# Logistic regression II

Slides: initially developed by Mine Çetinkaya-Rundel and Curry W. Hilton of OpenIntro

# The logistic regression model

In the logistic regression model: logit(p) = log( p / (1-p)) is **linear** in x_i’s.

$$y_i \sim \text{Binom}(p_i)$$

$$\eta = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n$$

$$\text{logit}(p) = \eta$$

Binom(p_i): “probability of success = p_i”.  From which we arrive at,

$$p_i = \frac{\exp(\beta_0 + \beta_1 x_{1,i} + \cdots + \beta_n x_{n,i})}{1 + \exp(\beta_0 + \beta_1 x_{1,i} + \cdots + \beta_n x_{n,i})}$$

# Donner party data

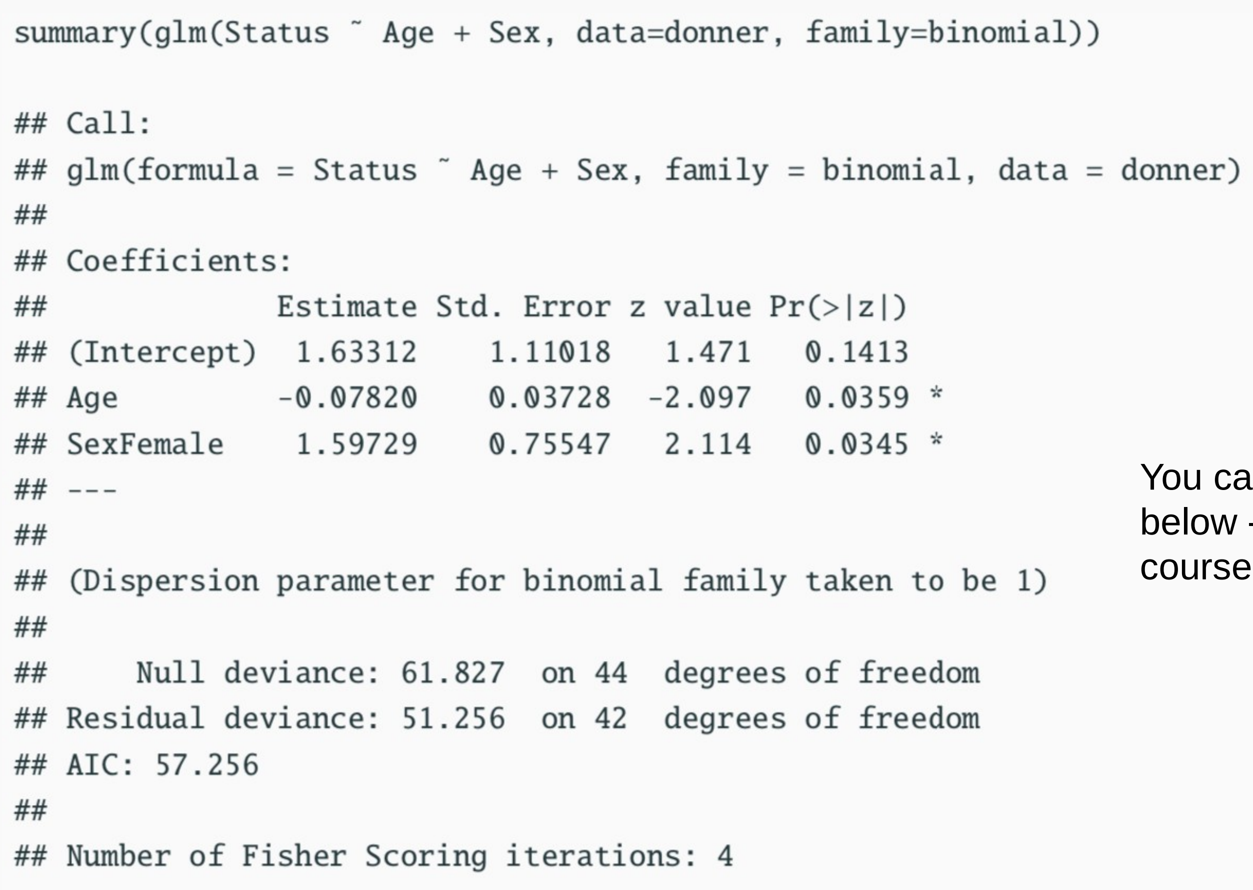
Let’s revisit the Donner party data from Monday

