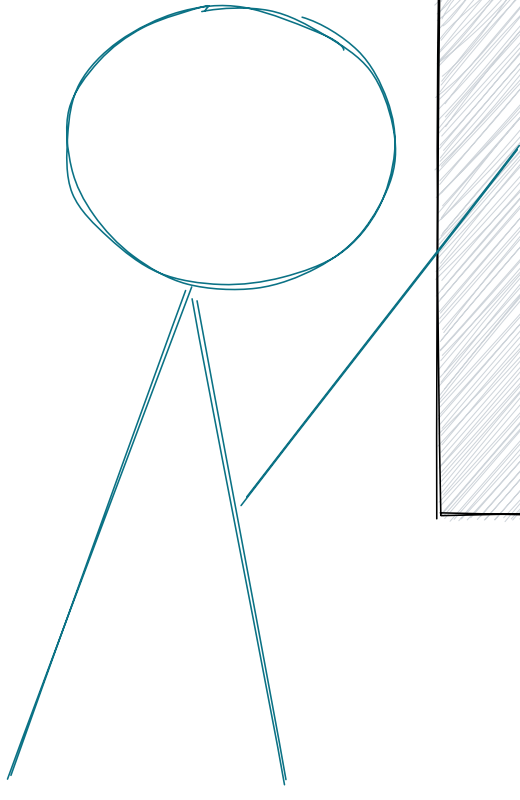


Multiple linear regression

EBA3500: Data analysis with programming

Lecturer: Jonas Moss
Contact: jonas.moss@bi.no

Office hours: 10-11 on Tuesdays.

Agenda

1. What's a regression model?
2. Multiple regression
3. Live coding multiple regression
4. Quadratic regression + live coding
5. Summary

For readers: This lecture include live coding; please see the lecture notes.

Definition: Regression model

> Let X be covariates, which may be *vector-valued*, and Y a real-valued response. Then a regression model for Y is

$$Y = f(X; \beta) + \epsilon,$$

where $\beta = (\beta_0, \beta_1, ..., \beta_p)$ is an unknown parameter vector ϵ is an error term, and f is the regression function.

We use the word *model* in two related sense.

- Like Bob, we may believe the model is actually true.
- Like Alice, we may *pretend* that it is true. The we minimize some distance such as the squared error or a proper scoring rule, or use maximum likelihood.

A regression model is true if the data (X,Y) has been generated according to the regression model!

Example: Simple linear regression

In the simple linear regression model we have

$$Y = a + bX + u,$$

which may be written as

$$Y = f(x; \beta) + \epsilon,$$

with

- $f(x; \beta) = \beta_0 + \beta_1 x,$
- $\epsilon = u,$
- $a = \beta_0, b = \beta_1.$

Example: Simple non-linear regression

Recall the non-linear least squares routine, where we minimized

$$\sum_{i=1}^{n}(y_i - f(x_i;\beta))^2$$

for some function $f(x_i;\beta)$ and $x_1, ..., x_n$ were real-valued data. The regression model associatiated with this problem is $Y = f(X;\beta) + \epsilon$.

In some cases, such as $f(x;\beta) = \frac{1}{1+e^{-\beta_0-\beta_1 x}}$, the non-linear regression model has its own name, simple *logistic regression* in this instance.

Data generating mechanisms and so on.

```
# Do not run!
import numpy as np
rng = np.random.default_rng(313)
x = np.linspace(-1, 1, 100) # Doesn't have to be this!
epsilon = rng.uniform(-1, 1, 100) # Doesn't have to be this!
y = f(x, beta) + epsilon
```

Can be ANY function of beta and x!

Multiple linear regression

These are linear regressions with more than one covariate, i.e., multiple covariates.

Definition: Multiple linear regression

Let $x_1, x_2, \ldots, x_p$ be p covariates. The multiple linear regression model of Y on x is

$$Y = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p + \epsilon.$$

The parameters β are often called *regression coefficients*.

