Logistic Regression: Univariate and Multivariate

Events and Logistic Regression

- Logisitic regression is used for modelling event probabilities.
- Example of an event: Mrs. Smith had a myocardial infarction between 1/1/2000 and 31/12/2009.
- The occurrence of an event is a binary (dichotomous) variable. There are two possibilities: the event occurs or it does not occur.
- For this reason, event occurrence variables can always be coded with $0, 1$ e.g.

  $Y_i = 1 \iff$ person i became pregnant in 2011.
  $Y_i = 0 \iff$ person i did not become pregnant in 2011.

Measuring the Probability of an Event

- There are many equivalent ways of measuring the probability of an event.
- We will use three:
  1. probability of the event
  2. odds in favour of the event
  3. log-odds in favour of the event
- These are equivalent in the sense that if you know the value of one measure for an event you can compute the value of the other two measures for the same event

  cf. measuring a distance in kilometres, statute miles or nautical miles

The Probability of an Event

- This is a number π between 0 and 1. We write

  $$\pi = \mathbb{P}(Y = 1)$$

  to mean π is the probability that $Y = 1$.

- $\pi = 1$ means we know the event is certain to occur.

- $\pi = 0$ means we know the event is certain **not** to occur.

- Values between 0 and 1 represent intermediate states of certainty, ordered monotonically.

- Because we are certain one of $Y = 1$ and $Y = 0$ is true and because they cannot be true simultaneously:

  $$\mathbb{P}(Y = 0) = 1 - \mathbb{P}(Y = 1) = 1 - \pi.$$

Odds in Favour of an Event

- The odds in favour of an event is defined as the probability the event occurs divided by the probability the event does not occur.
- The odds in favour of $Y = 1$ is defined as:

$$\text{ODDS}(Y = 1) = \frac{\mathbb{P}(Y = 1)}{\mathbb{P}(Y \neq 1)} = \frac{\mathbb{P}(Y = 1)}{\mathbb{P}(Y = 0)} = \frac{\pi}{1 - \pi}.$$

- Note:

$$\text{ODDS}(Y = 0) = \frac{1}{\text{ODDS}(Y = 1)} = \frac{1 - \pi}{\pi}.$$

so

$$\text{ODDS}(Y = 1) \times \text{ODDS}(Y = 0) = 1.$$

Interpreting the Odds in Favour of an Event

- An odds is a number between 0 and ∞.
- An odds of 0 means we are certain the event does **not** occur.
- An increased odds corresponds to increased belief in the occurrence of the event.
- An odds of 1 corresponds to a probability of 1/2.
- An odds of ∞ corresponds to certainty the event occurs.

Log-odds in Favour of an Event

- The log odds in favour of an event is defined as the log of the odds in favour of the event:

$$\log \text{ODDS}(Y = 1) = \log \frac{\mathbb{P}(Y = 1)}{\mathbb{P}(Y = 0)} = \log \frac{\pi}{1 - \pi}.$$

- Note

$$\log \text{ODDS}(Y = 1) = -\log \text{ODDS}(Y = 0) = \log \frac{1 - \pi}{\pi}$$

Interpreting the Log-odds in Favour of an Event

- A log-odds is a number between $-\infty$ and ∞.
- A log odds of $-\infty$ means we are certain the event does **not** occur.
- An increased log-odds corresponds to increased belief in the occurrence of the event.
- A log-odds of 0 corresponds to a probability of $1/2$.
- A log-odds of ∞ corresponds to certainty the event occurs.

Moving between Probability, Odds and Log-odds

- You can use the following table to compute one measure of probability from another:

	$\mathbb{P}$	ODDS	$\log$ ODDS
$\mathbb{P}(Y=1)=\pi$		$\frac{\pi}{1-\pi}$	$\log\frac{\pi}{1-\pi}$
$\text{ODDS}(Y=1)=o$	$\frac{o}{1+o}$		$\log o$
$\log\text{ODDS}(Y=1)=x$	$\frac{e^x}{1+e^x}$	e^x	

- Choose the row corresponding to the quantity you start with and the column corresponding to the quantity you want to compute.
- $\log\frac{\pi}{1-\pi}$ is often written $\text{logit}(\pi)$.
- $\frac{\exp(x)}{1+\exp(x)}$ is often written $\text{inv.logit}(x)$ (sometimes $\text{expit}(x)$).

Motivation for (Multivariate) Logistic Regression

- We want to model $\mathbb{P}(Y = 1)$ in terms of a set of predictor variables X_1, X_2,... X_p (for univariate regression $p = 1$).
- In linear regression we use the regression equation

$$\mathbb{E}(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p \quad (1)$$

- However, for a binary Y (0 or 1), $\mathbb{E}(Y) = \mathbb{P}(Y = 1)$.
- We cannot now use equation (**??**), because the left hand side is a number between 0 and 1 while the right hand side is potentially a number between $-\infty$ and ∞.
- Solution: replace the LHS with $\operatorname{logit} \mathbb{E}Y$:

$$\operatorname{logit} \mathbb{E}(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p$$

Logistic Regression Equation Written on Three Scales

- We defined the regression equation on the logit or log ODDS scale:

$$\log \text{ODDS}(Y = 1) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p$$

