Binary data – logistic regression

Fundamental Rule of Data Analysis

Different types of data require different statistical analyses.

Type of Outcome/Exposure Variables

OUTCOME VARIABLE

EXPOSURE VARIABLE

	Continuous	Binary
1 group	One-group t-test	Exact binomial test [or] normal approximation test
2 groups	Two-group t-test	χ^2 test [or] Fisher’s exact test
>2 groups	ANOVA	χ^2 test [or] Fisher’s exact test
Continuous	Linear regression	Logistic Regression

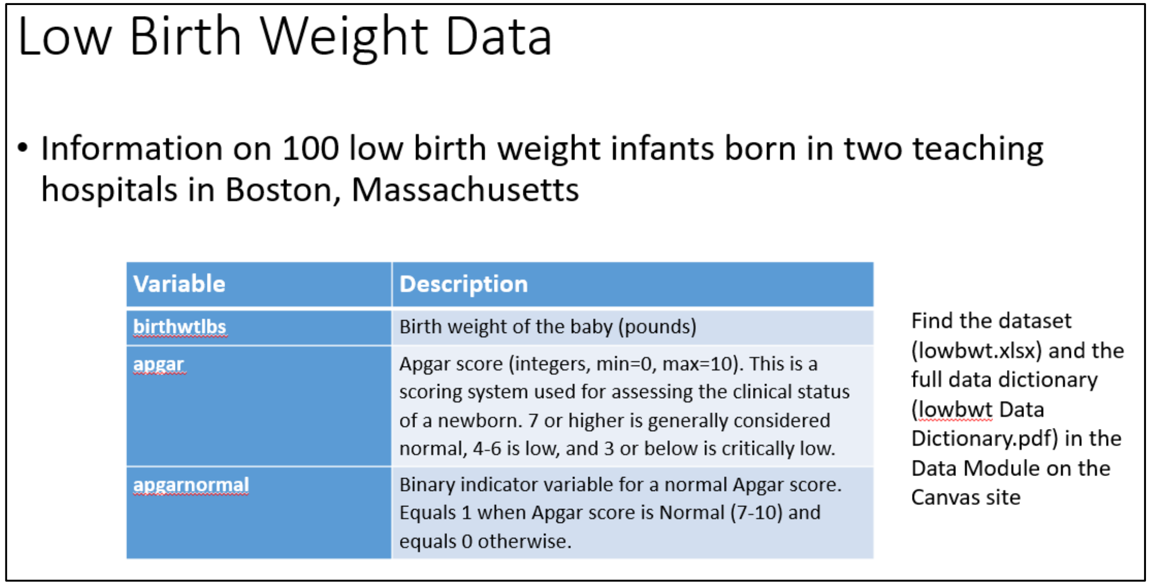
Regression with a Binary Outcome

- Why can’t we use linear regression with a binary outcome variable?

Motivating Example

- Let’s look at the association between birth weight and having a normal Apgar score

- Outcome variable: *apgarnormal*
  - 1 if Apgar score is normal
  - 0 if Apgar score is low
- Exposure/predictor variable: *birthwtlbs*
  - Birth weight (measured in pounds)



Low Birth Weight Data

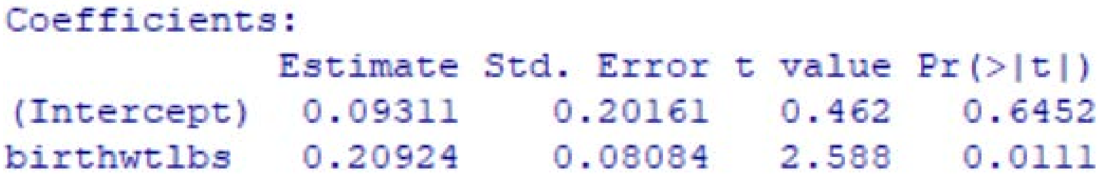
- Information on 100 low birth weight infants born in two teaching hospitals in Boston, Massachusetts

Variable	Description
birthwtlbs	Birth weight of the baby (pounds)
apgar	Apgar score (integers, min=0, max=10). This is a scoring system used for assessing the clinical status of a newborn. 7 or higher is generally considered normal, 4-6 is low, and 3 or below is critically low.
apgarnormal	Binary indicator variable for a normal Apgar score. Equals 1 when Apgar score is Normal (7-10) and equals 0 otherwise.