Estimated coefficients

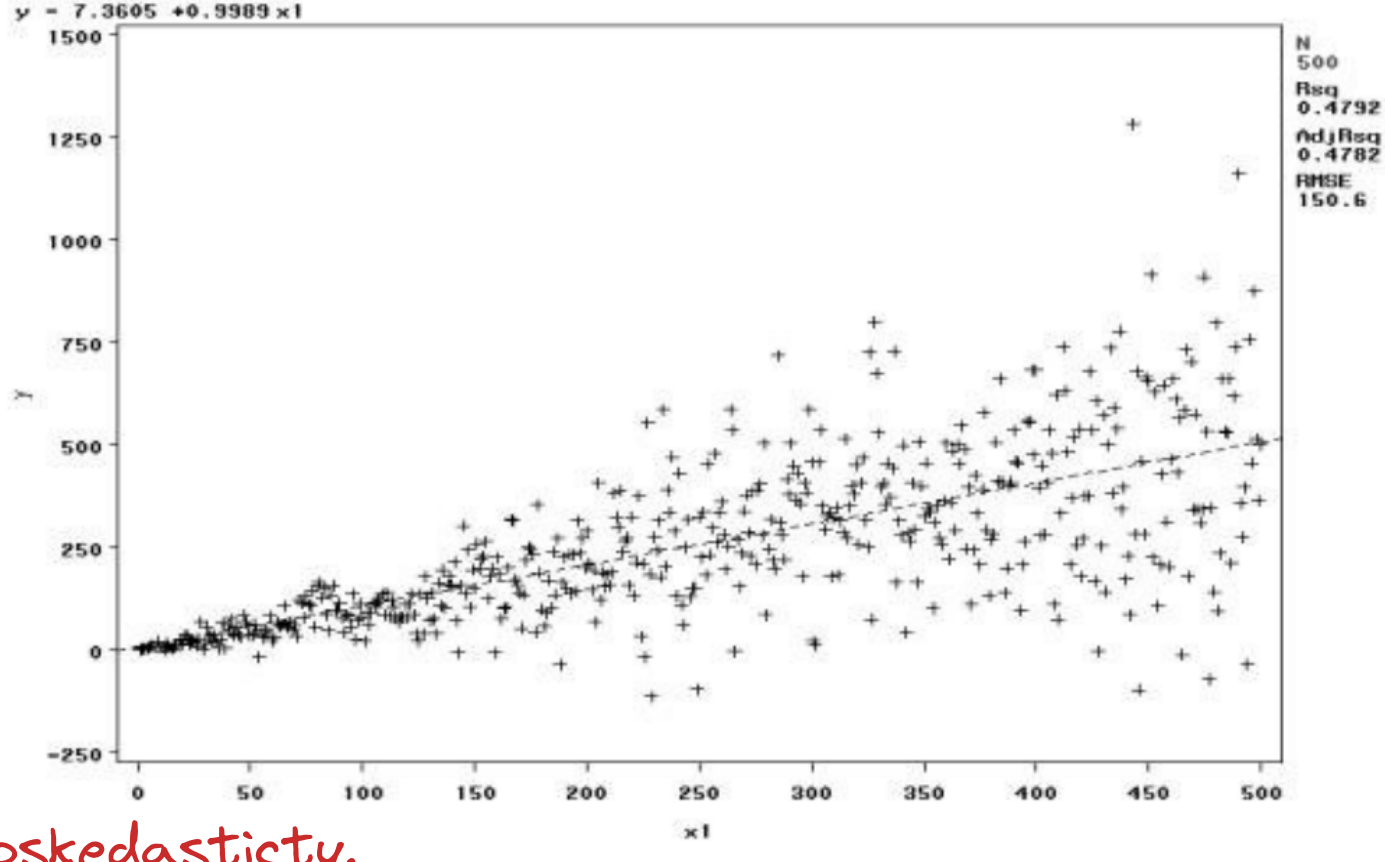
* Can estimate coeffients using e.g. least squares.
* These coefficients are called estimated regression coefficients.
* They are denoted with a hat, i.e., $\hat{\beta}_i$.

So we have both population regression coefficients and estimated coefficients!

