Review Session 3

Non-linearities in Linear Regression

Ben Berger

February 10, 2023

Today

- Interactions
- Binary dependent variables
  - Exercise on predicting political ideology and making regression tables
- Quadratics
- Logarithms
  - Exercise on plotting relationship between life expectancy and GDP.

Following along with the R exercises

My recommended workflow:

- Download the entire Review Session 3 Dropbox folder
- Unzip the file.
- Open the `.Rproj` file. This will boot a new RStudio session and you'll automatically be in the right directory to load the data.
- Open a new blank script. Use this to load the data and perform your analysis.

Interactions

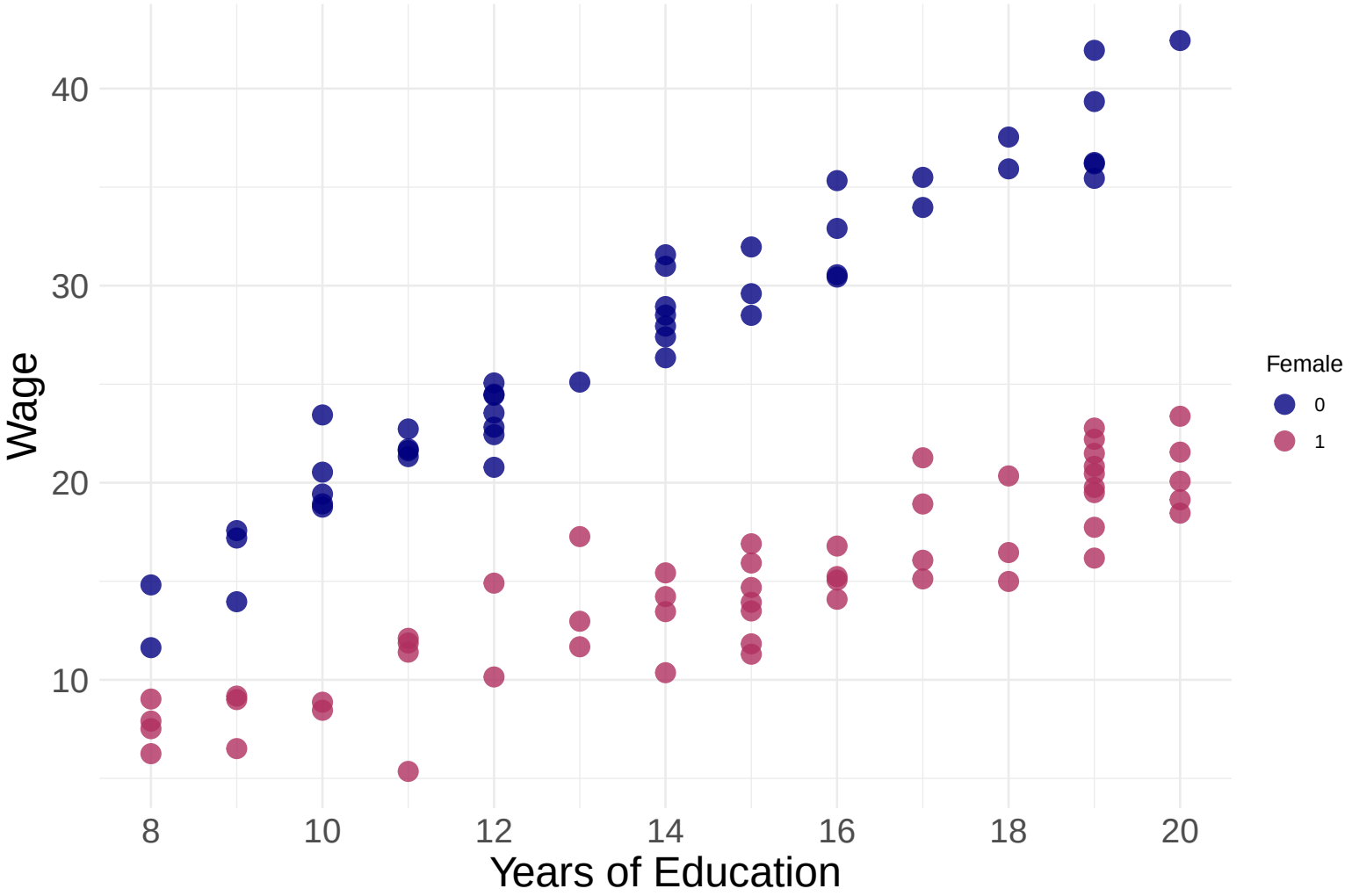
Suppose we fit the regression from the last TF session that **doesn't** include interactions. Note again that we are using fake data for this part.

$$wage_i = \beta_0 + \beta_1 female_i + \beta_2 educ_i + u$$

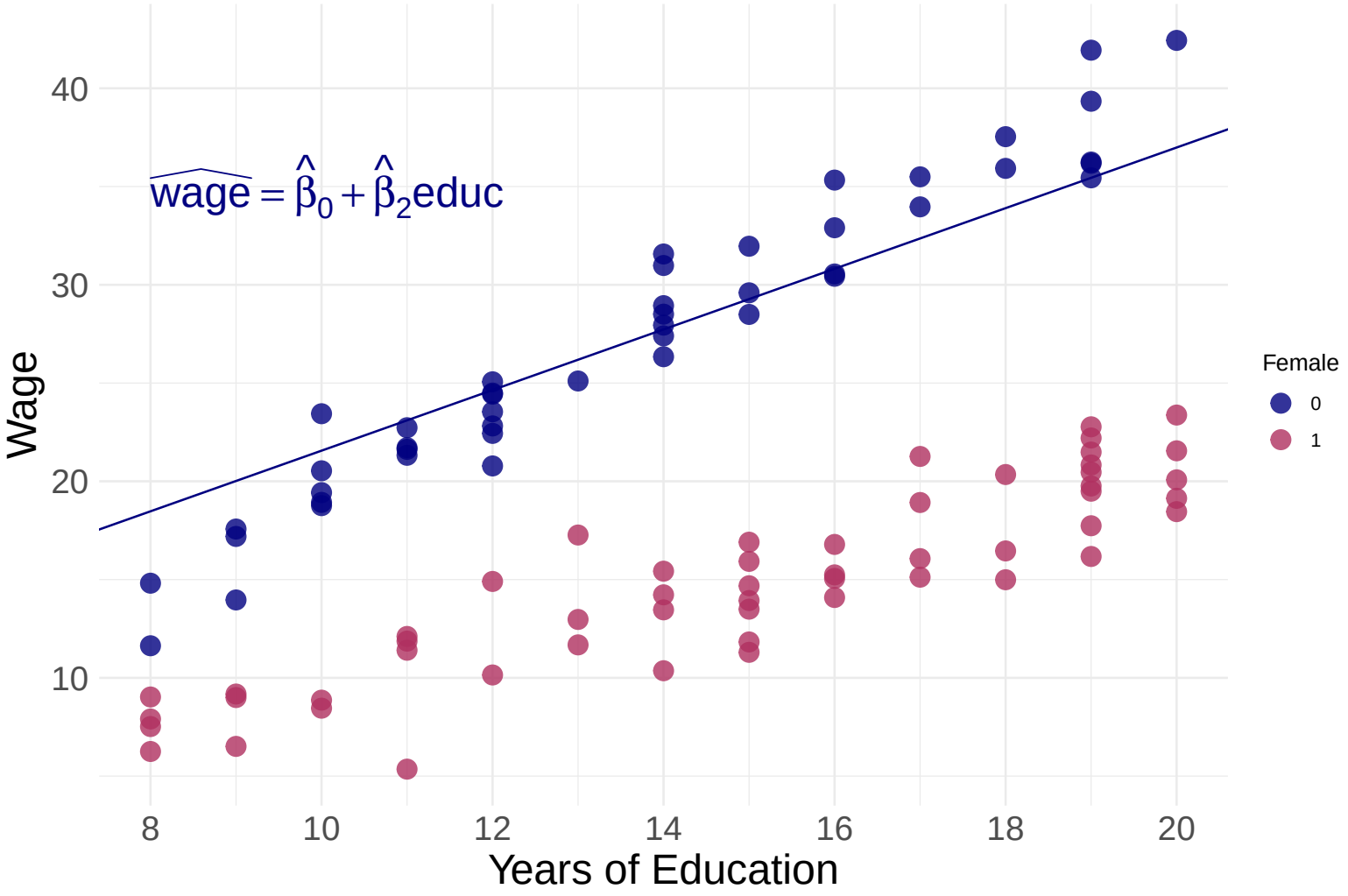
```
fit_wage <- feols(wage ~ female + educ, data, vcov = "hetero")
summary(fit_wage)
```

```
OLS estimation, Dep. Var.: wage
Observations: 100
Standard-errors: Heteroskedasticity-robust
             Estimate Std. Error   t value   Pr(>|t|)
(Intercept)   6.13266   1.317638   4.65428 1.0308e-05 ***
female      -14.28905   0.567063 -25.19837  < 2.2e-16 ***
educ          1.54262   0.091830  16.79860  < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
RMSE: 2.67085   Adj. R2: 0.903332
```

