SOC6707 Intermediate Data Analysis

Monica Alexander

Week 8: Logistic Regression

Notes

- Finish at 1230 today
- Confirming EDA extension (Wednesday)
- A3 out this week (short)
- Remainder of class
  - logistic regression
  - multinomial regression

Motivation

What if we are interested in modeling a binary response variable as a function of continuous and/or categorical explanatory variables?

- A binary response variable is an indicator variable that is coded 1 to indicate that an observation is a member of a particular group/category, and 0 otherwise
  - e.g. high income yes/no
  - has bachelor or higher yes/no
  - at least good self-reported health yes/no
  - life sentence yes/no
- Today we will see how we can build a regression model with a binary outcome as the dependent/response variable

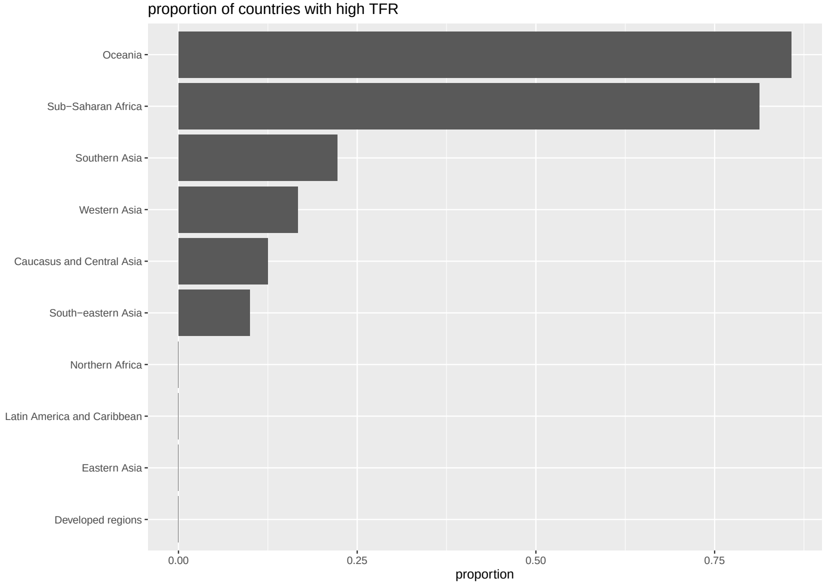
Binary dependent variable

- For example, let's use the country indicators dataset again.
- $Y_i = 1$ if a country has a high TFR (i.e. TFR > 3.5) and $Y_i = 0$ otherwise.
- Note that we have to create this variable using the `ifelse` function:

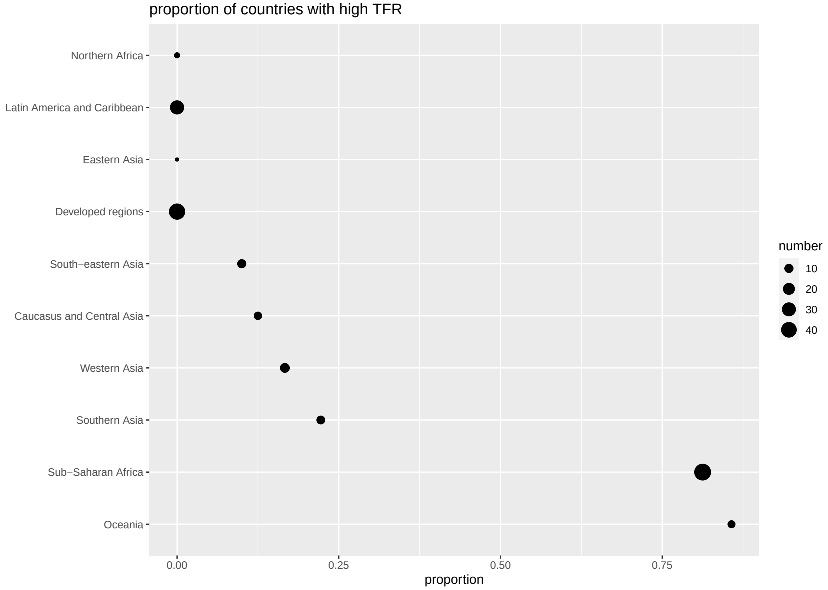
```
country_ind_2017 <- country_ind_2017 %>%
  mutate(high_tfr = ifelse(tfr>3.5, 1, 0))
head(country_ind_2017 %>% select(country, region, tfr, high_tfr))
```

```
## # A tibble: 6 x 4
##   country             region                       tfr high_tfr
##   <chr>               <chr>                      <dbl>    <dbl>
## 1 Afghanistan         Southern Asia               4.63        1
## 2 Albania             Developed regions           1.64        0
## 3 Algeria             Northern Africa             3.04        0
## 4 Angola              Sub-Saharan Africa          5.60        1
## 5 Antigua and Barbuda Latin America and Caribbean  2.00        0
## 6 Argentina           Latin America and Caribbean  2.28        0
```

