Biostatistics
Week IX

Ege Ülgen, M.D.

2 December 2021

ACIBADEM
MEHMET ALİ AYDINLAR
ÜNİVERSİTESİ

Regression Analysis

- Regression can be used to
  - Understand the relationship between variables
  - Predict the value of one variable based on other variables
- Examples:
  - Quantifying the relative impacts of age, gender, and diet on BMI
  - Predicting whether the treatment will be successful or not based on age, tumor stage, tumor volume, …

Regression Analysis

- The variable to be predicted is called the **dependent variable**
  - Also called the **response variable**
- The value of this variable depends on the value of the **independent variable(s)**
  - Also called the **explanatory** or **predictor variable(s)**

Dependent variable = Independent variable + Independent variable + … + Independent variable

Linear Regression

E.g., quantifying the relative impacts of age, gender, and diet on BMI

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon$$

where Y is the dependent variable, X_1 to X_p are p independent variables, β_0 to β_p are the coefficients, and ε is the error term

Example - Prognostic factors for body fat - Multiple Linear Regression

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-60.045	5.365	-11.192	0.000
bmi	0.123	0.236	0.519	0.605
abdomen	0.438	0.105	4.183	0.000
waist_hip_ratio	38.468	10.262	3.749	0.000

$R^2 = 0.681$, $\boxed{R^2_{\mathrm{adj}} = 0.677}$ the proportion of the variation in the dependent variable that is predictable from the independent variable

$$Estimated\ Body\ Fat = -60.045 + 0.123 * bmi + 0.438 * abdomen + 38.468 * waist\_hip\_ratio$$

Example - Prognostic factors for body fat - Multiple Linear Regression

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-60.045	5.365	-11.192	0.000
bmi	0.123	0.236	0.519	0.605
abdomen	0.438	0.105	4.183	0.000
waist_hip_ratio	38.468	10.262	3.749	0.000

- For a person with bmi = 0, abdomen = 0, waist_hip_ratio = 0, the boy fat is estimated to be -60.045 ($p < 0.001$)
- (Keeping all other variables the same) with one unit increase in bmi, body fat increases by 0.123 (not significant since $p > 0.05$)
- With 95% confidence, it can be stated that with one unit increase in abdomen, body fat increases by 0.438 ($p < 0.001$)
- With one unit increase in waist_hip_ratio, body fat increases by 38.468 ($p < 0.001$)

Example II

- We’ll analyze the prostate cancer dataset
- The main aim of collecting this data set was to inspect the associations between **prostate-specific antigen (PSA)** and **prognostic clinical measurements** in men advanced prostate cancer
- Data were collected on 97 men who were about to undergo radical prostectomies

**PSA was transformed to logPSA for “normalization”*

Example II – Model 1

$$logPSA = 1.8 + 0.07 \ * \boldsymbol{vol} + 0.77 \ * I(\boldsymbol{invasion = 1})$$

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.8035	0.1141	15.81	<0.001
vol	0.0725	0.0133	5.43	<0.001
invasion1	0.7755	0.2541	3.05	0.003

Adjusted R-squared: 0.472

Example II – Model 2

$$logPSA = 1.67 + 0.1021 \ * \boldsymbol{vol} + 1.326 \ * I(\boldsymbol{invasion} = \mathbf{1}) - 0.056 * I(\boldsymbol{invasion} = \mathbf{1}) * \boldsymbol{vol}$$

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.6673	0.1289	12.94	<0.001
vol	0.1021	0.0191	5.35	<0.001
invasion1	1.326	0.3588	3.7	<0.001
vol:invasion1	-0.056	0.0262	-2.13	0.0354

Adjusted R-squared: 0.491

For a patient with invasion, there is an additional -0.056 change in PSA when vol changes one unit
= For a patient with invasion, one unit change in volume results in (0.1021 – 0.056) change in PSA

Example II – Model 3

$$logPSA = 1.55 + 0.076 \; * \boldsymbol{vol} + 0.45 \; * I(\boldsymbol{Gleason} = \mathbf{7}) + 0.9 * I(\boldsymbol{Gleason} = \mathbf{8})$$

(compared to **Gleason = 6**)

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.5523	0.1548	10.02	< 2e-16
vol	0.0758	0.0131	5.79	9.30E-08
Gleason7	0.4521	0.1928	2.34	0.0212
Gleason8	0.9043	0.2747	3.29	0.0014

Adjusted R-squared: 0.48

Logistic Regression

- Logistic regression is a specialized form of regression used when the dependent variable is **binary outcome**
  - Having a binary outcome (dependent variable) violates the assumption of linearity in linear regression

- The goal of logistic regression is to find the best fitting model to describe the relationship between the binary outcome and a set of independent variables
  - e.g., predicting whether the treatment will be successful or not, the presence/absence of a disease, etc.