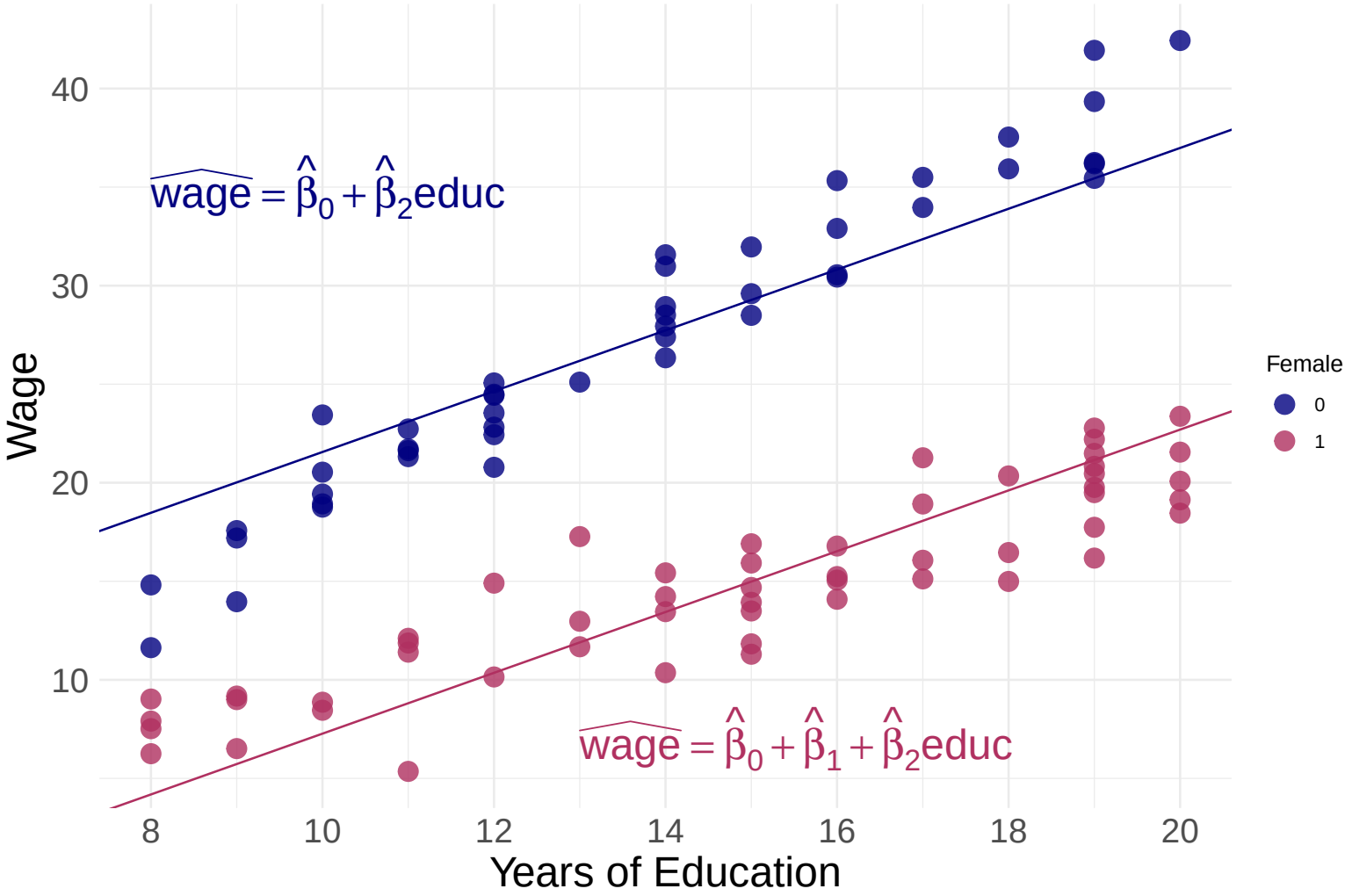
Interactions



Interactions



Interactions



Interactions

Now consider the model that interacts $female_i$ with $educ_i$.

$$wage_i = \beta_0 + \beta_1 female_i + \beta_2 educ_i + \beta_3 female_i \times educ_i + u$$

```
fit_wage_int <- feols(wage ~ female + educ + female:educ,
                      data, vcov = "hetero")
summary(fit_wage_int)
```

```
OLS estimation, Dep. Var.: wage
Observations: 100
Standard-errors: Heteroskedasticity-robust
             Estimate Std. Error   t value   Pr(>|t|)
(Intercept) -2.425627   1.236519 -1.961657 5.2699e-02 .
female       0.552979   1.589054  0.347992 7.2861e-01
educ         2.171501   0.091074 23.843137  < 2.2e-16 ***
female:educ -1.050926   0.112669 -9.327540 4.1438e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
RMSE: 1.93133   Adj. R2: 0.948926
```

Interactions

Model:

