Business Statistics: 41000

Section 4: Predictive Analytics
Week 7: Multiple and Logistic Regression

Nick Polson
The University of Chicago Booth School of Business

http://faculty.chicagobooth.edu/nicholas.polson/teaching/41000/

Suggested Reading
OpenIntro Statistics, Chapter 8
Multiple Regression

Many problems involve more than one independent (explanatory) variable or factor which affects the dependent or response variable.

- **Multi-factor asset pricing models (APT).** Stock returns, book-to-market ratios, Interest rates
- **Demand** for a product given prices of competing brands, advertising, household attributes (to formulate pricing strategies)
- **Internet Analytics** What do I like? Suggestions instead of Search! Alexa “book my Xmas vacation,” “buy my best friend a birthday present”
### R Regression Commands

Given input-output vectors $x$ and $y$ `cor( ... )` computes correlation table

```r
model = lm(y ~ x) for linear model (a.k.a regression)
model = glm(y ~ x) for logistic regression
model = lm(y ~ x1+ ... + xp) for linear multiple regression model
```

```r
plot(model) diagnostics
plot(cooks.distance(model)) influential points
rstudent(model) outliers
summary(model) provides a summary analysis of our model
```

```r
newdata = data.frame( ... ) constructs a new input variable
predict.lm(model,newdata) provides a prediction at a new input
```

Regression in Excel

`linest(yrange,xrange)` and `slope(yrange,xrange)`
Regression Model

\[ Y = \text{response or outcome variable} \quad X_1, \ldots, X_p = \text{explanatory or input variable} \]

The general relationship is given by

\[ Y = f(X_1, \ldots, X_p) + e \]

And a linear relationship is written

\[ Y = b_0 + b_1 X_1 + b_2 X_2 + \ldots + b_p X_p + e \]
MLR Assumptions

The Multiple Linear Regression (MLR) model

\[ Y = b_0 + b_1 X_1 + b_2 X_2 + \ldots + b_p X_p + e \]

assumptions follow those of simple linear regression:

1. The conditional mean of \( Y \) is linear in the \( X_j \) variables

2. The errors are normal \( N(0, \sigma^2) \). We write

\[ Y \mid X_1, \ldots, X_p \sim N\left(b_0 + b_1 X_1 + \ldots + b_p X_p, \sigma^2\right) \]
When looking at the $\beta$ coefficients there are two issues

1. **Statistical Significance**: The $t$-ratios of the $\beta$’s

2. **Economic Significance**: The magnitudes of the $\beta$’s If $X_i$ increases by one unit **holding the other $X$’s constant** Then $Y$ will react by $\beta_i$ units. They are called **marginal effects**

At the end of the day use your judgment!
plot(model) provides diagnostics before model building!

There are many possible caveats

1. Running simple regressions gives you the wrong answer!

2. Multiple regression takes into account the correlation between the factors and the independent variable. It does all the work for you.

3. A variable might be insignificant once we have incorporated a more important predictor variable.

A common sense approach usually works well. If a variable never seems to be significant it typically isn’t. Model Prediction is the great equalizer!!
Models with Interactions

In many situations, $X_1$ and $X_2$ interact when predicting $Y$

**Interaction Model:** run the regression

$$Y = a + b_1 X_1 + b_2 X_2 + b_3 X_1 X_2 + \epsilon$$

In R: `model = lm(y ~ x1 * x2)` gives $X_1 + X_2 + X_1 X_2$

In R: `model = lm(y ~ x1 : x2)` gives only $X_1 X_2$

The coefficients $b_1$ and $b_2$ are marginal effects **IF** $b_3$ is significant there’s an interaction effect. We leave $b_1$ and $b_2$ in the model whether they are significant or not.
Models with Interactions and Dummies

$X_1$ and $D$ dummy

- $X_2 = D$ is a dummy variable with values of zero or one.
- **Model**: typically we run a regression of the form

$$Y = a + b_1 X_1 + b_2 X_1 \times D + \epsilon$$

- The coefficient $b_1 + b_2$ is the effect of $X_1$ when $D = 1$. The coefficient $b_1$ is the effect when $D = 0$. 
Example: Newfood Data

Goal of Experiment

- A six month market test has been performed on the Newfood product. A breakfast cereal.
- Build a multiple regression model that gives us good sales forecasts.
- This dataset is the outcome of a controlled experiment in which the values of the independent variables which affect sales are chosen by the analyst.
Example: Newfood Data

Analyses the factors which contribute to sales of a new breakfast cereal. Quantify the effects of business decisions such as choice of advertising level, location in store and pricing.

