Utrecht University

11 February 2022

Lecture 2: Prediction, effect sizes

Peter van der Heijden

Professor of statistics for social and behavioural sciences,
Methods and Statistics, UU

Utrecht University

Outline

1. Linear regression

- Confidence interval for mean value of y

- Confidence interval for individual predictions of y

2. Classification by logistic regression

- Introduction

- Confidence interval for $\text{logit}(\pi[x_o])$

3. Effect sizes

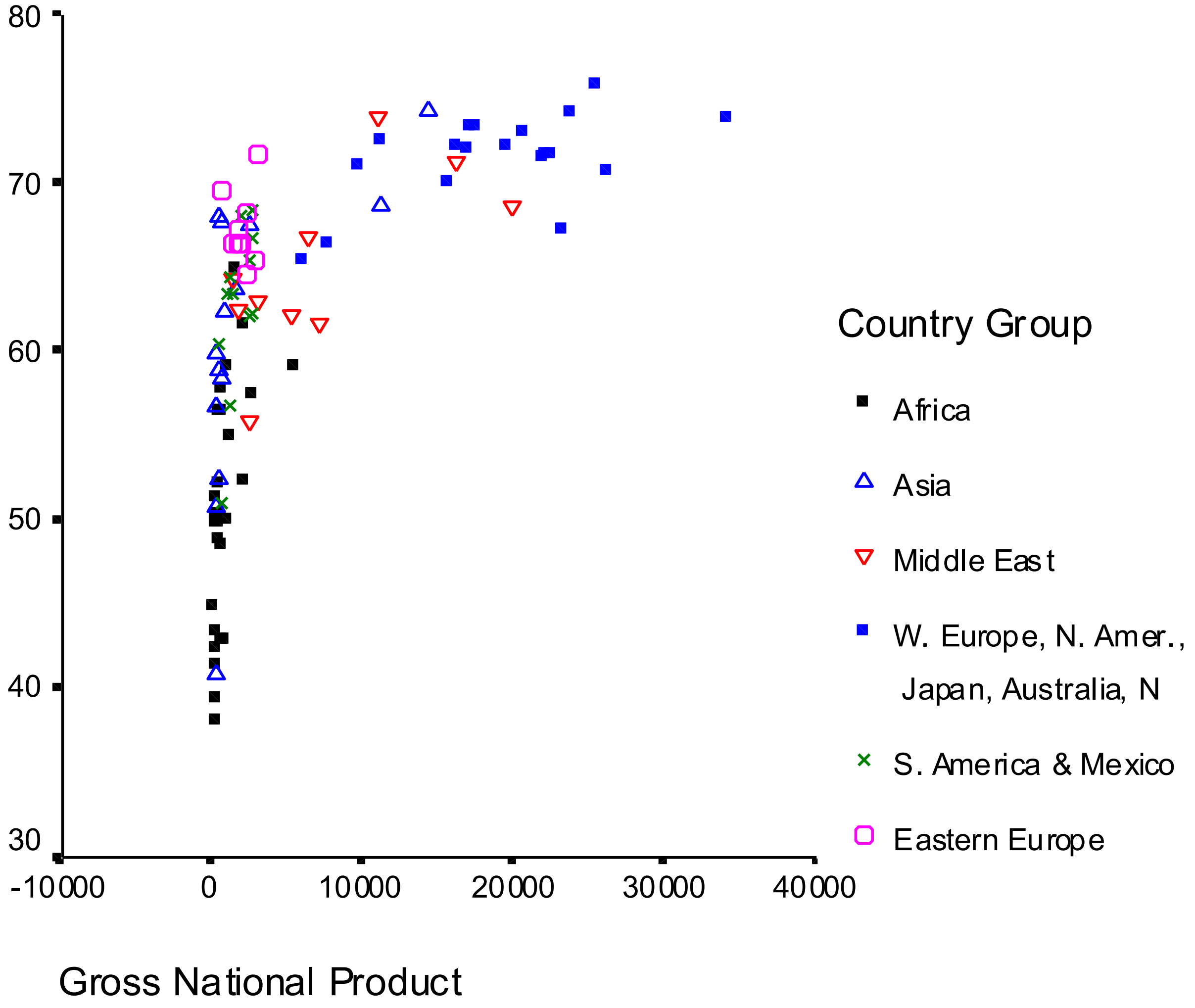
Example: GNP predicting male life expectancy

