HEADS
PhD PROGRAMME IN HEALTH DATA SCIENCE

Simple Linear Regression

1. Regression analysis
2. Regression line
3. Least squares
4. Simple linear regression

Basics of Health Intelligent Data Analysis

PhD Programme in Health Data Science

Cláudia Camila Dias

Pedro Pereira Rodrigues

U.PORTO FMUP FACULDADE DE MEDICINA UNIVERSIDADE DO PORTO

MEDCIDS DEPARTAMENTO DE MEDICINA DA COMUNIDADE, INFORMAÇÃO E DECISÃO EM SAÚDE

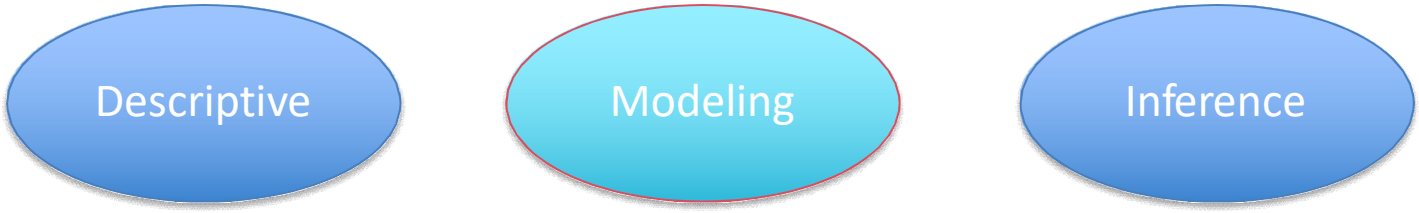
CINTESIS Health. Research.

Introduction

- There are several steps that we follow when we analyze data
- (Data checking and cleaning)
- We should start by describing and performing a graphical inspection of the data
- The next step usually involves univariate/bivariate inferential methods to identify association between two variables
- Finally, we might be interested in studying more complex associations in the data involving several variables at the same time

Introduction

Broad division of statistical methods



Summarize and simplify the information

Create models to understand complex relations in the data

Evaluate the precision and generalize the results

Introduction

The Descriptive methods aim to summarize information

Summary measures:

Single variable
***Frequency**: counts, percentage, rate, risk, odds, prevalence, incidence*
***Central tendency and position**: mean, median, minimum, maximum, percentiles*
***Dispersion**: standard deviation (variance), range (e.g. interquartile range)*
Two variables
***Association**: correlation, relative risk, odds ratio, hazard ratio, mean difference, risk difference*
***Agreement**: kappa statistics, intraclass correlation, sensitivity, specificity, area under the ROC curve*

Introduction

With methods for inference we draw conclusions about the population using the sample results

- Confidence intervals
- Hypothesis Testing
  - Parametric: t-test, anova
  - Non-parametric: Mann-Whitney, Wilcoxon,…

Introduction

- The choice of descriptive and inferential methods depend:
  - Research question
  - Design type
  - Type of variable collected
  - Distributional assumptions

Introduction

- Regression analysis is a broad term for statistical models of the form

$$g(Y) = f(X, \beta)$$

- Where Y stands for the outcome (dependent) variable(s), X is vector of covariates, β is the vector of regression parameters, f is some function (known or unknown) and g is a known function.

- Linear regression, logistic regression, Poisson regression, Cox regression

Introduction

- Typically, we use statistical models to describe the relation of one outcome (dependent variable) with multiple variables (covariates or independent variables)

- We refer to these methods as **multivariable** methods (**multivariate** refers to multiple outcomes and multiple covariates)

- The choice of the model will depend
  - Research question
  - Design type
  - Type of outcome
  - Distributional assumptions

Introduction

Why do we care about modeling?

1. Descriptive - describe strength of the association between outcome and factors of interest eliminating “noise”

2. Adjustment - for covariates/confounders

e.g. compare mortality between hospitals adjusting for the patients’ case-mix

3. Predictors - to determine important risk factors affecting the outcome

e.g. identify risk factors associated with cardiovascular disease

4. Prediction - prognostic/diagnostic

e.g. severity scores (APACHE, SAPS; PRIMS, MPM,…)

Introduction

Yanosky et al. Environmental Health 2014, 13:63
http://www.ehjournal.net/content/13/1/63

ENVIRONMENTAL HEALTH

RESEARCH Open Access

Spatio-temporal modeling of particulate air pollution in the conterminous United States using geographic and meteorological predictors

Jeff D Yanosky[1*], Christopher J Paciorek[2], Francine Laden[3,4], Jaime E Hart[3,4], Robin C Puett[5], Duanping Liao[1] and Helen H Suh[6]

Abstract

Background: Exposure to atmospheric particulate matter (PM) remains an important public health concern, although it remains difficult to quantify accurately across large geographic areas with sufficiently high spatial resolution. Recent epidemiologic analyses have demonstrated the importance of spatially- and temporally-resolved exposure estimates, which show larger PM-mediated health effects as compared to nearest monitor or county-specific ambient concentrations.

Methods: We developed generalized additive mixed models that describe regional and small-scale spatial and temporal gradients (and corresponding uncertainties) in monthly mass concentrations of fine ($PM_{2.5}$), inhalable (PM_{10}), and coarse mode particle mass ($PM_{2.5-10}$) for the conterminous United States (U.S.). These models expand our previously developed models for the Northeastern and Midwestern U.S. by virtue of their larger spatial domain, their inclusion of an additional 5 years of PM data to develop predictions through 2007, and their use of refined geographic covariates for population density and point-source PM emissions. Covariate selection and model validation were performed using 10-fold cross-validation (CV).

Results: The $PM_{2.5}$ models had high predictive accuracy (CV R^2=0.77 for both 1988–1998 and 1999–2007). While model performance remained strong, the predictive ability of models for PM_{10} (CV R^2=0.58 for both 1988–1998 and 1999–2007) and $PM_{2.5-10}$ (CV R^2=0.46 and 0.52 for 1988–1998 and 1999–2007, respectively) was somewhat lower. Regional variation was found in the effects of geographic and meteorological covariates. Models generally performed well in both urban and rural areas and across seasons, though predictive performance varied somewhat by region (CV R^2=0.81, 0.81, 0.83, 0.72, 0.69, 0.50, and 0.60 for the Northeast, Midwest, Southeast, Southcentral, Southwest, Northwest, and Central Plains regions, respectively, for $PM_{2.5}$ from 1999–2007).

Conclusions: Our models provide estimates of monthly-average outdoor concentrations of $PM_{2.5}$, PM_{10}, and $PM_{2.5-10}$ with high spatial resolution and low bias. Thus, these models are suitable for estimating chronic exposures of populations living in the conterminous U.S. from 1988 to 2007.

Keywords: Particulate matter, Spatio-temporal models, Land use regression, Spatial smoothing, Penalized splines, Generalized additive mixed model

* Correspondence: jyanosky@phs.psu.edu
[1]Department of Public Health Sciences, The Pennsylvania State University College of Medicine, Hershey, PA, USA
Full list of author information is available at the end of the article

Abstract

Background: Exposure to atmospheric particulate matter (PM) remains an important public health concern, although it remains difficult to quantify accurately across large geographic areas with sufficiently high spatial resolution. Recent epidemiologic analyses have demonstrated the importance of spatially- and temporally-resolved exposure estimates, which show larger PM-mediated health effects as compared to nearest monitor or county-specific ambient concentrations.

Methods: We developed generalized additive mixed models that describe regional and small-scale spatial and temporal gradients (and corresponding uncertainties) in monthly mass concentrations of fine ($PM_{2.5}$), inhalable (PM_{10}), and coarse mode particle mass ($PM_{2.5-10}$) for the conterminous United States (U.S.). These models expand our previously developed models for the Northeastern and Midwestern U.S. by virtue of their larger spatial domain, their inclusion of an additional 5 years of PM data to develop predictions through 2007, and their use of refined geographic covariates for population density and point-source PM emissions. Covariate selection and model validation were performed using 10-fold cross-validation (CV).

Results: The $PM_{2.5}$ models had high predictive accuracy (CV R^2=0.77 for both 1988–1998 and 1999–2007). While model performance remained strong, the predictive ability of models for PM_{10} (CV R^2=0.58 for both 1988–1998 and 1999–2007) and $PM_{2.5-10}$ (CV R^2=0.46 and 0.52 for 1988–1998 and 1999–2007, respectively) was somewhat lower. Regional variation was found in the effects of geographic and meteorological covariates. Models generally performed well in both urban and rural areas and across seasons, though predictive performance varied somewhat by region (CV R^2=0.81, 0.81, 0.83, 0.72, 0.69, 0.50, and 0.60 for the Northeast, Midwest, Southeast, Southcentral, Southwest, Northwest, and Central Plains regions, respectively, for $PM_{2.5}$ from 1999–2007).

Conclusions: Our models provide estimates of monthly-average outdoor concentrations of $PM_{2.5}$, PM_{10}, and $PM_{2.5-10}$ with high spatial resolution and low bias. Thus, these models are suitable for estimating chronic exposures of populations living in the conterminous U.S. from 1988 to 2007.

Keywords: Particulate matter, Spatio-temporal models, Land use regression, Spatial smoothing, Penalized splines, Generalized additive mixed model

Introduction

Open access Research

BMJ Open

Fatty liver and mortality: a cohort population study in South Italy

Maria Gabriella Caruso,[1,2] Nicola Veronese,[3] Maria Notarnicola,[1] Anna Maria Cisternino,[2] Rosa Reddavide,[2] Rosa Inguaggiato,[2] Vito Guerra,[4] Rossella Donghia,[4] Antonio Logroscino,[5] Ornella Rotolo,[2] Marisa Chiloiro,[6] Gioacchino Leandro,[7] Giampiero De Leonardis,[1,2] Valeria Tutino,[1] G Misciagna,[8] Caterina Bonfiglio,[9] Rocco Guerra,[9] Alberto Osella[9]

To cite: Caruso MG, Veronese N, Notarnicola M, *et al*. Fatty liver and mortality: a cohort population study in South Italy. *BMJ Open* 2019;**9**:e027379. doi:10.1136/bmjopen-2018-027379

► Prepublication history for this paper is available online. To view these files, please visit the journal online (http://dx.doi.org/10.1136/bmjopen-2018-027379).

Received 19 October 2018
Revised 8 January 2019
Accepted 4 April 2019

For numbered affiliations see end of article.