- On the ODDS scale the same equation may be written:

$$\text{ODDS}(Y = 1) = \exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p)$$

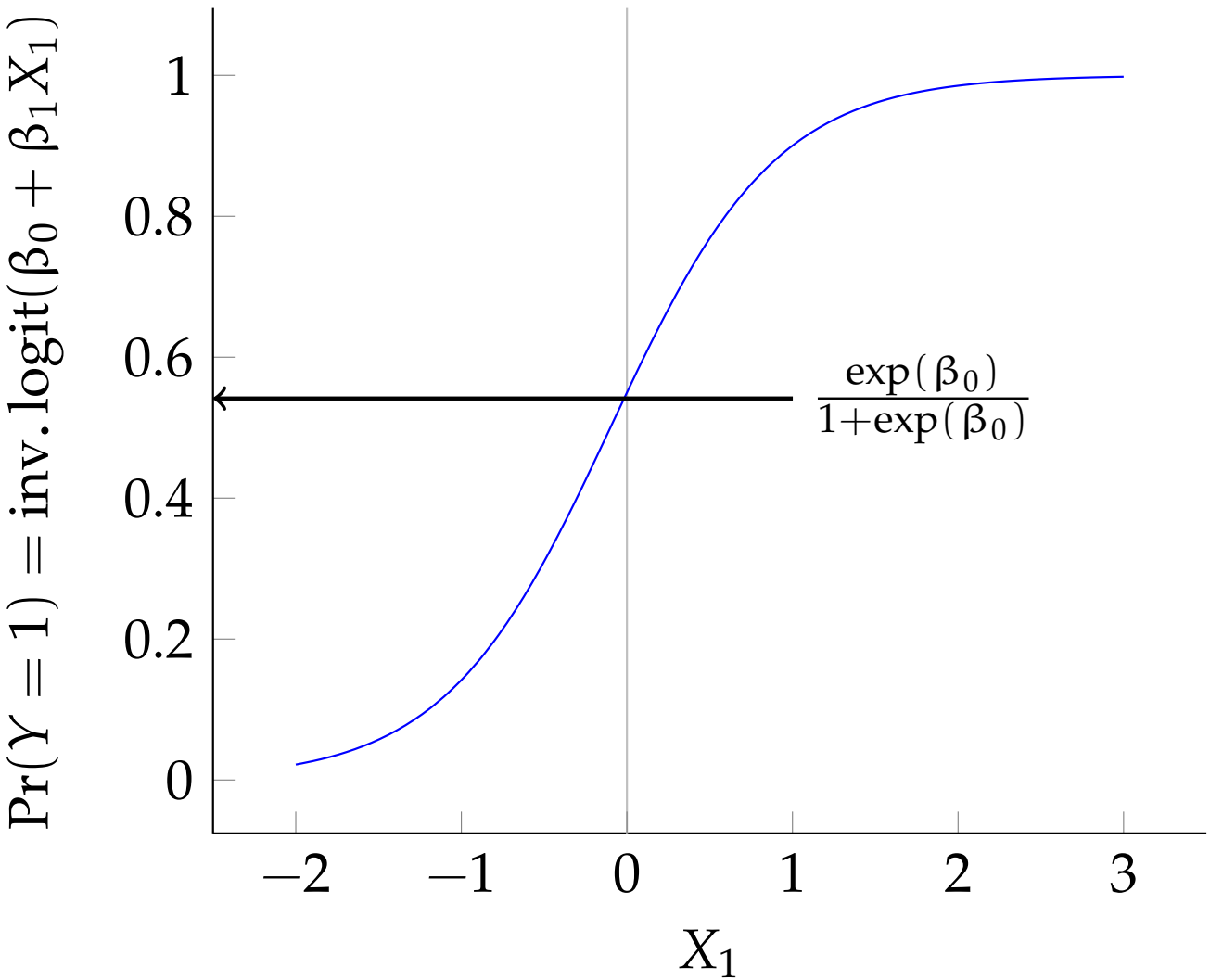
- On the probability scale the equation may be written:

$$\mathbb{P}(Y = 1) = \frac{\exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p)}{1 + \exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p)}$$

