAND AUSTRALIA

(b) The amount of lead leached from lead pipes depends on the chemistry of the water running through the pipes. The more acidic the water (that is, the lower its pH), the more lead is leached. Run a regression of `infrate` on `lead`, `ph`, and the interaction term `lead` $\times$ `ph`.

i. The regression includes four coefficients (the intercept and the three coefficients multiplying the regressors). Explain what each coefficient measures.

Table 1: Lead and Infant Mortality

	(1)	(2)	(3)
(Intercept)	0.381^{***}	0.919^{***}	0.919^{***}
	(0.020)	(0.150)	(0.150)
lead	0.022	0.462^{*}	0.092^{**}
	(0.024)	(0.208)	(0.033)
ph		-0.075^{***}	-0.075^{***}
		(0.021)	(0.021)
lead_ph		-0.057^{*}	
		(0.028)	
lead_ph_65			-0.057^{*}
			(0.028)
R^2	0.005	0.272	0.272
Adj. R^2	-0.001	0.259	0.259
Num. obs.	172	172	172
RMSE	0.151	0.130	0.130

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

THE UNIVERSITY OF QUEENSLAND AUSTRALIA

i. The regression includes four coefficients (the intercept and the three coefficients multiplying the regressors). Explain what each coefficient measures.

i The first coefficient is the intercept, which shows the level of $\texttt{infrate}$ when $\texttt{lead} = 0$ and $\texttt{ph} = 0$. It dictates the level of the regression line. The second and fourth coefficients measure the effect of lead on the infant mortality rate. Comparing 2 cities, one with lead pipes ($\texttt{lead} = 1$) and one without lead pipes ($\texttt{lead} = 0$), but the same of $\texttt{ph}$, the difference in predicted infant mortality rate is

$$\widehat{\texttt{infrate}}|_{\texttt{lead}=1} - \widehat{\texttt{infrate}}|_{\texttt{lead}=0} = 0.462 - 0.057 \times \texttt{ph}$$

The third and fourth coefficients measure the effect of $\texttt{ph}$ on the infant mortality rate. Comparing 2 cities, one with a $\texttt{ph} = 6$ and the other with $\texttt{ph} = 5$, but the same of $\texttt{lead}$, the difference in predicted infant mortality rate is

$$\widehat{\texttt{infrate}}|_{\texttt{ph}=6} - \widehat{\texttt{infrate}}|_{\texttt{ph}=5} = -0.075 - 0.057 \times \texttt{lead}$$

so the difference is -0.075 for cities without lead pipes and -0.132 for cities with lead pipes.

ii. Plot the estimated regression function relating `infrate` to `ph` for `lead` = 0 and for `lead` = 1. Describe the differences in the regression functions and relate these differences to the coefficients you discussed in (i).

```
fig8.1 <- ggplot(LM, aes(x = ph, y = infrate, col = as.factor(lead))) +
  labs(title = "Figure 1: Infant Mortality Rate and pH Value",
       x = "PH Value", y = "Infant Mortality Rate") +
  geom_smooth(data = LM, method = "lm", se = FALSE, size = 1) +
  theme(axis.title = element_text(family = "serif"),
        plot.title = element_text(hjust = 0.5, size = 12, family = "serif",
                                  face = "bold"),
        legend.position = c(0.9, 0.8),
        legend.title = element_blank(),
        legend.text = element_text(family = "serif", face = "bold"),
        legend.key = element_rect(color = "transparent"),
        legend.background = element_rect(fill = "lightgrey",
                                         size = 0.8,
                                         linetype="solid")) +
  scale_color_discrete(name = "Lead", labels = c(" lead = 0", " lead = 1"))
print(fig8.1)
```

ii. Plot the estimated regression function relating `infrate` to `ph` for `lead` = 0 and for `lead` = 1. Describe the differences in the regression functions and relate these differences to the coefficients you discussed in (i).

ii (See Figure 1 above) The infant mortality rate is higher for cities with lead pipes, but the difference declines as the pH level increases. For example:

The 10th percentile of pH is 6.4. At this level, the difference in infant mortality rates is

$$\widehat{\texttt{infrate}}|_{\texttt{lead=1,ph=6.4}} - \widehat{\texttt{infrate}}|_{\texttt{lead=0,ph=6.4}} = 0.462 - 0.057 \times 6.4 = 0.097$$

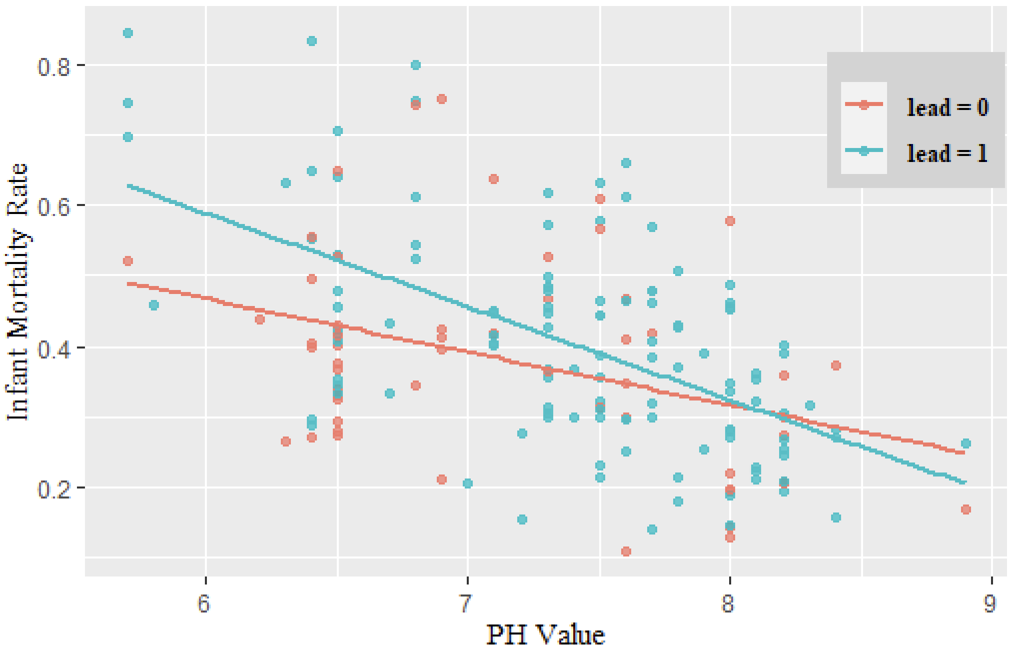
The 50th percentile of pH is 7.5. At this level, the difference in infant mortality rates is