<table>
<thead>
<tr>
<th>variable</th>
<th>description</th>
</tr>
</thead>
<tbody>
<tr>
<td>sales</td>
<td>new cereal sales</td>
</tr>
<tr>
<td>price</td>
<td>price</td>
</tr>
<tr>
<td>adv</td>
<td>low or high advertising (0 or 1)</td>
</tr>
<tr>
<td>locat</td>
<td>bread or breakfast section (0 or 1)</td>
</tr>
<tr>
<td>inc</td>
<td>neighborhood income</td>
</tr>
<tr>
<td>svol</td>
<td>size of store</td>
</tr>
</tbody>
</table>
Example: Newfood

1. What happens when you regress sales on price, adv, locat?

2. Run the “kitchen-sink” regression. Perform Diagnostic checks.

3. Which variables should we transform?

4. Run the new model. Perform diagnostics and variable selection.

5. What’s the largest cooks distance?

6. Provide a summary of coefficients and statistical significance

7. Predict sales when price = 30, adv = 1, income = 8 and svol = 34.

   What happens when you predict at the median values of the characteristics?
Example: Newfood

First we examine the correlation matrix:

<table>
<thead>
<tr>
<th></th>
<th>sales</th>
<th>price</th>
<th>adv</th>
<th>locat</th>
<th>income</th>
</tr>
</thead>
<tbody>
<tr>
<td>price</td>
<td>-0.658</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>adv</td>
<td>0.001</td>
<td>0.000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>locat</td>
<td>-0.001</td>
<td>0.000</td>
<td>0.000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>income</td>
<td>0.163</td>
<td>-0.131</td>
<td>-0.746</td>
<td>0.000</td>
<td></td>
</tr>
<tr>
<td>svol</td>
<td>0.375</td>
<td>-0.179</td>
<td>-0.742</td>
<td>-0.040</td>
<td>0.809</td>
</tr>
</tbody>
</table>

Remember: correlations are not $b$'s!!
Total sales volume is negatively correlated to advertising.

Income is negatively correlated with advertising as well.

How is the negative correlation apt to affect the advertising effects?

<table>
<thead>
<tr>
<th></th>
<th>sales</th>
<th>price</th>
<th>adv</th>
</tr>
</thead>
<tbody>
<tr>
<td>price</td>
<td>-0.658</td>
<td></td>
<td></td>
</tr>
<tr>
<td>adv</td>
<td>0.001</td>
<td>0.000</td>
<td></td>
</tr>
<tr>
<td>locat</td>
<td>-0.001</td>
<td>0.000</td>
<td>0.000</td>
</tr>
</tbody>
</table>

There’s no correlation in the X’s by design!
Let’s start by only including price, adv, locat

\[ \text{sales} = 562 - 12.8 \text{ price} + 0.2 \text{ adv} - 0.2 \text{ locat} \]

Coefficients:

|                | Estimate | Std. Error | t value | P(>|t|) |
|----------------|----------|------------|---------|---------|
| (intercept)    | 562.31   | 53.14      | 10.58   | 0.000   |
| price          | -12.812  | 1.780      | -7.20   | 0.000   |
| adv            | 0.22     | 14.54      | 0.02    | 0.988   |
| locat          | -0.22    | 14.54      | -0.02   | 0.988   |

Why is the marketer likely to be upset by this regression?!

Why is the economist happy?

Let’s add income and svol to the regression!
Transformation

Power model: transform with log-log

\[ \log(\text{sales}) = 8.41 - 1.74 \log(\text{price}) + 0.150 \text{adv} + 0.0010 \text{locat} - 0.524 \log(\text{inc}) + 1.03 \log(\text{svol}) \]

Coefficients:

| Estimate  | Std. Error | t value | P(>|t|) |
|-----------|------------|---------|---------|
| (intercept) | 8.407 | 1.387 | 6.06 | 0.000 |
| logprice   | -1.7430 | 0.2207 | -7.90 | 0.000 |
| adv        | 0.1496 | 0.1005 | 1.49 | 0.141 |
| locat      | 0.00100 | 0.06088 | 0.02 | 0.987 |
| loginc     | -0.5241 | 0.4958 | -1.06 | 0.294 |
| logsvol    | 1.0308 | 0.2553 | 4.04 | 0.000 |

Why no logs for adv and locat variables?

