Chapter 8: Multiple and logistic regression

OpenIntro Statistics, 3rd Edition

Slides developed by Mine Çetinkaya-Rundel of OpenIntro.

Introduction to multiple regression

Multiple regression

- Simple linear regression: Bivariate - two variables: y and x
- Multiple linear regression: Multiple variables: y and $x_1, x_2, \cdots$

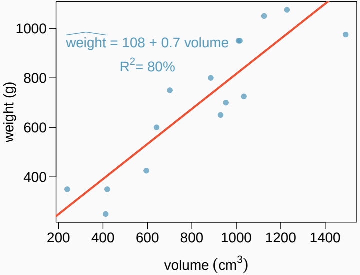
Weights of books

	weight (g)	volume (cm^3)	cover
1	800	885	hc
2	950	1016	hc
3	1050	1125	hc
4	350	239	hc
5	750	701	hc
6	600	641	hc
7	1075	1228	hc
8	250	412	pb
9	700	953	pb
10	650	929	pb
11	975	1492	pb
12	350	419	pb
13	950	1010	pb
14	425	595	pb
15	725	1034	pb

From: Maindonald, J.H. and Braun, W.J. (2nd ed., 2007) “Data Analysis and Graphics Using R”

Weights of books (cont.)

The scatterplot shows the relationship between weights and volumes of books as well as the regression output. Which of the below is correct?



(a) Weights of 80% of the books can be predicted accurately using this model.

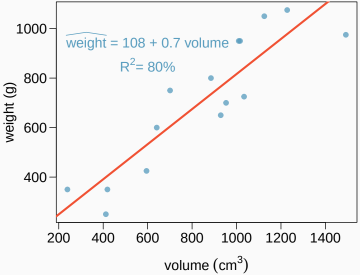
(b) Books that are 10 cm^3 over average are expected to weigh 7 g over average.

(c) The correlation between weight and volume is $R = 0.80^2 = 0.64$.

(d) The model underestimates the weight of the book with the highest volume.

Weights of books (cont.)

The scatterplot shows the relationship between weights and volumes of books as well as the regression output. Which of the below is correct?



(a) Weights of 80% of the books can be predicted accurately using this model.

(b) *Books that are 10 cm^3 over average are expected to weigh 7 g over average.*

(c) The correlation between weight and volume is $R = 0.80^2 = 0.64$.

(d) The model underestimates the weight of the book with the highest volume.

Modeling weights of books using volume

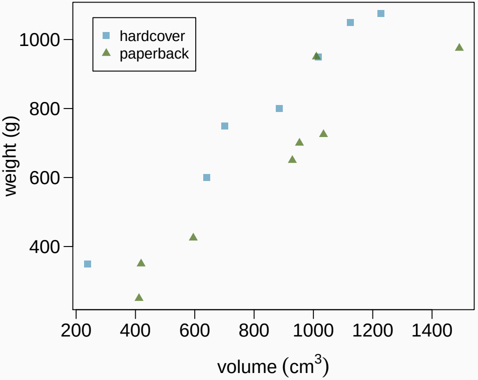
somewhat abbreviated output...

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 107.67931   88.37758   1.218    0.245
volume        0.70864    0.09746   7.271 6.26e-06


Residual standard error: 123.9 on 13 degrees of freedom
Multiple R-squared: 0.8026,Adjusted R-squared: 0.7875
F-statistic: 52.87 on 1 and 13 DF,  p-value: 6.262e-06
```

Weights of hardcover and paperback books

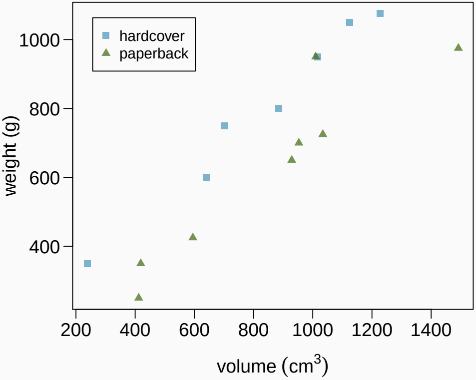
Can you identify a trend in the relationship between volume and weight of hardcover and paperback books?



Weights of hardcover and paperback books

Can you identify a trend in the relationship between volume and weight of hardcover and paperback books?

Paperbacks generally weigh less than hardcover books after controlling for the bookÕs volume.



Modeling weights of books using volume <u>and</u> cover type

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  197.96284   59.19274   3.344 0.005841 **
volume         0.71795    0.06153  11.669  6.6e-08 ***
cover:pb    -184.04727   40.49420  -4.545 0.000672 ***

Residual standard error: 78.2 on 12 degrees of freedom
Multiple R-squared: 0.9275,Adjusted R-squared: 0.9154
F-statistic: 76.73 on 2 and 12 DF,  p-value: 1.455e-07
```

Determining the reference level

Based on the regression output below, which level of `cover` is the reference level? Note that `pb`: paperback.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	197.9628	59.1927	3.34	0.0058
volume	0.7180	0.0615	11.67	0.0000
cover:pb	-184.0473	40.4942	-4.55	0.0007

(a) paperback

(b) hardcover

Determining the reference level

Based on the regression output below, which level of `cover` is the reference level? Note that `pb`: paperback.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	197.9628	59.1927	3.34	0.0058
volume	0.7180	0.0615	11.67	0.0000
cover:pb	-184.0473	40.4942	-4.55	0.0007

(a) paperback

(b) *hardcover*

Determining the reference level

Which of the below correctly describes the roles of variables in this regression model?

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	197.9628	59.1927	3.34	0.0058
volume	0.7180	0.0615	11.67	0.0000
cover:pb	-184.0473	40.4942	-4.55	0.0007

(a) response: weight, explanatory: volume, paperback cover

(b) response: weight, explanatory: volume, hardcover cover

(c) response: volume, explanatory: weight, cover type

(d) response: weight, explanatory: volume, cover type

Determining the reference level

Which of the below correctly describes the roles of variables in this regression model?

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	197.9628	59.1927	3.34	0.0058
volume	0.7180	0.0615	11.67	0.0000
cover:pb	-184.0473	40.4942	-4.55	0.0007

(a) response: weight, explanatory: volume, paperback cover

(b) response: weight, explanatory: volume, hardcover cover

(c) response: volume, explanatory: weight, cover type

(d) *response: weight, explanatory: volume, cover type*

Linear model

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	197.96	59.19	3.34	0.01
volume	0.72	0.06	11.67	0.00
cover:pb	-184.05	40.49	-4.55	0.00

Linear model

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	197.96	59.19	3.34	0.01
volume	0.72	0.06	11.67	0.00
cover:pb	-184.05	40.49	-4.55	0.00

$$\widehat{weight} = 197.96 + 0.72\ volume - 184.05\ cover:pb$$

Linear model

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	197.96	59.19	3.34	0.01
volume	0.72	0.06	11.67	0.00
cover:pb	-184.05	40.49	-4.55	0.00

$$\widehat{weight} = 197.96 + 0.72\ volume - 184.05\ cover:pb$$

1. For *hardcover* books: plug in *0* for `cover`

$$\widehat{weight} = 197.96 + 0.72\ volume - 184.05 \times 0$$

Linear model

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	197.96	59.19	3.34	0.01
volume	0.72	0.06	11.67	0.00
cover:pb	-184.05	40.49	-4.55	0.00

$$\widehat{weight} = 197.96 + 0.72\ volume - 184.05\ cover : pb$$

1. For *hardcover* books: plug in *0* for `cover`

$$\begin{aligned}\widehat{weight} &= 197.96 + 0.72\ volume - 184.05 \times 0 \\ &= 197.96 + 0.72\ volume\end{aligned}$$

Linear model

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	197.96	59.19	3.34	0.01
volume	0.72	0.06	11.67	0.00
cover:pb	-184.05	40.49	-4.55	0.00

$$\widehat{weight} = 197.96 + 0.72\ volume - 184.05\ cover : pb$$

1. For *hardcover* books: plug in *0* for `cover`

$$\begin{aligned}\widehat{weight} &= 197.96 + 0.72\ volume - 184.05 \times 0 \\ &= 197.96 + 0.72\ volume\end{aligned}$$

2. For *paperback* books: plug in *1* for `cover`

$$\widehat{weight} = 197.96 + 0.72\ volume - 184.05 \times 1$$

Linear model

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	197.96	59.19	3.34	0.01
volume	0.72	0.06	11.67	0.00
cover:pb	-184.05	40.49	-4.55	0.00

$$\widehat{weight} = 197.96 + 0.72\ volume - 184.05\ cover : pb$$

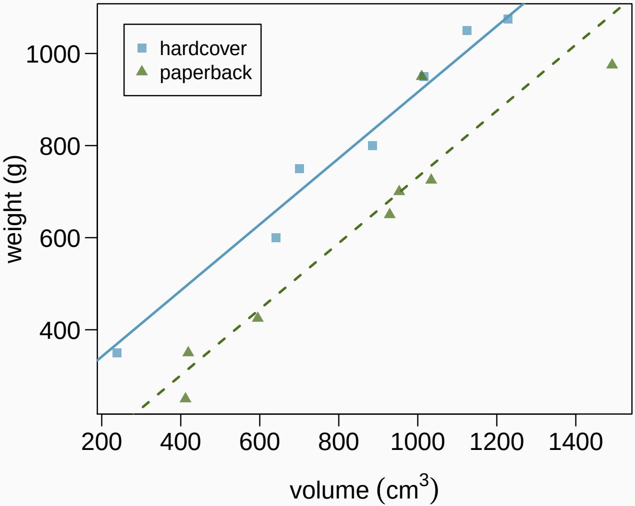
1. For *hardcover* books: plug in *0* for `cover`

$$\begin{aligned} \widehat{weight} &= 197.96 + 0.72\ volume - 184.05 \times 0 \\ &= 197.96 + 0.72\ volume \end{aligned}$$

2. For *paperback* books: plug in *1* for `cover`

$$\begin{aligned} \widehat{weight} &= 197.96 + 0.72\ volume - 184.05 \times 1 \\ &= 13.91 + 0.72\ volume \end{aligned}$$

Visualising the linear model



Interpretation of the regression coefficients

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	197.96	59.19	3.34	0.01
volume	0.72	0.06	11.67	0.00
cover:pb	-184.05	40.49	-4.55	0.00

Interpretation of the regression coefficients

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	197.96	59.19	3.34	0.01
volume	0.72	0.06	11.67	0.00
cover:pb	-184.05	40.49	-4.55	0.00

- *Slope of volume:* All else held constant, books that are 1 more cubic centimeter in volume tend to weigh about 0.72 grams more.

Interpretation of the regression coefficients

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	197.96	59.19	3.34	0.01
volume	0.72	0.06	11.67	0.00
cover:pb	-184.05	40.49	-4.55	0.00

- *Slope of volume:* All else held constant, books that are 1 more cubic centimeter in volume tend to weigh about 0.72 grams more.
- *Slope of cover:* All else held constant, the model predicts that paperback books weigh 184 grams lower than hardcover books.

Interpretation of the regression coefficients

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	197.96	59.19	3.34	0.01
volume	0.72	0.06	11.67	0.00
cover:pb	-184.05	40.49	-4.55	0.00

- *Slope of volume:* All else held constant, books that are 1 more cubic centimeter in volume tend to weigh about 0.72 grams more.
- *Slope of cover:* All else held constant, the model predicts that paperback books weigh 184 grams lower than hardcover books.
- *Intercept:* Hardcover books with no volume are expected on average to weigh 198 grams.

Interpretation of the regression coefficients

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	197.96	59.19	3.34	0.01
volume	0.72	0.06	11.67	0.00
cover:pb	-184.05	40.49	-4.55	0.00

- *Slope of volume:* All else held constant, books that are 1 more cubic centimeter in volume tend to weigh about 0.72 grams more.
- *Slope of cover:* All else held constant, the model predicts that paperback books weigh 184 grams lower than hardcover books.
- *Intercept:* Hardcover books with no volume are expected on average to weigh 198 grams.
  - Obviously, the intercept does not make sense in context. It only serves to adjust the height of the line.

Prediction

Which of the following is the correct calculation for the predicted weight of a paperback book that is 600 cm^3?

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	197.96	59.19	3.34	0.01
volume	0.72	0.06	11.67	0.00
cover:pb	-184.05	40.49	-4.55	0.00

(a) 197.96 + 0.72 * 600 - 184.05 * 1

(b) 184.05 + 0.72 * 600 - 197.96 * 1

(c) 197.96 + 0.72 * 600 - 184.05 * 0

(d) 197.96 + 0.72 * 1 - 184.05 * 600

Prediction

Which of the following is the correct calculation for the predicted weight of a paperback book that is 600 cm^3?

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	197.96	59.19	3.34	0.01
volume	0.72	0.06	11.67	0.00
cover:pb	-184.05	40.49	-4.55	0.00

(a) *197.96 + 0.72 * 600 - 184.05 * 1* $= 445.91$ *grams*

(b) 184.05 + 0.72 * 600 - 197.96 * 1

(c) 197.96 + 0.72 * 600 - 184.05 * 0

(d) 197.96 + 0.72 * 1 - 184.05 * 600

Another example: Modeling kid's test scores

Predicting cognitive test scores of three- and four-year-old children using characteristics of their mothers. Data are from a survey of adult American women and their children - a subsample from the National Longitudinal Survey of Youth.

	kid_score	mom_hs	mom_iq	mom_work	mom_age
1	65	yes	121.12	yes	27
⋮					
5	115	yes	92.75	yes	27
6	98	no	107.90	no	18
⋮					
434	70	yes	91.25	yes	25

Gelman, Hill. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. (2007) Cambridge University Press.

Interpreting the slope

What is the correct interpretation of the slope for mom's IQ?

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	19.59	9.22	2.13	0.03
mom_hs:yes	5.09	2.31	2.20	0.03
mom_iq	0.56	0.06	9.26	0.00
mom_work:yes	2.54	2.35	1.08	0.28
mom_age	0.22	0.33	0.66	0.51

, kids with mothers whose IQs are one point higher tend to score on average 0.56 points higher.