$$\widehat{\texttt{infrate}}|_{\texttt{lead=1,ph=7.5}} - \widehat{\texttt{infrate}}|_{\texttt{lead=0,ph=7.5}} = 0.462 - 0.057 \times 7.5 = 0.035$$

The 90th percentile of pH is 8.2. At this level, the difference in infant mortality rates is

$$\widehat{\texttt{infrate}}|_{\texttt{lead=1,ph=8.2}} - \widehat{\texttt{infrate}}|_{\texttt{lead=0,ph=8.2}} = 0.462 - 0.057 \times 8.2 = -0.01$$

Figure 1: Infant Mortality Rate and pH Value



iii. Does `lead` have a statistically significant effect on infant mortality? Explain.

```
> linearHypothesis(reg2, c("lead=0", "lead_ph = 0"), test=c("F"))
Linear hypothesis test

Hypothesis:
lead = 0
lead_ph = 0

Model 1: restricted model
Model 2: infrate ~ lead + ph + lead_ph

  Res.Df Df     F  Pr(>F)
1    170
2    168  2 3.936 0.02135 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

iii The F-statistic for the coefficient on `lead` and the interaction term is 3.94, which has a p-value of 0.021, so the coefficients are jointly statistically significantly different from zero at the 5% but not the 1% significance level.

iv. Does the effect of `lead` on `infrate` depend on `ph`? Is this dependence statistically significant?

```
> summary(reg2)

Call:
lm_robust(formula = infrate ~ lead + ph + lead_ph, se_type = "stata")

Standard error type:  HC1

Coefficients:
            Estimate Std. Error t value  Pr(>|t|) CI Lower CI Upper  DF
(Intercept)  0.91890    0.15049   6.106 6.866e-09  0.62180  1.21601 168
lead         0.46180    0.20761   2.224 2.746e-02  0.05193  0.87167 168
ph          -0.07518    0.02095  -3.588 4.369e-04 -0.11654 -0.03381 168
lead_ph     -0.05686    0.02808  -2.025 4.448e-02 -0.11230 -0.00142 168

Multiple R-squared:  0.2719 ,   Adjusted R-squared:  0.2589
F-statistic: 20.97 on 3 and 168 DF,  p-value: 1.366e-11
```

iv The interaction term has a t-statistic of -2.02, so the coefficient is significant at the 5% but not the 1% significance level.

v. What is the average value of `ph` in the sample? At this pH level, what is the estimated effect of `lead` on infant mortality? What is the standard deviation of pH? Suppose that the pH level is one standard deviation lower than the average level of pH in the sample: what is the estimated effect of `lead` on infant mortality? What if pH level is one standard deviation higher than the average value?

v The mean of pH is 7.32. At this level, the difference in infant mortality rates is

$$\widehat{\texttt{infrate}}|_{\texttt{lead}=1,\texttt{ph=7.32}} - \widehat{\texttt{infrate}}|_{\texttt{lead}=0,\texttt{ph=7.32}} = 0.462 - 0.057 \times 7.32 = 0.045$$

The standard deviation of pH is 0.69 (`sd(ph)`), so that the mean plus 1 standard deviation is 8.01 and the mean minus 1 standard deviation is 6.63. The infant mortality rates at the pH levels are:

$$\widehat{\texttt{infrate}}|_{\texttt{lead}=1,\texttt{ph=8.01}} - \widehat{\texttt{infrate}}|_{\texttt{lead}=0,\texttt{ph=8.01}} = 0.462 - 0.057 \times 8.01 = 0.005$$

$$\widehat{\texttt{infrate}}|_{\texttt{lead}=1,\texttt{ph=6.63}} - \widehat{\texttt{infrate}}|_{\texttt{lead}=0,\texttt{ph=6.63}} = 0.462 - 0.057 \times 6.63 = 0.084$$

vi. Construct a 95% confidence interval for the effect of **lead** on infant mortality when pH = 6.5.

```
> confint(glht(reg2, linfct = c("lead + 6.5*lead_ph=0")))

	 Simultaneous Confidence Intervals

Fit: lm_robust(formula = infrate ~ lead + ph + lead_ph, se_type = "stata")