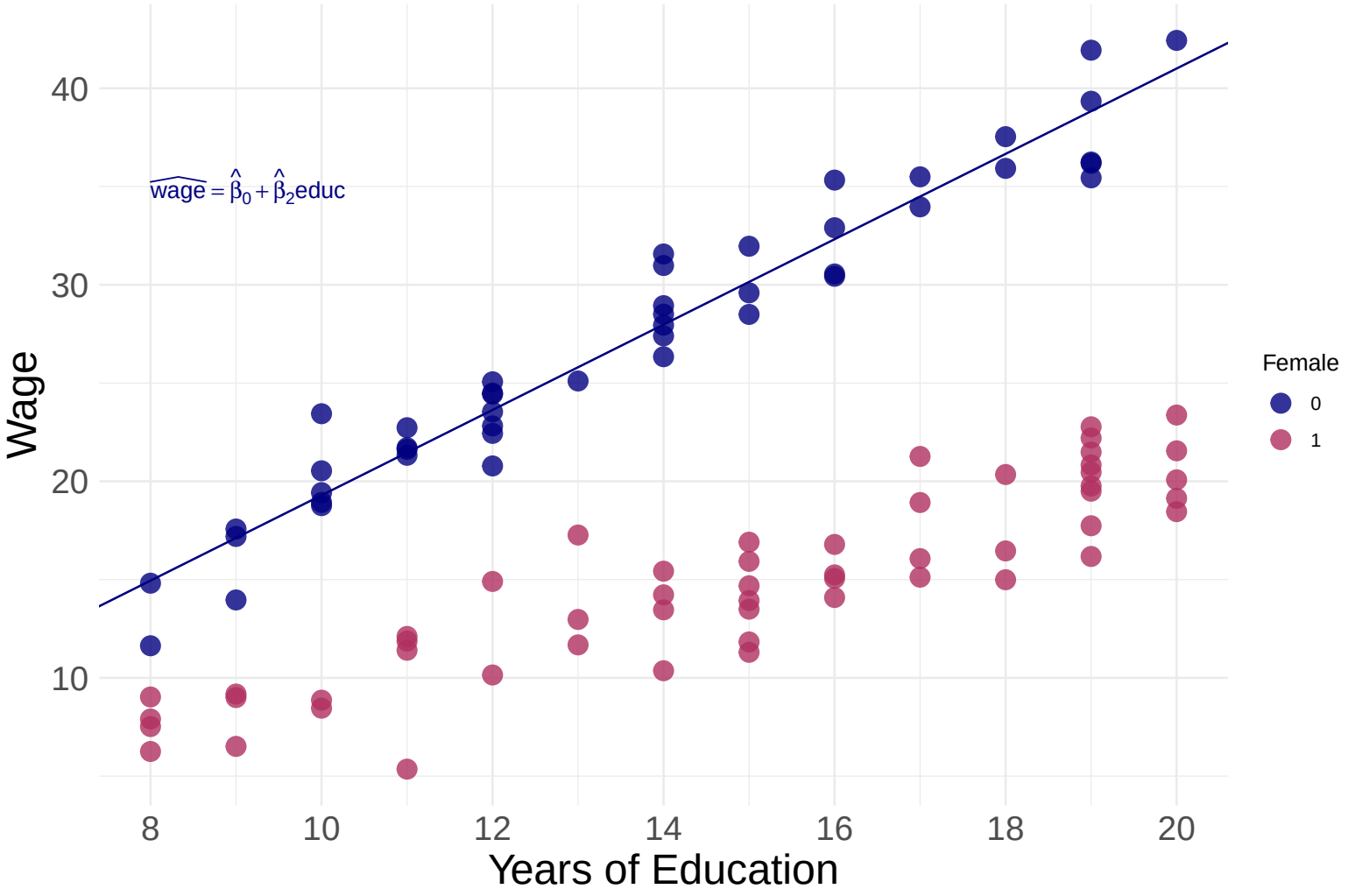
$$wage_i = \beta_0 + \beta_1 female_i + \beta_2 educ_i + \beta_3 female_i \times educ_i + u_i$$

What does this model reduce to in the case that $female = 0$? $female = 1$?

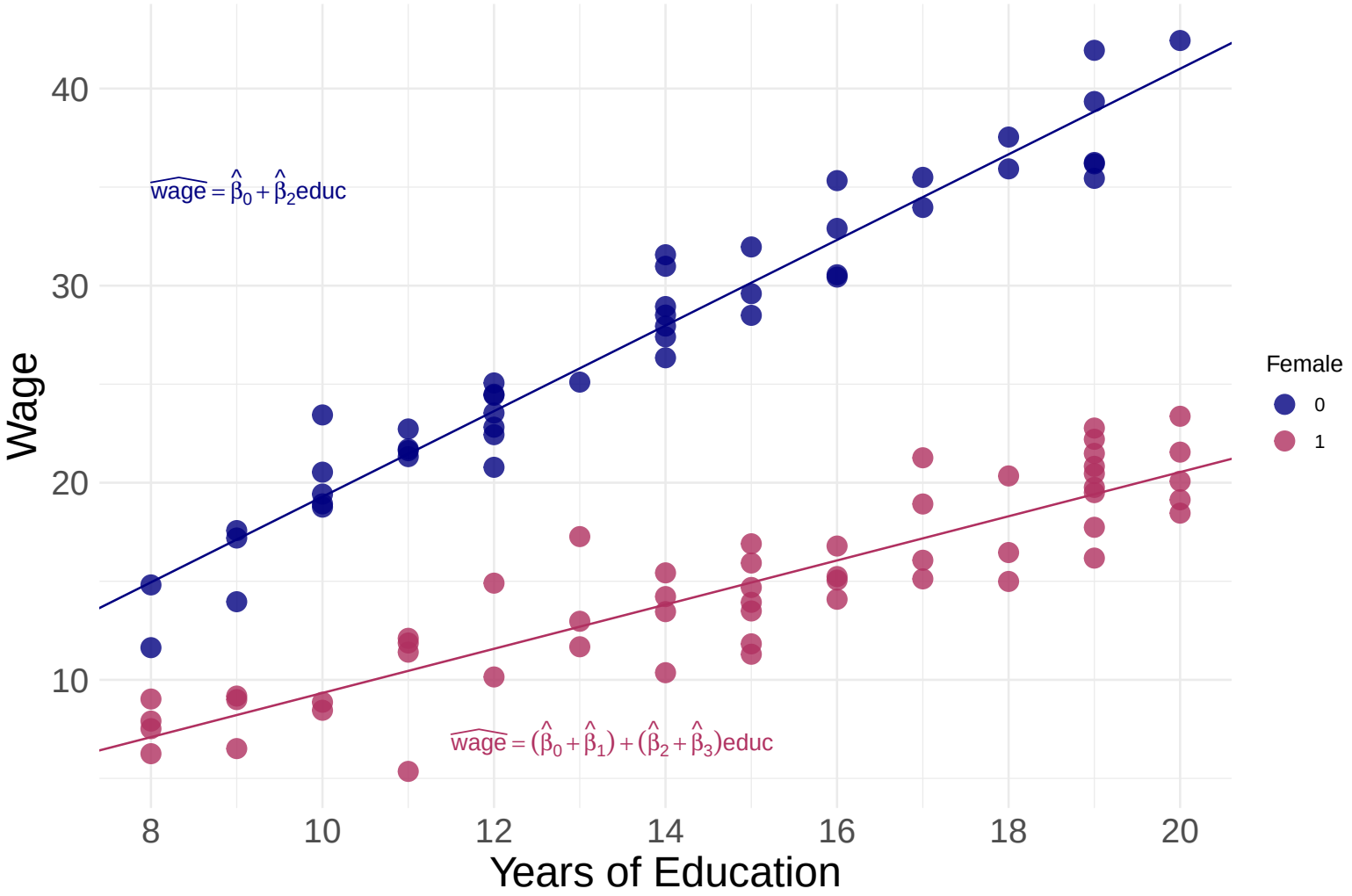
$$\begin{aligned} female_i = 0 \implies wage_i &= \beta_0 + \beta_2 educ_i + u_i \\ female_i = 1 \implies wage_i &= \beta_0 + \beta_1 + \beta_2 educ_i + \beta_3 educ_i + u_i \\ &= (\beta_0 + \beta_1) + (\beta_2 + \beta_3) educ_i + u_i \end{aligned}$$

The interacted model nests two different bivariate regressions, one for males and one for females.

