MIS 431 Data Mining

Logistic Regression

David Svancer – George Mason University School of Business

GEORGE MASON UNIVERSITY

School of Business

Multiple Linear Regression

Advertising Data Analysis – Putting It All Together

Is there a relationship between sales revenue and advertising budget?

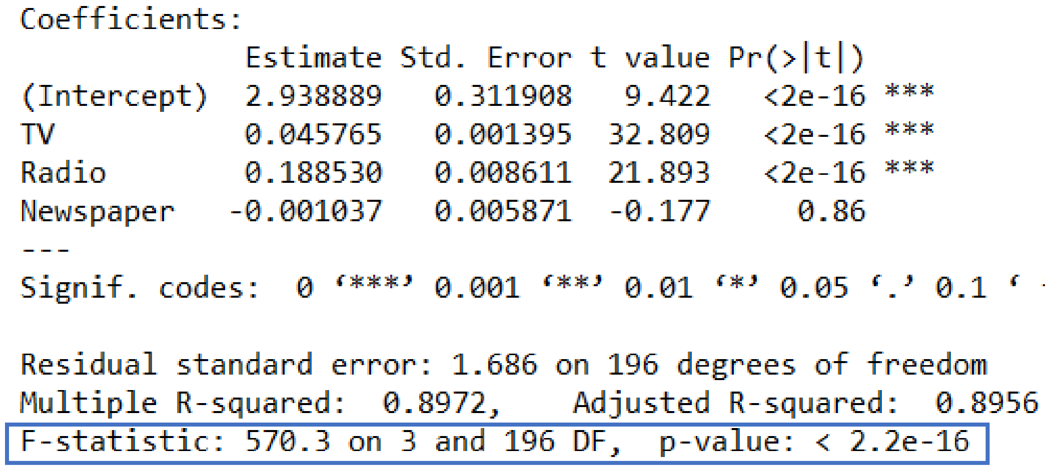
We can answer this question by fitting a multiple linear regression of ***Sales*** using *TV*, *Radio*, and *Newspaper* as predictor variables and testing the following hypothesis:

$$H_0\colon \beta_1 = \beta_2 = \beta_3 = 0 \quad \textbf{vs} \quad H_{a:} \text{ At least one } \beta_j \text{ is non} - \text{zero}$$

From the output on the right, we have an F statistic of **570.3** with a highly significant *p-value*.

This provides strong evidence ***against*** the null hypothesis, and we conclude that *Sales* revenue is associated with ***at least one advertising type***.

$$Sales = \beta_0 + \beta_1 TV + \beta_2 Radio + \beta_3 Newspaper$$



```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  2.938889   0.311908   9.422   <2e-16 ***
TV           0.045765   0.001395  32.809   <2e-16 ***
Radio        0.188530   0.008611  21.893   <2e-16 ***
Newspaper   -0.001037   0.005871  -0.177     0.86    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘

Residual standard error: 1.686 on 196 degrees of freedom
Multiple R-squared:  0.8972,	Adjusted R-squared:  0.8956 
F-statistic: 570.3 on 3 and 196 DF,  p-value: < 2.2e-16
```

Multiple Linear Regression

Advertising Data Analysis – Putting It All Together

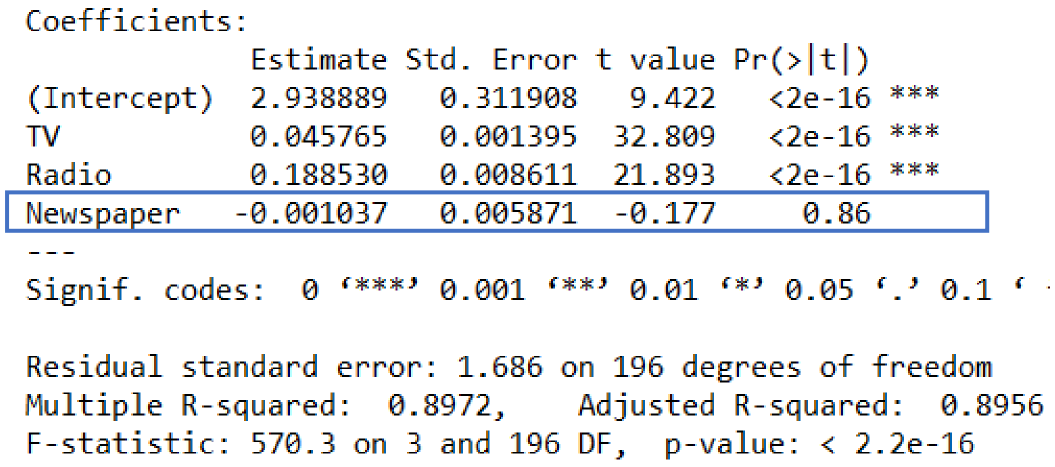
Which advertising media contribute to sales revenue?

