Multiple Linear Regression

EC 339

Marcio Santetti

Fall 2022

Motivation

Beyond simple regression

Simple regression models may not be **sufficient** to describe the relationships we are interested in.

A few reasons:

- Avoiding **bias** due to *omitted variables*;
- More consistency with **economic theory**;
- Usually, relationships we study are a product of **several different events**.

Multiple regression models

In **standard** notation:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \ldots + \beta_k x_{ki} + u_i$$
$$\forall\ i = 1, 2, 3, \ldots, n$$

- From last week...

$$wage_i = \beta_0 + \beta_1 educ_i + u_i$$

- And now...

$$wage_i = \beta_0 + \beta_1 educ_i + \beta_2 exper_i + \beta_3 tenure_i + \beta_4 gender_i + u_i$$

Important: even if we are only interested in the effect of *educ* on *wage*, the model above is more consistent with theoretical priors.

An example

```
                        #>
#> ===============================================
#>                          Dependent variable:
#>                      ---------------------------
#>                                 wage
#> -----------------------------------------------
#> educ                          0.541***
#>                               (0.053)
#>
#> Constant                      -0.905
#>                               (0.685)
#>
#> -----------------------------------------------
#> Observations                    526
#> R2                             0.165
#> Adjusted R2                    0.163
#> Residual Std. Error      3.378 (df = 524)
#> F Statistic          103.363*** (df = 1; 524)
#> ===============================================
#> Note:               *p<0.1; **p<0.05; ***p<0.01
```

An example

```
                         #>
#> ================================================
#>                           Dependent variable:
#>                       ---------------------------
#>                                  wage
#> ------------------------------------------------
#> educ                           0.572***
#>                                (0.049)
#> exper                          0.025**
#>                                (0.012)
#> tenure                         0.141***
#>                                (0.021)
#> female                        -1.811***
#>                                (0.265)
#> Constant                      -1.568**
#>                                (0.725)
#> ------------------------------------------------
#> Observations                     526
#> R2                              0.364
#> Adjusted R2                     0.359
#> Residual Std. Error       2.958 (df = 521)
#> F Statistic           74.398*** (df = 4; 521)
#> ================================================
#> Note:                 *p<0.1; **p<0.05; ***p<0.01
```

Interpreting multiple coefficients

The *ceteris paribus* assumption

When **interpreting** multiple regression models, we **isolate** the effect of one independent variable on the dependent variable.

Therefore, the estimated **slope parameters** $(\hat{\beta}_1, \ldots, \hat{\beta}_k)$ inform the change in y resulting from a one-unit change in x_i, *holding all other independent variables constant.*

Mathematically speaking...

$$wage_i = \beta_0 + \beta_1 educ_i + \beta_2 exper_i + \beta_3 tenure_i + \beta_4 gender_i + u_i$$

$$\frac{\partial wage_i}{\partial educ_i} = \beta_1$$

$$\frac{\partial wage_i}{\partial exper_i} = \beta_2$$

Goodness-of-fit

Goodness-of-fit

As more variables are added our model, R^2 increases in a **mechanical** fashion.

- **Problem!**

Simple regression wage model
0.16

Multiple regression wage model
0.36

Goodness-of-fit

- Let us add a `construc` indicator variable, including it into our previous model.

- `construc = 1` if working in the construction sector;
- `construc = 0` otherwise.

$$wage_i = \beta_0 + \beta_1 educ_i + \beta_2 exper_i + \beta_3 tenure_i + \beta_4 gender_i + \beta_5 construc_i + u_i$$

Goodness-of-fit

```
                        #>
#> ===============================================
#>                          Dependent variable:
#>                      ---------------------------
#>                                 wage
#> -----------------------------------------------
#> educ                          0.577***
#>                               (0.050)
#> exper                         0.026**
#>                               (0.012)
#> tenure                        0.141***
#>                               (0.021)
#> female                       -1.788***
#>                               (0.266)
#> construc                       0.563
#>                               (0.626)
#> Constant                     -1.685**
#>                               (0.736)
#> -----------------------------------------------
#> Observations                    526
#> R2                             0.365
#> Adjusted R2                    0.358
#> Residual Std. Error      2.958 (df = 520)
#> F Statistic          59.658*** (df = 5; 520)
#> ===============================================
#> Note:               *p<0.1; **p<0.05; ***p<0.01
```

Goodness-of-fit

Before, the R^2 was **.364**! Why?

