

Chapter 8: Multiple and logistic regression

OpenIntro Statistics, 3rd Edition

Slides developed by Mine Çetinkaya-Rundel of OpenIntro.

The slides may be copied, edited, and/or shared via the CC BY-SA license.

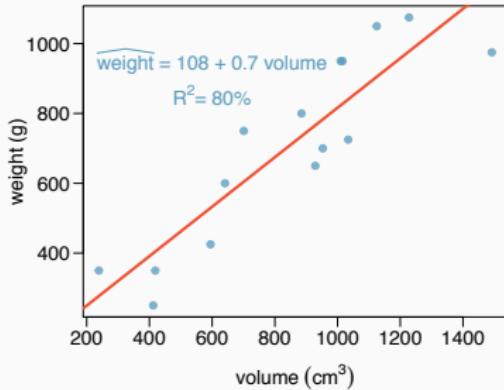
Some images may be included under fair use guidelines (educational purposes).

Multiple regression

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

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 cm³ 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

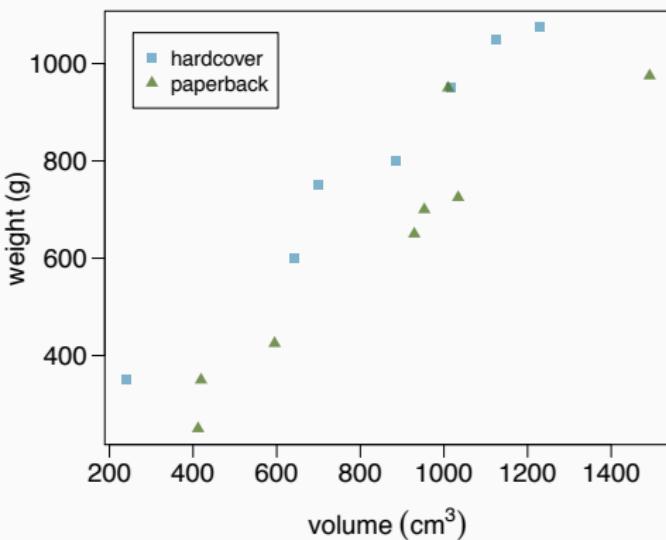
Residual standard error: 123.9 on 13 degrees of freedom

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?



Modeling weights of books using volume and 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

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

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

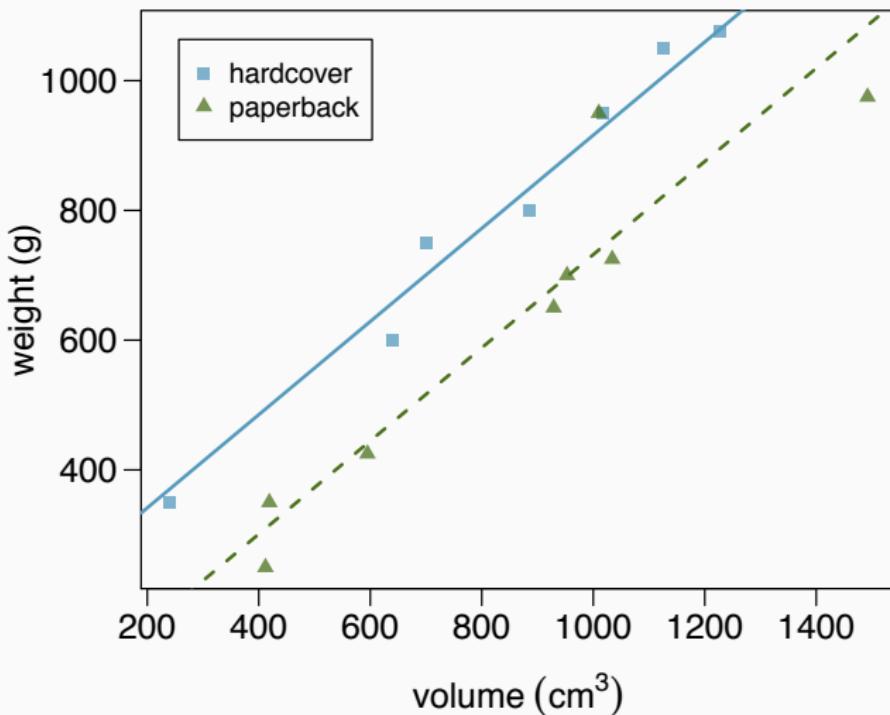
1. For *hardcover* books: plug in *0* for cover

