Foundations of Data Science: Regression and inference – Generalised linear models

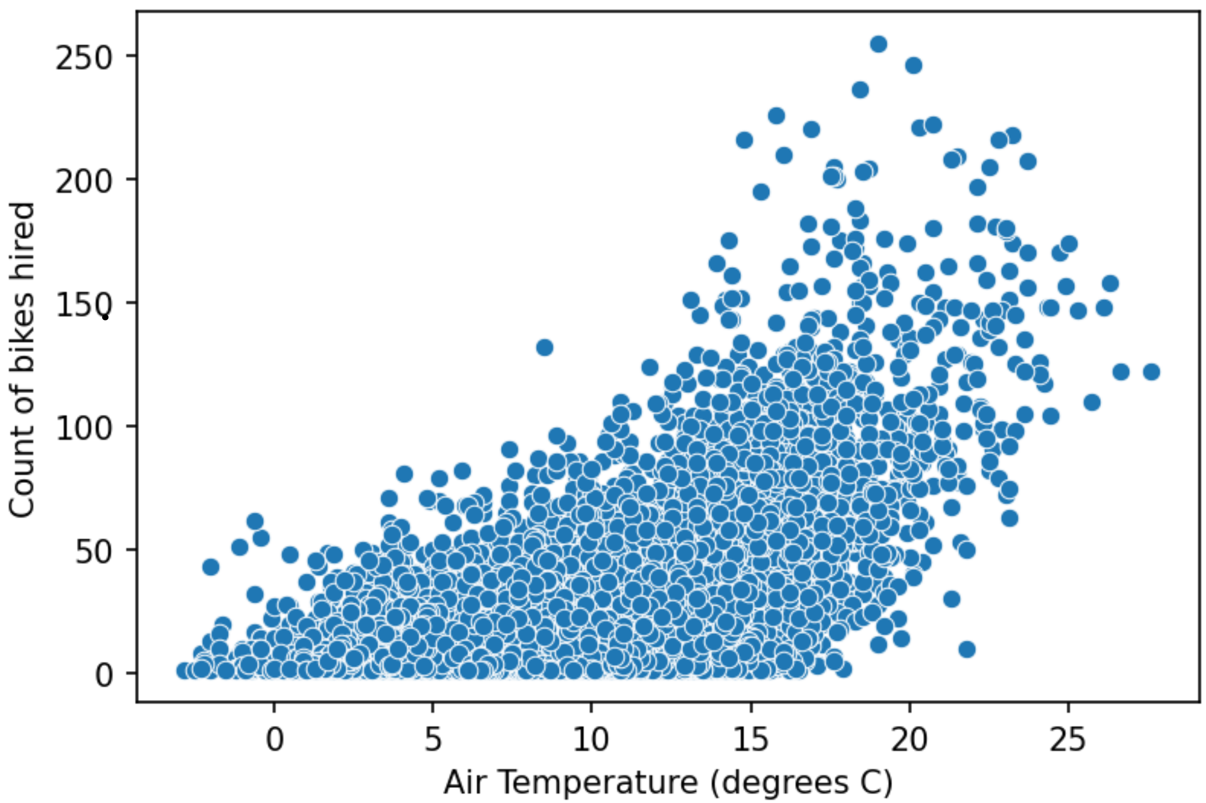
THE UNIVERSITY of EDINBURGH
informatics

FOUNDATIONS OF DATA SCIENCE

We want to investigate the relationship between the number of bikes hired in an hour and the mean temperature during that hour

Is there a problem with using ordinary least squares linear regression to do this?

Data sources:
- Edinburgh Just Eat Bikes data 2020
- Edinburgh temperature observations, Met Office via MIDAS



Overview

Monday

1. The maximum likelihood principle
2. Application of maximum likelihood principle to a simple example
3. Application of maximum likelihood principle to linear regression

Today

0. Recap + prediction uncertainty
1. Max likelihood with non-normal distributions
2. Poisson regression
3. Generalised linear regresion

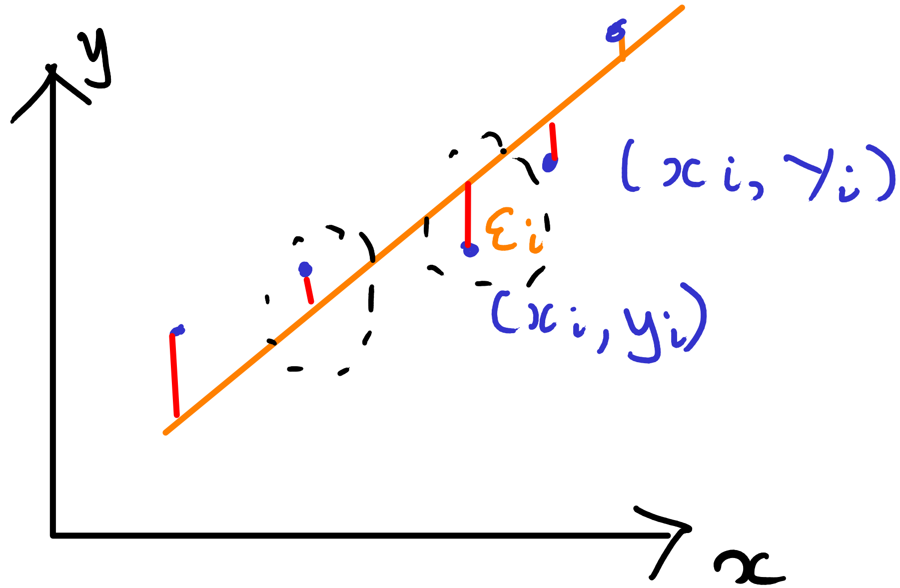
Foundations of Data Science: Regression and inference – Recap of max likelihood applied to linear regression

