36-309/749
Experimental Design for Behavioral and Social Sciences

Sep. 29, 2015
Lecture 5: Multiple Regression
Review of ANOVA & Simple Regression

- Both
  - Quantitative outcome
  - Independent, Gaussian errors with equal variance
  - Group assignment assumed correct (fixed-x)

- One way (between-subjects) ANOVA
  - Categorical IV (k levels) with means $\mu_1$ through $\mu_k$
  - Best prediction: $\hat{Y}_i = \bar{Y}_j$ for subject $i$ in group $j$

- Simple (one IV) regression
  - Quantitative IV
  - Coefficient parameters are $\beta_0$ and $\beta_1$
  - True mean outcome at each $x$ is $\text{E}(Y|x)=\beta_0+\beta_1x$ (linearity)
  - Best prediction: $\hat{Y}_i = b_0 + b_1x$
Example

Team Problem Solving
Multiple Regression

New Idea #1: extend the means model

- IVs are \( x_1, x_2, \ldots \)
- Means model: \( E(Y | x_1, x_2, \ldots) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots \)
- Prediction: \( \hat{Y} = b_0 + b_1 x_1 + b_2 x_2 + \ldots \)
- Consequences:
  - \( \beta_0 \) is the mean of the DV when all IVs equal 0
  - \( \beta_1 \) is the change in the mean of the DV when \( x_1 \) goes up by one and all other \( x \)'s are held constant.
  - E.g., with 2 \( x \)'s and \( x_2 \) fixed at \( c \), \( Y \) vs. \( x_1 \) is a line:
    \[
    E(Y | x_1, x_2 = c) = \beta_0 + \beta_1 x_1 + \beta_2 c = (\beta_0 + \beta_2 c) + \beta_1 x_1
    
    \]
    So \( Y \) vs. \( x_1 \) forms parallel lines at various fixed \( x_2 \) values
Multiple Regression: dummies

New idea #2: Dummy variables

- Multiple regression can accommodate categorical IVs but **only if** they are coded appropriately
- **Indicator variable**: A categorical variable (factor) with 2 levels should be named for one level and coded with: 1=named level, 0=other level, e.g., a “Female” variable “F” is coded 0=Male, 1=Female.
- E.g., $x_1=$Age, $x_2=$Female: $E(Y)=\beta_0+\beta_A A+\beta_F F$ is a means model of parallel lines:
  
  Males: $E(Y) = \beta_0 + \beta_A A$
  
  Females: $E(Y) = (\beta_0+\beta_F F) + \beta_A A = (\beta_0+\beta_F) + \beta_A A$
Multiple Regression: dummies, cont.

- Coding of a categorical IV with k>2 levels
  - Choose an arbitrary baseline (e.g., “control”)
  - Create *indicator variables* for all *non-baseline* levels
  - Throw away the original variable
  - Example:

<table>
<thead>
<tr>
<th>Color (code)</th>
<th>Color (“value”)</th>
<th>Red</th>
<th>Blue</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>Red</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>Blue</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>Blue</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>Green</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

- “Green” is the arbitrary “baseline”. “Red” and “Blue” are the IVs used in the regression.
Generally the term **ANCOVA** (analysis of covariance) refers to multiple regression with one quantitative IV ("covariate") and one categorical IV of primary interest coded as dummy variables.

**Example:** covariate "Age" and factor "Color" (baseline=green)

\[
E(Y|\text{Age, Color}) = E(Y|A,B,R) = \beta_0 + \beta_A A + \beta_B B + \beta_R R
\]

\[
E(Y|\text{Age, Color=Green}) = E(Y|A, B=0, R=0) = \beta_0 + \beta_A A
\]

\[
E(Y|\text{Age, Red}) = \beta_0 + \beta_A A + \beta_B 0 + \beta_R 1 = (\beta_0 + \beta_R) + \beta_A A
\]

\[
E(Y|\text{Age, Blue}) = \beta_0 + \beta_A A + \beta_B 1 + \beta_R 0 = (\beta_0 + \beta_B) + \beta_A A
\]
In ANCOVA as regression, dummy variables’ “slopes” reflect different intercept offsets from the intercept of the baseline category. As opposed to individual regressions, inference for comparing lines is provided.
Fear and Anger Example

This is loosely based on *Constraints for emotion specificity in fear and anger: The context counts* by Stemmler, et al., *Psychophysiology*, 38, 275–291 (2001). One hundred and sixty-nine adult female subjects were randomized to a control condition or to induction of fear or anger. The outcome of interest is the subjects’ combined ratings on three 0-10 point scales of “negativity”. The “covariate” is a quantitative measure called heart-period-variability (HPV), which is measured before the emotion induction and is taken as a measure of individual physiological sensitivity to one’s surroundings.