Interactions



Interactions



Interpreting Dummy Variables

Model:

$$wage_i = \beta_0 + \beta_1 female_i + \beta_2 educ_i + \beta_3 female_i \times educ_i + u_i$$

- β_0 : the expected wage of males with 0 years of education.
- β_1 : the expected difference in wage between females and males with 0 years of education.
- β_2 : the change in wage associated with one more year of education for males.
- β_3 : the additional change in wage associated with one more year of education for females relative to the association for males.

Interpreting Dummy Variables

```
OLS estimation, Dep. Var.: wage
Observations: 100
Standard-errors: Heteroskedasticity-robust
             Estimate Std. Error   t value   Pr(>|t|)
(Intercept) -2.425627   1.236519 -1.961657 5.2699e-02 .
female       0.552979   1.589054  0.347992 7.2861e-01
educ         2.171501   0.091074 23.843137  < 2.2e-16 ***
female:educ -1.050926   0.112669 -9.327540 4.1438e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
RMSE: 1.93133   Adj. R2: 0.948926
```

Given this regression output, how would you test the following null hypotheses?

a. The association between education and wage among males is 0.
   Use p-value on `educ`.
b. The female-male difference in the association between education and wage is 0.
   Use p-value on `female:educ`.

Interpreting Dummy Variables

```
OLS estimation, Dep. Var.: wage
Observations: 100
Standard-errors: Heteroskedasticity-robust
             Estimate Std. Error   t value   Pr(>|t|)
(Intercept) -2.425627   1.236519 -1.961657 5.2699e-02 .
female       0.552979   1.589054  0.347992 7.2861e-01
educ         2.171501   0.091074 23.843137  < 2.2e-16 ***
female:educ -1.050926   0.112669 -9.327540 4.1438e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
RMSE: 1.93133   Adj. R2: 0.948926
```

- The association between education and wage among *females* is 0.

We know that the sample association between education and wage for `females` is $\hat{\beta}_2 + \hat{\beta}_3$. To test this against 0, we need to know $se(\hat{\beta}_2 + \hat{\beta}_3)$, which we can't derive from the regression summary alone because it depends on the covariance of $\hat{\beta}_2$ and $\hat{\beta}_3$.

Interpreting Dummy Variables

```
OLS estimation, Dep. Var.: wage
Observations: 100
Standard-errors: Heteroskedasticity-robust
             Estimate Std. Error   t value   Pr(>|t|)
(Intercept) -2.425627   1.236519 -1.961657 5.2699e-02 .
female       0.552979   1.589054  0.347992 7.2861e-01
educ         2.171501   0.091074 23.843137  < 2.2e-16 ***
female:educ -1.050926   0.112669 -9.327540 4.1438e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
RMSE: 1.93133   Adj. R2: 0.948926
```

Solutions:

- Calculate $se(\hat{\beta}_2 + \hat{\beta}_3) = \sqrt{Var(\hat{\beta}_2) + Var(\hat{\beta}_3) + 2Cov(\hat{\beta}_2, \hat{\beta}_3)}$ using the regression's variance-covariance matrix: `vcov(fit_wage_int)`.
- Estimate the regression of wage on `male`, `educ`, and `male:educ` and report the coefficient on `educ`.

Dummy-Dummy Interactions

Now suppose that we have a variable $black_i$ which equals 1 if person i is black and 0 otherwise. We estimate the following model:

$$wage_i = \beta_0 + \beta_1 female_i + \beta_2 black_i + \beta_3 female_i \times black_i + u_i$$

- Non-black male, $female = 0$, $black = 0$: $\widehat{wage}_i = \beta_0$
- Non-black female, $female = 1$, $black = 0$: $\widehat{wage}_i = \beta_0 + \beta_1$
- Black male, $female = 0$, $black = 1$: $\widehat{wage}_i = \beta_0 + \beta_2$
- Black female, $female = 1$, $black = 1$: $\widehat{wage}_i = \beta_0 + \beta_1 + \beta_2 + \beta_3$

Dummy-Dummy Interactions

Model:

$$wage_i = \beta_0 + \beta_1 female_i + \beta_2 black_i + \beta_3 female_i \times black_i + u_i$$

- β_0 : the average wage of non-black males.
- β_1 : the difference in average wage of non-black females compared to non-black males.
- β_2 : the difference in average wage of black males compared to non-black males.
- β_3 : the additional difference in average wage of black females compared to black males relative to the difference between non-black females and non-black males

Linear Probability Model

Basic Idea: The linear probability model (LPM) is the same old linear regression model we've been working with all along, just with a binary dependent variable:

$$Y = \beta_0 + \beta_1 X_1 + ...\beta_k X_k + u, \text{ where } Y = 1 \text{ or } 0$$

Interpretation of β_1: A one-unit increase in X_1 is associated with a $\beta_1 \times$ 100 percentage point increase in the probability of the outcome corresponding to $Y = 1$, controlling for other variables.

Suppose we estimate the following SRF:

$$Covid_i = 0.0093 - 0.0088 Vax_i + \hat{u}_i$$

Individuals who receive the vaccine ($Vax_i = 1$) are 0.88 **percentage points** less likely to be infected with Covid-19 relative to the unvaccinated.

What is the relative risk of the vaccinated?

$$\frac{0.0093 - 0.0088}{0.0093} = 5\% \text{ as likely to be infected}$$

Linear Probability Model

Key issue with LPM: it is possible to get predicted probabilities less than 0 or greater than 1.

- Prediction: this is usually a significant issue because we can definitely do better than predicting probabilities outside of $[0, 1]$. Use logit or probit model instead (wait for the prediction unit).
- Causal Inference: this issue rarely matters. Average marginal effect generally very close to estimates from logit or probit.

Exercise 1: Predicting Liberals

Let's estimate some models to predict whether individuals identify as liberal (dep. var. `lib`) using a dummy variable `female`, educational achievement `colgrad`, and their interaction.

```
library(tidyverse)
library(haven)
library(fixest)
# Load data
nes <- read_dta("nes2012edit.dta")
```

Load the NES data (`nes2012edit.dta` — this is different from the dataset used last week), and create a new variable called `female_colgrad` that is the interaction of `female` and `colgrad`.

Exercise 1: Predicting Liberals

```
# Create an interaction term
nes <- mutate(nes, female_colgrad = female * colgrad)
```

Now estimate the following models

$$lib_i = \alpha_0 + \alpha_1 female_i + u_i \tag{1}$$
$$lib_i = \beta_0 + \beta_1 female_i + \beta_2 colgrad_i + v_i \tag{2}$$
$$lib_i = \gamma_0 + \gamma_1 female_i + \gamma_2 colgrad_i + \gamma_3 female_i \times colgrad_i + \varepsilon_i \tag{3}$$

Exercise 1: Predicting Liberals

```
# Estimate models
m1 <- feols(lib ~ female,
            nes, vcov = "hetero")
m2 <- feols(lib ~ female + colgrad,
            nes, vcov = "hetero")
m3 <- feols(lib ~ female + colgrad + female_colgrad,
            nes, vcov = "hetero")
```

Tip: instead of creating a variable for the interaction term, you can use `female:colgrad` in the regression formula to include an interaction.