Recall we have variables:

- Survived (categorical)
- Age
- Sex
- Name

Let’s formulate a logistic regression model for predicting survival using Age and Sex. And then let’s examine a hypothesis test for whether or not Age is significant for predicting the log-odds of survival.

# Hypothesis test for the whole model

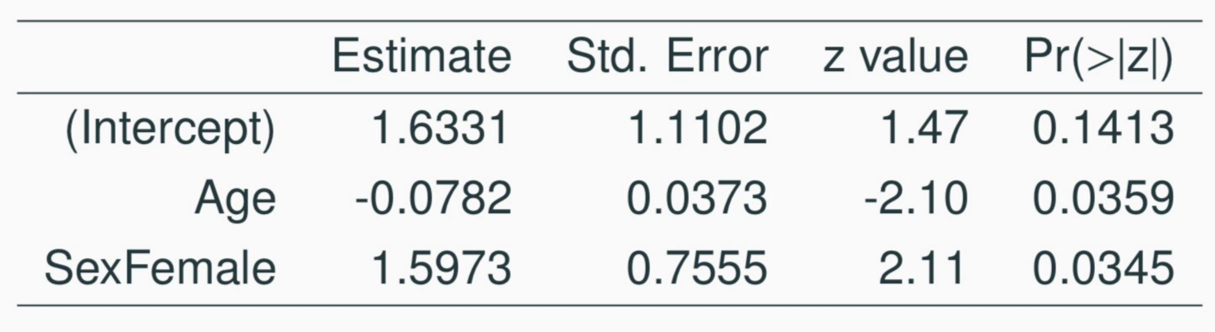


```
summary(glm(Status ~ Age + Sex, data=donner, family=binomial))

## Call:
## glm(formula = Status ~ Age + Sex, family = binomial, data = donner)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.63312    1.11018   1.471   0.1413
## Age         -0.07820    0.03728  -2.097   0.0359 *
## SexFemale    1.59729    0.75547   2.114   0.0345 *
## ---
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 61.827  on 44  degrees of freedom
## Residual deviance: 51.256  on 42  degrees of freedom
## AIC: 57.256
##
## Number of Fisher Scoring iterations: 4
```

You can ignore the stuff below ---, beyond scope of course.

# Hypothesis tests for a coefficient



| | Estimate | Std. Error | z value | Pr(>\|z\|) |
|---|---|---|---|---|
| (Intercept) | 1.6331 | 1.1102 | 1.47 | 0.1413 |
| Age | -0.0782 | 0.0373 | -2.10 | 0.0359 |
| SexFemale | 1.5973 | 0.7555 | 2.11 | 0.0345 |

We are still able to perform inference on individual coefficients, the basic setup is exactly the same as what we've seen before except we use a Z-test.

Note: Beyond the scope of this course to describe how standard error is calculated.

# Testing for the slope of Age

| | Estimate | Std. Error | z value | Pr(>\|z\|) |
|---|---|---|---|---|
| (Intercept) | 1.6331 | 1.1102 | 1.47 | 0.1413 |
| Age | -0.0782 | 0.0373 | -2.10 | 0.0359 |
| SexFemale | 1.5973 | 0.7555 | 2.11 | 0.0345 |