The \( \log(\text{svol}) \) coefficient is close to one!

\[ R^2 = 60\% \]
On the transformed scale,

\[ \log \text{sales} = 8.41 - 1.74 \log \text{price} + 0.150 \text{adv} + 0.001 \text{locat} - 0.524 \log \text{inc} + 1.03 \log \text{svol} \]

On the un-transformed scale,

\[ \text{sales} = e^{8.41 (\text{price}) - 1.74 e^{0.15 \text{adv}} e^{0.001 \text{locat}} (\text{inc}) - 0.524 (\text{svol})^{1.03}} \]

sales/price, income and svol are a power sales/adv, locat are exponential
Interpret your regression model as follows

- **Price elasticity** is $\hat{\beta}_{\text{price}} = -1.74$. A 1% increase in price will drop sales by 1.74%.

- **adv = 1 increases sales** by a factor of $e^{0.15} = 1.16$. That's a 16% improvement.

**Variable Selection**: delete `locat` as it is *statistically insignificant.*
Prediction

`predict.lm` provides a $\hat{Y}$-prediction given a new $X_f$

```
# predict.lm at newdata
> predict.lm(modelnew,newdata,se.fit=T,interval="prediction")
$fit
   fit  lwr   upr
 1 5.259691 4.739762 5.77962
$se.fit
[1] 0.05560662
```

Exponentiate-back to find $sales = e^{5.2596} = 192.40$.

`newdata=data.frame(price=30,adv=1,income=8,svol=34)`
Example: Golf Performance Data

Dave Pelz has written two best-selling books for golfers, *Dave Pelz's Short Game Bible*, and *Dave Pelz's Putting Bible*.

▶ Dave Pelz was formerly a “rocket scientist” (literally) Data analytics helped him refine his analysis It’s the short-game that matters!

▶ The optimal speed for a putt Best chance to make the putt is one that will leave the ball 17 inches past the hole, if it misses.
Golf Data

Year-end performance data on 195 players from the 2000 PGA Tour.

1. $\text{nevents}$, the number of official PGA events included in the statistics
2. $\text{money}$, the official dollar winnings of the player
3. $\text{drivedist}$, the average number of yards driven on par 4 and par 5 holes
4. $\text{gir}$, greens in regulation, measured as the percentage of time that the first (tee) shot on a par 3 hole ends up on the green, or the second shot on a par 4 hole ends up on the green, or the third shot on a par 5 hole ends up on the green
5. $\text{avgputts}$, which is the average number of putts per round.

Analyze these data to see which of $\text{nevents}$, $\text{rivedist}$, $\text{gir}$, $\text{avgputts}$ is most important for winning money.
Regression of Money on all explanatory variables:

\[ \text{lm(formula = money} \sim \text{nevents} + \text{drivedist} + \text{gir} + \text{avgputts, data = d00)} \]

Coefficients:

| Estimate  | Std. Error | t value | Pr(>|t|) |
|-----------|------------|---------|----------|
| (Intercept) | 14856638 | 4206466 | 3.532 | 0.000518 *** |
| nevents   | -30066 | 11183 | -2.689 | 0.007815 ** |
| drivedist | 21310 | 6913 | 3.083 | 0.002358 ** |
| gir       | 120855 | 17429 | 6.934 | 6.22e-11 *** |
| avgputts  | -15203045 | 2000905 | -7.598 | 1.33e-12 *** |

\[ R^2 = 50\% \]
Residuals

Standardized Residuals Frequency

-2 0 2 4 6 8

0 10 20 30 40 50

Tiger Woods
Regression

Transform with $\log(Money)$ as it has much better residual diagnostic plots.

```r
lm(formula = log(money) ~ nevents + drivedist + gir + avgputts, data = d00)
```

Coefficients:

|                     | Estimate | Std. Error | t value | Pr(>|t|) |
|---------------------|----------|------------|---------|----------|
| (Intercept)         | 36.14928| 3.577630   | 10.104  | <2e-16   *** |
| nevents             | -0.00899| 0.009511   | -0.945  | 0.3459   |
| drivedist           | 0.014091| 0.005880   | 2.397   | 0.0175   * |
| gir                 | 0.165672| 0.014824   | 11.176  | <2e-16   *** |
| avgputts            | -21.128752| 1.701784  | -12.416 | <2e-16   *** |

$R^2 = 67\%$. There's still 33% of variation to go
Residuals for $\log(\text{Money})$

![Histogram of Residuals](image)