Conditions for multiple linear regression

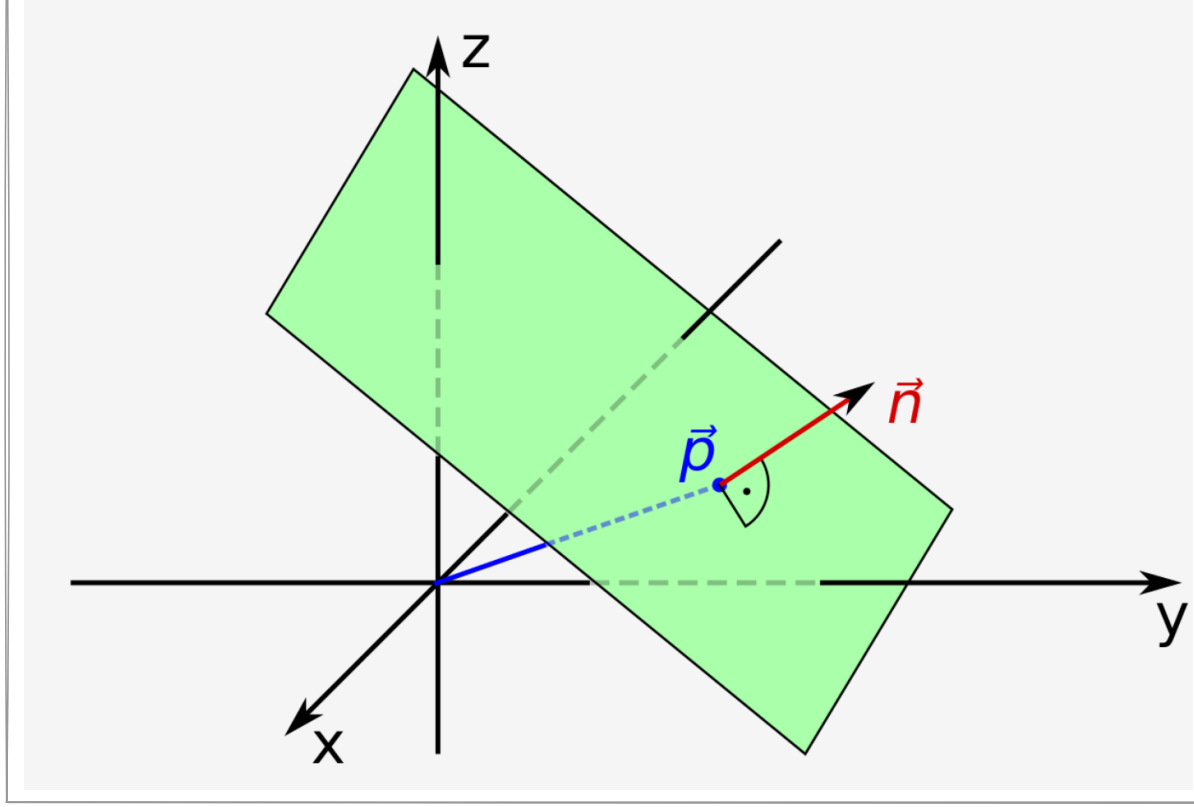
1.) The linear model is true.
2.) The errors are independent of each other.
3.) The errors have constant variance -- this is called homoskedasticity.

If the errors do not have constant variance we're dealing with heteroskedasticity.



Heteroskedasticty.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$



This defines a plane in two d dimensions!

We'll get to a practical example soon.

(With one covariate it's a line; with more than two, a hyperplane.)

Definition of plane: "A flat, two-dimensional surface that extends infinitely far."

We care about the geometric intuition since it helps us understand what we are doing!

Multiple regression with two covariates:

Find the plane that fits the data best.

Time for live coding!

* Will use statsmodels.formula.api
* The data set "marketing" from the R package "datarium".
* Will use a variety of techniques here.

Data description

"A data frame containing the impact of three advertising medias (youtube, facebook and newspaper) on sales. Data are the advertising budget in thousands of dollars along with the sales (in thousands of units). The advertising experiment has been repeated 200 times.