Correspondence to
Dr Maria Gabriella Caruso; gabriella.caruso@irccsdebellis.it

ABSTRACT

Objective Alcoholic fatty liver (AFLD) and non-alcoholic fatty liver (NAFLD) are two common conditions. However, if they can increase the risk of death is poorly explored. We therefore aimed to investigate the potential association between the presence and severity of liver steatosis and mortality in a large sample of older people.

Design Prospective.

Setting Community.

Participants Women and men randomly sampled from the electoral rolls of the population of Castellana Grotte, a town in Southern Italy (Apulia region) between 2005 and 2006. Among 1942 initially contacted, 1708 (=87.9%) participated to the baseline survey (Multicentrica Colelitiasi III (MICOL III)). This specific study included 1445 older participants (mean age=65.2 years, females=44.2%).

Exposure NAFLD or AFLD.

Primary and secondary outcomes Mortality (all-cause and specific-cause).

Results After a median of 12 years, 312 participants (=21.6%) died. After adjusting for nine potential confounders, the presence of steatosis was not associated with any increased risk of death in both NAFLD and AFLD. The severity of liver steatosis was not associated with any increased risk of mortality in NAFLD, while in AFLD, the presence of moderate steatosis significantly increased the risk of overall (HR=2.16; 95% CI 1.19 to 3.91) and cancer-specific (HR=3.54; 95% CI 1.16 to 10.87) death.

Conclusions Liver steatosis is not associated with any increased risk of death in NAFLD, while moderate steatosis could be a risk factor for mortality (particularly due to cancer) in people affected by AFLD.

Strengths and limitations of this study

- The long follow-up (12 years) is a strength of our research.
- The use of liver echography for the diagnosis of liver steatosis (contrary to most studies using biohumoral parameters) is another strength of our work.
- The number of people with severe grade of either AFLD/NALFD is small and this can introduce an important bias in our findings.
- No liver biopsy (the gold standard method for the diagnosis of liver steatosis) is available.
- Comorbidities are self-reported and this could introduce another bias.

INTRODUCTION

Fatty liver seems to be a common condition in Western countries,[1] probably affecting about half of the individuals in some studies.[2] It is estimated that nearly one in every three patients with persistently elevated alanine transaminase might have a fatty liver disease.[3]

Several risk factors have been recognised for the development of fatty liver. Among them, alcohol is probably the most important. Heavy alcohol intake is associated with fatty liver and afterwards with fibrosis, indicating a form called alcoholic fatty liver (AFLD).[4] In Western countries, however, fatty liver is usually associated with abdominal obesity and with metabolic syndrome features[1] underlining a picture called non-alcoholic fatty liver (NAFLD).[1]

Even if the prevalence of the NAFLD varies according to the definition and the country, a recent systematic review and meta-analysis reported an overall prevalence of 25%.[5] Independently from the origin, liver steatosis seems to be a significant predictor of several medical conditions including cardiovascular diseases (CVDs),[6] type 2 diabetes[7] and liver cirrhosis.[1] Metabolic conditions associated with NAFLD included obesity, present in more than half of the people, type 2 diabetes, hyperlipidemia, hypertension and metabolic syndrome.[5] Liver-specific mortality and overall mortality in NAFLD patients is estimated in 0.77 persons per 1000.[5] However, the role of potential confounders seems to be important for explaining the association between the presence of liver steatosis and these medical conditions. For example, the association between liver steatosis and the onset of CVDs disappeared after the adjustment for potential confounders (eg, the

Abstract

Objective Alcoholic fatty liver (AFLD) and non-alcoholic fatty liver (NAFLD) are two common conditions. However, if they can increase the risk of death is poorly explored. We therefore aimed to investigate the potential association between the presence and severity of liver steatosis and mortality in a large sample of older people.

Design Prospective.

Setting Community.

Participants Women and men randomly sampled from the electoral rolls of the population of Castellana Grotte, a town in Southern Italy (Apulia region) between 2005 and 2006. Among 1942 initially contacted, 1708 (=87.9%) participated to the baseline survey (Multicentrica Colelitiasi III (MICOL III)). This specific study included 1445 older participants (mean age=65.2 years, females=44.2%).

Exposure NAFLD or AFLD.

Primary and secondary outcomes Mortality (all-cause and specific-cause).

Results After a median of 12 years, 312 participants (=21.6%) died. After adjusting for nine potential confounders, the presence of steatosis was not associated with any increased risk of death in both NAFLD and AFLD. The severity of liver steatosis was not associated with any increased risk of mortality in NAFLD, while in AFLD, the presence of moderate steatosis significantly increased the risk of overall (HR=2.16; 95% CI 1.19 to 3.91) and cancer-specific (HR=3.54; 95% CI 1.16 to 10.87) death.

Conclusions Liver steatosis is not associated with any increased risk of death in NAFLD, while moderate steatosis could be a risk factor for mortality (particularly due to cancer) in people affected by AFLD.

Introduction

PLOS ONE

RESEARCH ARTICLE

Predictors of sudden cardiac death in atrial fibrillation: The Atherosclerosis Risk in Communities (ARIC) study

Ryan J. Koene[1]*, Faye L. Norby[2], Ankit Maheshwari[1], Mary R. Rooney[2], Elsayed Z. Soliman[3], Alvaro Alonso[4], Lin Y. Chen[1]

1 Cardiovascular Division, Department of Medicine, University of Minnesota Medical School, Minneapolis, Minnesota, United States of America, **2** Division of Epidemiology and Community Health, School of Public Health, University of Minnesota, Minneapolis, Minnesota, United States of America, **3** Epidemiological Cardiology Research Center (EPICARE), Department of Epidemiology and Prevention, and Department of Internal Medicine-Cardiology, Wake Forest School of Medicine, Winston Salem, North Carolina, United States of America, **4** Department of Epidemiology, Rollins School of Public Health, Emory University, Atlanta, Georgia, United States of America

* koene030@umn.edu

OPEN ACCESS

Citation: Koene RJ, Norby FL, Maheshwari A, Rooney MR, Soliman EZ, Alonso A, et al. (2017) Predictors of sudden cardiac death in atrial fibrillation: The Atherosclerosis Risk in Communities (ARIC) study. PLoS ONE 12(11): e0187659. https://doi.org/10.1371/journal.pone.0187659

Editor: Chunhua Song, Pennsylvania State University, UNITED STATES

Received: June 18, 2017

Accepted: October 24, 2017

Published: November 8, 2017

Data Availability Statement: Study participants did not consent to have their data publicly available and freely accessible and, therefore, we cannot publicly share the data. However, the data underlying our work can be obtained through two mechanisms. First, interested investigators can contact the ARIC Coordinating Center at the University of North Carolina – Chapel Hill. Details about the procedures for data request can be found in the following website: http://www2.cscc.unc.edu/aric/distribution-agreements. Second, most

Abstract

We previously reported that incident atrial fibrillation (AF) is associated with an increased risk of sudden cardiac death (SCD) in the general population. We now aimed to identify predictors of SCD in persons with AF from the Atherosclerosis Risk in Communities (ARIC) study, a community-based cohort study. We included all participants who attended visit 1 (1987–89) and had no prior AF (n = 14,836). Incident AF was identified from study electrocardiograms and hospitalization discharge codes through 2012. SCD was physician-adjudicated. We used cause-specific Cox proportional hazards models, followed by stepwise selection (backwards elimination, removing all variables with p>0.10) to identify predictors of SCD in participants with AF. AF occurred in 2321 (15.6%) participants (age 45–64 years, 58% male, 18% black). Over a median of 3.3 years, SCD occurred in 110 of those with AF (4.7%). Predictors of SCD in AF included higher age, body mass index (BMI), coronary heart disease, hypertension, diabetes, current smoker, left ventricular hypertrophy, increased heart rate, and decreased albumin. Predictors associated only with SCD and not other cardiovascular (CV) death included increased BMI (HR per 5-unit increase, 1.15, 95% CI, 0.97–1.36, p = 0.10), increased heart rate (HR per SD increase, 1.18, 95% CI 0.99–1.41, p = 0.07), and low albumin (HR per SD decrease 1.23, 95% CI 1.02–1.48, p = 0.03). In the ARIC study, predictors of SCD in AF that are not associated with non-sudden CV death included increased BMI, increased heart rate, and low albumin. Further research to confirm these findings in larger community-based cohorts and to elucidate the underlying mechanisms to facilitate prevention is warranted.

Abstract

We previously reported that incident atrial fibrillation (AF) is associated with an increased risk of sudden cardiac death (SCD) in the general population. We now aimed to identify predictors of SCD in persons with AF from the Atherosclerosis Risk in Communities (ARIC) study, a community-based cohort study. We included all participants who attended visit 1 (1987–89) and had no prior AF (n = 14,836). Incident AF was identified from study electrocardiograms and hospitalization discharge codes through 2012. SCD was physician-adjudicated. We used cause-specific Cox proportional hazards models, followed by stepwise selection (backwards elimination, removing all variables with p>0.10) to identify predictors of SCD in participants with AF. AF occurred in 2321 (15.6%) participants (age 45–64 years, 58% male, 18% black). Over a median of 3.3 years, SCD occurred in 110 of those with AF (4.7%). Predictors of SCD in AF included higher age, body mass index (BMI), coronary heart disease, hypertension, diabetes, current smoker, left ventricular hypertrophy, increased heart rate, and decreased albumin. Predictors associated only with SCD and not other cardiovascular (CV) death included increased BMI (HR per 5-unit increase, 1.15, 95% CI, 0.97–1.36, p = 0.10), increased heart rate (HR per SD increase, 1.18, 95% CI 0.99–1.41, p = 0.07), and low albumin (HR per SD decrease 1.23, 95% CI 1.02–1.48, p = 0.03). In the ARIC study, predictors of SCD in AF that are not associated with non-sudden CV death included increased BMI, increased heart rate, and low albumin. Further research to confirm these findings in larger community-based cohorts and to elucidate the underlying mechanisms to facilitate prevention is warranted.

Introduction

NIH Public Access
Author Manuscript
J Am Coll Surg. Author manuscript; available in PMC 2014 November 01.

Published in final edited form as:
J Am Coll Surg. 2013 November ; 217(5): 833–842.e3. doi:10.1016/j.jamcollsurg.2013.07.385.