Quantile = 1.96
95% family-wise confidence level


Linear Hypotheses:
                          Estimate lwr     upr    
lead + 6.5 * lead_ph == 0 0.09219  0.02779 0.15660
```

vi Write the regression as

$$\texttt{infrate} = \beta_0 + \beta_1 \texttt{lead} + \beta_2 \texttt{ph} + \beta_3 \texttt{lead} \times \texttt{ph} + u$$

so the effect of **lead** on **infrate** is $\beta_1 + \beta_3 \times \texttt{ph}$. Thus, we want to construct a 95% confidence interval (CI) for $\beta_1 + 6.5\beta_3$. This CI can be easily computed using the functions **confint** and **glht**. The resulting CI is $[0.028, 0.157]$. Equivalently, we can also use method 2 of Section 7.3, add and subtract $6.5\beta_3$ lead to the regression to obtain:

$$\begin{aligned}\texttt{infrate} &= \beta_0 + (\beta_1 + 6.5\beta_3) \times \texttt{lead} + \beta_2 \times \texttt{ph} + \beta_3 \times (\texttt{lead} \times \texttt{ph} - 6.5 \times \texttt{lead}) + u \\ &= \beta_0 + \gamma \times \texttt{lead} + \beta_2 \times \texttt{ph} + \beta_3 \texttt{lead} \times (\texttt{ph} - 6.5) + u\end{aligned}$$

So, the 95% CI for $\beta_1 + 6.5\beta_3$ is equal to the 95% CI for γ, which can be constructed using results presented in column (3) of Table 1.

(c) The analysis in (b) may suffer from omitted variable bias because it neglects factors that affect infant mortality and that might potentially be correlated with `lead` and `ph`. Investigate this concern, using the other variables in the dataset. [self-study, no solution]

E8.2 In this exercise, you will investigate the relationship between a worker's age and earnings using the sample `cps12.csv`, which contains data for full-time, full-year workers, ages 25–34, with a high school diploma or B.A./B.S. as their highest degree; see also `CPS12_Description.pdf`.

	A	B	C	D	E
1	year	ahe	bachelor	female	age
2	2012	19.23077	0	0	30
3	2012	17.54808	0	0	29
4	2012	8.547009	0	0	27
5	2012	16.82692	0	1	25
6	2012	16.34615	1	1	27
7	2012	16.10577	1	0	30
8	2012	15.81197	0	0	31
9	2012	3.525641	1	0	29
10	2012	14.42308	0	0	29

Series in Data Set:

FEMALE: 1 if female; 0 if male

YEAR: Year

AHE : Average Hourly Earnings

BACHELOR: 1 if worker has a bachelor's degree;
0 if worker has a high school degree

E8.2 In this exercise, you will investigate the relationship between a worker's age and earnings using the sample `cps12.csv`, which contains data for full-time, full-year workers, ages 25–34, with a high school diploma or B.A./B.S. as their highest degree; see also `CPS12_Description.pdf`.

```
rm(list = ls())
setwd("/Users/uqdkim7/Dropbox/Teaching/R tutorials/Data")
CPS12 <- read_csv("cps12.csv") %>%
  mutate(ln_ahe = log(ahe),
         ln_age = log(age),
         age2 = age*age,
         fem_bac = female*bachelor,
         fem_age = female*age,
         fem_age2= female*age2,
         bac_age = bachelor*age,
         bac_age2= bachelor*age2)
attach(CPS12)
```

(a) 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?

```
reg1 = lm_robust(ahe ~ age + female + bachelor, data = CPS12, se_type = "stata")
reg2 = lm_robust(ln_ahe ~ age + female + bachelor, data = CPS12, se_type = "stata")
reg3 = lm_robust(ln_ahe ~ ln_age + female + bachelor, data = CPS12, se_type = "stata")
reg4 = lm_robust(ln_ahe ~ age + age2 + female + bachelor,
                 data = CPS12, se_type = "stata")
reg5 = lm_robust(ln_ahe ~ age + age2 + female + bachelor + fem_bac,
                 data = CPS12, se_type = "stata")
reg6 = lm_robust(ln_ahe ~ age + age2 + fem_age + fem_age2 + female +
                  bachelor + fem_bac, data = CPS12, se_type = "stata")
reg7 = lm_robust(ln_ahe ~ age + age2 + bac_age + bac_age2 + female +
                  bachelor + fem_bac, data = CPS12, se_type = "stata")

reg8 = lm_robust(ln_ahe ~ age + age2 + fem_age + fem_age2
                 + bac_age + bac_age2 + female + bachelor + fem_bac,
                 data = CPS12, se_type = "stata")

