Statistics for Life Sciences

Nathan Harmston

2022-03-24

Linear models and anovas

```
    library(datasets)

data <- esoph
head(data)
model <- aov(ncases ~ agegp*alcgp, data = data)
summary(model)
model$coefficients

modellm <- lm(ncases ~ agegp*alcgp, data = data)
summary(modellm)
```

Linear regression

So, we've talked about linear regression, which covers the linear relationship between a continuous dependent/response variable and a continuous or categorical independent/predictor variable.

```
?lm
```

we're trying to ask whether a independent variable explains a significant proportion of the variance in the dependent variable

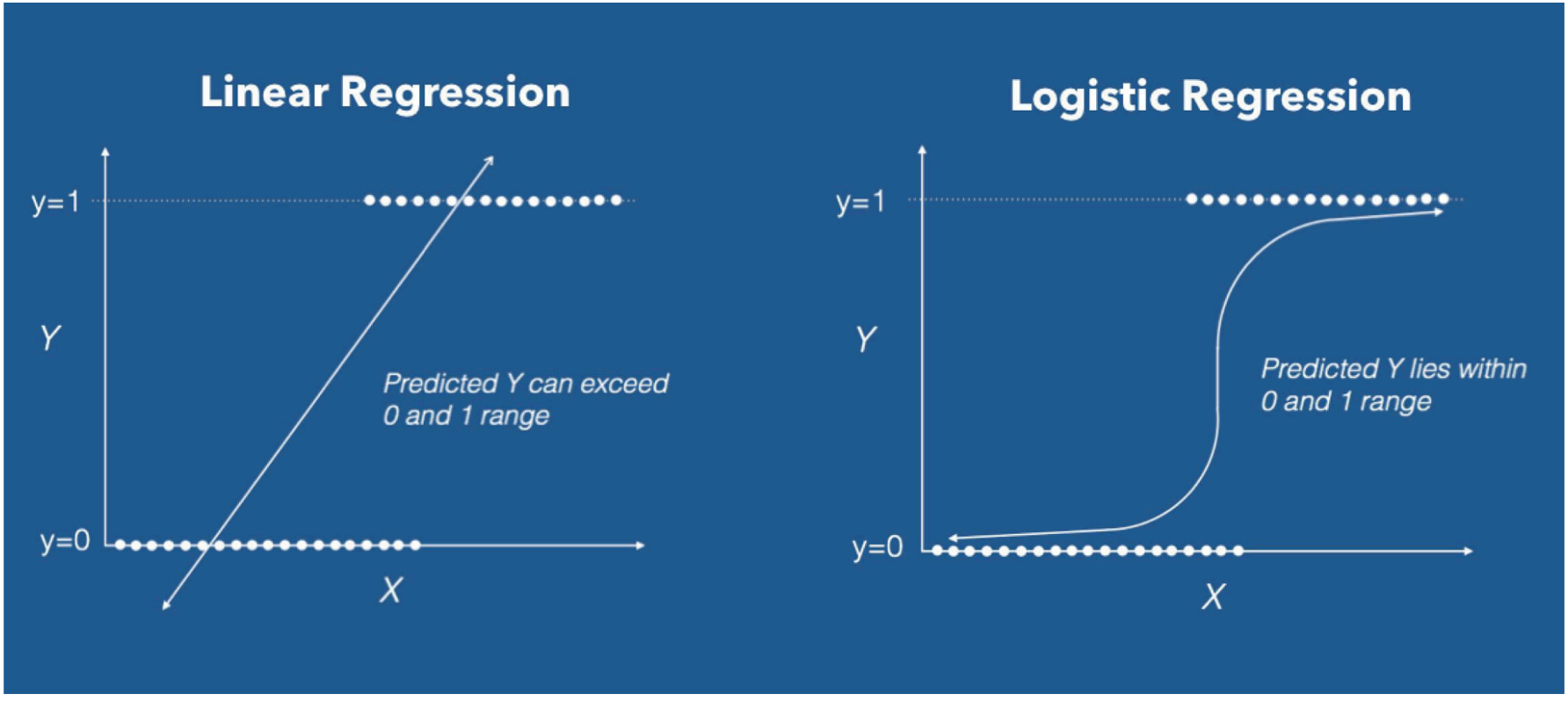
or we're trying to model - predict wgt given smoking status and gest