Exercise 1: Predicting Liberals

Now we can call on `etable()` from `fixest` to make a table of results.

```
etable(m1, m2, m3)
```

```
                                        m1
Dependent Var.:                    Liberal

Constant                 0.3147*** (0.0088)
Female                   0.0517*** (0.0125)
College Graduate
Female x College Graduate
------------------------ ------------------
S.E. type                Heteroskedas.-rob.
Observations                         5,722
R2                                 0.00296
Adj. R2                            0.00279

                                        m2
Dependent Var.:                    Liberal

Constant                 0.2874*** (0.0098)
Female                   0.0548*** (0.0125)
College Graduate         0.0809*** (0.0137)
Female x College Graduate
------------------------ ------------------
S.E. type                Heteroskedas.-rob.
Observations                         5,722
R2                                 0.00925
```

Exercise 1: Predicting Liberals

It would be nice to put the standard errors below the coefficient estimates.

```
etable(m1, m2, m3, se.below = TRUE)
```

```
                                 m1          m2          m3
Dependent Var.:             Liberal     Liberal     Liberal

Constant                  0.3147***   0.2874***   0.2913***
                         (0.0088)    (0.0098)    (0.0106)
Female                    0.0517***   0.0548***   0.0475**
                         (0.0125)    (0.0125)    (0.0149)
College Graduate                      0.0809***   0.0694***
                                     (0.0137)    (0.0189)
Female x College Graduate                         0.0229
                                                 (0.0273)
------------------------ ---------- ---------- ----------
S.E. type                Hete.-rob. Hete.-rob. Hete.-rob.
Observations                 5,722       5,722       5,722
R2                         0.00296     0.00925     0.00938
Adj. R2                    0.00279     0.00891     0.00886
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Exercise 1: Predicting Liberals

It would also be nice to format the table by replacing independent and dependent variable names with descriptive labels. We can do this by first using the function `setFixest_dict()` to assign labels to variable names.

```
setFixest_dict(c(lib = "Liberal",
                 female = "Female",
                 colgrad = "College Graduate",
                 female_colgrad = "Female x College Graduate"
))
```

Exercise 1: Predicting Liberals

Now `etable()` will automatically replace variable names with labels.

```
etable(m1, m2, m3, se.below = T)
```

```
                                    m1          m2          m3
Dependent Var.:                Liberal     Liberal     Liberal

Constant                     0.3147***   0.2874***   0.2913***
                            (0.0088)    (0.0098)    (0.0106)
Female                       0.0517***   0.0548***   0.0475**
                            (0.0125)    (0.0125)    (0.0149)
College Graduate                         0.0809***   0.0694***
                                        (0.0137)    (0.0189)
Female x College Graduate                            0.0229
                                                    (0.0273)
------------------------- ---------- ---------- ----------
S.E. type                  Hete.-rob. Hete.-rob. Hete.-rob.
Observations                    5,722      5,722      5,722
R2                            0.00296    0.00925    0.00938
Adj. R2                       0.00279    0.00891    0.00886
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Exercise 1: Predicting Liberals