texreg(list(reg1, reg2, reg3, reg4, reg5, reg6, reg7, reg8),
       include.ci = F, caption.above = T,
      digits = 3, caption = "Earnings and Age, 2012",
       custom.model.names = c("(1)", "(2)", "(3)", "(4)", "(5)", "(6)", "(7)", "(8)"))
```

(a) 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?

Table 2: Earnings and Age, 2012

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
(Intercept)	1.866	1.941***	0.150	0.792	0.804	1.987*	1.810	2.715*
	(1.175)	(0.059)	(0.195)	(0.671)	(0.671)	(0.883)	(0.949)	(1.069)
age	0.510***	0.026***		0.104*	0.104*	0.020	0.037	−0.027
	(0.040)	(0.002)		(0.046)	(0.046)	(0.060)	(0.065)	(0.073)
female	−3.810***	−0.192***	−0.192***	−0.192***	−0.242***	−2.949*	−0.242***	−2.672
	(0.224)	(0.011)	(0.011)	(0.011)	(0.017)	(1.356)	(0.017)	(1.367)
bachelor	8.319***	0.438***	0.438***	0.437***	0.400***	0.401***	−1.529	−1.223
	(0.224)	(0.011)	(0.011)	(0.011)	(0.015)	(0.015)	(1.340)	(1.351)
ln_age			0.753***					
			(0.058)					
age2				−0.001	−0.001	0.000	−0.000	0.001
				(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
fem_bac					0.090***	0.089***	0.090***	0.089***
					(0.023)	(0.023)	(0.023)	(0.023)
fem_age						0.193*		0.174
						(0.092)		(0.093)
fem_age2						−0.003*		−0.003*
						(0.002)		(0.002)
bac_age							0.128	0.106
							(0.091)	(0.092)
bac_age2							−0.002	−0.002
							(0.002)	(0.002)
R^2	0.180	0.196	0.197	0.197	0.198	0.199	0.199	0.200
Adj. R^2	0.180	0.196	0.196	0.196	0.198	0.199	0.198	0.199
Statistic	539.537	623.312	624.311	469.238	382.921	275.770	273.657	214.586
Num. obs	7440	7440	7440	7440	7440	7440	7440	7440
RMSE	9.678	0.478	0.478	0.478	0.478	0.477	0.478	0.478

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

(a) 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?

	(1)
(Intercept)	1.866
	(1.175)
age	0.510***
	(0.040)
female	−3.810***
	(0.224)
bachelor	8.319***
	(0.224)

Table 2: Earnings and Age, 2012

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
(Intercept)	1.866	1.941***	0.150	0.792	0.804	1.987*	1.810	2.715*
	(1.175)	(0.059)	(0.195)	(0.671)	(0.671)	(0.883)	(0.949)	(1.069)
age	0.510***	0.026***		0.104*	0.104*	0.020	0.037	−0.027
	(0.040)	(0.002)		(0.046)	(0.046)	(0.060)	(0.065)	(0.073)
female	−3.810***	−0.192***	−0.192***	−0.192***	−0.242***	−2.949*	−0.242***	−2.672
	(0.224)	(0.011)	(0.011)	(0.011)	(0.017)	(1.356)	(0.017)	(1.367)
bachelor	8.319***	0.438***	0.438***	0.437***	0.400***	0.401***	−1.529	−1.223
	(0.224)	(0.011)	(0.011)	(0.011)	(0.015)	(0.015)	(1.340)	(1.351)
ln_age			0.753***					
			(0.058)					
age2				−0.001	−0.001	0.000	−0.000	0.001
				(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
fem_bac					0.090***	0.089***	0.090***	0.089***
					(0.023)	(0.023)	(0.023)	(0.023)
fem_age						0.193*		0.174
						(0.092)		(0.093)
fem_age2						−0.003*		−0.003*
						(0.002)		(0.002)
bac_age							0.128	0.106
							(0.091)	(0.092)
bac_age2							−0.002	−0.002
							(0.002)	(0.002)
R^2	0.180	0.196	0.197	0.197	0.198	0.199	0.199	0.200
Adj. R^2	0.180	0.196	0.196	0.196	0.198	0.199	0.198	0.199
Statistic	539.537	623.312	624.311	469.238	382.921	275.770	273.657	214.586
Num. obs	7440	7440	7440	7440	7440	7440	7440	7440
RMSE	9.678	0.478	0.478	0.478	0.478	0.477	0.478	0.478

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

The regression results for this question are shown in column (1) of Table 2. If `age` increases from 25 to 26, earnings are predicted to increase by \$0.510 per hour. If `age` increases from 33 to 34, earnings are predicted to increase by \$0.510 per hour. These values are the same because the regression is a linear function relating `ahe` and `age`.

(b) Run a regression of the logarithm of 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 34 to 35, how are earnings expected to change?

Case II: Log-linear population regression function

- Compute Y "before" and "after" changing X:

$$\ln(Y) = \beta_0 + \beta_1 X \qquad \text{("before")}$$

- New change X:

$$\ln(Y + \Delta Y) = \beta_0 + \beta_1(X + \Delta X) \qquad \text{("after")}$$

- Subtract ("after") - ("before"):

$$\underbrace{\ln(Y + \Delta Y) - \ln(Y)}_{\approx \Delta Y / Y} = \beta_1 \Delta X \Longrightarrow \frac{\Delta Y}{Y} \approx \beta_1 \Delta X$$

- If X changes by one unit, i.e. $\Delta X = 1$, then $\frac{\Delta Y}{Y}$ changes by β_1.
- A one-unit change in X is associated with a $\beta_1 \times 100\%$ change in Y

23 / 40

(b) Run a regression of the logarithm of 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 34 to 35, how are earnings expected to change?

T

	(1)	(2)
(Intercept)	1.866	1.941***
	(1.175)	(0.059)
age	0.510***	0.026***
	(0.040)	(0.002)
female	−3.810***	−0.192***
	(0.224)	(0.011)
bachelor	8.319***	0.438***
	(0.224)	(0.011)

Table 2: Earnings and Age, 2012

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
(Intercept)	1.866	1.941***	0.150	0.792	0.804	1.987*	1.810	2.715*
	(1.175)	(0.059)	(0.195)	(0.671)	(0.671)	(0.883)	(0.949)	(1.069)
age	0.510***	0.026***		0.104*	0.104*	0.020	0.037	−0.027
	(0.040)	(0.002)		(0.046)	(0.046)	(0.060)	(0.065)	(0.073)
female	−3.810***	−0.192***	−0.192***	−0.192***	−0.242***	−2.949*	−0.242***	−2.672
	(0.224)	(0.011)	(0.011)	(0.011)	(0.017)	(1.356)	(0.017)	(1.367)
bachelor	8.319***	0.438***	0.438***	0.437***	0.400***	0.401***	−1.529	−1.223
	(0.224)	(0.011)	(0.011)	(0.011)	(0.015)	(0.015)	(1.340)	(1.351)
ln_age			0.753***					
			(0.058)					
age2				−0.001	−0.001	0.000	−0.000	0.001
				(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
fem_bac					0.090***	0.089***	0.090***	0.089***
					(0.023)	(0.023)	(0.023)	(0.023)
fem_age						0.193*		0.174
						(0.092)		(0.093)
fem_age2						−0.003*		−0.003*
						(0.002)		(0.002)
bac_age							0.128	0.106
							(0.091)	(0.092)
bac_age2							−0.002	−0.002
							(0.002)	(0.002)
R^2	0.180	0.196	0.197	0.197	0.198	0.199	0.199	0.200
Adj. R^2	0.180	0.196	0.196	0.196	0.198	0.199	0.198	0.199
Statistic	539.537	623.312	624.311	469.238	382.921	275.770	273.657	214.586
Num. obs	7440	7440	7440	7440	7440	7440	7440	7440
RMSE	9.678	0.478	0.478	0.478	0.478	0.477	0.478	0.478

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