Development and Evaluation of the Universal ACS NSQIP Surgical Risk Calculator: A Decision Aide and Informed Consent Tool for Patients and Surgeons

Karl Y Bilimoria, MD, MS, FACS[1,2], **Yaoming Liu, PhD**[1], **Jennifer L Paruch, MD**[1], **Lynn Zhou, PhD**[1], **Thomas E Kmiecik, PhD**[2], **Clifford Y Ko, MD, MS, MSHS, FACS**[1,3], and **Mark E Cohen, PhD**[1]

[1] Division of Research and Optimal Patient Care, American College of Surgeons, Chicago, IL

[2] Surgical Outcomes and Quality Improvement Center, Department of Surgery, Feinberg School of Medicine, Northwestern University, Chicago, IL

[3] Department of Surgery, University of California, Los Angeles (UCLA) and VA Greater Los Angeles Healthcare System, Los Angeles, CA

Abstract

BACKGROUND—Accurately estimating surgical risks is critical for shared decision making and informed consent. The Centers for Medicare and Medicaid Services may soon put forth a measure requiring surgeons to provide patients with patient-specific, empirically-derived estimates of postoperative complications. Our objectives were (1) to develop a universal surgical risk estimation tool, (2) to compare performance of the universal vs. prior procedure-specific Surgical Risk Calculators, and (3) to allow surgeons to empirically adjust the estimates of risk.

STUDY DESIGN—Using standardized clinical data from 393 ACS NSQIP hospitals, a web-based tool was developed to allow surgeons to easily enter 21 preoperative factors (demographics, comorbidities, procedure). Regression models were developed to predict 8 outcomes based on the preoperative risk factors. The universal model was compared to procedure-specific models. To incorporate surgeon input, a subjective Surgeon Adjustment Score, allowing risk estimates to vary within the estimate's confidence interval, was introduced and tested with 80 surgeons using 10 case scenarios.

RESULTS—Based on 1,414,006 patients encompassing 1,557 unique CPT codes, a universal Surgical Risk Calculator model was developed which had excellent performance for mortality (c-statistic=0.944; Brier=0.011[where scores approaching zero are better]), morbidity (c-statistic=0.816, Brier=0.069), and 6 additional complications (c-statistics>0.8). Predictions were similarly robust for the universal calculator vs. procedure-specific calculators (e.g., colorectal). Surgeons demonstrated considerable agreement on the case scenario scoring (80-100% agreement), suggesting reliable score assignment between surgeons.

Correspondence address: Karl Y. Bilimoria, MD, MS Division of Research and Optimal Patient Care American College of Surgeo 633 N. St. Clair Street, 22nd Floor Chicago, IL 60611 kbilimoria@facs.org Office: (312) 202-560 Fax: (312) 202-5062.

Disclosure Information: Nothing to disclose.

Presented at the 2013 Annual ACS NSQIP Conference, San Diego, CA, July 2013.

Abstract

BACKGROUND—Accurately estimating surgical risks is critical for shared decision making and informed consent. The Centers for Medicare and Medicaid Services may soon put forth a measure requiring surgeons to provide patients with patient-specific, empirically-derived estimates of postoperative complications. Our objectives were (1) to develop a universal surgical risk estimation tool, (2) to compare performance of the universal vs. prior procedure-specific Surgical Risk Calculators, and (3) to allow surgeons to empirically adjust the estimates of risk.

STUDY DESIGN—Using standardized clinical data from 393 ACS NSQIP hospitals, a web-based tool was developed to allow surgeons to easily enter 21 preoperative factors (demographics, comorbidities, procedure). Regression models were developed to predict 8 outcomes based on the preoperative risk factors. The universal model was compared to procedure-specific models. To incorporate surgeon input, a subjective Surgeon Adjustment Score, allowing risk estimates to vary within the estimate's confidence interval, was introduced and tested with 80 surgeons using 10 case scenarios.

RESULTS—Based on 1,414,006 patients encompassing 1,557 unique CPT codes, a universal Surgical Risk Calculator model was developed which had excellent performance for mortality (c-statistic=0.944; Brier=0.011[where scores approaching zero are better]), morbidity (c-statistic=0.816, Brier=0.069), and 6 additional complications (c-statistics>0.8). Predictions were similarly robust for the universal calculator vs. procedure-specific calculators (e.g., colorectal). Surgeons demonstrated considerable agreement on the case scenario scoring (80-100% agreement), suggesting reliable score assignment between surgeons.

CONCLUSIONS—The ACS NSQIP Surgical Risk Calculator is a decision-support tool based on reliable multi-institutional clinical data which can be used to estimate the risks of most operations. The ACS NSQIP Surgical Risk Calculator will allow clinicians and patients to make decisions using empirically derived, patient-specific postoperative risks.

Introduction

We will introduce (review) the simplest model – **simple linear regression** – used for continuous outcomes (with some distributional assumptions)

The "**simple**" linear regression refers to the use of a **single covariate**

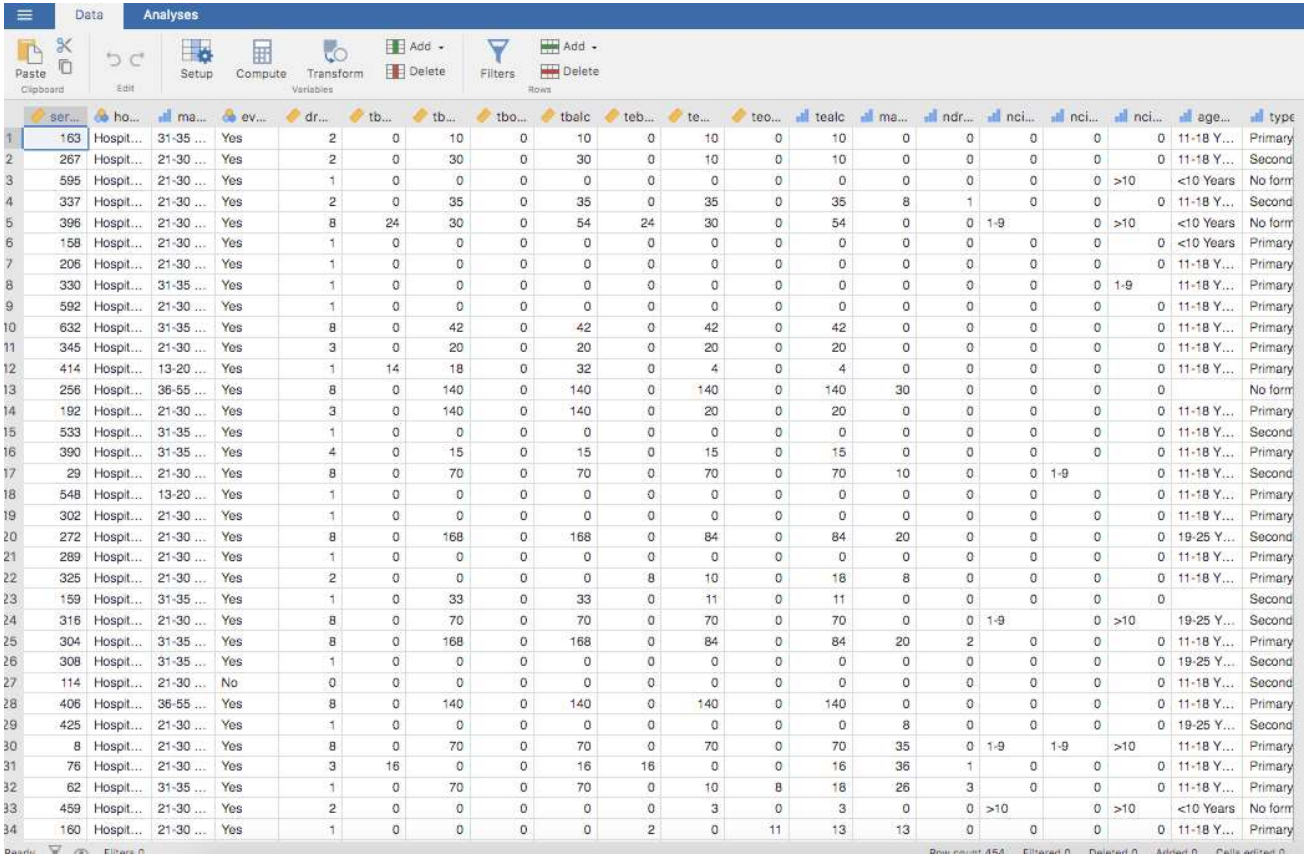
Then we will proceed to **multiple linear regression**, **simple logistic regression** and **multiple logistic regression**.

Simple Linear Regression

modeling a continuous variable

Motivation - Example

As a motivation example we will use a dataset with information regarding newborns (n=454)



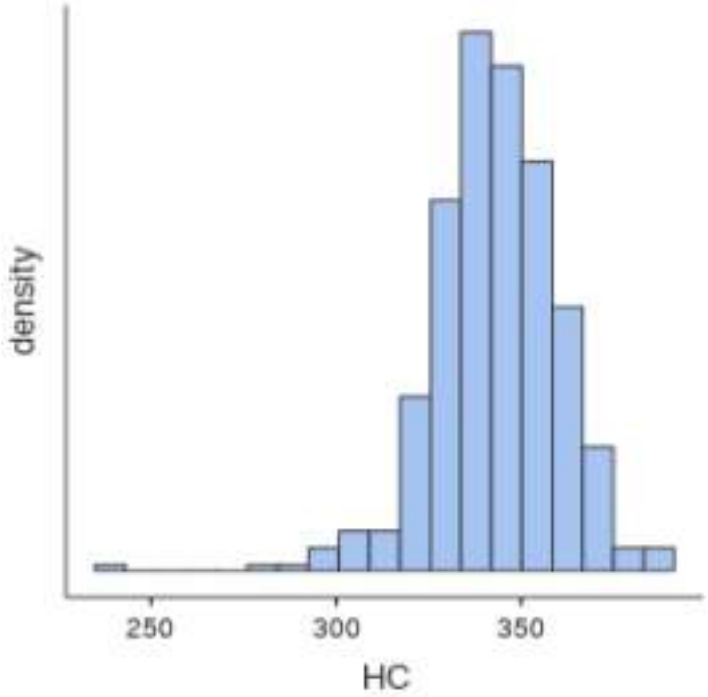
Data | Analyses

Paste | Clipboard | Edit | Setup | Compute | Transform | Add | Delete | Variables | Filters | Add | Delete | Rows