Interpreting the Intercept

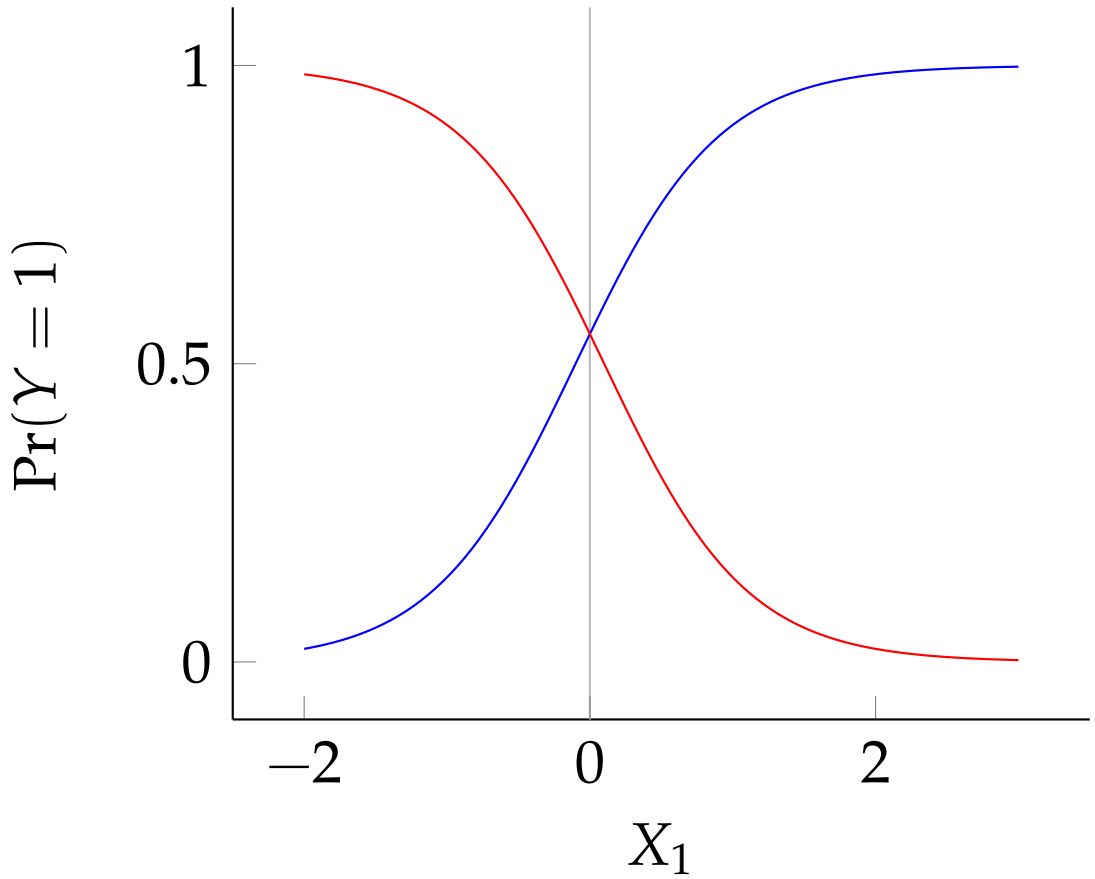
- In order to obtain a simple interpretation of the intercept we need to find a situation in which the other parameters $(\beta_1, ..., \beta_p)$ vanish.
- This happens when $X_1, X_2..., X_p$ are all equal to 0.
- Consequently we can interpret β_0 in 3 equivalent ways:
  1. β_0 is the log-odds in favour of $Y = 1$ when $X_1 = X_2... = X_p = 0$.
  2. β_0 is such that $\exp(\beta_0)$ is the odds in favour of $Y = 1$ when $X_1 = X_2... = X_p = 0$.
  3. β_0 is such that $\frac{\exp(\beta_0)}{1+\exp(\beta_0)}$ is the probability that $Y = 1$ when $X_1 = X_2... = X_p = 0$.
- You can choose any one of these three interpretations when you make a report.

Univariate Picture: Intercept



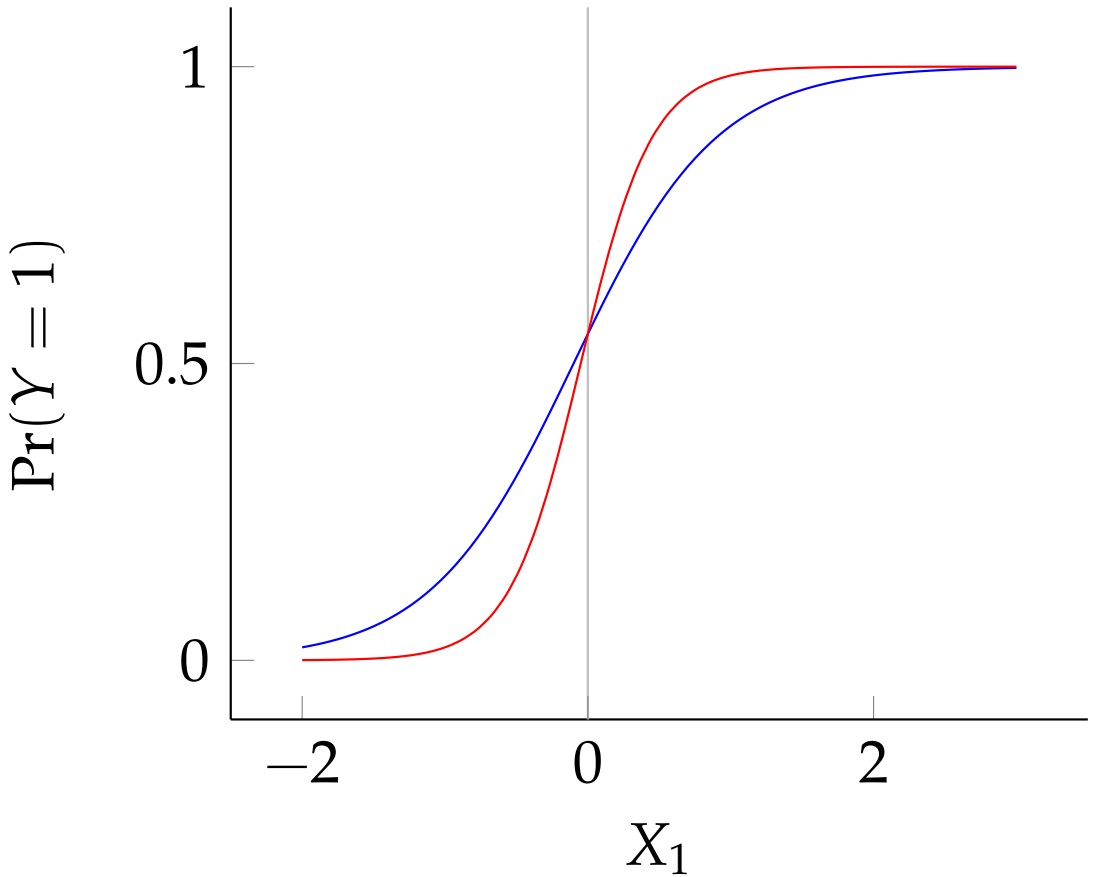
- $\mathbb{P}(Y = 1)$ *vs.* X_1 when $p = 1$ (univariate regression).

