

# Lecture 2 – Linear Regression and Extensions

Martin Spindler

2016-04-26

# Linear Regression

We start with a linear regression model:

$$y_i = x_i' \beta + \varepsilon_i, i = 1, \dots, n,$$

where  $x_i$  is a  $p$ -dimensional vector of regressors for observation  $i$ ,  $\beta$  a  $p$ -dimensional coefficient vector, and  $\varepsilon_i$  iid error terms with  $\mathbb{E}[\varepsilon_i | x_i] = 0$ .

The ordinary least squares (ols) estimator for  $\beta$  is defined as

$$\hat{\beta} = \arg \min_{\beta \in \mathbb{R}^p} \sum_{i=1}^n (y_i - x_i' \beta)^2.$$

# Linear Regression

If the Gram matrix  $\sum_{i=1}^n x_i x_i'$  is of full rank, the ols estimate is given by

$$\hat{\beta} = \left( \sum_{i=1}^n x_i x_i' \right)^{-1} \left( \sum_{i=1}^n x_i y_i \right).$$

The residuals  $\hat{\varepsilon}_i$  are defined as

$$\hat{\varepsilon}_i = y_i - x_i' \hat{\beta}.$$

For an observation  $x$  the *fitted* or *predicted* values are given by

$$\hat{y} = x' \hat{\beta}.$$

# Linear Regression

In matrix notation we can write

$$Y = X\beta + \varepsilon$$

with  $Y = (y_1 \dots y_n)$ ,  $\varepsilon = (\varepsilon_1 \dots \varepsilon_n)$  and  $X$  is a  $n \times p$ -matrix with observation  $i$  forming the  $i$ th row of the matrix  $X$ . The ols estimate  $\hat{\beta}$  can then be written as

$$\hat{\beta} = (X'X)^{-1}X'y.$$

# Linear Regression

Under homoscedastic errors, i.e.  $\mathbb{V}\mathcal{D} \setminus \varepsilon_i = \sigma^2$ , we have that

$$\mathbb{V}\mathcal{D} \setminus (\hat{\beta}) = (X'X)^{-1}\sigma^2.$$

Asymptotically, the ols estimate is normal distributed:

$$\hat{\beta} \sim N(\beta, (X'X)^{-1}\sigma^2).$$

This can be used for testing hypotheses and construction of confidence intervals.

# Linear Regression

$$z_j = \frac{\hat{\beta}_j}{\hat{\sigma}^2 \sqrt{v_j}}$$

where  $v_j$  is the  $j$ th diagonal element of  $(X'X)^{-1}$ .

Under the null hypothesis  $\beta_j = 0$  the *Z-score* / *t-statistic*  $z_j$  is  $t_{n-p-1}$ -distributed.

# Linear Regression

Remark: In the high-dimensional-setting, i.e.  $p \gg n$  the Gram Matrix is rank deficient and the ols estimate is not uniquely defined and the variance of the parameter estimate is unbounded.

# Extensions

- ▶ Polynomial Regression
- ▶ Step Functions
- ▶ Basis Functions
- ▶ Regression Splines
- ▶ Smoothing Splines



## Extensions | Remarks

- ▶ Although the linear regression model looks quite simple, it can be extended / modified to model complex relations.
- ▶ For the extensions we consider without loss of generality univariate regressions:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

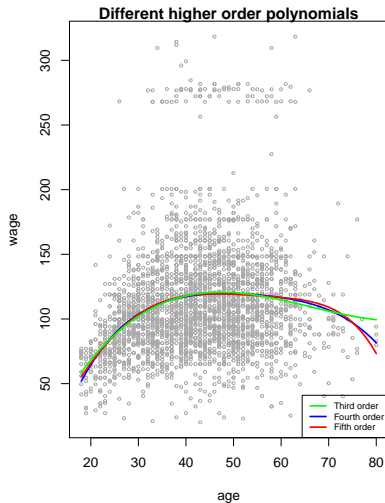
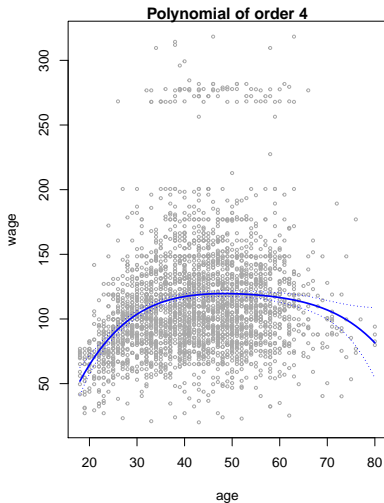
## Extensions | Polynomial Regression