$$H_0 : \beta_{age} = 0$$

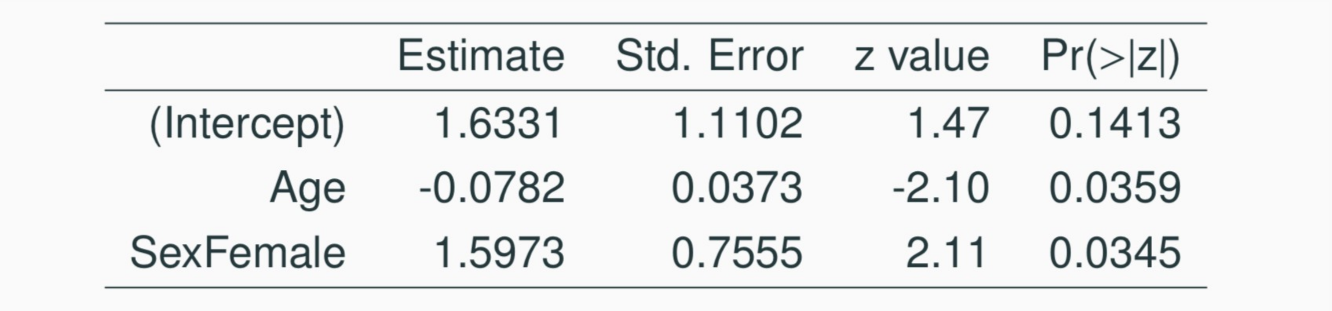
$$H_A : \beta_{age} \neq 0$$

$$Z = \frac{\hat{\beta}_{age} - \beta_{age}}{SE_{age}} = \frac{-0.0782 - 0}{0.0373} = -2.10$$

$$\text{p-value} = P(|Z| > 2.10) = P(Z > 2.10) + P(Z < -2.10)$$

$$= 2 \times 0.0178 = 0.0359$$

# Confidence interval for age slope coefficient



| | Estimate | Std. Error | z value | Pr(>\|z\|) |
|---|---|---|---|---|
| (Intercept) | 1.6331 | 1.1102 | 1.47 | 0.1413 |
| Age | -0.0782 | 0.0373 | -2.10 | 0.0359 |
| SexFemale | 1.5973 | 0.7555 | 2.11 | 0.0345 |

Remember, the interpretation for a slope is the change in log odds ratio per unit change in the predictor.

We can create confidence interval using point est. +/- margin of error.

Log odds ratio CI if we want 95% confidence (z* = 1.96):

$$CI = PE \pm CV \times SE = -0.0782 \pm 1.96 \times 0.0373 = (-0.1513, -0.0051)$$

Odds ratio:

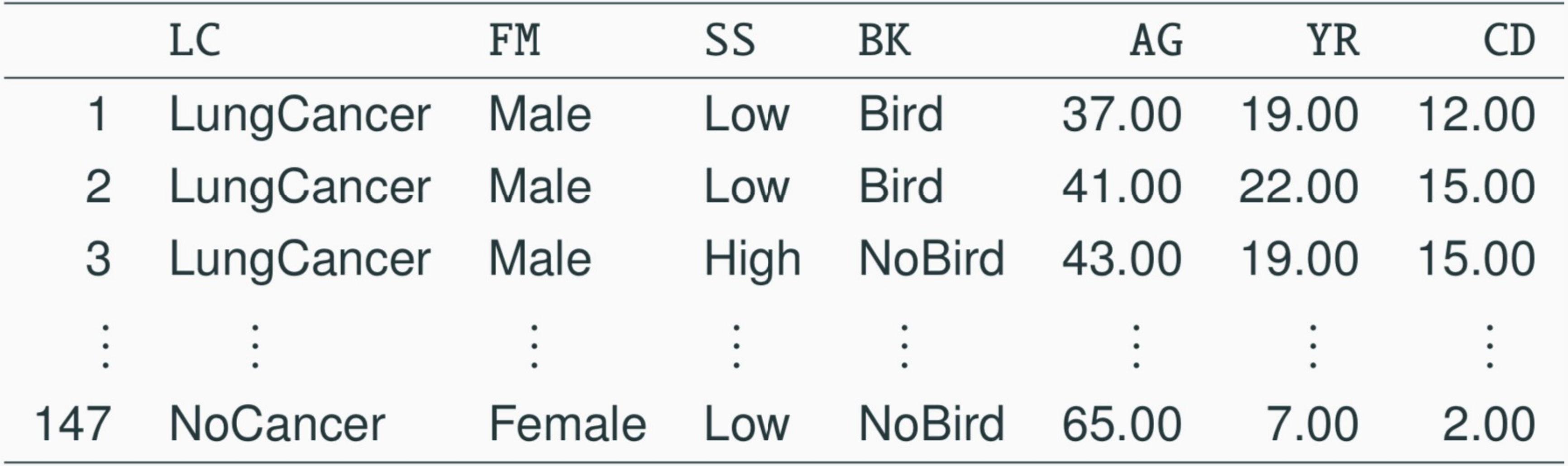
$$\exp(CI) = (\exp(-0.1513), \exp(-0.0051)) = (0.8596, 0.9949)$$