Interpreting the slope

What is the correct interpretation of the slope for mom's IQ?

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	19.59	9.22	2.13	0.03
mom_hs:yes	5.09	2.31	2.20	0.03
mom_iq	0.56	0.06	9.26	0.00
mom_work:yes	2.54	2.35	1.08	0.28
mom_age	0.22	0.33	0.66	0.51

All else held constant, kids with mothers whose IQs are one point higher tend to score on average 0.56 points higher.

Interpreting the slope

What is the correct interpretation of the intercept?

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	19.59	9.22	2.13	0.03
mom_hs:yes	5.09	2.31	2.20	0.03
mom_iq	0.56	0.06	9.26	0.00
mom_work:yes	2.54	2.35	1.08	0.28
mom_age	0.22	0.33	0.66	0.51

Interpreting the slope

What is the correct interpretation of the intercept?

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	19.59	9.22	2.13	0.03
mom_hs:yes	5.09	2.31	2.20	0.03
mom_iq	0.56	0.06	9.26	0.00
mom_work:yes	2.54	2.35	1.08	0.28
mom_age	0.22	0.33	0.66	0.51

Kids whose moms haven't gone to HS, did not work during the first three years of the kid's life, have an IQ of 0 and are 0 yrs old are expected on average to score 19.59. Obviously, the intercept does not make any sense in context.

Interpreting the slope

What is the correct interpretation of the slope for `mom_work`?

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	19.59	9.22	2.13	0.03
mom_hs:yes	5.09	2.31	2.20	0.03
mom_iq	0.56	0.06	9.26	0.00
mom_work:yes	2.54	2.35	1.08	0.28
mom_age	0.22	0.33	0.66	0.51

All else being equal, kids whose moms worked during the first three year's of the kid's life

(a) are estimated to score 2.54 points lower

(b) are estimated to score 2.54 points higher

than those whose moms did not work.

Interpreting the slope

What is the correct interpretation of the slope for `mom_work`?

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	19.59	9.22	2.13	0.03
mom_hs:yes	5.09	2.31	2.20	0.03
mom_iq	0.56	0.06	9.26	0.00
mom_work:yes	2.54	2.35	1.08	0.28
mom_age	0.22	0.33	0.66	0.51

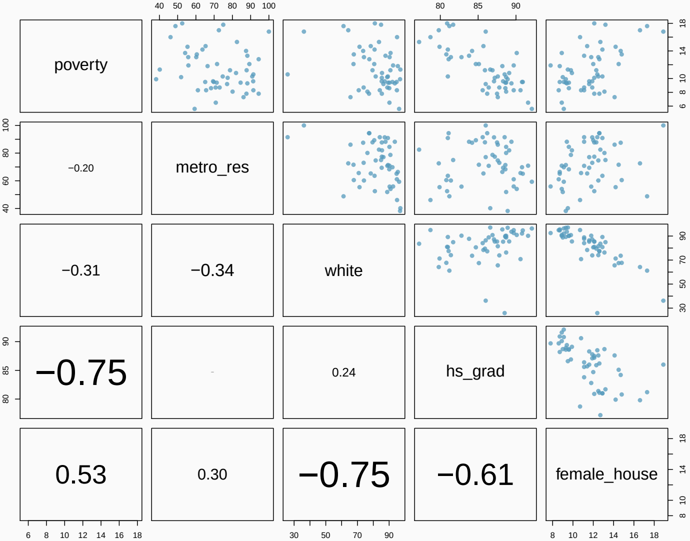
All else being equal, kids whose moms worked during the first three year's of the kid's life

(a) are estimated to score 2.54 points lower

(b) *are estimated to score 2.54 points higher*

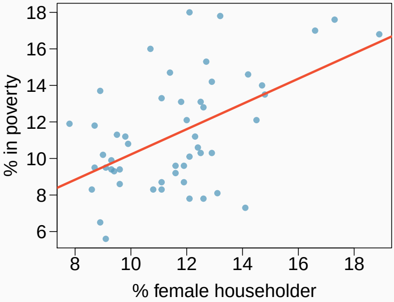
than those whose moms did not work.

Revisit: Modeling poverty



Predicting poverty using % female householder

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3.31	1.90	1.74	0.09
female_house	0.69	0.16	4.32	0.00



$$R = 0.53$$

$$R^2 = 0.53^2 = 0.28$$

Another look at R^2

R^2 can be calculated in three ways:

Another look at R^2

R^2 can be calculated in three ways:

1. square the correlation coefficient of x and y (how we have been calculating it)

Another look at R^2

R^2 can be calculated in three ways:

1. square the correlation coefficient of x and y (how we have been calculating it)
2. square the correlation coefficient of y and $\hat{y}$

Another look at R^2

R^2 can be calculated in three ways:

1. square the correlation coefficient of x and y (how we have been calculating it)
2. square the correlation coefficient of y and $\hat{y}$
3. based on definition:

$$R^2 = \frac{\text{explained variability in } y}{\text{total variability in } y}$$

Another look at R^2

R^2 can be calculated in three ways:

1. square the correlation coefficient of x and y (how we have been calculating it)
2. square the correlation coefficient of y and $\hat{y}$
3. based on definition:

$$R^2 = \frac{\text{explained variability in } y}{\text{total variability in } y}$$

Using *ANOVA* we can calculate the explained variability and total variability in y.

Sum of squares

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
female_house	1	132.57	132.57	18.68	0.00
Residuals	49	347.68	7.10		
Total	50	480.25			

Sum of squares

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
female_house	1	132.57	132.57	18.68	0.00
Residuals	49	347.68	7.10		
Total	50	480.25			

Sum of squares of y: $SS_{Total} \quad = \quad \sum (y - \bar{y})^2 = 480.25$ → *total variability*

Sum of squares

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
female_house	1	132.57	132.57	18.68	0.00
Residuals	49	347.68	7.10		
Total	50	480.25			

Sum of squares of y: $SS_{Total} = \sum (y - \bar{y})^2 = 480.25$ → *total variability*

Sum of squares of residuals: $SS_{Error} = \sum e_i^2 = 347.68$ → *unexplained variability*

Sum of squares

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
female_house	1	132.57	132.57	18.68	0.00
Residuals	49	347.68	7.10		
Total	50	480.25			

Sum of squares of y: $SS_{Total} = \sum (y - \bar{y})^2 = 480.25$ → *total variability*

Sum of squares of residuals: $SS_{Error} = \sum e_i^2 = 347.68$ → *unexplained variability*

Sum of squares of x: $SS_{Model} = SS_{Total} - SS_{Error}$ → *explained variability*

$= 480.25 - 347.68 = 132.57$

Sum of squares

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
female_house	1	132.57	132.57	18.68	0.00
Residuals	49	347.68	7.10		
Total	50	480.25			

$$\begin{aligned}
\text{Sum of squares of } y\text{: } SS_{Total} &= \sum (y - \bar{y})^2 = 480.25 \rightarrow \textit{total variability} \\
\text{Sum of squares of residuals: } SS_{Error} &= \sum e_i^2 = 347.68 \rightarrow \textit{unexplained variability} \\
\text{Sum of squares of } x\text{: } SS_{Model} &= SS_{Total} - SS_{Error} \rightarrow \textit{explained variability} \\
&= 480.25 - 347.68 = 132.57
\end{aligned}$$

$$R^2 = \frac{\text{explained variability}}{\text{total variability}} = \frac{132.57}{480.25} = 0.28 \checkmark$$

Why bother?

Why bother with another approach for calculating R^2 when we had a perfectly good way to calculate it as the correlation coefficient squared?

Why bother?

Why bother with another approach for calculating R^2 when we had a perfectly good way to calculate it as the correlation coefficient squared?

- *For single-predictor linear regression, having three ways to calculate the same value may seem like overkill.*
- *However, in multiple linear regression, we can't calculate R^2 as the square of the correlation between x and y because we have multiple xs.*
- *And next we'll learn another measure of explained variability, adjusted R^2, that requires the use of the third approach, ratio of explained and unexplained variability.*

Predicting poverty using % female hh + % white

Linear model:	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-2.58	5.78	-0.45	0.66
female_house	0.89	0.24	3.67	0.00
white	0.04	0.04	1.08	0.29

ANOVA:	Df	Sum Sq	Mean Sq	F value	Pr(>F)
female_house	1	132.57	132.57	18.74	0.00
white	1	8.21	8.21	1.16	0.29
Residuals	48	339.47	7.07		
Total	50	480.25			

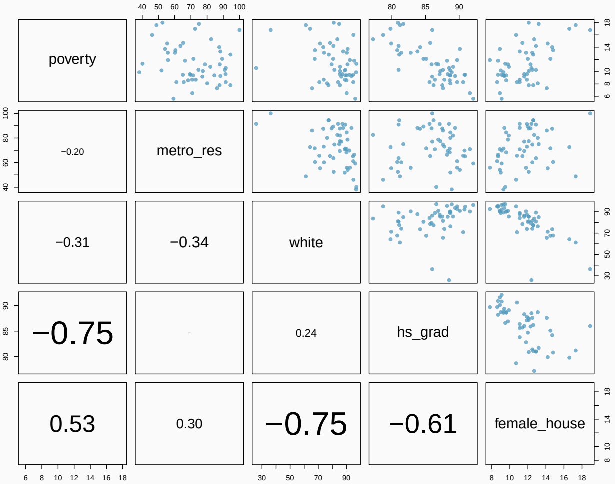
Predicting poverty using % female hh + % white

Linear model:	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-2.58	5.78	-0.45	0.66
female_house	0.89	0.24	3.67	0.00
white	0.04	0.04	1.08	0.29

ANOVA:	Df	Sum Sq	Mean Sq	F value	Pr(>F)
female_house	1	132.57	132.57	18.74	0.00
white	1	8.21	8.21	1.16	0.29
Residuals	48	339.47	7.07		
Total	50	480.25			

$$R^2 = \frac{\text{explained variability}}{\text{total variability}} = \frac{132.57 + 8.21}{480.25} = 0.29$$

Does adding the variable `white` to the model add valuable information that wasn't provided by `female_house`?



Collinearity between explanatory variables

poverty vs. % female head of household

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3.31	1.90	1.74	0.09
female_house	0.69	0.16	4.32	0.00

poverty vs. % female head of household and % female hh

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-2.58	5.78	-0.45	0.66
female_house	0.89	0.24	3.67	0.00
white	0.04	0.04	1.08	0.29

Collinearity between explanatory variables

poverty vs. % female head of household

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3.31	1.90	1.74	0.09
female_house	*0.69*	0.16	4.32	0.00

poverty vs. % female head of household and % female hh

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-2.58	5.78	-0.45	0.66
female_house	*0.89*	0.24	3.67	0.00
white	0.04	0.04	1.08	0.29

Collinearity between explanatory variables (cont.)

- Two predictor variables are said to be collinear when they are correlated, and this *collinearity* complicates model estimation.

  Remember: Predictors are also called explanatory or independent variables. Ideally, they would be independent of each other.

Collinearity between explanatory variables (cont.)

- Two predictor variables are said to be collinear when they are correlated, and this *collinearity* complicates model estimation.

  Remember: Predictors are also called explanatory or <u>independent</u> variables. Ideally, they would be independent of each other.

- We don't like adding predictors that are associated with each other to the model, because often times the addition of such variable brings nothing to the table. Instead, we prefer the simplest best model, i.e. *parsimonious* model.

Collinearity between explanatory variables (cont.)

- Two predictor variables are said to be collinear when they are correlated, and this *collinearity* complicates model estimation.

  Remember: Predictors are also called explanatory or <u>independent</u> variables. Ideally, they would be independent of each other.

- We don't like adding predictors that are associated with each other to the model, because often times the addition of such variable brings nothing to the table. Instead, we prefer the simplest best model, i.e. *parsimonious* model.
- While it's impossible to avoid collinearity from arising in observational data, experiments are usually designed to prevent correlation among predictors.

R^2 vs. adjusted R^2

	R^2	Adjusted R^2
Model 1 (Single-predictor)	0.28	0.26
Model 2 (Multiple)	0.29	0.26

R^2 vs. adjusted R^2

	R^2	Adjusted R^2
Model 1 (Single-predictor)	0.28	0.26
Model 2 (Multiple)	0.29	0.26

- When <u>any</u> variable is added to the model R^2 increases.

R^2 vs. adjusted R^2

	R^2	Adjusted R^2
Model 1 (Single-predictor)	0.28	0.26
Model 2 (Multiple)	0.29	0.26

- When <u>any</u> variable is added to the model R^2 increases.
- But if the added variable doesn't really provide any new information, or is completely unrelated, adjusted R^2 does not increase.

Adjusted R^2

Adjusted R^2

$$R^2_{adj} = 1 - \left(\frac{SS_{Error}}{SS_{Total}} \times \frac{n-1}{n-p-1} \right)$$

where n is the number of cases and p is the number of predictors (explanatory variables) in the model.

- Because p is never negative, R^2_{adj} will always be smaller than R^2.
- R^2_{adj} applies a penalty for the number of predictors included in the model.
- Therefore, we choose models with higher R^2_{adj} over others.

