Lecture 7

2023-08-11

```
library(gapminder)
lm(lifeExp ~ gdpPercap, data = gapminder)
```

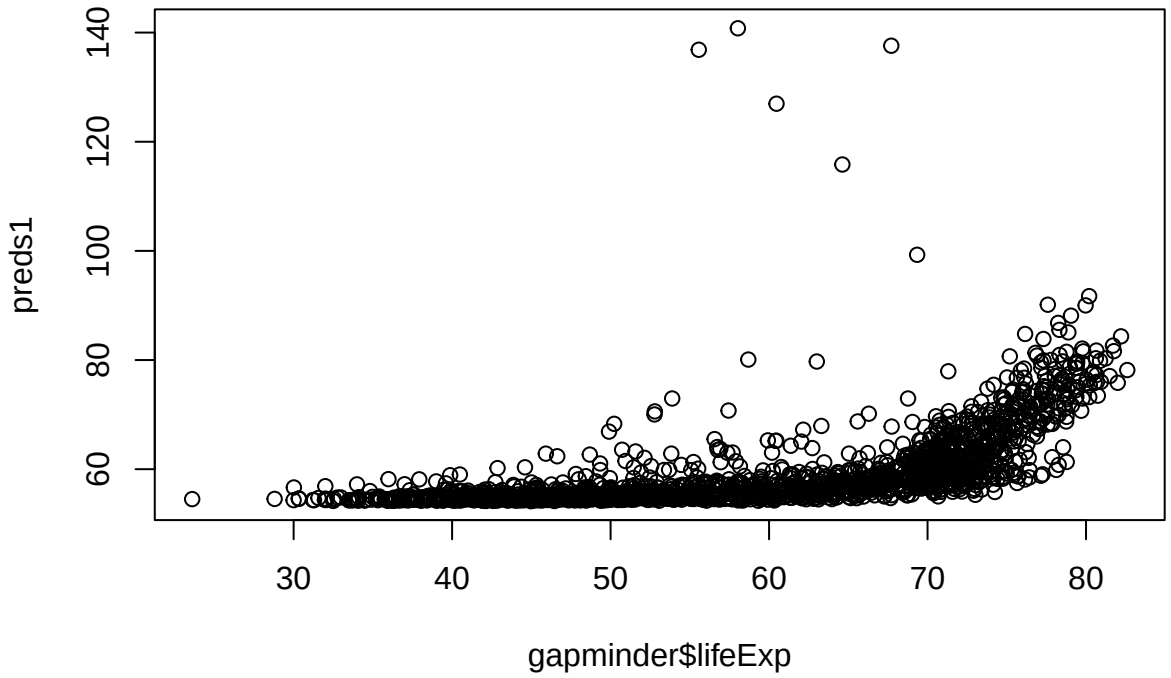
```
##
## Call:
## lm(formula = lifeExp ~ gdpPercap, data = gapminder)
##
## Coefficients:
## (Intercept)    gdpPercap
##   5.396e+01    7.649e-04
```

So we know

$$\text{lifeExp} \approx \text{gpdPercap} \times 7.649 \times 10^{-4} + 5.396 \times 10^{1}$$

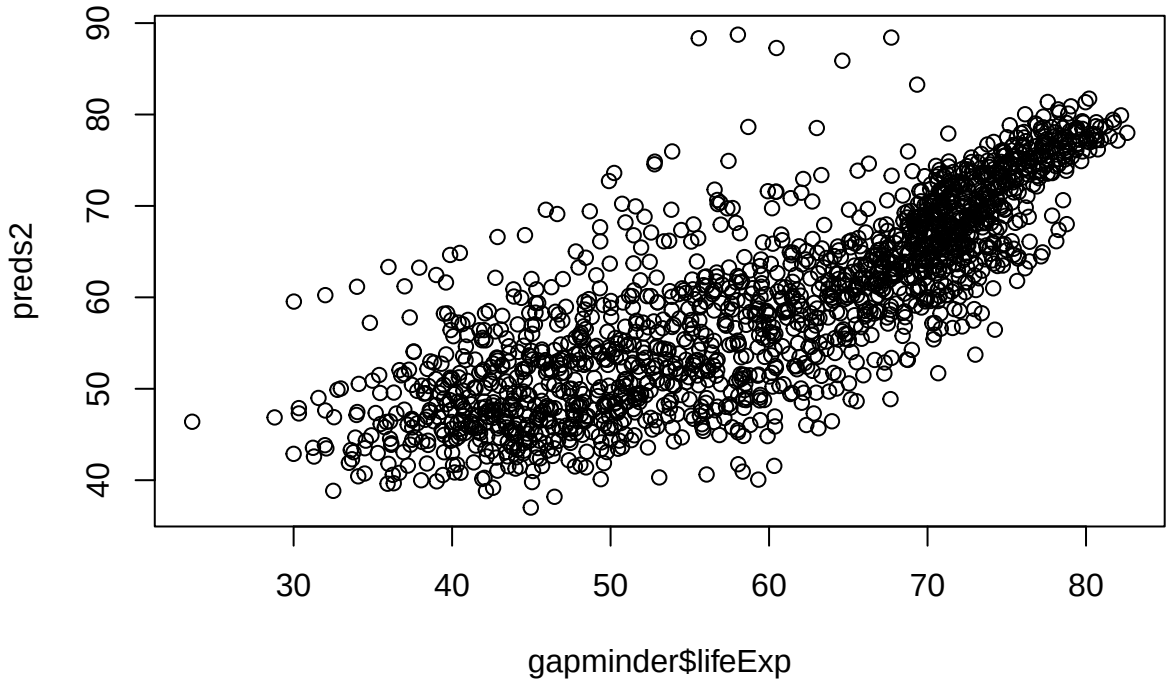
We can compare our predictions with the actual data.

```
model1 <- lm(lifeExp ~ gdpPercap, data = gapminder)
preds1 <- predict(model1, newdata = gapminder)
plot(gapminder$lifeExp, preds1)
```