Lastly, we may want to rename the models. To do this, we supply `etable()` with a named list of models like so:

```
etable(list("Bivariate" = m1, "Multivariate" = m2, "Interacted" = m3),
       se.below = T)
```

```
                          Bivariate Multivar.. Interacted
Dependent Var.:             Liberal    Liberal    Liberal

Constant                   0.3147***  0.2874***  0.2913***
                          (0.0088)   (0.0098)   (0.0106)
Female                     0.0517***  0.0548***  0.0475**
                          (0.0125)   (0.0125)   (0.0149)
College Graduate                      0.0809***  0.0694***
                                     (0.0137)   (0.0189)
Female x College Graduate                        0.0229
                                                (0.0273)

------------------------- ---------- ---------- ----------
S.E. type                 Hete.-rob. Hete.-rob. Hete.-rob.
Observations                   5,722      5,722      5,722
R2                           0.00296    0.00925    0.00938
Adj. R2                      0.00279    0.00891    0.00886
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Exercise 1: Predicting Liberals

	Bivariate	Multivar..	Interacted
Dependent Var.:	Liberal	Liberal	Liberal
Constant	0.3147***	0.2874***	0.2913***
	(0.0088)	(0.0098)	(0.0106)
Female	0.0517***	0.0548***	0.0475**
	(0.0125)	(0.0125)	(0.0149)
College Graduate		0.0809***	0.0694***
		(0.0137)	(0.0189)
Female x College Graduate			0.0229
			(0.0273)
Observations	5,722	5,722	5,722
R2	0.00296	0.00925	0.00938
Adj. R2	0.00279	0.00891	0.00886

Interpret the coefficients on the following terms:

- Bivariate: Intercept and Female
- Multivariate: College Graduate
- Interacted: College Graduate and Female $\times$ College Graduate

Exercise 1: Predicting Liberals

- Bivariate:
  - Intercept: 31.5 percent of males identify as liberal
  - Female: females are 5.2 percentage points more likely than men to identify as liberal
- Multivariate:
  - College Graduate: college graduates are 8.1 percentage points more likely to identify as liberal than non-grads, controlling for gender.
- Interacted:
  - College Graduate: Male college graduates are 6.9 percentage points more likely to be liberal than male non-grads.
  - Female $\times$ College Graduate: Female college graduates are an additional 2.3 percentage points more likely to be liberal than female non-grads relative to the difference between male grads and male non-grads (although this isn't a significant difference at 5% significance level).

Quadratics

- So far, we have considered regression models where marginal effects are constant or vary with the value of a different variable.
- But what if we suspect that the change in Y for a change in X varies with the value of X?

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + u$$

- This is like a variable interacting with itself: the marginal effect of X varies as a function of X.

Quadratics

Model:

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + u$$

- Marginal effect:

$$\frac{\partial Y}{\partial X} = \beta_1 + 2\beta_2 X$$

- In the PRF, the coefficients do not have a straightforward interpretation. It's impossible to change one covariate (X) without also changing another (X^2).
- β_0 is still the Y intercept
- β_1 is the slope at $X = 0$
- β_2 is the amount of concavity/convexity

Quadratics

Hypothesis tests to consider with quadratic terms:

1. If you just want to know whether there is evidence of a nonlinear relationship between X and Y:
   - do a t-test on the coefficient of the quadratic term.
2. If you want to know if X is a significant predictor of Y
   - do a joint hypothesis test on the coefficients on all the powers of X with an F-test.

Logarithms

Logarithms, why would we ever want to use those?

They allows us to express relationships in terms of percent changes, rather than absolute changes, which is useful in many policy contexts (e.g. elasticities).

- When Δx is small: $\log(x + \Delta x) - \log(x) \approx \frac{\Delta x}{x}$
- e.g. $\log(100 + 5) - \log(100) = 0.049 \approx 5\%$.

Big limitation: Cannot be used when a variable takes non-positive values.

Logarithms – Interpretation

The interpretation of the regression coefficients depends on whether X, Y, or both are in logarithms (three cases):

1. **Level-log (X in logs, Y is not)**

$$Y = \beta_0 + \beta_1 \log X + u$$

"A 1% increase in X is associated with a $0.01 \times \beta_1$ change in Y"

2. **Log-level (Y in logs, X is not)**

$$\log Y = \beta_0 + \beta_1 X + u$$

"A one unit increase in X is associated with a $\beta_1 \times 100\%$ change in Y"

3. **Log-log (X in logs, Y in logs)**

$$\log Y = \beta_0 + \beta_1 \log X + u$$

"A 1% increase in X is associated with a $\beta_1\%$ change in Y"

Exercise 2: Health-Income Gradient

Health-income gradient is a good example of a "level-log" relationship (at least when measuring health by life-expectancy).

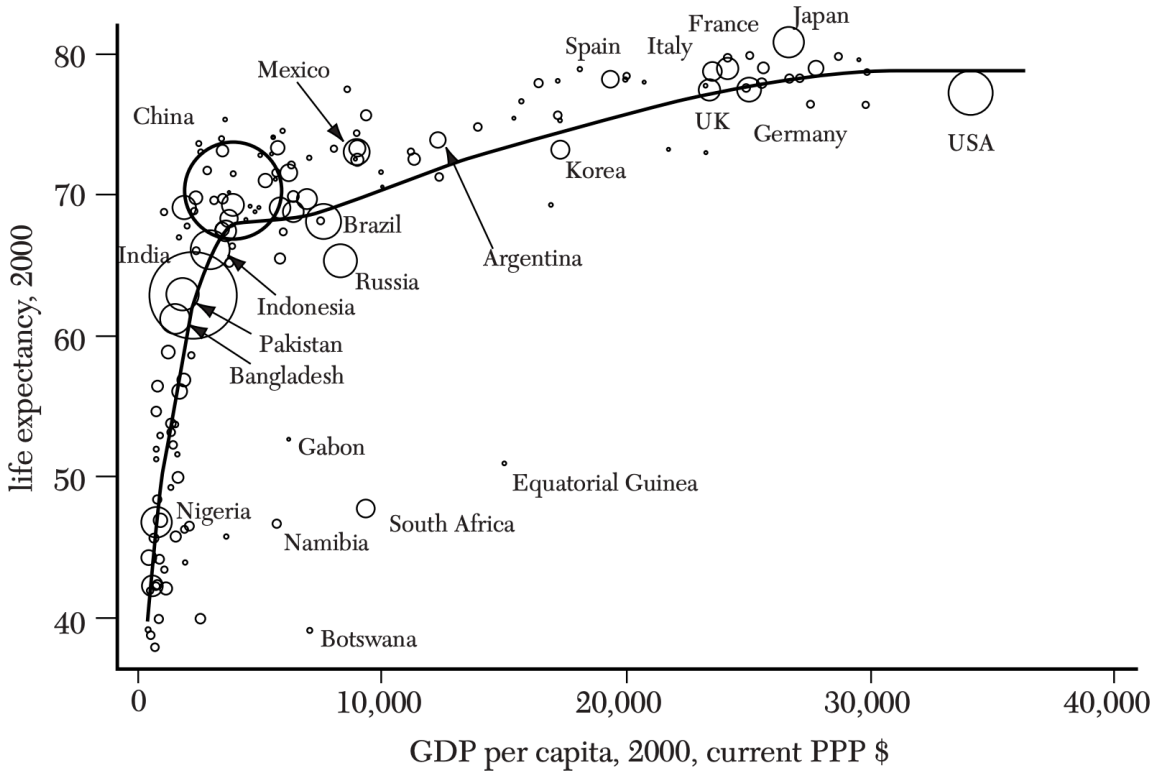


Figure 1. The Preston Curve: Life Expectancy versus GDP Per Capita

Source: World Development Indicators CD-ROM, World Bank (2002)
Note: Circles are proportional to population and some of the largest (or most interesting) countries are labeled. The solid line is a plot of a population-weighted nonparametric regression. Luxembourg, with per capita GDP of $50,061 and life expectancy of 77.04 years, is excluded.

Exercise 2: Health-Income Gradient

Let's reconstruct the Preston curve using 2015 GDP per capita and life expectancy. Open `section_3.Rproj`, start a new script, and load the data. Then use `ggplot` to plot `rgdppc` on the x-axis and `life_exp` on the y-axis.

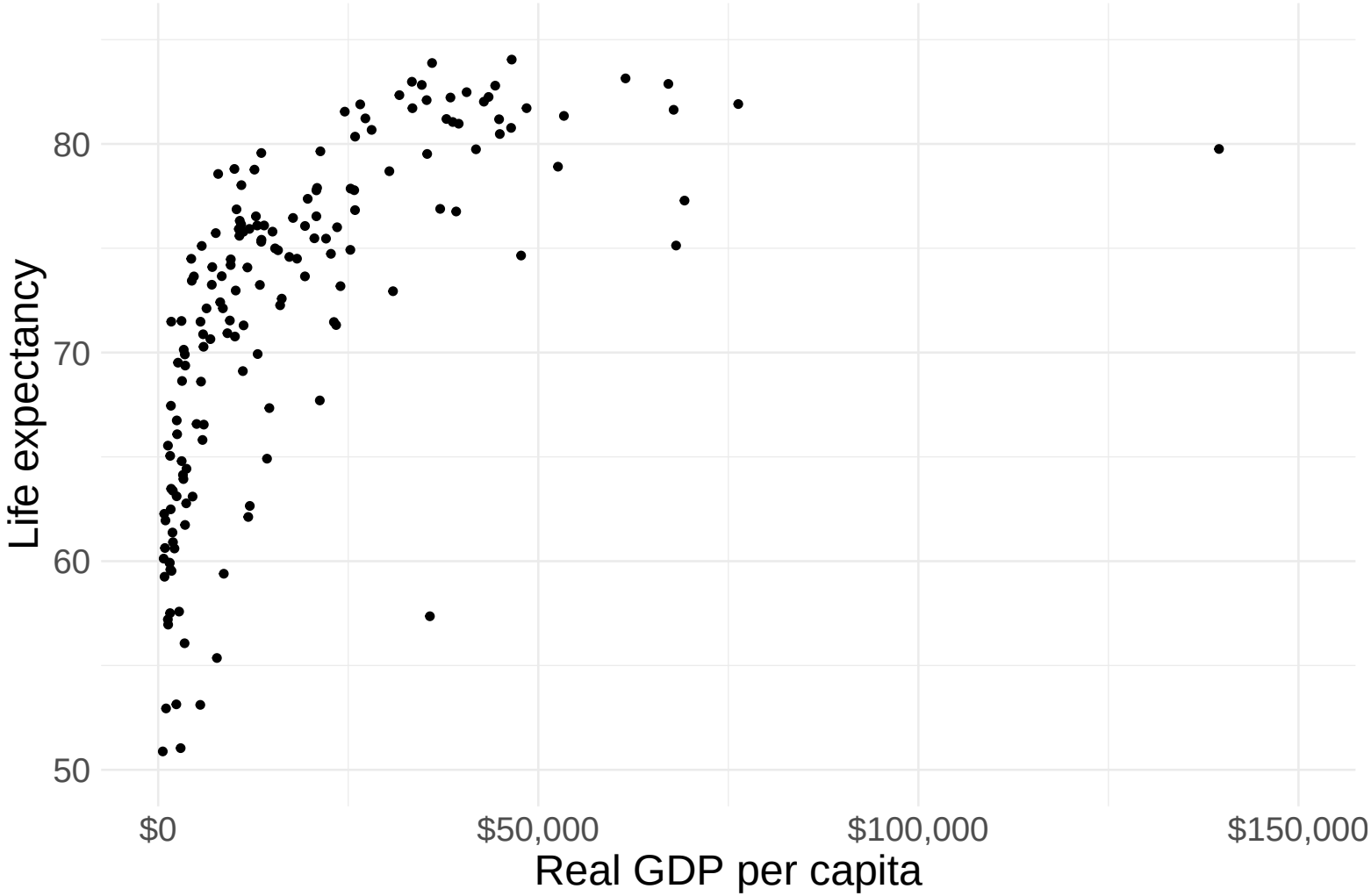
```
library(tidyverse)
library(haven)
library(fixest)