Calculate adjusted R^2

ANOVA:	Df	Sum Sq	Mean Sq	F value	Pr(>F)
female_house	1	132.57	132.57	18.74	0.0001
white	1	8.21	8.21	1.16	0.2868
Residuals	48	339.47	7.07		
Total	50	480.25			

$$R^2_{adj} = 1 - \left(\frac{SS_{Error}}{SS_{Total}} \times \frac{n-1}{n-p-1} \right)$$

Calculate adjusted R^2

ANOVA:	Df	Sum Sq	Mean Sq	F value	Pr(>F)
female_house	1	132.57	132.57	18.74	0.0001
white	1	8.21	8.21	1.16	0.2868
Residuals	48	339.47	7.07		
Total	50	480.25			

$$
\begin{aligned}
R^2_{adj} &= 1 - \left(\frac{SS_{Error}}{SS_{Total}} \times \frac{n-1}{n-p-1} \right) \\
&= 1 - \left(\frac{339.47}{480.25} \times \frac{51-1}{51-2-1} \right)
\end{aligned}
$$

Calculate adjusted R^2

ANOVA:	Df	Sum Sq	Mean Sq	F value	Pr(>F)
female_house	1	132.57	132.57	18.74	0.0001
white	1	8.21	8.21	1.16	0.2868
Residuals	48	339.47	7.07		
Total	50	480.25			

$$\begin{aligned}
R^2_{adj} &= 1 - \left(\frac{SS_{Error}}{SS_{Total}} \times \frac{n-1}{n-p-1}\right) \\
&= 1 - \left(\frac{339.47}{480.25} \times \frac{51-1}{51-2-1}\right) \\
&= 1 - \left(\frac{339.47}{480.25} \times \frac{50}{48}\right)
\end{aligned}$$

Calculate adjusted R^2

ANOVA:	Df	Sum Sq	Mean Sq	F value	Pr(>F)
female_house	1	132.57	132.57	18.74	0.0001
white	1	8.21	8.21	1.16	0.2868
Residuals	48	339.47	7.07		
Total	50	480.25			

$$
\begin{aligned}
R^2_{adj} &= 1 - \left(\frac{SS_{Error}}{SS_{Total}} \times \frac{n-1}{n-p-1} \right) \\
&= 1 - \left(\frac{339.47}{480.25} \times \frac{51-1}{51-2-1} \right) \\
&= 1 - \left(\frac{339.47}{480.25} \times \frac{50}{48} \right) \\
&= 1 - 0.74
\end{aligned}
$$

Calculate adjusted R^2

ANOVA:	Df	Sum Sq	Mean Sq	F value	Pr(>F)
female_house	1	132.57	132.57	18.74	0.0001
white	1	8.21	8.21	1.16	0.2868
Residuals	48	339.47	7.07		
Total	50	480.25			

$$
\begin{aligned}
R^2_{adj} &= 1 - \left(\frac{SS_{Error}}{SS_{Total}} \times \frac{n-1}{n-p-1} \right) \\
&= 1 - \left(\frac{339.47}{480.25} \times \frac{51-1}{51-2-1} \right) \\
&= 1 - \left(\frac{339.47}{480.25} \times \frac{50}{48} \right) \\
&= 1 - 0.74 \\
&= 0.26
\end{aligned}
$$

Model selection

Beauty in the classroom

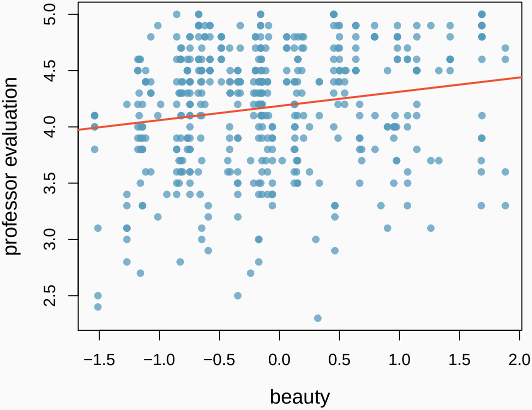
- Data: Student evaluations of instructors’ beauty and teaching quality for 463 courses at the University of Texas.
- Evaluations conducted at the end of semester, and the beauty judgements were made later, by six students who had not attended the classes and were not aware of the course evaluations (2 upper level females, 2 upper level males, one lower level female, one lower level male).

Hamermesh & Parker. (2004)“Beauty in the classroom: instructorsÕ pulchritude and putative pedagogical productivityÓ

Economics Education Review.

Professor rating vs. beauty

Professor evaluation score (higher score means better) vs. beauty score (a score of 0 means average, negative score means below average, and a positive score above average):



Which of the below is correct based on the model output?

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	4.19	0.03	167.24	0.00
beauty	0.13	0.03	4.00	0.00

R^2 = 0.0336

(a) Model predicts 3.36% of professor ratings correctly.

(b) Beauty is not a significant predictor of professor evaluation.

(c) Professors who score 1 point above average in their beauty score are tend to also score 0.13 points higher in their evaluation.

(d) 3.36% of variability in beauty scores can be explained by professor evaluation.

(e) The correlation coefficient could be $\sqrt{0.0336} = 0.18$ or -0.18, we can't tell which is correct.

Which of the below is correct based on the model output?

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	4.19	0.03	167.24	0.00
beauty	0.13	0.03	4.00	0.00

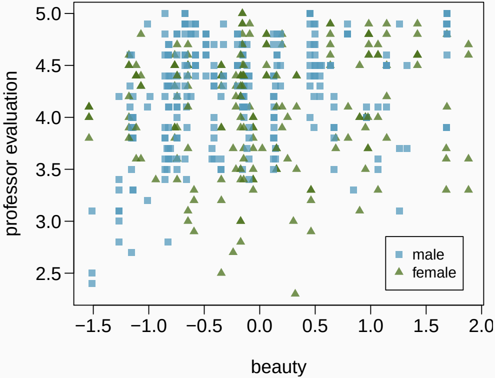
$R^2 = 0.0336$

(a) Model predicts 3.36% of professor ratings correctly.
(b) Beauty is not a significant predictor of professor evaluation.
(c) *Professors who score 1 point above average in their beauty score are tend to also score 0.13 points higher in their evaluation.*
(d) 3.36% of variability in beauty scores can be explained by professor evaluation.
(e) The correlation coefficient could be $\sqrt{0.0336} = 0.18$ or -0.18, we can't tell which is correct.

Exploratory analysis

Any interesting features?

For a given beauty score, are male professors evaluated higher, lower, or about the same as female professors?

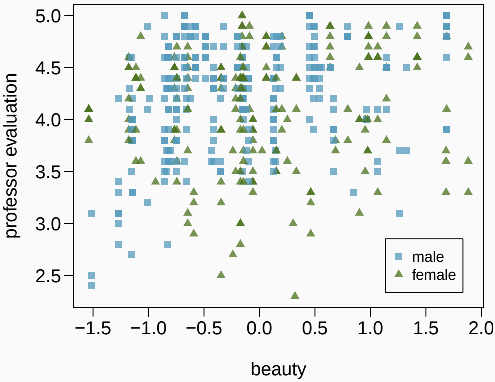


Exploratory analysis

Any interesting features?

Few females with very low beauty scores.

For a given beauty score, are male professors evaluated higher, lower, or about the same as female professors?



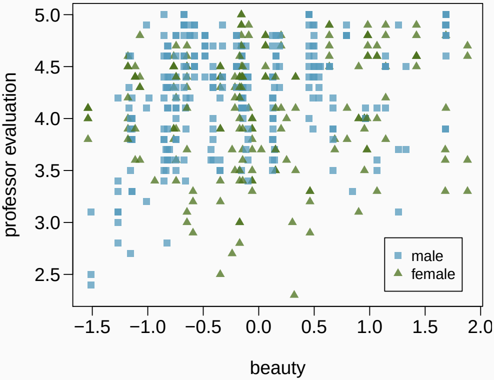
Exploratory analysis

Any interesting features?

Few females with very low beauty scores.

For a given beauty score, are male professors evaluated higher, lower, or about the same as female professors?

Difficult to tell from this plot only.



Professor rating vs. beauty + gender

For a given beauty score, are male professors evaluated higher, lower, or about the same as female professors?

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	4.09	0.04	107.85	0.00
beauty	0.14	0.03	4.44	0.00
gender.male	0.17	0.05	3.38	0.00
$R^2_{adj} = 0.057$				

(a) higher

(b) lower

(c) about the same

Professor rating vs. beauty + gender

For a given beauty score, are male professors evaluated higher, lower, or about the same as female professors?

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	4.09	0.04	107.85	0.00
beauty	0.14	0.03	4.44	0.00
gender.male	0.17	0.05	3.38	0.00
$R^2_{adj} = 0.057$				

(a) *higher $\rightarrow$ Beauty held constant, male professors are rated 0.17 points higher on average than female professors.*

(b) lower

(c) about the same

Full model

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	4.6282	0.1720	26.90	0.00
beauty	0.1080	0.0329	3.28	0.00
gender.male	0.2040	0.0528	3.87	0.00
age	-0.0089	0.0032	-2.75	0.01
formal.yes [1]	0.1511	0.0749	2.02	0.04
lower.yes [2]	0.0582	0.0553	1.05	0.29
native.non english	-0.2158	0.1147	-1.88	0.06
minority.yes	-0.0707	0.0763	-0.93	0.35
students [3]	-0.0004	0.0004	-1.03	0.30
tenure.tenure track [4]	-0.1933	0.0847	-2.28	0.02
tenure.tenured	-0.1574	0.0656	-2.40	0.02

[1] `formal`: picture wearing tie&jacket/blouse, levels: `yes`, `no`

[2] `lower`: lower division course, levels: `yes`, `no`

[3] `students`: number of students

[4] `tenure`: tenure status, levels: `non-tenure track`, `tenure track`, `tenured`

Hypotheses

Just as the interpretation of the slope parameters take into account all other variables in the model, the hypotheses for testing for significance of a predictor also takes into account all other variables.

$H_0 : B_i = 0$ when other explanatory variables are included in the model.

$H_A : B_i \neq 0$ when other explanatory variables are included in the model.

Assessing significance: numerical variables

The p-value for age is 0.01. What does this indicate?

	Estimate	Std. Error	t value	Pr(>|t|)
...				
age	-0.0089	0.0032	-2.75	0.01
...				

(a) Since p-value is positive, higher the professor's age, the higher we would expect them to be rated.

(b) If we keep all other variables in the model, there is strong evidence that professor's age is associated with their rating.

(c) Probability that the true slope parameter for age is 0 is 0.01.

(d) There is about 1% chance that the true slope parameter for age is -0.0089.

Assessing significance: numerical variables

The p-value for age is 0.01. What does this indicate?

	Estimate	Std. Error	t value	Pr(>|t|)
...				
age	-0.0089	0.0032	-2.75	0.01
...				

(a) Since p-value is positive, higher the professor's age, the higher we would expect them to be rated.

(b) *If we keep all other variables in the model, there is strong evidence that professor's age is associated with their rating.*

(c) Probability that the true slope parameter for age is 0 is 0.01.

(d) There is about 1% chance that the true slope parameter for age is -0.0089.

Assessing significance: categorical variables

Tenure is a categorical variable with 3 levels: non tenure track, tenure track, tenured. Based on the model output given, which of the below is false?

	Estimate	Std. Error	t value	Pr(>|t|)
...				
tenure.tenure track	-0.1933	0.0847	-2.28	0.02
tenure.tenured	-0.1574	0.0656	-2.40	0.02

(a) Reference level is non tenure track.

(b) All else being equal, tenure track professors are rated, on average, 0.19 points lower than non-tenure track professors.

(c) All else being equal, tenured professors are rated, on average, 0.16 points lower than non-tenure track professors.

(d) All else being equal, there is a significant difference between the average ratings of tenure track and tenured professors.

Assessing significance: categorical variables

Tenure is a categorical variable with 3 levels: non tenure track, tenure track, tenured. Based on the model output given, which of the below is false?

	Estimate	Std. Error	t value	Pr(>|t|)
...				
tenure.tenure track	-0.1933	0.0847	-2.28	0.02
tenure.tenured	-0.1574	0.0656	-2.40	0.02

(a) Reference level is non tenure track.

(b) All else being equal, tenure track professors are rated, on average, 0.19 points lower than non-tenure track professors.

(c) All else being equal, tenured professors are rated, on average, 0.16 points lower than non-tenure track professors.

(d) *All else being equal, there is a significant difference between the average ratings of tenure track and tenured professors.*

Assessing significance

Which predictors do not seem to meaningfully contribute to the model, i.e. may not be significant predictors of professor's rating score?

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	4.6282	0.1720	26.90	0.00
beauty	0.1080	0.0329	3.28	0.00
gender.male	0.2040	0.0528	3.87	0.00
age	-0.0089	0.0032	-2.75	0.01
formal.yes	0.1511	0.0749	2.02	0.04
lower.yes	0.0582	0.0553	1.05	0.29
native.non english	-0.2158	0.1147	-1.88	0.06
minority.yes	-0.0707	0.0763	-0.93	0.35
students	-0.0004	0.0004	-1.03	0.30
tenure.tenure track	-0.1933	0.0847	-2.28	0.02
tenure.tenured	-0.1574	0.0656	-2.40	0.02

Model selection strategies

Based on what we've learned so far, what are some ways you can think of that can be used to determine which variables to keep in the model and which to leave out?