Let's try predicting the logarithm of `lifeExp`.

```
model2 <- lm(lifeExp ~ log(gdpPercap), data = gapminder)
preds2 <- predict(model2, newdata = gapminder)
plot(gapminder$lifeExp, preds2)
```



This seems to be much better! Let's see which minimises error squared.

```
sum((gapminder$lifeExp - preds1)**2)
```

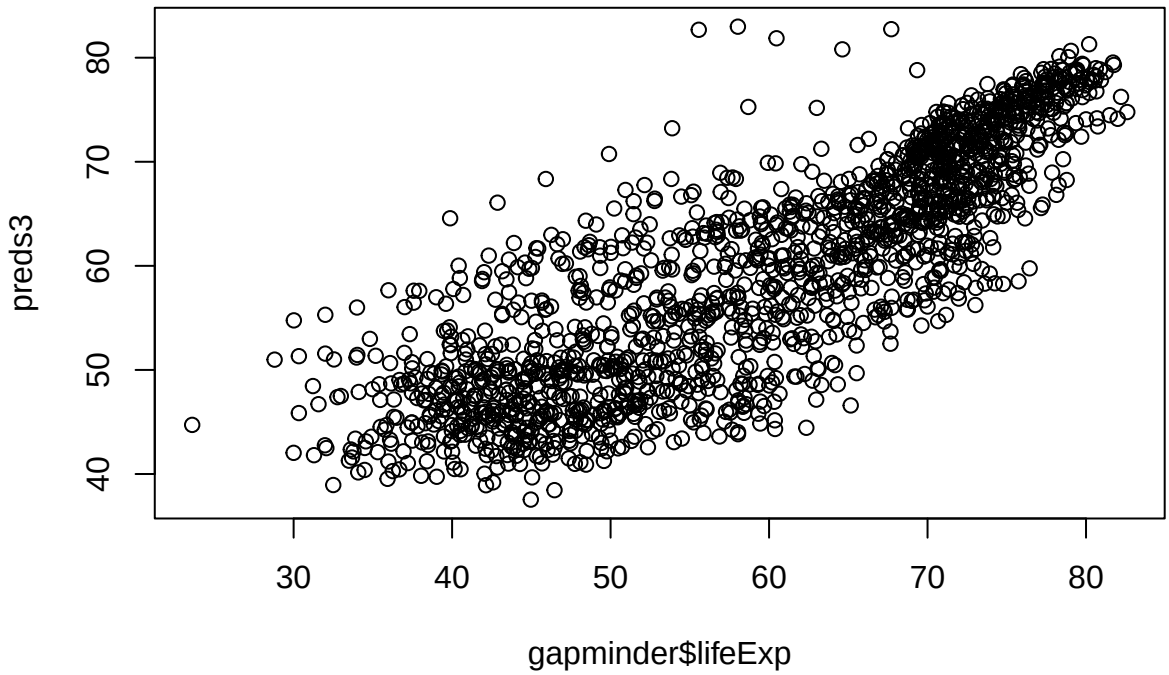
```
## [1] 187335.3
```

```
sum((gapminder$lifeExp - preds2)**2)
```

```
## [1] 98813.56
```

Let's make this even better. We can account for continent.

```
model3 <- lm(lifeExp ~ log(gdpPercap) + continent, data = gapminder)
preds3 <- predict(model3, newdata = gapminder)
plot(gapminder$lifeExp, preds3)
```




```
sum((gapminder$lifeExp - preds3)**2)
```

```
## [1] 84112.54
```

We can even sort by country, but the fit may be too specific, i.e. you can't use it to predict data of a new country not included in the gapminder dataset.

Categorial Variables

Here, `continent` is a categorial variable. It is not a number. In generating a linear model, we look for

$$y = ax + f(b)$$

where b is the category and x is the independent variable.

Titanic Dataset

```
library(titanic)
```

We use 0 to represent a dead passenger, and 1 for an alive one. Since the predicted value is unlikely to be either, and may often exeed the range, we use the *sigmoid function* `plogis()` to normalise the data.

$$\sigma(y) = \frac{1}{1 + e^{-y}}$$

The probability is then

$$p = \sigma(a_0 + a_1 x)$$

```
model_titanic <- glm(Survived ~ Sex + Age + Pclass, family = "binomial", data = titanic_train)
mean((plogis(predict(model_titanic, newdata = titanic_train)) > 0.5) == titanic_train$Survived, na.rm =
```

```
## [1] 0.7885154
```

Now the death rate is

```
1 - mean(titanic_train$Survived)
```

```
## [1] 0.6161616
```

which is less than the accuracy of our model. This is good, or else a better model is just to assume everyone dies.

For perfect prediction, you only need the names of the passengers, but then your model is useless and has no applications and you should be ashamed of yourself.