Binary variables: looking at proportions is useful



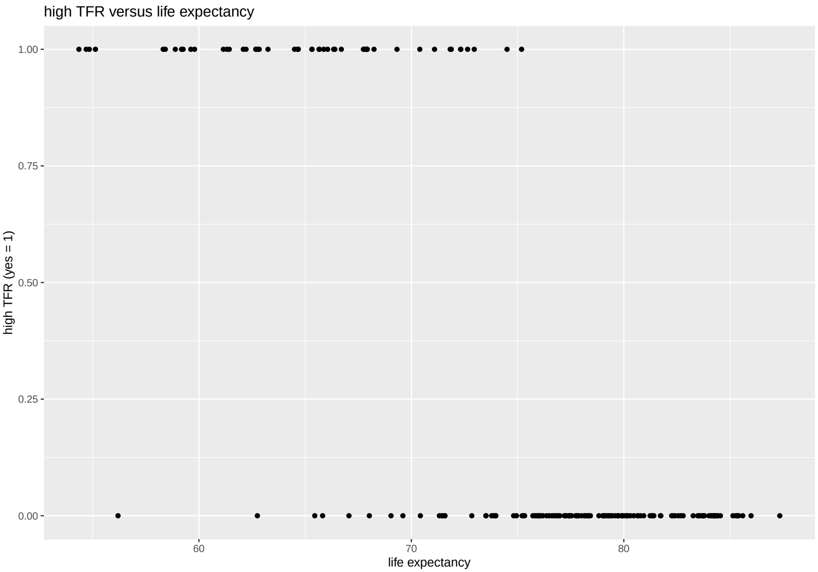
Another option



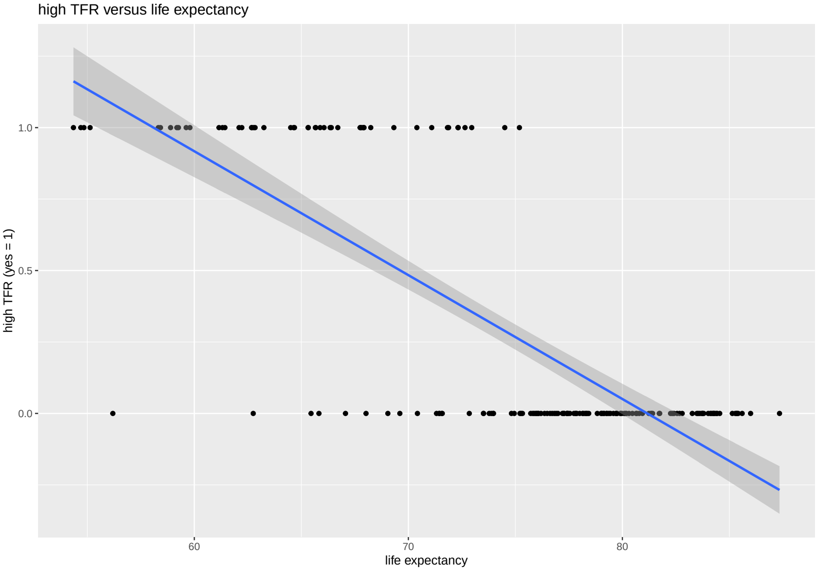
Binary dependent variable

- $Y_i = 1$ if a country has a high TFR (i.e. TFR > 3.5) and $Y_i = 0$ otherwise.
- We are interested in exploring how high TFR is associated with life expectancy and gross domestic product (GDP)
- What does this actually mean, given "high TFR" is 1 or 0 (yes or no)?