Univariate Picture: Sign of β_1



- When $\beta_1 > 0$, $\mathbb{P}(Y = 1)$ increases with X_1 (blue curve).
- When $\beta_1 < 0$, $\mathbb{P}(Y = 1)$ decreases with X_1 (red curve).

Univariate Picture: Magnitude of β_1



- $\beta_1 = 2$ (blue curve), $\beta_1 = 4$ (red curve).
- When $|\beta_1|$ is greater, changes in X_1 more strongly influence the probability that the event occurs.

Interpreting β_1: Univariate Logistic Regression

- To obtain a simple interpretation of β_1 we need to find a way to remove β_0 from the regression equation.
- On the log-odds scale we have the regression equation:

$$\log \text{ODDS}(Y = 1) = \beta_0 + \beta_1 X_1$$

- This suggests we could consider looking at the difference in the log odds at different values of X_1, say $t + z$ and t.

$$\log \text{ODDS}(Y = 1 | X_1 = t + z) - \log \text{ODDS}(Y = 1 | X_1 = t)$$

which is equal to

$$\beta_0 + \beta_1(t + z) - (\beta_0 + \beta_1 t) = z\beta_1.$$

Interpreting β_1: Univariate Logistic Regression

- By putting $z = 1$ we arrive at the following interpretation of β_1:

  β_1 *is the additive change in the log-odds in favour of* $Y = 1$ *when* X_1 *increases by* 1 *unit.*

- We can write an equivalent second interpretation on the odds scale:

  $\exp(\beta_1)$ *is the multiplicative change in the odds in favour of* $Y = 1$ *when* X_1 *increases by* 1 *unit.*

β_1 as a Log-odds Ratio

- The first interpretation of β_1 expresses the equation:

$$\log \frac{\text{ODDS}(Y = 1|X_1 = t + z)}{\text{ODDS}(Y = 1|X_1 = t)} = z\beta_1$$

  whilst the second interpretation expresses the equation:

$$\frac{\text{ODDS}(Y = 1|X_1 = t + z)}{\text{ODDS}(Y = 1|X_1 = t)} = \exp(z\beta_1).$$

- The quantity $\frac{\text{ODDS}(Y=1|X_1=t+z)}{\text{ODDS}(Y=1|X_1=t)}$ is the odds-ratio in favour of $Y = 1$ for $X_1 = t + z$ *vs.* $X_1 = t$.

Interpreting Coefficients in Multivariate Logistic Regression

- The interpretation of regression coefficients in multivariate logistic regression is similar to the interpretation in univariate regression.
- We dealt with β_0 previously.
- In general the coefficient β_k (corresponding to the variable X_k) can be interpreted as follows:

  β_k *is the additive change in the log-odds in favour of* $Y = 1$ *when* X_k *increases by* 1 *unit, while the other predictor variables remain unchanged.*
- As in the univariate case, an equivalent interpretation can be made on the odds scale.

Fitting a Logistic Regression in R

- We fit a logistic regression in R using the `glm` function:

```
> output <- glm(sta ~ sex, data=icu1.dat, family=binomial)
```

- This fits the regression equation

$$\text{logit}\,\mathbb{P}(\text{sta} = 1) = \beta_0 + \beta_1 \times \text{sex}.$$

- `data=icu1.dat` tells `glm` the data are stored in the data frame `icu1.dat`.
- `family=binomial` tells `glm` to fit a logistic model.
- As an aside, we can use `glm` as an alternative to `lm` to fit a linear model, by specifying `family=gaussian`.