MLE vs GNP not linear

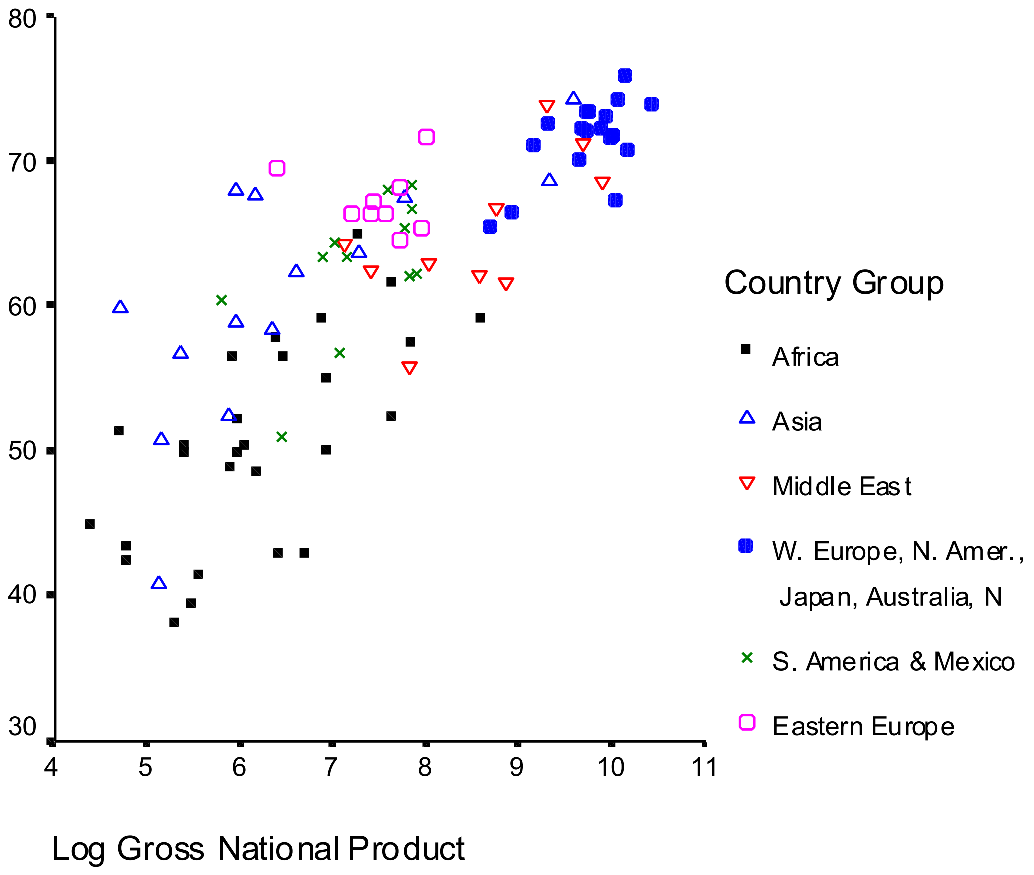
Fitting straight line not good idea

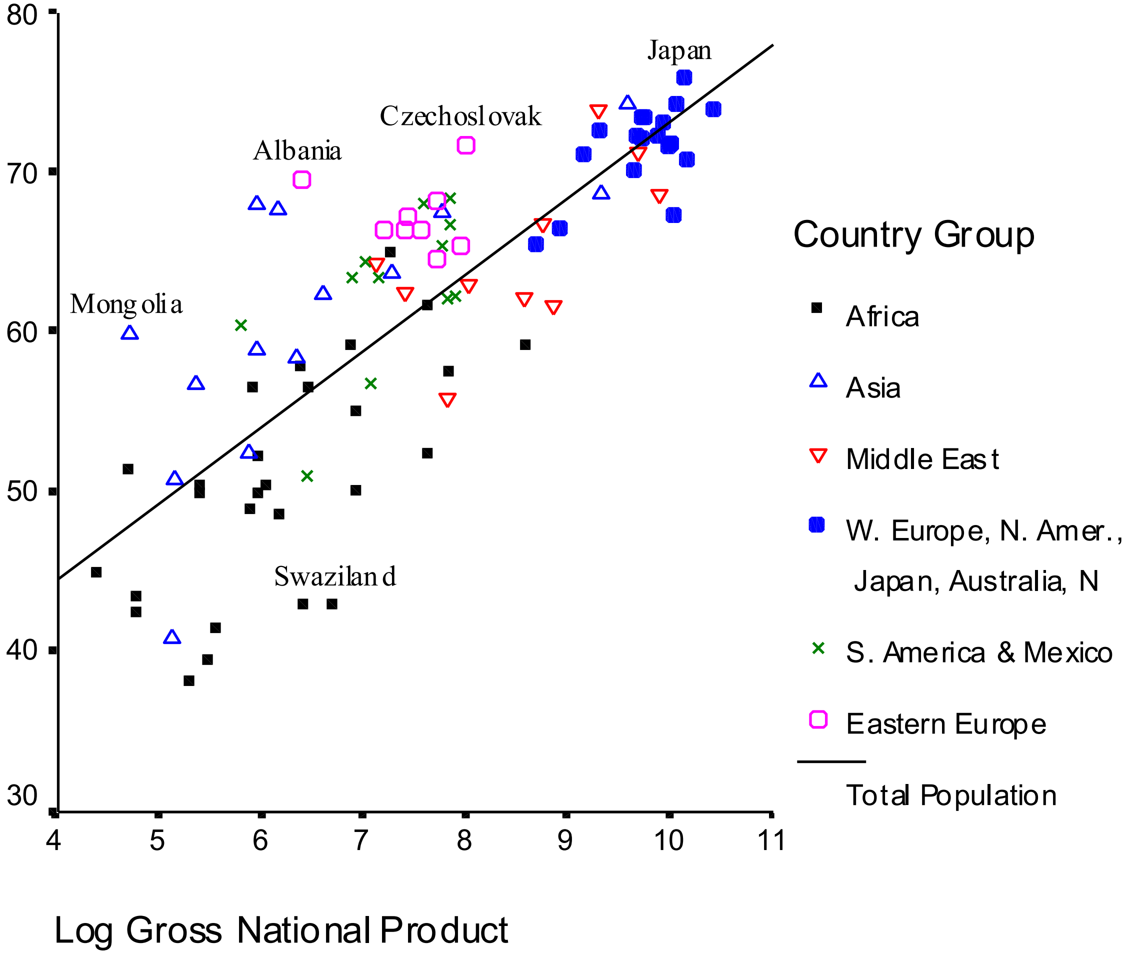
We take log

Often good idea with financial data, data more normally distributed



Straight line appears to be OK -> linear regression

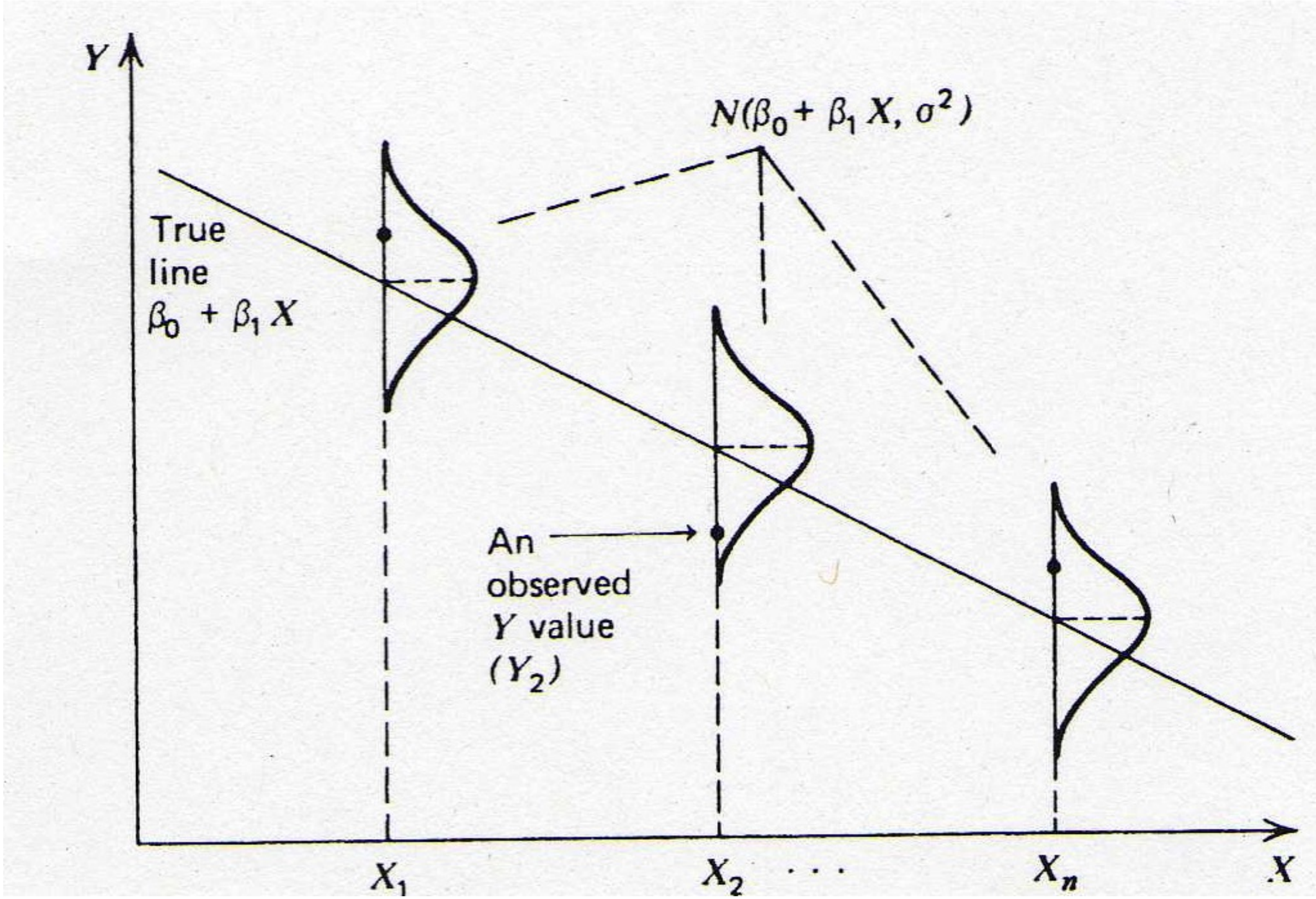




male life expectancy = 25.5 + 4.8 × log(GNP)

Utrecht University

Model assumptions:


Y
$N(\beta_0 + \beta_1 X, \sigma^2)$
True line $\beta_0 + \beta_1 X$
An observed Y value (Y_2)
X_1
X_2
. . .
X_n
X

- Linearity of relation
- Observed y normally distributed around regression line
- Normal distribution constant over regression line
- Independent observations

So we want to find $\hat{y}_0 \quad = \quad \hat{\beta}_o + \hat{\beta}_1 x_0$

i.e. estimates for β_o, β_1, their standard errors and σ,

σ estimated using departures from regression line

Utrecht University

Confidence Intervals for β_o and β_1

The confidence intervals for β_o and β_1 at the (1 - α)% confidence level and using the t-distribution are

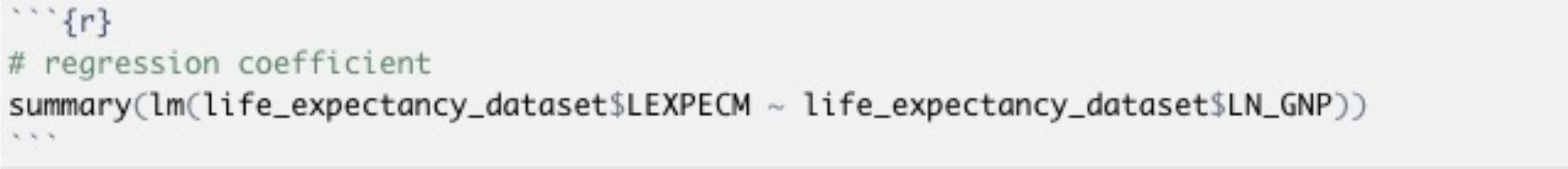
$$\hat{\beta}_o \quad \pm \quad t_v(1-\alpha/2)\,\hat{se}(\hat{\beta}_o)$$