Application of max likelihood to linear regression

$$Y_i = \beta_0 + \beta_1 x_i + \underbrace{\varepsilon_i}_{\text{error term}}$$

$$\varepsilon_i \sim N(0, \sigma^2)$$

↑ residual



OR

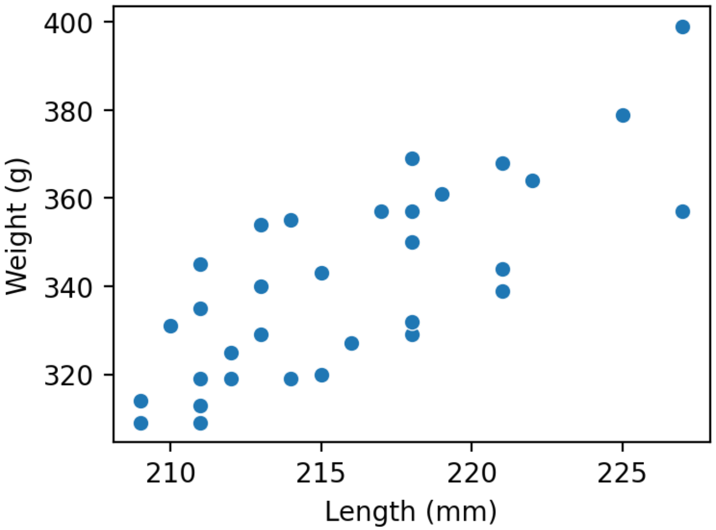
$$Y_i \sim N(\overbrace{\beta_0 + \beta_1 x_i}^{\mu}, \sigma^2)$$

$$\ln p(\underline{Y} = y_1, \dots, y_n ; x_1, \dots, x_n \mid \overbrace{\beta_0, \beta_1}^{\mu}, \sigma^2)$$

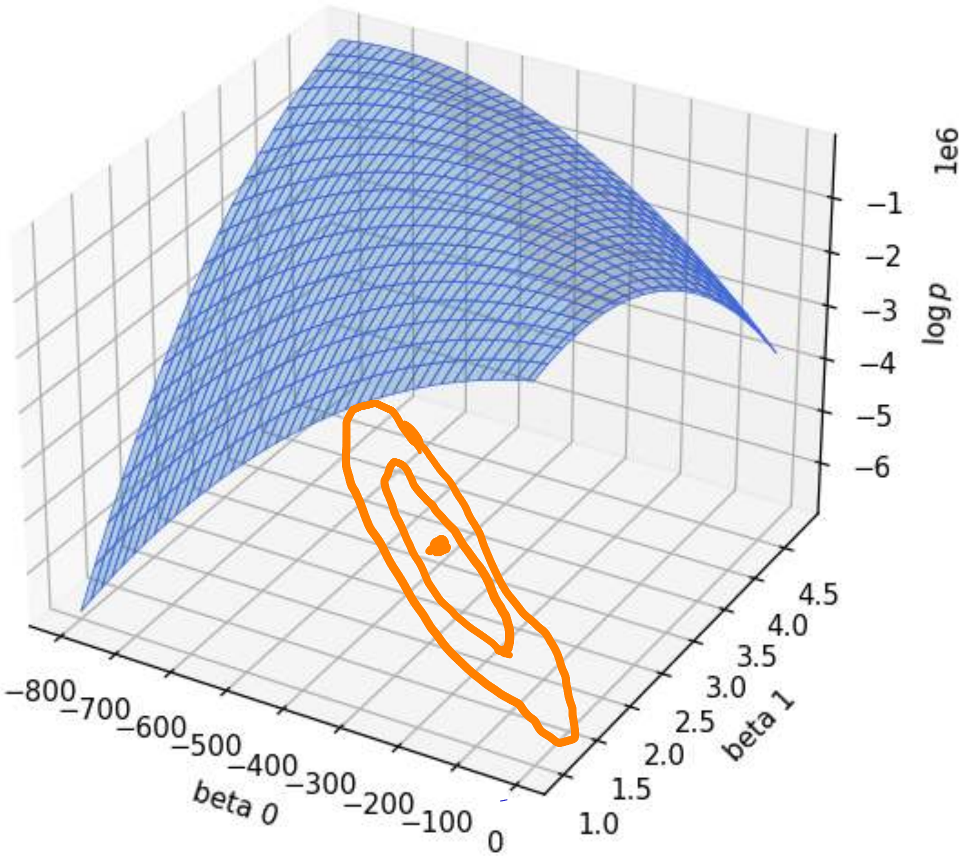
$$= \sum_{i=1}^{n} \left(-\frac{1}{2} \ln 2\pi\sigma^2 - \frac{1}{2} \left(\frac{y_i - \beta_0 - \beta_1 x_i}{\sigma^2} \right)^2 \right)$$