To answer this question, we must examine the **partial F-test results** in the summary output. We find that the coefficient of Newspaper is not statistically significant. This corresponds to the following hypothesis:

$$H_0\text{: } \beta_3 = 0 \ \ \textbf{vs} \ \ H_{a:} \ \beta_3 \neq 0$$

Interpretation: ***given that I am using TV and Radio*** as predictors, *Newspaper* does not provide increased accuracy to the multiple linear regression model. This suggests that *TV* and *Radio* are the primary drivers of *Sales* revenue

$$Sales = \beta_0 + \beta_1 TV + \beta_2 Radio + \beta_3 Newspaper$$



```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  2.938889   0.311908   9.422   <2e-16 ***
TV           0.045765   0.001395  32.809   <2e-16 ***
Radio        0.188530   0.008611  21.893   <2e-16 ***
Newspaper   -0.001037   0.005871  -0.177     0.86    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘

Residual standard error: 1.686 on 196 degrees of freedom
Multiple R-squared:  0.8972,    Adjusted R-squared:  0.8956 
F-statistic: 570.3 on 3 and 196 DF,  p-value: < 2.2e-16
```

Multiple Linear Regression

Advertising Data Analysis – Putting It All Together

How large is the effect of each advertising type on Sales revenue?

The estimated coefficients associated with *TV* and *Radio* advertising budgets are:

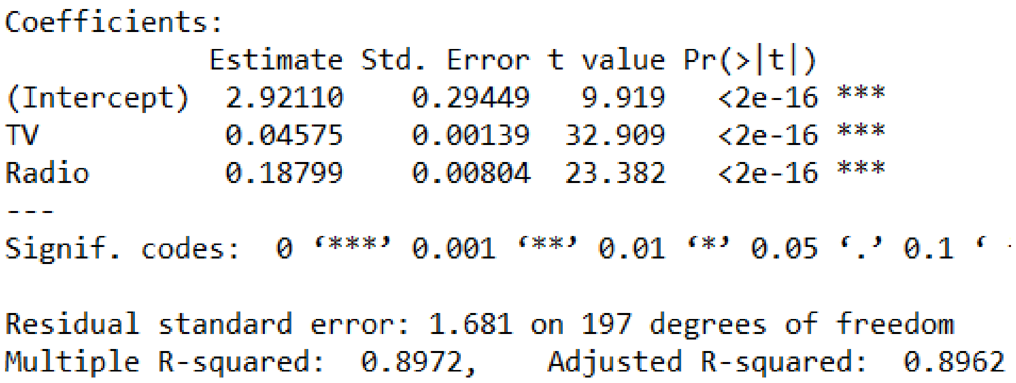
0.045 and 0.19 [*remember that all units in the Advertising data set are in **thousands***]

Interpretation: For a $1,000 increase in *TV* advertising budget, we estimate that the increase in ***average*** *Sales* revenue will be ***$45 for a fixed budget of Radio***

For a $1,000 increase in *Radio* advertising budget, we estimate that the increase in ***average*** *Sales* revenue will be ***$190, for a fixed budget of TV***

Overall, the effect of *Radio* advertising on average *Sales* is nearly **5 times greater** than that of *TV* advertising

$$Sales = \beta_0 + \beta_1 TV + \beta_2 Radio$$



```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  2.92110    0.29449   9.919   <2e-16 ***
TV           0.04575    0.00139  32.909   <2e-16 ***
Radio        0.18799    0.00804  23.382   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ 

Residual standard error: 1.681 on 197 degrees of freedom
Multiple R-squared:  0.8972,	Adjusted R-squared:  0.8962
```

Multiple Linear Regression

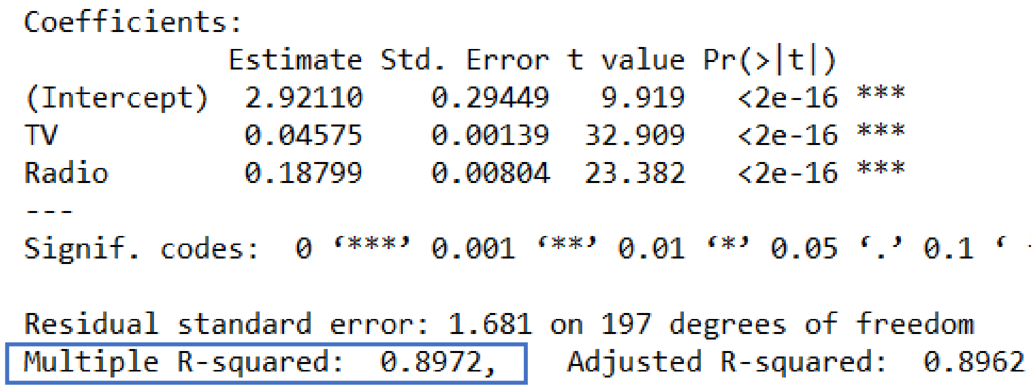
Advertising Data Analysis – Putting It All Together

How strong is the relationship between sales revenue and advertising budgets for TV and Radio?