$$\hat{\beta}_1 \quad \pm \quad t_v\,(1-\alpha/2)\,\hat{se}(\hat{\beta}_1)$$

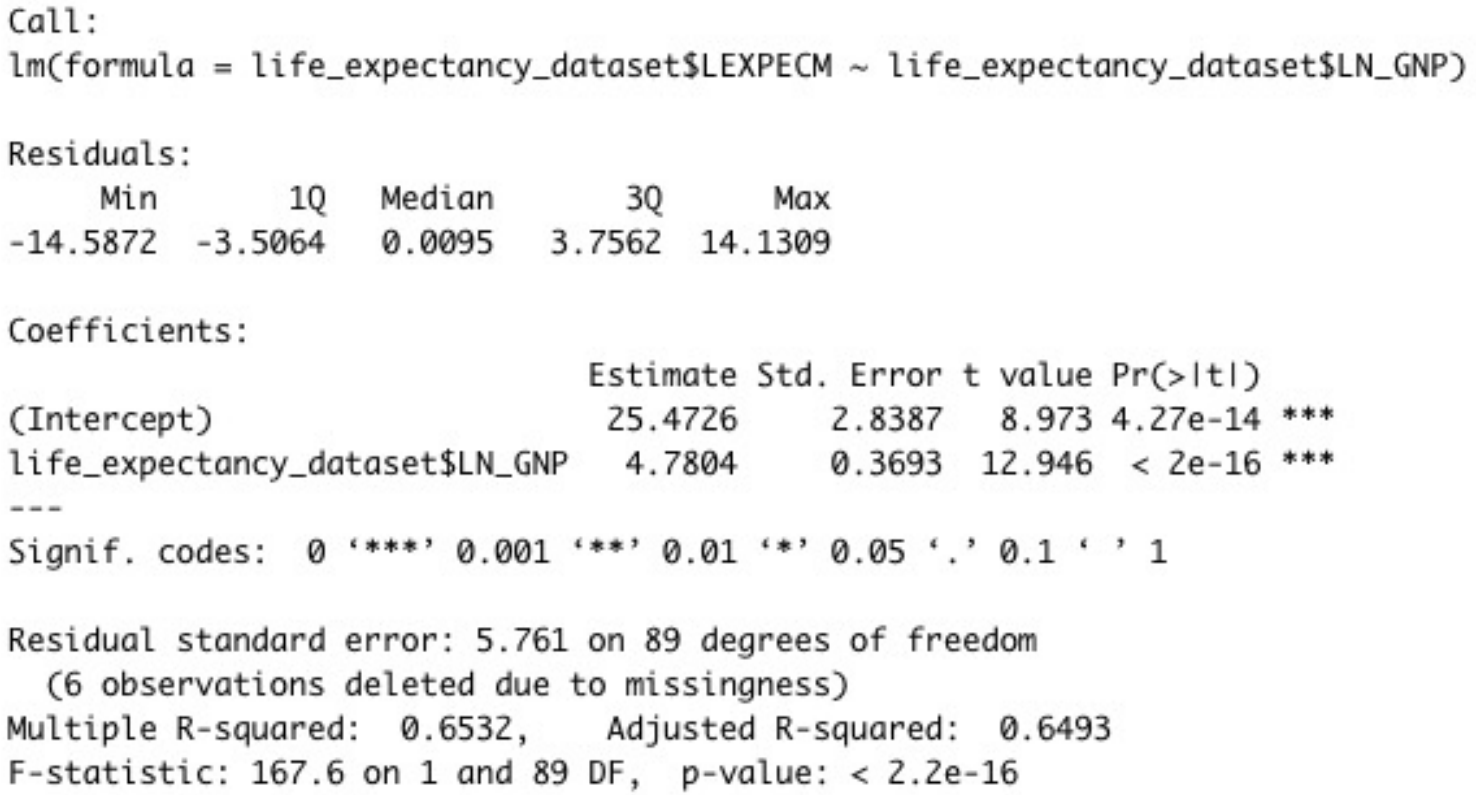
where t_v(1- α/2) is the upper quantile of the t-distribution with v = **n** - 2 degrees of freedom.

Utrecht University

Example: Life Expectancy



~~~
```{r}
# regression coefficient
summary(lm(life_expectancy_dataset$LEXPECM ~ life_expectancy_dataset$LN_GNP))
```
~~~



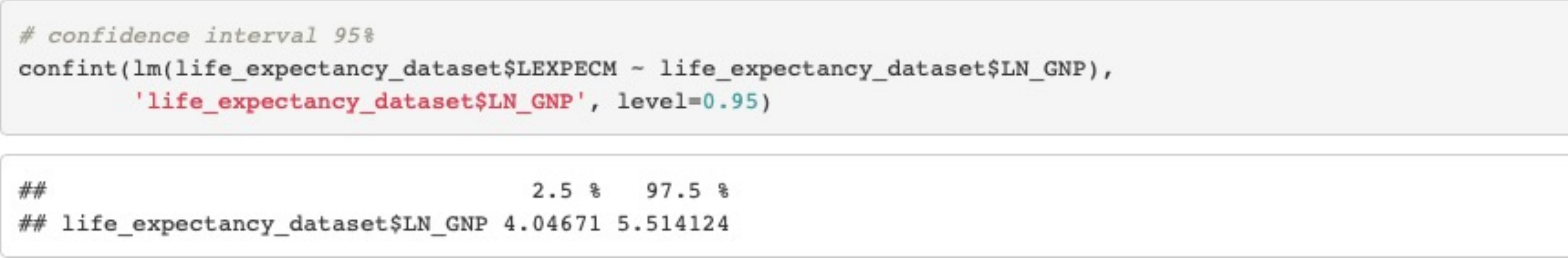
```
Call:
lm(formula = life_expectancy_dataset$LEXPECM ~ life_expectancy_dataset$LN_GNP)

Residuals:
     Min       1Q   Median       3Q      Max 
-14.5872  -3.5064   0.0095   3.7562  14.1309 

Coefficients:
                               Estimate Std. Error t value Pr(>|t|)    
(Intercept)                     25.4726     2.8387   8.973 4.27e-14 ***
life_expectancy_dataset$LN_GNP   4.7804     0.3693  12.946  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.761 on 89 degrees of freedom
  (6 observations deleted due to missingness)
Multiple R-squared:  0.6532,	Adjusted R-squared:  0.6493 
F-statistic: 167.6 on 1 and 89 DF,  p-value: < 2.2e-16
```

Example: Life Expectancy

95% CI for β_1: $4.780 \pm 1.987 \times 0.369 = (4.05, 5.51)$

95% CI for β_0: $25.473 \pm 1.987 \times 2.839 = (19.83, 31.11)$

$t_{89,\,0.025} = 1.987$



```
# confidence interval 95%
confint(lm(life_expectancy_dataset$LEXPECM ~ life_expectancy_dataset$LN_GNP),
        'life_expectancy_dataset$LN_GNP', level=0.95)
```

```
##                                  2.5 %   97.5 %
## life_expectancy_dataset$LN_GNP 4.04671 5.514124
```

Confidence Interval for Mean Value of y

Best estimate of **mean** y value at a **specific** value $x = x_0$ is obtained by using the prediction from the fitted regression line

$$\hat{y}_0 \quad = \quad \hat{\beta}_o + \hat{\beta}_1 x_0$$

The estimated standard error is

$$\hat{se}(\hat{y}_0) \quad = \quad s\left\{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right\}^{1/2}$$

Note the greater the distance x_0 from mean, the larger the error, i.e. best predictions in the ‘middle’ of the range and worse predictions at the ‘ends’.

Predictions become worse the further away from the observed range of x.

So, the 100(1-α)% confidence interval for the **mean** value of y when x = x_0 is

$$(\hat{\beta}_0 + \hat{\beta}_1 x_0) \ \pm \ t_{n-2}(1-\alpha/2) \times \hat{s}e(\hat{y}_0)$$

Utrecht University

Example: Life Expectancy

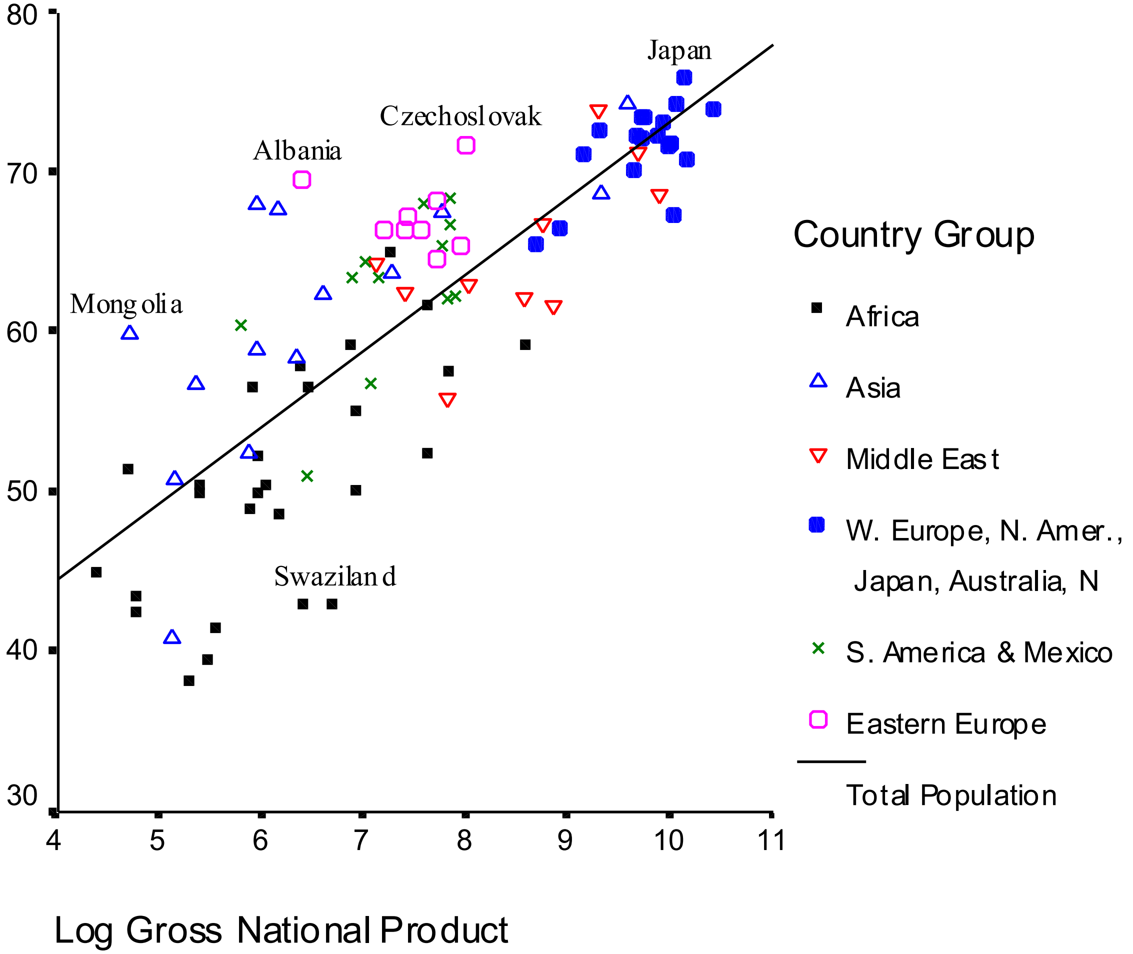
So a 95% CI for the mean life expectancy of countries with log(GNP) = 10 is

$$(25.473 + 4.780 \times 10) \pm 1.987 \times 5.761 \sqrt{\frac{1}{91} + \frac{(10 - 7.512)^2}{243.45}}$$

$$= 73.273 \pm 2.184$$

= (**71.09**, **75.46**).