Tiger Woods
Regression

Variable selection: $t$-stats for nevents is $< 1.5$.

```
lm(formula = log(money) ~ drivedist + gir + avgputts, data = d00)
```

Residuals:

<table>
<thead>
<tr>
<th></th>
<th>Min</th>
<th>1Q</th>
<th>Median</th>
<th>3Q</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Residuals</td>
<td>-1.48002</td>
<td>-0.37038</td>
<td>0.00079</td>
<td>0.40227</td>
<td>1.96546</td>
</tr>
</tbody>
</table>

Coefficients:

|                     | Estimate | Std. Error | $t$ value | Pr(>|$t$|) |
|---------------------|----------|------------|-----------|----------|
| (Intercept)         | 36.17370 | 3.57653    | 10.114    | <2e-16 *** |
| drivedist           | 0.01463  | 0.00585    | 2.501     | 0.0132 *  |
| gir                 | 0.16577  | 0.01482    | 11.186    | <2e-16 *** |
| avgputts            | -21.36844 | 1.68230    | -12.702   | <2e-16 *** |

The fewer the putts the better golfer you are. Duh!

avgputts per round is hard to decrease by one!
Evaluating the Coefficients

1. **Greens in Regulation (GIR)** has a $\hat{\beta} = 0.17$. If I can increase my GIR by one, I'll earn $e^{0.17} = 1.18\%$ An extra 18%

2. **DriveDis** has a $\hat{\beta} = 0.014$. A 10 yard improvement, I'll earn $e^{0.014\times10} = e^{0.14} = 1.15\%$ An extra 15%

**Caveat:** Everyone has gotten better since 2000!
Summary

Tiger was 9 standard deviations better than the model.

- Taking logs of Money helps the residuals!
- An exponential model seems to fit well. The residual diagnostics look good
- The t-ratios for events are under 1.5.
**Over-Performers**

**Outliers:** biggest over and under-performers in terms of money winnings, compared with the performance statistics.

Woods, Mickelson, and Els won major championships by playing well when big money prizes were available.

<table>
<thead>
<tr>
<th>Name</th>
<th>Money</th>
<th>Predicted</th>
<th>Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tiger Woods</td>
<td>9,188,321</td>
<td>3,584,241</td>
<td>5,604,080</td>
</tr>
<tr>
<td>Phil Mickelson</td>
<td>4,746,457</td>
<td>2,302,171</td>
<td>2,444,286</td>
</tr>
<tr>
<td>Ernie Els</td>
<td>3,469,405</td>
<td>1,633,468</td>
<td>1,835,937</td>
</tr>
<tr>
<td>Hal Sutton</td>
<td>3,061,444</td>
<td>1,445,904</td>
<td>1,615,540</td>
</tr>
</tbody>
</table>
Under-Performers

Underperformers are given by large negative residuals Glasson and Stankowski should win more money.

<table>
<thead>
<tr>
<th>Name</th>
<th>Money</th>
<th>Predicted</th>
<th>Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kenny Perry</td>
<td>889,381</td>
<td>1,965,740</td>
<td>–1,076,359</td>
</tr>
<tr>
<td>Paul Stankowski</td>
<td>669,709</td>
<td>1,808,690</td>
<td>–1,138,981</td>
</tr>
<tr>
<td>Bill Glasson</td>
<td>552,795</td>
<td>1,711,530</td>
<td>–1,158,735</td>
</tr>
<tr>
<td>Jim McGovern</td>
<td>266,647</td>
<td>1,397,818</td>
<td>–1,131,171</td>
</tr>
</tbody>
</table>
Lets look at 2018 data

Highest earners are

<table>
<thead>
<tr>
<th>name</th>
<th>nevents</th>
<th>money</th>
<th>drivenist</th>
<th>gir</th>
<th>avgputts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Justin Thomas</td>
<td>23</td>
<td>8,694,821</td>
<td>311.800</td>
<td>68.770</td>
<td>1.714</td>
</tr>
<tr>
<td>Dustin Johnson</td>
<td>20</td>
<td>8,457,352</td>
<td>314</td>
<td>70.570</td>
<td>1.699</td>
</tr>
<tr>
<td>Justin Rose</td>
<td>18</td>
<td>8,130,678</td>
<td>303.500</td>
<td>69.950</td>
<td>1.732</td>
</tr>
<tr>
<td>Bryson DeChambeau</td>
<td>26</td>
<td>8,094,489</td>
<td>305.700</td>
<td>69.650</td>
<td>1.758</td>
</tr>
<tr>
<td>Brooks Koepka</td>
<td>17</td>
<td>7,094,047</td>
<td>313.400</td>
<td>68.280</td>
<td>1.747</td>
</tr>
<tr>
<td>Bubba Watson</td>
<td>24</td>
<td>5,793,748</td>
<td>313.100</td>
<td>68.210</td>
<td>1.773</td>
</tr>
</tbody>
</table>
### Overperformers