The R^2 of the multiple linear regression on the right is **0.90** when rounded to 2 decimal places

Interpretation: *TV* and *Radio* advertising budgets explain approximately 90% of the total variance in *Sales* revenue, indicating a strong relationship between *Sales* revenue and advertising budgets

$$Sales = \beta_0 + \beta_1 TV + \beta_2 Radio$$



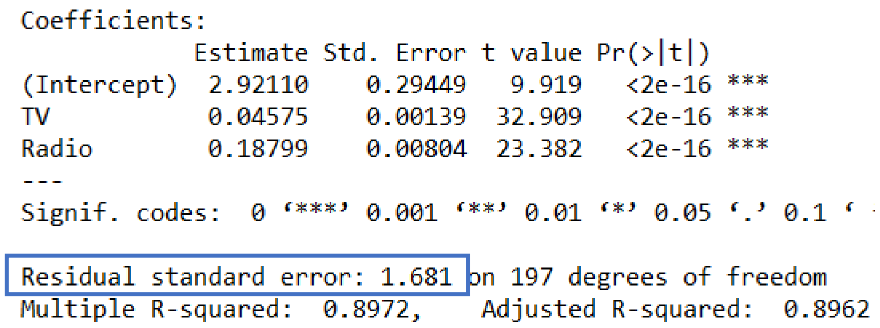
```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  2.92110    0.29449   9.919   <2e-16 ***
TV           0.04575    0.00139  32.909   <2e-16 ***
Radio        0.18799    0.00804  23.382   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘

Residual standard error: 1.681 on 197 degrees of freedom
Multiple R-squared:  0.8972,	Adjusted R-squared:  0.8962
```

Multiple Linear Regression

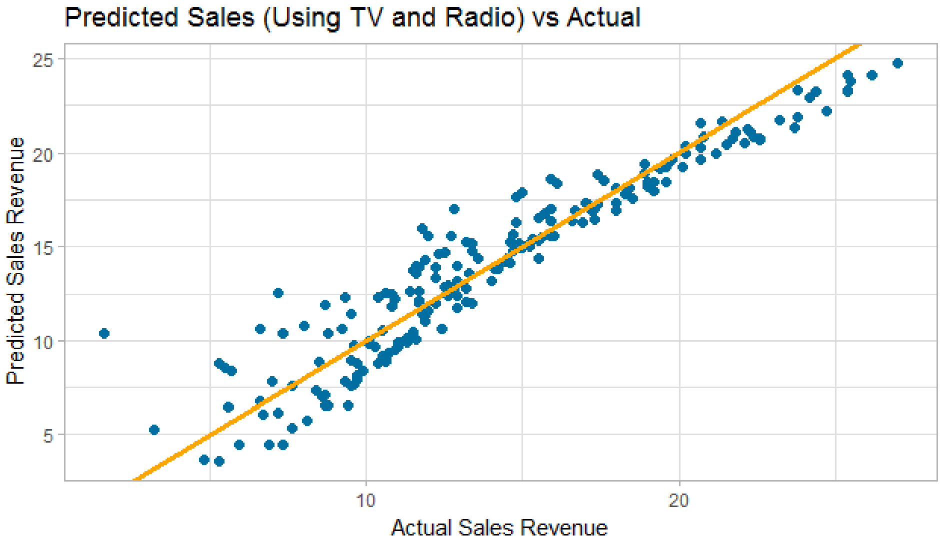
Advertising Data Analysis – Putting It All Together

$$Sales = \beta_0 + \beta_1 TV + \beta_2 Radio$$



```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  2.92110    0.29449   9.919   <2e-16 ***
TV           0.04575    0.00139  32.909   <2e-16 ***
Radio        0.18799    0.00804  23.382   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘

Residual standard error: 1.681 on 197 degrees of freedom
Multiple R-squared:  0.8972,	Adjusted R-squared:  0.8962
```



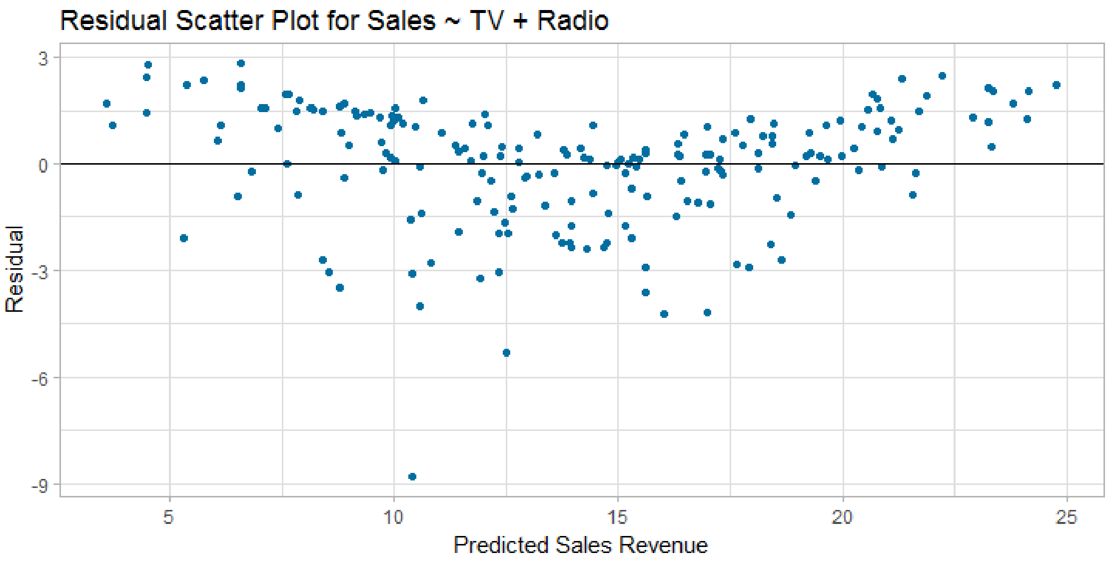
How accurately can we predict future Sales revenue?