Backward-elimination

1. R^2_{adj} approach:
   - Start with the full model
   - Drop one variable at a time and record R^2_{adj} of each smaller model
   - Pick the model with the highest increase in R^2_{adj}
   - Repeat until none of the models yield an increase in R^2_{adj}
2. p-value approach:
   - Start with the full model
   - Drop the variable with the highest p-value and refit a smaller model
   - Repeat until all variables left in the model are significant

Backward-elimination: R^2_{adj} approach

Step	Variables included	R^2_{adj}
Full	beauty + gender + age + formal + lower + native + minority + students + tenure	*0.0839*

Backward-elimination: R^2_{adj} approach

Step	Variables included	R^2_{adj}
Full	beauty + gender + age + formal + lower + native + minority + students + tenure	*0.0839*
Step 1	gender + age + formal + lower + native + minority + students + tenure	0.0642
	beauty + age + formal + lower + native + minority + students + tenure	0.0557
	beauty + gender + formal + lower + native + minority + students + tenure	0.0706
	beauty + gender + age + lower + native + minority + students + tenure	0.0777
	beauty + gender + age + formal + native + minority + students + tenure	0.0837
	beauty + gender + age + formal + lower + minority + students + tenure	0.0788
	beauty + gender + age + formal + lower + native + students + tenure	*0.0842*
	beauty + gender + age + formal + lower + native + minority + tenure	0.0838
	beauty + gender + age + formal + lower + native + minority + students	0.0733

Backward-elimination: R^2_{adj} approach

Step	Variables included	R^2_{adj}
Full	beauty + gender + age + formal + lower + native + minority + students + tenure	*0.0839*
Step 1	gender + age + formal + lower + native + minority + students + tenure	0.0642
	beauty + age + formal + lower + native + minority + students + tenure	0.0557
	beauty + gender + formal + lower + native + minority + students + tenure	0.0706
	beauty + gender + age + lower + native + minority + students + tenure	0.0777
	beauty + gender + age + formal + native + minority + students + tenure	0.0837
	beauty + gender + age + formal + lower + minority + students + tenure	0.0788
	beauty + gender + age + formal + lower + native + students + tenure	*0.0842*
	beauty + gender + age + formal + lower + native + minority + tenure	0.0838
	beauty + gender + age + formal + lower + native + minority + students	0.0733
Step 2	gender + age + formal + lower + native + students + tenure	0.0647
	beauty + age + formal + lower + native + students + tenure	0.0543
	beauty + gender + formal + lower + native + students + tenure	0.0708
	beauty + gender + age + lower + native + students + tenure	0.0776
	beauty + gender + age + formal + native + students + tenure	*0.0846*
	beauty + gender + age + formal + lower + native + tenure	0.0844
	beauty + gender + age + formal + lower + native + students	0.0725

Backward-elimination: R^2_{adj} approach

Step	Variables included	R^2_{adj}
Full	beauty + gender + age + formal + lower + native + minority + students + tenure	*0.0839*
Step 1	gender + age + formal + lower + native + minority + students + tenure	0.0642
	beauty + age + formal + lower + native + minority + students + tenure	0.0557
	beauty + gender + formal + lower + native + minority + students + tenure	0.0706
	beauty + gender + age + lower + native + minority + students + tenure	0.0777
	beauty + gender + age + formal + native + minority + students + tenure	0.0837
	beauty + gender + age + formal + lower + minority + students + tenure	0.0788
	beauty + gender + age + formal + lower + native + students + tenure	*0.0842*
	beauty + gender + age + formal + lower + native + minority + tenure	0.0838
	beauty + gender + age + formal + lower + native + minority + students	0.0733
Step 2	gender + age + formal + lower + native + students + tenure	0.0647
	beauty + age + formal + lower + native + students + tenure	0.0543
	beauty + gender + formal + lower + native + students + tenure	0.0708
	beauty + gender + age + lower + native + students + tenure	0.0776
	beauty + gender + age + formal + native + students + tenure	*0.0846*
	beauty + gender + age + formal + lower + native + tenure	0.0844
	beauty + gender + age + formal + lower + native + students	0.0725
Step 3	gender + age + formal + native + students + tenure	0.0653
	beauty + age + formal + native + students + tenure	0.0534
	beauty + gender + formal + native + students + tenure	0.0707
	beauty + gender + age + native + students + tenure	0.0786
	beauty + gender + age + formal + students + tenure	0.0756
	beauty + gender + age + formal + native + tenure	*0.0855*
	beauty + gender + age + formal + native + students	0.0713

Backward-elimination: R^2_{adj} approach

Step	Variables included	R^2_{adj}
Full	beauty + gender + age + formal + lower + native + minority + students + tenure	*0.0839*
Step 1	gender + age + formal + lower + native + minority + students + tenure	0.0642
	beauty + age + formal + lower + native + minority + students + tenure	0.0557
	beauty + gender + formal + lower + native + minority + students + tenure	0.0706
	beauty + gender + age + lower + native + minority + students + tenure	0.0777
	beauty + gender + age + formal + native + minority + students + tenure	0.0837
	beauty + gender + age + formal + lower + minority + students + tenure	0.0788
	beauty + gender + age + formal + lower + native + students + tenure	*0.0842*
	beauty + gender + age + formal + lower + native + minority + tenure	0.0838
	beauty + gender + age + formal + lower + native + minority + students	0.0733
Step 2	gender + age + formal + lower + native + students + tenure	0.0647
	beauty + age + formal + lower + native + students + tenure	0.0543
	beauty + gender + formal + lower + native + students + tenure	0.0708
	beauty + gender + age + lower + native + students + tenure	0.0776
	beauty + gender + age + formal + native + students + tenure	*0.0846*
	beauty + gender + age + formal + lower + native + tenure	0.0844
	beauty + gender + age + formal + lower + native + students	0.0725
Step 3	gender + age + formal + native + students + tenure	0.0653
	beauty + age + formal + native + students + tenure	0.0534
	beauty + gender + formal + native + students + tenure	0.0707
	beauty + gender + age + native + students + tenure	0.0786
	beauty + gender + age + formal + students + tenure	0.0756
	beauty + gender + age + formal + native + tenure	*0.0855*
	beauty + gender + age + formal + native + students	0.0713
Step 4	gender + age + formal + native + tenure	0.0667
	beauty + age + formal + native + tenure	0.0553
	beauty + gender + formal + native + tenure	0.0723
	beauty + gender + age + native + tenure	0.0806

```
Call:
lm(formula = profevaluation ~ beauty + gender + age + formal +
    native + tenure, data = d)

Coefficients:
       (Intercept)              beauty          gendermale
          4.628435            0.105546            0.208079
               age           formalyes   nativenon english
         -0.008844            0.132422           -0.243003
tenuretenure track      tenuretenured
         -0.206784           -0.175967
```

step function in R

```
Call:
lm(formula = profevaluation ~ beauty + gender + age + formal +
    native + tenure, data = d)

Coefficients:
       (Intercept)               beauty           gendermale
          4.628435             0.105546             0.208079
               age            formalyes   nativenon english
         -0.008844             0.132422            -0.243003
tenuretenure track      tenuretenured
         -0.206784            -0.175967
```

Best model: beauty + gender + age + formal + native + tenure

Backward-elimination: $p - value$ approach

Step	Variables included & p-value									
Full	beauty	gender male	age	formal yes	lower yes	native non english	minority yes	students	tenure tenure track	tenure tenured
	0.00	0.00	0.01	0.04	0.29	0.06	*0.35*	0.30	0.02	0.02

Backward-elimination: $p - value$ approach

Step	Variables included & p-value									
Full	beauty	gender male	age	formal yes	lower yes	native non english	minority yes	students	tenure tenure track	tenure tenured
	0.00	0.00	0.01	0.04	0.29	0.06	*0.35*	0.30	0.02	0.02
Step 1	beauty	gender male	age	formal yes	lower yes	native non english		students	tenure tenure track	tenure tenured
	0.00	0.00	0.01	0.04	*0.38*	0.03		0.34	0.02	0.01

Backward-elimination: $p - value$ approach

Step	Variables included & p-value									
Full	beauty	gender	age	formal	lower	native	minority	students	tenure	tenure
		male		yes	yes	non english	yes		tenure track	tenured
	0.00	0.00	0.01	0.04	0.29	0.06	*0.35*	0.30	0.02	0.02
Step 1	beauty	gender	age	formal	lower	native		students	tenure	tenure
		male		yes	yes	non english			tenure track	tenured
	0.00	0.00	0.01	0.04	*0.38*	0.03		0.34	0.02	0.01
Step 2	beauty	gender	age	formal		native		students	tenure	tenure
		male		yes		non english			tenure track	tenured
	0.00	0.00	0.01	0.05		0.02		*0.44*	0.01	0.01

Backward-elimination: $p - value$ approach

Step	Variables included & p-value									
Full	beauty	gender	age	formal	lower	native	minority	students	tenure	tenure
		male		yes	yes	non english	yes		tenure track	tenured
	0.00	0.00	0.01	0.04	0.29	0.06	*0.35*	0.30	0.02	0.02
Step 1	beauty	gender	age	formal	lower	native		students	tenure	tenure
		male		yes	yes	non english			tenure track	tenured
	0.00	0.00	0.01	0.04	*0.38*	0.03		0.34	0.02	0.01
Step 2	beauty	gender	age	formal		native		students	tenure	tenure
		male		yes		non english			tenure track	tenured
	0.00	0.00	0.01	0.05		0.02		*0.44*	0.01	0.01
Step 3	beauty	gender	age	formal		native			tenure	tenure
		male		yes		non english			tenure track	tenured
	0.00	0.00	0.01	*0.06*		0.02			0.01	0.01

Backward-elimination: $p - value$ approach

Step	Variables included & p-value									
Full	beauty	gender	age	formal	lower	native	minority	students	tenure	tenure
		male		yes	yes	non english	yes		tenure track	tenured
	0.00	0.00	0.01	0.04	0.29	0.06	*0.35*	0.30	0.02	0.02
Step 1	beauty	gender	age	formal	lower	native		students	tenure	tenure
		male		yes	yes	non english			tenure track	tenured
	0.00	0.00	0.01	0.04	*0.38*	0.03		0.34	0.02	0.01
Step 2	beauty	gender	age	formal		native		students	tenure	tenure
		male		yes		non english			tenure track	tenured
	0.00	0.00	0.01	0.05		0.02		*0.44*	0.01	0.01
Step 3	beauty	gender	age	formal		native			tenure	tenure
		male		yes		non english			tenure track	tenured
	0.00	0.00	0.01	*0.06*		0.02			0.01	0.01
Step 4	beauty	gender	age			native			tenure	tenure
		male				non english			tenure track	tenured
	0.00	0.00	0.01			*0.06*			0.01	0.01

Backward-elimination: $p - value$ approach

Step	**Variables included & p-value**									
Full	beauty	gender	age	formal	lower	native	minority	students	tenure	tenure
		male		yes	yes	non english	yes		tenure track	tenured
	0.00	0.00	0.01	0.04	0.29	0.06	*0.35*	0.30	0.02	0.02
Step 1	beauty	gender	age	formal	lower	native		students	tenure	tenure
		male		yes	yes	non english			tenure track	tenured
	0.00	0.00	0.01	0.04	*0.38*	0.03		0.34	0.02	0.01
Step 2	beauty	gender	age	formal		native		students	tenure	tenure
		male		yes		non english			tenure track	tenured
	0.00	0.00	0.01	0.05		0.02		*0.44*	0.01	0.01
Step 3	beauty	gender	age	formal		native			tenure	tenure
		male		yes		non english			tenure track	tenured
	0.00	0.00	0.01	*0.06*		0.02			0.01	0.01
Step 4	beauty	gender	age			native			tenure	tenure
		male				non english			tenure track	tenured
	0.00	0.00	0.01			*0.06*			0.01	0.01
Step 5	beauty	gender	age						tenure	tenure
		male							tenure track	tenured
	0.00	0.00	0.01						0.01	0.01

Backward-elimination: $p - value$ approach

Step	Variables included & p-value									
Full	beauty	gender	age	formal	lower	native	minority	students	tenure	tenure
		male		yes	yes	non english	yes		tenure track	tenured
	0.00	0.00	0.01	0.04	0.29	0.06	*0.35*	0.30	0.02	0.02
Step 1	beauty	gender	age	formal	lower	native		students	tenure	tenure
		male		yes	yes	non english			tenure track	tenured
	0.00	0.00	0.01	0.04	*0.38*	0.03		0.34	0.02	0.01
Step 2	beauty	gender	age	formal		native		students	tenure	tenure
		male		yes		non english			tenure track	tenured
	0.00	0.00	0.01	0.05		0.02		*0.44*	0.01	0.01
Step 3	beauty	gender	age	formal		native			tenure	tenure
		male		yes		non english			tenure track	tenured
	0.00	0.00	0.01	*0.06*		0.02			0.01	0.01
Step 4	beauty	gender	age			native			tenure	tenure
		male				non english			tenure track	tenured
	0.00	0.00	0.01			*0.06*			0.01	0.01
Step 5	beauty	gender	age						tenure	tenure
		male							tenure track	tenured
	0.00	0.00	0.01						0.01	0.01

Best model: beauty + gender + age + tenure

Forward-selection

1. R^2_{adj} approach:
    - Start with regressions of response vs. each explanatory variable
    - Pick the model with the highest R^2_{adj}
    - Add the remaining variables one at a time to the existing model, and once again pick the model with the highest R^2_{adj}
    - Repeat until the addition of any of the remanning variables does not result in a higher R^2_{adj}
2. $p - value$ approach:
    - Start with regressions of response vs. each explanatory variable
    - Pick the variable with the lowest significant p-value
    - Add the remaining variables one at a time to the existing model, and pick the variable with the lowest significant p-value
    - Repeat until any of the remaining variables does not have a significant p-value

In forward-selection the p-value approach isn't any simpler

Selected model

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	4.6284	0.1673	27.66	0.00
beauty	0.1055	0.0328	3.21	0.00
gender.male	0.2081	0.0519	4.01	0.00
age	-0.0088	0.0032	-2.75	0.01
formal.yes	0.1324	0.0714	1.85	0.06
native:non english	-0.2430	0.1080	-2.25	0.02
tenure:tenure track	-0.2068	0.0839	-2.46	0.01
tenure:tenured	-0.1760	0.0641	-2.74	0.01

Checking model conditions using graphs

Modeling conditions

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

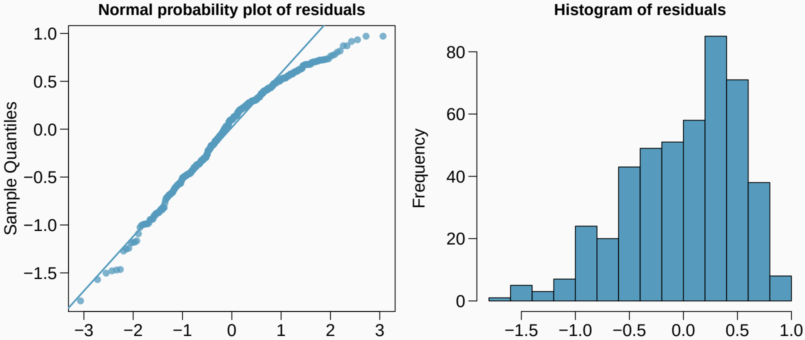
The model depends on the following conditions

1. residuals are nearly normal (primary concern relates to residuals that are outliers)
2. residuals have constant variability
3. residuals are independent
4. each variable is linearly related to the outcome

We often use graphical methods to check the validity of these conditions, which we will go through in detail in the following slides.