$$Wgt = -2389.573 + 143 Gest - 244.544 Smoke \quad (1)$$

Logistic Regression

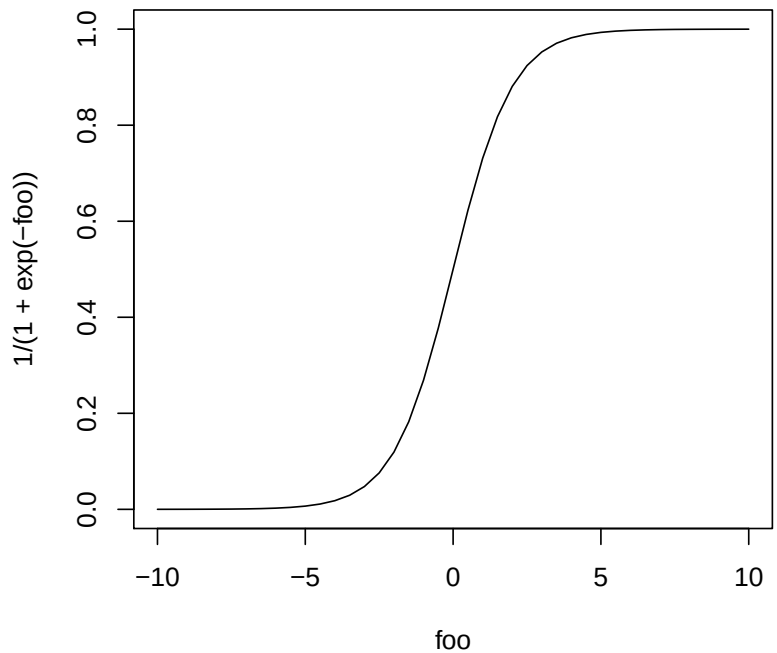
Now, say that we're interested in a categorical dependent variable, how do we look for a relationship between that variable and a continuous or categorical independent variable?

Logistic Regression



Logistic regression

the dependent/response/outcome is 0 or 1 (true or false, heads or tails, disease or not disease, obese or not obese), not continuous



Ovarian Cancer

	Age	Resid Disease	Rx	ECOG	BP	Chol
1	72.33	yes	A	good	117.83	13.58
2	74.49	yes	A	good	114.00	7.78
3	66.47	yes	A	bad	117.55	10.95
4	74.50	yes	A	bad	113.50	22.50
5	43.14	yes	A	good	139.19	22.11
6	63.22	no	B	bad	124.80	8.46
7	64.42	yes	B	good	118.09	23.19
8	58.31	no	B	good	130.09	26.51
.	.	.	.	.	.	.
.	.	.	.	.	.	.
25	44.21	yes	B	good	138.34	26.25
26	59.59	no	B	bad	129.18	22.93

Logistic Regression - Ovarian Cancer

In this cohort of ovarian cancer patients,

- is there a linear relationship between age and residual disease?

In this instance, age is the independent variable. Age can influence residual disease, but residual disease is not going to change a subjects age.

Logistic Regression - Ovarian Cancer

Logistic Regression

Logistic regression is an instance of classification technique that you can use to predict a qualitative response. More specifically, logistic regression models the probability that residual disease belongs to a particular category.

That means that, if you are trying to do residual disease classification, where the residual disease response falls into one of the two categories, yes or no, you'll use logistic regression models to estimate the probability that residual disease belongs to a particular category based on another factor, like age.

For example, the probability of residual disease given age can be written as:

$$Pr(\text{residual disease} = yes|age)$$

The values of $Pr(\text{residual disease} = yes|age)$) will range between 0 and 1. Then, for any given value of age, a prediction can be made for residual disease. The notation $Pr(\text{residual disease})$ is the probability of there being residual disease at a certain age.