# Read data
preston_data <- read_dta("preston.dta")
```

Exercise 2: Health-Income Gradient

```
# Define plot
preston_plot <- ggplot(preston_data,
                       aes(x = rgdppc, y = life_exp)) +
  geom_point() +
  coord_cartesian(xlim = c(0, 150000), ylim = c(50, 85)) +
  scale_x_continuous(labels = scales::dollar) +
  labs(x = "Real GDP per capita", y = "Life expectancy")
```

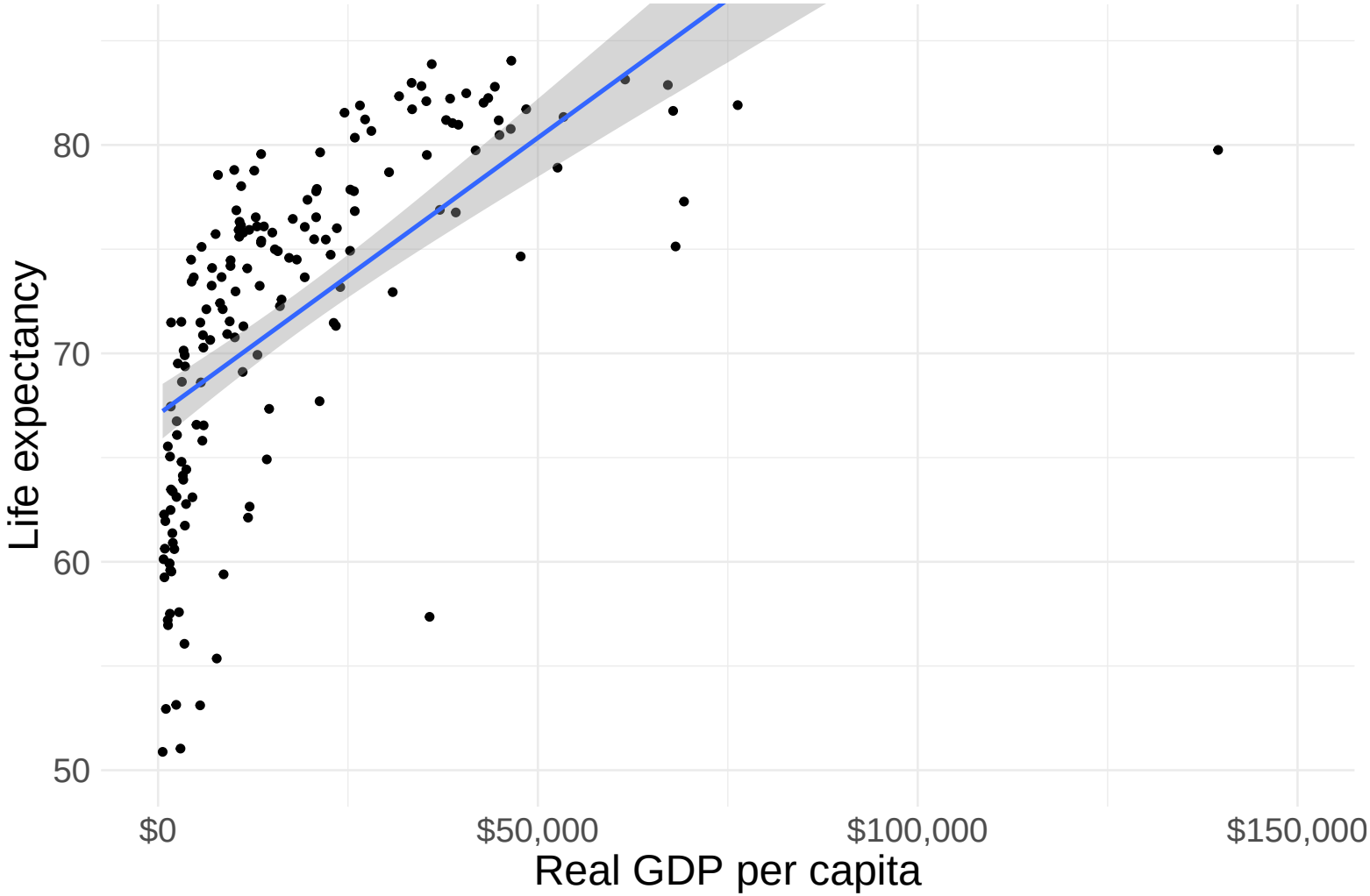
Exercise 2: Health-Income Gradient

```
preston_plot
```



Exercise 2: Health-Income Gradient

```
# Add linear fit
preston_plot + geom_smooth(method = "lm")
```



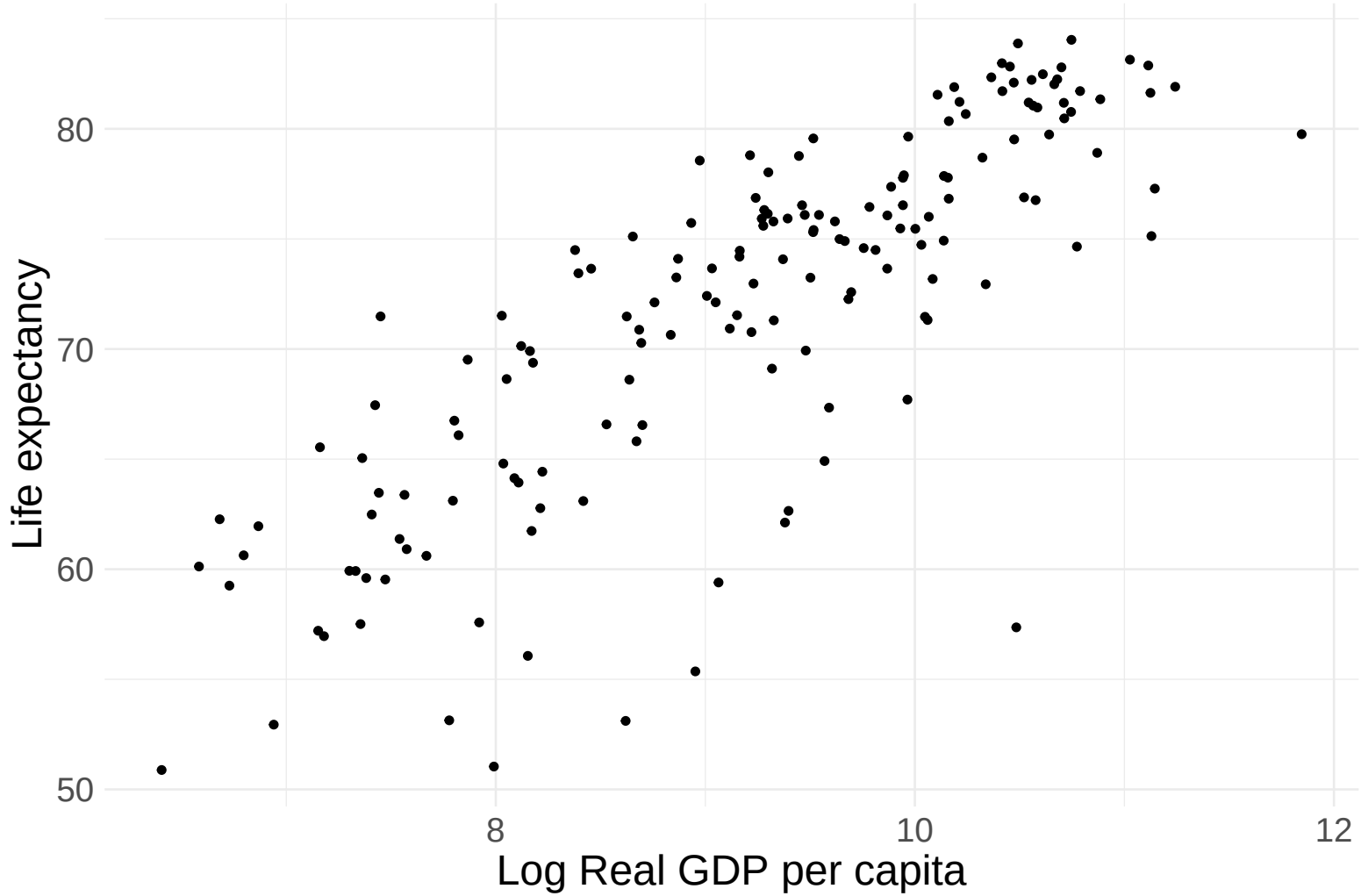
Exercise 2: Health-Income Gradient

Now, create a new variable called `log_rgdppc` and plot the relationship between `life_exp` and `log_rgdppc`.