Note: R calculates the estimated standard error for the predicted **mean** y value and the CI at every **observed** value of x in your dataset.



For Log GNP = 10,

CI of mean is (71.09, 75.46).

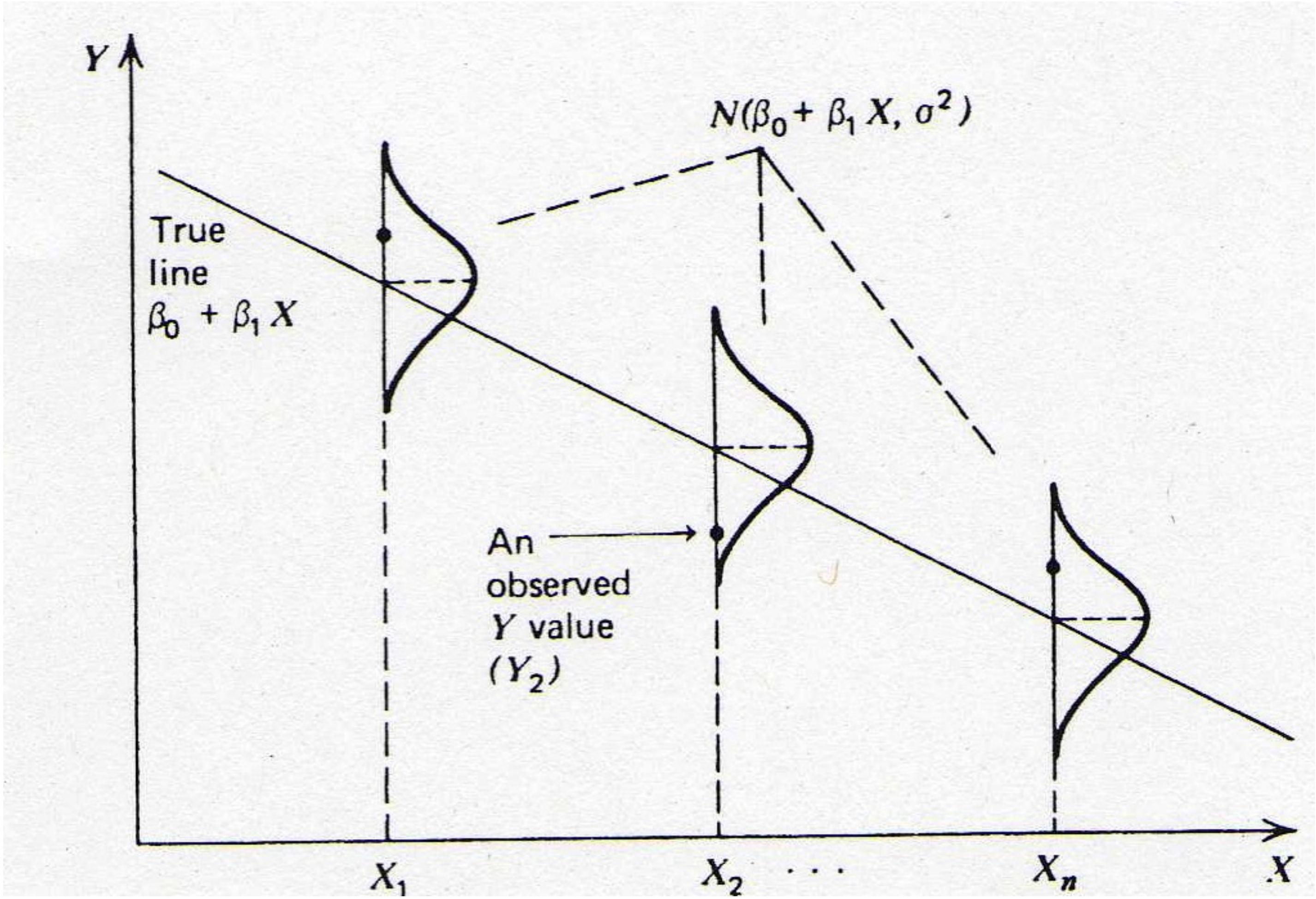
male life expectancy = 25.5 + 4.8 × log(GNP)

Utrecht University

Prediction: Individual Values of y

We can also predict a future or unknown value of y given x = x_0 and construct an interval for this prediction. The predicted value is again given by

$$\hat{y}_0 \quad = \quad \hat{\beta}_o + \hat{\beta}_1 x_0$$


Y
$N(\beta_0 + \beta_1 X, \sigma^2)$
True line
$\beta_0 + \beta_1 X$
An observed Y value (Y_2)
X_1
X_2
. . .
X_n
X

But, **the interval is not the same** as that for the mean of y at x_0 since a future value of y is assumed to be normally distributed about the line and so the extra variation about the line must be taken into account.

The estimated standard error for mean

$$\hat{\mathrm{s}}\mathrm{e}(\hat{\mathrm{y}}_0) \quad = \quad \mathrm{s}\left\{\frac{1}{\mathrm{n}} + \frac{(\mathrm{x}_0 - \bar{\mathrm{x}})^2}{\sum_{\mathrm{i}=1}^{\mathrm{n}}(\mathrm{x}_\mathrm{i} - \bar{\mathrm{x}})^2}\right\}^{1/2}$$

The estimated standard error for individual prediction is

$$\hat{\mathrm{s}}\mathrm{e}(\hat{\mathrm{y}}_0) \quad = \quad \mathrm{s}\left\{1 + \frac{1}{\mathrm{n}} + \frac{(\mathrm{x}_0 - \bar{\mathrm{x}})^2}{\sum_{\mathrm{i}=1}^{\mathrm{n}}(\mathrm{x}_\mathrm{i} - \bar{\mathrm{x}})^2}\right\}^{1/2}$$

Utrecht University

Example: Life Expectancy

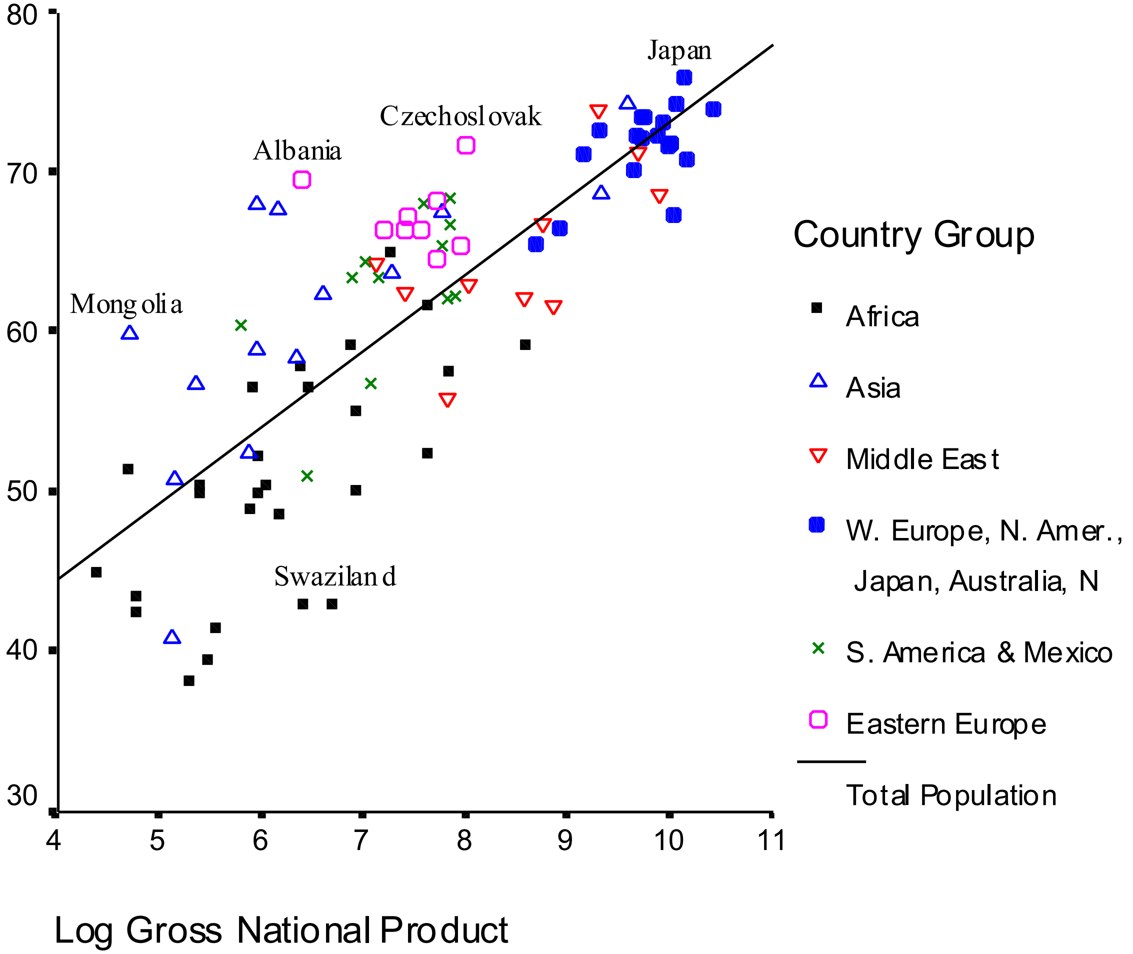
A 95% CI for the life expectancy of a particular country with log(GNP) = 10 is