Logistic Regression

Remember with linear regression, our fitted line follows

$$y = \beta_0 + \beta_1 x + \epsilon$$

In logistic regression, our fitted line follows

$$\frac{p(X)}{1-p(X)} = e^{\beta_0+\beta_1 X} \quad \text{or} \quad \frac{p(\text{residual disease})}{1-p(\text{residual disease})} = e^{\beta_0+\beta_1 age}$$

and

$$\frac{p(\text{residual disease})}{1-p(\text{residual disease})}$$

is called the odds ratio, and can take on any value between 0 and ∞. Values of the odds ratio close to 0 and ∞ indicate very low and very high probabilities of p(residual disease), respectively.

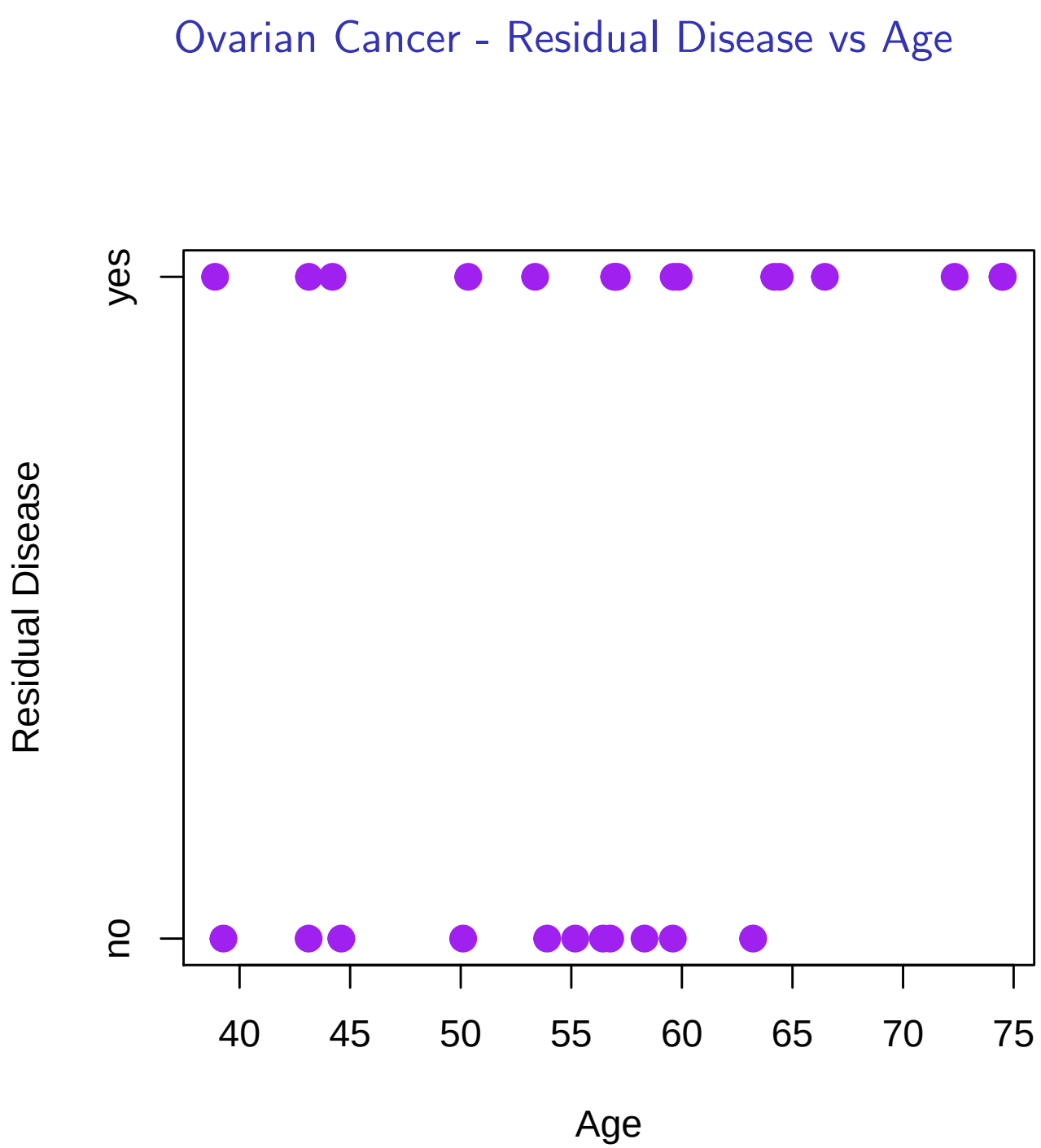
Logistic Regression - Logit Function

The odds ratio can tell us about the relationship between the residual disease and age. If we take the log

$$log(\frac{p(\text{residual disease})}{1-p(\text{residual disease})}) = \beta_0 + \beta_1 age$$

the left-hand side is called the logit. In this logistic regression model, increasing age by one unit changes the logit by β_0. The amount that p(residual disease) changes due to a one-unit change in age will depend on the current value of age.

But regardless of the value of age, if $\beta 1$ is positive then increasing age will be associated with increasing p(residual disease), and if β_1 is negative then increasing age will be associated with decreasing p(residual disease).


Ovarian Cancer - Residual Disease vs Age
Residual Disease
yes
no
40
45
50
55
60
65
70
75
Age

Ovarian Cancer - Residual Disease vs Age

	age	probability
1	72.33	0.80
2	74.49	0.82
3	66.47	0.73
4	74.50	0.82
5	43.14	0.38
6	63.22	0.69
7	64.42	0.71
8	58.31	0.62
.	.	.
.	.	.
25	44.21	0.39
26	59.59	0.64

Ovarian Cancer - Residual Disease vs Age

	age	probability	prediction
1	72.33	0.80	yes
2	74.49	0.82	yes
3	66.47	0.73	yes
4	74.50	0.82	yes