	ser...	ho...	ma...	ev...	dr...	tb...	tb...	tbo...	tbalc	teb...	te...	teo...	tealc	ma...	ndr...	nci...	nci...	nci...	age...	type
1	163	Hospit...	31-35 ...	Yes	2	0	10	0	10	0	10	0	10	0	0	0	0	0	11-18 Y...	Primary
2	267	Hospit...	21-30 ...	Yes	2	0	30	0	30	0	10	0	10	0	0	0	0	0	11-18 Y...	Second
3	595	Hospit...	21-30 ...	Yes	1	0	0	0	0	0	0	0	0	0	0	0	0	>10	<10 Years	No form
4	337	Hospit...	21-30 ...	Yes	2	0	35	0	35	0	35	0	35	8	1	0	0	0	11-18 Y...	Second
5	396	Hospit...	21-30 ...	Yes	8	24	30	0	54	24	30	0	54	0	0	1-9	0	>10	<10 Years	No form
6	158	Hospit...	21-30 ...	Yes	1	0	0	0	0	0	0	0	0	0	0	0	0	0	<10 Years	Primary
7	206	Hospit...	21-30 ...	Yes	1	0	0	0	0	0	0	0	0	0	0	0	0	0	11-18 Y...	Primary
8	330	Hospit...	31-35 ...	Yes	1	0	0	0	0	0	0	0	0	0	0	0	0	1-9	11-18 Y...	Primary
9	592	Hospit...	21-30 ...	Yes	1	0	0	0	0	0	0	0	0	0	0	0	0	0	11-18 Y...	Primary
10	632	Hospit...	31-35 ...	Yes	8	0	42	0	42	0	42	0	42	0	0	0	0	0	11-18 Y...	Primary
11	345	Hospit...	21-30 ...	Yes	3	0	20	0	20	0	20	0	20	0	0	0	0	0	11-18 Y...	Primary
12	414	Hospit...	13-20 ...	Yes	1	14	18	0	32	0	4	0	4	0	0	0	0	0	11-18 Y...	Primary
13	256	Hospit...	36-55 ...	Yes	8	0	140	0	140	0	140	0	140	30	0	0	0	0		No form
14	192	Hospit...	21-30 ...	Yes	3	0	140	0	140	0	20	0	20	0	0	0	0	0	11-18 Y...	Primary
15	533	Hospit...	31-35 ...	Yes	1	0	0	0	0	0	0	0	0	0	0	0	0	0	11-18 Y...	Second
16	390	Hospit...	31-35 ...	Yes	4	0	15	0	15	0	15	0	15	0	0	0	0	0	11-18 Y...	Primary
17	29	Hospit...	21-30 ...	Yes	8	0	70	0	70	0	70	0	70	10	0	0	1-9	0	11-18 Y...	Second
18	548	Hospit...	13-20 ...	Yes	1	0	0	0	0	0	0	0	0	0	0	0	0	0	11-18 Y...	Primary
19	302	Hospit...	21-30 ...	Yes	1	0	0	0	0	0	0	0	0	0	0	0	0	0	11-18 Y...	Primary
20	272	Hospit...	21-30 ...	Yes	8	0	168	0	168	0	84	0	84	20	0	0	0	0	19-25 Y...	Second
21	289	Hospit...	21-30 ...	Yes	1	0	0	0	0	0	0	0	0	0	0	0	0	0	11-18 Y...	Primary
22	325	Hospit...	21-30 ...	Yes	2	0	0	0	0	8	10	0	18	8	0	0	0	0	11-18 Y...	Primary
23	159	Hospit...	31-35 ...	Yes	1	0	33	0	33	0	11	0	11	0	0	0	0	0		Second
24	316	Hospit...	21-30 ...	Yes	8	0	70	0	70	0	70	0	70	0	0	1-9	0	>10	19-25 Y...	Second
25	304	Hospit...	31-35 ...	Yes	8	0	168	0	168	0	84	0	84	20	2	0	0	0	11-18 Y...	Primary
26	308	Hospit...	31-35 ...	Yes	1	0	0	0	0	0	0	0	0	0	0	0	0	0	19-25 Y...	Second
27	114	Hospit...	21-30 ...	No	0	0	0	0	0	0	0	0	0	0	0	0	0	0	11-18 Y...	Second
28	406	Hospit...	36-55 ...	Yes	8	0	140	0	140	0	140	0	140	0	0	0	0	0	11-18 Y...	Primary
29	425	Hospit...	21-30 ...	Yes	1	0	0	0	0	0	0	0	0	8	0	0	0	0	19-25 Y...	Second
30	8	Hospit...	21-30 ...	Yes	8	0	70	0	70	0	70	0	70	35	0	1-9	1-9	>10	11-18 Y...	Primary
31	76	Hospit...	21-30 ...	Yes	3	16	0	0	16	16	0	0	16	36	1	0	0	0	11-18 Y...	Primary
32	62	Hospit...	31-35 ...	Yes	1	0	70	0	70	0	10	8	18	26	3	0	0	0	11-18 Y...	Primary
33	459	Hospit...	21-30 ...	Yes	2	0	0	0	0	0	3	0	3	0	0	>10	0	>10	<10 Years	No form
34	160	Hospit...	21-30 ...	Yes	1	0	0	0	0	2	0	11	13	13	0	0	0	0	11-18 Y...	Primary

Motivation - Example

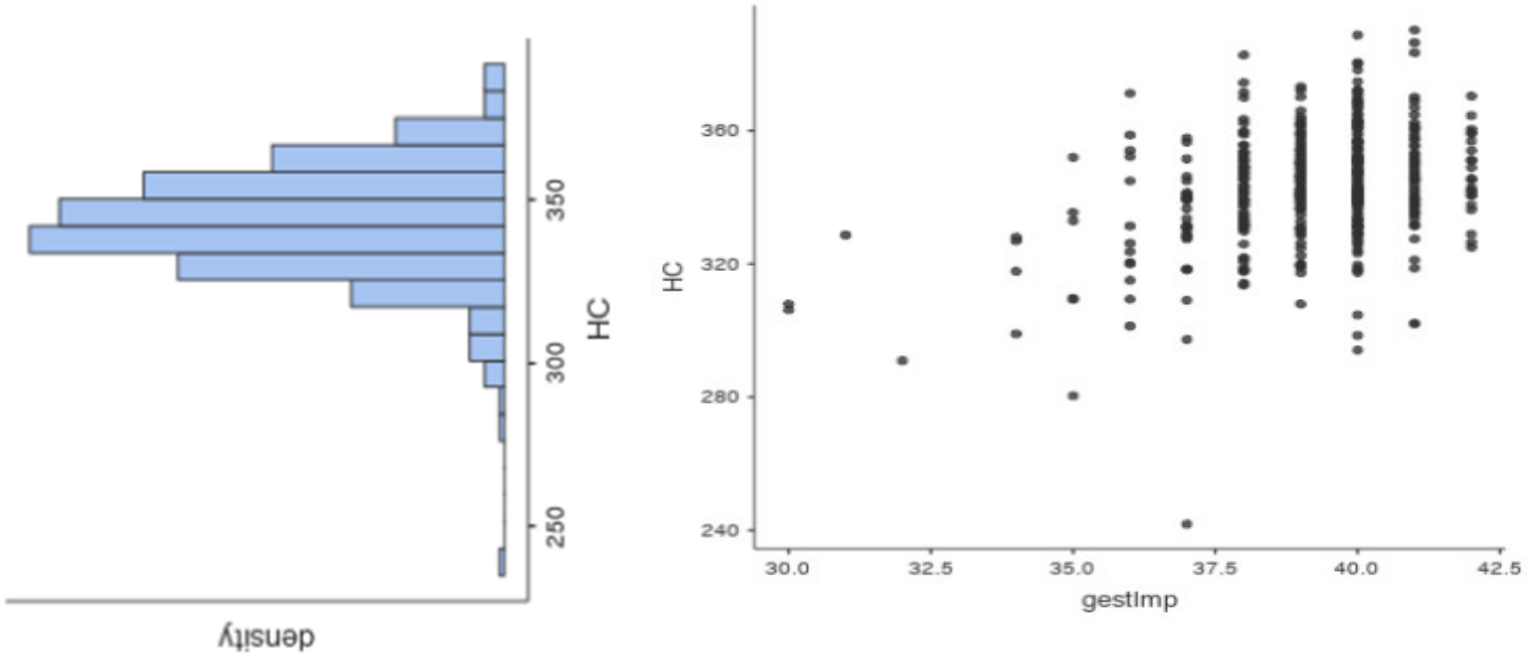
In particular we are interested in studying several characteristics of the newborns and pregnancy with their **head circumference**



Motivation - Example

For example the relation between gestational age and head circumference (HC)

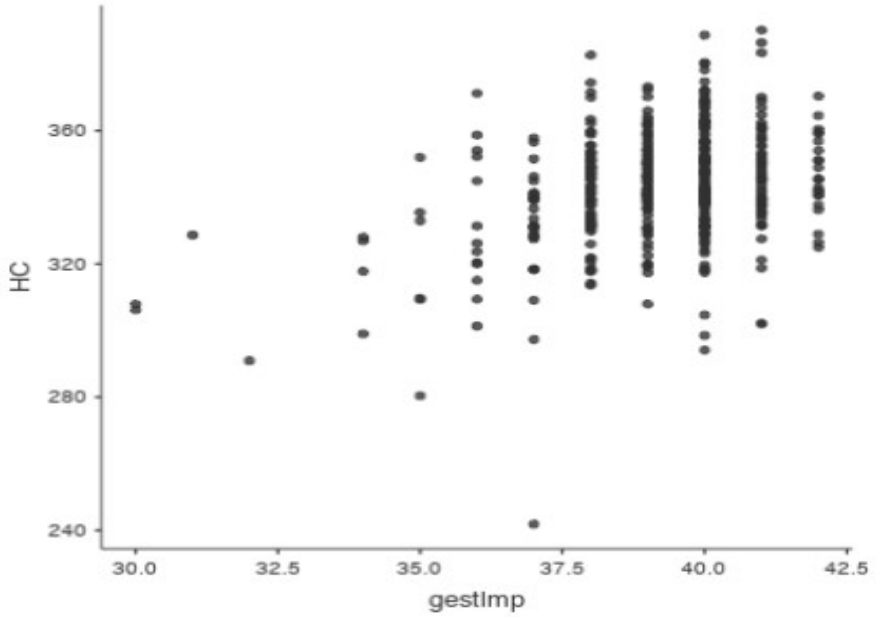
Let’s start by representing both variables in a scatter plot