Log likelihood of coefficients

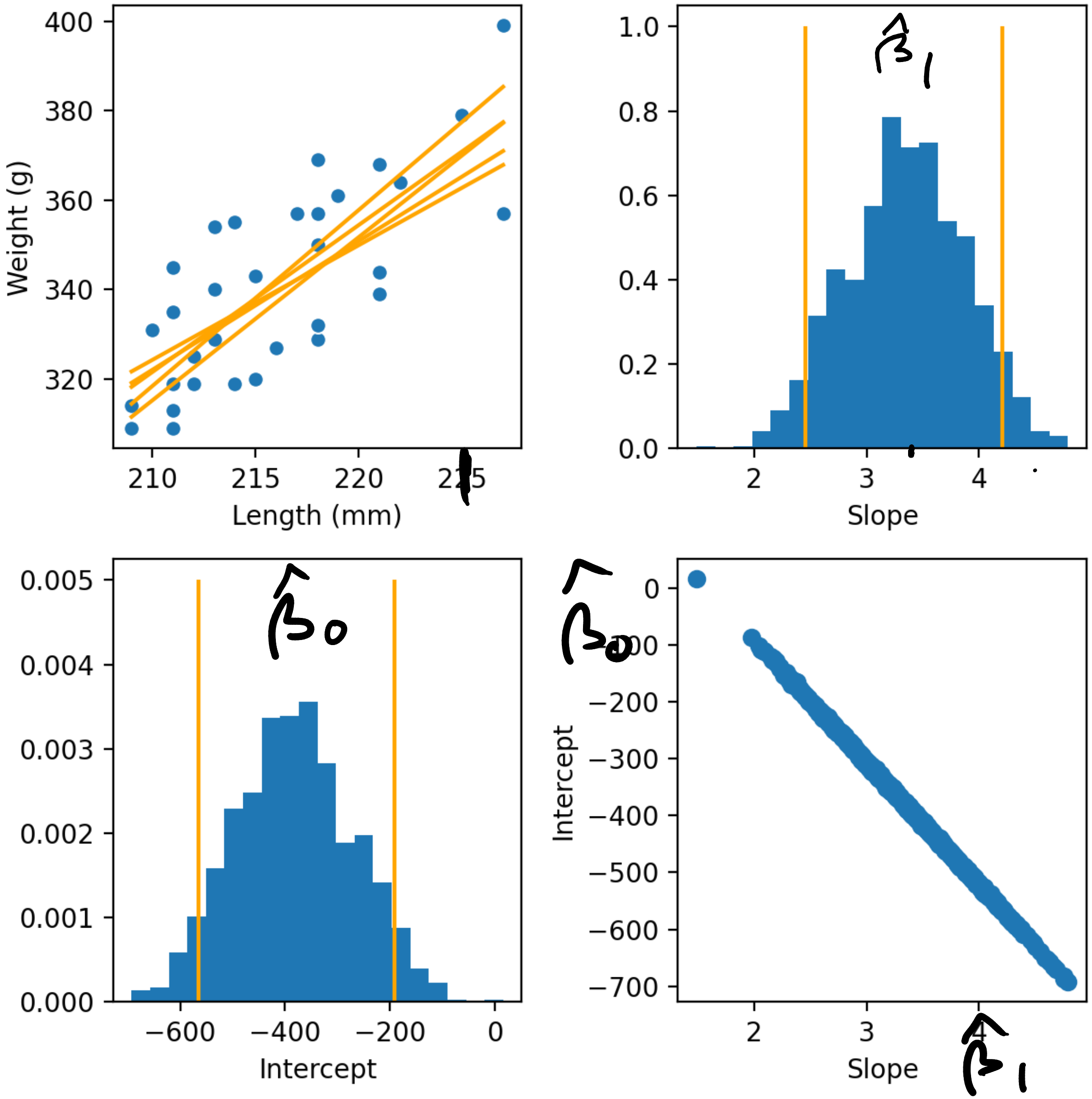
Peter Trimming, Wikimedia Commons,



Data from Wauters and Dhondt 1989

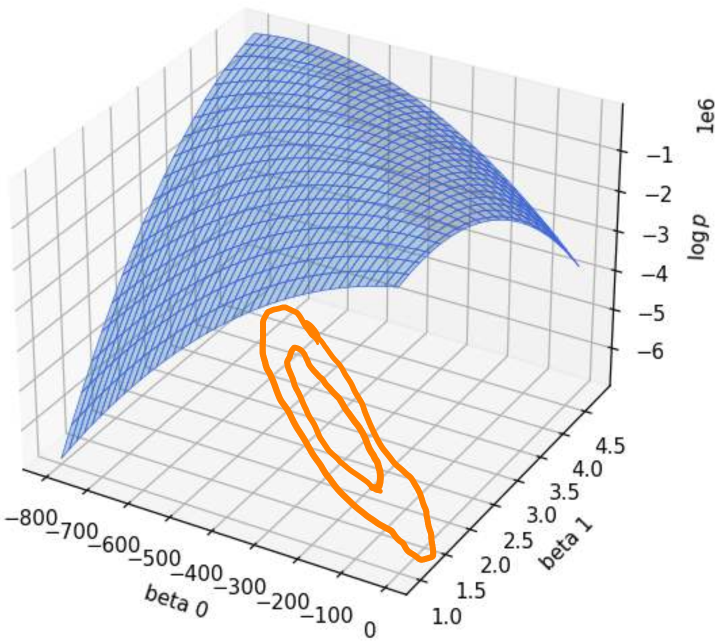


Bootstrap inference of coefficients

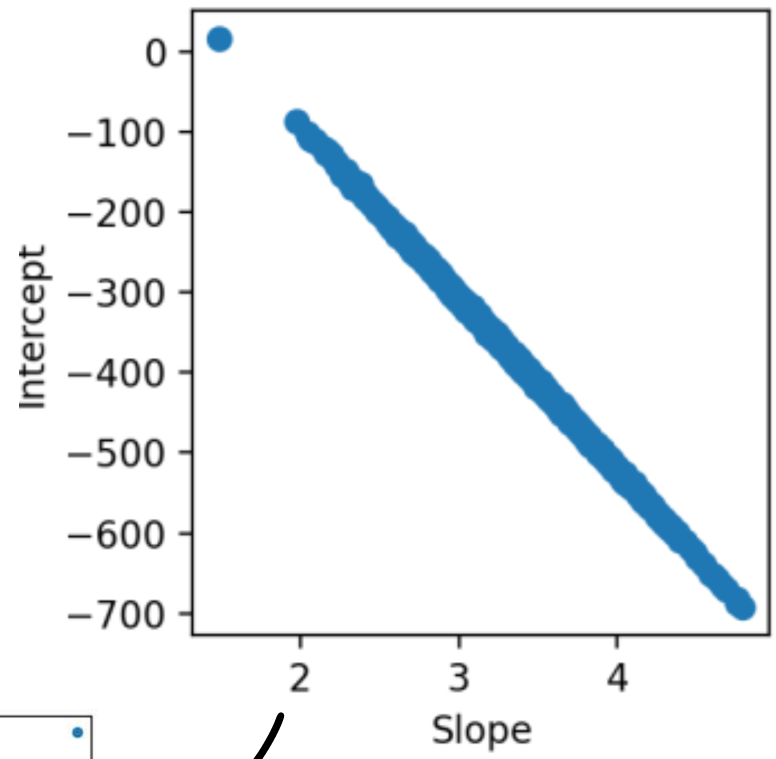