Let us have a closer look at its **formula**:

$$R^2 = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum_{i=1}^{n} \hat{u}_i^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

- The **denominator** will remain the same, but the **numerator** will, at most, remain the same.

- **Solution**: the *adjusted* R^2, $\overline{R}^2$:

$$\bar{R}^2 = 1 - \frac{\sum_{i=1}^{n} \hat{u}_i^2/(n-k-1)}{\sum_{i=1}^{n}(y_i - \bar{y})^2/(n-1)}$$

- $k =$ # independent variables;
- $(n-k-1) =$ # degrees-of-freedom.

Goodness-of-fit

Multiple regression model **without** *construc*:

R-squared	Adjusted R-squared
0.36354	0.35865

Multiple regression model **with** *construc*:

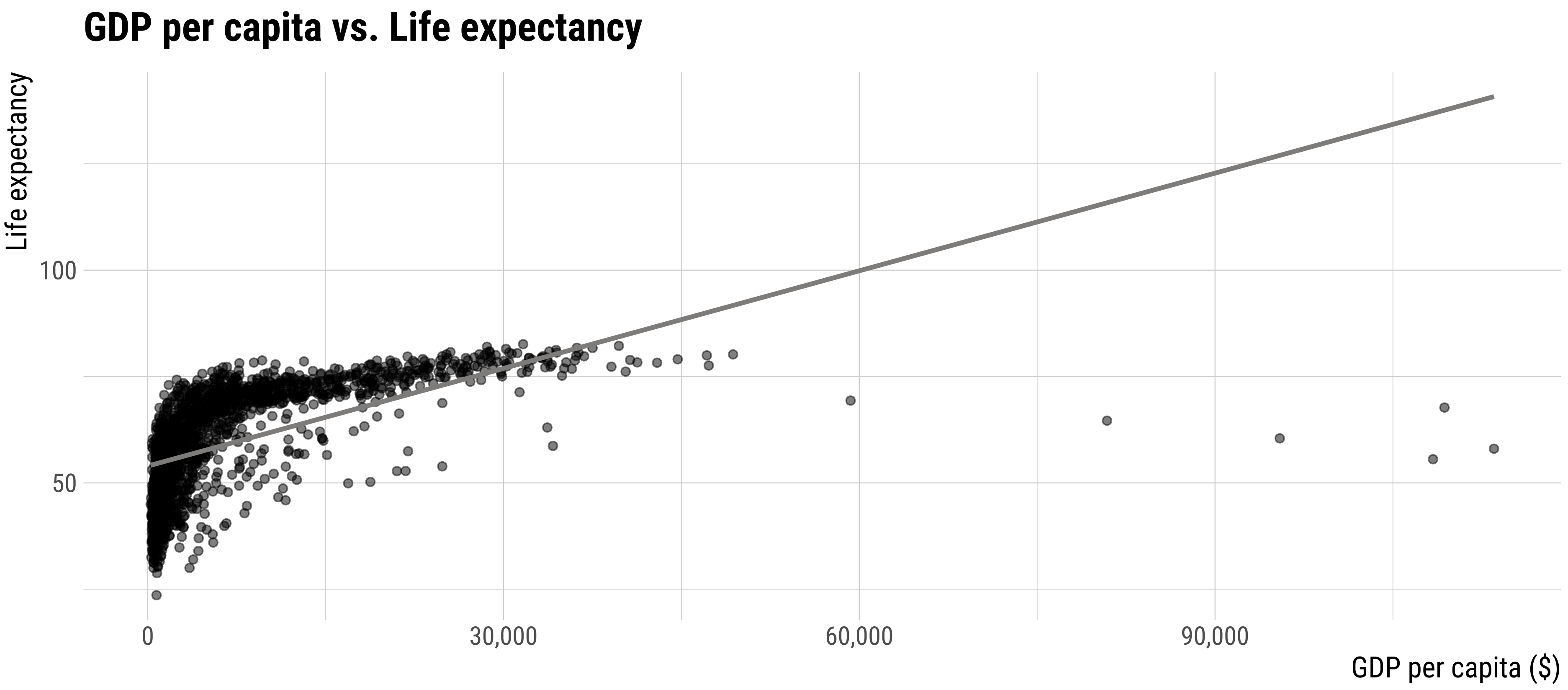
R-squared	Adjusted R-squared
0.36453	0.35842