# Example - Birdkeeping and Lung Cancer

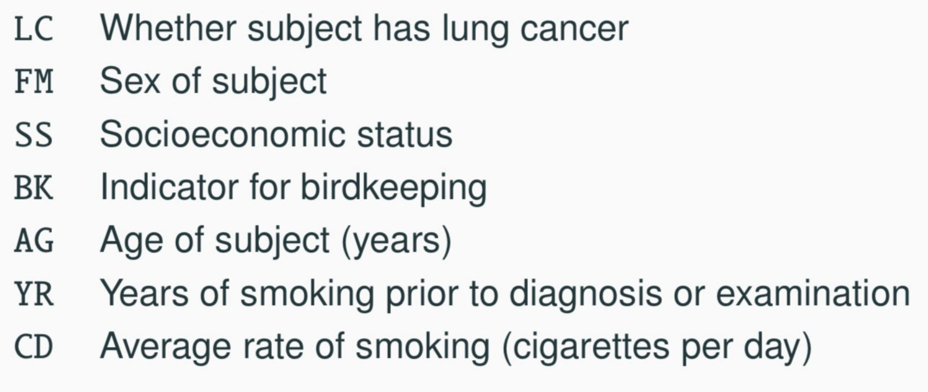
A 1972 - 1981 health survey in The Hague, Netherlands, discovered an association between keeping pet birds and increased risk of lung cancer. To investigate birdkeeping as a risk factor, researchers conducted a case-control study of patients in 1985 at four hospitals in The Hague (population 450,000). They identified 49 cases of lung cancer among the patients who were registered with a general practice, who were age 65 or younger and who had resided in the city since 1965. They also selected 98 controls from a population of residents having the same general age structure.

From *Ramsey, F.L. and Schafer, D.W. (2002). The Statistical Sleuth: A Course in Methods of Data Analysis (2nd ed)*