Motivation - Example

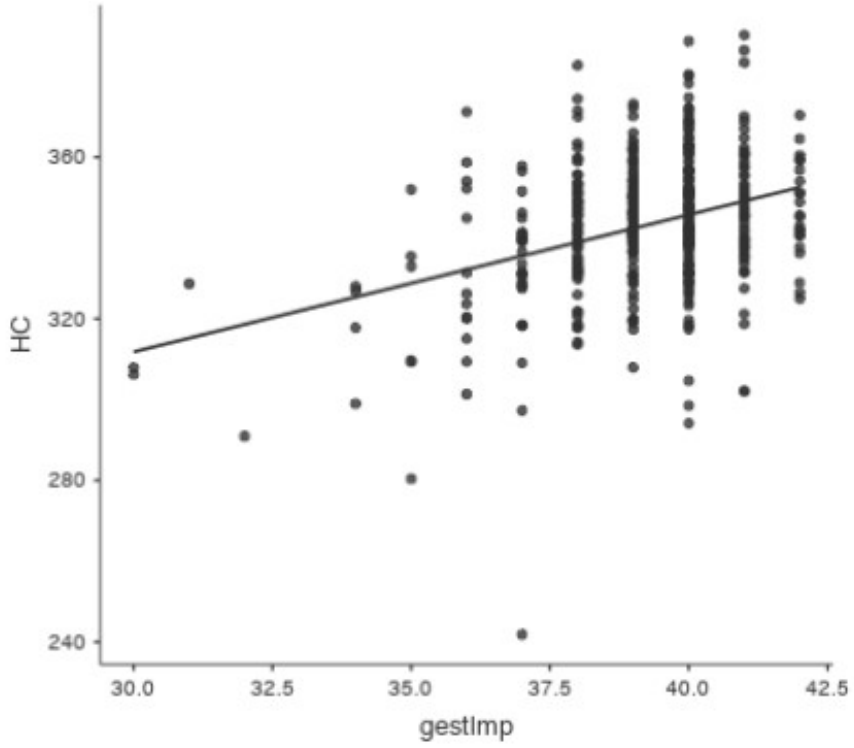
There is the visual **suggestion that the HC increases with higher gestational age**

This becomes clearer if we represented the averages of HC for the different gestational ages.



Regression Line

Furthermore, this increase seems to be approximately linear (following a straight line)



Regression Line

- We may suggest a **model for the mean of *y*** (head circumference) **as a function of *x*** (gestational age)

- As we have seen, the previous scatter plot suggests that the **mean of *y* increases linearly with *x***, i.e., the relation between the two variables can be approximated by a **straight line**

y – head circumference,
x – gestational age,
$\mu_{y|x}$ – mean of ***y*** for a value of ***x***

Equation for a straight line:

$$\mu_{y|x} = \alpha + \beta x$$

Regression Line

Equation for a straight line:

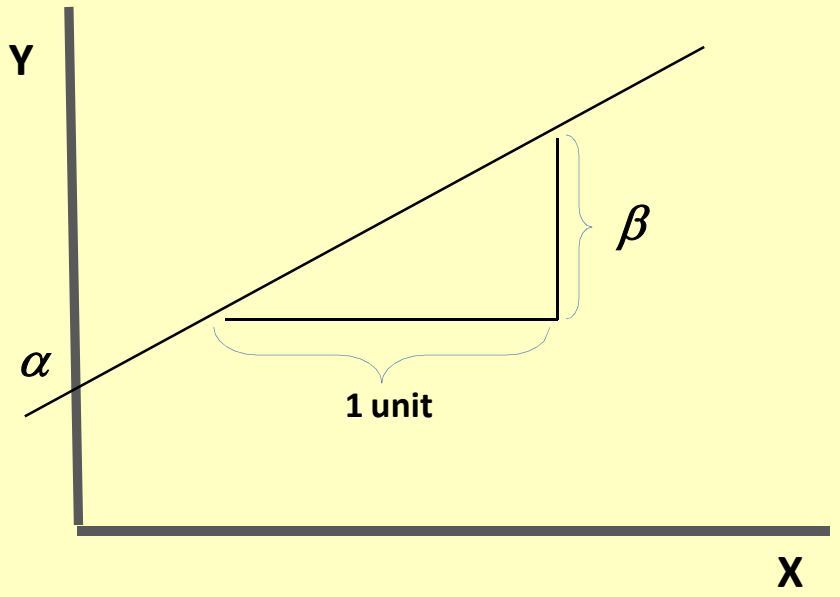
$$\mu_{y|x} = \alpha + \beta x$$

α – intercept

value of $\mu_{y|x}$ for $x = 0$

β – slope

change in $\mu_{y|x}$ for the increase of 1 unit in x



Regression Line

Equation for a straight line:

$$\mu_{y|x} = \alpha + \beta x$$

α – intercept

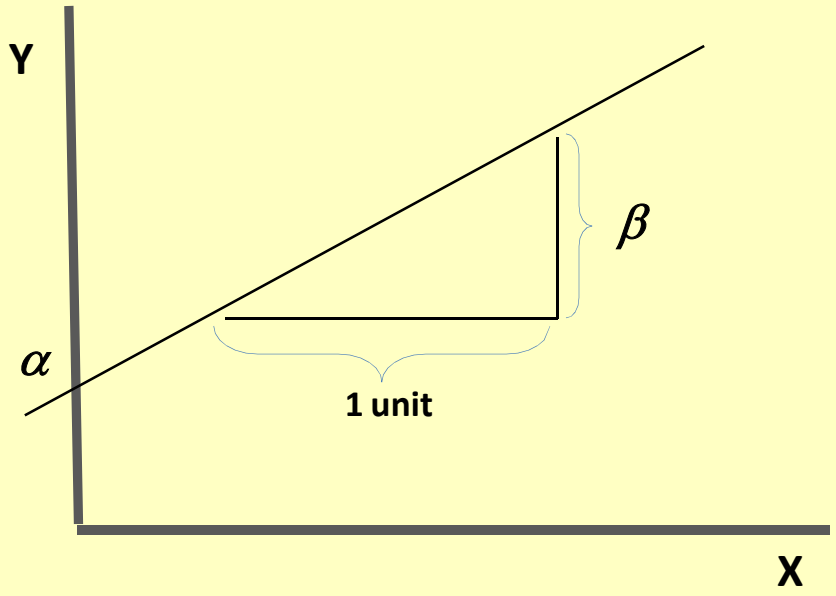
value of $\mu_{y|x}$ for $x = 0$

β – slope

change in $\mu_{y|x}$ for the increase of 1 unit in x

If X increases => y increases

β **positive**



Regression Line

Equation for a straight line:

$$\mu_{y|x} = \alpha + \beta x$$

α – intercept

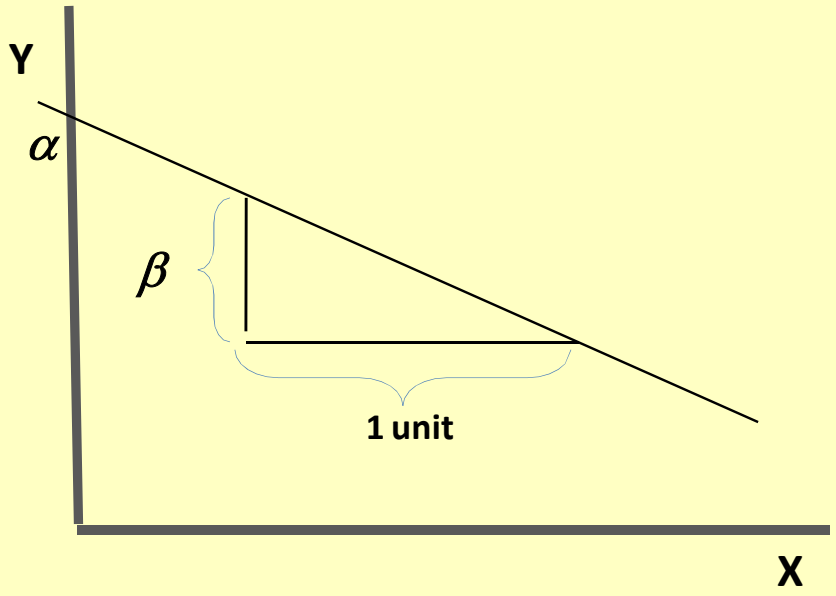
value of $\mu_{y|x}$ for $x = 0$

β – slope

change in $\mu_{y|x}$ for the increase of 1 unit in x

If X increases => y decreases

β negative



Regression Line

Model Coefficients - HC

Predictor	Estimate	SE	t	p
Intercept	210.04	17.043	12.32	<.001
gestlmp	3.39	0.434	7.81	<.001

$$\mu_{y|x} = 210 + 3.4x$$

```
> lm(ofc ~ gestlmp, data=alcohol)

Call:
lm(formula = ofc ~ gestlmp, data = alcohol)

Coefficients:
(Intercept)      gestlmp
    210.042        3.392
```

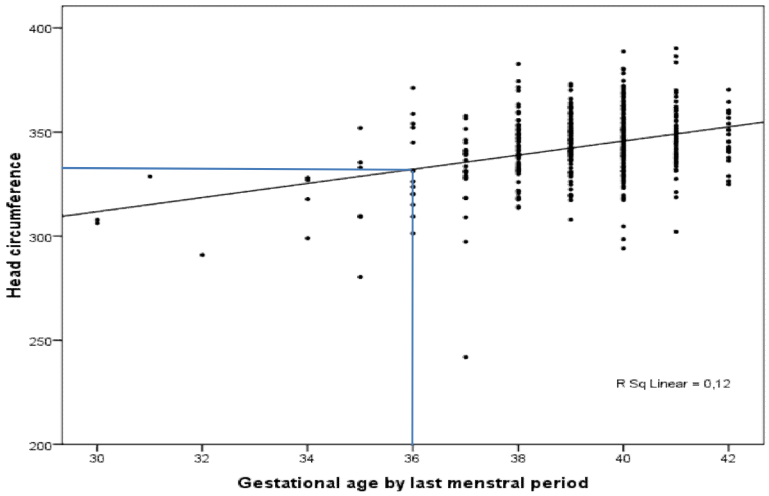
Regression Line

$$\mu_{y|x} = 210 + 3.4\,x$$

For example, for 36 weeks the model predicts an average head circumference of 332mm

(332 = 210 + 3.4 * 36)