What happened?

Functional forms

Nonlinear relationships

Many times, the relationships we are interested in **do not** follow a linear pattern.

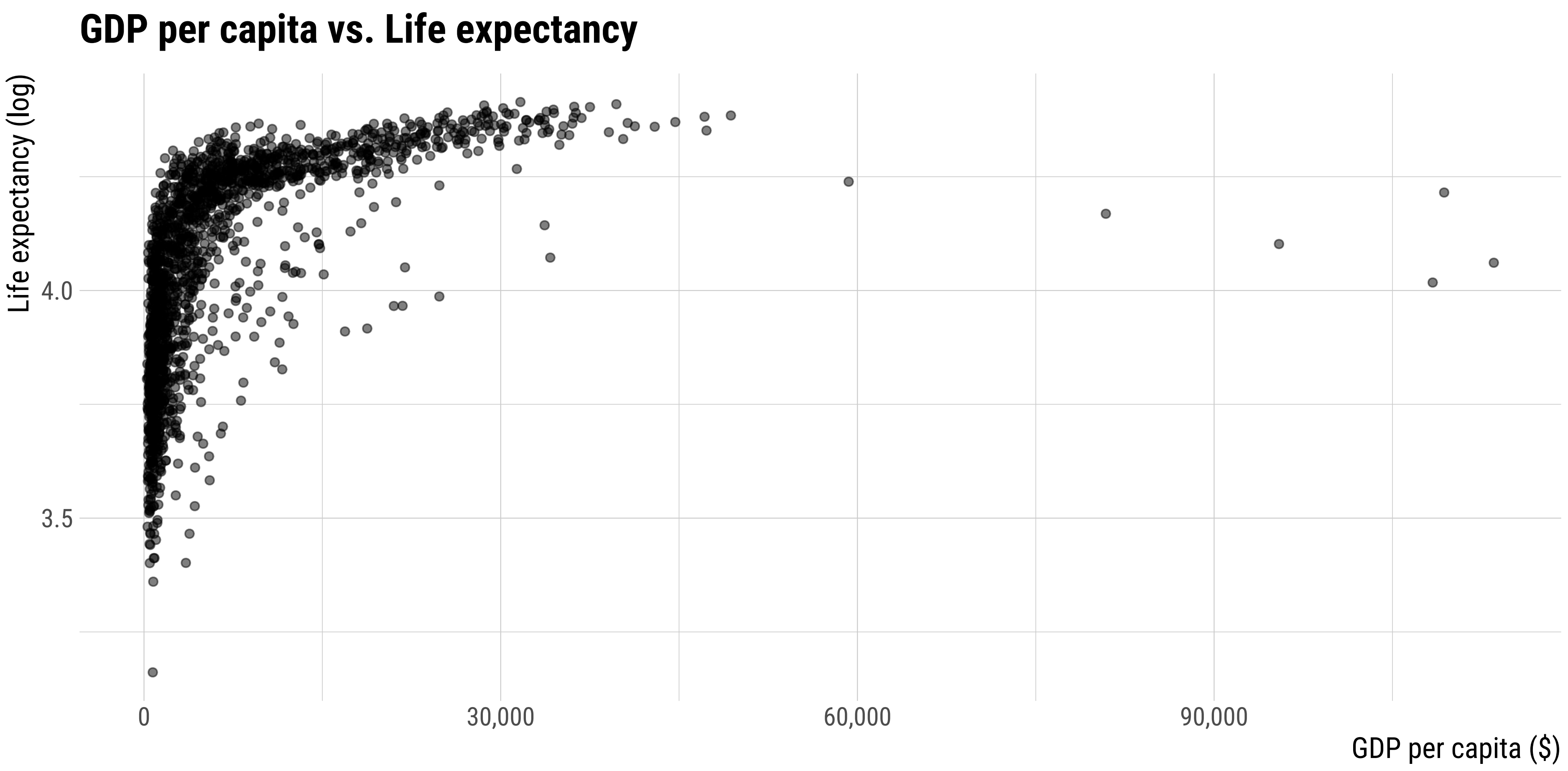


A level-level model

term	estimate	std.error	statistic	p.value
(Intercept)	53.955561	0.314995	171.29025	0
gdpPercap	0.000765	0.000026	29.65766	0

- **Interpretation:**
- A 10,000-dollar increase in GDP per capita ***increases*** life expectancy by 7.65 years.