The regression results for this question are shown in column(2) of Table 2. If **age** increases from 25 to 26, ln(**ahe**) is predicted to increase by 0.026, so earnings are predicted to increase by 2.6%. If **age** increases from 34 to 35, ln(**ahe**) is predicted to increase by 0.026, earnings are predicted to increase by 2.6%. These values, in percentage terms, are the same because the regression is a linear function relating ln(**ahe**) and **age**.

(c) Run a regression of the logarithm of 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 34 to 35, how are earnings expected to change?

Case III: Log-log population regression function

- Compute Y "before" and "after" changing X:

$$\ln(Y) = \beta_0 + \beta_1 \ln(X) \qquad (\text{"before"})$$

- New change X:

$$\ln(Y + \Delta Y) = \beta_0 + \beta_1 \ln(X + \Delta X) \qquad (\text{"after"})$$

- Subtract ("after") - ("before"):

$$\underbrace{\ln(Y + \Delta Y) - \ln(Y)}_{\approx \Delta Y / Y} = \beta_1 \underbrace{[\ln(X + \Delta X) - \ln(X)]}_{\approx \Delta X / X}$$

- So,

$$\beta_1 \approx \frac{\Delta Y / Y}{\Delta X / X} = \frac{\text{percentage change in } Y}{\text{percentage change in } X} = \text{Elasticity of } Y \text{ to } X$$

(c) Run a regression of the logarithm of 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 34 to 35, how are earnings expected to change?

Table 2: Earn

	(1)	(2)	(3)
(Intercept)	1.866	1.941***	0.150
	(1.175)	(0.059)	(0.195)
age	0.510***	0.026***	
	(0.040)	(0.002)	
female	−3.810***	−0.192***	−0.192***
	(0.224)	(0.011)	(0.011)
bachelor	8.319***	0.438***	0.438***
	(0.224)	(0.011)	(0.011)
ln_age			0.753***
			(0.058)

Table 2: Earnings and Age, 2012

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
(Intercept)	1.866	1.941***	0.150	0.792	0.804	1.987*	1.810	2.715*
	(1.175)	(0.059)	(0.195)	(0.671)	(0.671)	(0.883)	(0.949)	(1.069)
age	0.510***	0.026***		0.104*	0.104*	0.020	0.037	−0.027
	(0.040)	(0.002)		(0.046)	(0.046)	(0.060)	(0.065)	(0.073)
female	−3.810***	−0.192***	−0.192***	−0.192***	−0.242***	−2.949*	−0.242***	−2.672
	(0.224)	(0.011)	(0.011)	(0.011)	(0.017)	(1.356)	(0.017)	(1.367)
bachelor	8.319***	0.438***	0.438***	0.437***	0.400***	0.401***	−1.529	−1.223
	(0.224)	(0.011)	(0.011)	(0.011)	(0.015)	(0.015)	(1.340)	(1.351)
ln_age			0.753***					
			(0.058)					
age2				−0.001	−0.001	0.000	−0.000	0.001
				(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
fem_bac					0.090***	0.089***	0.090***	0.089***
					(0.023)	(0.023)	(0.023)	(0.023)
fem_age						0.193*		0.174
						(0.092)		(0.093)
fem_age2						−0.003*		−0.003*
						(0.002)		(0.002)
bac_age							0.128	0.106
							(0.091)	(0.092)
bac_age2							−0.002	−0.002
							(0.002)	(0.002)
R^2	0.180	0.196	0.197	0.197	0.198	0.199	0.199	0.200
Adj. R^2	0.180	0.196	0.196	0.196	0.198	0.199	0.198	0.199
Statistic	539.537	623.312	624.311	469.238	382.921	275.770	273.657	214.586
Num. obs	7440	7440	7440	7440	7440	7440	7440	7440
RMSE	9.678	0.478	0.478	0.478	0.478	0.477	0.478	0.478

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

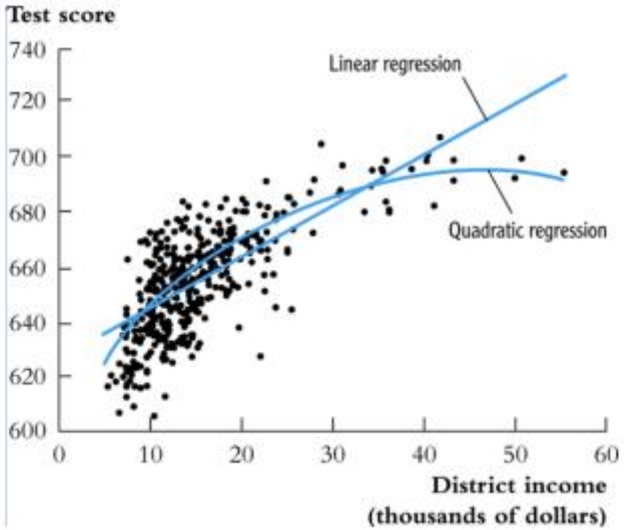
The regression results for this question are shown in column (3) of Table 2. If **age** increases from 25 to 26, then ln(**age**) has increased by $\ln(26) - \ln(25) = 0.0392$ (or 3.92%). The predicted increase in ln(**ahe**) is $0.75 \times 0.0392 = 0.029$. This means that earnings are predicted to increase by 2.9%. If **age** increases from 34 to 35, then ln(**age**) has increased by $\ln(35) - \ln(34) = 0.0290$ (or 2.90%). The predicted increase in ln(**ahe**) is $0.75 \times 0.0290 = 0.021$. This means that earnings are predicted to increase by 2.1%.

(d) Run a regression of the logarithm of 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 34 to 35, how are earnings expected to change?

Interpreting the estimated regression function:

(a) Plot the predicted values;

$$\underset{}{TestScore_i} = \underset{(2.9)}{607.3} + \underset{(0.27)}{3.85 Income_i} - \underset{(0.0048)}{0.0423 (Income_i)^2}$$



12 / 40

(d) Run a regression of the logarithm of 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 34 to 35, how are earnings expected to change?

Table 2: Earnings and Age

	(1)	(2)	(3)	(4)
(Intercept)	1.866	1.941***	0.150	0.792
	(1.175)	(0.059)	(0.195)	(0.671)
age	0.510***	0.026***		0.104*
	(0.040)	(0.002)		(0.046)
female	−3.810***	−0.192***	−0.192***	−0.192***
	(0.224)	(0.011)	(0.011)	(0.011)
bachelor	8.319***	0.438***	0.438***	0.437***
	(0.224)	(0.011)	(0.011)	(0.011)
ln_age			0.753***	
			(0.058)	
age2				−0.001
				(0.001)

The regression results for this question are shown in column (4) of Table 2. When age increases from 25 to 26, the predicted change in ln(ahe) is

$$(0.104 \times 26 - 0.0013 \times 26^2) - (0.104 \times 25 - 0.0013 \times 25^2) = 0.036.$$

This means that earnings are predicted to increase by 3.6%. When age increases from 34 to 35, the predicted change in ln(ahe) is

$$(0.104 \times 35 - 0.0013 \times 35^2) - (0.104 \times 34 - 0.0013 \times 34^2) = 0.012.$$

This means that earnings are predicted to increase by 1.2%.

(e) Do you prefer the regression in (c) to the regression in (b)? Explain.

The regressions differ in their choice of one of the regressors. The two sets of estimates are overall very similar. They can be compared on the basis of the $\overline{R}^2$. The regression in (c) has a (marginally) higher $\overline{R}^2$ (0.1962 vs. 0.1961), so it is preferred.

(f) Do you prefer the regression in (d) to the regression in (b)? Explain.

The regression in (d) adds the variable $\texttt{age}^2$ to regression (b). The coefficient on $\texttt{age}^2$ is not statistically significant at the 5% level and the estimated coefficient is very close to zero. This suggests that the regression in (b) is preferred to that in (d). However, the regressions are so similar that either may be used.

(g) Do you prefer the regression in (d) to the regression in (c)? Explain.

The regressions differ in their choice of the regressors ($\ln(\texttt{age})$ in (c) and $\texttt{age}$ and $\texttt{age}^2$ in (d)). They can be compared on the basis of the $\overline{R}^2$. The regression in (d) has a (marginally) higher $\overline{R}^2$ (0.1963 vs. 0.1962), so it is preferred.

(h) Plot the regression relation between age and ln(ahe) from (b), (c), and (d) 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?