- Model? Null hypotheses? Alternative hypotheses?
### Regression output

<table>
<thead>
<tr>
<th>Model</th>
<th>R</th>
<th>R Square</th>
<th>Adjusted R Square</th>
<th>Std. Error of the Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.891&lt;sup&gt;a&lt;/sup&gt;</td>
<td>.794</td>
<td>.790</td>
<td>3.181</td>
</tr>
</tbody>
</table>

a. Predictors: (Constant), Anger induction, Heart period variability, Fear induction

b. Dependent Variable: Feelings of negativity

<table>
<thead>
<tr>
<th>Model</th>
<th>Sum of Squares</th>
<th>df</th>
<th>Mean Square</th>
<th>F</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Regression</td>
<td>6438.971</td>
<td>3</td>
<td>2146.324</td>
<td>212.119</td>
</tr>
<tr>
<td></td>
<td>Residual</td>
<td>1669.550</td>
<td>165</td>
<td>10.118</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>8108.521</td>
<td>168</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Example, cont.
Example, cont.

<table>
<thead>
<tr>
<th></th>
<th>Unstandardized Coefficients</th>
<th></th>
<th></th>
<th>95% Confidence Interval for B</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>B</td>
<td>Std. Error</td>
<td>t</td>
<td>Sig.</td>
</tr>
<tr>
<td>(Constant)</td>
<td>2.707</td>
<td>.691</td>
<td>3.920</td>
<td>.000</td>
</tr>
<tr>
<td>Heart period variability</td>
<td>1.442</td>
<td>.128</td>
<td>11.263</td>
<td>.000</td>
</tr>
<tr>
<td>Fear induction</td>
<td>13.233</td>
<td>.618</td>
<td>21.397</td>
<td>.000</td>
</tr>
<tr>
<td>Anger induction</td>
<td>12.118</td>
<td>.602</td>
<td>20.141</td>
<td>.000</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Lower Bound</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Upper Bound</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.343</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>4.071</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.189</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.694</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>12.012</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>14.454</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>10.930</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>13.306</td>
</tr>
</tbody>
</table>

**Prediction equations:**

\[ \hat{Y}_i = 2.71 + 1.44 \text{ HPV}_i + 13.23 \text{ Fear}_i + 12.12 \text{ Anger}_i \]

Controls (Fear=0, Anger=0): \[ \hat{Y}_i = 2.71 + 1.44 \text{ HPV}_i \]

Fear subjects: \[ \hat{Y}_i = 2.71 + 1.44 \text{ HPV}_i + 13.23 \text{ (1)} = 15.94 + 1.44 \text{ HPV}_i \]

Anger subjects: \[ \hat{Y}_i = 2.71 + 1.44 \text{ HPV}_i + 12.12(1) = 14.83 + 1.44 \text{ HPV}_i \]

**Hidden assumption of the (non-interaction) ANCOVA means model:**
Example, cont.

- **Standardized coefficients**: coefficients from running regression on standardized x’s and Y:
  \[ x_{ij}^* = \frac{x_{ij} - \bar{x}_j}{s_{x_j}} \]
  \[ Y_i^* = \frac{(Y_i - \bar{Y})}{s_Y} \]

- **Residual plots** for assumption checking
  - Residual = obs – exp = \( Y_i - \hat{Y}_i \) (estimated error)
  - Residual quantile normal plot: random scatter around reference line \( \rightarrow \) Normality OK

![Normal Q-Q Plot of Residual](image)
Example: residual analysis, cont.

**Residual plots** for assumption checking

- Residual vs. fit (predicted) plot
  y-axis: residuals, x-axis: fitted values
  Smile or frown suggests non-linearity (bad means model)
  Funneling suggests unequal variance
Example: skipped EDA
Multiple Regression: interaction

- An interaction between two IVs in their effect on the DV implies non-additivity. The effect of a one unit increase in $x_1$ depends on the level of $x_2$ (and vice versa).

- Interaction is coded by adding a new IV which is the product of the two original IVs. If $x_1$ and $x_2$ are both quantitative there is one new IV, $x_1 \times x_2$. If one is a k-level factor there are k-1 new IVs.