To make the linear specification more flexible, we might include higher-order polynomials:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \dots \beta_p x_i^p + \varepsilon_i$$

- ▶ Estimation by ols
- ▶ Quite flexible, but usually  $p=3$  or  $p=4$
- ▶ Higher order polynomials ( $p > 5$ ) might lead to strange fits (overfitting), especially at the boundary.

# Extensions | Polynomial Regression - Example



## Extensions | Step Functions

- ▶ Definition: Step functions are functions which are constant on each part of a partition of the domain.
- ▶ Univariate Regression: choosing  $K$  cut points  $c_1, \dots, c_K$  and defining new auxiliary variables:  
 $C_0(x) = 1(x < c_1)$ ,  $C_1(x) = 1(c_1 \leq x < c_2)$ ,  $\dots$ ,  
 $C_K(x) = 1(c_K \leq x)$
- ▶  $1(\cdot)$  is the so-called indicator function which is 1 if the condition is true and 0 otherwise.
- ▶ This gives us the following regression:

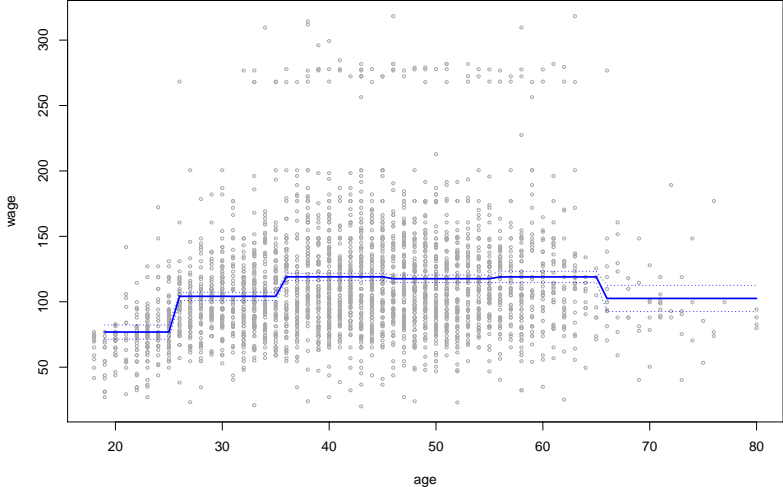
$$y_i = \beta_0 + \beta_1 * C_1(x_i) + \dots + \beta_K * C_K(x_i) + \varepsilon_i$$

## ## Extensions | Step Functions

- ▶ Note:  $C_0(x) + \dots + C_K(x) = 1$  and hence we drop  $C_0 = (\cdot)$  to avoid multicollinearity.
- ▶ Interpretation  $\beta_0$
- ▶ Example: wage regression

# Extensions | Step Functions

regression with step functions



## Extension | Basis Functions

- ▶ Idea: family of functions or transformations that can be applied to a variable:  $b_1(x), \dots, b_K(x)$  (basis functions)
- ▶ Regression:

$$y_i = \beta_0 + \beta_1 b_1(x_i) + \dots + \beta_K b_K(x_i) + \varepsilon_i$$

- ▶ Examples
- ▶ Polynomial regression:  $b_j(x_i) = x_i^j$
- ▶ Piecewise constant functions (step functions):  
 $b_j(x_i) = 1(c_j \leq x_i < c_{j+1})$
- ▶ Regressions splines (coming next)