Find the dataset (lowbwt.xlsx) and the full data dictionary (lowbwt Data Dictionary.pdf) in the Data Module on the Canvas site

Motivating Example

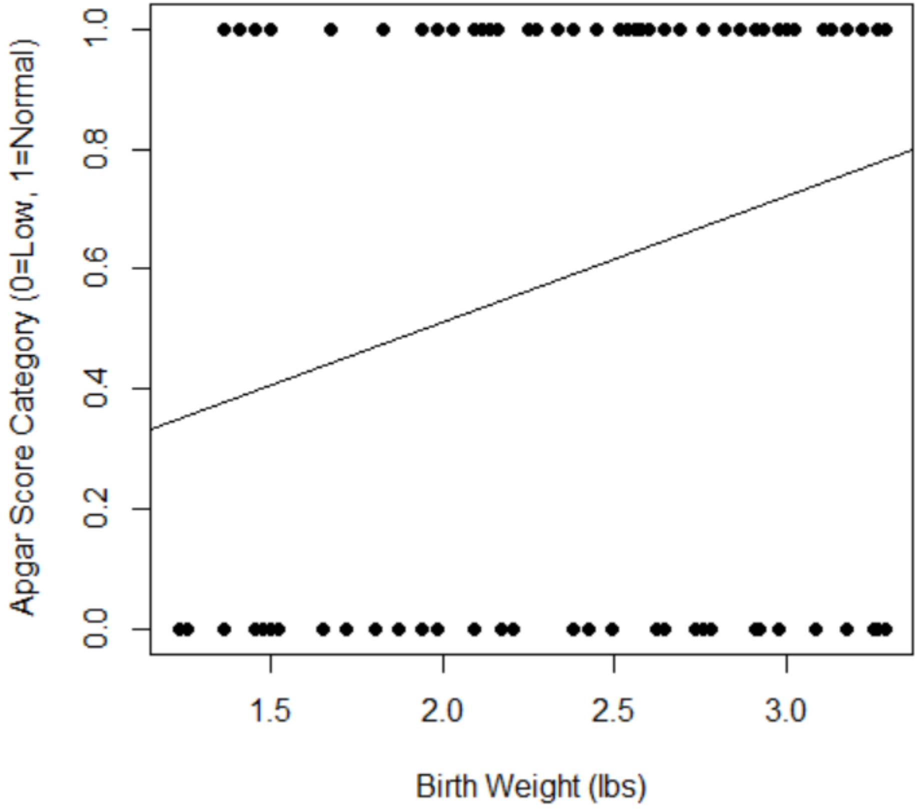
Linear regression model (outcome is *apgarnormal*):



```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.09311    0.20161   0.462   0.6452
birthwtlbs   0.20924    0.08084   2.588   0.0111
```

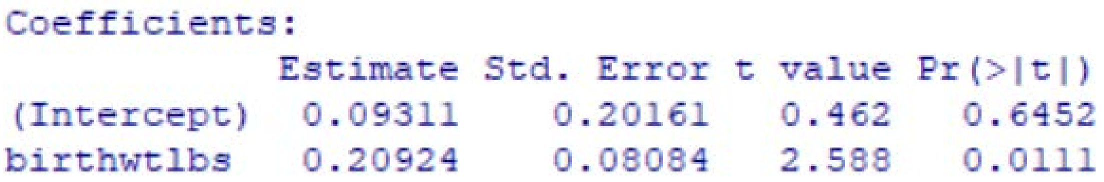
- On average, a 1 pound increase in birth weight is associated with an increase of 0.21 in the indicator for a normal Apgar score.

…but the indicator can only be 0 or 1!