5	43.14	0.38	no
6	63.22	0.69	yes
7	64.42	0.71	yes
8	58.31	0.62	yes
.	.	.	.
.	.	.	.
25	44.21	0.39	no
26	59.59	0.64	yes

Ovarian Cancer - Residual Disease vs Age

	age	probability	prediction	actual
1	72.33	0.80	yes	yes
2	74.49	0.82	yes	yes
3	66.47	0.73	yes	yes
4	74.50	0.82	yes	yes
5	43.14	0.38	no	yes
6	63.22	0.69	yes	no
7	64.42	0.71	yes	yes
8	58.31	0.62	yes	no
.	.	.	.	.
.	.	.	.	.
25	44.21	0.39	no	yes
26	59.59	0.64	yes	no

Ovarian Cancer - Residual Disease vs Age

We can look at the odds ratio for this logistic regression to see how the expected change in odds of having residual disease ($ResidDisease = yes$) given an increase of one unit of age.

In our example the odds ratio $= 1.07$, meaning the expected change in the odds is $1 \times$ odds, i.e. there is no expected change in residual disease as age increases.

Summary - Logistic Regression

- Models the relationship between a categorical dependent variable continuous or categorical independent variable
- Odds ratio calculated from the model fitted βs can inform on the relationship between the dependent and independent variables
- The log of the odds ratio tells the increase in odds of the dependent variable with respect to a unit increase of the dependent variable, i.e. an age increase of one year would result in an estimated 2-fold increase in the quality of life or presence of residual disease

Confusion matrix

	Predict positive	Predict negative
Positive example	True Positive (TP)	False negative (FN)
Negative example	False Positive (FP)	True negative (TN)

Performance metrics for classifications

Precision, Positive Predictive Value: $\frac{TP}{TP+FP}$

Recall, Sensitivity, True Positive Rate: $\frac{TP}{TP+FN}$

Specificity, True Negative Rate: $\frac{TN}{TN+FP}$

False positive rate: $\frac{FP}{FP+TN}$

So how well are did we do?