$$25.473 \; + \; 4.780{\times}10 \;\; \pm \;\; 1.987 \times 5.761 \sqrt{1 + \frac{1}{91} + \frac{(10 - 7.512)^2}{243.45}}$$

$$= 73.273 \pm 11.654$$

= (**61.62**, **84.93**).

Compare this with (**71.09**, **75.46**) for the mean value at log(GNP) = 10.



For Log GNP = 10,

CI of mean is (71.09, 75.46),

CI of individual prediction is (61.62, 84.93).

male life expectancy = 25.5 + 4.8 × log(GNP)

Conclusions of 1:

- There is a difference between CI for predicted mean and CI for predicted individual value,
- Where the latter is much larger,
- And with larger n, CI for predicted mean vanishes as

$$\hat{\mathrm{se}}(\hat{\mathrm{y}}_0) = \mathrm{s}\left\{\frac{1}{\mathrm{n}} + \frac{(\mathrm{x}_0 - \bar{\mathrm{x}})^2}{\sum_{\mathrm{i}=1}^{\mathrm{n}}(\mathrm{x}_\mathrm{i} - \bar{\mathrm{x}})^2}\right\}^{1/2}$$

- CI depends on distance x-value from mean of x
- The further away from the mean, the larger the CI’s,
- As regression line wiggles more at extremes

- Do not use regression line to predict outside range of x’s

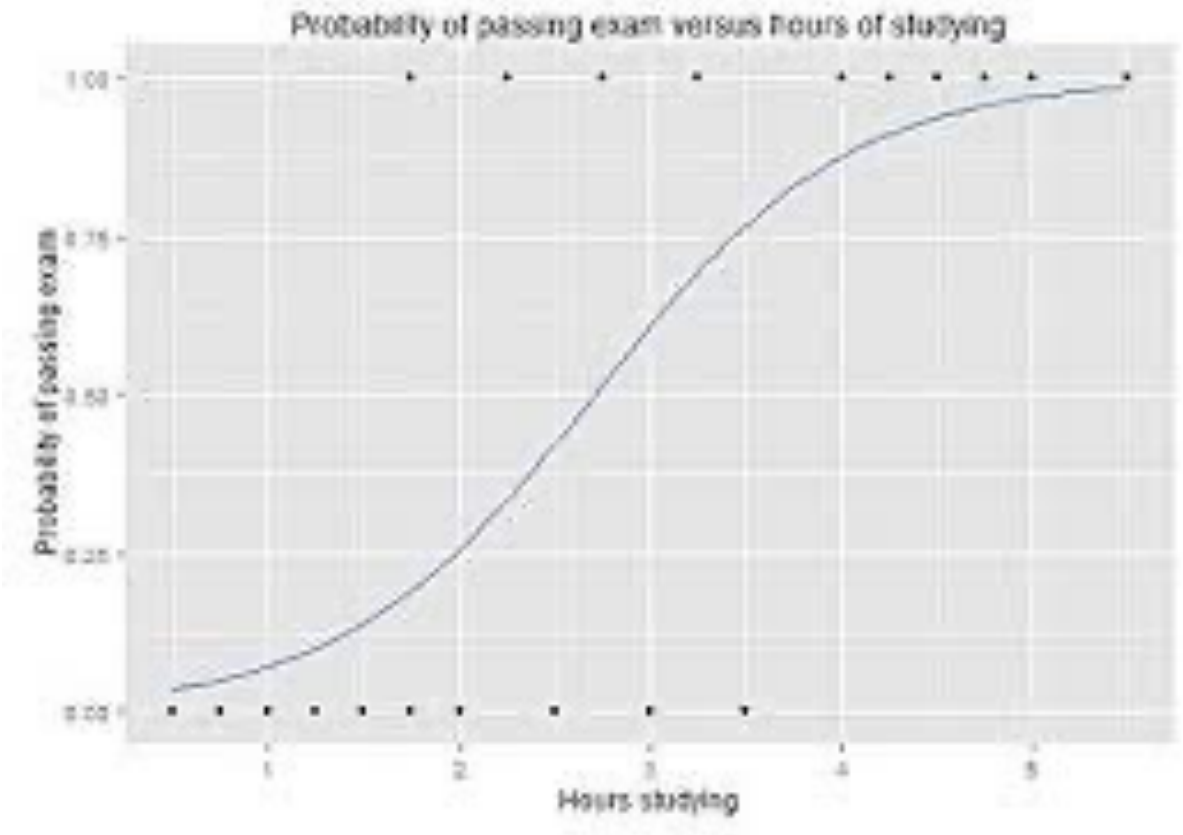
2. Classification by logistic regression

- Introduction
- Parameter interpretation
- Classification
- Prediction

Introduction

Predict dichotomous outcome

Non-linear function to predict probability of success



Three equivalent ways to formulate logistic regression

(1) Linear model for logit scale

$$\log_{\mathrm{e}}\left(\frac{\pi}{1-\pi}\right) = \mathrm{logit}(\pi)$$

$$= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k$$

(2) Model for odds

$$\frac{\pi}{1-\pi} = e^{(\beta_0+\beta_1 X_1+\beta_2 X_2+\ldots+\beta_k X_k)}$$

(3) Model for probability

$$\pi = \frac{e^{(\beta_0+\beta_1 X_1+\beta_2 X_2+\ldots+\beta_k X_k)}}{1+e^{(\beta_0+\beta_1 X_1+\beta_2 X_2+\ldots+\beta_k X_k)}}$$

Example : Birth weight

Is the probability of a normal birth weight, as opposed to a low birth weight related to gestational age of the human foetus (number of weeks from conception to birth)?

Dependent variable is BWGHT, coded 1 = normal ; 0 = low

Independent variable (continuous) is GAGE

Data for 24 babies (7 of whom were classified as having low weight at birth).