```
# Add log of rgddpc
preston_data <- mutate(preston_data, log_rgdppc = log(rgdppc))

# Plot level log
preston_logplot <- ggplot(preston_data,
                          aes(x = log_rgdppc, y = life_exp))  +
  geom_point() +
  labs(x = "Log Real GDP per capita",
       y = "Life expectancy")
```

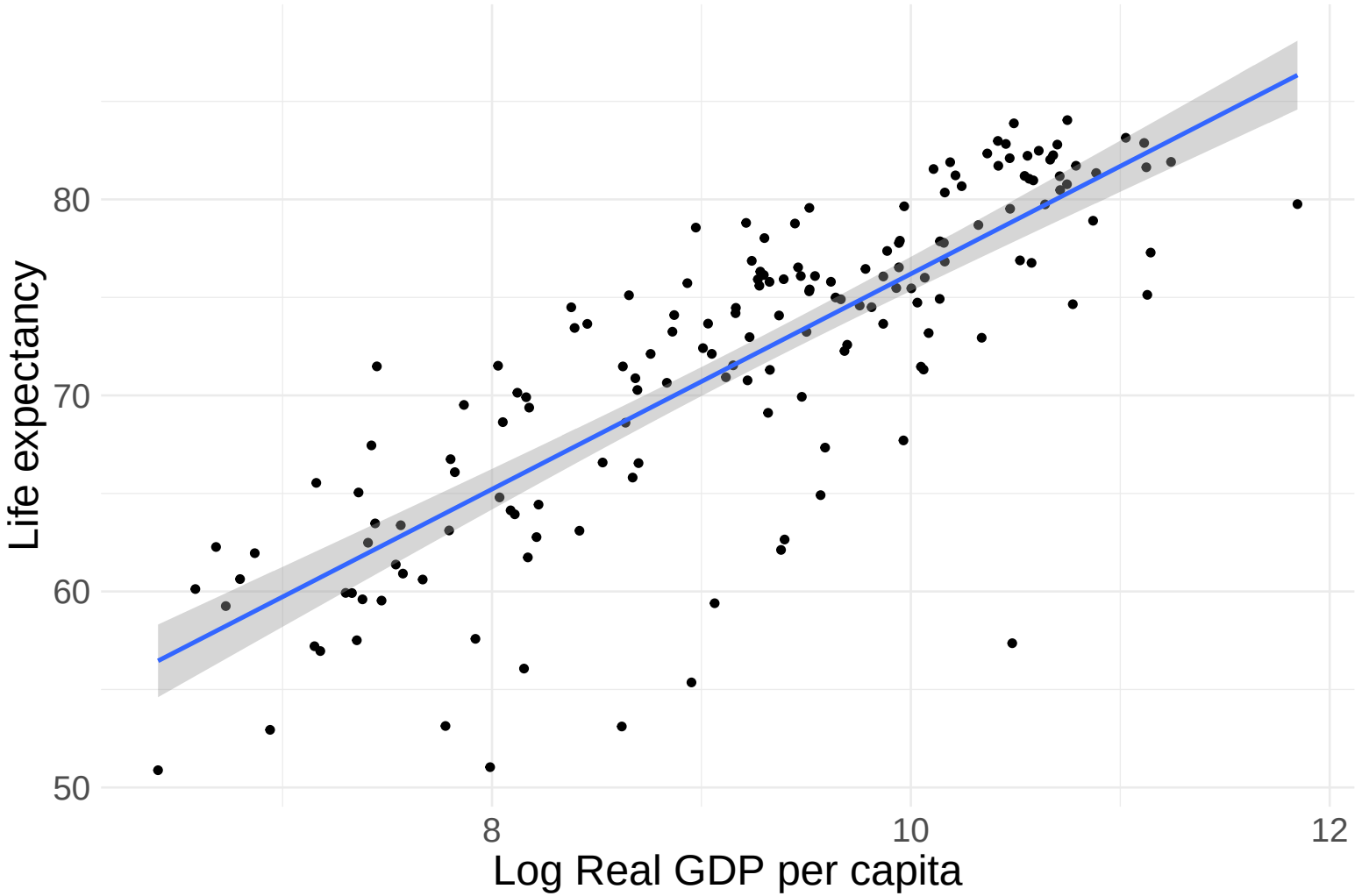
Exercise 2: Health-Income Gradient

```
preston_logplot
```



Exercise 2: Health-Income Gradient

```
# Add linear fit
preston_logplot + geom_smooth(method = "lm")
```



Exercise 2: Health-Income Gradient

```r
# Estimate model
m_levlog <- feols(life_exp ~ log_rgdppc, preston_data, vcov = "hetero")
summary(m_levlog, robust = TRUE)
```

```
OLS estimation, Dep. Var.: life_exp
Observations: 166
Standard-errors: Heteroskedasticity-robust
            Estimate Std. Error  t value   Pr(>|t|)
(Intercept) 21.28639   2.714307  7.84229 5.3481e-13 ***
log_rgdppc   5.49126   0.288221 19.05230  < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
RMSE: 4.71736   Adj. R2: 0.658696
```

Interpretation? A one percent increase in a country's real GDP Per capita is associated with 0.055 additional years of life expectancy.

Exercise 2: Health-Income Gradient

We can also plot the non-linear relationship between life expectancy and real GDP per capita by creating predictions for each value of `rgdppc`.

```
# Extract coefficients
coefs <- m_levlog$coefficients

# Create dataset of predicted values for RGDPPC
pred_data <- tibble(
  rgdppc = seq(100, 150000, 100),
  pred = coefs[1] + coefs[2] * log(rgdppc)
)
```

```
# Plot with predictions
preston_plot +
  geom_line(aes(y = pred, x = rgdppc), data = pred_data)
```

Exercise 2: Health-Income Gradient