The sample average for 36 weeks is 333 mm



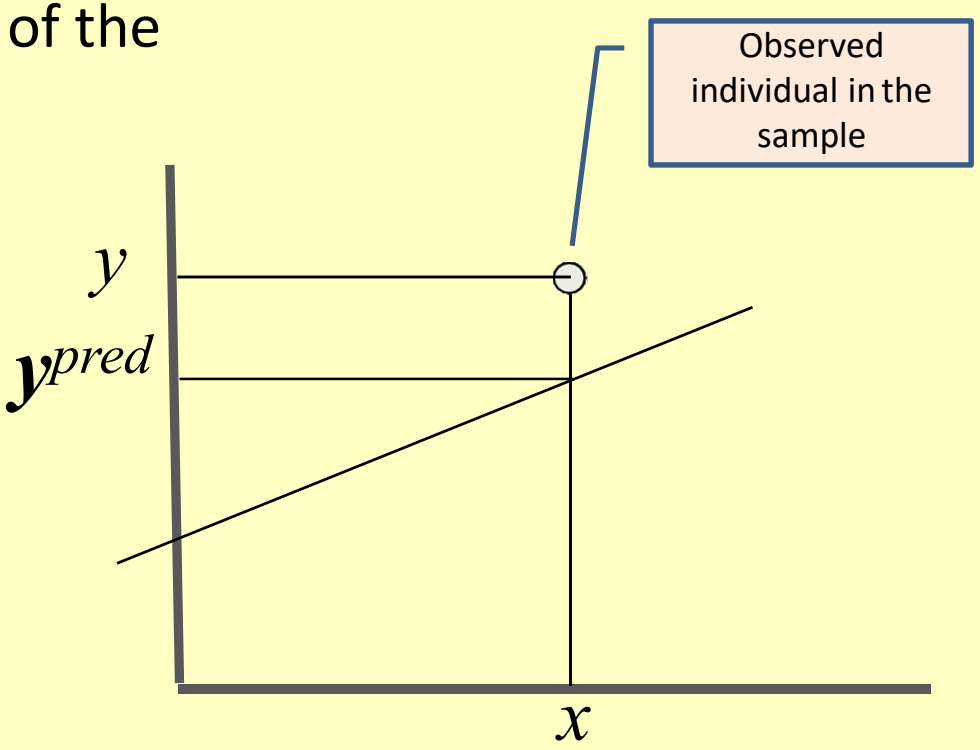
Regression Line

Another representation of the regression model

$$y = \alpha + \beta x + \varepsilon$$

ε – error
α – intercept
β – slope

$\varepsilon \sim N(0, \sigma_{y|x})$



Regression Line

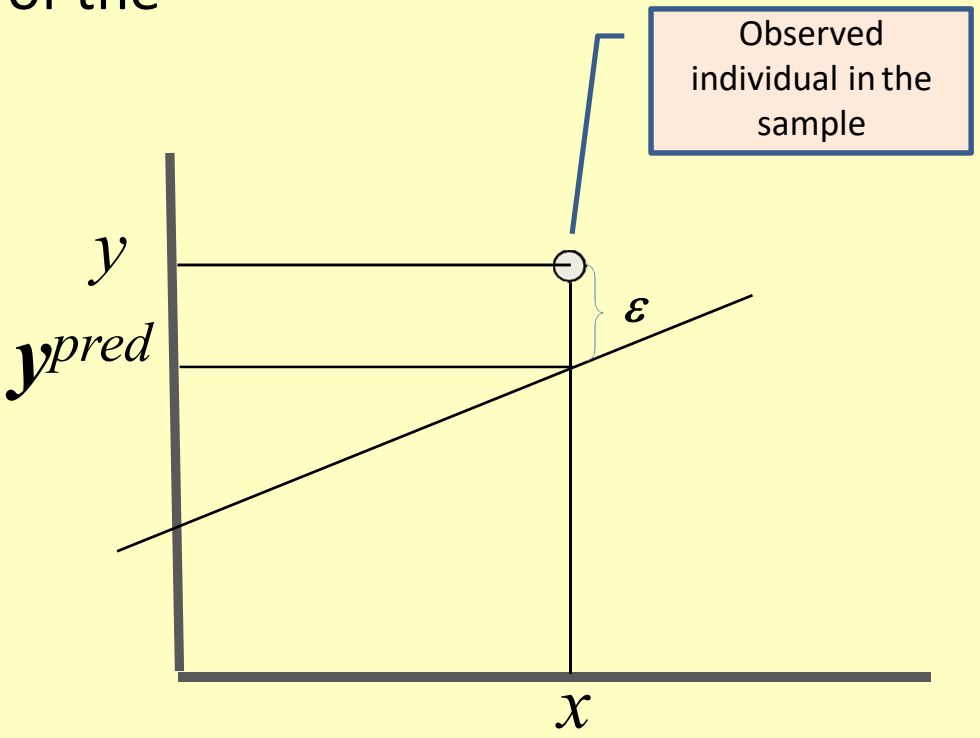
Another representation of the regression model

$$y = \alpha + \beta x + \varepsilon$$

For a specific x, the model predicts:

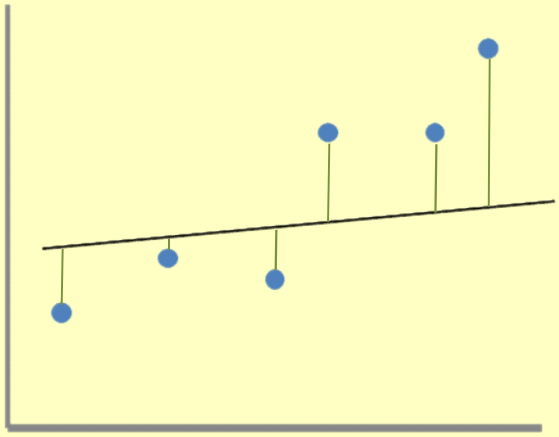
$$y^{pred} = \alpha + \beta x$$

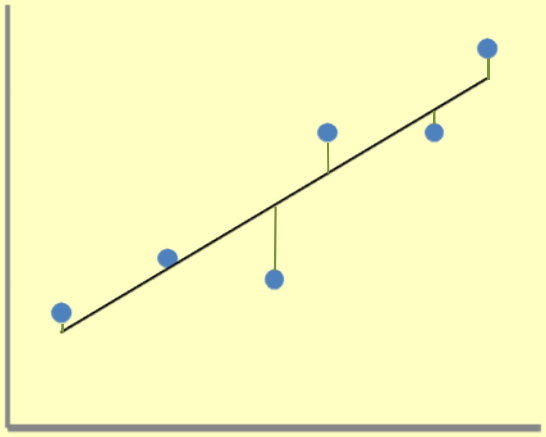
So, $y - y^{pred} = \varepsilon$



Least squares

- How do we find the line that **best fits** the data?
- In other words, how do we **estimate** α and β?
- The line that we are looking for is the line that **minimizes the errors** (ε)!





Least squares

- We have defined: $\varepsilon = y - y^{pred}$
- For each observation (individual) i we have:

$$\varepsilon_i = y_i - y_i^{pred} = y_i - \alpha - \beta x_i$$

This represents the value of x for the individual i

- The main idea is to minimize the sum of the squared errors

$$\text{sum of squares} = \sum_i \varepsilon_i^2 = \sum_i (y_i - \alpha - \beta x_i)^2$$

- The reason we square the errors is to “eliminate” the signs of the errors

Least squares

- To minimize the sum of squares we first find the zero of the first derivatives:

$$\frac{\partial \text{ sum of squares}}{\partial \alpha} = 0 \Leftrightarrow \sum -2(y_i - \alpha - \beta x_i) = 0$$

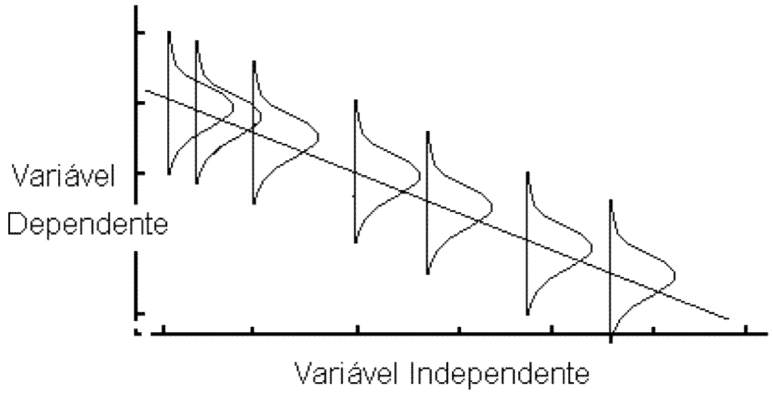
$$\frac{\partial \text{ sum of squares}}{\partial \beta} = 0 \Leftrightarrow \sum -2x_i(y_i - \alpha - \beta x_i) = 0$$

- We can show that this solution is a minimum and it is known as the ***ordinary least squares estimator (OLS)*** for the regression parameters.

Least squares

The OLS estimator is the best estimator under the following assumptions:

- The association of *x* and *y* is linear, i.e., the model for the mean of *y* is correctly specified
- The observations are independent
- Fixing *x*, *y* is normally distributed (i.e. the errors, or residuals, are normally distributed)
- Fixing *x*, the standard deviation for *y* is the same for all *x*'s (homoscedasticity)



Least squares

- The estimated regression parameters for the example

Model Coefficients - HC

Predictor	Estimate	SE	t	p
Intercept	210.04	17.043	12.32	<.001
gestimp	3.39	0.434	7.81	<.001

$\hat{\alpha}$

$\hat{\beta}$

- Note that $\hat{\alpha}$ and $\hat{\beta}$ are estimators of the true α and β (population parameters) based in the sample
- Usually, we want to make inference about β

Inference

The typical question after fitting the model is about the existence of a statistically significant "effect" of the covariate x on the outcome y

- **The "effect" of x on y is given by β**

This corresponds to test the null hypothesis H_0: $\beta = 0$

The observed data allow us to test this hypothesis. How?