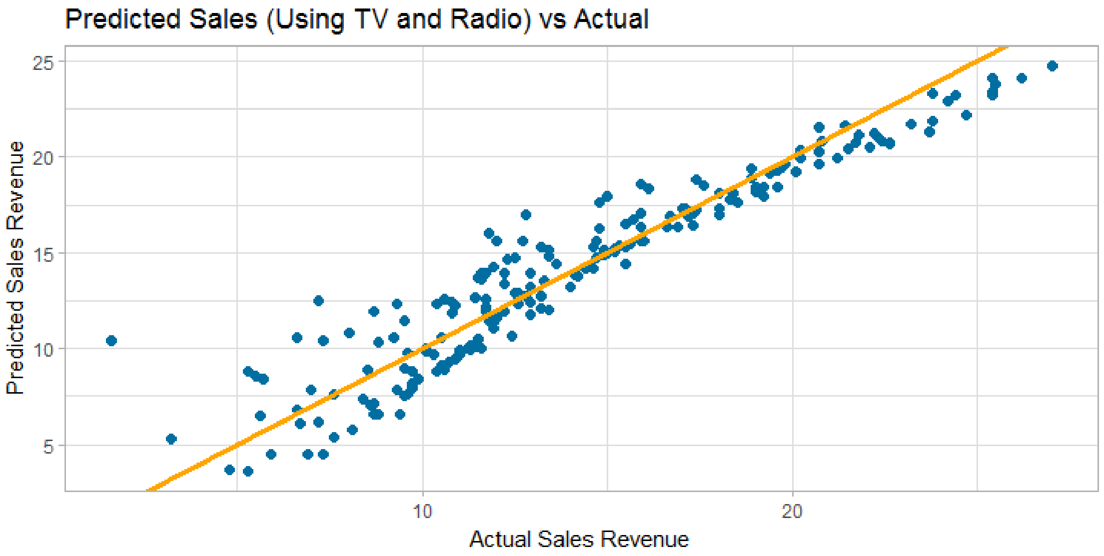
The residual standard error (**RSE**) for this model is 1.681. As a proportion of the average *Sales* revenue in the data (***14.02***), the RSE represents approximately 12%.

Interpretation: Roughly speaking, we can expect *12% prediction error, on average*. We also note from the visualization of R^2 that the highest prediction accuracy occurs for *Sales* values between approximately $14,000 and $22,000

Multiple Linear Regression

Advertising Data Analysis – Putting It All Together





Is the relationship between average Sales revenue and advertising budgets linear?

We see a ***non-linear*** relationship in both the residual plot and the R^2 visualization of predicted *Sales* versus actual *Sales*. The non-linearity mainly occurs at the lower and upper bounds of *Sales* revenue.

We conclude that the true relationship between sales revenue and advertising budgets may not be perfectly linear, however, the multiple linear regression model provides a reasonable approximation that is easy to interpret. To correct for non-linearity, we could try transforming the response and predictor variables as well as adding quadratic terms to the model.