Logistic Regression: `glm` Output in R

```
Call:
glm(formula = sta ~ sex, family = binomial, data = icu1.dat)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.6876  -0.6876  -0.6559  -0.6559   1.8123

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.4271     0.2273  -6.278 3.42e-10 ***
sex1          0.1054     0.3617   0.291    0.771
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- Summary of the distribution of the deviance residuals.
- Deviance residuals measure how well the observations fit the model. The closer a residual to 0 the better the fit of the observation.

Logistic Regression: `glm` Output in R

```
Call:
glm(formula = sta ~ sex, family = binomial, data = icu1.dat)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.6876  -0.6876  -0.6559  -0.6559   1.8123

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.4271     0.2273  -6.278 3.42e-10 ***
sex1          0.1054     0.3617   0.291    0.771
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- $\hat{\beta}_0$, the maximum likelihood estimate of the intercept coefficient β_0.
- $\frac{\exp(\hat{\beta}_0)}{1+\exp(\hat{\beta}_0)}$ is an estimate of $\mathbb{P}(\text{sta} = 1)$ when $\text{sex} = 0$

Logistic Regression: `glm` Output in R

```
Call:
glm(formula = sta ~ sex, family = binomial, data = icu1.dat)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.6876  -0.6876  -0.6559  -0.6559   1.8123

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.4271     0.2273  -6.278 3.42e-10 ***
sex1          0.1054     0.3617   0.291    0.771
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- $SE(\hat{\beta}_0)$, the standard error of the maximum likelihood estimate of β_0.

Logistic Regression: `glm` Output in R

```
Call:
glm(formula = sta ~ sex, family = binomial, data = icu1.dat)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.6876  -0.6876  -0.6559  -0.6559   1.8123

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.4271     0.2273  -6.278 3.42e-10 ***
sex1          0.1054     0.3617   0.291    0.771
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- z-value for a Wald-statistic, $z = \hat{\beta}_0 / SE(\hat{\beta}_0)$
- p-value for test of null hypothesis $\beta_0 = 0$ via the Wald-test.
- $p = 2\Phi(z)$, where Φ is the cdf of the normal distribution.

Logistic Regression: `glm` Output in R

```
Call:
glm(formula = sta ~ sex, family = binomial, data = icu1.dat)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.6876  -0.6876  -0.6559  -0.6559   1.8123

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.4271     0.2273  -6.278 3.42e-10 ***
sex1          0.1054     0.3617   0.291    0.771
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- Significance codes for p-values.
- List of p-value thresholds (the critical values) corresponding to significance codes.

Logistic Regression: `glm` Output in R

```
Call:
glm(formula = sta ~ sex, family = binomial, data = icu1.dat)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.6876  -0.6876  -0.6559  -0.6559   1.8123

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.4271     0.2273  -6.278 3.42e-10 ***
sex1          0.1054     0.3617   0.291    0.771
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- All entries are as for intercept row but apply to β_1 rather than to β_0.

Computing a 95% Confidence Interval from `glm`

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.4271     0.2273  -6.278 3.42e-10 ***
 sex1         0.1054     0.3617   0.291    0.771
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- We can compute a 95% confidence interval for a regression coefficient using a normal approximation:

$$\hat{\beta}_k - 1.96 \times SE(\hat{\beta}_k) < \beta_k < \hat{\beta}_k + 1.96 \times SE(\hat{\beta}_k)$$

- Plugging in the numbers for β_1:

$$0.105 - 1.96 \times 0.362 < \beta_1 < 0.105 + 1.96 \times 0.362$$
$$-0.603 < \beta_1 < 0.814$$

Computing a 95% Confidence Interval on Odds Scale

- We can compute a 95% confidence interval for the odds-ratio parameter $\exp(\beta_1)$ by transforming the limits to the new scale (see table above).
- Start with the log-odds scale interval:

$$-0.603 < \beta_1 < 0.814$$

- Transform the limits:

$$\exp(-0.603) < \exp(\beta_1) < \exp(0.814)$$
$$0.547 < \exp(\beta_1) < 2.257$$

Logistic Regression with Dummy Variables

- A dummy variable is a $0/1$ representation of a dichotomous catagorical variable.
- Such a numeric representation allows us to use categorical variables as predictors in a regression model.
- For example the dichotomous variable sex can be coded

$$\text{sex}_i = 0 \text{ means individual } i \text{ is male}$$
$$\text{sex}_i = 1 \text{ means individual } i \text{ is female}$$

Logistic Regression with Dummy Variables

- Suppose we fit the regression specified by the equation

$$\operatorname{logit} \mathbb{P}(Y_i = 1) = \beta_0 + \beta_1 \mathsf{sex}_i.$$

- Recall one interpretation of β_1:

  $\exp(\beta_1)$ *is the multiplicative change in the odds in favour of* $Y = 1$ *as* sex *increases by* 1 *unit.*

- The only unit increase possible is from 0 to 1, so we can write an interpretation in terms of male/female:

  $\exp(\beta_1)$ *is multiplicative change of the odds in favour of* $Y = 1$ *as a male becomes a female.*

- A bit ridiculous, so better to say:

  $\exp(\beta_1)$ *is the odds-ratio (in favour of* $Y = 1$*) for females vs. males.*

Multivariate Logistic Regression Example

- Data on admisssions to an intensive care unit (ICU).
- sta - outcome variable, status on leaving: dead=1, alive=0.
- loc - level of consciousness: no coma/stupor=0, deep stupor=1, coma=2.
- sex - male=0, female=1.
- ser - service at ICU: medical=0, surgical=1.
- ser and sex are dummy variables
- loc is a categorical/factor variable with 3 levels.