???


high TFR versus life expectancy
high TFR (yes = 1)
1.00
0.75
0.50
0.25
0.00
60
70
80
life expectancy

???


high TFR versus life expectancy
high TFR (yes = 1)
1.0
0.5
0.0
60
70
80
life expectancy

Binary dependent variable

- We are interested in exploring how high TFR is associated with life expectancy and gross domestic product (GDP)
- What does this actually mean, given "high TFR" is 1 or 0 (yes or no)?
- We are interested to see if the **probability** of high fertility is associated with life expectancy and GDP

But how do we model the probability in a regression framework?

FANCY AVERAGES: REDUX

The expectation of a binary variable

- Recall that the regression models we've looked at so far (SLR and MLR) are models for the **conditional expectation** $E(Y|X)$
- Conditional expectations are just fancy averages (averages conditioning on other variables of interest)
- So if we want to model a binary outcome as a dependent variable in a regression model, we first need to find the conditional expectation

The expectation of a binary variable

- Recall that for a discrete random variable, Y, with a known probability distribution $P(Y_i)$ and where Y_i is the ith outcome in the set of k simple events:

$$E(Y_i) = Y_1 \times P(Y_1) + Y_2 \times P(Y_2) + \ldots + Y_k \times P(Y_k) = \sum_{i=1}^{k} Y_i \times P(Y_i)$$

- So the expected value of a binary variable is

$$\begin{aligned} E(Y_i) &= (0)(1-p) + (1)p \\ &= p \end{aligned}$$

- That is, the expectation of a binary variable is equal to the probability that the variable is equal to one
- So the thing that we're interested in (probability) is actually the expected value of our outcome!

Conditional Expectation

- By extension, the conditional expectation of a binary variable is equal to the conditional probability that the variable is equal to one—that is,

$$E\left(Y_i \mid X_{i1}, \ldots, X_{ik}\right) = P\left(Y_i = 1 \mid X_{i1}, \ldots, X_{ik}\right)$$

So we know what the conditional expectation of our outcome of interest is, so can we just do linear regression now?

A complication

- The regression models discussed previously were direct models for the conditional expectation. But there's a complication here in that

$$E\left(Y_i \mid X_{i1}, \ldots, X_{ik}\right) = P\left(Y_i = 1 \mid X_{i1}, \ldots, X_{ik}\right)$$

  is bounded between values zero and one. It's a probability!
- We can get around this by first **transforming** the conditional expectation to be unbounded
- I.e. want to go from $y =$ probabilities to $y =$ function(probabilities) where the function lets y be any real number.

Logarithms

$$\log_b x$$

- The logarithm of a positive real number x with respect to base b is the exponent by which b must be raised to yield x.
- It is the inverse function to exponentiation
- The natural logarithm (often just written $\log x$) is to the base e, the mathematical constant $e \approx 2.718$

$$y = \log x$$

implies

$$x = e^y = \exp y$$

- You can think of taking the natural logarithm of x as transforming x to be on a different scale

The logit function

- The logit function takes a probability as its argument and then returns a value between negative infinity and positive infinity
- In other words, the logit transformation of a probability is unbounded even though the probability is bounded by the unit interval, [0,1]
- It is also called log-odds