Machine Learning Methods

Supervised Learning - Classification

Classification

Supervised learning methods used to predict *categorical* response variables

Example

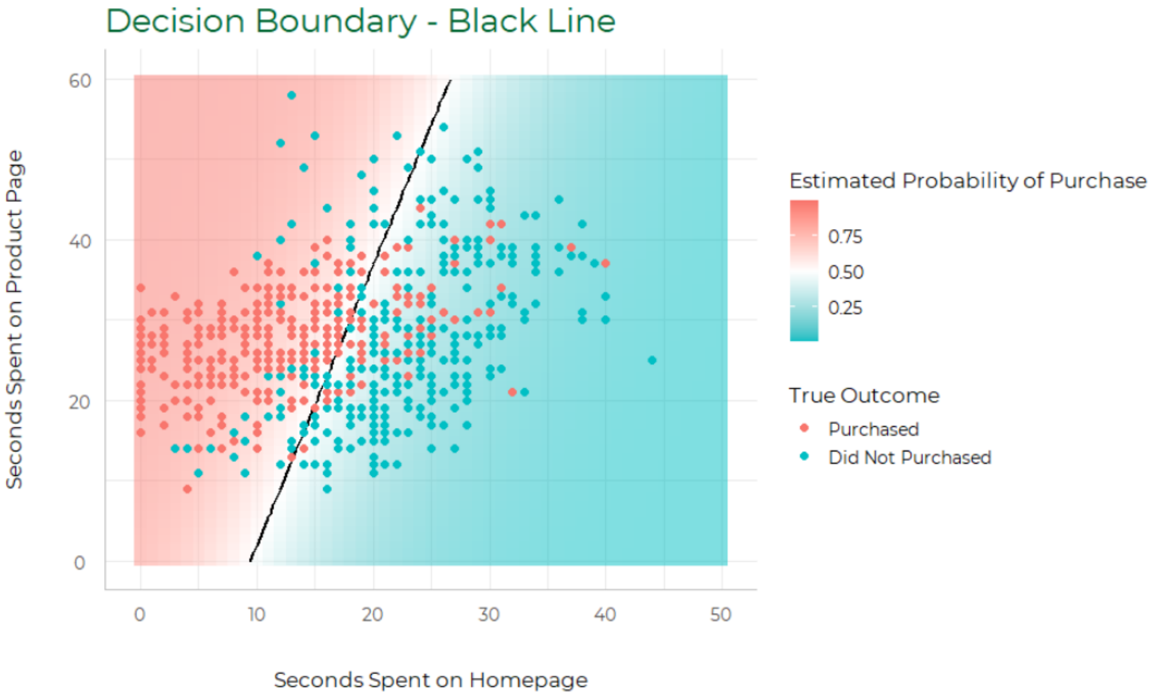
- Predict whether a customer will purchase a product based on the seconds they have spent browsing a company's homepage and product page

Response

Predictors

Outcome	Seconds Homepage	Seconds Product Page
Did Not Purchase	4	30
Purchased	32	43
Did Not Purchase	2	22
Purchased	24	36

Segmenting the predictor values into distinct, non-overlapping regions to predict a category



Classification

Predicting Categorical Outcomes

An example of classification would include predicting whether a patient will develop heart disease (Yes/No) using the Heart Disease data set on the right

- There are many classification techniques, or *classifiers*, that can be used to predict categorical response variables
- This lecture will focus on ***logistic regression***
  - Logistic regression is used to predict **dichotomous** response variables – these are categorical variables with two levels
  - The *HeartDisease* variable on the right is dichotomous

heart_disease	age	chest_pain	resting_bp	cholesterol
No	63	typical	145	233
Yes	67	asymptomatic	160	286
Yes	67	asymptomatic	120	229
No	37	nonanginal	130	250
No	41	nontypical	130	204

Logistic Regression
The Bernoulli Distribution

The ***Bernoulli*** distribution plays an important role in *logistic regression*

A Bernoulli random variable can be used to model the probabilistic behavior of dichotomous sample spaces

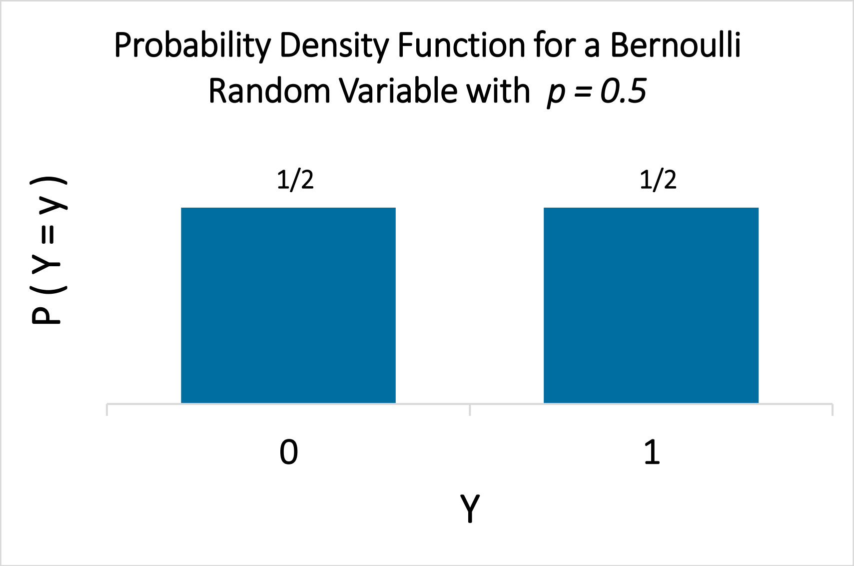
- An example would be tossing a fair coin
- Here $S = \{Heads, Tails\}$

Generally, the *event of interest is mapped to 1* (called the **Positive class** in classification) and the other event is mapped to 0

- If we were interested in studying whether a coin lands on "Heads", then we would define a Bernoulli random variable **Y** that equals 1 when the outcome is "Heads" and 0 otherwise

The Bernoulli distribution is indexed by a parameter **p**, which represents $P(Y = 1)$, and has the following probability function:

$$p^y(1 - p)^{(1-y)}$$



Logistic Regression

The Logistic Regression Setting

In *logistic regression*, we are predicting a **dichotomous response variable Y**