# Example - Birdkeeping and Lung Cancer - Data

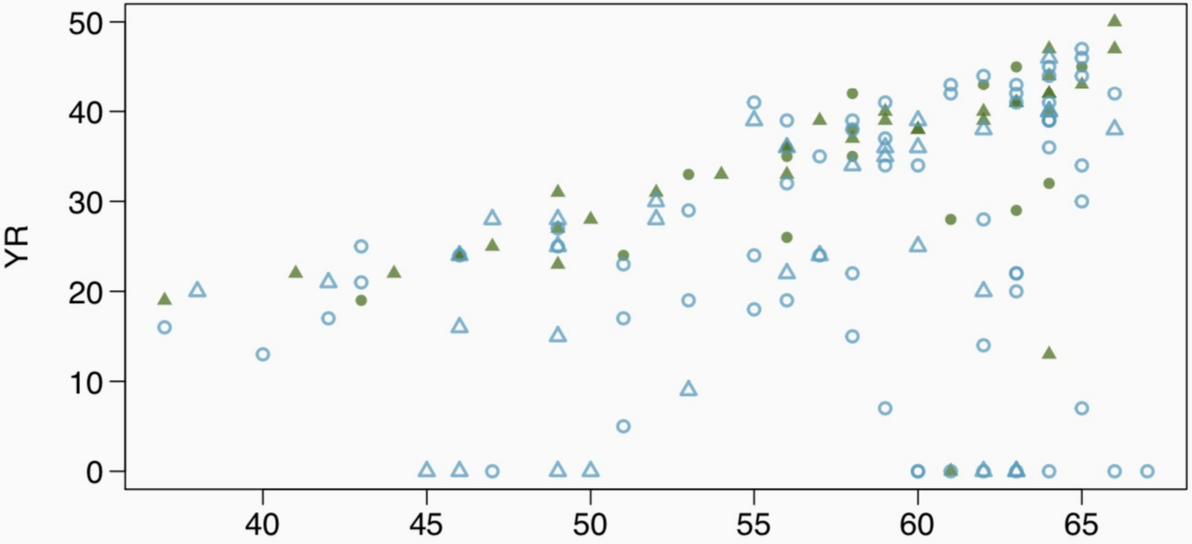


|  | LC | FM | SS | BK | AG | YR | CD |
|---|---|---|---|---|---|---|---|
| 1 | LungCancer | Male | Low | Bird | 37.00 | 19.00 | 12.00 |
| 2 | LungCancer | Male | Low | Bird | 41.00 | 22.00 | 15.00 |
| 3 | LungCancer | Male | High | NoBird | 43.00 | 19.00 | 15.00 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| 147 | NoCancer | Female | Low | NoBird | 65.00 | 7.00 | 2.00 |



| | |
|---|---|
| LC | Whether subject has lung cancer |
| FM | Sex of subject |
| SS | Socioeconomic status |
| BK | Indicator for birdkeeping |
| AG | Age of subject (years) |
| YR | Years of smoking prior to diagnosis or examination |
| CD | Average rate of smoking (cigarettes per day) |

# Example - Birdkeeping and Lung Cancer - EDA

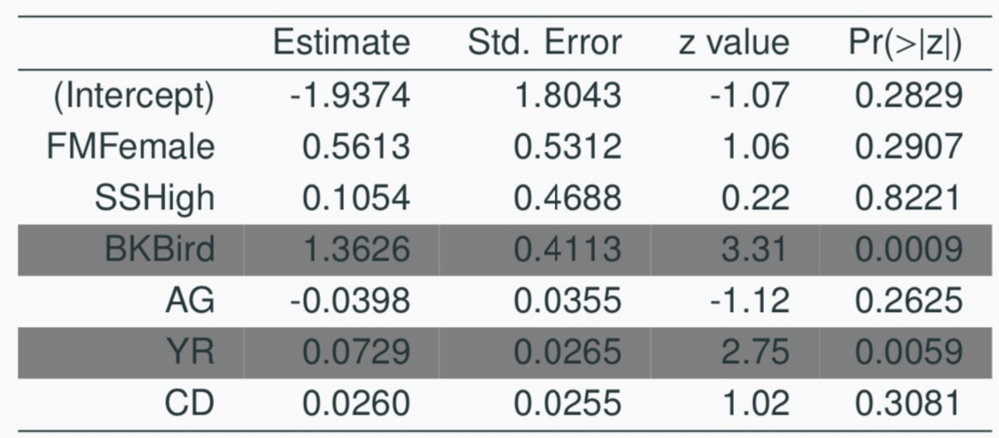


| | Bird | No Bird |
|---|---|---|
| Lung Cancer | ▲ | ● |
| No Lung Cancer | △ | ○ |

# Example - Birdkeeping and Lung Cancer - Model

```
summary(glm(LC ~ FM + SS + BK + AG + YR + CD, data=bird, family=binomial))
## Call:
## glm(formula = LC ˜ FM + SS + BK + AG + YR + CD, family = binomial,
##     data = bird)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.93736  1.80425    -1.074 0.282924
## FMFemale     0.56127  0.53116     1.057 0.290653
## SSHigh       0.10545  0.46885     0.225 0.822050
## BKBird       1.36259  0.41128     3.313 0.000923 ***
## AG          -0.03976  0.03548    -1.120 0.262503
## YR           0.07287  0.02649     2.751 0.005940 **
## CD           0.02602  0.02552     1.019 0.308055
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 187.14  on 146  degrees of freedom
## Residual deviance: 154.20  on 140  degrees of freedom
## AIC: 168.2
```