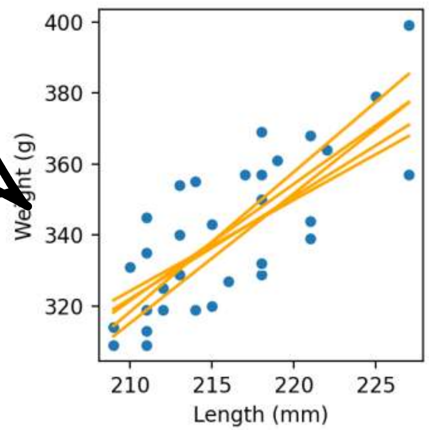
Data from Wauters and Dhondt 1989

(Log) Likelihood function

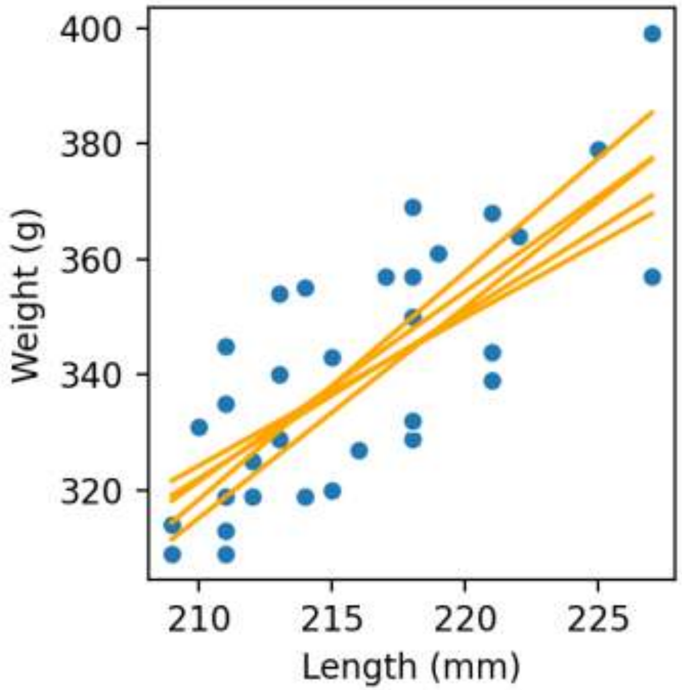


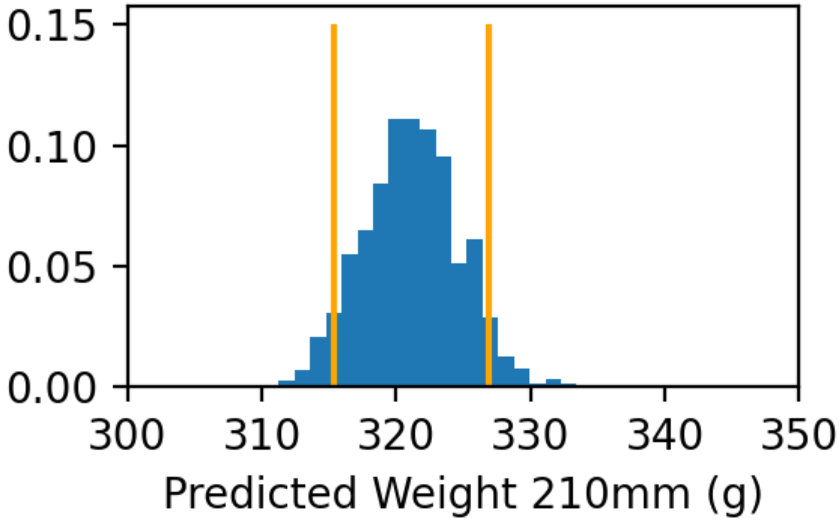
Bootstrap samples

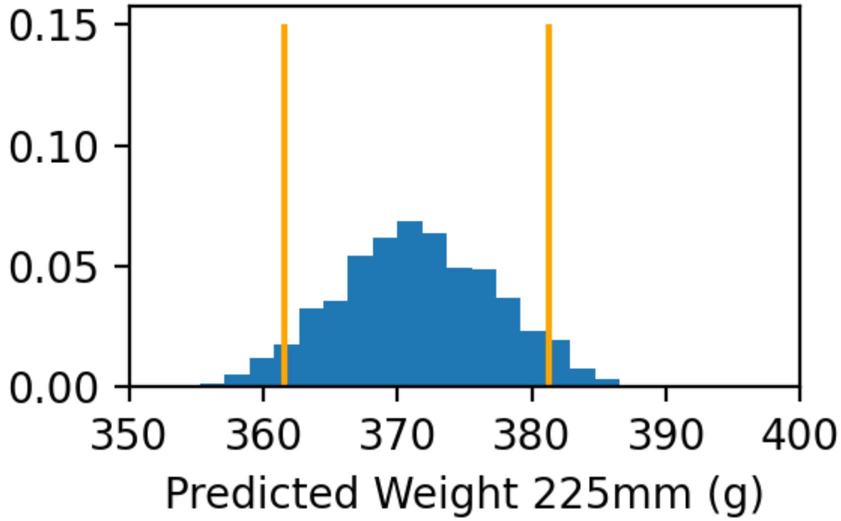




Uncertainty in predictions (with Bootstrap)