Multivariate Logistic Regression ICU Example

- Summarise the data:

```
> summary(icu1.dat)
      sta       loc      sex      ser
 Min.   :0.0   0:185   0:124   0: 93
 1st Qu.:0.0   1:  5   1: 76   1:107
 Median :0.0   2: 10
 Mean   :0.2
 3rd Qu.:0.0
 Max.   :1.0
```

- 20% leave ICU dead.
- Categories 1 and 2 of loc are rare, not many people arrive in a stupor/deep coma. This variable may not be very informative.
- sex and ser are reasonably well balanced.

Multivariate Logistic Regression ICU Example

- Take an initial look at the 2-way tables cross classifying the outcome with each predictor variable in turn.
- vital status (rows) *vs.* sex (columns):

```
> table(icu1.dat$sta, icu1.dat$sex)
     0   1
  0 100  60
  1  24  16
```

- Observed death rate in males: $24/124 = 0.19$
- Observed death rate in females: $16/76 = 0.21$
- Without doing a formal test, looks significantly different.

Multivariate Logistic Regression ICU Example

- vital status (rows) *vs.* service type at ICU (columns):

```
> table(icu1.dat$sta, icu1.dat$ser)

    0  1
0  67 93
1  26 14
```

- Observed death rate at medical unit (ser=0): $26/93 = 0.28$
- Observed death rate at surgical unit (ser=1): $14/107 = 0.13$

Multivariate Logistic Regression ICU Example

- vital status (rows) *vs.* level of consciousness (columns):

```
> table(icu1.dat$sta, icu1.dat$loc)

     0   1   2
0  158   0   2
1   27   5   8
```

- Few observations but higher death rate amongst those in a stupor or coma.

Multivariate Logistic Regression ICU Example

- Take an initial look at the 2-way tables cross classifying each pair of predictors.
- sex (rows) *vs.* service type (columns):

```
> table(icu1.dat$sex, icu1.dat$ser)

    0  1
0  54 70
1  39 37
```

- Rate of admission to SU in males: $70/124 = 0.56$
- Rate of admission to SU in females: $37/76 = 0.48$
- Some correlation to be aware of but confounding of ser by sex seems unlikely given weak effect of sex.

Multivariate Logistic Regression ICU Example

- sex (rows) *vs.* level of consciousness (columns):

```
> table(icu1.dat$sex, icu1.dat$loc)

     0   1   2
0 116   3   5
1  69   2   5
```

- Hard to say much, maybe females have higher levels of loc.

Multivariate Logistic Regression ICU Example

- Service type (rows) *vs.* level of consciousness (columns):

```
> table(icu1.dat$ser, icu1.dat$loc)

     0   1   2
0   84   2   7
1  101   3   3
```

- Hard to say much.
- loc may not to be a useful variable due to low variability.

Multivariate Logistic Regression ICU Example

- Now look at univariate regressions.

```
glm(formula = sta ~ sex, family = binomial, data = icu1.dat)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.4271     0.2273  -6.278 3.42e-10 ***
sex1          0.1054     0.3617   0.291    0.771
---
$intercept.ci
[1] -1.8726220 -0.9816107

$slopes.ci
[1] -0.6035757  0.8142967

$OR
    sex1
1.111111

$OR.ci
[1] 0.5468528 2.2575874
```

- Wide confidence interval for sex including $OR = 1$.

Multivariate Logistic Regression ICU Example

```
glm(formula = sta ~ ser, family = binomial, data = icu1.dat)
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.9466     0.2311  -4.097 4.19e-05 ***
ser1         -0.9469     0.3682  -2.572   0.0101 *
---
$intercept.ci
[1] -1.3994574 -0.4937348

$slopes.ci
[1] -1.6685958 -0.2252964

$OR
     ser1
0.3879239

$OR.ci
[1] 0.1885116 0.7982796
```

- $OR < 1$ so being in surgical unit may lower risk of death.
- CI implies at least 20% effect.

Multivariate Logistic Regression ICU Example

```
Call:
glm(formula = sta ~ loc, family = binomial, data = icu1.dat)
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   -1.7668     0.2082  -8.484  < 2e-16 ***
loc1          18.3328  1073.1090   0.017 0.986370
loc2           3.1531     0.8175   3.857 0.000115 ***
---
$intercept.ci
[1] -2.174912 -1.358605

$slopes.ci
              [,1]        [,2]
[1,] -2084.922247 2121.587900
[2,]     1.550710    4.755395
```

- Huge *SE*, should be wary of using this variable.

Multivariate Logistic Regression ICU Example

Summary of univariate analyses:

- Vital status not significantly associated with sex.
- Vital status associated with service type at 5% level.
- Admission to surgical unit associated with reduced death rate.
- loc variable not very useful, will now drop.