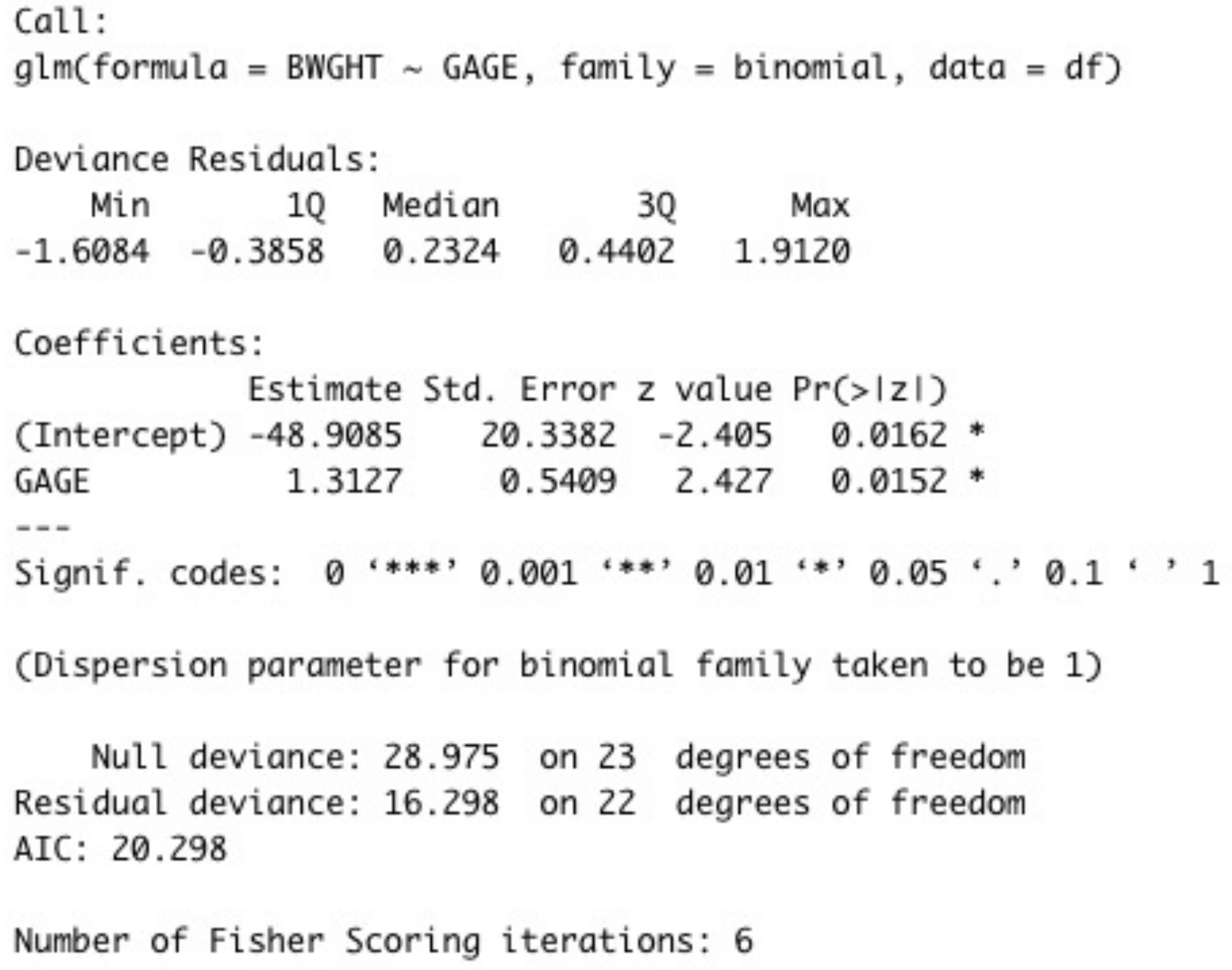
Utrecht University

Model with GAGE

R output

Our fitted regression model is

$$\text{logit}(\hat{\pi}) = -48.91 + 1.31\,GAGE$$



```
Call:
glm(formula = BWGHT ~ GAGE, family = binomial, data = df)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6084  -0.3858   0.2324   0.4402   1.9120

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -48.9085    20.3382  -2.405   0.0162 *
GAGE          1.3127     0.5409   2.427   0.0152 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 28.975  on 23  degrees of freedom
Residual deviance: 16.298  on 22  degrees of freedom
AIC: 20.298

Number of Fisher Scoring iterations: 6
```

Interpretation on the logit scale

A unit increase in X increases logit(p) by β_1

Birth weight

for every extra week of gestational age, the *logit* of the probability of a normal birth weight increase by 1.31 units, on average

OR

a unit increase in GAGE increases the log odds of normal birth weight by 1.313, on average.

Interpretation on the odds scale

- On the odds scale:

$$\text{odds} = \frac{\hat{\pi}}{1-\hat{\pi}} = e^{-48.908+1.313GAGE}$$

$$= e^{-48.908} e^{1.313GAGE}$$

as exp (a+b) = exp(a)exp(b). Interpret as:

A one week increase in GAGE changes the odds of normal birth weight multiplicatively by a factor equal to $e^{1.31}$, i.e. by 3.716.

Utrecht University

Interpretation on the odds scale

Odds scale measures effects on a percentage scale.

So can interpret as the percentage change in the odds of a success for a unit change in *X*

A one week increase in gestational age increases the odds of a normal birth weight by

$$100 \times [e^{1.31} - 1]\% = 271\%$$

Interpretation on the odds scale

Percentage increase if $e^{\beta_1} > 1 \qquad (\beta_1 > 0)$

Percentage decrease if $e^{\beta_1} < 1 \qquad (\beta_1 < 0)$

Examples: $\beta_1 = 1 \Rightarrow e^{\beta_1} = 2.72$

or a 172% increase in the odds of a success

$\beta_1 = -0.21 \Rightarrow e^{\beta_1} = 0.81$

or a 19% decrease in the odds of a success

Interpretation on the probability scale

The ‘baseline’ probability $\hat{\pi} = \dfrac{\mathrm{e}^{\beta_0}}{1 + \mathrm{e}^{\beta_0}}$

is the probability when *x* equals zero.

For the birth weight example

$$\hat{\pi} = \frac{\exp(\text{-}48.9 + 1.31 \times \mathrm{GAGE})}{1 + \exp(\text{-}48.9 + 1.31 \times \mathrm{GAGE})}$$

Interpretation on the probability scale

The estimated probability of a baby with gestational age 39 weeks having a normal birth weight is

$$\hat{\pi} = \frac{\exp(\text{-}48.9 + 1.31 \times 39)}{1 + \exp(\text{-}48.9 + 1.31 \times 39)}$$
$$= 0.90$$

Baseline probability does not have a sensible interpretation for the birth weight example

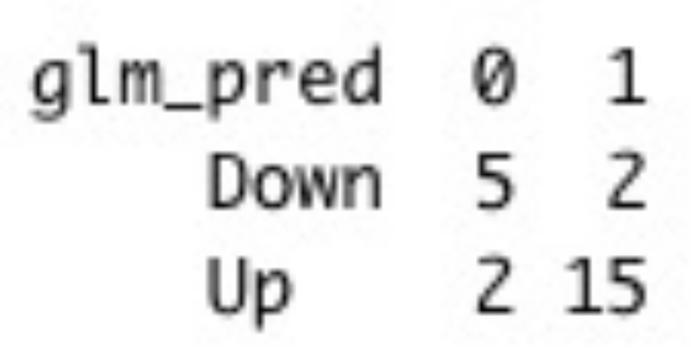
Proportion with CHD by age

GAGE	BWGHT	$\hat{\pi}$	Pr	GAGE	BWGHT	$\hat{\pi}$	Pr
35	0	0.050	0	39	1	0.909	1
36	0	0.162	0	39	1	0.909	1
36	0	0.162	0	40	1	0.974	1
36	**1**	0.162	**0**	40	1	0.974	1
37	0	0.419	0	40	1	0.974	1
37	0	0.419	0	40	1	0.974	1
37	**1**	0.419	**0**	40	1	0.974	1
38	**0**	0.728	**1**	40	1	0.974	1
38	**0**	0.728	**1**	40	1	0.974	1
38	1	0.728	1	40	1	0.974	1
38	1	0.728	1	41	1	0.993	1
38	1	0.728	1	42	1	0.998	1

0 = low

1 = normal

R Classification Table Summary



```
glm_pred  0  1
    Down  5  2
    Up    2 15
```

0 = low
1 = normal

5/7 = sensitivity
15/17 = specificity

The higher the overall %age of correct predictions (in this case 20/24 = **83%, is accuracy**) the better the model.

But note, without any predictor, if we place all infants in 1, we have 17/24 = 71%

Prediction in logistic regression

Observed scores are 0, 1, predicted is probability

$$\log_e\left(\frac{\pi}{1-\pi}\right) \quad = \quad \beta_0 + \beta_1 X$$

Variance at $X = X_0$: variance of sum equals

sum of variances plus twice covariances

$$\text{var}\,(\beta_0 + \beta_1 X_0) = \text{var}\,(\beta_0) + X_0\text{var}\,(\beta_1) + 2X_0\text{covar}\,(\beta_0, \beta_1)$$

$$\text{var}\,(\beta_0+\beta_1 X_0)=\text{var}\,(\beta_0)+X_0\text{var}\,(\beta_1)+2X_0\text{covar}\,(\beta_0,\beta_1)$$