# Example - Birdkeeping and Lung Cancer - Interpretation



|  | Estimate | Std. Error | z value | Pr(>\|z\|) |
|---|---|---|---|---|
| (Intercept) | -1.9374 | 1.8043 | -1.07 | 0.2829 |
| FMFemale | 0.5613 | 0.5312 | 1.06 | 0.2907 |
| SSHigh | 0.1054 | 0.4688 | 0.22 | 0.8221 |
| BKBird | 1.3626 | 0.4113 | 3.31 | 0.0009 |
| AG | -0.0398 | 0.0355 | -1.12 | 0.2625 |
| YR | 0.0729 | 0.0265 | 2.75 | 0.0059 |
| CD | 0.0260 | 0.0255 | 1.02 | 0.3081 |

Keeping all other predictors constant then,

- The odds ratio of getting lung cancer for bird keepers vs non-bird keepers is exp(1.3626) = 3.91.
- The odds ratio of getting lung cancer for an additional year of smoking is exp(0.0729) = 1.08.

# What do the numbers not mean ...

The most common mistake made when interpreting logistic regression is to treat an odds ratio as a ratio of probabilities.

Bird keepers are not 4x more likely to develop lung cancer than non-bird keepers.

This is the difference between **relative risk (RR)** and **an odds ratio (OR)**.

$$RR = \frac{P(\text{disease}|\text{exposed})}{P(\text{disease}|\text{unexposed})}$$

$$OR = \frac{P(\text{disease}|\text{exposed})/[1 - P(\text{disease}|\text{exposed})]}{P(\text{disease}|\text{unexposed})/[1 - P(\text{disease}|\text{unexposed})]}$$

# Back to the birds

What is probability of lung cancer in a bird keeper if we knew that P(lung cancer|no birds) = 0.05?

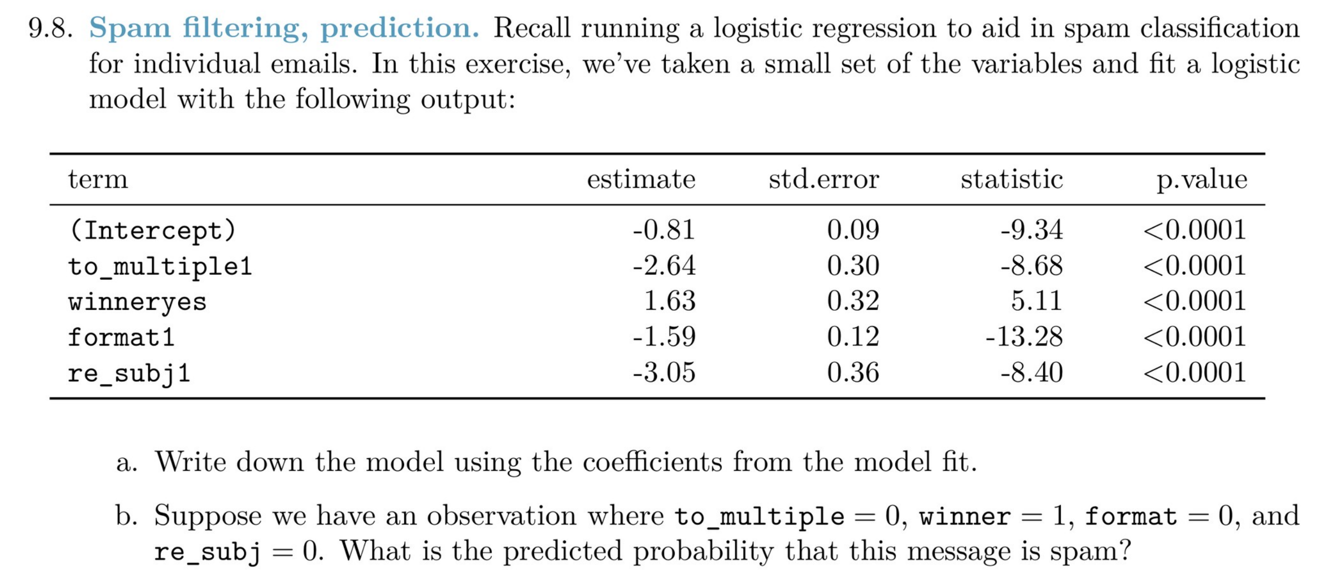
$$OR = \frac{P(\text{lung cancer}|\text{birds})/[1 - P(\text{lung cancer}|\text{birds})]}{P(\text{lung cancer}|\text{no birds})/[1 - P(\text{lung cancer}|\text{no birds})]}$$