(1) nearly normal residuals

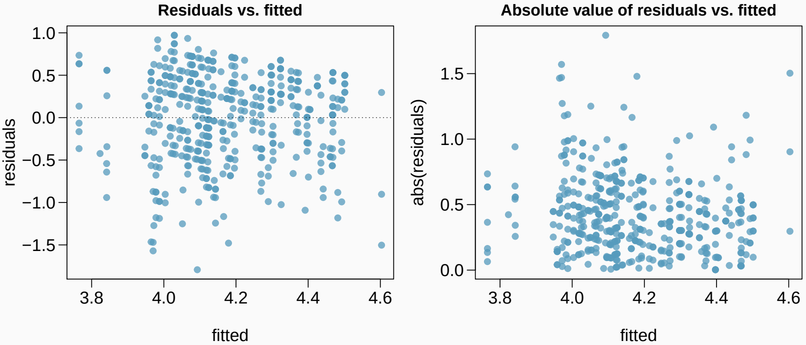
normal probability plot and/or histogram of residuals:



Does this condition appear to be satisfied?

(2) constant variability in residuals

scatterplot of residuals and/or absolute value of residuals vs. fitted (predicted):



Does this condition appear to be satisfied?

Checking constant variance - recap

- When we did simple linear regression (one explanatory variable) we checked the constant variance condition using a plot of *residuals vs. x*.
- With multiple linear regression (2+ explanatory variables) we checked the constant variance condition using a plot of *residuals vs. fitted*.

Why are we using different plots?

Checking constant variance - recap

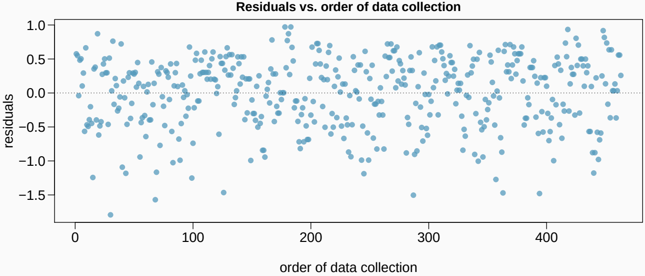
- When we did simple linear regression (one explanatory variable) we checked the constant variance condition using a plot of *residuals vs. x*.
- With multiple linear regression (2+ explanatory variables) we checked the constant variance condition using a plot of *residuals vs. fitted*.

Why are we using different plots?

In multiple linear regression there are many explanatory variables, so a plot of residuals vs. one of them wouldn't give us the complete picture.

(3) independent residuals

scatterplot of residuals vs. order of data collection:



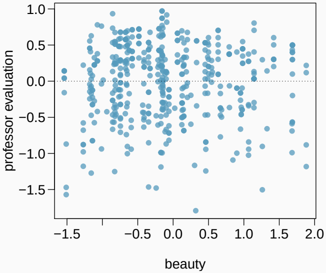
Does this condition appear to be satisfied?

More on the condition of independent residuals

- Checking for independent residuals allows us to indirectly check for independent observations.
- If observations and residuals are independent, we would not expect to see an increasing or decreasing trend in the scatterplot of residuals vs. order of data collection.
- This condition is often violated when we have time series data. Such data require more advanced time series regression techniques for proper analysis.

(4) linear relationships

scatterplot of residuals vs. each (numerical) explanatory variable:



Does this condition appear to be satisfied?

Note: We use residuals instead of the predictors on the y-axis so that we can still check for linearity without worrying about other possible violations like collinearity between the predictors.

Logistic regression

Regression so far ...

At this point we have covered:

- Simple linear regression
  - Relationship between numerical response and a numerical or categorical predictor

Regression so far ...

At this point we have covered:

- Simple linear regression
  - Relationship between numerical response and a numerical or categorical predictor
- Multiple regression
  - Relationship between numerical response and multiple numerical and/or categorical predictors

Regression so far ...

At this point we have covered:

- Simple linear regression
    - Relationship between numerical response and a numerical or categorical predictor
- Multiple regression
    - Relationship between numerical response and multiple numerical and/or categorical predictors

What we haven't seen is what to do when the predictors are weird (nonlinear, complicated dependence structure, etc.) or when the response is weird (categorical, count data, etc.)

Odds

Odds are another way of quantifying the probability of an event, commonly used in gambling (and logistic regression).

Odds

For some event E,

$$\text{odds}(E) = \frac{P(E)}{P(E^c)} = \frac{P(E)}{1 - P(E)}$$

Similarly, if we are told the odds of E are x to y then

$$\text{odds}(E) = \frac{x}{y} = \frac{x/(x+y)}{y/(x+y)}$$

which implies

$$P(E) = x/(x+y), \quad P(E^c) = y/(x+y)$$

Example - Donner Party

In 1846 the Donner and Reed families left Springfield, Illinois, for California by covered wagon. In July, the Donner Party, as it became known, reached Fort Bridger, Wyoming. There its leaders decided to attempt a new and untested rote to the Sacramento Valley. Having reached its full size of 87 people and 20 wagons, the party was delayed by a difficult crossing of the Wasatch Range and again in the crossing of the desert west of the Great Salt Lake. The group became stranded in the eastern Sierra Nevada mountains when the region was hit by heavy snows in late October. By the time the last survivor was rescued on April 21, 1847, 40 of the 87 members had died from famine and exposure to extreme cold.

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

Example - Donner Party - Data

	Age	Sex	Status
1	23.00	Male	Died
2	40.00	Female	Survived
3	40.00	Male	Survived
4	30.00	Male	Died
5	28.00	Male	Died
⋮	⋮	⋮	⋮
43	23.00	Male	Survived
44	24.00	Male	Died
45	25.00	Female	Survived

Example - Donner Party - EDA

Status vs. Gender:

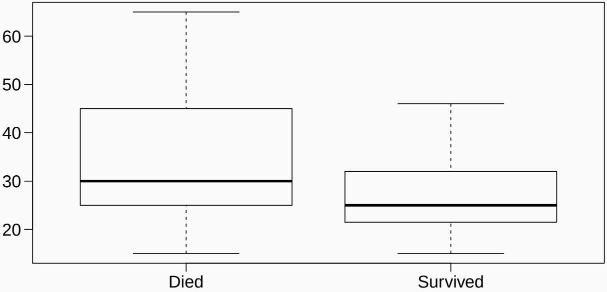
	Male	Female
Died	20	5
Survived	10	10

Example - Donner Party - EDA

Status vs. Gender:

	Male	Female
Died	20	5
Survived	10	10

Status vs. Age:



Example - Donner Party

It seems clear that both age and gender have an effect on someone's survival, how do we come up with a model that will let us explore this relationship?

Example - Donner Party

It seems clear that both age and gender have an effect on someone's survival, how do we come up with a model that will let us explore this relationship?

Even if we set Died to 0 and Survived to 1, this isn't something we can transform our way out of - we need something more.

Example - Donner Party

It seems clear that both age and gender have an effect on someone's survival, how do we come up with a model that will let us explore this relationship?

Even if we set Died to 0 and Survived to 1, this isn't something we can transform our way out of - we need something more.

One way to think about the problem - we can treat Survived and Died as successes and failures arising from a binomial distribution where the probability of a success is given by a transformation of a linear model of the predictors.

Generalized linear models

It turns out that this is a very general way of addressing this type of problem in regression, and the resulting models are called generalized linear models (GLMs). Logistic regression is just one example of this type of model.

Generalized linear models

It turns out that this is a very general way of addressing this type of problem in regression, and the resulting models are called generalized linear models (GLMs). Logistic regression is just one example of this type of model.

All generalized linear models have the following three characteristics:

1. A probability distribution describing the outcome variable
2. A linear model
    - $\eta = \beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n$
3. A link function that relates the linear model to the parameter of the outcome distribution
    - $g(p) = \eta$ or $p = g^{-1}(\eta)$

Logistic Regression

Logistic regression is a GLM used to model a binary categorical variable using numerical and categorical predictors.

We assume a binomial distribution produced the outcome variable and we therefore want to model p the probability of success for a given set of predictors.

Logistic Regression

Logistic regression is a GLM used to model a binary categorical variable using numerical and categorical predictors.

We assume a binomial distribution produced the outcome variable and we therefore want to model p the probability of success for a given set of predictors.

To finish specifying the Logistic model we just need to establish a reasonable link function that connects η to p. There are a variety of options but the most commonly used is the logit function.

Logit function

$$logit(p) = \log\left(\frac{p}{1-p}\right), \text{ for } 0 \le p \le 1$$

Properties of the Logit

The logit function takes a value between 0 and 1 and maps it to a value between $-\infty$ and ∞.

Inverse logit (logistic) function

$$g^{-1}(x) = \frac{\exp(x)}{1 + \exp(x)} = \frac{1}{1 + \exp(-x)}$$

The inverse logit function takes a value between $-\infty$ and ∞ and maps it to a value between 0 and 1.

This formulation also has some use when it comes to interpreting the model as logit can be interpreted as the log odds of a success, more on this later.

The logistic regression model

The three GLM criteria give us:

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

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

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

From which we arrive at,

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

Example - Donner Party - Model

In R we fit a GLM in the same was as a linear model except using `glm` instead of `lm` and we must also specify the type of GLM to fit using the `family` argument.

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

```
## Call:
## glm(formula = Status ~ Age, family = binomial, data = donner)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.81852    0.99937   1.820   0.0688 .
## Age         -0.06647    0.03222  -2.063   0.0391 *
##
##     Null deviance: 61.827  on 44  degrees of freedom
## Residual deviance: 56.291  on 43  degrees of freedom
## AIC: 60.291
##
```

Example - Donner Party - Prediction

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.8185	0.9994	1.82	0.0688
Age	-0.0665	0.0322	-2.06	0.0391

Model:

$$\log\left(\frac{p}{1-p}\right) = 1.8185 - 0.0665 \times \text{Age}$$

Example - Donner Party - Prediction

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.8185	0.9994	1.82	0.0688
Age	-0.0665	0.0322	-2.06	0.0391

Model:

$$\log\left(\frac{p}{1-p}\right) = 1.8185 - 0.0665 \times \text{Age}$$

Odds / Probability of survival for a newborn (Age=0):

Example - Donner Party - Prediction

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.8185	0.9994	1.82	0.0688
Age	-0.0665	0.0322	-2.06	0.0391

Model:

$$\log\left(\frac{p}{1-p}\right) = 1.8185 - 0.0665 \times \text{Age}$$

Odds / Probability of survival for a newborn (Age=0):

$$\log\left(\frac{p}{1-p}\right) = 1.8185 - 0.0665 \times 0$$

$$\frac{p}{1-p} = \exp(1.8185) = 6.16$$

$$p = 6.16/7.16 = 0.86$$

Example - Donner Party - Prediction (cont.)

Model:

$$\log\left(\frac{p}{1-p}\right) = 1.8185 - 0.0665 \times \text{Age}$$

Odds / Probability of survival for a 25 year old:

Example - Donner Party - Prediction (cont.)

Model:

$$\log\left(\frac{p}{1-p}\right) = 1.8185 - 0.0665 \times \text{Age}$$

Odds / Probability of survival for a 25 year old:

$$\log\left(\frac{p}{1-p}\right) = 1.8185 - 0.0665 \times 25$$

$$\frac{p}{1-p} = \exp(0.156) = 1.17$$

$$p = 1.17/2.17 = 0.539$$

Example - Donner Party - Prediction (cont.)

Model:

$$\log\left(\frac{p}{1-p}\right) = 1.8185 - 0.0665 \times \text{Age}$$

Odds / Probability of survival for a 25 year old:

$$\log\left(\frac{p}{1-p}\right) = 1.8185 - 0.0665 \times 25$$

$$\frac{p}{1-p} = \exp(0.156) = 1.17$$

$$p = 1.17/2.17 = 0.539$$

Odds / Probability of survival for a 50 year old:

Example - Donner Party - Prediction (cont.)

Model:

$$\log\left(\frac{p}{1-p}\right) = 1.8185 - 0.0665 \times \text{Age}$$

Odds / Probability of survival for a 25 year old:

$$\log\left(\frac{p}{1-p}\right) = 1.8185 - 0.0665 \times 25$$

$$\frac{p}{1-p} = \exp(0.156) = 1.17$$

$$p = 1.17/2.17 = 0.539$$

Odds / Probability of survival for a 50 year old:

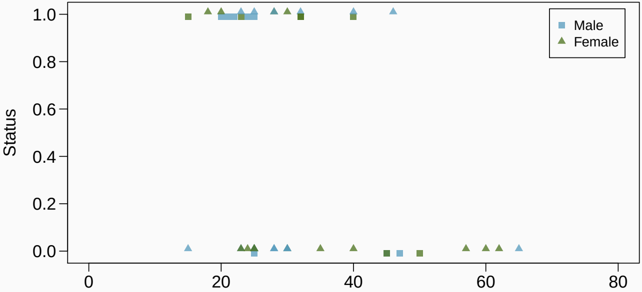
$$\log\left(\frac{p}{1-p}\right) = 1.8185 - 0.0665 \times 0$$

$$\frac{p}{1-p} = \exp(-1.5065) = 0.222$$

$$p = 0.222/1.222 = 0.181$$

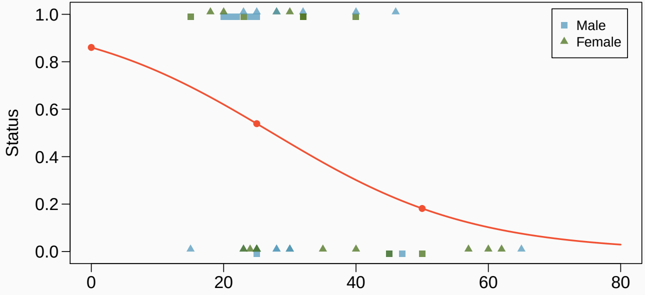
Example - Donner Party - Prediction (cont.)

$$\log\left(\frac{p}{1-p}\right) = 1.8185 - 0.0665 \times \text{Age}$$



Example - Donner Party - Prediction (cont.)

$$\log\left(\frac{p}{1-p}\right) = 1.8185 - 0.0665 \times \text{Age}$$



Example - Donner Party - Interpretation

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.8185	0.9994	1.82	0.0688
Age	-0.0665	0.0322	-2.06	0.0391

Simple interpretation is only possible in terms of log odds and log odds ratios for intercept and slope terms.