<table>
<thead>
<tr>
<th>name</th>
<th>money</th>
<th>Predicted</th>
<th>Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Justin Thomas</td>
<td>8,694,821</td>
<td>5,026,220</td>
<td>3,668,601</td>
</tr>
<tr>
<td>Dustin Johnson</td>
<td>8,457,352</td>
<td>6,126,775</td>
<td>2,330,577</td>
</tr>
<tr>
<td>Justin Rose</td>
<td>8,130,678</td>
<td>4,392,812</td>
<td>3,737,866</td>
</tr>
<tr>
<td>Bryson DeChambeau</td>
<td>8,094,489</td>
<td>3,250,898</td>
<td>4,843,591</td>
</tr>
<tr>
<td>Brooks Koepka</td>
<td>7,094,047</td>
<td>4,219,781</td>
<td>2,874,266</td>
</tr>
<tr>
<td>Bubba Watson</td>
<td>5,793,748</td>
<td>3,018,004</td>
<td>2,775,744</td>
</tr>
<tr>
<td>Webb Simpson</td>
<td>5,376,417</td>
<td>2,766,988</td>
<td>2,609,429</td>
</tr>
<tr>
<td>Francesco Molinari</td>
<td>5,065,842</td>
<td>2,634,466</td>
<td>2,431,376</td>
</tr>
<tr>
<td>Patrick Reed</td>
<td>5,006,267</td>
<td>2,038,455</td>
<td>2,967,812</td>
</tr>
<tr>
<td>Satoshi Kodaira</td>
<td>1,471,462</td>
<td>-1,141,085</td>
<td>2,612,547</td>
</tr>
</tbody>
</table>
## Underperformers

<table>
<thead>
<tr>
<th>name</th>
<th>money</th>
<th>Predicted</th>
<th>Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trey Mullinax</td>
<td>1,184,245</td>
<td>3,250,089</td>
<td>-2,065,844</td>
</tr>
<tr>
<td>J.T. Poston</td>
<td>940,661</td>
<td>3,241,369</td>
<td>-2,300,708</td>
</tr>
<tr>
<td>Tom Lovelady</td>
<td>700,783</td>
<td>2,755,854</td>
<td>-2,055,071</td>
</tr>
<tr>
<td>Michael Thompson</td>
<td>563,972</td>
<td>2,512,330</td>
<td>-1,948,358</td>
</tr>
<tr>
<td>Matt Jones</td>
<td>538,681</td>
<td>2,487,139</td>
<td>-1,948,458</td>
</tr>
<tr>
<td>Hunter Mahan</td>
<td>457,337</td>
<td>2,855,898</td>
<td>-2,398,561</td>
</tr>
<tr>
<td>Cameron Percy</td>
<td>387,612</td>
<td>3,021,278</td>
<td>-2,633,666</td>
</tr>
<tr>
<td>Ricky Barnes</td>
<td>340,591</td>
<td>3,053,262</td>
<td>-2,712,671</td>
</tr>
<tr>
<td>Brett Stegmaier</td>
<td>305,607</td>
<td>2,432,494</td>
<td>-2,126,887</td>
</tr>
</tbody>
</table>
Findings

Here’s three interesting effects:

- Tiger Woods is 8 standard deviations better!
- Increasing driving distance by 10 yards makes you 15% more money
- Increasing GIR by one makes you 18% more money.
- Detect Under- and Over-Performers

Go Play!!
Regression

1. Input and Plot Data In R: `plot` and `summary` commands

2. “Kitchen-Sink” Regression `lm` command with all variables

3. Residual Diagnostics and `plot(model)` Fitted values and Standardised residuals. Outliers and Influence

4. Transformation?
   
   Correct the 4-in-1 plots and assumptions.
Regression Strategy

1. Variable Selection $t$-state and $p$-values from `summary(model)`

2. Final Regression Re-run the model. Interpret the coefficients `summary(model)`. Economic and Statistical Significance

3. Prediction `predict.lm`. Out-of-sample forecasting A model is only as good as its predictions!!
Business Statistics: 41000