This is a simulated data (sic!)."

R^2 again

The R^2 for multiple regression is sometimes called the multiple R^2. It is defined as

$$1 - \frac{\sum_{i=1}^{n}(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \ldots + \hat{\beta}_p x_{pi}))^2}{\sum_{i=1}^{n}(y_i - \overline{x})^2}.$$

The interpretation should be familiar by now.

Interpretation of coefficients

1. "How much does Y increase when I increase xi by 1 unit, provided the rest is constant?"

$$\beta_0 + \beta_1 x_1 + \beta_2 x_2 - (\beta_0 + \beta_1 x_1 + \beta_2 x_2') = \beta_2 (x_2' - x_2).$$

2. The derivative of the regression function with respect to xi.

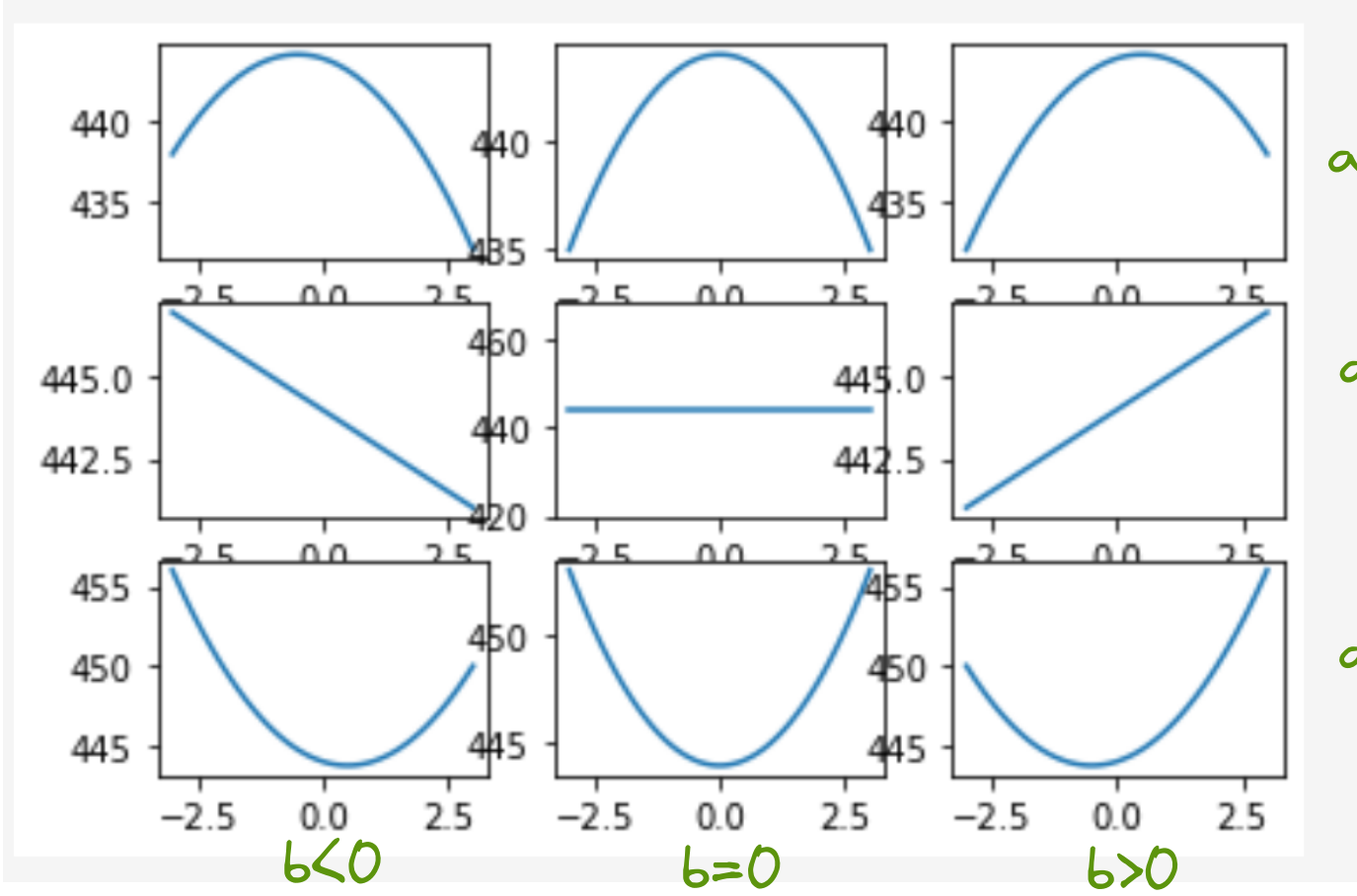
$$\frac{\partial f(x, \beta)}{\partial x_i} = \beta_i.$$