Logistic Regression

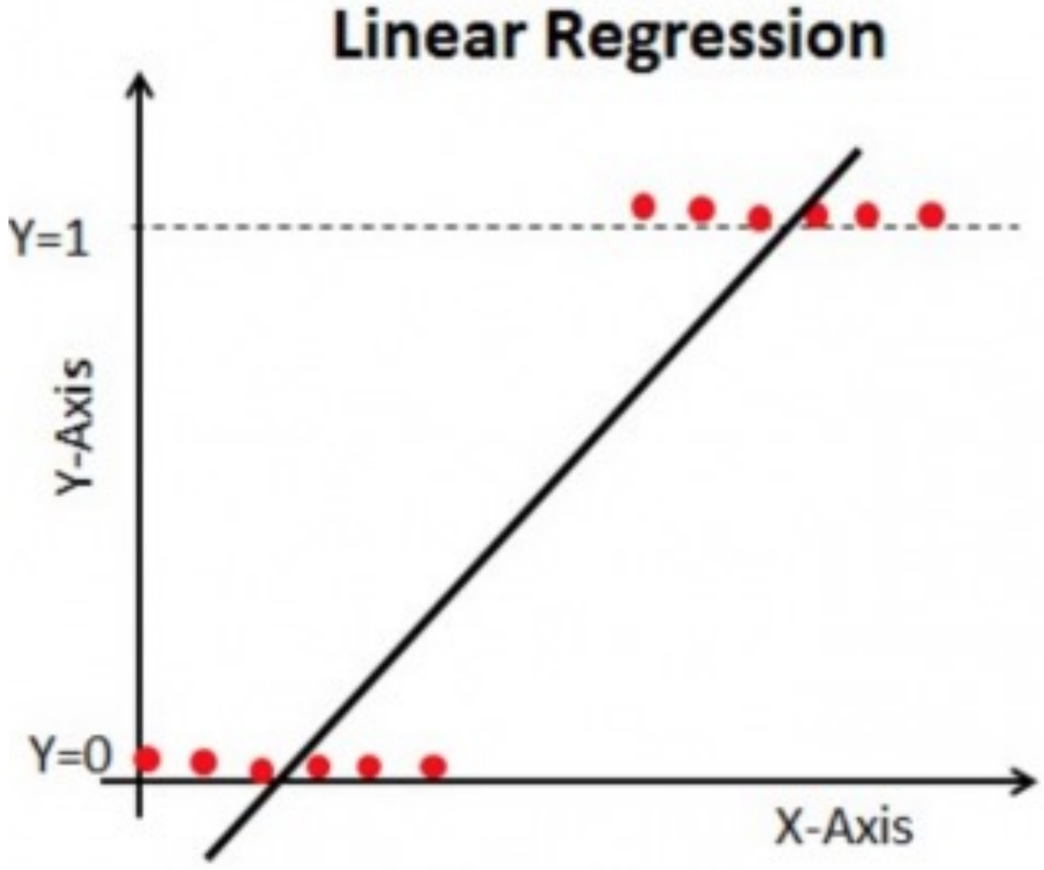
- Logistic regression generates the coefficients of the following formula to predict a **logit transformation** of the probability of presence of the outcome:

$$logit(P(Y=1)) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

where *P(Y = 1)* indicates the probability that the outcome is 1 (where the binary outcome variable is encoded as 0 and 1)

- *logit* is in fact the log of odds:

$$logit(p) = ln\left(\frac{p}{1-p}\right)$$



Science O-OD. Logistic regression with python [Internet]. Medium. 2019 [cited 2021 Oct 9].
Available from: https://medium.com/@ODSC/logistic-regression-with-python-ede39f8573c7

Logistic Regression – Example

- Identification of risk factors for lymph node metastases with prostate cancer
- n = 52 patients
- y = nodal metastases (0 = none, 1 = metastases)
- x = phosphatase, age , X-ray result, tumor size, tumor grade
  - The first two variables are continuous, the rest are binary

https://www.ebpi.uzh.ch/dam/jcr:ffffffff-c1f2-5119-ffff-ffff9de0876f/slides-logistic.pdf

Lymph node metastases – **Univariate** Models

	Estimate	Std. Error	z value	Pr(>|z|)	OR
$\log_2$(phosph)	2.4198	0.8778	2.76	0.0058	11.2
Age	-0.0448	0.0468	-0.96	0.3379	1.0
X-ray	2.1466	0.6984	3.07	0.0021	8.6
Size	1.6094	0.6325	2.54	0.0109	5.0
Grade	1.1389	0.5972	1.91	0.0565	3.1

https://www.ebpi.uzh.ch/dam/jcr:ffffffff-c1f2-5119-ffff-ffff9de0876f/slides-logistic.pdf

