Regression analysis

Nisheeth
Recommender system example

\[ p(i|u) = \sum_{g,t} p(i|g)p(u|t)p(g|t)p(t) \]

Try now
Problem

Too Much Information
Typical human-computer interface

Behavior

Var₁ Var₂ Var₃ Var₄ Var₅ Var₆ Var₇

Insight
Stored variables are a function of business requirements.

How Does Facebook Choose What To Show In News Feed?

News Feed Visibility = C x P x T x R

- **Creator**: Interest of the user in the creator
- **Post**: This post’s performance amongst other users
- **Type**: Type of post (status, photo, link) user prefers
- **Recency**: How new is the post

*This is a simplified equation. Facebook also looks at roughly 100,000 other high-personalized factors when determining what’s shown.*
Important scientific discoveries made tangentially
What we want - I

Behavior

Var_1  Var_2  Var_3  Var_4  Var_5  Var_6  Var_7

Insight
We’d get

Behavior

Var₁

Var₄

Var₇

Insight

Machine learners call this the ‘feature selection’ problem. We will focus on one particular way of solving it - regression analysis.
What we want - II

This is causality inference
Regression model

• Regression model estimates the nature of relationship between the independent and dependent variables.
  – Change in dependent variables that results from changes in independent variables, i.e. size of the relationship.
  – Strength of the relationship.
  – Statistical significance of the relationship.
Bivariate and multivariate models

Bivariate or simple regression model

(Education) \( x \) \( \rightarrow \) \( y \) (Income)

Multivariate or multiple regression model

(Education) \( x_1 \)
(Sex) \( x_2 \)
(Experience) \( x_3 \)
(Age) \( x_4 \)
Bivariate or simple linear regression

- $x$ is the independent variable
- $y$ is the dependent variable
- The regression model is
  \[ y = w_0 + w_1 x + \epsilon \]
- Two parameters to estimate – the slope of the line $w_1$ and the $y$-intercept $w_0$
- $\epsilon$ is the unexplained, random, or error component.
Fitting the regression model

- Any choice of $w$ gives us predictions for the dependent variable $f_i$ for each $x_i$
- Residual $e_i = y_i - f_i$
- Good fit = minimize $\sum_i e_i^2$
- Easy to derive estimators for coefficients using basic calculus
- $\min_w \sum_i (y_i - w_0 - w_1 x_i)^2$
  
  - $w_1 = \frac{Cov(x,y)}{Var(x)}$
  - $w_0 = \bar{y} - w_1 \bar{x}$
Assessing goodness of fit

- Sanity check $E[e] = 0 \rightarrow E[f] = E[y]$
- $SST = \sum_i (y_i - \bar{y})^2$
- $SSR = \sum_i (y_i - f_i)^2$
- $R^2 = 1 - \frac{SSR}{SST}$
Uses of vanilla regression

- Amount of change in a dependent variable that results from changes in the independent variable(s)
- Attempt to determine causes of phenomena.
- Prediction and forecasting
- Support or negate theoretical model.
- Modify and improve theoretical models and explanations of phenomena.
Used to tell stories that make a big difference
Multiple regression

• Everything works the same way as in simple regression

\[ y = w_0 + w_1 x_1 + w_2 x_2 + ... + w_n x_n \]

• Estimated as in the univariate case by minimizing

\[ \sum_{i=1}^{n} \left( y_i - w_0 - \sum_{j=1}^{p} w_j x_{ij} \right)^2 \]

• Each independent variable affects the dependent variable linearly in isolation

• Can also use modifications of the same variable, e.g. \( x^2 \) in place of a new variable
Non-linear relationship

Correlation = +0.12.
All multiple linear regressions

**LINEAR**

Multiple R-squared: 0.7044
\[ Y = 30.53 + 3.05X \]

**QUADRATIC**

Multiple R-squared: 0.7559
\[ Y = 29.90 + 6.48X - 3.22X^2 \]