$y = \beta_0^{(1)} + \beta_1^{(1)} \cdot 210$ Bootstrap sample #1

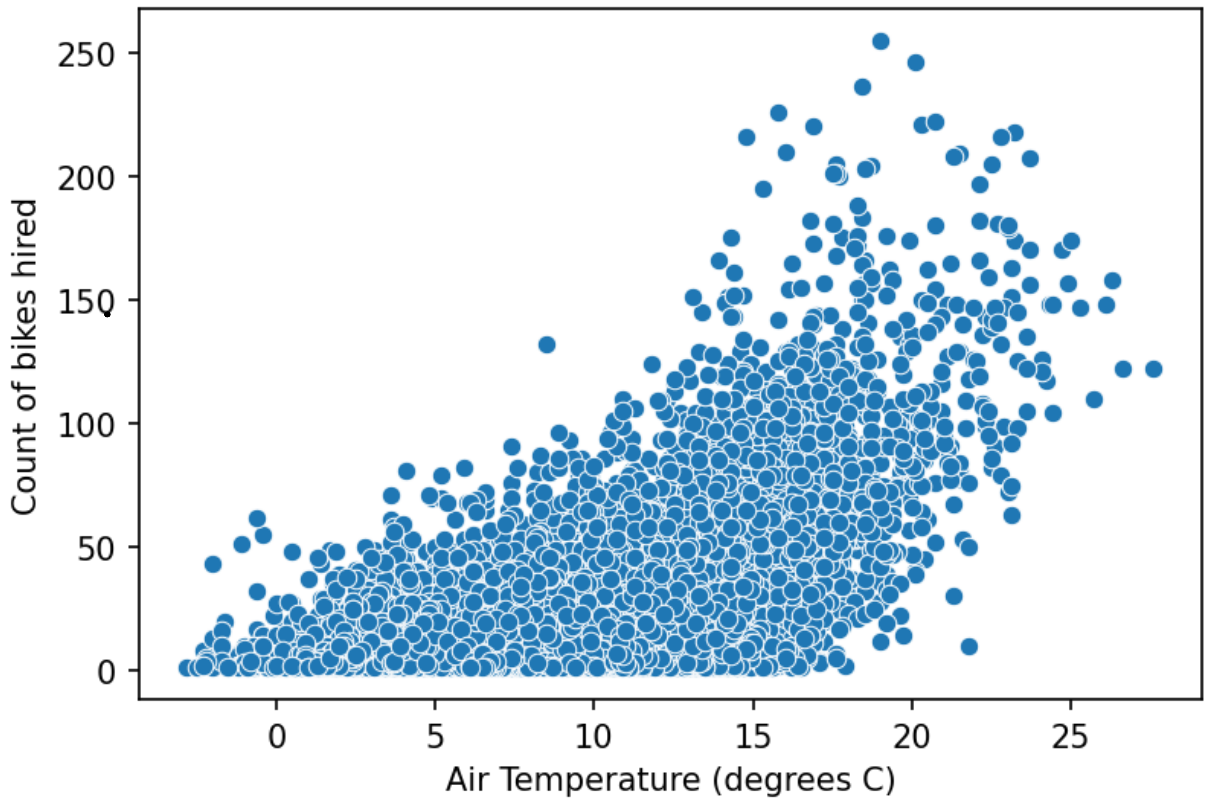
$y = \beta_0^{(2)} + \beta_1^{(2)} \cdot 210$ " " #2

$\vdots$

We want to investigate the relationship between the number of bikes hired in an hour and the mean temperature during that hour

Is there a problem with using ordinary least squares linear regression to do this?

Are there any techniques described in the course so far that could fit the data?

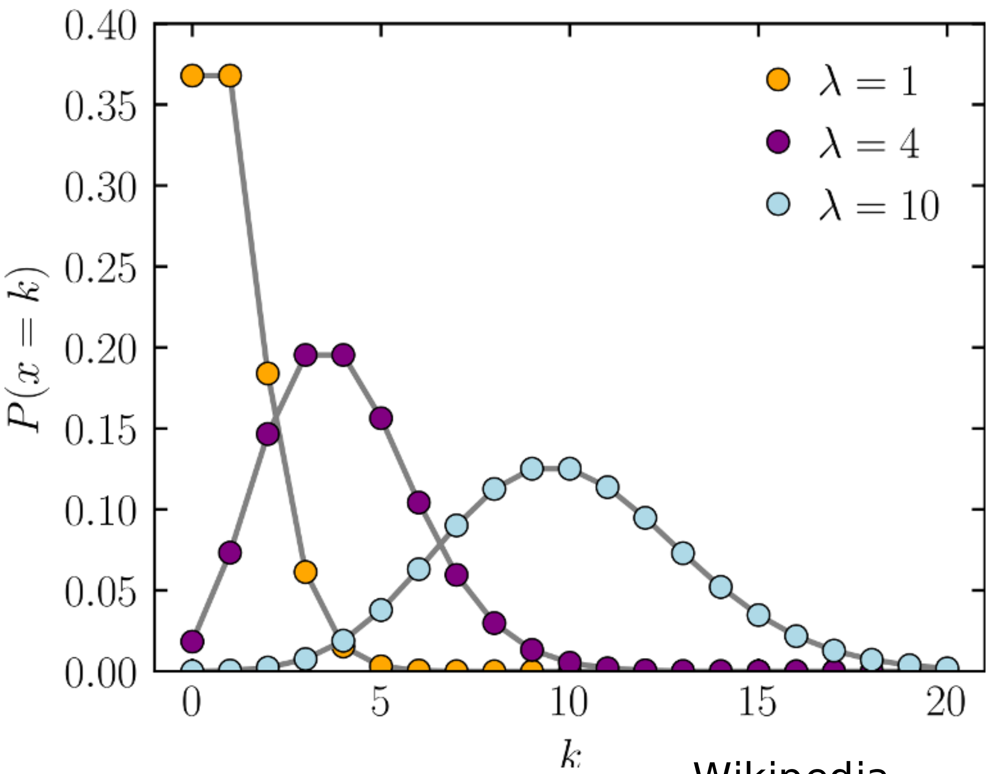


Foundations of Data Science: Regression and inference – Max likelihood of univariate non-normal distributions

Max likelihood for models other than the normal

We don't have to assume the data is normally distributed.