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

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

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

Visualising the linear model

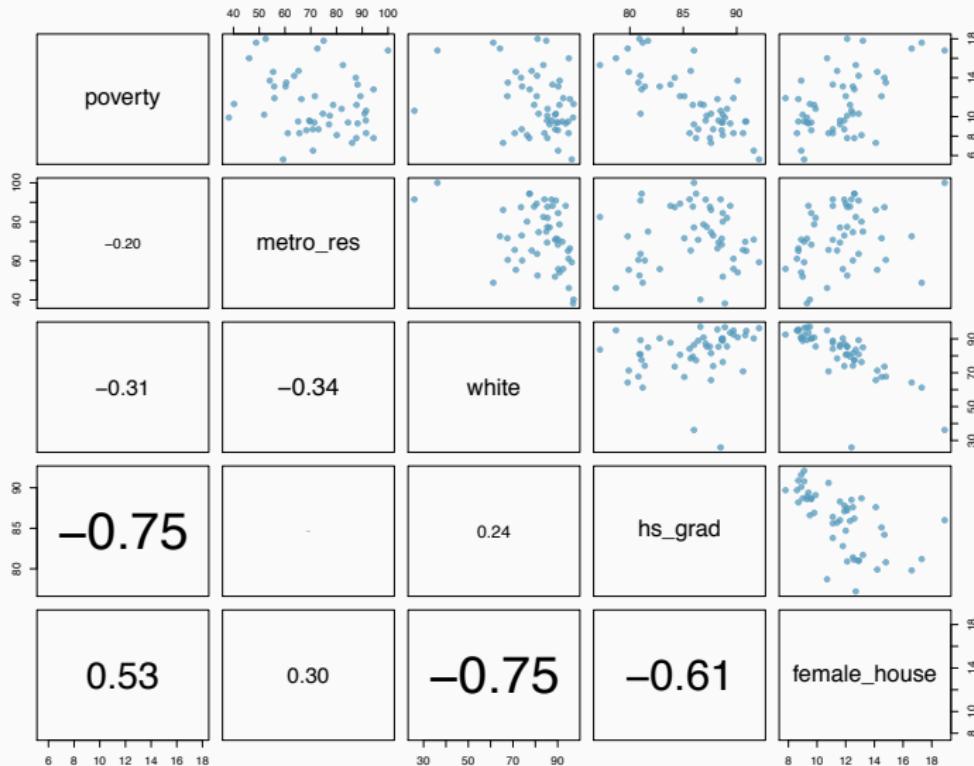


Interpretation of the regression coefficients

	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

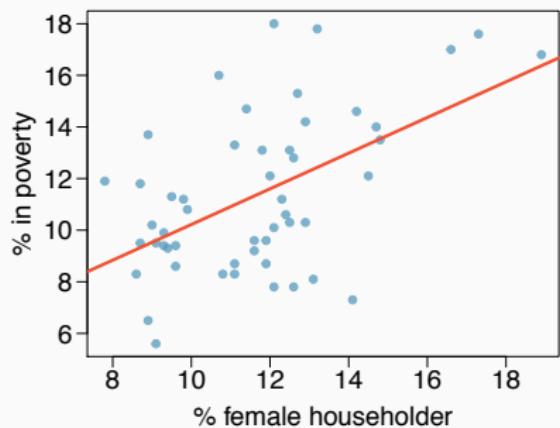
- *Slope of volume:* All else held constant, books that are 1 more cubic centimeter in volume tend to weigh about 0.72 grams more.
- *Slope of cover:* All else held constant, the model predicts that paperback books weigh 184 grams lower than hardcover books.
- *Intercept:* Hardcover books with no volume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

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



$$R = 0.53$$

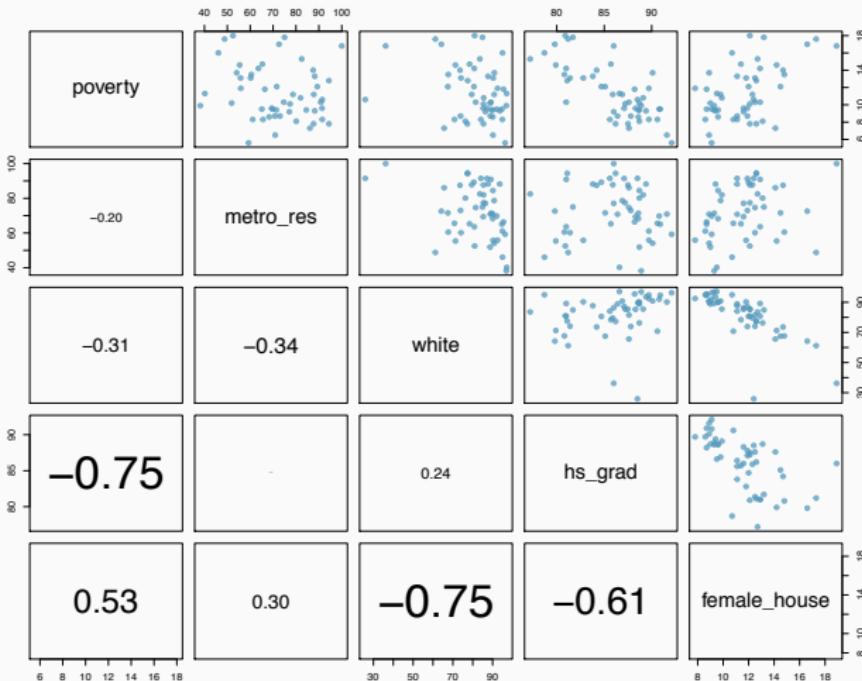
$$R^2 = 0.53^2 = 0.28$$

Predicting poverty using % female hh + % white

<i>Linear model:</i>	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

<i>ANOVA:</i>	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			

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 (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.