Gives 95 % CI for mean for logit [$\pi(x_0)$], being

*($\beta_0+\beta_1 X$) +/- 1.96*SE*

Then use inverse transformation, where you plug in each endpoint into

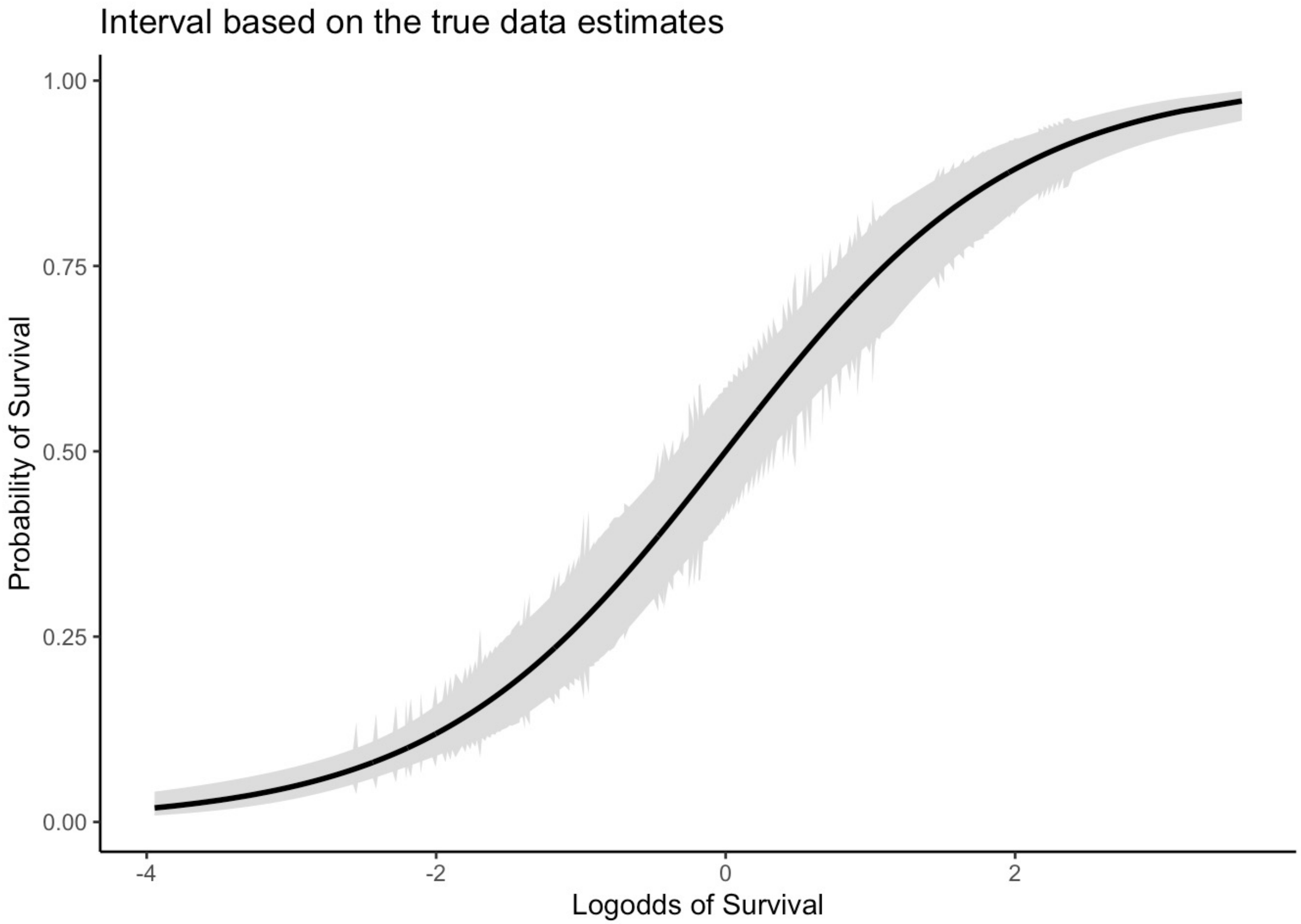
$\pi(x_0)$ = exp(logit)/[1 + exp(logit)]

and this gives interval for $\pi(x_0)$

Example plot

This is for prediction of mean,

for prediction of individual values, see Workshop



Utrecht University

Discussion

So we have prediction intervals for means.

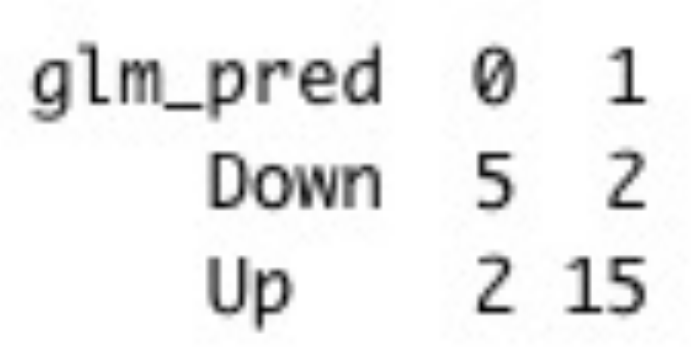
How about prediction intervals for individual values?
Is this a reasonable wish, given that observed values are either 0 or 1?

I would argue it is. For predictions of individual values it may be that Estimate of probability is .45, in CI of .35 - .60. Good to know that the cut off value of .50 is in interval.

But not everyone agrees. See Workshop for how to obtain these CI’s

Utrecht University

R Classification Table Summary revisited



```
glm_pred  0  1
    Down  5  2
    Up    2 15
```

0 = low
1 = normal

5/7 = sensitivity
15/17 = specificity

Imagine this would have been data where classification variable is unknown, should you take into account the confidence interval for each prediction?

What is the cost of a false positive, and what is cost of false negative?
Can you postpone judgement?

For instance, for individual estimated probability of fraud is .40, with 95 % CI between .30 and .55.

If individual is fraud-case, then deciding no fraud policy leads to false negative
If individual is non-fraud-case, then deciding fraud policy leads to false positive.

Decision theory:

Probability mass >.50, and cost of false negative, true negative
Probability mass < .50, and cost of true positive, false positive.

95% CI or 99% CI? Or 90% CI?

Larger sample sizes make CI smaller, and hypotheses are easily rejected, even when they do not depart much from the nul-hypothesis

Smaller sample sizes make CI larger, and hypotheses are not easily rejected, even when departure from nul hypothesis is large. For smaller sample sizes one often takes smaller CI, for example 90 % CI

Conclusions of 2. Classification by logistic regression

- Probabilities used for classification have a confidence interval that may or may not be taken into account

- Consider errors in terms of false positives and false negatives

- Is it possible to say something about costs?

3. Effect sizes: effect of individual variables

Generally: when equation is

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2$$

then **the effect size of *x₁*** is $$\frac{d(\hat{y})}{d(x_1)} = b_1$$

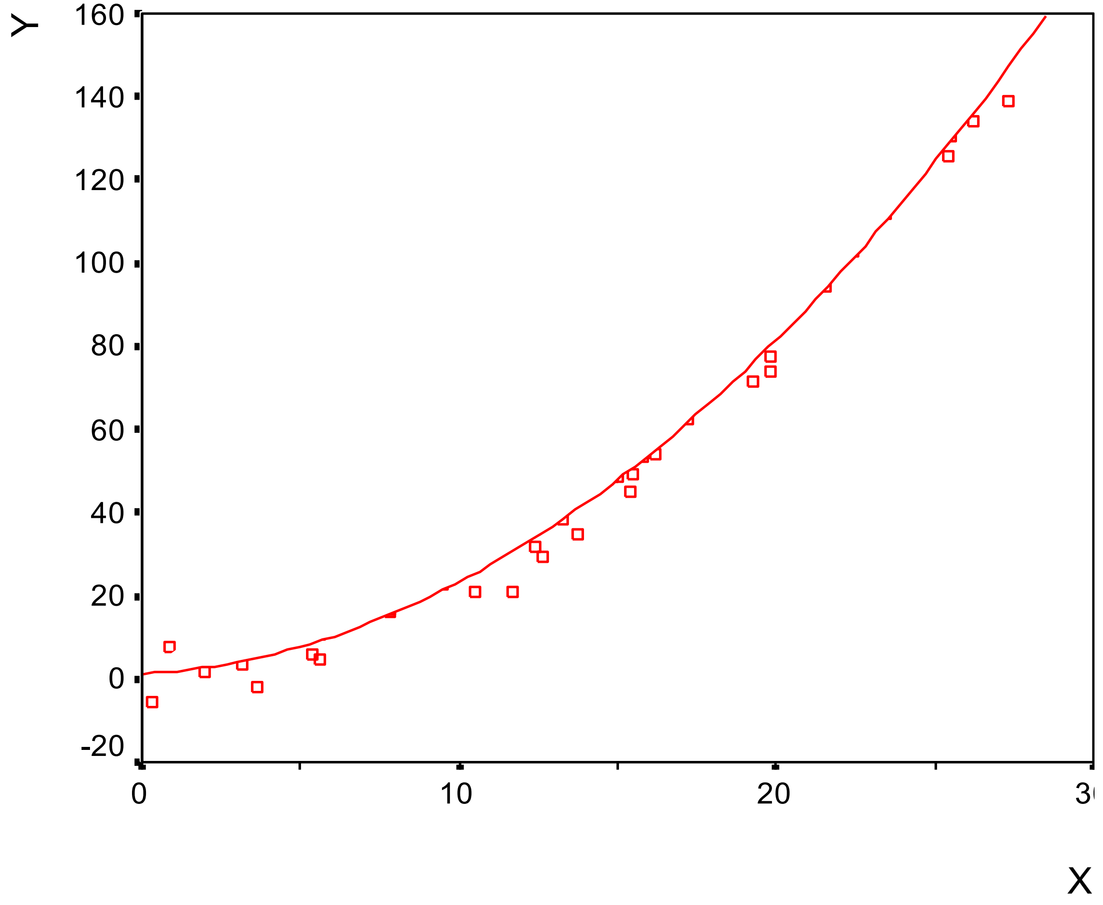
Quadratic regression

$$\hat{y} = b_0 + b_1 x + b_2 x^2$$

then **effect size of *x*** is $\dfrac{d(\hat{y})}{d(x)} = b_1 + 2b_2 x$

So effect of x depends on x itself!