- We map the event of interest to Y = 1 and call it the **Positive** class
- The event corresponding to Y = 0 is the **Negative** class

We assume that each individual observation of **Y** follows a **Bernoulli distribution**

Given predictor variables $X_1, X_2, \dots, X_p$, we assume that

$$E\left(Y_i \middle| X_1 = x_1, \dots, X_p = x_p\right) = p_i$$

heart_disease	age	chest_pain	resting_bp	cholesterol
No	63	typical	145	233
Yes	67	asymptomatic	160	286
Yes	67	asymptomatic	120	229
No	37	nonanginal	130	250
No	41	nontypical	130	204

Logistic Regression

The Logistic Regression Setting

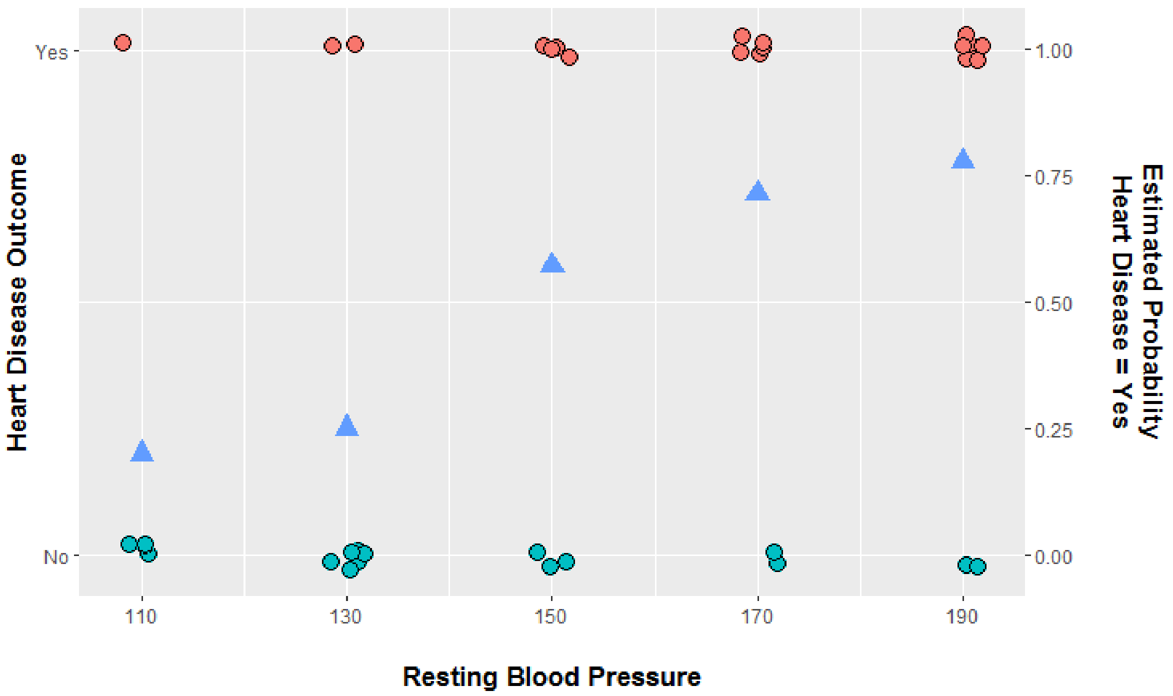
$$E(Y_i | X_1 = x_1, \ldots, X_p = x_p) = p_i$$

In most textbooks, the above is denoted as $p(x)$ and represents the probability that Y = 1 given the values of the predictor variable(s)

In *logistic regression*, we are modeling the relationship between $p(x)$ and the predictor variable values

We are interested in estimating $p(x)$ as a continuous function of the predictor variable values

Resting Blood Pressure	Heart Disease Yes	Heart Disease No	Estimated Probability of (Heart Disease = Yes)
110	1	4	0.20
130	2	6	0.25
150	4	3	0.57
170	5	2	0.71
190	7	2	0.78



Logistic Regression

Why Not Linear Regression?

How should we estimate $p(x)$?

For the case of one predictor variable X, why not use linear regression? This would be represented by the following model