**CUBIC**

Multiple R-squared: 0.7623
\[ Y = 30.17 + 3.61X + 3.71X^2 - 4.48X^3 \]
Example

- Infant mortality
- Regional wealth levels
- Medical resource availability
Predictable relationship
Regression result

<table>
<thead>
<tr>
<th>Dep. Variable</th>
<th>infant_mortality</th>
<th>R-squared:</th>
<th>0.524</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model:</td>
<td>OLS</td>
<td>Adj. R-squared:</td>
<td>0.490</td>
</tr>
<tr>
<td>Method:</td>
<td>Least Squares</td>
<td>F-statistic:</td>
<td>15.43</td>
</tr>
<tr>
<td>Date:</td>
<td>Wed, 11 Jan 2017</td>
<td>Prob (F-statistic):</td>
<td>3.04e-05</td>
</tr>
<tr>
<td>No. Observations:</td>
<td>31</td>
<td>AIC:</td>
<td>234.8</td>
</tr>
<tr>
<td>Df Residuals:</td>
<td>28</td>
<td>BIC:</td>
<td>239.1</td>
</tr>
<tr>
<td>Df Model:</td>
<td>2</td>
<td>Covariance Type:</td>
<td>nonrobust</td>
</tr>
</tbody>
</table>

|                         | coef       | std err   | t        | P>|t|    | [95.0% Conf. Int.] |
|-------------------------|------------|-----------|----------|--------|-------------------|
| Intercept               | 5.5161     | 23.306    | 0.237    | 0.815  | -42.224 -53.257   |
| NSDP_per_capita         | -0.0064    | 0.002     | -2.936   | 0.007  | -0.011 -0.002     |
| np.log(patients_per_bed)| 6.1056     | 2.825     | 2.162    | 0.039  | 0.320 11.892      |

| Omnibus:                | 1.664      | Durbin-Watson: | 2.644 |
| Prob(Omnibus):          | 0.435      | Jarque-Bera (JB): | 1.086 |
| Skew:                   | -0.116     | Prob(JB):       | 0.581 |
| Kurtosis:               | 2.113      | Cond. No.       | 2.97e+04 |
Looks decent too

$$\text{IMR} = -0.0064 \times (\text{NSDP per capita}) + 6.11 \times (\log(\text{patients per bed})) + 5.51$$
Interpreting the coefficients

<table>
<thead>
<tr>
<th>Dep. Variable:</th>
<th>infant_mortality</th>
<th>R-squared:</th>
<th>0.524</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model:</td>
<td>CLS</td>
<td>Adj. R-squared:</td>
<td>0.490</td>
</tr>
<tr>
<td>Method:</td>
<td>Least Squares</td>
<td>F-statistic:</td>
<td>15.43</td>
</tr>
<tr>
<td>Date:</td>
<td>Thu, 12 Jan 2017</td>
<td>Prob (F-statistic):</td>
<td>3.04e-05</td>
</tr>
<tr>
<td>Time:</td>
<td>10:50:49</td>
<td>Log-Likelihood:</td>
<td>-114.39</td>
</tr>
<tr>
<td>No. Observations:</td>
<td>31</td>
<td>AIC:</td>
<td>234.8</td>
</tr>
<tr>
<td>Df Residuals:</td>
<td>28</td>
<td>BIC:</td>
<td>239.1</td>
</tr>
<tr>
<td>Df Model:</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Covariance Type:</td>
<td>nonrobust</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

|          | coef  | std err | t     | P>|t|  | [95.0% Conf. Int.] |
|----------|-------|---------|-------|-----|-------------------|
| Intercept| 36.2581 | 1.831 | 19.801 | 0.000 | 32.507 40.009 |
| st_NSIPD | -6.5784 | 2.241 | -2.936 | 0.007 | -11.169 -1.988 |
| st_ppb   | 4.8437 | 2.241 | 2.162 | 0.039 | 0.254 9.434 |