Intercept: The log odds of survival for a party member with an age of 0. From this we can calculate the odds or probability, but additional calculations are necessary.

Slope: For a unit increase in age (being 1 year older) how much will the log odds ratio change, not particularly intuitive. More often then not we care only about sign and relative magnitude.

Example - Donner Party - Interpretation - Slope

$$\begin{aligned}
\log\left(\frac{p_1}{1-p_1}\right) &= 1.8185 - 0.0665(x+1) \\
&= 1.8185 - 0.0665x - 0.0665 \\
\log\left(\frac{p_2}{1-p_2}\right) &= 1.8185 - 0.0665x
\end{aligned}$$

$$\begin{aligned}
\log\left(\frac{p_1}{1-p_1}\right) - \log\left(\frac{p_2}{1-p_2}\right) &= -0.0665 \\
\log\left(\frac{p_1}{1-p_1}\bigg/\frac{p_2}{1-p_2}\right) &= -0.0665 \\
\frac{p_1}{1-p_1}\bigg/\frac{p_2}{1-p_2} &= \exp(-0.0665) = 0.94
\end{aligned}$$

Example - Donner Party - Age and Gender

```
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
```


Gender slope: When the other predictors are held constant this is

Example - Donner Party - Gender Models

Just like MLR we can plug in gender to arrive at two status vs age models for men and women respectively.

General model:

$$\log\left(\frac{p_1}{1-p_1}\right) = 1.63312 + -0.07820 \times \text{Age} + 1.59729 \times \text{Sex}$$

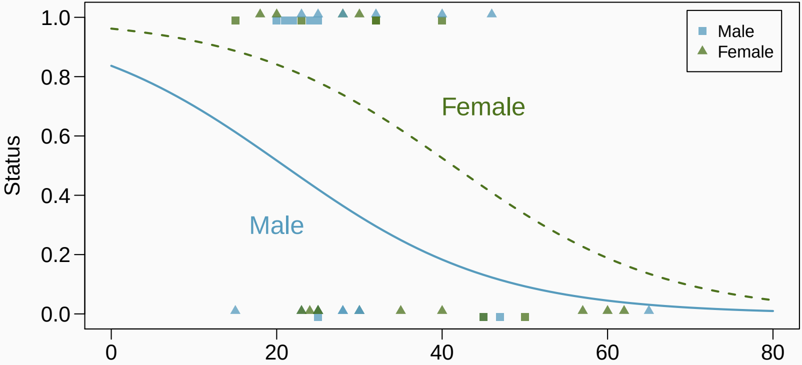
Male model:

$$\begin{aligned}\log\left(\frac{p_1}{1-p_1}\right) &= 1.63312 + -0.07820 \times \text{Age} + 1.59729 \times \mathit{0} \\ &= 1.63312 + -0.07820 \times \text{Age}\end{aligned}$$

Female model:

$$\begin{aligned}\log\left(\frac{p_1}{1-p_1}\right) &= 1.63312 + -0.07820 \times \text{Age} + 1.59729 \times \mathit{1} \\ &= 3.23041 + -0.07820 \times \text{Age}\end{aligned}$$

Example - Donner Party - Gender Models (cont.)



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
```

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
```

Note: The model output does not include any F-statistic, as a general rule there are not single model hypothesis tests for GLM models.

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 however 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: The only tricky bit, which is way beyond the scope of this course, is how the 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

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

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

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

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.

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.

Log odds ratio:

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

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.

Log odds ratio:

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

Odds ratio:

$$\exp(CI) = (\exp -0.1513, \exp -0.0051) = (0.85960.9949)$$

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.

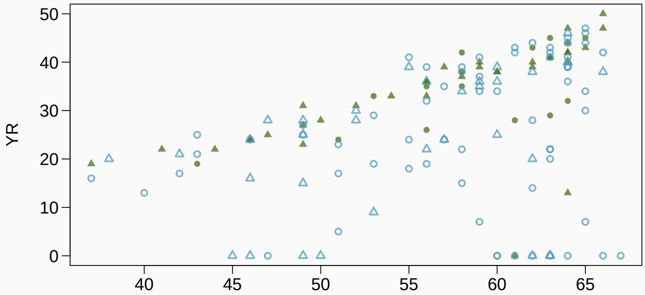
From *Ramsey, F.L. and Schafer, D.W. (2002). The Statistical Sleuth: A Course in Methods of Data Analysis (2nd ed)*

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

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,

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 $\exp(1.3626) = 3.91$.

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.

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 <u>not</u> 4x more likely to develop lung cancer than non-bird keepers.

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 <u>not</u> 4x more likely to develop lung cancer than non-bird keepers.

This is the difference between relative risk and an odds ratio.

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

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

Back to the birds

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

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

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

Back to the birds

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

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

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

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

Back to the birds

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

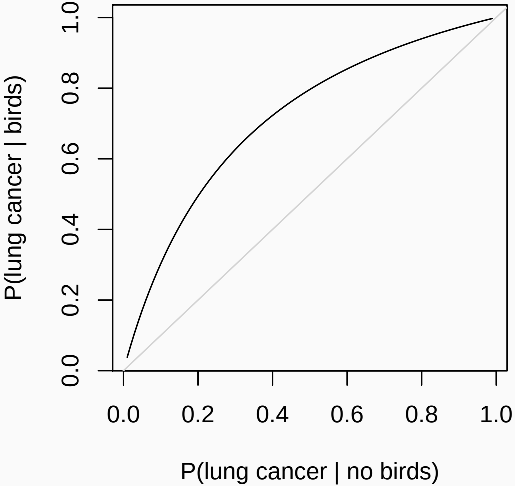
$$OR = \frac{P(\text{lung cancer}|\text{birds})/[1 - P(\text{lung cancer}|\text{birds})]}{P(\text{lung cancer}|\text{no birds})/[1 - P(\text{lung cancer}|\text{no birds})]}$$

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

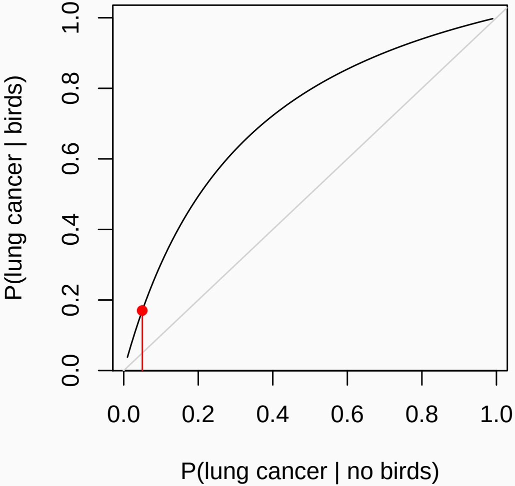
$$P(\text{lung cancer}|\text{birds}) = \frac{3.91 \times \frac{0.05}{0.95}}{1 + 3.91 \times \frac{0.05}{0.95}} = 0.171$$

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

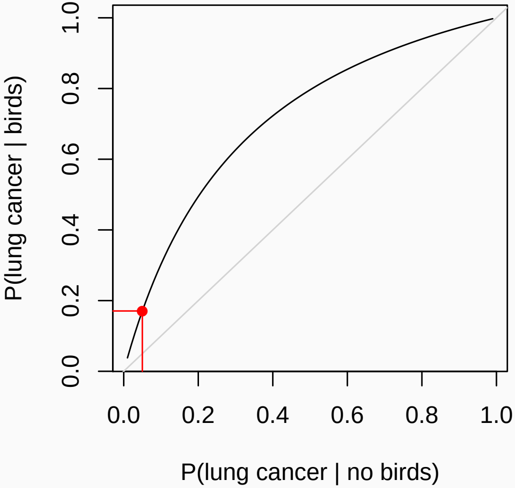
Bird OR Curve



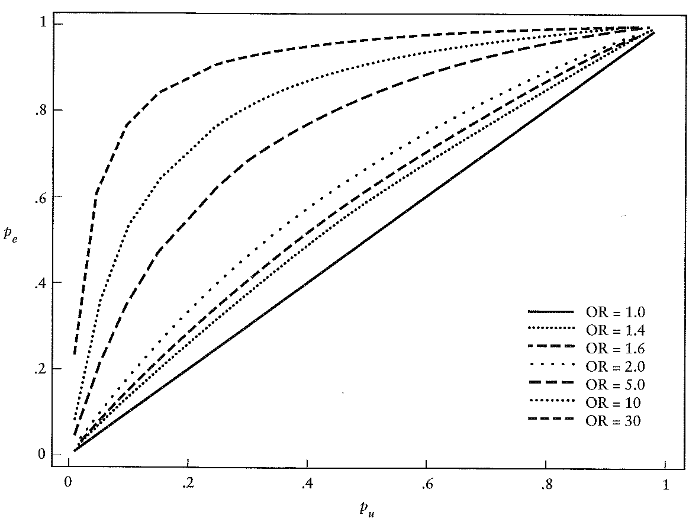
Bird OR Curve



Bird OR Curve



OR Curves



(An old) Example - House

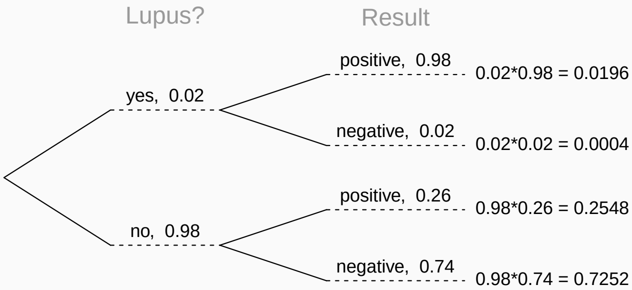
If you've ever watched the TV show House on Fox, you know that Dr. House regularly states, "It's never lupus."

Lupus is a medical phenomenon where antibodies that are supposed to attack foreign cells to prevent infections instead see plasma proteins as foreign bodies, leading to a high risk of blood clotting. It is believed that 2% of the population suffer from this disease.

The test for lupus is very accurate if the person actually has lupus, however is very inaccurate if the person does not. More specifically, the test is 98% accurate if a person actually has the disease. The test is 74% accurate if a person does not have the disease.

Is Dr. House correct even if someone tests positive for Lupus?

(An old) Example - House



$$\begin{aligned} P(\text{Lupus}|+) &= \frac{P(+, \text{Lupus})}{P(+, \text{Lupus}) + P(+, \text{No Lupus})} \\ &= \frac{0.0196}{0.0196 + 0.2548} = 0.0714 \end{aligned}$$

Testing for lupus

It turns out that testing for Lupus is actually quite complicated, a diagnosis usually relies on the outcome of multiple tests, often including: a complete blood count, an erythrocyte sedimentation rate, a kidney and liver assessment, a urinalysis, and or an antinuclear antibody (ANA) test.

It is important to think about what is involved in each of these tests (e.g. deciding if complete blood count is high or low) and how each of the individual tests and related decisions plays a role in the overall decision of diagnosing a patient with lupus.

Testing for lupus

At some level we can view a diagnosis as a binary decision (lupus or no lupus) that involves the complex integration of various explanatory variables.

The example does not give us any information about how a diagnosis is made, but what it does give us is just as important - the sensitivity and the specificity of the test. These values are critical for our understanding of what a positive or negative test result actually means.

Sensitivity and Specificity

Sensitivity - measures a tests ability to identify positive results.

$$P(\text{Test } + \mid \text{Conditon } +) = P(+|\text{lupus}) = 0.98$$

Specificity - measures a tests ability to identify negative results.

$$P(\text{Test } - \mid \text{Condition } -) = P(-|\text{no lupus}) = 0.74$$

Sensitivity and Specificity

Sensitivity - measures a tests ability to identify positive results.

$$P(\text{Test } + \mid \text{Conditon } +) = P(+|\text{lupus}) = 0.98$$

Specificity - measures a tests ability to identify negative results.

$$P(\text{Test } - \mid \text{Condition } -) = P(-|\text{no lupus}) = 0.74$$

It is illustrative to think about the extreme cases - what is the sensitivity and specificity of a test that always returns a positive result? What about a test that always returns a negative result?

Sensitivity and Specificity (cont.)

	Condition Positive	Condition Negative
Test Positive	True Positive	False Positive (Type I error)
Test Negative	False Negative (Type II error)	True Negative

Sensitivity and Specificity (cont.)

	Condition Positive	Condition Negative
Test Positive	True Positive	False Positive (Type I error)
Test Negative	False Negative (Type II error)	True Negative

$$\textit{Sensitivity} = P(\text{Test } + \mid \text{Condition } +)$$

Sensitivity and Specificity (cont.)

	Condition Positive	Condition Negative
Test Positive	True Positive	False Positive (Type I error)
Test Negative	False Negative (Type II error)	True Negative

$$\text{Sensitivity} = P(\text{Test } + \mid \text{Condition } +) = TP/(TP + FN)$$

Sensitivity and Specificity (cont.)

	Condition Positive	Condition Negative
Test Positive	True Positive	False Positive (Type I error)
Test Negative	False Negative (Type II error)	True Negative

Sensitivity $= P(\text{Test } + \mid \text{Condition } +) = TP/(TP + FN)$

Specificity $= P(\text{Test } - \mid \text{Condition } -)$

Sensitivity and Specificity (cont.)

	Condition Positive	Condition Negative
Test Positive	True Positive	False Positive (Type I error)
Test Negative	False Negative (Type II error)	True Negative

$$\textit{Sensitivity} = P(\text{Test } + \mid \text{Condition } +) = TP/(TP + FN)$$

$$\textit{Specificity} = P(\text{Test } - \mid \text{Condition } -) = TN/(FP + TN)$$

Sensitivity and Specificity (cont.)

	Condition Positive	Condition Negative
Test Positive	True Positive	False Positive (Type I error)
Test Negative	False Negative (Type II error)	True Negative

$$\textit{Sensitivity} = P(\text{Test } + \mid \text{Condition } +) = TP/(TP + FN)$$

$$\textit{Specificity} = P(\text{Test } - \mid \text{Condition } -) = TN/(FP + TN)$$

$$\textit{False negative rate } (\beta) = P(\text{Test } - \mid \text{Condition } +)$$

Sensitivity and Specificity (cont.)

