Course Business

- We *will* have class this coming Monday (Labor Day)
- Lab materials on Canvas
- Follow-up on fixed effects – experimental vs. observational variables
  - No difference in how we include them in R script
  - Difference is just in conclusions we can draw
Week 3.2: Fixed Effect Interactions

- Residuals
  - Predicted Values & Residuals
    - Using Residuals to Detect Outliers
- Interactions
  - In the Model Formula
  - Interpretation
- Lab
Predicted Values

- Last time, we focused on testing the significance of effects.
- But, model implies an predicted or fitted value for each observation.

\[ E(Y_i(j)) = Y_{00} + Y_{10}X_1i(j) + Y_{20}X_2i(j) \]

- Minutes taken to assemble phone
- Baseline
- Newcomers
- Years of Experience
Predicted Values

• Last time, we focused on testing the significance of effects
• But, model implies an predicted or fitted value for each observation ("y hat"):

\[ \hat{Y}_{i(j)} = Y_{00} + Y_{10}x_{1i(j)} + Y_{20}x_{2i(j)} \]

- \( Y_{00} \): Baseline
- \( Y_{10} \): Newcomers
- \( Y_{20} \): Years of Experience

Minutes taken to assemble phone
Predicted Values

• Last time, we focused on testing the significance of effects
• But, model implies an predicted or fitted value for each observation ("y hat"): 

\[ \hat{Y}_{i(j)} = 38.25 + 2.16x_{1i(j)} - 0.10x_{2i(j)} \]

Minutes taken to assemble phone
Baseline
Newcomers
Years of Experience

• What do we predict for a team with 1 newcomer \((x_1=1)\) and 2 years experience \((x_2=2)\)?
Predicted Values

- Last time, we focused on *testing* the significance of effects.
- But, model implies an *predicted* or *fitted* value for each observation ("y hat"): 

\[ \hat{Y}_{i(j)} = 38.25 + 2.16 + (-0.20) \]

<table>
<thead>
<tr>
<th>Minutes taken to assemble phone</th>
<th>Baseline</th>
<th>Newcomers</th>
<th>Years of Experience</th>
</tr>
</thead>
<tbody>
<tr>
<td>What do we predict for a team with 1 newcomer (x_1=1) and 2 years experience (x_2=2)?</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Predicted Values

• Last time, we focused on testing the significance of effects
• But, model implies an predicted or fitted value for each observation (“y hat”):

\[ \hat{Y}_{i(j)} = 40.21 \]

Minutes taken to assemble phone

• What do we predict for a team with 1 newcomer \((x_1=1)\) and 2 years experience \((x_2=2)\)?
Predicted Values

• We can use our model to generate predictions for new data with the same variables
  • `model1 %>% predict(newdataframe)`

• We can also find out what our model predicted for each of our current observations
  • `problemsolving %>%
    mutate(PredictedScore=fitted(model1))
    -> problemsolving`

  Function that gets the fitted/predicted values from a model
  • Adds a `PredictedScore` column to the `problemsolving` dataframe

We need to specify a `particular model` here. These are the values predicted by a specific regression model.
Predicted Values

• We can use our model to generate predictions for new data with the same variables
  • `model1 %>% predict(newdataframe)`

• We can also find out what our model predicted for each of our current observations

This prediction was a little less accurate
Residuals

• How far off are our individual predictions?
• **Residuals**: Difference between predicted & actual for a specific observation

• “2% or 3% [market share] is what Apple might get.”
  — former Microsoft CEO Steve Ballmer on the iPhone
• Actual iPhone market share (2014): **43%**
• Residual: **40 to 41** percentage points
Residuals

- How far off are our individual predictions?

**Residuals**: Difference between predicted & actual for a specific observation

- + residual: Real observation was higher
- - residual: Real observation was lower

```r
problemsolving %>%
mutate(Residual = resid(model1))
```

Measured in original units (here, minutes)
Week 3.2: Fixed Effect Interactions

- Residuals
- Predicted Values & Residuals
- Using Residuals to Detect Outliers
- Interactions
  - In the Model Formula
  - Interpretation
- Lab
Outliers

• One reason to be interested in residuals is that they can help us detect **outliers**
• Concern that a few unusual observations may distort our results

No outliers! Nice scatterplot with best-fitting regression line
\[ y = 3.00 + 0.5x \ (r = .82) \]

Also best fit by:
\[ y = 3.00 + 0.5x \ (r = .82) \]
But driven by one datapoint!