Motivating Example

Linear regression model (outcome is *apgarnormal*):

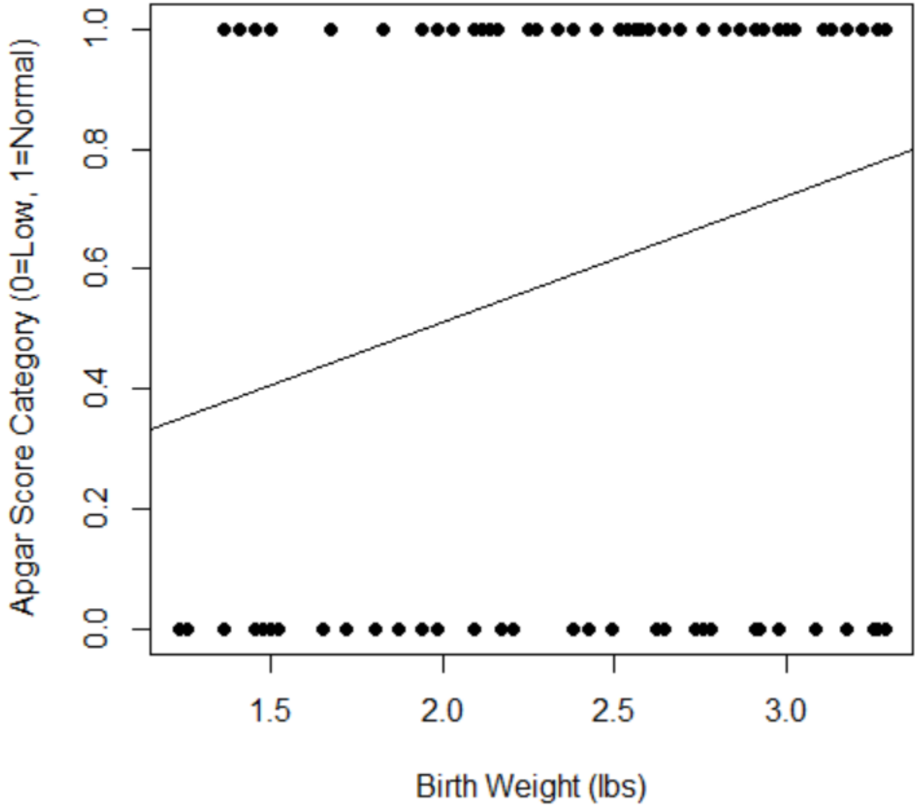


```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.09311    0.20161   0.462   0.6452
birthwtlbs    0.20924    0.08084   2.588   0.0111
```

- Predicted indicator of a normal Apgar score for an infant weighing 2.5 lbs:

$$\begin{aligned}\hat{y} &= 0.093 + 0.209(2.5)\\ &= 0.62\end{aligned}$$

…but the indicator can only be 0 or 1!



Regression with a Binary Outcome

- Why can’t we use linear regression with a binary outcome variable?
    - Coefficient interpretations don’t make sense
    - Predicted values of the outcome are impossible
    - Violates linear regression assumptions

Regression with a Binary Outcome

- Instead of modeling the 0/1 outcome itself, what if we use linear regression to model the *probability* of having the outcome of interest?

Let y be a 0/1 binary variable. $\hat{y} = \beta_0 + \beta_1 x \longleftarrow$ Already showed why this is problematic

Let p be the probability of $y = 1$. $p = \beta_0 + \beta_1 x$

Regression with a Binary Outcome

- Instead of modeling the 0/1 outcome itself, what if we use linear regression to model the *probability* of having the outcome of interest?
  - Advantage: Probability is continuous instead of binary
  - Problem: Probabilities can’t be less than 0 or greater than 1… and a linear regression model could give us predictions outside of that range

Regression with a Binary Outcome

- Instead of modeling the 0/1 outcome itself, what if we use linear regression to model the *odds* of having the outcome of interest?