- We can show that the standard error of $\hat{\beta}$ is given by:

$$se\left(\hat{\beta}\right) = \frac{\sigma_{y|x}}{\sqrt{\sum_{i=1}^{n}\left(x_i - x\right)^2}}$$

Standard deviaton of the error term ε

Inference

Typically, we do not know $\boldsymbol{\sigma}_{y|x}$ but we can estimate $se(\hat{\beta})$ using an estimator for $\boldsymbol{\sigma}_{y|x}$

So, to test the null hypothesis H_0: β=0 in the example

Model Coefficients - HC

Predictor	Estimate	SE	t	p
Intercept	210.04	17.043	12.32	<.001
gestlmp	3.39	0.434	7.81	<.001

$\hat{\beta}$

$\text{Se}(\hat{\beta})$

$\frac{\hat{\beta}}{se(\hat{\beta})}$

```
> summary(model)

Call:
lm(formula = ofc ~ gestlmp, data = alcohol)

Residuals:
<Labelled double>: Head circumference
    Min      1Q  Median      3Q     Max
-93.678  -8.908   0.112   9.684  43.769

Labels:
 value   label
   999 Missing

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 210.0417    17.0427  12.324  < 2e-16 ***
gestlmp       3.3921     0.4345   7.807 4.17e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15.84 on 448 degrees of freedom
  (4 observations deleted due to missingness)
Multiple R-squared:  0.1198,    Adjusted R-squared:  0.1178
F-statistic: 60.95 on 1 and 448 DF,  p-value: 4.171e-14
```

Inference

- So, to test the null hypothesis H_0: β=0

Model Coefficients - HC

Predictor	Estimate	SE	t	p
Intercept	210.04	17.043	12.32	<.001
gestimp	3.39	0.434	7.81	<.001

- In the example, the p-value for H_0: β=0 is <0.001

- We conclude that the true β is not zero, i.e., there is a significant association between gestational age and head circumference ("more age", "bigger head")

- The test for α is usually of no interest

Inference

- Having the standard error for the regression parameters estimates we can also compute a confidence interval (CI) for α and β
- For example, the 95% CI for β

$$CI_{95\%}(\beta) = \hat{\beta} \pm 2 \times se(\hat{\beta})$$

$$CI_{95\%}(\beta) = 3.4 \pm 2 \times 0.4$$

$$CI_{95\%}(\beta) = [2.6; 4.2]$$

Inference

- What about y^{pred} ? The predicted value for y (given the covariate x) is based on the parameters estimates! So it is also an estimated value of the "true" y^{pred}

$$\hat{y}^{pred} = \hat{\alpha} + \hat{\beta}x$$

- Therefore, it is also possible to construct a confidence interval for $\hat{y}^{pred}$
- However, we have to be specific about the meaning of y^{pred}
- Notice that if we want to estimate $\mu_{y|x}$ the expression is the same but the meaning is different:

$$\hat{\mu}_{y|x} = \hat{\alpha} + \hat{\beta}x$$

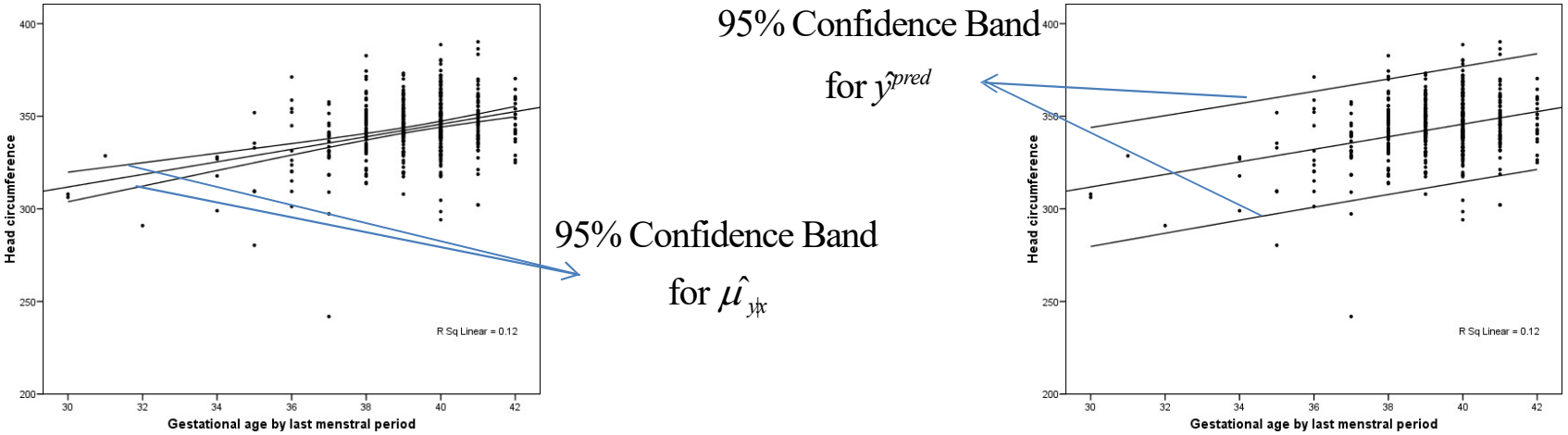
Inference

- y^{pred} is the prediction of y for **an individual** that has a certain x
- $\mu_{y|x}$ is the **average of** y for **individuals** with a certain x
- The estimates for both quantities are the same but the confidence intervals are very different.
- The reason for this has to do with the standard errors (se) of the estimates:

$$\hat{se}\left(\hat{\mu}_{y|x}\right) = s_{y|x}\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{\sum_i (x_i - x)^2}} \qquad \hat{se}\left(\hat{y}^{pred}\right) = \sqrt{s^2_{y|x} + \hat{se}\left(\hat{\mu}_{y|x}\right)}$$

Inference

- y^{pred} is the prediction of y for **an individual** that has a certain x
- $\mu_{y|x}$ is the **average of** y for **individuals** with a certain x
- Since the confidence interval for both estimates depends on x we can plot the CI for "all" x's
- The intervals obtained are called the confidence bands

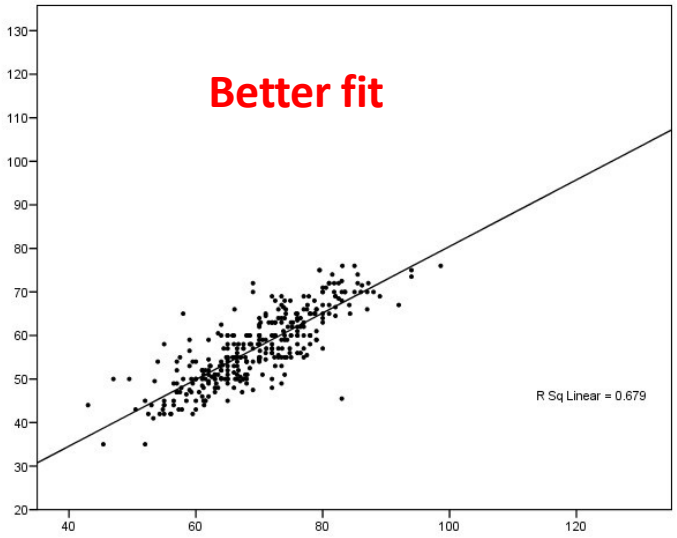


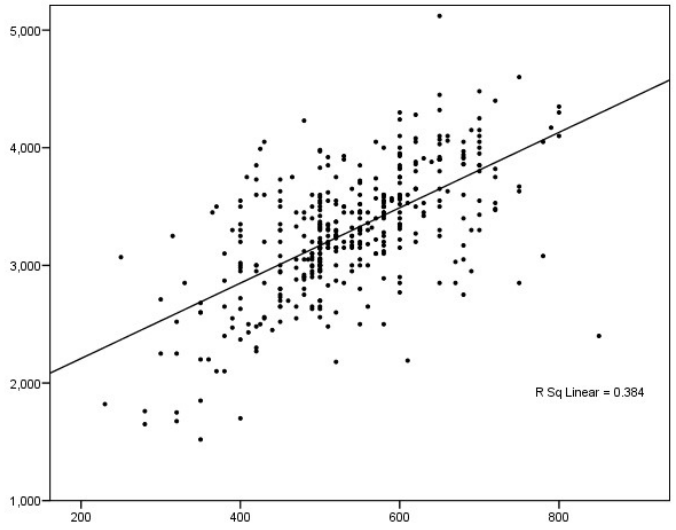
Model Evaluation

- Once we have the regression line, we should check
  - how well the model fits the data (goodness of fit), and
  - the model assumptions:
    - The model for the mean is correctly specified
    - The distribution of the residuals is normally distributed
      - Or equivalently, $y|x$ is normally distributed
    - Homoscedasticity
      - Equal variance of y for every x

Model Evaluation

- Goodness of fit
  - How well the model fits the data or
  - How well x predicts y or
  - How much of the variance of y is explained by x or
  - How good is the linear relation between x and y





A.

Model Evaluation

- We know a statistics that measures the linear relation between x and y

Pearson's correlation (r)

- r^2 gives the amount of variation on y that is explained by x

- Another way of deriving this statistics (r^2) is to write the ANOVA table

Model Evaluation

- The variation of *y* can be expressed by the sum of squares of *y*

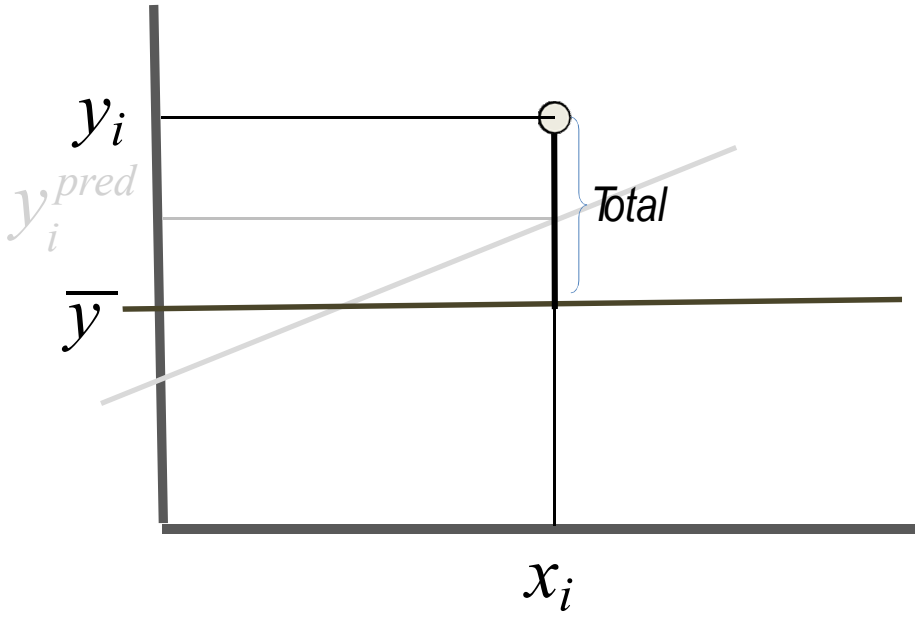
$$\text{Total sum of squares} = \sum_i \left(y_i - \bar{y}\right)^2$$

- The total sum of squares can be decomposed as:

$$\underbrace{\sum_i \left(y_i - \bar{y}\right)^2}_{\text{Total sum of squares}} = \underbrace{\sum_i \left(y_i^{pred} - \bar{y}\right)^2}_{\text{Explained sum of squares}} + \underbrace{\sum_i \left(y_i - y_i^{pred}\right)^2}_{\text{Residual sum of squares}}$$