Nonlinear relationships

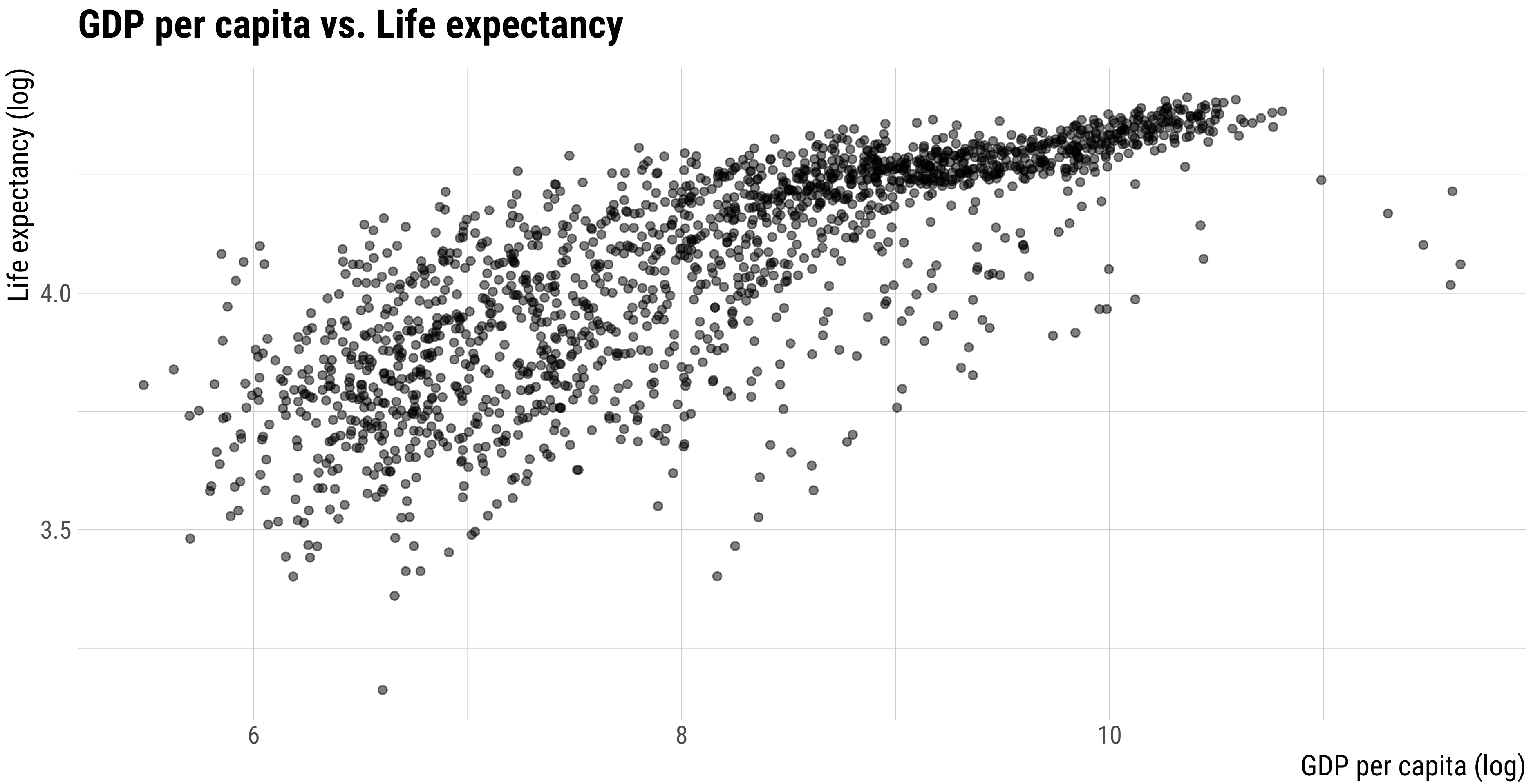


A log-level model

term	estimate	std.error	statistic	p.value
(Intercept)	3.9666387	0.0058346	679.85339	0
gdpPercap	0.0000129	0.0000005	27.03958	0

- **Interpretation:**
  - A one-unit increase in the explanatory variable increases the dependent variable by approximately $\beta_1 \times 100$ percent, on average.
  - A 1,000-dollar increase in GDP per capita ***increases*** life expectancy by 1.29%.