Week 8: Predictive Analytics
Logistic Regression

Nick Polson
The University of Chicago Booth School of Business

http://faculty.chicagobooth.edu/nicholas.polson/teaching/41000/
Predictive Analytics

General Introduction

Predictive Analytics is the most widely used tool for high dimensional input-output

\[ Y = F(X) \text{ where } X = (X_1, \ldots, X_p) \]

- Consumer Demand (Amazon, Airbnb, ...)
- Maps (Bing, Uber)
- Pricing
- Healthcare

The applications are endless ....
### Kaggle: Predictive Culture

#### 353 Competitions

<table>
<thead>
<tr>
<th>Competition</th>
<th>Prize</th>
<th>Teams</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Passenger Screening Algorithm Challenge</strong></td>
<td>$1,500,000</td>
<td>518 teams</td>
</tr>
<tr>
<td>Improve the accuracy of the Department of Homeland Security's threat recognition algorithms</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Featured</strong> &lt;br&gt; 2 years ago &lt;br&gt; terrorism, image data, object detection</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Zillow Prize: Zillow’s Home Value Prediction (Zestimate)</strong></td>
<td>$1,200,000</td>
<td>3,775 teams</td>
</tr>
<tr>
<td>Can you improve the algorithm that changed the world of real estate?</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Featured</strong> &lt;br&gt; 2 years ago &lt;br&gt; real estate, housing</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Data Science Bowl 2017</strong></td>
<td>$1,000,000</td>
<td>1,972 teams</td>
</tr>
<tr>
<td>Can you improve lung cancer detection?</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Featured</strong> &lt;br&gt; 2 years ago &lt;br&gt; healthcare, binary classification, image data</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Heritage Health Prize</strong></td>
<td>$500,000</td>
<td>1,351 teams</td>
</tr>
<tr>
<td>Identify patients who will be admitted to a hospital within the next year using historical claims data.</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Featured</strong> &lt;br&gt; 6 years ago</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Target and other retailers use predictive analytics to study consumer purchasing behaviour to see what type of coupons or promotions you might like.

Here’s a famous story about a father and his daughter. Target predicted that his daughter was pregnant from her purchasing behaviour long before they were buying diapers.

Here’s the original link...

Target and Pregnancy
Walmart began using predictive analytics in 2004. Mining trillions of bytes’ worth of sales data from recent hurricanes.

Determine what customers most want to purchase leading up to a storm.

Strawberry Pop-Tarts are one of the most purchased food items, especially after storms, as they require no heating and can be eaten at any meal.

Walmart and Hurricanes
Airbnb New User Bookings Prediction Competition

New users on Airbnb can book a place to stay in 34,000+ cities across 190+ countries.

Accurately predict where a new user will book their first travel experience.

Airbnb can then personalized content, decrease the average time to first booking, and better forecast demand.

12 classes—major destinations, and a did not book category.
Airbnb

List of users, demographics, web session records, and content data

Winner has the best out-of-sample prediction!!
Germany’s Otto

Predicting Consumer Demand

Customers were less likely to return merchandise if it arrived within two days.

If an item takes longer to arrive, it gives customers more time to spot the product in a shop for less money and buy it, forcing Otto to forgo the sale and eat the shipping costs.

While customers are less likely to return merchandise that arrives quickly, also they prefer to receive everything at once.
As the number and kinds of available experiences on TripAdvisor grew rapidly in the last couple of years, a personalized website can increase user satisfaction significantly by providing travelers with an easy way to find experiences that are relevant to them.

▶ Approach

1. Training Data Collection
2. Entity Embeddings
3. RFY Model Architecture

▶ Result Analysis

1. Offline Evaluation
2. Online A/B test
Germany’s Otto

Otto sells other brands, does not stock those goods itself, hard to avoid one of the two evils: shipping delays until all the orders are ready for fulfilment, or lots of boxes arriving at different times.

- Analyze around 3 billion past transactions and 200 variables—past sales, searches on Otto’s site and weather information. They predict what customers will buy a week before they order. This system has proved so reliable, predicting with 90% accuracy what will be sold within 30 days, that Otto allows it automatically to purchase around 200,000 items a month from third-party brands with no human intervention.

Economist

Germany’s Otto
Stitch Fix asks customers for insights and feedback alongside their size and color preference for items, even the ones customers didn’t like or buy, in exchange for a clear value proposition.

The breadth and depth of their data are valuable.