Lymph node metastases – Final Model

	Estimate	Std. Error	z value	Pr(>|z|)	OR
(Intercept)	-0.5418	0.8298	-0.65	0.5138	
$\log_2$(phosph)	2.3645	1.0267	2.30	0.0213	10.6
X-ray	1.9704	0.8207	2.40	0.0163	7.2
Size	1.6175	0.7534	2.15	0.0318	5.0

https://www.ebpi.uzh.ch/dam/jcr:ffffffff-c1f2-5119-ffff-ffff9de0876f/slides-logistic.pdf

Interpretation

	Estimate	Std. Error	z value	Pr(>|z|)	OR
(Intercept)	-0.5418	0.8298	-0.65	0.5138	
$\log_2$(phosph)	2.3645	1.0267	2.30	0.0213	10.6
X-ray	1.9704	0.8207	2.40	0.0163	7.2
Size	1.6175	0.7534	2.15	0.0318	5.0

- With 95% confidence, it could be said that a patient with $\log_2$(phosphatase) = 0, negative X-ray result, size = 0 was equally-likely in terms of having nodal metastases ($p = 0.5138$)
- With 95% confidence, it could be said that $\log_2$(phosphatase) and having nodal metastases are associated ($p = 0.0213$)
  - A one unit increase in $\log_2$(phosphatase) was associated with approximately 963.87% increase in the odds of having nodal metastases
  - (exp(2.3645) – 1) * 100 = 963.87
- …

Poisson Regression

- Linear regression was for continuous outcome, whereas logistic regression for binary outcome
- For **count** outcome, Poisson regression can be used

Poisson Regression - Example

- For 59 epilepsy patients the following data were collected:
  - **treatment:** the **treatment group**, a factor with levels placebo and Progabide
  - **base**: the **number of seizures** collected during 8-week period **before** the trial started
  - **age**: the **age of the patient**
  - **seizure rate:** the **number of seizures** occurred during the 2-week period **after** the trial was started

Poisson Regression – Example (cont.)

- First 10 patients:

treatment	base	age	seizure.rate	subject
placebo	11	31	5	1
placebo	11	30	3	2
placebo	6	25	2	3
placebo	8	36	4	4
placebo	66	22	7	5
placebo	27	29	5	6
placebo	12	31	6	7
placebo	52	42	40	8
placebo	23	37	5	9
placebo	10	28	14	10

Poisson Regression – Example (cont.)

- A Poisson regression with treatment group, previous seizures and age are related to the mean number of of seizure for patient i, λ_i, is given by:

$$log(\lambda_i) = \beta_0 + \beta_1 * I(treatment = Progabide) + \beta_2 * (base - 6) + \beta_3(age - 18)$$