Let y be a 0/1 binary variable. $\hat{y} = \beta_0 + \beta_1 x$ ⟵ Already showed why this is problematic

Let p be the probability of $y = 1$. $p = \beta_0 + \beta_1 x$ ⟵ Already showed why this is problematic

$$\frac{p}{1-p} = \beta_0 + \beta_1 x$$

Regression with a Binary Outcome

- Instead of modeling the 0/1 outcome itself, what if we use linear regression to model the *odds* of having the outcome of interest?
    - Advantage: Odds are continuous and range from 0 to ∞
    - Problem: Linear regression model could give us predictions that are negative

Regression with a Binary Outcome

- Instead of modeling the 0/1 outcome itself, what if we use linear regression to model the *log of the odds* of having the outcome of interest?

Let y be a 0/1 binary variable. $\hat{y} = \beta_0 + \beta_1 x$ ← Already showed why this is problematic 0 or 1

Let p be the probability of $y = 1$. $p = \beta_0 + \beta_1 x$ ← Already showed why this is problematic 0 to 1

$$\frac{p}{1-p} = \beta_0 + \beta_1 x$$ ← Already showed why this is problematic 0 to infinity

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x$$ -infinity to infinity

Regression with a Binary Outcome

- Instead of modeling the 0/1 outcome itself, what if we use linear regression to model the *log of the odds* of having the outcome of interest?
  - Advantage: Odds are continuous, and taking the log transformation means the log odds can range from -∞ to ∞. The regression model won't give us impossible values.
  - Problem: None! This is called **logistic regression**.

Logistic Regression

- Logistic regression relates the predictor variable(s) to the **log odds of the outcome** (log odds of being in the category of interest)

If y is the binary outcome variable, then let p be the probability that y equals the category of interest. The logistic regression model is:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x$$

Notation & Algebra Reminder

- In statistics, we use the natural log (base e)
- $\exp(x)$ means e^x
- If $\log(y) = x$, then $y = \exp(x)$
- If $\log\left(\frac{y}{1-y}\right) = x$, then $y = \frac{\exp(x)}{1+\exp(x)}$

FYI:

- $\log\left(\frac{y}{1-y}\right)$ is also called the “logit” of y and may be written as $\text{logit}(y)$
- $\frac{\exp(x)}{1+\exp(x)}$ is also called the “expit” of x and may be written as $\text{expit}(x)$

Logistic Regression: Predicted Values

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x$$

p is defined as P(y=1)
where y is a 0/1 variable

- Plugging in the x value for a new subject, we can calculate their:
  - Log odds of the outcome = $\beta_0 + \beta_1 x$
  - Odds of the outcome = $\exp(\beta_0 + \beta_1 x)$
  - Probability of the outcome = $\frac{\exp(\beta_0+\beta_1 x)}{1+\exp(\beta_0+\beta_1 x)}$

Example: Birth Weight/Apgar Score

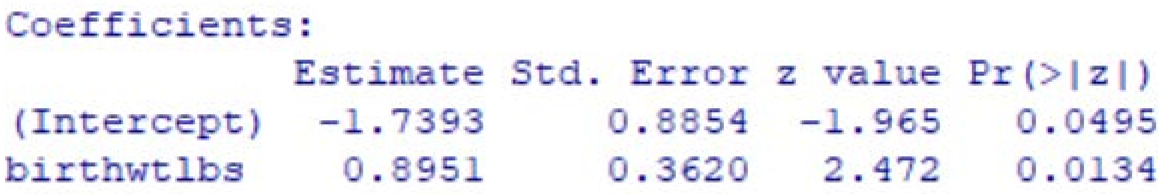
- We're interested in the association between birth weight and having a normal Apgar score.
  - Outcome variable: *apgarnormal* = y
    - 1 if Apgar score is normal
    - 0 if Apgar score is low
  - Exposure/predictor variable: *birthwtlbs* = y
    - Birth weight (measured in pounds)

Let p be the probability of having a normal apgar score (y=1)

- What is the predicted probability of having a normal Apgar score for an infant weighing 2.5 pounds?