Nonlinear relationships

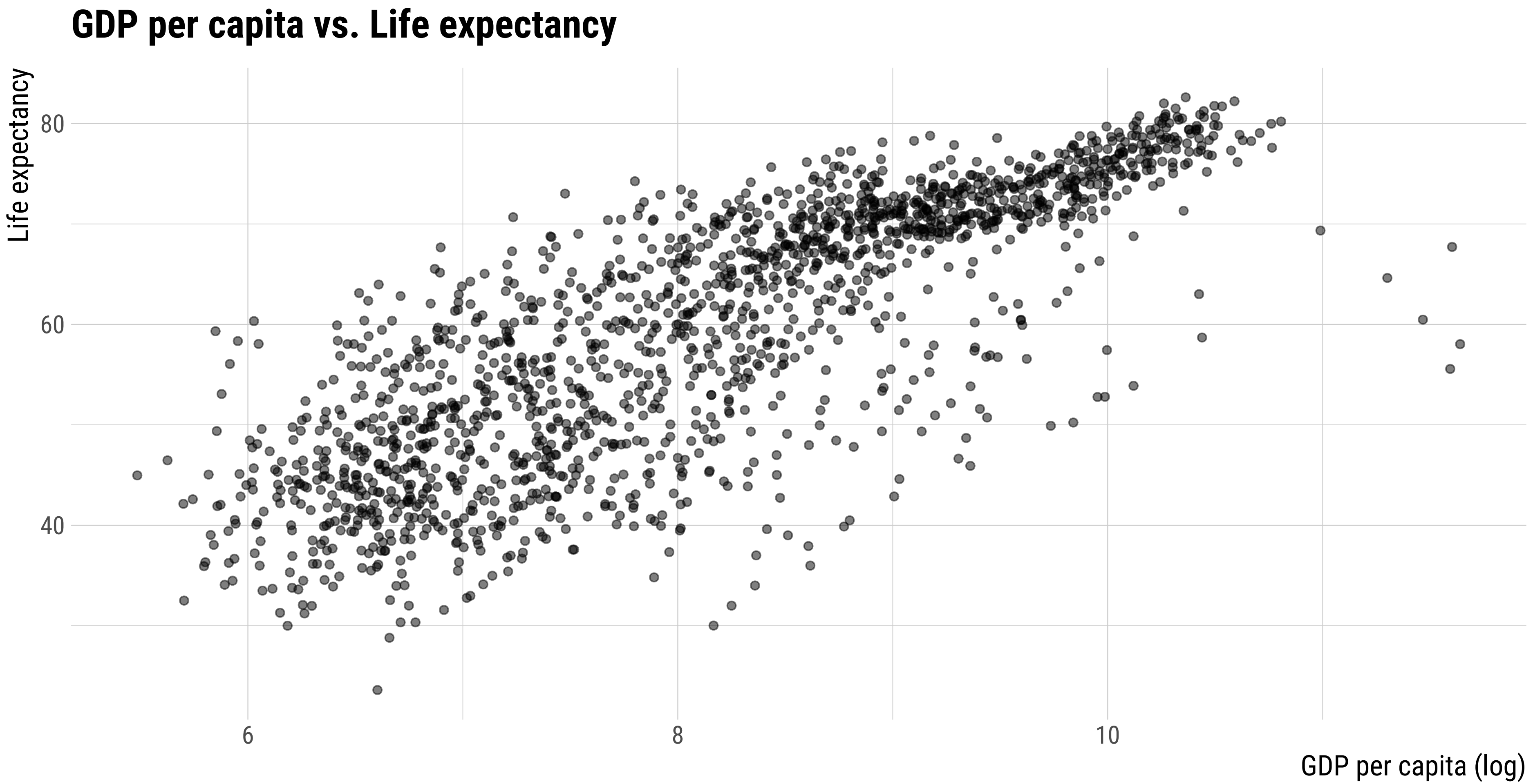


A log-log model

term	estimate	std.error	statistic	p.value
(Intercept)	2.864177	0.0232827	123.01718	0
log(gdpPercap)	0.146549	0.0028213	51.94452	0

- **Interpretation:**
  - A one-percent increase in the independent variable results in a β_1 percent change in the dependent variable, on average.
  - A 1 % increase in GDP per capita ***increases*** life expectancy by 0.147 %.

Nonlinear relationships



A level-log model

term	estimate	std.error	statistic	p.value
(Intercept)	-9.100889	1.227674	-7.413117	0
log(gdpPercap)	8.405085	0.148762	56.500206	0

- **Interpretation:**
  - A one-percent change in the independent variable leads to a $\beta_1 \div 100$ change in the dependent variable, on average.
  - A 1 % increase in GDP per capita ***increases*** life expectancy by 0.0841 years.

Quick summary

A nice interpretation reference*

Model's functional form	How to interpret β_1?
Level-level $y_i = \beta_0 + \beta_1 x_i + u_i$	$\Delta y = \beta_1 \cdot \Delta x$ *A one-unit increase in x leads to a β_1-unit increase in y*