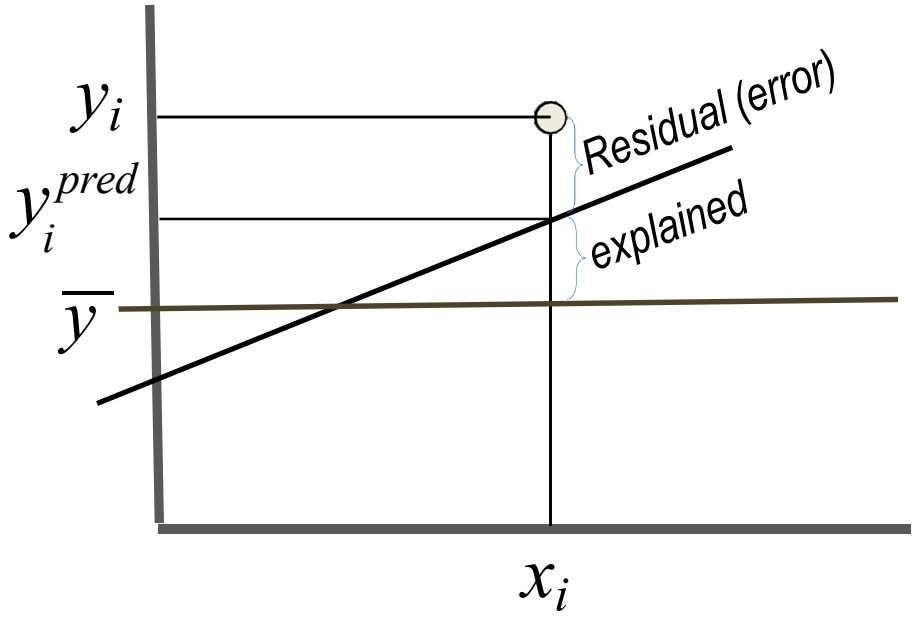
Model Evaluation

$$\underbrace{\sum_i \left(y_i - \bar{y}\right)^2}_{\text{Total sum of squares}} = \underbrace{\sum_i \left(y_i^{pred} - \bar{y}\right)^2}_{\text{Explained sum of squares}} + \underbrace{\sum_i \left(y_i - y_i^{pred}\right)^2}_{\text{Residual sum of squares}}$$



Model Evaluation

$$\underbrace{\sum_i \left(y_i - \bar{y}\right)^2}_{\text{Total sum of squares}} = \underbrace{\sum_i \left(y_i^{pred} - \bar{y}\right)^2}_{\text{Explained sum of squares}} + \underbrace{\sum_i \left(y_i - y_i^{pred}\right)^2}_{\text{Residual sum of squares}}$$



Model Evaluation

```
> anova(model)
Analysis of Variance Table

Response: ofc
           Df Sum Sq Mean Sq F value    Pr(>F)
gestlmp     1  15291 15290.8  60.955 4.171e-14 ***
Residuals 448 112383   250.9
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Model Evaluation

$$\underbrace{\sum_i \left(y_i - \bar{y}\right)^2}_{\text{Total sum of squares}} = \underbrace{\sum_i \left(y_i^{pred} - \bar{y}\right)^2}_{\text{Explained sum of squares}} + \underbrace{\sum_i \left(y_i - y_i^{pred}\right)^2}_{\text{Residual sum of squares}}$$

Omnibus ANOVA Test

	Sum of Squares	df	Mean Square	F	p
gestimp	15291	1	15291	61.0	<.001
Residuals	112383	448	251		

Note. Type 3 sum of squares

Model Evaluation

- How much of the variance of y is explained by x ?

$$r^2 = \frac{\text{Explained sum of squares}}{\text{Total sum of squares}}$$

$$r^2 = \frac{15\ 290}{127\ 674} = 0.12$$

Model Fit Measures

Model	R	R^2
1	0.346	0.120

```
> with(anova(model), get("Sum Sq")[1] / sum(get("Sum Sq")))
[1] 0.1197645
```

- So, r^2 may be used as a mesure of **goodness of fit**

Model Evaluation

- Note:
  - An inference question could be formulated regarding the sum squares:
    - Is the amount of variation explained by the model (explained sum of squares) significantly different from zero?

Omnibus ANOVA Test

	Sum of Squares	df	Mean Square	F	p
gestImp	15291	1	15291	61.0	<.001
Residuals	112383	448	251		

Note. Type 3 sum of squares

```
> anova(model)
Analysis of Variance Table

Response: ofc
           Df Sum Sq Mean Sq F value    Pr(>F)
gestImp     1  15291 15290.8  60.955 4.171e-14 ***
Residuals 448 112383   250.9
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- For the example, $p<0.001$ so we can conclude that the amount of variation explained by the model is significantly different from 0

Model Evaluation

- Note:
  - Intuitively the last question may seem related to a previous inference question about the statistically significant "effect" of the covariate x on the outcome y:

$$H_0: \beta=0$$

  - In fact both tests (testing β and testing the explained sum of squares) are equivalent for the case of simple linear regression

Regression line Model Evaluation

- Note:

Omnibus ANOVA Test

	Sum of Squares	df	Mean Square	F	p
gestimp	15291	1	15291	61.0	<.001
Residuals	112383	448	251		

Note. Type 3 sum of squares

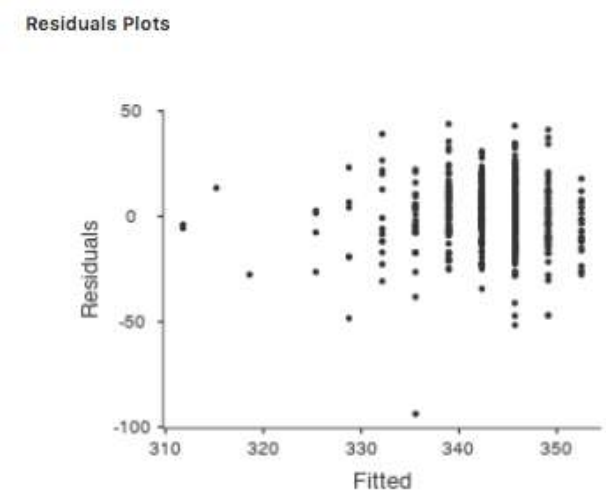
Model Coefficients - HC

			95% Confidence Interval			
Predictor	Estimate	SE	Lower	Upper	t	p
Intercept	210.04	17.043	176.55	243.54	12.32	<.001
gestimp	3.39	0.434	2.54	4.25	7.81	<.001

$t^2=F \quad => \quad 7.81^2 = 61$

Model Evaluation

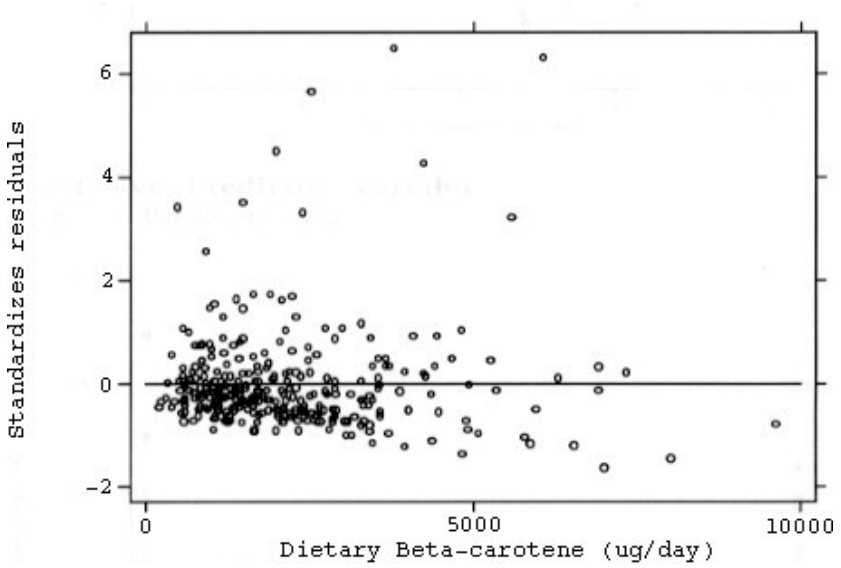
- A common way of checking the model assumptions is to look at the residuals



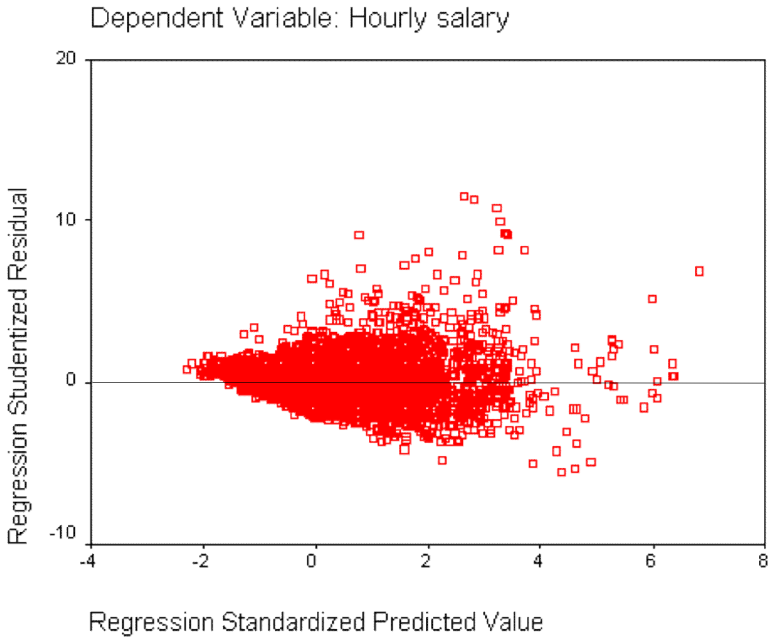
- The points should scatter around zero with no clear pattern and with similar dispersion.

Model Evaluation

- Violation of assumptions:



Violation of normality



Violation of homoscedasticity

HEADS

PhD PROGRAMME IN HEALTH DATA SCIENCE

Cheers!

Title

Simple Linear Regression

Acknowledgments

Armando Teixeira Pinto, University of Sydney, Australia

Basics of Health Intelligent Data Analysis

PhD Programme in Health Data Science

Porto, 2nd of December, 2019

Cláudia Camila Dias

Pedro Pereira Rodrigues

U.PORTO FMUP FACULDADE DE MEDICINA UNIVERSIDADE DO PORTO

MEDCIDS DEPARTAMENTO DE MEDICINA DA COMUNIDADE, INFORMAÇÃO E DECISÃO EM SAÚDE

CINTESIS Health. Research.