- 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

- When any 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.

Model selection

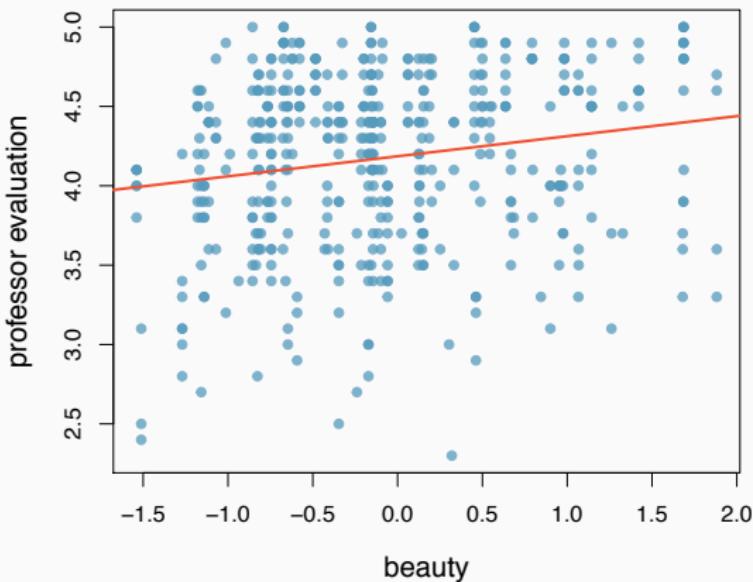
Beauty in the classroom

- 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: instructors' pulchritude and putative pedagogical productivity" *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

$$R^2 = 0.0336$$

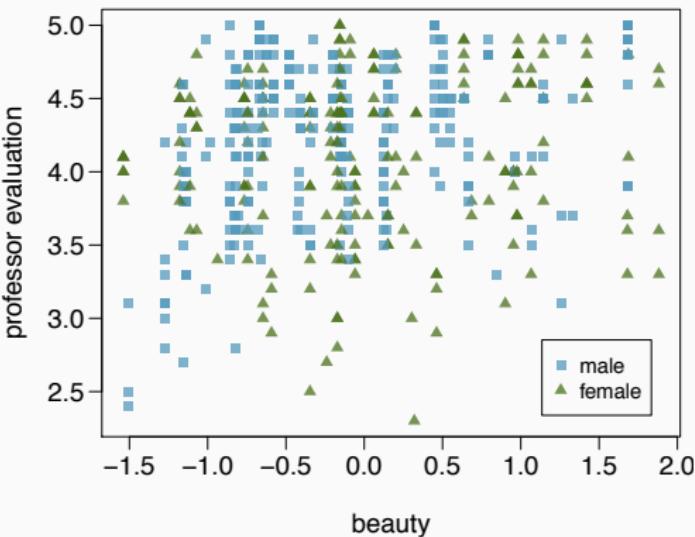
- (a) Model predicts 3.36% of professor ratings correctly.
- (b) Beauty is not a significant predictor of professor evaluation.
- (c) Professors who score 1 point above average in their beauty score are tend to also score 0.13 points higher in their evaluation.
- (d) 3.36% of variability in beauty scores 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.

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?



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

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: 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
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

step function in R

Call:

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

Coefficients:

	beauty	gendermale
(Intercept)		
4.628435	0.105546	0.208079
age	formalyes	native non english
-0.008844	0.132422	-0.243003
tenure	tenure tenured	
tenure track		
-0.206784	-0.175967	

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

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

Logistic regression

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.)

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 route 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:

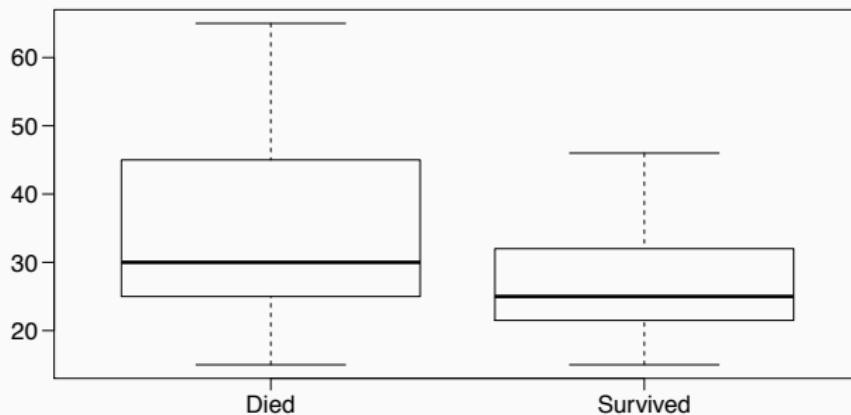
	Male	Female
Died	20	5
Survived	10	10

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.

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.

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 - 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:

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

Female model:

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

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.