Multivariate Logistic Regression ICU Example

- Multivariate analysis:

```
Call:
glm(formula = sta ~ sex+ser, family = binomial, data = icu1.dat)
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.96129    0.27885  -3.447 0.000566 ***
sex1         0.03488    0.36896   0.095 0.924688
ser1        -0.94442    0.36915  -2.558 0.010516 *
---
$intercept.ci
[1] -1.5078281 -0.4147469

$slopes.ci
           [,1]      [,2]
[1,] -0.6882692  0.758025
[2,] -1.6679299 -0.220904

$OR
     sex1      ser1
1.0354933 0.3889063
```

Multivariate Logistic Regression ICU Example

Main Conclusions:

- Univariate and multivariate parameter models show same pattern of significance.
- Direction of association of service variable the same.
- Admission to surgical unit associated with reduced death rate ($OR = 0.39,\ 95\%\ CI = (0.19, 0.80)$.

Prediction In Logistic Regression

- Suppose we fit a logistic regression model and obtain coefficient estimates $\hat{\beta}_0, \hat{\beta}_1, ...\hat{\beta}_p$.
- Suppose we observe a set of predictor variables $X_{i1}, X_{i2}, ...X_{ip}$ for a new individual i.
- If Y_i is unobserved, we can estimate the log-odds in favour of $Y_i = 1$ using the following formula:

$$\text{logit}\,\frac{\hat{\pi}_i}{1-\hat{\pi}_i} = \hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \hat{\beta}_2 X_{i2} + ... + \hat{\beta}_p X_{ip}$$

- Equivilently an estimate of the probability that $Y_i = 1$:

$$\hat{\pi}_i = \frac{\exp(\hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \hat{\beta}_2 X_{i2} + ... + \hat{\beta}_p X_{ip})}{1 + \exp(\hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \hat{\beta}_2 X_{i2} + ... + \hat{\beta}_p X_{ip})}$$

- $\hat{\pi}_i$ can be thought of as a prediction of Y_i.

Prediction In Logistic Regression Using R

- We can use the `predict` function to calculate $\hat{\pi}_i$

```
> output <- glm(sta ~ sex, data=icu1.dat, family=binomial)
> newdata <- data.frame(sex=as.factor(c(0,0,1,1)),
                        ser=as.factor(c(0,1,0,1)))

> newdata
  sex ser
1   0   0
2   0   1
3   1   0
4   1   1
```

- Predict on the log-odds scale (i.e. $\log \frac{\hat{\pi}_i}{1-\hat{\pi}_i}$) :

```
> predict(output, newdata=newdata)
         1          2          3          4
-0.9612875 -1.9057045 -0.9264096 -1.8708266
```

- Predict on the probability scale (i.e. $\hat{\pi}_i$) :

```
> predict(output, newdata=newdata, type="response")
        1         2         3         4
0.2766205 0.1294642 0.2836537 0.1334461
```

Multivariate Logistic Regression Example

- Return to ICU example and consider additional variables age and typ.
- sta - outcome variable, status on leaving: dead=1, alive=0.
- sex - male=0, female=1.
- ser - service at ICU: medical=0, surgical=1.
- age - in years
- typ - type of admission: elective=0, emergency=1.

Multivariate Logistic Regression ICU Example

- Look at the joint distribution of the new predictors and the outcome:
- vital status (rows) *vs.* admission type (columns):

```
> table(icu2.dat$sta, icu2.dat$typ)

    0   1
0  51 109
1   2  38
```

- Observed death rate for elective admissions: $2/53 = 0.04$
- Observed death rate for emergencies: $38/147 = 0.25$
- Much higher risk of death for admission as an emergency.

Multivariate Logistic Regression ICU Example

- Look at the joint distribution of ser and typ:
- service at ICU (rows) *vs.* admission type (columns):

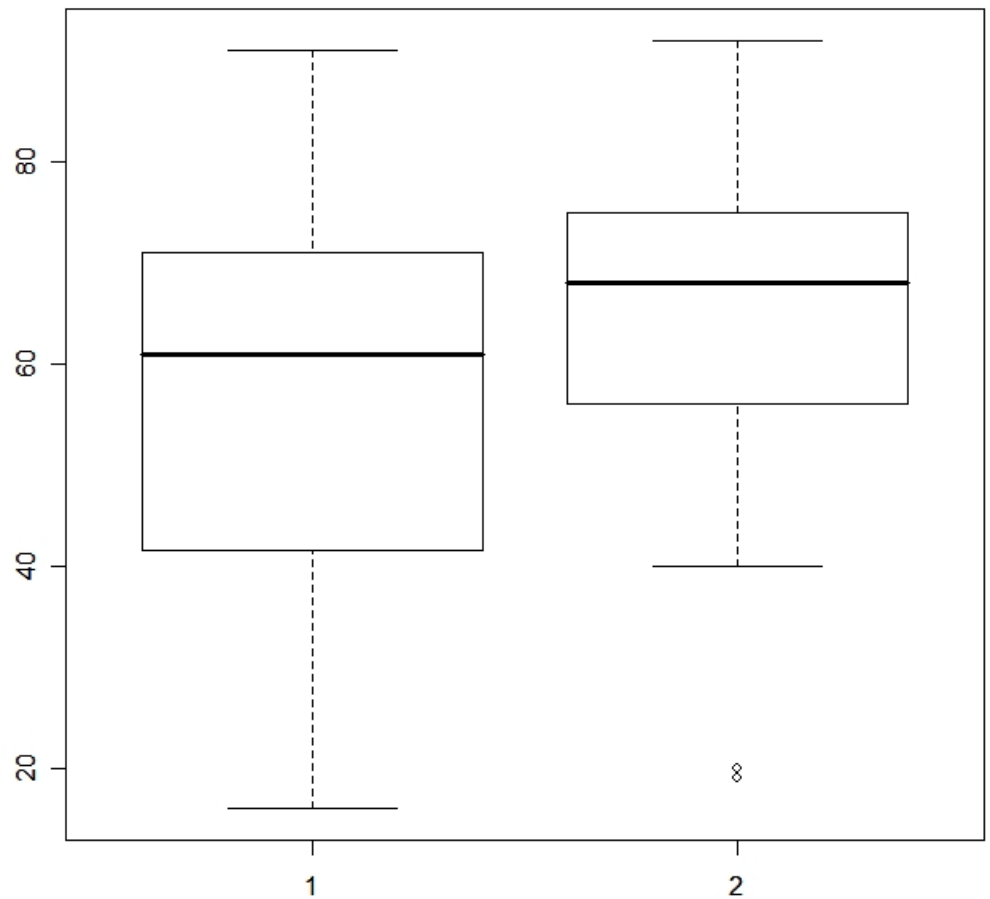
```
> table(icu2.dat$ser, icu2.dat$typ)

    0  1
0   1 92
1  52 55
```