Example: Birth Weight/Apgar Score

Logistic regression model (outcome is log odds of *apgarnormal*):



```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.7393     0.8854  -1.965   0.0495
birthwtlbs    0.8951     0.3620   2.472   0.0134
```

Predicted log odds of normal Apgar score for infant weighing 2.5 lbs:

$$\log(p/1-p) = -1.74 + 0.8951(x)$$
$$= -1.74 + 0.8951(2.5)$$
$$= 0.498$$

Predicted odds of normal Apgar score for infant weighing 2.5 lbs:

$$p/1-p = \exp(0.498)$$
$$=1.646$$

Predicted probability of normal Apgar score for infant weighing 2.5 lbs:

$$p = \exp(-1.74 + 0.8951(x)) / 1+\exp(-1.74+0.8951(x))$$
$$= \exp(-1.74+0.8951(2.5))/1+\exp(-1.74+0.8951(2.5))$$
$$= 0.622$$

Logistic Regression: Coefficient Interpretation

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x$$

p is defined as P(y=1)
where y is a 0/1 variable

- β_0: The log odds of the outcome is β_0 when the predictor variable is 0 units.
- β_1: A one unit increase in the predictor variable is associated with a β_1 increase in the log odds of the outcome.

Logistic Regression: Coefficient Interpretation

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x$$

p is defined as P(y=1)
where y is a 0/1 variable

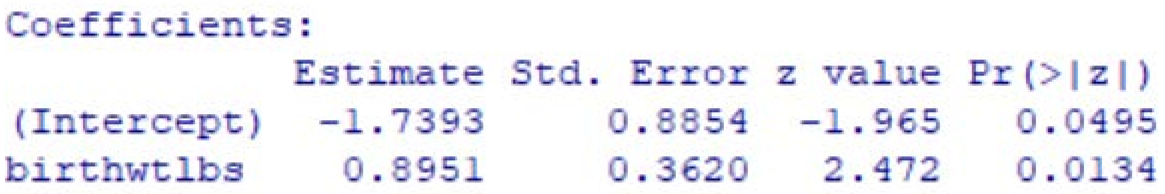
- $\exp(\beta_0)$: The odds of the outcome is $\exp(\beta_0)$ when the predictor variable is 0 units.
- $\exp(\beta_1)$: A one unit increase in the predictor variable is associated with a $\exp(\beta_1)$ times increase in the odds of the outcome.
  [or]
  $\exp(\beta_1)$ is the odds ratio for the outcome that is associated with a one unit increase in the predictor variable.

Example: Birth Weight/Apgar Score

- We’re interested in the association between birth weight and having a normal Apgar score.
    - Outcome variable: *apgarnormal*
        - 1 if Apgar score is normal
        - 0 if Apgar score is low
    - Exposure/predictor variable: *birthwtlbs*
        - Birth weight (measured in pounds)
- Interpret the estimated model coefficients (intercept and slope) in context of the problem.

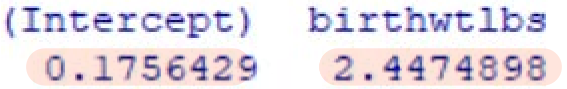
Example: Birth Weight/Apgar Score

Logistic regression model (outcome is log odds of *apgarnormal*):



```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.7393     0.8854  -1.965   0.0495
birthwtlbs    0.8951     0.3620   2.472   0.0134
```

Exponentiated coefficient estimates:



```
(Intercept)  birthwtlbs
  0.1756429   2.4474898
```

exp(B_0) exp(B_1)

The odds of having a normal Apgar score is 0.18 for an infant weighing 0 pounds.

A one pound increase in weight is associated with a 2.45 times increase in the odds of having a normal Apgar score.

Types of Predictor Variables

- Like linear regression, logistic regression can accommodate any type of predictor (exposure) variable
- Logistic regression can also accommodate multiple predictor variables
  - Each slope coefficient is interpreted “holding the other variable(s) constant” or “within levels of the other variable(s)”

Example: Apgar Score

- How are birth weight and having a germinal matrix hemorrhage associated with whether or not the infant has a normal Apgar score?
  - Interpret the regression coefficients
  - Calculate the predicted probability of having a normal Apgar score for an infant weighing 2 pounds who has a germinal matrix hemorrhage as well as for an infant weighing 2 pounds who does not have a germinal matrix hemorrhage.