Log-level $\log(y_i) = \beta_0 + \beta_1 x_i + u_i$	$\%\Delta y = 100 \cdot \beta_1 \cdot \Delta x$ *A one-unit increase in x leads to a $\beta_1 \cdot 100$-percent increase in y*
Log-log $\log(y_i) = \beta_0 + \beta_1 \log(x_i) + u_i$	$\%\Delta y = \beta_1 \cdot \%\Delta x$ *A one-percent increase in x leads to a β_1-percent increase in Y*
Level-log $y_i = \beta_0 + \beta_1 \log(x_i) + u_i$	$\Delta y = (\beta_1 \div 100) \cdot \%\Delta x$ *A one-percent increase in x leads to a $\beta_1 \div 100$-unit increase in y*

The meaning of linear regression

If we are able to use these nonlinear functional forms, what does *linear* regression mean after all?

- As long as the model remains **linear in parameters**, it will be linear.
- This means that we cannot **mess around** with our β coefficients!
- **Examples**:

$$log(wage_i) = \beta_0 + \beta_1 educ_i + \beta_2 exper_i + \beta_3 tenure_i + \beta_4 gender_i + u_i$$

$$log(wage_i) = \beta_0 + log(\beta_1) educ_i + \beta_2 exper_i + \beta_3^2 tenure_i + \beta_4 gender_i + u_i$$

- Which one is **not** linear in parameters?

Next time: Multiple Regression in practice