	Predict positive	Predict negative
Positive example		
Negative example		

Receiver operating characteristic

compare the TPR against the FPR as you vary some parameter θ

why?

what is TPR? .... its the probability

what is FPR? ... its the probability

Receiver operating characteristic

compare the TPR against the FPR as you vary some parameter θ

why?

what is TPR? .... its the probability of saying something is labelled C when it is really from C

what is FPR? ... its the probability

Receiver operating characteristic

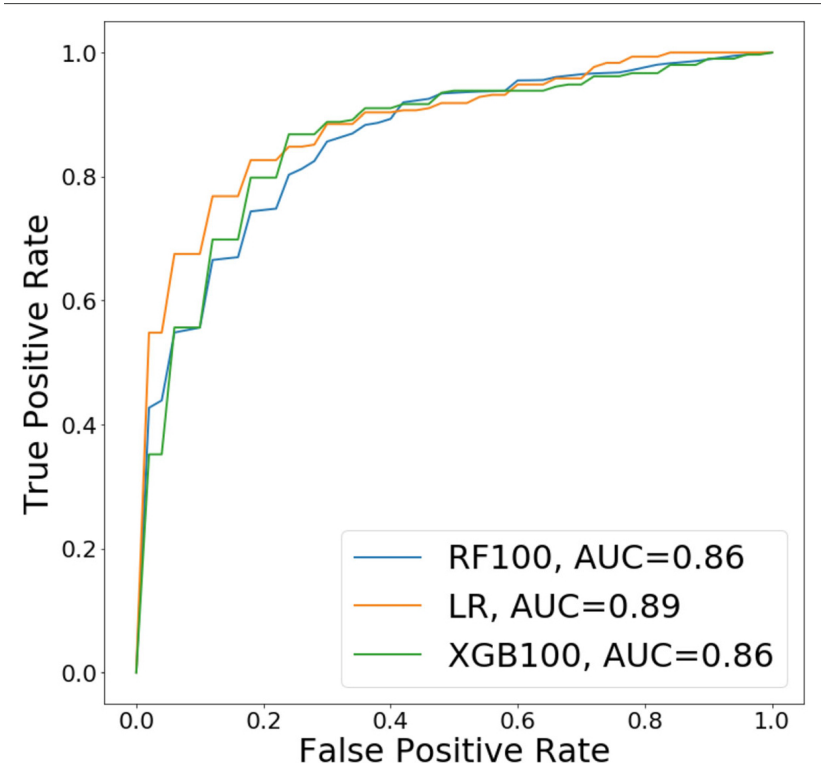
compare the TPR against the FPR as you vary some parameter θ

why?

what is TPR? .... its the probability of saying something is labelled C when it is really from C

what is FPR? ... its the probability of saying something is labelled C when it is not labelled C

Receiver operating characteristic



plot the TPR against the FPR

AUC - area under the curve reported as a measure of performance to rank different classifiers

Class imbalance

What is an unbalanced dataset?

majority of samples/genes/X belongs to one class

lots of real world datasets are imbalanced -

I build a classifier that predicts + for everything (n=100, +=95, -=5) it sees

$$TP = 95,\ FP = 5;\ TN = 0,\ FN = 0.$$

$$TPR = 95 / 95 = 1$$

Class imbalance

What is an unbalanced dataset?

majority of samples/genes/X belongs to one class

lots of real world datasets are imbalanced -

I build a classifier that predicts + for everything (n=100, +=95, -=5) it sees

$$TP = 95,\ FP = 5;\ TN = 0,\ FN = 0.$$

$$TPR = 95 / 95 = 1$$

wow aren't we doing well

Performance metrics for classifications

Precision, Positive Predictive Value: $\frac{TP}{TP+FP}$

Recall, Sensitivity, True Positive Rate: $\frac{TP}{TP+FN}$

Specificity, True Negative Rate: $\frac{TN}{TN+FP}$

False positive rate: $\frac{FP}{FP+TN}$

Class imbalance

In lots of biology, we often have very sparse dataset with many negative instances and few positive instances. Therefore, we prefer to avoid the involvement of true negatives in our prediction score.