Their model relies on a combination of data science – machine learning, AI and natural language processing – and human stylists; on top of complex customer profiles built by data, stylists can layer the nuances of buying and wearing clothes.
Bayes predicts where you’re going to be dropped off.

Uber constructs prior probabilities for riders, Uber cars, and popular places.

Combine to construct a joint probability table

Then calculate the posterior probability of destination for each person and pool travellers together
Logistic Regression: Classification

When the $Y$ we are trying to predict is *categorical* (or *qualitative*) we say that we have a *classification* problem.

For a numeric (or *quantitative*) $Y$ we predict its value.

For a binary output we predict the probability its going to happen

$$p(Y = 1 \mid X = x)$$

where $X$ is our usual list of predictors, $X_1, \ldots, X_p$
Logistic Regression

Suppose that we have a binary response, $Y$ taking the value 0 or 1

- Win or lose
- Sick or healthy
- Buy or not buy
- Pay or default

The goal is to predict the probability that $Y$ equals 1

You can then do classification and categorize a new data-point
Example: Default Data

Here’s a typical problem

Assessing credit risk and default data ...

- **Y**: whether or not a customer defaults on their credit card (No or Yes)
- **X**: The average balance that customer has remaining on their credit card after making their monthly payment.

... plus as many other features you think might predict **Y** ...
Logistic Regression

$Y$ is an indicator: $Y = 0$ or $1$.

$X$ is our usual set of predictors/covariates

We need to model the probability that $Y = 1$ as

$$p(Y = 1 \mid X_1, \ldots, X_p) = f(b_1 X_1 + \ldots + b_p X_p)$$

where $f$ is increasing and $0 < f(X) < 1$ The logit-transform is given by

$$f(x) = e^x / (1 + e^x)$$
Logistic Regression

The logistic regression model is linear in log-odds

$$\log \left( \frac{p(Y = 1 | X)}{1 - p(Y = 1 | X)} \right) = b_0 + b_1 X_1 + \ldots + b_p X_p$$

These model are easy to fit in R:

```r
 glm(Y ~ X1 + X2, family = binomial)
```

▶ “g” is for generalized; binomial indicates $Y = 0$ or 1

▶ “glm” has a bunch of other options.
Example: NBA point spread

Does the Vegas point spread predict whether the favourite wins or not?

Histogram of the data by group:

<table>
<thead>
<tr>
<th>spread</th>
<th>favwin=1</th>
<th>favwin=0</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>20</td>
<td>40</td>
</tr>
<tr>
<td>120</td>
<td>80</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>20</td>
<td>40</td>
</tr>
</tbody>
</table>

spread
favwin
R: Logistic Regression

In R: the output gives us ...

```
nbareg = glm(favwin~spread-1, family=binomial)
summary(nbareg)
Call:
  glm(formula = favwin ~ spread - 1, family = binomial)
Coefficients:
  Estimate  Std. Error   z value     Pr(>|z|)  
 spread  0.15600     0.01377     11.33   <2e-16 ***
```
# prediction
```
newweek=c(8,4)
```

The $\beta$ measures how our log-odds change! $\beta = 0.156$
“Plug-in” the values for the new game into our logistic regression

\[ P(\text{favwin} \mid \text{spread}) = \frac{e^{\beta x}}{1 + e^{\beta x}} \]

Check that when \( \beta = 0 \) we have \( p = \frac{1}{2} \).

- Given our new values spread= 8 or spread= 4,
  - The win probabilities are 77% and 65%, respectively.

  0.7769474, 0.6511238

Clearly, the bigger spread means a higher chance of winning.
The Gambler Who Cracked the Horse-Racing Code

Bill Benter did the impossible: He wrote an algorithm that couldn’t lose at the track, close to a billion dollars later.

Benter’s model required his undivided attention. It monitored only about 20 inputs—just a fraction of the infinite factors that influence a horse’s performance, from wind speed to what it ate for breakfast, and the Jockey Club’s publicly available betting odds.
Horse race prediction

We use the run.csv data from Kaggle.
We want to use individual variables to predict the chance of winning of a horse.

For the simplicity of computation, we only consider horses with id ≤ 500, and train the model with $\ell_1$-regularized logistic regression.

And we include lengths_behind, horse_age, horse_country, horse_type, horse_rating, horse_gear, declared_weight, actual_weight, draw, win_odds, place_odds as predicting variables in our model.
Horse Race Predictions

Since most of the variables, such as country, gear, type, are categorical, after spanning them into binary indicators, we have more than 800 columns in the design matrix.