	Condition Positive	Condition Negative
Test Positive	True Positive	False Positive (Type I error)
Test Negative	False Negative (Type II error)	True Negative

$$\text{Sensitivity} = P(\text{Test } + \mid \text{Condition } +) = TP/(TP + FN)$$

$$\text{Specificity} = P(\text{Test } - \mid \text{Condition } -) = TN/(FP + TN)$$

$$\text{False negative rate } (\beta) = P(\text{Test } - \mid \text{Condition } +) = FN/(TP + FN)$$

Sensitivity and Specificity (cont.)

	Condition Positive	Condition Negative
Test Positive	True Positive	False Positive (Type I error)
Test Negative	False Negative (Type II error)	True Negative

Sensitivity $= P(\text{Test } + \mid \text{Condition } +) = TP/(TP + FN)$

Specificity $= P(\text{Test } - \mid \text{Condition } -) = TN/(FP + TN)$

False negative rate $(\beta) = P(\text{Test } - \mid \text{Condition } +) = FN/(TP + FN)$

False positive rate $(\alpha) = P(\text{Test } + \mid \text{Condition } -)$

Sensitivity and Specificity (cont.)

	Condition Positive	Condition Negative
Test Positive	True Positive	False Positive (Type I error)
Test Negative	False Negative (Type II error)	True Negative

$$\textit{Sensitivity} = P(\text{Test } + \mid \text{Condition } +) = TP/(TP + FN)$$

$$\textit{Specificity} = P(\text{Test } - \mid \text{Condition } -) = TN/(FP + TN)$$

$$\textit{False negative rate } (\beta) = P(\text{Test } - \mid \text{Condition } +) = FN/(TP + FN)$$

$$\textit{False positive rate } (\alpha) = P(\text{Test } + \mid \text{Condition } -) = FP/(FP + TN)$$

Sensitivity and Specificity (cont.)

	Condition Positive	Condition Negative
Test Positive	True Positive	False Positive (Type I error)
Test Negative	False Negative (Type II error)	True Negative

$$\textit{Sensitivity} = P(\text{Test } + \mid \text{Condition } +) = TP/(TP + FN)$$

$$\textit{Specificity} = P(\text{Test } - \mid \text{Condition } -) = TN/(FP + TN)$$

$$\textit{False negative rate } (\beta) = P(\text{Test } - \mid \text{Condition } +) = FN/(TP + FN)$$

$$\textit{False positive rate } (\alpha) = P(\text{Test } + \mid \text{Condition } -) = FP/(FP + TN)$$

$$\textit{Sensitivity} = 1 - \textit{False negative rate} = \text{Power}$$

$$\textit{Specificity} = 1 - \textit{False positive rate}$$

So what?

Clearly it is important to know the Sensitivity and Specificity of test (and or the false positive and false negative rates). Along with the incidence of the disease (e.g. $P(\text{lupus})$) these values are necessary to calculate important quantities like $P(\text{lupus}|+)$.

Additionally, our brief foray into power analysis before the first midterm should also give you an idea about the trade offs that are inherent in minimizing false positive and false negative rates (increasing power required either increasing α or n).

How should we use this information when we are trying to come up with a decision?

Back to Spam

In lab this week, we examined a data set of emails where we were interesting in identifying the spam messages. We examined different logistic regression models to evaluate how different predictors influenced the probability of a message being spam.

These models can also be used to assign probabilities to incoming messages (this is equivalent to prediction in the case of SLR / MLR). However, if we were designing a spam filter this would only be half of the battle, we would also need to use these probabilities to make a decision about which emails get flagged as spam.

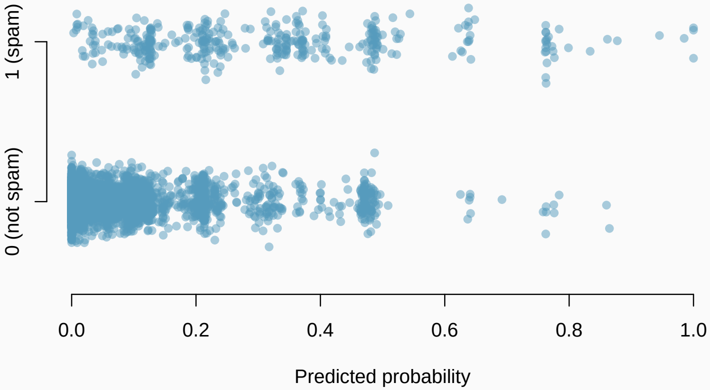
Back to Spam

In lab this week, we examined a data set of emails where we were interesting in identifying the spam messages. We examined different logistic regression models to evaluate how different predictors influenced the probability of a message being spam.

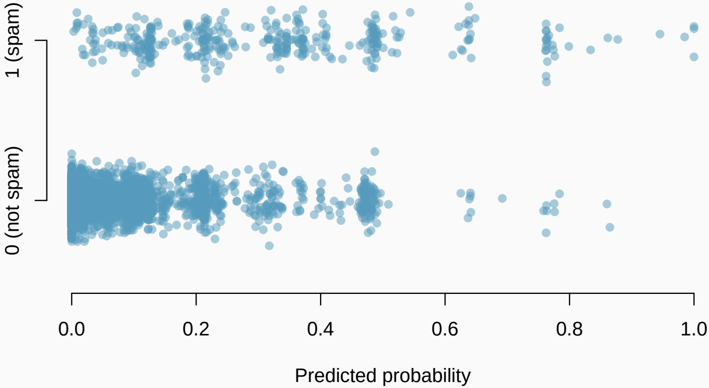
These models can also be used to assign probabilities to incoming messages (this is equivalent to prediction in the case of SLR / MLR). However, if we were designing a spam filter this would only be half of the battle, we would also need to use these probabilities to make a decision about which emails get flagged as spam.

While not the only possible solution, we will consider a simple approach where we choose a threshold probability and any email that exceeds that probability is flagged as spam.

Picking a threshold

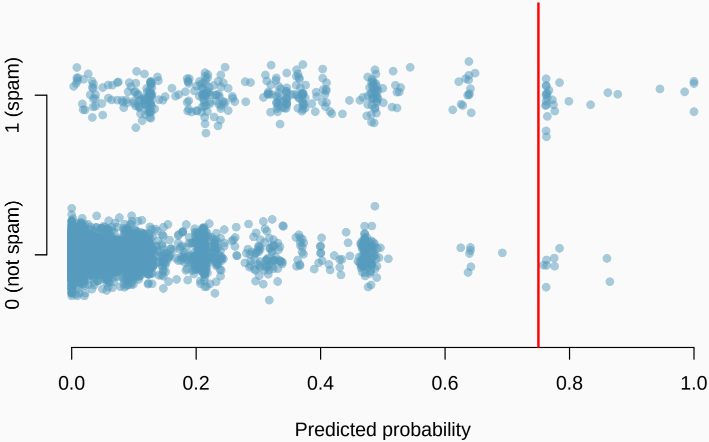


Picking a threshold



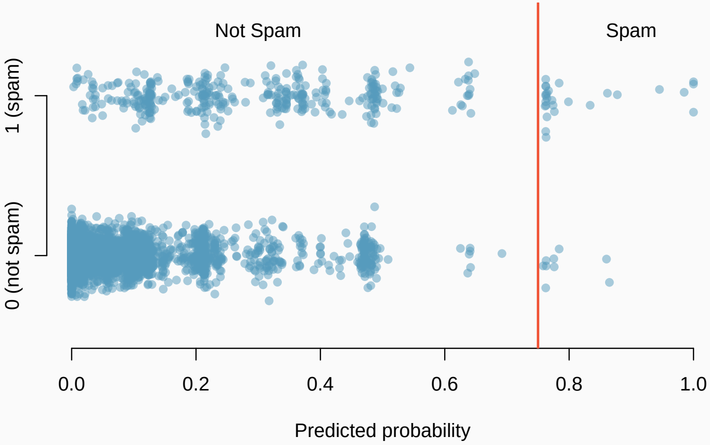
Lets see what happens if we pick our threshold to be *0.75*.

Picking a threshold



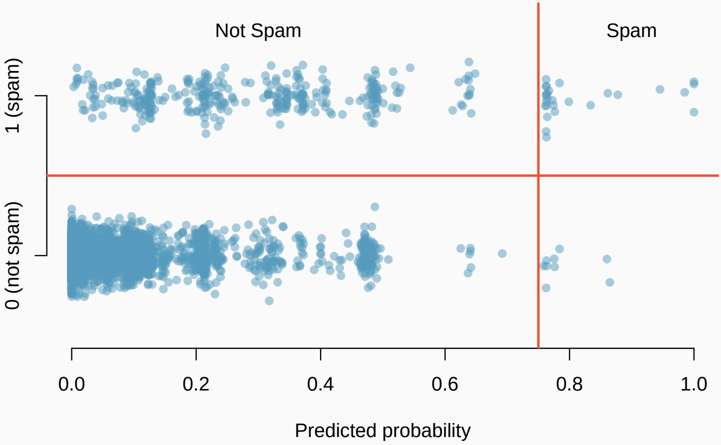
Lets see what happens if we pick our threshold to be *0.75*.

Picking a threshold



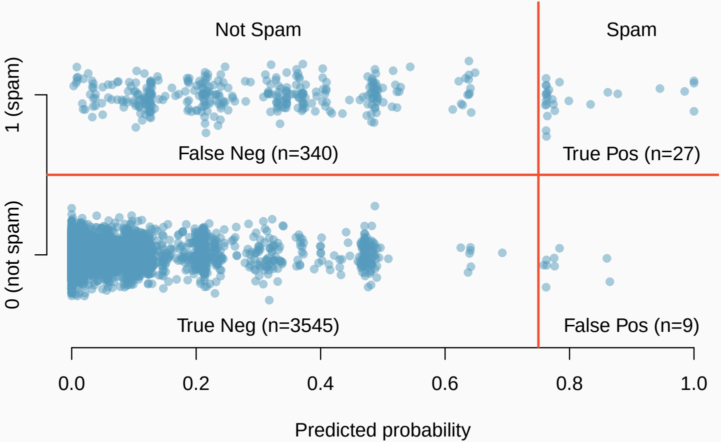
Lets see what happens if we pick our threshold to be *0.75*.

Picking a threshold



Lets see what happens if we pick our threshold to be *0.75*.

Picking a threshold



Lets see what happens if we pick our threshold to be *0.75*.

Consequences of picking a threshold

For our data set picking a threshold of 0.75 gives us the following results:

$$FN = 340 \qquad TP = 27$$
$$TN = 3545 \qquad FP = 9$$

Consequences of picking a threshold

For our data set picking a threshold of 0.75 gives us the following results:

$$\begin{aligned} FN &= 340 & TP &= 27 \\ TN &= 3545 & FP &= 9 \end{aligned}$$

What are the sensitivity and specificity for this particular decision rule?

Consequences of picking a threshold

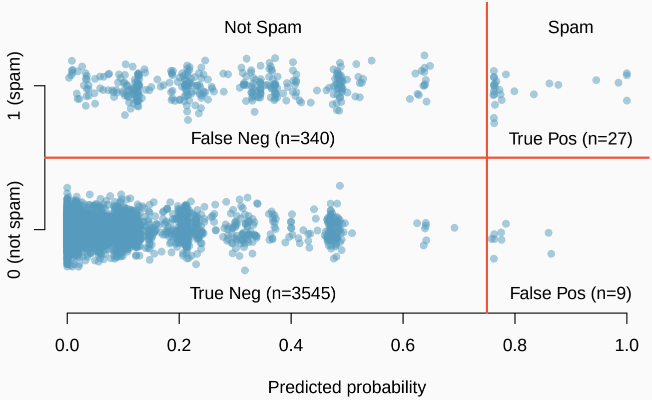
For our data set picking a threshold of 0.75 gives us the following results:

$$FN = 340 \qquad TP = 27$$
$$TN = 3545 \qquad FP = 9$$

What are the sensitivity and specificity for this particular decision rule?