$$E(Y|X = x) = p(x) = \beta_0 + \beta_1 x + \varepsilon$$

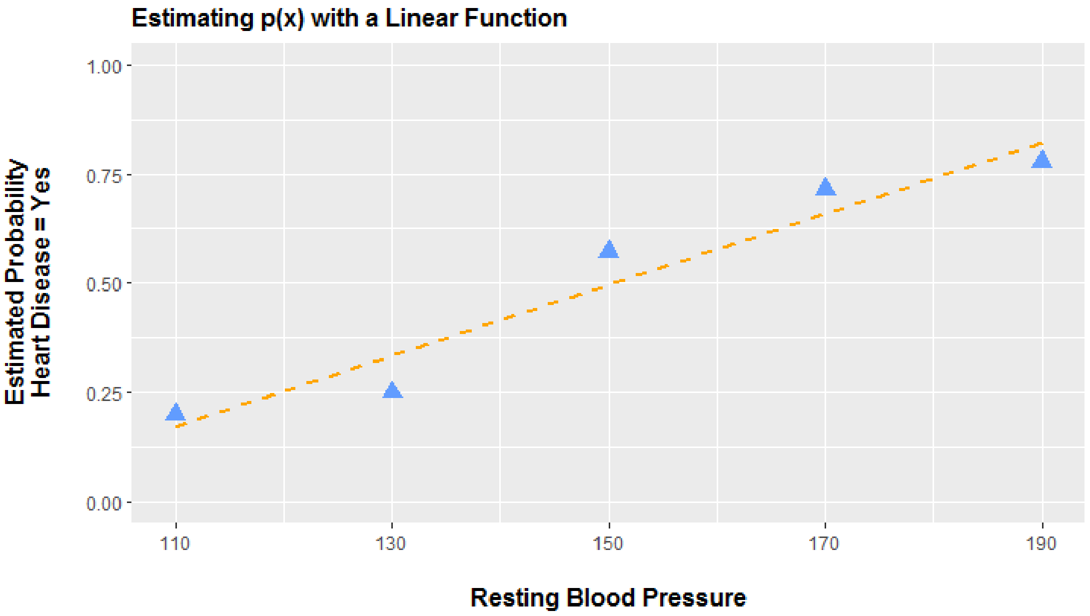
For our example on the right, this gives us an estimated regression line of

$$p(x) = -0.71 + 0.008(Resting\ Blood\ Pressure)$$

Resting Blood Pressure	Heart Disease Yes	Heart Disease No	Estimated Probability of (Heart Disease = Yes)
110	1	4	0.20
130	2	6	0.25
150	4	3	0.57
170	5	2	0.71
190	7	2	0.78

Problems with this model

- For resting blood pressure of 70, the estimated probability that a patient will develop heart disease is -0.15
- In linear regression the ε are assumed to have the same common variance, but by the properties of the Bernoulli distribution make this impossible



Logistic Regression

The Logistic Function

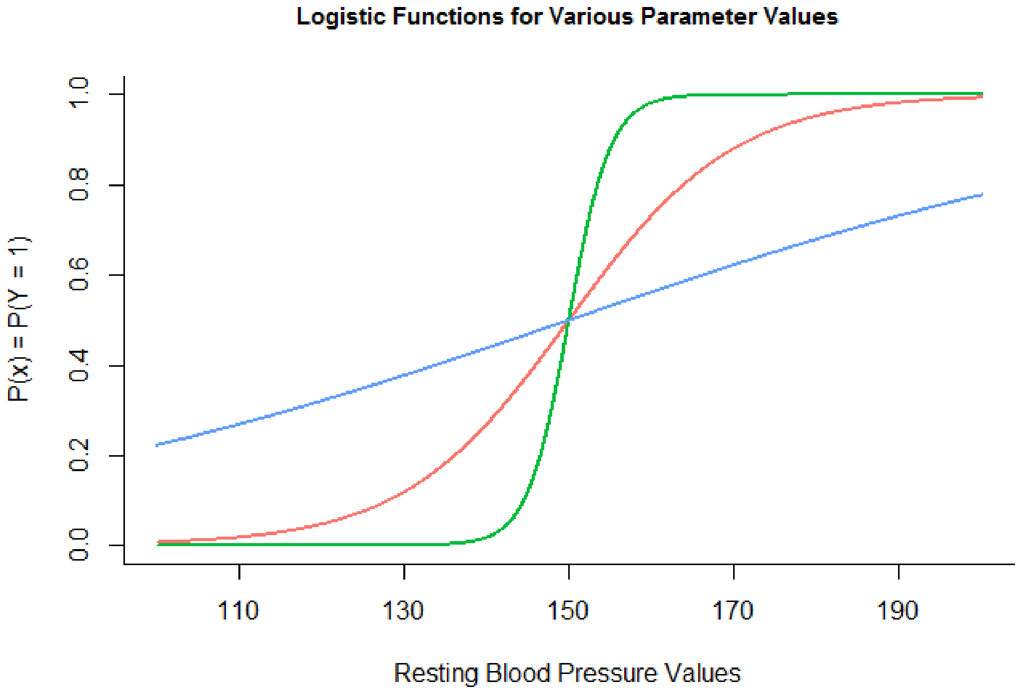
To avoid the problems we encountered on the previous slide, we must model $p(x)$ using a function that gives outputs between 0 and 1

In logistic regression, we use the ***logistic function***. For the case of one predictor variable X, the logistic function takes the form below

$$p(x) = \frac{e^{(\beta_0+\beta_1 x)}}{1 + e^{(\beta_0+\beta_1 x)}}$$

Three logistic curves are plotted to the right, using various values of β_0 and β_1

Estimating $p(x) = P(Y = 1|X = x)$ with the logistic function is a good choice since the logistic curve can take various shapes, from almost linear to extremely "S" shaped



