Going Beyond Data Visualization
With SAS® Visual Statistics

Varsha Chawla
Sr. Solutions Architect
SAS
@varshachawla
Approachable Analytics Journey

- Descriptive
- Diagnostic
- Predictive
- Prescriptive
The Why

• Predict:
  ▪ whether a patient has a disease based on factors such as age, weight, BMI
  ▪ likelihood to churn / purchase / convert
  ▪ what affects a ratings, satisfaction scores
  ▪ voting preferences
  ▪ failure / non-failure
Regression is a statistical technique that tries to explain the relationship between a dependent variable and one or more independent variables.
Your poll will show here

1. Install the app from pollev.com/app
2. Make sure you are in Slide Show mode

Still not working? Get help at pollev.com/app/help or Open poll in your web browser
Let’s learn regression.
Types of Models

- **Logistic**: predicted probability of event, and the predicted label
- **Linear**: predicted value and residual (difference between observed value and actual / true value)
- **Generalized Linear Model**: predicted value and residual
Types of Variables

• Independent / predictor variables:
  ▪ **Continuous** (Continuous Effects)
  ▪ **Categorical** (Classification Effects)
  ▪ **Interaction terms** (Interaction Effects)
Hypothesis testing

• **Why**: Determine whether there’s enough evidence in sample data to infer that a certain condition is true for entire population.

• **How**: Null and alternative hypothesis

• **What**: You use a p-value to make this determination. You reject or fail to reject the null hypothesis.
Hypothesis testing - examples

Null hypothesis: The population mean of all movie ratings is zero

<table>
<thead>
<tr>
<th>Condition to test</th>
<th>Alternative Hypothesis</th>
</tr>
</thead>
<tbody>
<tr>
<td>The population mean is less than the target</td>
<td>one-sided: $\mu &lt; 0$</td>
</tr>
<tr>
<td>The population mean is greater than the target</td>
<td>one-sided: $\mu &gt; 0$</td>
</tr>
<tr>
<td>The population mean differs from the target</td>
<td>two-sided: $\mu \neq 0$</td>
</tr>
</tbody>
</table>
What’s next?

• Collect data
• Determine significance level (\(\alpha \) – alpha = 0.05)
• Perform hypothesis test
• Compare p-value to the significance level
• Determine to reject or fail to reject your null hypothesis:

  - strongly reject
  - reject
  - marginally reject
  - fail to reject

0.01  0.05  0.10
Remember this?

\[ Y = mx + b \]

- **Response / target / dependent variable**
- **Coefficient / rate / slope**
- **Independent variable / Effect (continuous or classification)**
- **Y-intercept**
Hypothesis testing – testing the m

\[ Y = mx + b \]

- \( m = 0 \rightarrow \text{no relationship} \)
- \( m \neq 0 \rightarrow \text{relationship exists} \)
The graph shows the fit summary for the IMDB rating model. The R-Square value is 0.8538 and the observations used are 397. The p-values for the variables are shown on the x-axis. The p-values range from 0.1 to <0.00001.

- Variables such as rt_score, audience_freshness, rt_audience_score, adjusted, worldwide_gross, and length are strongly rejected (p < 0.01).
- Variables with p-values between 0.01 and 0.05 are marginally rejected.
- Variables with p-values between 0.05 and 0.10 are rejected.
- Variables with p-values greater than 0.1 are fail to reject.
What’s a good model?

Good

Good

Bad
R-Square

• How close is that data to the fitted regression line?
• Explained variation / Total variation
• Between 0 and 100%
• In general, the higher the R-squared, the better the model fits your data
• Key point: we don’t look at it in isolation
Your poll will show here

1. Install the app from pollev.com/app
2. Make sure you are in Slide Show mode

Still not working? Get help at pollev.com/app/help or Open poll in your web browser
## Actual Ratings

<table>
<thead>
<tr>
<th>Movie</th>
<th>IMBD rating</th>
</tr>
</thead>
<tbody>
<tr>
<td>The Matrix</td>
<td>8.7</td>
</tr>
<tr>
<td>Back to the Future</td>
<td>8.5</td>
</tr>
<tr>
<td>Toy Story 3</td>
<td>8.4</td>
</tr>
<tr>
<td>Jurassic Park</td>
<td>8</td>
</tr>
<tr>
<td>Mission Impossible III</td>
<td>6.8</td>
</tr>
</tbody>
</table>
What we're working with

- Movie titles
- Genre information (three different columns dedicated to genre)
- Run time
- IMDB rating
- Rotten tomato audience / critic ratings
- Box office receipts (adjusted for inflation)
K-means clustering

- Midlife Crisis
- Happily Retired
- Midlife Success
- Young Beginners
Online. Everyday.

“I always learn something new when I post in this forum. Just what I needed…”

SAS Visual Analytics Community

Communities.sas.com/visual-analytics