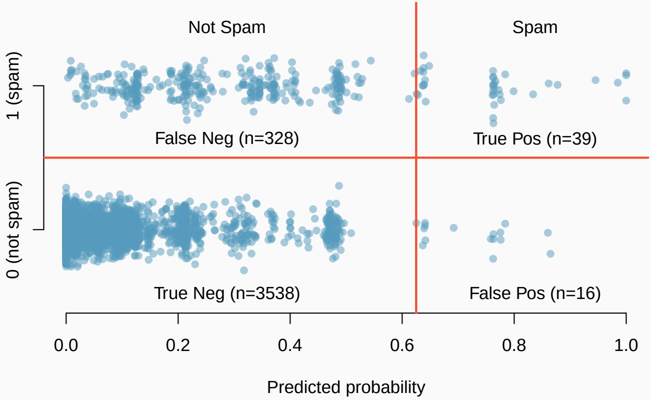
$$\textit{Sensitivity} = TP/(TP + FN) = 27/(27 + 340) = 0.073$$
$$\textit{Specificity} = TN/(FP + TN) = 3545/(9 + 3545) = 0.997$$

Trying other thresholds



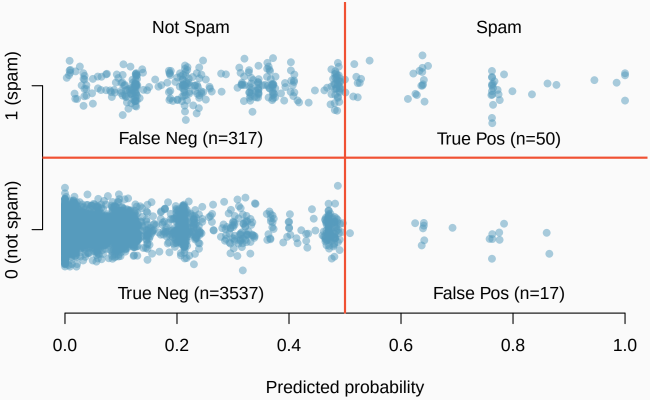
Threshold	0.75	0.625	0.5	0.375	0.25
Sensitivity	0.074				
Specificity	0.997				

Trying other thresholds



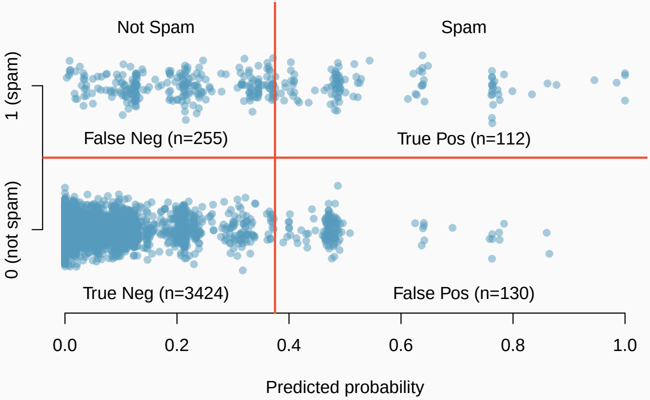
Threshold	0.75	0.625	0.5	0.375	0.25
Sensitivity	0.074	0.106			
Specificity	0.997	0.995			

Trying other thresholds



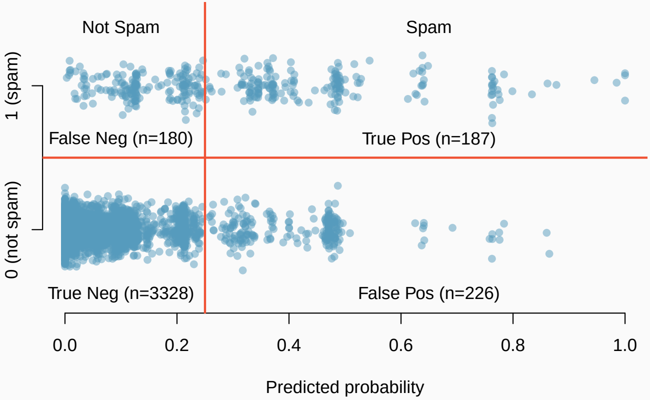
Threshold	0.75	0.625	0.5	0.375	0.25
Sensitivity	0.074	0.106	0.136		
Specificity	0.997	0.995	0.995		

Trying other thresholds



Threshold	0.75	0.625	0.5	0.375	0.25
Sensitivity	0.074	0.106	0.136	0.305	
Specificity	0.997	0.995	0.995	0.963	

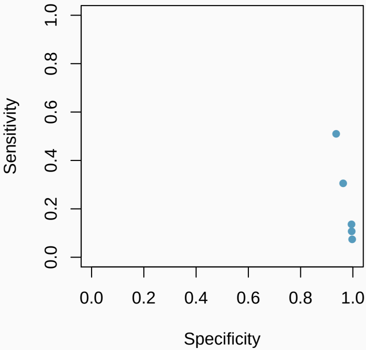
Trying other thresholds



Threshold	0.75	0.625	0.5	0.375	0.25
Sensitivity	0.074	0.106	0.136	0.305	0.510
Specificity	0.997	0.995	0.995	0.963	0.936

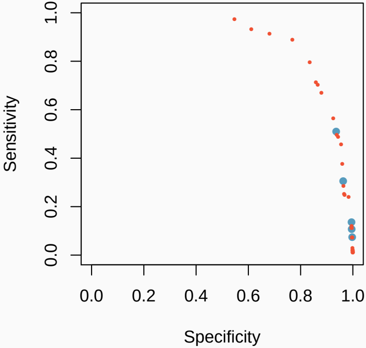
Relationship between Sensitivity and Specificity

Threshold	0.75	0.625	0.5	0.375	0.25
Sensitivity	0.074	0.106	0.136	0.305	0.510
Specificity	0.997	0.995	0.995	0.963	0.936



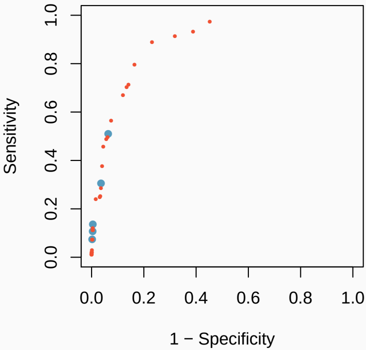
Relationship between Sensitivity and Specificity

Threshold	0.75	0.625	0.5	0.375	0.25
Sensitivity	0.074	0.106	0.136	0.305	0.510
Specificity	0.997	0.995	0.995	0.963	0.936

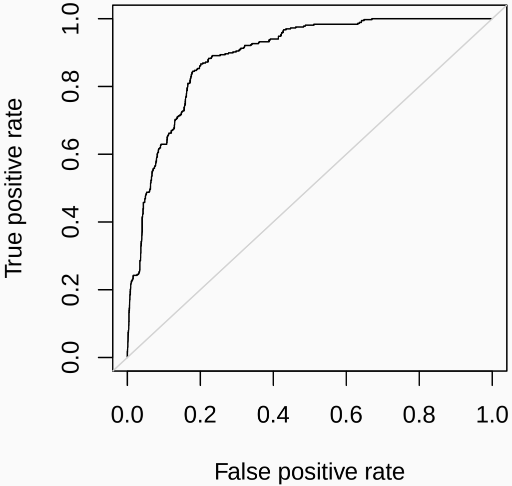


Relationship between Sensitivity and Specificity

Threshold	0.75	0.625	0.5	0.375	0.25
Sensitivity	0.074	0.106	0.136	0.305	0.510
Specificity	0.997	0.995	0.995	0.963	0.936



Receiver operating characteristic (ROC) curve



Receiver operating characteristic (ROC) curve (cont.)

Why do we care about ROC curves?

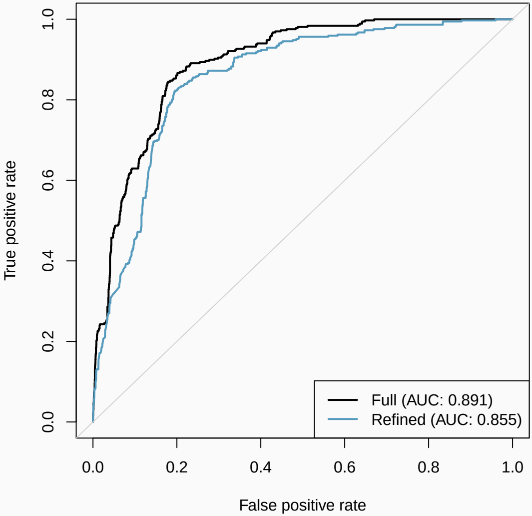
- Shows the trade off in sensitivity and specificity for all possible thresholds.
- Straight forward to compare performance vs. chance.
- Can use the area under the curve (AUC) as an assessment of the predictive ability of a model.

Refining the Spam model

```
g_refined = glm(spam ~ to_multiple+cc+image+attach+winner
                      +password+line_breaks+format+re_subj
                      +urgent_subj+exclaim_mess,
                data=email, family=binomial)
summary(g_refined)
```

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.7594	0.1177	-14.94	0.0000
to_multipleyes	-2.7368	0.3156	-8.67	0.0000
ccyes	-0.5358	0.3143	-1.71	0.0882
imageyes	-1.8585	0.7701	-2.41	0.0158
attachyes	1.2002	0.2391	5.02	0.0000
winneryes	2.0433	0.3528	5.79	0.0000
passwordyes	-1.5618	0.5354	-2.92	0.0035
line_breaks	-0.0031	0.0005	-6.33	0.0000
formatPlain	1.0130	0.1380	7.34	0.0000
re_subjyes	-2.9935	0.3778	-7.92	0.0000
urgent_subjyes	3.8830	1.0054	3.86	0.0001
exclaim_mess	0.0093	0.0016	5.71	0.0000

Comparing models



Utility Functions

There are many other reasonable quantitative approaches we can use to decide on what is the “best” threshold.

If you’ve taken an economics course you have probably heard of the idea of utility functions, we can assign costs and benefits to each of the possible outcomes and use those to calculate a utility for each circumstance.

Utility function for our spam filter

To write down a utility function for a spam filter we need to consider the costs / benefits of each out.

Outcome	Utility
True Positive	
True Negative	
False Positive	
False Negative	

Utility function for our spam filter

To write down a utility function for a spam filter we need to consider the costs / benefits of each out.

Outcome	Utility
True Positive	1
True Negative	
False Positive	
False Negative	

Utility function for our spam filter

To write down a utility function for a spam filter we need to consider the costs / benefits of each out.

Outcome	Utility
True Positive	1
True Negative	1
False Positive	
False Negative	

Utility function for our spam filter

To write down a utility function for a spam filter we need to consider the costs / benefits of each out.

Outcome	Utility
True Positive	1
True Negative	1
False Positive	-50
False Negative	

Utility function for our spam filter

To write down a utility function for a spam filter we need to consider the costs / benefits of each out.

Outcome	Utility
True Positive	1
True Negative	1
False Positive	-50
False Negative	-5

Utility function for our spam filter

To write down a utility function for a spam filter we need to consider the costs / benefits of each out.

Outcome	Utility
True Positive	1
True Negative	1
False Positive	-50
False Negative	-5

$$U(p) = TP(p) + TN(p) - 50 \times FP(p) - 5 \times FN(p)$$

Utility for the 0.75 threshold

For the email data set picking a threshold of 0.75 gives us the following results:

$$\begin{aligned} FN &= 340 \qquad & TP &= 27 \\ TN &= 3545 \qquad & FP &= 9 \end{aligned}$$

Utility for the 0.75 threshold

For the email data set picking a threshold of 0.75 gives us the following results:

$$
\begin{aligned}
FN &= 340 \qquad & TP &= 27 \\
TN &= 3545 \qquad & FP &= 9
\end{aligned}
$$

$$
\begin{aligned}
U(p) &= TP(p) + TN(p) - 50 \times FP(p) - 5 \times FN(p) \\
&= 27 + 3545 - 50 \times 9 - 5 \times 340 = 1422
\end{aligned}
$$

Utility for the 0.75 threshold

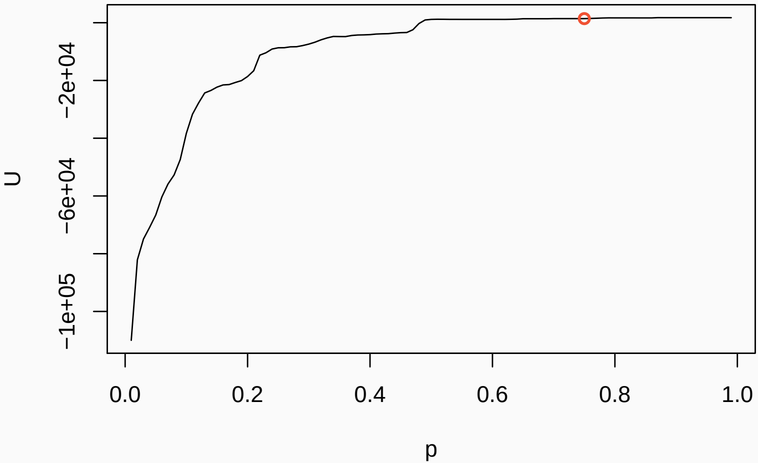
For the email data set picking a threshold of 0.75 gives us the following results:

$$FN = 340 \qquad TP = 27$$
$$TN = 3545 \qquad FP = 9$$

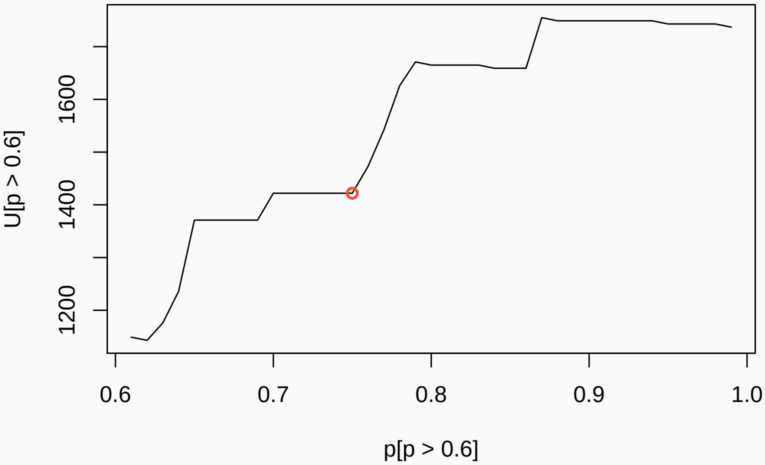
$$\begin{aligned} U(p) &= TP(p) + TN(p) - 50 \times FP(p) - 5 \times FN(p) \\ &= 27 + 3545 - 50 \times 9 - 5 \times 340 = 1422 \end{aligned}$$

Not useful by itself, but allows us to compare with other thresholds.

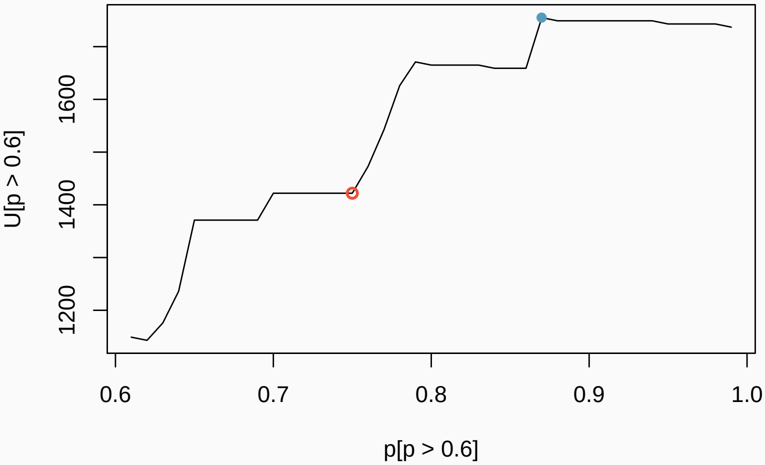
Utility curve



Utility curve (zoom)



Utility curve (zoom)



Maximum Utility