$$= \frac{P(\text{lung cancer}|\text{birds})/[1 - P(\text{lung cancer}|\text{birds})]}{0.05/[1 - 0.05]} = 3.91$$

$$P(\text{lung cancer}|\text{birds}) = \frac{3.91 \times \frac{0.05}{0.95}}{1 + 3.91 \times \frac{0.05}{0.95}} = 0.171$$

$$RR = P(\text{lung cancer}|\text{birds})/P(\text{lung cancer}|\text{no birds}) = 0.171/0.05 = 3.41$$

# Practice:



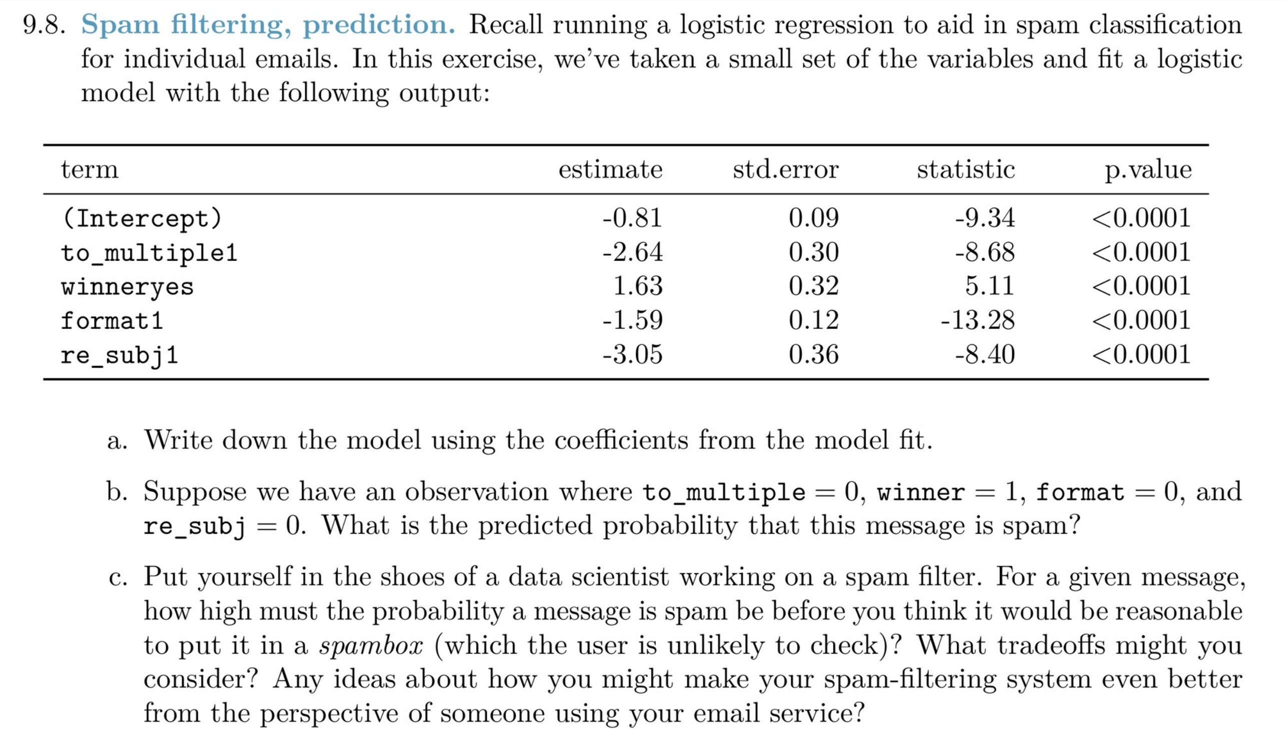
9.8. **Spam filtering, prediction.** Recall running a logistic regression to aid in spam classification for individual emails. In this exercise, we've taken a small set of the variables and fit a logistic model with the following output:

| term | estimate | std.error | statistic | p.value |
|---|---:|---:|---:|---:|
| `(Intercept)` | -0.81 | 0.09 | -9.34 | $<0.0001$ |
| `to_multiple1` | -2.64 | 0.30 | -8.68 | $<0.0001$ |
| `winneryes` | 1.63 | 0.32 | 5.11 | $<0.0001$ |
| `format1` | -1.59 | 0.12 | -13.28 | $<0.0001$ |
| `re_subj1` | -3.05 | 0.36 | -8.40 | $<0.0001$ |

a. Write down the model using the coefficients from the model fit.

b. Suppose we have an observation where `to_multiple` $= 0$, `winner` $= 1$, `format` $= 0$, and `re_subj` $= 0$. What is the predicted probability that this message is spam?

# Practice:



9.8. **Spam filtering, prediction.** Recall running a logistic regression to aid in spam classification for individual emails. In this exercise, we've taken a small set of the variables and fit a logistic model with the following output:

| term | estimate | std.error | statistic | p.value |
|---|---|---|---|---|
| `(Intercept)` | -0.81 | 0.09 | -9.34 | <0.0001 |
| `to_multiple1` | -2.64 | 0.30 | -8.68 | <0.0001 |
| `winneryes` | 1.63 | 0.32 | 5.11 | <0.0001 |
| `format1` | -1.59 | 0.12 | -13.28 | <0.0001 |
| `re_subj1` | -3.05 | 0.36 | -8.40 | <0.0001 |

a. Write down the model using the coefficients from the model fit.

b. Suppose we have an observation where `to_multiple` $= 0$, `winner` $= 1$, `format` $= 0$, and `re_subj` $= 0$. What is the predicted probability that this message is spam?

c. Put yourself in the shoes of a data scientist working on a spam filter. For a given message, how high must the probability a message is spam be before you think it would be reasonable to put it in a *spambox* (which the user is unlikely to check)? What tradeoffs might you consider? Any ideas about how you might make your spam-filtering system even better from the perspective of someone using your email service?

Suppose we had data for attending grad school as a function of GPA and number of year to graduate, and we formed a logistic regression model, which had the following (incomplete!) table.

| Term | Estimate | Std Error | Z value | P(>\|z\|) |
|---|---|---|---|---|
| (intercept) | 0.52 | 0.02 | | |
| GPA | 3.0 | 0.7 | | |
| YearsToGrad | -1.5 | 0.5 | | |
| | | | | |

Can you fill out the rest of the table? Do you need additional information/tools? Describe how to fill as much as you can, and describe any issues you run into.

# Final exam topics

* Mutate, group by, summarize
* regex, including str_remove, str_replace, str_replace_all
* pivot_longer, pivot_wider
* joining (left_join, anti_join, inner_join, full_join )
* plotting: ggplot, geom_{bar,point,histogram,boxplot}, color / shape / groupings
* functions – incl. using {{embrace}}, default values, `across()`, `if_any()`, `if_all()`
* linear regression - residuals, correlation, least squares line, `lm()`, R^2 and concepts of SST/SSE, categorical variables in linear regression, adjusted R^2 and model selection
* hypothesis tests / confidence intervals / p-values using randomization tests, bootstrap sampling, bootstrap confidence intervals, mathematical approaches (normal distribution / t distribution / F distribution)
  - estimating a single proportion
  - comparing two proportions
  - estimating a single mean
  - comparing two means
  - comparing multiple means
  - slope in linear regression
  - logistic regression