- ser and typ are highly correlated.
- We know both variables are associated with outcome
- One might be a confounder for the other

Multivariate Logistic Regression ICU Example

- Box showing distribution of age stratified by vital status

```
>   boxplot(list(icu2.dat$age[icu2.dat$sta==0],
          icu2.dat$age[icu2.dat$sta==1]))
```



Multivariate Logistic Regression ICU Example

- Multivariate analysis:

```
Call:
glm(formula = sta ~ sex + ser + age + typ, family = binomial,
    data = icu2.dat)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.2753  -0.7844  -0.3920  -0.2281   2.5072

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -5.26359    1.11678  -4.713 2.44e-06 ***
sex1        -0.20092    0.39228  -0.512  0.60851
ser1        -0.23891    0.41697  -0.573  0.56667
age          0.03473    0.01098   3.162  0.00156 **
typ1         2.33065    0.80238   2.905  0.00368 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
65050
```

- There is now no significant difference between medical and surgical service types: (ser) has lost its significance.

Multivariate Logistic Regression ICU Example

- Multivariate analysis on odds scale:

```
$OR
      sex1       ser1        age       typ1
 0.8179766  0.7874880  1.0353364 10.2846123

$OR.ci
           [,1]      [,2]
[1,] 0.3791710  1.764602
[2,] 0.3477894  1.783083
[3,] 1.0132920  1.057860
[4,] 2.1340289 49.565050
```

- age has a strong effect odds ratio of 1.035 for a 1 year change in age.
- Corresponds to an odds ratio of $1.035^{10} = 1.41$ for a 10 year change in age.

Multivariate Logistic Regression ICU Example

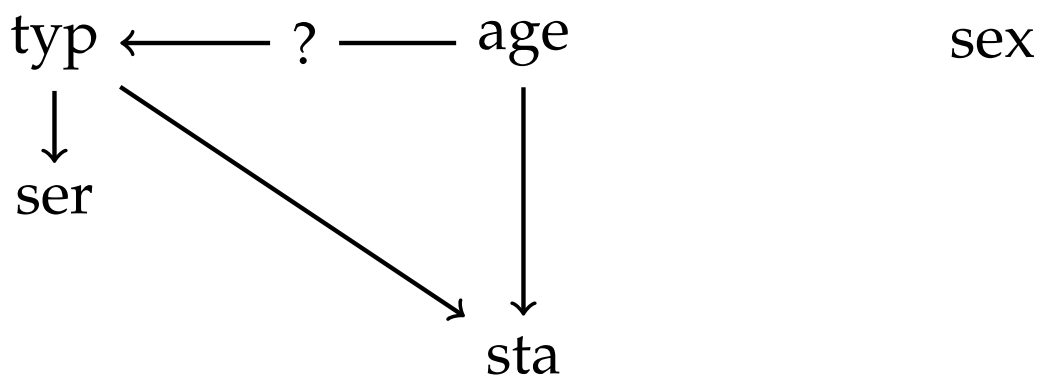
- Multivariate analysis on odds scale:

```
$OR
      sex1       ser1        age       typ1
 0.8179766  0.7874880  1.0353364 10.2846123

$OR.ci
           [,1]      [,2]
[1,] 0.3791710  1.764602
[2,] 0.3477894  1.783083
[3,] 1.0132920  1.057860
[4,] 2.1340289 49.565050
```

- age has a strong effect: odds ratio of 1.035 for a 1 year change in age.
- Corresponds to an odds ratio of $1.035^{10} = 1.41$ for a 10 year change in age.

Multivariate Logistic Regression ICU Example

- Draw a causal diagram (DAG)



- Arrow illustrates the direction of causality
- Causality (and so arrows) must obey temporal ordering
- Admission type (emergency/elective) determined before service type (medical/surgical)
- Further evidence that typ is the confounder: ser is not significant in the multivariate model