<table>
<thead>
<tr>
<th>Omnibus:</th>
<th>1.664</th>
<th>Durbin-Watson:</th>
<th>2.644</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prob(Omnibus):</td>
<td>0.435</td>
<td>Jarque-Bera (JB):</td>
<td>1.086</td>
</tr>
<tr>
<td>Skew:</td>
<td>-0.116</td>
<td>Prob(JB):</td>
<td>0.581</td>
</tr>
<tr>
<td>Kurtosis:</td>
<td>2.113</td>
<td>Cond. No.</td>
<td>1.93</td>
</tr>
</tbody>
</table>
Technical caveats

• Have to look at adjusted $R^2$ to assess model fit
  • $R^2$ can never decrease by adding an extra variable
  • Have to deflate it by number of variables used for fair comparison
    • $R_{adj}^2 = R^2 - \frac{p}{n-p-1}(1 - R^2)$

• Have to watch out for problems
  – Omitted variable bias
  – Multicollinearity
  – Dummy variable trap
  – Outliers
Omitted variable bias

We omit a variable from the analysis that is
- Correlated with at least one of the independent variables and
- Determinative for the response variable mechanistically.
Multicollinearity

- When two of the predictors are highly correlated
- Parameter estimation becomes unstable
- Results become suspect
- Rule of thumb: correlations higher than 0.7 between two variables → leave the less interesting one out of the analysis
Dummy variable trap

• How to handle categorical data?
• Create n dummy variables for an n category variable
• Never use all n in the analysis, leave one out
• Why? Multicollinearity
Outliers

• Rare, extreme values distort OLS fits.
  – Could be an error.
  – Could be a very important observation.
• Outlier: more than 3 standard deviations from the mean.
• Can discard, or use robust regression methods
• Caveat emptor
Pragmatic caveats

• Spurious correlations
• The kitchen sink problem
• Regularization
Spurious correlations

US spending on science, space, and technology correlates with

Suicides by hanging, strangulation and suffocation
Spurious correlations
Spurious correlations

Global Average Temperature Vs. Number of Pirates

[Graph showing a line chart with data points from 1820 to 2000, with years marked at 1820, 1860, 1880, 1920, 1940, 1980, and 2000. The x-axis represents the number of pirates (approximate) and the y-axis represents global average temperature (Celsius).]
Regularization in regression models

• Regularization = trying to keep your model simple

• Do this by adding a regularization term to the regression objective function, i.e. SSE + \( \lambda R \)

• Three basic forms in regression
  – Subset selection: \( R = |w|_0 = \sum_i^p I(w_i) \)
  – Lasso regression: \( R = |w|_1 = \sum_i^p |w_i| \)
  – Ridge regression: \( R = |w|_2 = \sum_i^p w_i^2 \)

• Larger \( \lambda \rightarrow \) simpler model, with fewer non-zero coefficients
Probabilistic intuition

- Assume that $p(y|x, w, \sigma^2) = N(y|f(x, w), \sigma^2)$
- Bayes inversion would give us $p(w|x, y, \sigma^2)$
  - If we have knowledge about the prior on $w$
  - Assume the prior is $N\left(w|0, \alpha = \frac{1}{\sigma_p^2}\right)$
- Find $w$ by maximizing the posterior probability
Probabilistic intuition

- Equivalent to minimizing the negative log posterior
  - \( \min \{- \log p(y|x, w, \sigma^2) - \log p(w|\alpha)\} \)
  - \( \min \left\{ \frac{1}{2\sigma^2} \sum_i (y_i - w_i x_i)^2 + \frac{\alpha}{2} \sum_j w_j^2 \right\} \)
- You could do a full Bayesian regression instead
- How? Why?
What we get

Behavior

Var_1 Var_2 Var_3 Var_4 Var_5 Var_6 Var_7