Logistic Regression

The *logit* Transformation

Note that our estimate of $p(x) = \frac{e^{(\beta_0+\beta_1 x)}}{1+ e^{(\beta_0+\beta_1 x)}}$ is not a linear function of the predictor variable X

However, using a ***logit transformation***, we can transform both sides of the equation to get a linear function of the predictor variable X

$$logit\big(p(x)\big) = \log\left(\frac{p(x)}{1 - p(x)}\right) = \boldsymbol{\beta_0} + \boldsymbol{\beta_1 x}$$

Logistic Regression

The *logit* Transformation

Once we have our estimated coefficients, we can obtain an estimated probability for the Positive class for any predictor value, **x**, with:

$$p(x) = \frac{e^{(\beta_0+\beta_1 x)}}{1+ e^{(\beta_0+\beta_1 x)}}$$

How do we predict the response categories?

- If our estimated probability for a given **x** is greater than or equal to 0.5
  - We predict the Positive class
  - Negative class otherwise

GEORGE MASON UNIVERSITY
School of Business

Logistic Regression
Multiple Logistic Regression

The logistic regression model can easily be extended to incorporate multiple predictor variables

Just like in the multiple regression setting, predictors can be quantitative or categorical

In this case

$$p(x) = \frac{e^{(\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_p X_p)}}{1+\ e^{(\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_p X_p)}}$$

and

$$logit(p(x)) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$$

Confusion Matrix

Evaluating Prediction Accuracy

A **confusion matrix** can be created for any classifier that is used to predict a dichotomous response variable

The positive class is associated with the level of the response variable representing our event of interest

In the Heart Disease data set, a positive (**+**) is associated with the "Yes" outcome for the ***heart_disease*** variable

Actual

Predicted

	+	-	Row Total
+	TP	FP	P*
-	FN	TN	N*
Column Total	P	N	

Key Performance Measures:

Metric	Meaning
True Positive (TP)	Predicted Positive – Truth is Positive
True Negative (TN)	Predicted Negative – Truth is Negative
False Positive (FP)	Predicted Positive – Truth is Negative
False Negative (FN)	Predicted Negative – Truth is Positive

Confusion Matrix

An Example Using the Heart Disease Data Set

```
conf_mat(test_results, truth = heart_disease, estimate = .pred_class)
```

```
          Truth
Prediction yes no
       yes  26  9
       no    8 32
```

Interpretation

Overall, 58 patients (77%) were correctly classified. We predicted that 8 patients would not develop heart disease when in fact they did develop heart disease (*false negative*).

		Actual		
		+	-	Row Total
Predicted	+	26	9	35
	-	8	32	40
	Column Total	34	41	

Performance Metrics

Precision, Recall, and F1 Score

Instead of having to look at both false positive and false negative rates, the F_1 score combines both metrics into one overall score

- The F_1 score gives ***equal weight*** to false positive and false negative rates
    - Precision – function of false positives
    - Recall – function of false negatives
- The F_1 score ranges from 0 (worst) to 1 (best)

Precision

$$\frac{TP}{TP+FP} = \frac{26}{26+9} = 0.74$$

Recall

$$\frac{TP}{TP+FN} = \frac{26}{26+8} = 0.76$$

F_1 Score

$$2\left(\frac{PR}{P+R}\right) = 2 * \frac{(0.74)(0.76)}{0.74+0.76} = 0.75$$

		Actual		
		+	-	Row Total
Predicted	+	26	9	35
	-	8	32	40
	Column Total	34	41	

Performance Metrics

Sensitivity and Specificity

There are two other common performance metrics

- **Sensitivity** – function of false negatives
  - Same as recall
  - Values near 1 are optimal
  - The proportion of actual positives that are correctly identified
- **Specificity** – function of false positives
  - Values near 1 are optimal
  - The proportion of actual negatives that are correctly identified

Sensitivity

$$\frac{TP}{TP+FN} = \frac{26}{26+8} = 0.76$$

Specificity

$$\frac{TN}{TN+FP} = \frac{32}{32+9} = 0.78$$

		Actual		
		+	-	Row Total
Predicted	+	26	9	35
	-	8	32	40
	Column Total	34	41	

Performance Metrics

ROC Curves and Area Under the ROC Curve

- ROC curve plots the **sensitivity** on the y-axis and (**1 – specificity**) on the x-axis for all possible probability cut-off values.
- The default probability cut-off value used by classification models is 0.5
  - Changing this can guard against either false positives or false negatives. The ROC curve plots all this information in one plot

What to look for

- The best ROC curve is as close as possible to the point (0, 1) that is at the top left corner of the plot. The closer the ROC curve is to that point throughout the entire range, the better the classification model

Area Under the ROC Curve (AUC)

Another common performance metric. Can be interpreted as a letter grade for model performance:

- 0.9 – 1 = A
- 0.8 – 0.89 = B
- 0.7 – 0.79 = C
- 0.6 – 0.69 = D
- Below 0.6 = F