E.g. Poisson distribution



Wikipedia

E.g. Number of goals in World Cup football matches

Wikpedia, CC-BY-SA 3.0

Expected number of goals in a match $\lambda = 2 \cdot 5$

Number of deaths by horse kicks in the Prussian army

Wikpedia, CC-BY 2.0

	75	76	77	78	79	80	81	82	83	84	85	86	87	88	89	90	91	92	93	94
G	—	2	2	1	—	—	1	1	—	3	—	2	1	—	—	1	—	1	—	1
I	—	—	—	2	—	3	—	2	—	—	—	1	1	1	—	2	—	3	1	—
II	—	—	—	2	—	2	—	—	1	1	—	—	2	1	1	—	—	2	—	—
III	—	—	—	1	1	1	2	—	2	—	—	—	1	—	1	2	1	—	—	—
IV	—	1	—	1	1	1	1	—	—	—	—	1	—	—	—	—	1	1	—	—
V	—	—	—	—	2	1	—	—	1	—	—	1	—	1	1	1	1	1	1	—
VI	—	—	1	—	2	—	—	1	2	—	1	1	3	1	1	1	—	3	—	—
VII	1	—	1	—	—	—	1	—	1	1	—	—	2	—	—	2	1	—	2	—
VIII	1	—	—	—	1	—	—	1	—	—	—	—	1	—	—	—	1	1	—	1
IX	—	—	—	—	—	2	1	1	1	—	2	1	1	—	1	2	—	1	—	—
X	—	—	1	1	—	1	—	2	—	2	—	—	—	—	2	1	3	—	1	1
XI	—	—	—	—	2	4	—	1	3	—	1	1	1	1	2	1	3	1	3	1
XIV	1	1	2	1	1	3	—	4	—	1	—	3	2	1	—	2	1	1	—	—
XV	—	1	—	—	—	—	—	1	—	1	1	—	—	—	2	2	—	—	—	—

Bortkewitsch 1898

$\underline{Y} = (0, 2, 2, 0, \dots)$

$y_1 \quad y_2 \quad \dots \quad y_{280}$

$$n_k = \sum_{i=1}^{n} I(y_i = k)$$

k	n_k
0	144
1	91
2	32
3	11
4	2

Log likelihood calculation of Poisson distribution

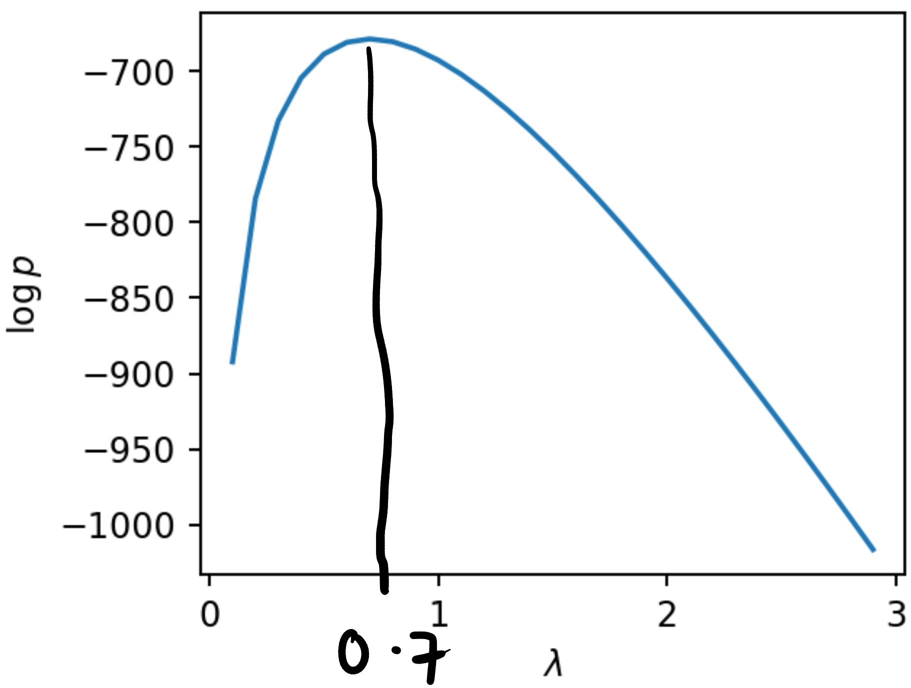
Log likelihood $\ell = \ln P(\underline{Y} = y_1, \ldots, y_n \mid \lambda)$

$$l = \ln P(\underline{Y} = y_1, \ldots, y_n) = \ln \lambda \sum_{i=1}^{n} y_i - n\lambda - \sum_{i=1}^{n} \ln y_i!$$