The fitted quadratic regression curve is



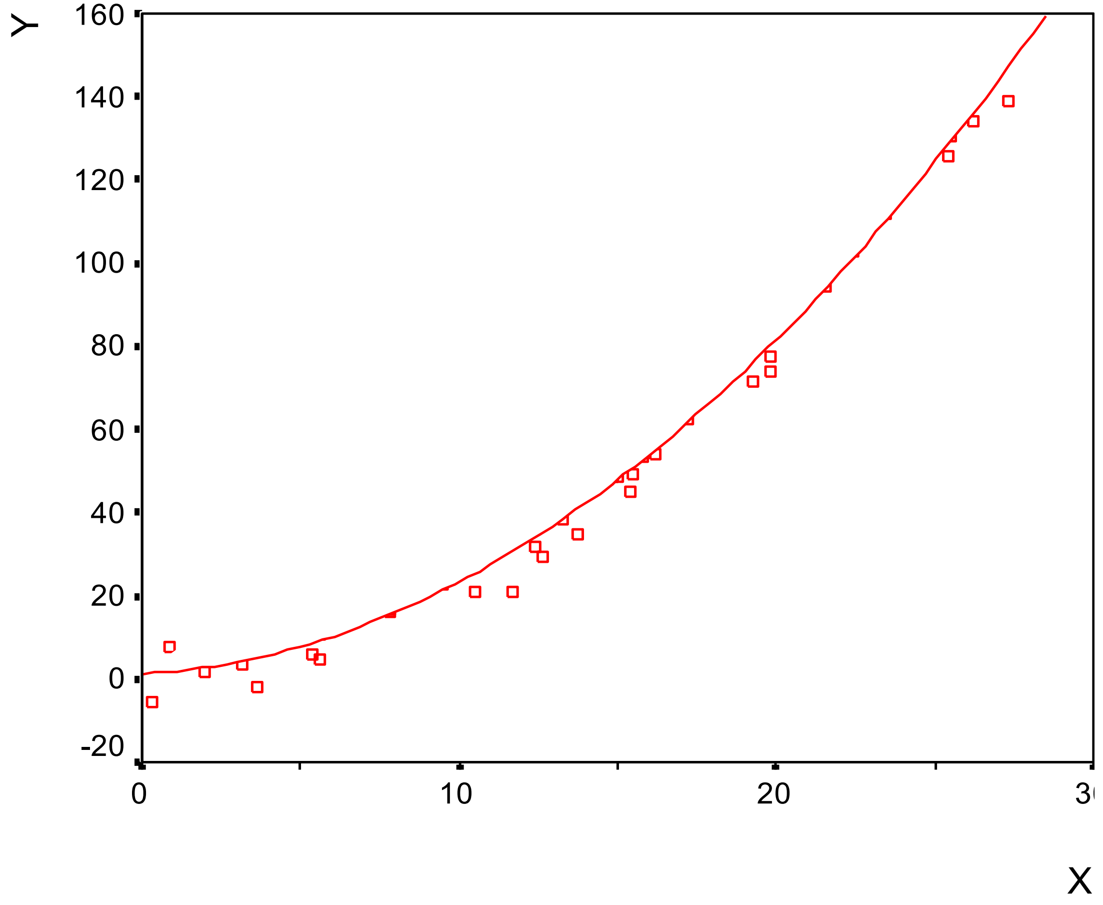
$$\hat{y} = 1.2788 + 0.3980x + 0.1808x^2$$

$$\text{effect of } x =$$

$$0.3980 + 0.3616x$$

Minimum lies outside range of x
(do you see this?)

The fitted quadratic regression curve is



$$\hat{y} = 1.2788 + 0.3980x + 0.1808x^2$$

$$\text{effect of } x =$$

$$0.3980 + 0.3616x$$

Minimum lies outside range of x

Set derivative equal to 0 and calculate x

Effect sizes for interaction

Interaction by multiplying variables.

$$E[y|x] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$$

Two ways in which this can be rewritten are

$$E[y|x] = (\beta_0 + \beta_2 x_2) + (\beta_1 + \beta_3 x_2) x_1$$

$$E[y|x] = (\beta_0 + \beta_1 x_1) + (\beta_2 + \beta_3 x_1) x_2$$

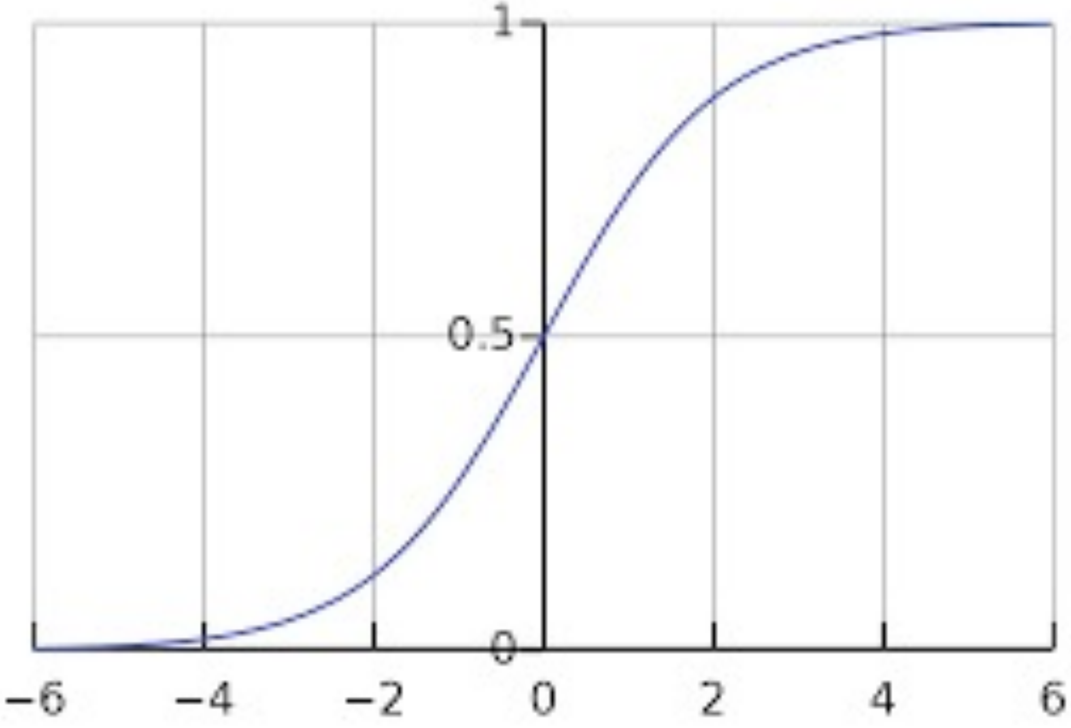
For $\mathbf{E[y|x]} = (\beta_0 + \beta_2 x_2) + (\beta_1 + \beta_3 x_2)\, x_1$

the effect size for x_1 is $\beta_1 + \beta_3 x_2$, i.e. a function of x_2.
Hence x_2 **moderates** the effect of x_1 on $\mathbf{E[y|x]}$.

For **example**: gender moderates effect of occupation mother on math score (or, which is mathematically identical: occupation mother score moderates effect of gender on math score)

Effect size in logistic regression: marginal effects

change in Pr(y=1|x,z) due to a change in *x* depends on *x* and on all other explanatory variables *z*, i.e. it depends on where an individual is on the logistic curve



Just like in linear regression we take derivatives, here derivative of p w.r.t. x

$$\frac{d\ \pi(x,z)}{d\ (x)} = \pi(1-\pi)b_x$$

We calculate for each observation (individual) separately the effect, and then average these effects.

Analysis of re-employment of data of unemployed men aged 40--55.
Finding work within one year after being unemployed as a function of age, health and educational level
Work = 1: work found within one year

```
LOGIST REGR work WITH age health edu
  /METHOD = ENTER age health edu
  /SAVE pred(pre).
Variables in the Equation
              B    S.E. Wald  df    Sig.    Exp(B)
age         -,128  ,031 17,1  1     ,000      ,879
health      -,319  ,158  4,9  1     ,043      ,727
edu          ,201  ,123  2,7  1     ,102     1,222
Constant    6,356 1,523 17,4  1     ,000   575,905
```

```
COMP d_health = pre*(1-pre)*(-.319).
COMP d_age    = pre*(1-pre)*(-.128).
COMP d_edu    = pre*(1-pre)*( .201).
MEANS d_age d_health d_edu.
```

```
Mean(d_age)     = -0.0280
Mean(d_health)  = -0.0697
```

So on average the effect of one year of age is 2.8% decrease in probability to find work, etc.

Useful addition to interpretational tools of logistic regression

Conclusions of 3. Effect sizes

- Where effect sizes are trivial in standard multiple regression, they provide helpful additions to the interpretation for quadratic regression, interactions and logistic regression

- In particular for logistic regression the addition is useful, as the standard interpretation of the parameter estimates in logistic regression is difficult, and the interpretation of marginal effects is easy

Thank you for your attention. Questions?

Utrecht University

Sharing science,
shaping tomorrow