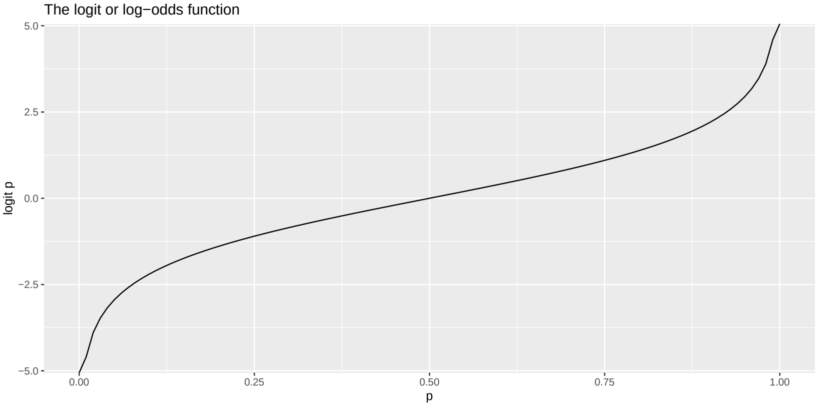
The logit function of probability p is

$$\text{logit } p = \log \frac{p}{1-p}$$

The logit function

For example,

$$\text{logit } 0.5 = \log \frac{0.5}{1 - 0.5} = \log 1 = 0$$



Aside: odds

Given probability p, odds are calculated as

$$\frac{p}{1-p}$$

- Odds provide a measure of the likelihood of a particular outcome. They are calculated as the ratio of the number of events that produce the outcome to the number that don't.
- Another way of expressing likelihood
- Often expressed as "1 to x"
- e.g. six sided die:
  - Probability rolling a $6 = ?$
  - Odds of rolling a $6 = ?$

Back to our problem

We want to model the conditional expectation

$$E\left(Y_i \mid X_{i1}, \ldots, X_{ik}\right) = P\left(Y_i = 1 \mid X_{i1}, \ldots, X_{ik}\right)$$

For simplicity/ease of reading, I'm going to use

$$p = P\left(Y_i = 1 \mid X_{i1}, \ldots, X_{ik}\right)$$

The logistic regression model

Logistic regression is a model for the conditional expectation of a binary response variable—that is, for the conditional probability that a binary response variable is equal to one.

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_k X_{ik}$$

where $\log\left(\frac{p}{1-p}\right)$ is known as the "log odds," or the "logit" transformation, and the β are unknown parameters to be estimated from data

The logistic regression model

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_k X_{ik}$$

We can rearrange this formula to get an expression for probability that $Y_i = 1$:

$$p = \frac{\exp\left(\beta_0 + \beta_1 X_{i1} + \cdots + \beta_k X_{ik}\right)}{1 + \exp\left(\beta_0 + \beta_1 X_{i1} + \cdots + \beta_k X_{ik}\right)}$$

- This is the inverse of the logit function
- The inverse of the logit link function is bounded by the unit interval (i.e., it falls between 0 and 1 for any value), which ensures that the conditional probabilities all fall within the logical range

The logistic regression model

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_k X_{ik}$$

To summarize:

- We transform probabilities to run a regression model that can have values anywhere on the real line
- We can then untransform these probabilities to get values back on the [0,1] scale

Interpreting logistic regression on the logit scale

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_k X_{ik}$$

What is β_0? Now I have to go back to the full notation:

$$\begin{aligned}\log\left(\tfrac{P(Y_i=1|X_{i1}=0,\ldots,X_{ik}=0)}{1-P(Y_i=1|X_{i1}=0,\ldots,X_{ik}=0)}\right) &= \beta_0 + \beta_1(0) + \cdots + \beta_k(0)\\ &= \beta_0\end{aligned}$$

β_0 is the log odds that $Y_i = 1$ given that all explanatory variables are equal to zero.

Interpreting logistic regression on the logit scale

What is β_1?

$$
\begin{aligned}
&\log\left(\frac{P\left(Y_i = 1 \mid X_{i1} = x_1^* + 1, X_{i2} = x_2^*, \ldots, X_{ik} = x_k^*\right)}{1 - P\left(Y_i = 1 \mid X_{i1} = x_1^* + 1, X_{i2} = x_2^*, \ldots, X_{ik} = x_k^*\right)}\right) \\
&\quad - \log\left(\frac{P\left(Y_i = 1 \mid X_{i1} = x_1^*, X_{i2} = x_2^*, \ldots, X_{ik} = x_k^*\right)}{1 - P\left(Y_i = 1 \mid X_{i1} = x_1^*, X_{i2} = x_2^*, \ldots, X_{ik} = x_k^*\right)}\right) \\
&\quad = \left(\beta_0 + \beta_1\left(x_1^* + 1\right) + \beta_2 x_2^* + \cdots + \beta_k x_k^*\right) - \left(\beta_0 + \beta_1 x_1^* + \beta_2 x_2^* + \cdots + \beta_k x_k^*\right) \\
&\quad = \beta_1
\end{aligned}
$$

β_1 is a log odds ratio, which gives the change in the log odds that $Y_i = 1$ associated with a unit increase in X_{i1}, holding other variables constant