$$\frac{dl}{d\lambda} = 0$$

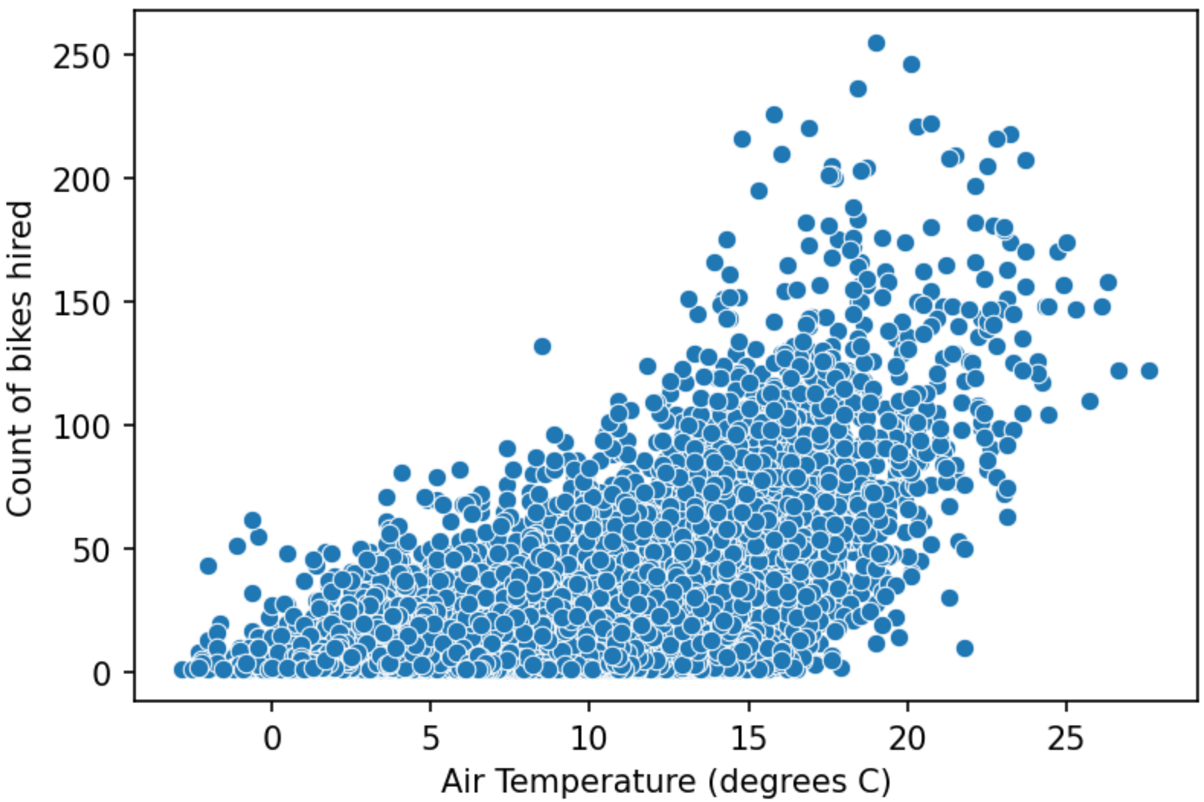
$\vdots$

$$\Rightarrow \lambda_{MLE} = \frac{1}{n} \sum_{i=1}^{n} y_i$$

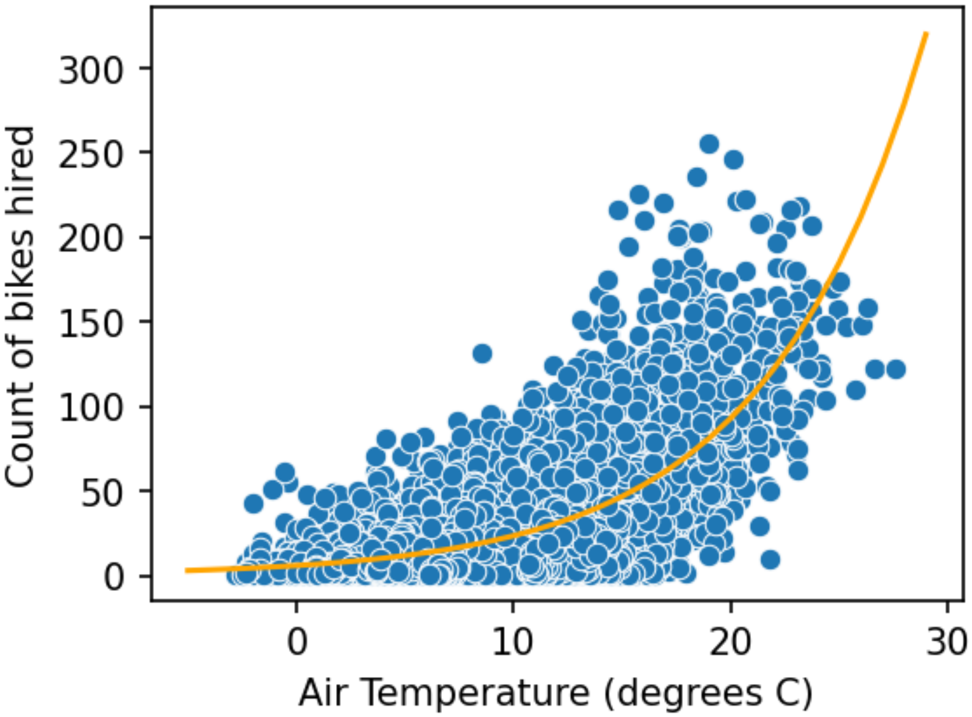


Foundations of Data Science: Regression and inference - Poisson regression

Poisson regression



Results with statsmodels GLM



Generalized Linear Model Regression Results

Dep. Variable:	count	**No. Observations:**	8301
Model:	GLM	**Df Residuals:**	8299
Model Family:	Poisson	**Df Model:**	1
Link Function:	Log	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-84533.
Date:	Wed, 01 Mar 2023	**Deviance:**	1.3111e+05
Time:	06:46:41	**Pearson chi2:**	1.40e+05
No. Iterations:	5	**Pseudo R-squ. (CS):**	1.000
Covariance Type:	nonrobust		

	coef	**std err**	**z**	**P>|z|**	**[0.025**	**0.975]**
const	1.7861	0.006	304.092	0.000	1.775	1.798
air_temperature	0.1373	0.000	323.057	0.000	0.136	0.138