Let's look at the interpretation
in the marketing data!

Quadratic regression

Short short review of quadratics

The function $f(x) = c + bx + ax^2$ is called a *quadratic function* or *second-degree polynomial*. It takes on four kinds of shapes, depending on the value of a, b, c.


a<0
a=0
a>0
b<0
b=0
b>0

Let the true data generating model be

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \epsilon.$$

This a *quadratic regression model*. It can be regarded as a kind of non-linear regression model, but it usually isn't.

Define $X_1 = X$ and $X_2 = X^2$. Then

$$\beta_0 + \beta_1 X + \beta_2 X^2 = \beta_0 + \beta_1 X_1 + \beta_2 X_2,$$

which implies that the quadratic regression model is actually a multiple linear regression model!

Quadratic regression allows you to be more flexible when modelling univariate relationssships!

But should you use it?

(Btw, it's super common to use quadratic regression for univariate relationssships in the social sciences.)

1. Quadratic functions have a very specific shape.

2. But few natural phenomena adhere to this shape - except if you know they do, as in physics.

3. Often used by social scientists when data isn't completely linear. And if the relationship between y and x isn't linear, a quadratic regression will always fit better in terms of e.g. the R^2.

4. Sometimes relationsships "flatten out", and the quadratic curve will give a wrong impression.

Let's take a look at some data from the following article, published in the highly prestiguous Psychological science.

> Swaab, R. I., Schaerer, M., Anicich, E. M., Ronay, R., & Galinsky, A. D. (2014). The too-much-talent effect: team interdependence determines when more talent is too much or not enough. Psychological Science, 25(8), 1581–1591. https://doi.org/10.1177/0956797614537280

This article is partly about football, reportedly "a simple game; 22 men chase a ball for 90 minutes and at the end, the Germans win."

> Five studies examined the relationship between talent and team performance. Two survey studies found that people believe there is a linear and nearly monotonic relationship between talent and performance: Participants expected that more talent improves performance and that this relationship never turns negative. However, building off research on status conflicts, we predicted that talent facilitates performance—but only up to a point, after which the benefits of more talent decrease and eventually become detrimental as intrateam coordination suffers.

So the authors claim there is no increasing relationsship between talent and performance at the top level. That seems plausible considering e.g. Martin Ødegaard!

More live coding!

Moral of the story?

1. Always look at the data. If it doesn't look like it supports your hypothesis, it probably does not.
2. Try out different models, some might fit much better the others.
3. Do not blindly trust the qualitative consequences of models. Even if the quadratic model had a better fit than the log-model, it wouldn't provide strong evidence for a U-shape - for a quadratic curve must be U-shaped!

It's often hard to do better than a log-transform!

1. A regression model is on the form $Y = f(X; \beta) + \epsilon$ for some regression function f.
2. In the multiple linear regression model, X is vector, and the regression function is $f(X; \beta) = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p$.
3. There are two interpretations of the regression coefficients β_i in a linear regression model.
4. The multiple linear regression function may be interpreted as the equation for a line when $p = 1$, a plane when $p = 2$, and a hyperplane when $p > 2$.
5. The quadratic regresion model $Y = c + bX + aX^2 + \epsilon$ is an example of a multiple linear regression model. It is frequently used due to its simplicity, but should be treated with suspicion.
6. `statsmodels.formula.api` may be used to calculate multiple linear regression models.

Summary!