Interpreting logistic regression on the odds scale

$$\log\left(\frac{P\left(Y_i = 1 \mid X_{i1}, \ldots, X_{ik}\right)}{1 - P\left(Y_i = 1 \mid X_{i1}, \ldots, X_{ik}\right)}\right) = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_k X_{ik}$$

What is $\exp \beta_0$?

$$\begin{aligned}\exp \beta_0 &= \exp\left(\log\left(\tfrac{P(Y_i=1|X_{i1}=0,\ldots,X_{ik}=0)}{1-P(Y_i=1|X_{i1}=0,\ldots,X_{ik}=0)}\right)\right)\\ &= \tfrac{P(Y_i=1|X_{i1}=0,\ldots,X_{ik}=0)}{1-P(Y_i=1|X_{i1}=0,\ldots,X_{ik}=0)}\end{aligned}$$

$\exp \beta_0$ is the odds that $Y_i = 1$ given that all explanatory variables are equal to zero.

Interpreting logistic regression on the odds scale

What is $\exp \beta_1$?

$$\exp\left(\beta_1\right) = \frac{P\left(Y_i = 1 \mid X_{i1} = x_1^* + 1, \ldots\right)}{1 - P\left(Y_i = 1 \mid X_{i1} = x_1^* + 1, \ldots\right)} \Big/ \frac{P\left(Y_i = 1 \mid X_{i1} = x_1^*, \ldots\right)}{1 - P\left(Y_i = 1 \mid X_{i1} = x_1^*, \ldots\right)}$$

$\exp \beta_1$ is a odds ratio, which ratio of the odds that $Y_i = 1$ associated with a unit increase in X_{i1}, holding other variables constant

Example in R

- Can run logistic regression in R using the `glm` function
- The additional `family` argument is related to the fact we are dealing with a binary response variable

```
lr_mod <- glm(high_tfr ~ life_expectancy + gdp,
              family = "binomial", data = country_ind_2017)
```

Example in R

```
summary(lr_mod)
```

```
##
## Call:
## glm(formula = high_tfr ~ life_expectancy + gdp, family = "binomial",
##     data = country_ind_2017)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -3.08570  -0.23518  -0.02127   0.18777   2.36080
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)     21.9815193  4.4298761   4.962 6.97e-07 ***
## life_expectancy -0.2960674  0.0627786  -4.716 2.40e-06 ***
## gdp             -0.0002081  0.0000654  -3.181  0.00147 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 211.886  on 175  degrees of freedom
## Residual deviance:  73.653  on 173  degrees of freedom
## AIC: 79.653
##
## Number of Fisher Scoring iterations: 8
```

Questions

Interpret

- β_1
- $\exp(\beta_1)$

```
coef(lr_mod)
```

```
##     (Intercept) life_expectancy             gdp
##   21.9815193435   -0.2960674332   -0.0002080662
```

```
exp(coef(lr_mod))
```

```
##     (Intercept) life_expectancy             gdp
##    3.519270e+09    7.437373e-01    9.997920e-01
```

Questions

- What is the probability of high TFR for a country with a life expectancy of 70 and a GDP of 9500?

```
beta0 <- coef(lr_mod)[[1]] # used double square brackets here to remove names (could use single)
beta1 <- coef(lr_mod)[[2]]
beta2 <- coef(lr_mod)[[3]]

estimated_log_odds <- beta0 + beta1*70 + beta2*9500

estimated_probability <- exp(estimated_log_odds)/(1+exp(estimated_log_odds))

estimated_log_odds
```

```
## [1] -0.7198301
```

```
estimated_probability
```

```
## [1] 0.3274304
```