```
age <- seq(25, 34, by = 1)
ln_ageb <- 1.941 + 0.0255*age
ln_agec <- 0.150 + 0.753*log(age)
ln_aged <- 0.792 + 0.104*age - 0.00133*age^2

datab <- data.frame(ln_age = ln_ageb, age = age, data = "b")
datac <- data.frame(ln_age = ln_agec, age = age, data = "c")
datad <- data.frame(ln_age = ln_aged, age = age, data = "d")
data.bcd <- rbind.data.frame(datab, datac, datad)
```

Table 2: Earnings and Age

	(1)	(2)	(3)	(4)
(Intercept)	1.866	1.941***	0.150	0.792
	(1.175)	(0.059)	(0.195)	(0.671)
age	0.510***	0.026***		0.104*
	(0.040)	(0.002)		(0.046)
female	−3.810***	−0.192***	−0.192***	−0.192***
	(0.224)	(0.011)	(0.011)	(0.011)
bachelor	8.319***	0.438***	0.438***	0.437***
	(0.224)	(0.011)	(0.011)	(0.011)
ln_age			0.753***	
			(0.058)	
age2				−0.001
				(0.001)

THE UNIVERSITY OF QUEENSLAND AUSTRALIA

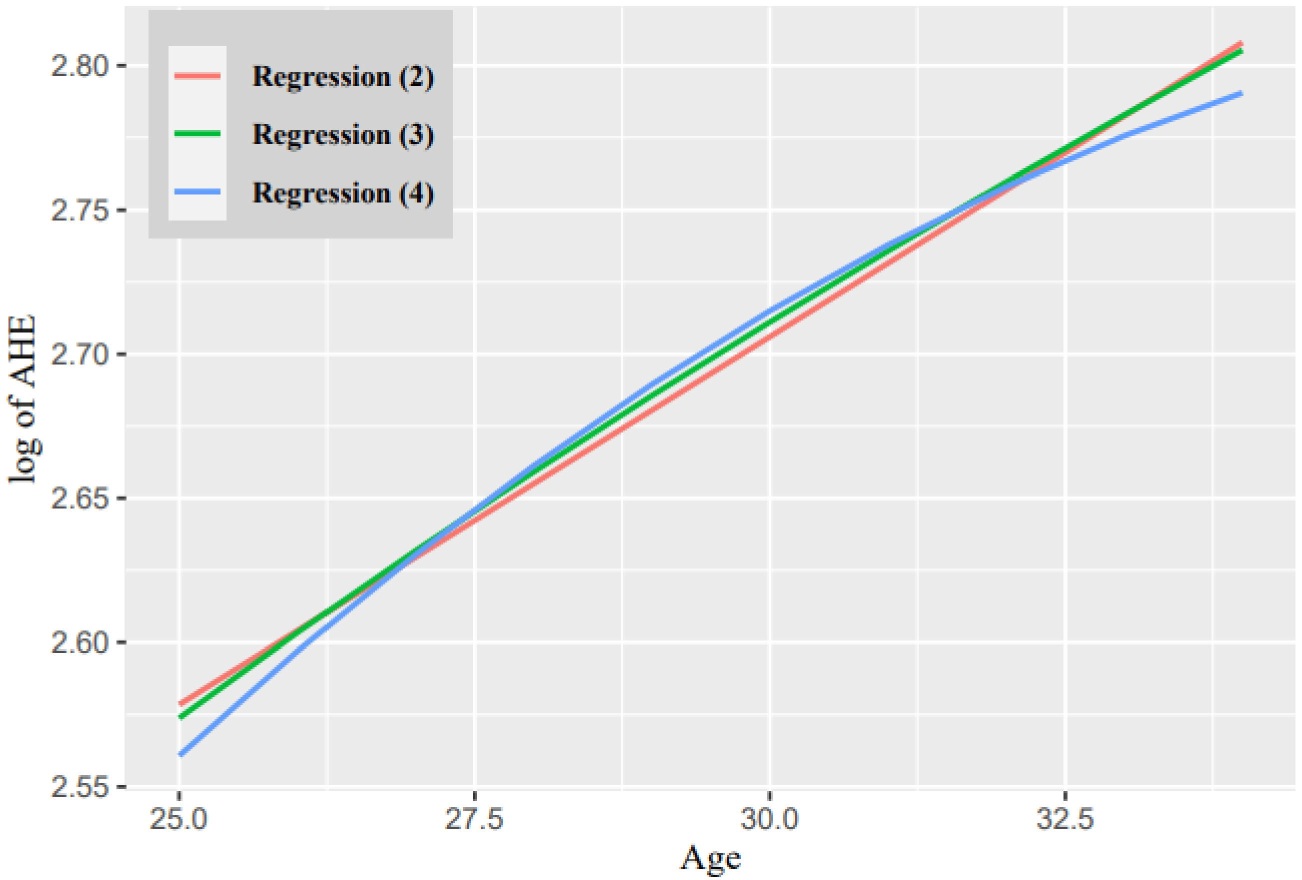
(h) Plot the regression relation between age and ln(ahe) from (b), (c), and (d) for males with a high school diploma. Describe the similarities and differences between the estimated regression functions. Would your answer change if you plotted the regression function for females with college degrees?