- ANCOVA with interaction
  - Structural model: $E(Y|x_1, x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2$
  - Prediction: $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_{12} x_1 x_2$
Example with interaction

\[ E(\text{Negativity} \mid \text{HPV, emotion}) = E(N \mid H, A, F) = \beta_0 + \beta_H H + \beta_A A + \beta_F F + \beta_{HA} HA + \beta_{HF} HF \]

Key step: simplification

Key concept: \( \beta \)'s are fixed; H, A, F are data values

Controls: \( E(N \mid H, A=0, F=0) = \beta_0 + \beta_H H \)

Anger: \( E(N \mid H, A=1, F=0) = \beta_0 + \beta_H H + \beta_A + \beta_{HA} H \)

\[ = (\beta_0 + \beta_A) + (\beta_H + \beta_{HA}) H \]

Fear: \( E(N \mid H, A=0, F=1) = (\beta_0 + \beta_F) + (\beta_H + \beta_{HF}) H \)
Example with Interaction: Results

<table>
<thead>
<tr>
<th>Model</th>
<th>R</th>
<th>R Square</th>
<th>Adjusted R Square</th>
<th>Std. Error of the Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.891</td>
<td>0.794</td>
<td>0.792</td>
<td>3.181</td>
</tr>
<tr>
<td>2</td>
<td>0.906</td>
<td>0.821</td>
<td>0.816</td>
<td>2.982</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Model</th>
<th>R Square Change</th>
<th>F Change</th>
<th>df1</th>
<th>df2</th>
<th>Sig. F Change</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.794</td>
<td>212.119</td>
<td>3</td>
<td>165</td>
<td>.000</td>
</tr>
<tr>
<td>2</td>
<td>.027</td>
<td>12.359</td>
<td>2</td>
<td>163</td>
<td>.000</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Model</th>
<th>Unstandardized Coefficients</th>
<th>Std. Error</th>
<th>t</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>(Constant)</td>
<td>2.707</td>
<td>.691</td>
<td>3.920</td>
</tr>
<tr>
<td></td>
<td>Heart period variability</td>
<td>1.442</td>
<td>.128</td>
<td>11.263</td>
</tr>
<tr>
<td></td>
<td>anger</td>
<td>12.118</td>
<td>.602</td>
<td>20.141</td>
</tr>
<tr>
<td></td>
<td>fear</td>
<td>13.233</td>
<td>.618</td>
<td>21.397</td>
</tr>
<tr>
<td>2</td>
<td>(Constant)</td>
<td>6.153</td>
<td>1.003</td>
<td>6.135</td>
</tr>
<tr>
<td></td>
<td>Heart period variability</td>
<td>.612</td>
<td>.220</td>
<td>2.780</td>
</tr>
<tr>
<td></td>
<td>anger</td>
<td>6.454</td>
<td>1.272</td>
<td>5.073</td>
</tr>
<tr>
<td></td>
<td>fear</td>
<td>9.807</td>
<td>1.350</td>
<td>7.264</td>
</tr>
<tr>
<td></td>
<td>HPV*Anger</td>
<td>1.439</td>
<td>.289</td>
<td>4.972</td>
</tr>
<tr>
<td></td>
<td>HPV*Fear</td>
<td>.825</td>
<td>.312</td>
<td>2.643</td>
</tr>
</tbody>
</table>

a. Dependent Variable: Feelings of negativity
Example with interaction: Diagnostics

QN plot: OK for Normality
Res. vs. Fit: OK for linearity and equal spread
Example: Subject Matter Conclusions

- There is a statistically significant interaction ($F_{\text{change}}=12.4$, df=2, 163, $p<0.0005$) between HPV and emotion in their effects on negativity (N).

- Heart period variability is positively associated with negativity ($t=2.78$, df=163, $p=0.006$) in controls, and the estimated mean change in N is a rise of 0.612 points for each 1 unit rise in HPV (95% CI=[0.18,1.05]).

- The estimated mean N when HPV=0 is 6.15 for controls, and is 6.45 higher for induced anger ($t=5.07$, df=163, $p<0.0005$) and 9.81 higher for induced fear ($t=7.26$, df=163, $p<0.0005$).

- The change in N associated with a 1 point rise in HPV is estimated to be 1.44 points greater for anger compared to control ($t=4.97$, df=163, $p<0.0005$) and 0.82 points greater for fear compared to control ($t=2.63$, df=163, $p=0.009$).

- Overall compared to control inducing fear and anger increases N, and the increase is greater when HPV is greater.
In multiple regression, the means model adds terms of the form $\beta_v V$ when variable $V$ is added.

Any $k$-level categorical variable must be replaced with $k-1$ indicator variables [or similar].

Without interaction, a “parallel” means model is produced: at each level of one IV the slope of the DV vs. the other IV is the same.

With interaction (adding product variables) different slopes are accommodated.

You can deduce the meanings of parameters by simplifying the means model for each category to a $Y = a + bX$ form where $a$ is the intercept and $b$ is the slope.

Continue the deduction by finding equations that differ only in a single parameter. The p-value for that parameter is the null hypothesis that that $\beta = 0$ which is equivalent to the two lines being the same.