Categorical Variables

- Response: Highway MPG
- Explanatory: Type of drive
  - All Wheel
  - Rear Wheel
  - Front Wheel

Indicator Variables

We have used indicator variables so that we can trick JMP into analyzing the data using multiple regression.

### Summary of Fit

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Ratio</th>
<th>Prob &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>2</td>
<td>466.152</td>
<td>16.5224</td>
<td>&lt;.0001*</td>
<td></td>
</tr>
<tr>
<td>Error</td>
<td>93</td>
<td>28.213</td>
<td>0.3016</td>
<td>&lt;.0001*</td>
<td></td>
</tr>
<tr>
<td>C. Total</td>
<td>99</td>
<td>99369.000</td>
<td>3669.000</td>
<td>&lt;.0001*</td>
<td></td>
</tr>
</tbody>
</table>

### Parameter Estimates

| Term      | Estimate | Std Error | t Ratio | Prob>|H| |
|-----------|----------|-----------|---------|-----|---|
| Intercept | 29.98333 | 0.665726  | 43.72   | <.0001**          |
| All Wheel | -7.374638| 1.3052648 | -5.66   | <.0001**          |
| Rear Wheel| -3.453922| 1.459395  | -2.37   | 0.0199*           |
Categorical Variables

There is a more straightforward analysis that can be done with categorical explanatory variables.

Categorical Variables

The analysis is an extension of the two independent sample analysis we did at the beginning of the semester.

Body mass index for men and women (Lectures 4 and 5).

JMP

Response, Y: Highway MPG (numerical)

Explanatory, X: Drive (categorical)

Fit Y by X
JMP Fit Y by X

- From the red triangle pull down menu next to Oneway select Means/Anova
- Display options
  - Uncheck Mean Diamonds
  - Check Mean Lines

### Oneway Analysis of Highway MPG By Drive Type

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Ratio</th>
<th>Prob &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Drive Type</td>
<td>2</td>
<td>932.3031</td>
<td>466.152</td>
<td>16.5224</td>
<td>&lt;.0001*</td>
</tr>
<tr>
<td>Error</td>
<td>97</td>
<td>2136.6969</td>
<td>28.213</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C. Total</td>
<td>99</td>
<td>3669.0000</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Means for Oneway Anova

<table>
<thead>
<tr>
<th>Level</th>
<th>Number</th>
<th>Mean</th>
<th>Std Error</th>
<th>Lower 95%</th>
<th>Upper 95%</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>23</td>
<td>22.5087</td>
<td>1.1076</td>
<td>20.411</td>
<td>24.507</td>
</tr>
<tr>
<td>Front</td>
<td>80</td>
<td>29.9833</td>
<td>0.8887</td>
<td>28.032</td>
<td>31.934</td>
</tr>
<tr>
<td>Rear</td>
<td>17</td>
<td>26.5294</td>
<td>1.2883</td>
<td>23.973</td>
<td>29.086</td>
</tr>
</tbody>
</table>

Std Error uses a pooled estimate of error variance
Analysis of Variance

- Response: numerical, Y
- Explanatory: categorical, X
- Total Sum of Squares

\[ SS_{Total} = \sum (y - \bar{y})^2 \]

Analysis of Variance

- Partition the Total Sum of Squares into two parts.
  - Due to differences among the sample means for the categories.
  - Due to variation within categories, i.e. error variation.

Sum of Squares Factor

\[ SS_{Factor} = \sum n_i (\bar{y}_i - \bar{y})^2 \]

\[ n_i = \text{number of observations in category } i \]

\[ \bar{y}_i = \text{sample mean for category } i \]
**Stat 301 – Lecture 34**

**Sum of Squares Error**

\[ SS_{Error} = \sum (n_i - 1)s_i^2 \]

- \( n_i \) = number of observations in category \( i \)
- \( s_i^2 \) = sample variance for category \( i \)

**Mean Square**

- A mean square is the sum of squares divided by its associated degrees of freedom.
- A mean square is an estimate of variability.

**Mean Square Factor**

- The mean square factor estimates the variability due to differences among category sample means.
- If the mean square factor is large, this indicates the category sample means are quite different.
Mean Square Error

- The mean square error estimates the naturally occurring variability, i.e. the error variance, $\sigma^2$.
- This is the ruler used to assess the statistical significance of the differences among category sample means.

Test of Hypothesis

- $H_0$: all the category population means are equal.
- $H_A$: some of the category population means are not equal.
- Similar to the test of model utility.

Test Statistic

- $F = \frac{MS_{Factor}}{MS_{Error}}$
- $P$-value = Prob > $F$
- If the $P$-value is small, reject $H_0$ and declare that at least two of the categories have different population means.
### Analysis of Variance

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Factor (Model)</td>
<td>k – 1</td>
<td>SS\text{Factor}</td>
<td>MS\text{Factor}</td>
<td>MS\text{Factor} / MS\text{Error}</td>
</tr>
<tr>
<td>Error</td>
<td>N – k</td>
<td>SS\text{Error}</td>
<td>MS\text{Error}</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>N – 1</td>
<td>SS\text{Total}</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Analysis of Variance

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Factor (Model)</td>
<td>2</td>
<td>932.3</td>
<td>466.15</td>
<td>16.52</td>
</tr>
<tr>
<td>Error</td>
<td>97</td>
<td>2736.7</td>
<td>28.213</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>99</td>
<td>3669.0</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Test of Hypothesis
- $F = 16.52$, P-value < 0.0001
- Because the P-value is so small, there are some categories that have different population means.
Test of Significance

- The test of significance, like the test of model utility, is very general. We know there are some categories with different population means but which categories are they?

Multiple Comparisons

- In ANOVA, a statistically significant F test is often followed up by a procedure for comparing the sample means of the categories.

Least Significant Difference

- One multiple comparison method is called Fisher's Least Significant Difference, LSD.
- This is the smallest difference in sample means that would be declared statistically significant.
Least Significant Difference

\[ \text{LSD} = t^* \times \text{RMSE} \times \sqrt{\frac{1}{n_i} + \frac{1}{n_j}} \]

where \( t^* \) is a value for 95% confidence and degrees of freedom = \( df_{\text{Error}} \). 
\( n_i \) and \( n_j \) are the sample sizes for categories \( i \) and \( j \).

---

Least Significant Difference

- When the number of observations in each category is the same, there is one value of LSD for all comparisons.
- When the numbers of observations in each category are different, there is one value of LSD for each comparison.

---

Compare All to Rear Wheel

- All Wheel: \( n_i = 23 \)
- Rear Wheel: \( n_j = 17 \)
- \( t^* = 1.98472 \)
- RMSE = 5.311626
Compare All to Rear Wheel

$LSD = 1.98472(5.311626)\sqrt{\frac{1}{23} + \frac{1}{17}}$

$LSD = 1.98472(1.698905)$

$LSD = 3.37185$

---

Compare All to Rear Wheel

- All Wheel: mean = 22.6087
- Rear Wheel: mean = 26.5294
- Difference in means = 3.9207
- 3.921 is bigger than the LSD = 3.372, therefore the difference between All Wheel and Rear Wheel is statistically significant.

---

JMP – Fit Y by X

- From the red triangle pull down menu next to Oneway select Compare Means – Each Pair Student’s t.
Regression vs ANOVA

- Note that the P-values for the comparisons are the same as the P-values for the slope estimates in the regression on indicator variables.

Regression vs ANOVA

- Multiple regression with indicator variables and ANOVA give you exactly the same analysis.