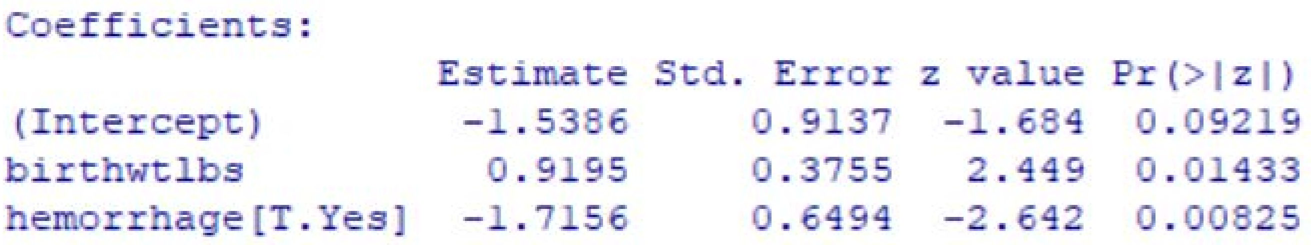
Let p be the probability of having a normal apgar score (y=1)

x1 be birthweight in pounds

x2 be having a germinal matrix hemorrhage

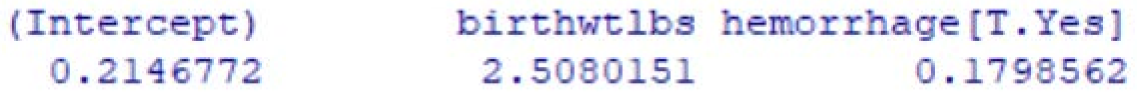
Example: Apgar Score

Logistic regression model (outcome is log odds of *apgarnormal*):



```
Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)         -1.5386     0.9137  -1.684  0.09219
birthwtlbs           0.9195     0.3755   2.449  0.01433
hemorrhage[T.Yes]   -1.7156     0.6494  -2.642  0.00825
```

Exponentiated coefficient estimates:



```
(Intercept)        birthwtlbs hemorrhage[T.Yes]
  0.2146772         2.5080151         0.1798562
```

The odds of having a normal Apgar score are 0.21 for an infant weighing 0 pounds who does not have germinal matrix hemorrhage.

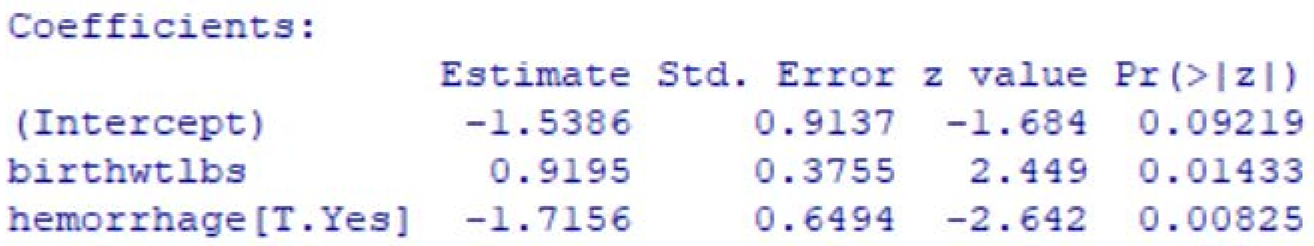
x1=0, x2=0 (gmh indicator)

odds = beta_hat_0 = 0.21

log(p/1-p)=-1.54+0.92(x1)-1.72(x2)

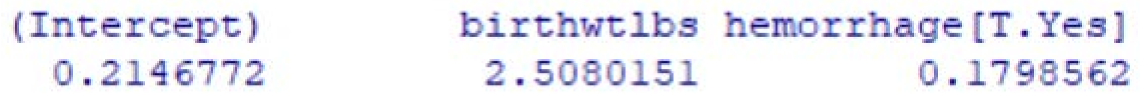
Example: Apgar Score

Logistic regression model (outcome is log odds of *apgarnormal*):



```
Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)        -1.5386     0.9137  -1.684  0.09219
birthwtlbs          0.9195     0.3755   2.449  0.01433
hemorrhage[T.Yes]  -1.7156     0.6494  -2.642  0.00825
```