We try two logistic regression model. The first one includes win_odds given by the gambling company. The second one does not include the win_odds and we use win_odds to test the power of our model. We tune both models with a 10-fold cross-validation to find the best penalty parameter $\lambda$. 
Predict with win_odds

In this model, we fit the logistic regression with full dataset. The best $\lambda$ we find is $5.699782e−06$. 

---

#variables w.r.t. $\lambda$  
Coefficient of top 20 variables
Predict with win_odds

In this model, we randomly partition the dataset into training(70%) and testing(30%) parts. We fit the logistic regression with training dataset. The best $\lambda$ we find is $4.792637 \times 10^{-6}$.

The out-of-sample mean squared error for win_odds is 0.0668.
LinkedIn Study: How to Become an Executive

Analyze the career paths of about 459,000 LinkedIn members who worked at a Top 10 consultancy between 1990 and 2010 and became a VP, CXO, or partner at a company with at least 200 employees.

About 64,000 members reached this milestone. \( \hat{p} = 0.1394 \).

- Look at their profiles – educational background, gender, work experience, and career transitions.
- Build a model to predict the probability of becoming an executive.

Conditional on making it into the database ....
Logistic Regression

Logistic regression with 8 key features (a.k.a. covariates):

$$\log \left( \frac{p}{1-p} \right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_8 X_8$$

- $p$: Probability of “success” – reach VP/CXO/Partner at a company with at least 200 employees.
- $X_i (i = 1, 2, ..., 8)$: Features to predict the “success” probability.
Features

Location Features: $X_1$ Metro region: whether a member has worked in one of the top 10 largest cities in the U.S. or globally.

Personal Features: $X_2$ Gender: Inferred from member names: ‘male’, or ‘female’.

Education Features: $X_3$ Graduate education type: whether a member has an MBA from a top U.S. program / a non-top program / a top non-U.S. program / another advanced degree.

$X_4$ Undergraduate education type: whether a member has attended a school from the U.S. News national university rankings / a top 10 liberal arts college / a top 10 non-U.S. school.
Features

Work Experience:

X5 **Company count**: # different companies in which a member has worked.

X6 **Function count**: # different job functions in which a member has worked.

X7 **Industry sector count**: # different industries in which a member has worked.

X8 **Years of experience**: # years of work experience, including years in consulting, for a member.
\( \hat{\beta} \)'s of Features\(^1\)

1. Location: Metro region: 0.28
2. Personal: Gender(Male): 0.31
3. Education: Graduate education type: 1.16, Undergraduate education type: 0.22
4. Work Experience: Company count: 0.14, Function count: 0.26, Industry sector count: -0.22, Years of experience: 0.09
Main Findings

1. Working across job functions, like marketing or finance, is good. Each additional job function provides a boost that, on average, is equal to three years of work experience. Switching industries has a slight negative impact. Learning curve? Lost network?

2. MBAs are worth the investment. But pedigree matters. *Top five program equivalent to 13 years of work experience!!!*

3. Location matters. NYC helps.
Examples

**Person A (p=6%)**: Male in Tulsa, Oklahoma, Undergraduate degree, 1 job function for 3 companies in 3 industries, 15-year experience.

**Person B (p=15%)**: Male in London, Undergraduate degree from top international school, Non-MBA Master, 2 different job functions for 2 companies in 2 industries, 15-year experience.

**Person C (p=63%)**: Female in New York City, Top undergraduate program, Top MBA program, 4 different job functions for 4 companies in 1 industry, 15-year experience.
Let's re-design Person B!!

Person B (p=15%): Male in London, Undergraduate degree from top international school, Non-MBA Master, 2 different job functions for 2 companies in 2 industries, 15-year experience.

1. Work in one industry rather than two. Increase 3%
2. Undergrad from top 10 US program rather than top international school. 3%
3. Worked for 4 companies rather than 2. Another 4%
4. Move from London to NYC. 4%
5. Four job functions rather than two. 8%. A 1.5X effect.
6. Worked for 10 more years. 15%. A 2X effect.
Choices and Impact (Person B)

Probability of Success vs. Choices

- Baseline: 15%
- Focused on single industry: 18%
- Got undergrad from top 10 U.S. undergrad: 18%
- Worked across more companies: 19%
- Moved to New York City: 19%
- Worked across more functions: 23%
- Worked for more years: 30%
- Got MBA from Top 5 U.S. program: 36%
- All Combined: 81%