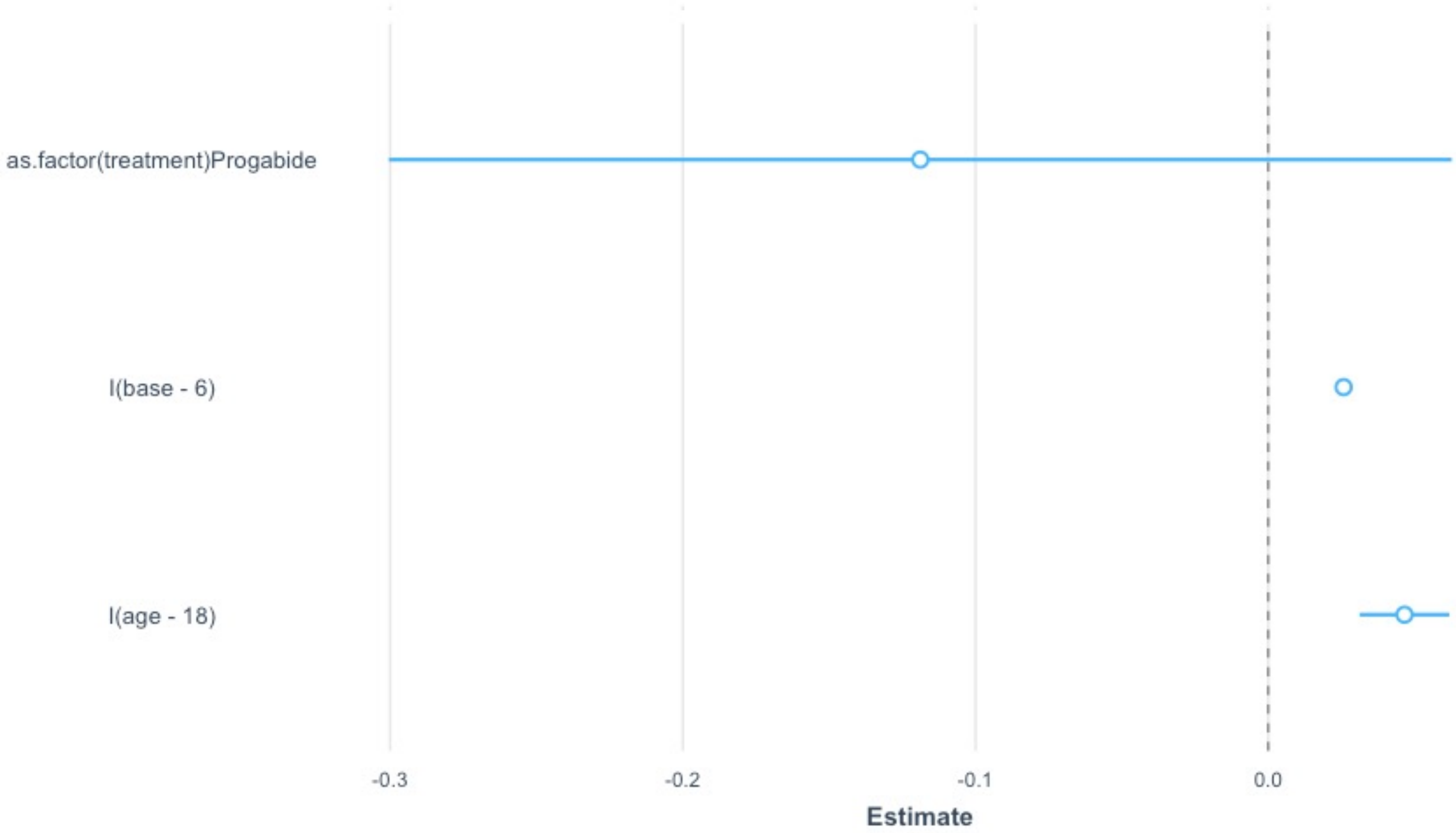
Poisson Regression – Example (cont.)

$$log(\lambda_i) = \beta_0 + \beta_1 * I(treatment = Progabide) + \beta_2 * (base - 6) + \beta_3(age - 18)$$

	Estimate	**Std. Error**	**z value**	**p**
(Intercept)	0.75	0.14	5.33	<0.001
treament = Progabide	-0.12	0.09	-1.28	0.20
base	0.03	0.00	26.37	<0.001
age	0.05	0.01	5.95	<0.001

Poisson Regression – Example (cont.)

$$log(\lambda_i) = \beta_0 + \beta_1 * I(treatment = Progabide) + \beta_2 * (base - 6) + \beta_3(age - 18)$$



Poisson Regression – Example (cont.)

	Estimate	Std. Error	z value	p
(Intercept)	0.75	0.14	5.33	<0.001
treament = Progabide	-0.12	0.09	-1.28	0.20
base	0.03	0.00	26.37	<0.001
age	0.05	0.01	5.95	<0.001

- A patient in placebo group, with 6 previous seizures, and aged 18 had approximately 2 seizures on average in the first two weeks after the trial was started
  - exp(0.75)
- With 95% confidence, it could be said that there was no difference between placebo and progabide (p-value = 0.199)
  - Negative estimate for β_1 indicates lowered mean number of seizures for progabide, but the difference from placebo was not significant

Poisson Regression – Example (cont.)

	Estimate	Std. Error	z value	p
(Intercept)	0.75	0.14	5.33	<0.001
treament = Progabide	-0.12	0.09	-1.28	0.20
base	0.03	0.00	26.37	<0.001
age	0.05	0.01	5.95	<0.001

- With 95% confidence, it could be said that previous number of seizures occurred in the 8-week interval prior to the study start and mean seizure rate was significantly associated (p-value < 0.001)
- One unit increase in previous seizure is associated with approximately 2.6% increase in the mean number of seizures in the first two weeks of the trial
  - (exp(0.03) - 1) * 100

Poisson Regression – Example (cont.)

	Estimate	Std. Error	z value	p
(Intercept)	0.75	0.14	5.33	<0.001
treament = Progabide	-0.12	0.09	-1.28	0.20
base	0.03	0.00	26.37	<0.001
age	0.05	0.01	5.95	<0.001

- With 95% confidence, it could be said that age sand mean seizure rate was significantly associated (p-value < 0.001)
- One unit increase in age is associated with approximately 4.8% increase in the mean number of seizures in the first two weeks of the trial
  - (exp(0.05) - 1) * 100

Brief Summary

Dependent Variable	Regression Model
Continuous	Linear Regression
Binary	Logistic Regression
Count	Poisson Regression