Exponentiated coefficient estimates:



```
(Intercept)          birthwtlbs hemorrhage[T.Yes]
  0.2146772           2.5080151         0.1798562
```

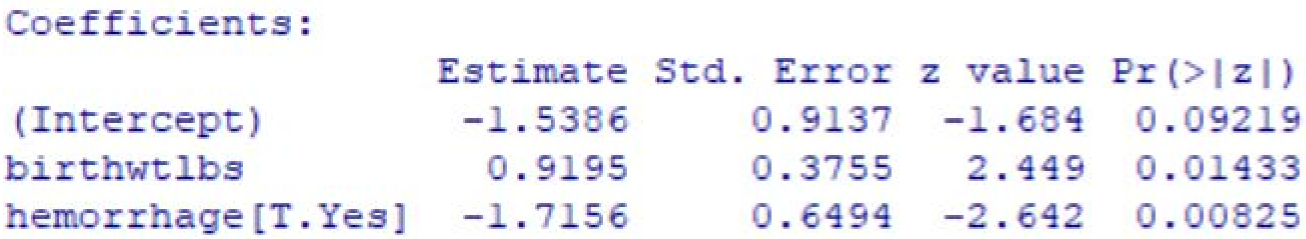
A one pound increase in weight is associated with a 2.51 times increase in the odds of having a normal Apgar score, holding germinal matrix hemorrhage constant.

The odds ratio of having a normal Apgar score is 2.51 for a one pound increase in birth weight, holding germinal matrix hemorrhage constant.

Consider two infants that are one pound apart in weight and have the same germinal matrix hemorrhage status (both Yes or both No). The infant who is one pound heavier has 2.51 times the odds of having a normal Apgar score compared to the infant who is one pound lighter.

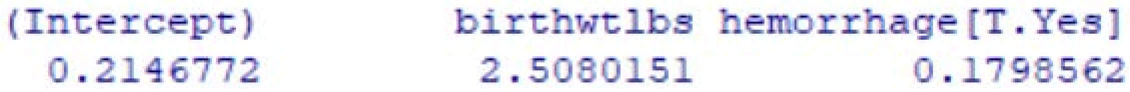
Example: Apgar Score

Logistic regression model (outcome is log odds of *apgarnormal*):



```
Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)          -1.5386     0.9137  -1.684  0.09219
birthwtlbs            0.9195     0.3755   2.449  0.01433
hemorrhage[T.Yes]    -1.7156     0.6494  -2.642  0.00825
```

Exponentiated coefficient estimates:



```
(Intercept)         birthwtlbs hemorrhage[T.Yes]
  0.2146772          2.5080151         0.1798562
```

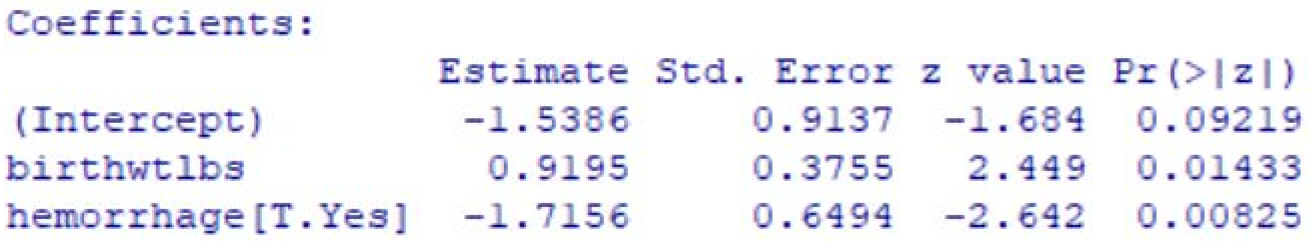
The odds of having a normal Apgar score for infants with germinal matrix hemorrhage are 0.18 times the odds of having a normal Apgar score for infants without germinal matrix hemorrhage, holding birth weight constant.

The odds ratio of having a normal Apgar score is 0.18 comparing infants with germinal matrix hemorrhage to those without germinal matrix hemorrhage, holding birth weight constant.