Including a categorical explanatory variable

```
lr_mod_2 <- glm(high_tfr ~ region,
              family = "binomial", data = country_ind_2017)
summary(lr_mod_2)
```

```
##
## Call:
## glm(formula = high_tfr ~ region, family = "binomial", data = country_ind_2017)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.97277  -0.00005  -0.00005   0.64442   2.14597
##
## Coefficients:
##                                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)                         -1.9459     1.0690  -1.820  0.06872 .
## regionDeveloped regions            -18.6202  2672.9540  -0.007  0.99444
## regionEastern Asia                 -18.6202 10236.6339  -0.002  0.99855
## regionLatin America and Caribbean  -18.6202  3184.4686  -0.006  0.99533
## regionNorthern Africa              -18.6202  8865.1850  -0.002  0.99832
## regionOceania                        3.7377     1.5197   2.459  0.01391 *
## regionSouth-eastern Asia            -0.2513     1.5013  -0.167  0.86706
## regionSouthern Asia                  0.6931     1.3363   0.519  0.60397
## regionSub-Saharan Africa             3.4122     1.1312   3.016  0.00256 **
## regionWestern Asia                   0.3365     1.3202   0.255  0.79882
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 211.886  on 175  degrees of freedom
## Residual deviance:  84.947  on 166  degrees of freedom
## AIC: 104.95
##
## Number of Fisher Scoring iterations: 19
```

Categorical explanatory variables

- The coefficient on Sub-Saharan Africa is 3.41. What does this mean?

```
exp(coef(lr_mod_2)[9])
```

```
## regionSub-Saharan Africa
##                 30.33333
```

Inference

Some brief comments on the sampling distribution estimators $\hat{\beta}$

- Thinking back to SLR and MLR, if we believed the 5 assumptions stated, we could write down the sampling distribution for the estimator $\hat{\beta}$
- (It was a Normal distribution)
- We could then use this property to make inferences about how likely $\hat{\beta}$ was to be different from zero, for example (hypothesis testing)
- We can use a similar approach here with logistic regression

Asymptotic distribution of $\hat{\beta}$

- It is known that the limiting distribution of $\hat{\beta}_k$ is normal with a mean β_k and some variance (related to the properties of the estimator)
- Because the probability distribution of $\hat{\beta}_k$ converges to a normal distribution as the sample size increases, we can use this fact to make approximate inferences about β_k
- It turns out that the standardized version

  $$Z_{\widehat{\beta}_k} = \frac{\widehat{\beta}_k - \beta_k}{se\left(\widehat{\beta}_k\right)}$$

  follows a standard normal distribution, which we can use to make inferences about β_k

Hypothesis testing

- The β_k parameters are unknown population quantities of interest, which we have estimated with data from a random sample of the population
- We can test hypotheses about these unknown population quantities based on the fact that their standardized estimates follow an approximately standard normal distribution in large samples
- With knowledge of the distribution of $Z_{\widehat{\beta}_k}$ we can make probabilistic statements about the chances of observing any particular value of $Z_{\widehat{\beta}_k}$ given a hypothesized value for the unknown parameter of interest
- As before, we are usually testing the null hypothesis that $\beta_k = 0$
- This test is called the Wald test

The Wald test

1. State your null and alternative hypotheses about β_k
2. Choose the level of type-I error, α
3. Compute the Wald test statistic $z_{\widehat{\beta}_k} = \frac{\widehat{\beta}_k - \beta_k}{se\left(\widehat{\beta}_k\right)}$
4. Compute the p-value, which gives the probability of observing a test statistic as or more extreme than $z_{\widehat{\beta}_k}$ under the assumption that the null hypothesis is true
5. Make a decision (reject the null if the p-value is less than α, and fail to reject otherwise)

Reminder: think of the p-value as a summary measure of 'evidence against the null hypothesis' (and *not* as evidence for the alternative hypothesis)

Example

```
summary(lr_mod)
```

```
##
## Call:
## glm(formula = high_tfr ~ life_expectancy + gdp, family = "binomial",
##     data = country_ind_2017)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -3.08570  -0.23518  -0.02127   0.18777   2.36080
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)     21.9815193  4.4298761   4.962 6.97e-07 ***
## life_expectancy -0.2960674  0.0627786  -4.716 2.40e-06 ***
## gdp             -0.0002081  0.0000654  -3.181  0.00147 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 211.886  on 175  degrees of freedom
## Residual deviance:  73.653  on 173  degrees of freedom
## AIC: 79.653
##
## Number of Fisher Scoring iterations: 8
```

Summary

- Logistic regression can be used when the outcome of interest is binary (yes/no)
  - Aside: why logistic?
- You can have one or more explanatory variables, which can be quantitative or categorical
- Practically, running logistic regression in R is very similar to linear regression
- Interpretation is usually a bit harder