```
fig8.2 <- ggplot(data.bcd, aes(x = age, y = ln_age)) +
  geom_line(aes(col = data), size = 0.8) +
  labs(title = "Figure 2: Regression Lines (2) - (4)",
       x = "Age", y = "log of AHE") +
  theme(axis.title = element_text(family = "serif"),
        plot.title = element_text(hjust = 0.5, size = 12, family = "serif",
                                  face = "bold"),
        legend.position = c(0.15, 0.85),
        legend.title = element_blank(),
        legend.text = element_text(family = "serif", face = "bold"),
        legend.key = element_rect(color = "transparent"),
        legend.background = element_rect(fill = "lightgrey",
                                         size = 0.8,
                                         linetype="solid")) +
  scale_color_discrete(name = "Model", labels = c(" Regression (2)",
                                                  " Regression (3)",
                                                  " Regression (4)"))

print(fig8.2)
```

(h) Plot the regression relation between `age` and ln(`ahe`) from (b), (c), and (d) 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 2: Regression Lines (2) – (4)



(h) Plot the regression relation between `age` and ln(`ahe`) from (b), (c), and (d) 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?

See Figure 2. The regression functions are very similar. The quadratic regression shows somewhat more curvature than the log-log regression, but the difference is small. The regression functions for a female with a high school diploma will look just like these, but they will be shifted by the amount of the coefficient on the binary regressor `female`. The regression functions for workers with a bachelor degree will also look just like these, but they would be shifted by the amount of the coefficient on the binary variable `bachelor`.

(i) Run a regression of ln(**ahe**) on **age**, $\texttt{age}^2$, **female**, **bachelor**, and the interaction term **female** × **bachelor**. 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 diploma. 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 diploma. What does the regression predict for his value of ln(**ahe**)? What is the predicted difference between Bob's and Jim's earnings?

Table 2: Earnings and Age, 2012

	(1)	(2)	(3)	(4)	(5)
(Intercept)	1.866	1.941***	0.150	0.792	0.804
	(1.175)	(0.059)	(0.195)	(0.671)	(0.671)
age	0.510***	0.026***		0.104*	0.104*
	(0.040)	(0.002)		(0.046)	(0.046)
female	−3.810***	−0.192***	−0.192***	−0.192***	−0.242***
	(0.224)	(0.011)	(0.011)	(0.011)	(0.017)
bachelor	8.319***	0.438***	0.438***	0.437***	0.400***
	(0.224)	(0.011)	(0.011)	(0.011)	(0.015)
ln_age			0.753***		
			(0.058)		
age2				−0.001	−0.001
				(0.001)	(0.001)
fem_bac					0.090***
					(0.023)

THE UNIVERSITY OF QUEENSLAND AUSTRALIA

(i) Run a regression of ln(**ahe**) on **age**, **age**^{2}, **female**, **bachelor**, and the interaction term **female** × **bachelor**. 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 diploma. 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 diploma. What does the regression predict for his value of ln(**ahe**)? What is the predicted difference between Bob's and Jim's earnings?