Consider two infants of the same birth weight, one having germinal matrix and one not having germinal matrix hemorrhage. The one with germinal matrix hemorrhage has odds of having a normal Apgar score that are about one-fifth the odds of having a normal Apgar score for the infant without germinal matrix hemorrhage.

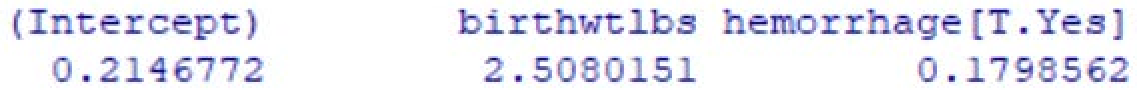
Example: Apgar Score

Logistic regression model (outcome is log odds of *apgarnormal*):



```
Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)         -1.5386     0.9137  -1.684  0.09219
birthwtlbs           0.9195     0.3755   2.449  0.01433
hemorrhage[T.Yes]   -1.7156     0.6494  -2.642  0.00825
```

Exponentiated coefficient estimates:



```
(Intercept)        birthwtlbs hemorrhage[T.Yes]
  0.2146772         2.5080151         0.1798562
```

Predicted probability of normal Apgar score for infant weighing 2 lbs and having germinal matrix hemorrhage:

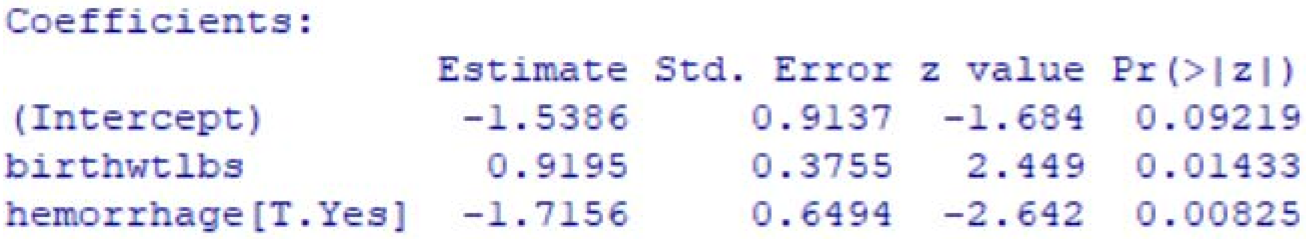
p=0.195

19.5% chance of having a normal apgar score if they weigh 2 pounds and have a germinal matrix hem.

$$\log\left(\frac{p}{1-p}\right) = -1.54 + 0.92(\underset{2}{birthwtlbs}) - 1.72(\underset{1}{hemorrhage})$$

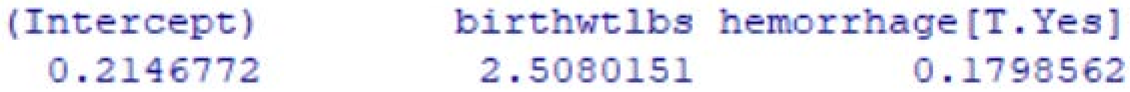
Example: Apgar Score

Logistic regression model (outcome is log odds of *apgarnormal*):



```
Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)       -1.5386     0.9137  -1.684  0.09219
birthwtlbs         0.9195     0.3755   2.449  0.01433
hemorrhage[T.Yes] -1.7156     0.6494  -2.642  0.00825
```

Exponentiated coefficient estimates:



```
(Intercept)        birthwtlbs hemorrhage[T.Yes]
  0.2146772         2.5080151         0.1798562
```

Predicted probability of normal Apgar score for infant weighing 2 lbs and not having germinal matrix hemorrhage:

p=0.575

$$\log\left(\frac{p}{1-p}\right) = -1.54 + 0.92(\underset{2}{birthwtlbs}) - 1.72(\underset{0}{hemorrhage})$$

Important Points

- Why linear regression is not appropriate for a binary outcome variable
- Logistic regression
  - Model setup/regression equation
  - Interpreting coefficients
  - Making predictions