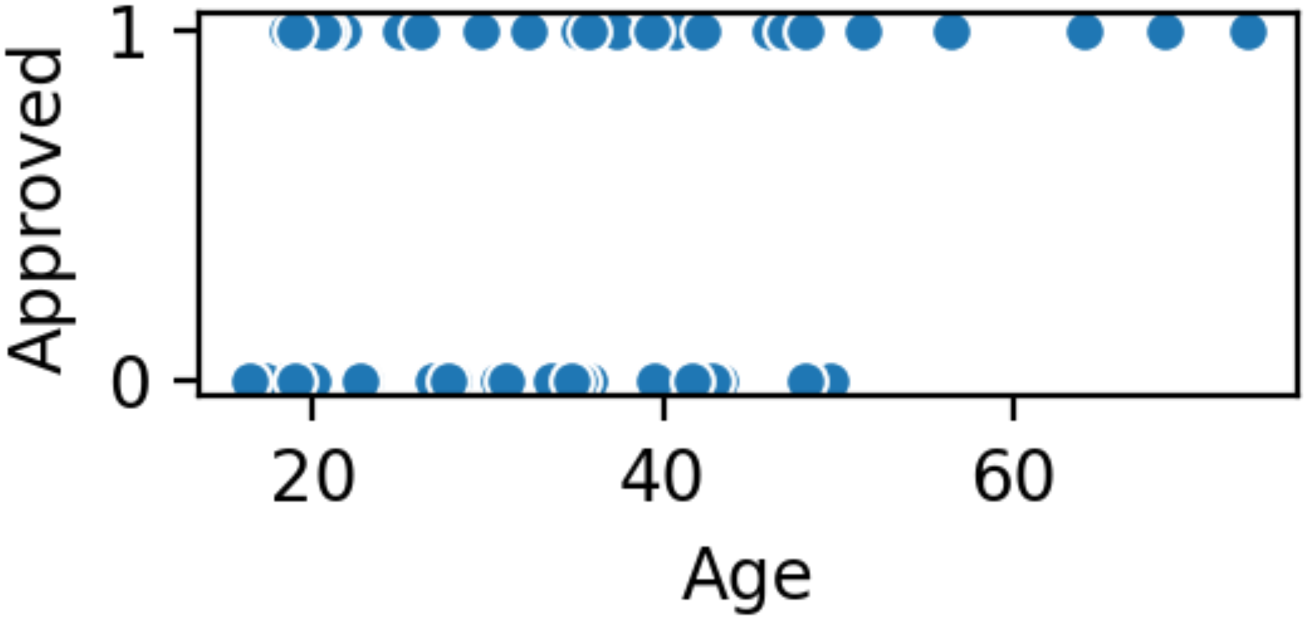
Poisson regression

$$l = \ln P(\underline{Y} = y_1, \ldots, y_n)$$

$$= \sum_{i=1}^{n} (\beta_0 + \beta_1 x_i) y_i - \sum_{i=1}^{n} e^{\beta_0 + \beta_1 x_i} - \sum_{i=1}^{n} \ln y_i !$$

Foundations of Data Science: Regression and inference – Generalised linear regression

Excercise



What distribution would we use to model the data here?

How would the parameter of that distribution depend on x (Age)?

Generalised linear regression

	Distribution	Link function
linear regression	Normal	$\mu = \beta_0 + \beta_1 x$, σ^2
Poisson regression	Poisson	$\ln \lambda = \beta_0 + \beta_1 x$
logistic regression	Bernoulli	$\ln \frac{p}{1-p} = \beta_0 + \beta_1 x$

Link functions

Expected value $\mu = E(Y|x)$ of a binomial dist is p

" " $\mu = E(Y|x)$ " " Poisson dist is λ

In general the link function is denoted $g(\mu)$ where $\mu = E(Y|x)$ for that distribution:

$$g(\mu) = \beta_0 + \beta_1 x$$

To make predictions, we invert the link function:

$$\mu = g^{-1}(\beta_0 + \beta_1 x)$$

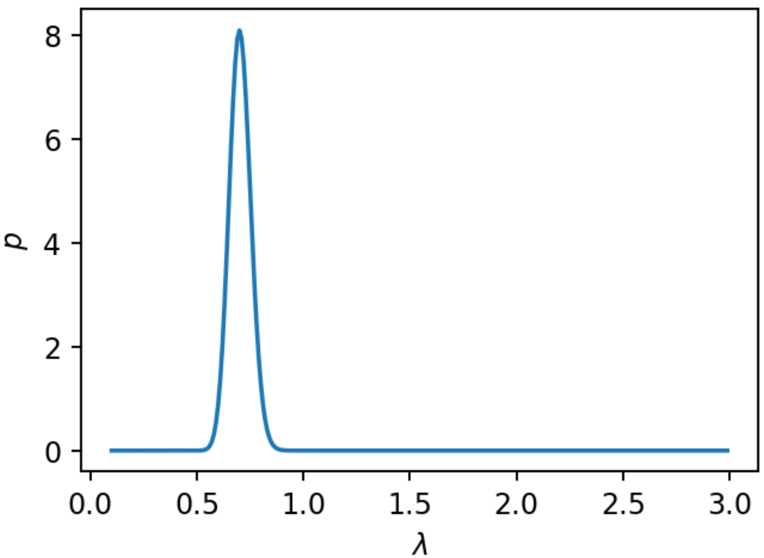
Max likelihood -> Bayesian Inference

Bayes Theorem:

Likelihood, Prior, Posterior, Evidence:

$$\underbrace{P(\vartheta \mid \underline{Y} = y)}_{\text{Posterior}} = \frac{\overbrace{P(\underline{Y} = y \mid \vartheta)}^{\text{Likelihood}}\ \overbrace{p(\vartheta)}^{\text{Prior}}}{\underbrace{P(Y = y)}_{\text{Evidence}}}$$

Horsekick posterior



$$P(Y = y) = \int_{-\infty}^{\infty} P(Y = y \mid \vartheta)\, p(\vartheta)\, d\vartheta$$

Summary

Motivated the probabalistic basis of inference using max likelihood

Important: think of what distribution should describe the data

Links to future courses:
- MLG (derivation of standard ML methods)
- MLPR (Bayesian approach; application to new problems)
- MCI (Causal inference)