```
> (Alexis <- predict(reg5, newdata = data.frame(
+   age = 30,
+   age2 = 30^2,
+   female = 1,
+   bachelor = 1,
+   fem_bac = 1
+ )))
       1
2.982923
> (Jane <- predict(reg5, newdata = data.frame(
+   age = 30,
+   age2 = 30^2,
+   female = 1,
+   bachelor = 0,
+   fem_bac = 0
+ )))
       1
2.492619
> Alexis - Jane
        1
0.4903034
```

```
> (Bob <- predict(reg5, newdata = data.frame(
+   age = 30,
+   age2 = 30^2,
+   female = 0,
+   bachelor = 1,
+   fem_bac = 0
+ )))
       1
3.135439
> (Jim <- predict(reg5, newdata = data.frame(
+   age = 30,
+   age2 = 30^2,
+   female = 0,
+   bachelor = 0,
+   fem_bac = 0
+ )))
       1
2.734993
> Bob - Jim
        1
0.4004463
```

(i) Run a regression of $\ln(\texttt{ahe})$ on $\texttt{age}$, $\texttt{age}^2$, $\texttt{female}$, $\texttt{bachelor}$, and the interaction term $\texttt{female} \times \texttt{bachelor}$. 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 $\ln(\texttt{ahe})$? Jane is a 30-year-old female with a high school diploma. What does the regression predict for her value of $\ln(\texttt{ahe})$? 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 $\ln(\texttt{ahe})$? Jim is a 30-year-old male with a high school diploma. What does the regression predict for his value of $\ln(\texttt{ahe})$? What is the predicted difference between Bob's and Jim's earnings?

This regression is shown in column (5) of Table 2. The coefficient on the interaction term $\texttt{female} \times \texttt{bachelor}$ shows the "extra effect" of $\texttt{bachelor}$ on $\ln(\texttt{ahe})$ for women relative the effect for men.

Predicted values of $\ln(\texttt{ahe})$:

$$Alexis: 0.104 \times 30 - 0.0013 \times 30^2 - 0.24 \times 1 + 0.40 \times 1 + 0.090 \times 1 + 0.80 = 3.00$$

$$Jane: 0.104 \times 30 - 0.0013 \times 30^2 - 0.24 \times 1 + 0.40 \times 0 + 0.090 \times 0 + 0.80 = 2.51$$

$$Bob: 0.104 \times 30 - 0.0013 \times 30^2 - 0.24 \times 0 + 0.40 \times 1 + 0.090 \times 0 + 0.80 = 3.15$$

$$Jim: 0.104 \times 30 - 0.0013 \times 30^2 - 0.24 \times 0 + 0.40 \times 0 + 0.090 \times 0 + 0.80 = 2.75$$

Difference in $\ln(\texttt{ahe})$: Alexis - Jane $= 3.00 - 2.51 = 0.49.\backslash$ Difference in $\ln(\texttt{ahe})$: Bob - Jim $= 3.15 - 2.75 = 0.40$

Notice that the difference in the differences of the predicted effects is $0.49 - 0.40 = 0.09$, which is the value of the coefficient on the interaction term.

THE UNIVERSITY OF QUEENSLAND AUSTRALIA

(j) Is the effect of age on earnings different for men than for women? Specify and estimate a regression that you can use to answer this question.

```
linearHypothesis(reg6, c("fem_age = 0", "fem_age2 = 0"), test=c("F"))
```

```
## Linear hypothesis test
##
## Hypothesis:
## fem_age = 0
## fem_age2 = 0
##
## Model 1: restricted model
## Model 2: ln_ahe ~ age + age2 + fem_age + fem_age2 + female + bachelor +
##     fem_bac
##
##   Res.Df Df     F  Pr(>F)
## 1   7434
## 2   7432  2 4.137 0.01601 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

This regression is shown in column (6) of Table 2, which includes two additional regressors: the interactions of female and the age variables, age and age^2. The F-statistic testing the restriction that the coefficients on these interaction terms is equal to zero is 4.14 with a p-value of 0.016. This implies that there is statistically significant evidence (at the 5% but not 1% level) that there is a different effect of age on ln(ahe) for men and women.

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

```
linearHypothesis(reg7, c("bac_age = 0", "bac_age2 = 0"), test=c("F"))
```

```
## Linear hypothesis test
##
## Hypothesis:
## bac_age = 0
## bac_age2 = 0
##
## Model 1: restricted model
## Model 2: ln_ahe ~ age + age2 + bac_age + bac_age2 + female + bachelor +
##     fem_bac
##
##   Res.Df Df      F Pr(>F)
## 1   7434
## 2   7432  2 1.3003 0.2725
```

This regression is shown in column (7) of Table 2, which includes two additional regressors that are interactions of bachelor and the age variables, age and age^2. The F-statistic testing the restriction that the coefficients on these interaction terms is zero is 1.30 with a p-value of 0.273. This implies that there is no statistically significant evidence (at the 10% level) that there is a different effect of age on ln(ahe) for high school and college graduates.

(l) After running all these regressions (and any others that you want to run), summarise the effect of age on earnings for young workers.

Table 3: Results using (8) from the 2012 Data

Gender, Education	Predicted Value of `ln(ahe)` at Age			Predicted Increase in `ln(ahe)` Percent per year	
	25	32	34	25 to 32	32 to 34
Female, High School	2.36	2.52	2.53	2.3	0.4
Male, High School	2.60	2.78	2.84	2.5	3.3
Female, BA	2.81	3.03	3.02	3.1	-0.4
Male, BA	2.96	3.19	3.24	3.3	2.6

The table3 summarizes the regressions predictions for increases in earnings as a person ages from 25 to 32 and 32 to 34.

The estimated regressions suggest that earnings increase as workers age from 25–35, the range of age studied in this sample. Gender and education are significant predictors of earnings, and there are statistically significant interaction effects between age and gender and between gender and and education.

Earnings for those with a college education are higher than those with a high school degree, and earnings of the college educated increase more rapidly early in their careers (age 25–34). Earnings for men are higher than those of women, and earnings of men increase more rapidly early in their careers (age 25–34). For all categories of workers (men/women, high school/college) earnings increase more rapidly from age 25–32 than from 32–34. While the percentage increase in women's earning is similar to the percentage increase for men from age 25–32, women's earning tend to stagnate from age 32–34, while men's continues to increase.

The University of Queensland
AUSTRALIA
CREATE CHANGE

Thank you

Francisco Tavares Garcia

Academic Tutor | School of Economics

tavaresgarcia.github.io

Reference

Stock, J. H., & Watson, M. W. (2019). Introduction to Econometrics, Global Edition, 4th edition. Pearson Education Limited.