Cancer is a rare event

make a classifier that says **not cancer**

overall I'm actually going to do pretty well on some performance measures - but is this really something you want to expose patients to?

Tip 8: Chicco 2017

Saito *et al.* 2015

How to deal with this?

		Positive	Negative
Prediction	Positive	236	103
	Negative	174	116,852

$$TPR = \frac{236}{236+174} = 0.576$$
$$FPR = \frac{103}{103+116{,}952} = 0.001$$

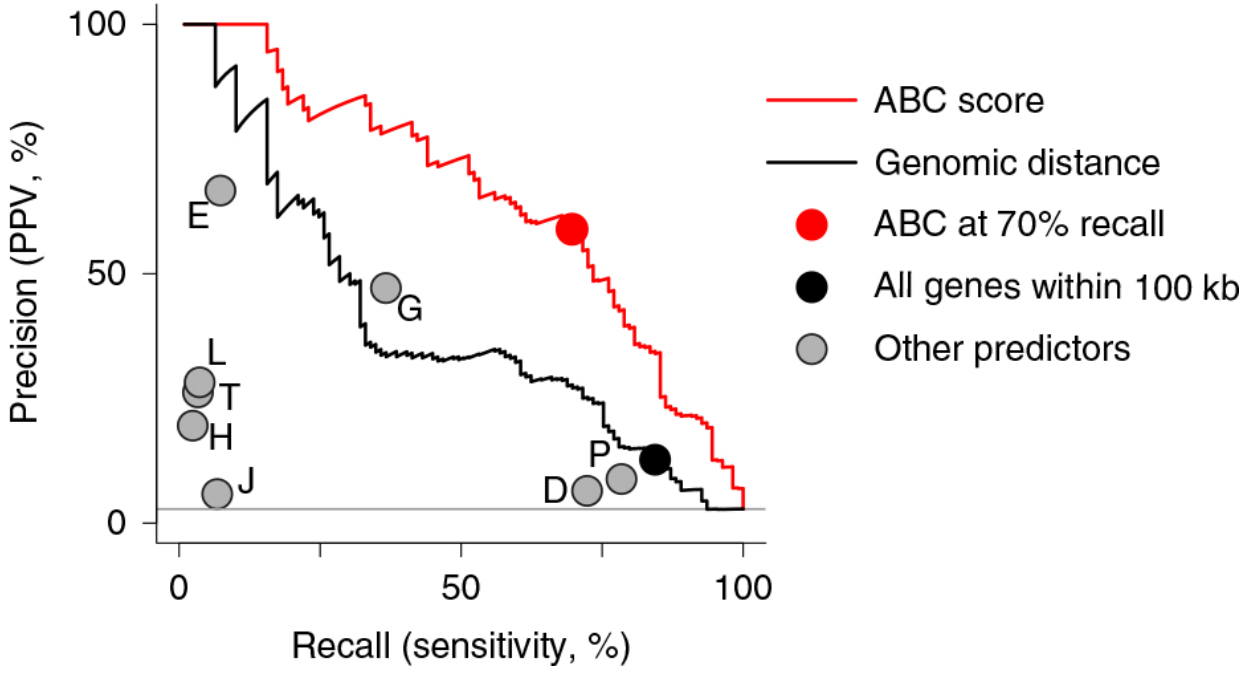
`https://davemcg.github.io/post/are-you-in-genomics-stop-using-roc-use-pr/`

Precision/PPV and TPR/sensitivity/recall

		Positive	Negative
Prediction	Positive	236	103
	Negative	174	116,852

$$PPV = \frac{236}{236+103} = 0.696$$
$$TPR = \frac{236}{236+174} = 0.576$$

Precision vs recall curves



Fulco *et al.* 2019

Next week....

- survival analysis
- chapter 5 and chapter 29

linear model problemset available over weekend