Still best fit by:
\[ y = 3.00 + 0.5x \ (r = .82) \]
But stronger relationship \( (r = 1.0!) \) is obscured by 1 outlier

Anscombe, 1973
Residuals

- Residuals can be one way to detect outlying observations
  - Outliers \textit{after} accounting for all of the variables of interest
    - \textbf{Multivariate outlier}
    - Long RT might not be an outlier if slowest participant on slowest item
Residuals

- What threshold should we use?
- Let’s put the residuals on a standard (z-scored) scale

```r
probsolving %>%
mutate(StdResidual = scale(resid(model1)))
```

<table>
<thead>
<tr>
<th>TeamID</th>
<th>Company</th>
<th>Newcomers</th>
<th>YearsExperience</th>
<th>MinutesTaken</th>
<th>PredictedScore</th>
<th>Residual</th>
<th>StdResidual</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Team1</td>
<td>Walmart</td>
<td>3</td>
<td>17</td>
<td>43.20</td>
<td>42.78028</td>
<td>0.419724</td>
</tr>
<tr>
<td>2</td>
<td>Team2</td>
<td>Walmart</td>
<td>5</td>
<td>17</td>
<td>43.83</td>
<td>47.10867</td>
<td>-3.278672</td>
</tr>
<tr>
<td>3</td>
<td>Team3</td>
<td>Walmart</td>
<td>2</td>
<td>13</td>
<td>44.49</td>
<td>41.00359</td>
<td>3.486406</td>
</tr>
<tr>
<td>4</td>
<td>Team4</td>
<td>Walmart</td>
<td>6</td>
<td>10</td>
<td>44.51</td>
<td>49.95102</td>
<td>-5.441022</td>
</tr>
<tr>
<td>5</td>
<td>Team5</td>
<td>Walmart</td>
<td>5</td>
<td>5</td>
<td>56.76</td>
<td>48.27122</td>
<td>8.488782</td>
</tr>
<tr>
<td>6</td>
<td>Team6</td>
<td>Walmart</td>
<td>5</td>
<td>1</td>
<td>47.14</td>
<td>48.65873</td>
<td>-1.518733</td>
</tr>
<tr>
<td>7</td>
<td>Team7</td>
<td>Walmart</td>
<td>5</td>
<td>14</td>
<td>52.77</td>
<td>47.39931</td>
<td>5.370692</td>
</tr>
<tr>
<td>8</td>
<td>Team8</td>
<td>Walmart</td>
<td>1</td>
<td>16</td>
<td>36.09</td>
<td>38.54876</td>
<td>-2.458759</td>
</tr>
<tr>
<td>9</td>
<td>Team9</td>
<td>Walmart</td>
<td>2</td>
<td>17</td>
<td>38.96</td>
<td>40.61608</td>
<td>-1.656078</td>
</tr>
<tr>
<td>10</td>
<td>Team10</td>
<td>Walmart</td>
<td>5</td>
<td>8</td>
<td>43.86</td>
<td>47.98058</td>
<td>-4.120581</td>
</tr>
</tbody>
</table>
Residuals

• What threshold should we use?
• Let’s put the residuals on a standard (z-scored) scale
  • `problemsolving %>%
    mutate(StdResidual = scale(resid(model1)))`
  • “Number of standard deviations” we were off

<table>
<thead>
<tr>
<th>TeamID</th>
<th>Company</th>
<th>Newcomers</th>
<th>YearsExperience</th>
<th>MinutesTaken</th>
<th>PredictedScore</th>
<th>Residual</th>
<th>StdResidual</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Team1</td>
<td>Walmart</td>
<td>3</td>
<td>17</td>
<td>43.20</td>
<td>42.78028</td>
<td>0.419724</td>
</tr>
<tr>
<td>2</td>
<td>Team2</td>
<td>Walmart</td>
<td>5</td>
<td>17</td>
<td>43.83</td>
<td>47.10867</td>
<td>-3.278672</td>
</tr>
<tr>
<td>3</td>
<td>Team3</td>
<td>Walmart</td>
<td>2</td>
<td>13</td>
<td>44.49</td>
<td>41.00359</td>
<td>3.486406</td>
</tr>
<tr>
<td>4</td>
<td>Team4</td>
<td>Walmart</td>
<td>6</td>
<td>10</td>
<td>44.51</td>
<td>49.95102</td>
<td>-5.441022</td>
</tr>
<tr>
<td>5</td>
<td>Team5</td>
<td>Walmart</td>
<td>5</td>
<td>5</td>
<td>56.76</td>
<td>48.27122</td>
<td>8.488782</td>
</tr>
<tr>
<td>6</td>
<td>Team6</td>
<td>Walmart</td>
<td>5</td>
<td>1</td>
<td>47.14</td>
<td>48.65873</td>
<td>-1.518733</td>
</tr>
<tr>
<td>7</td>
<td>Team7</td>
<td>Walmart</td>
<td>5</td>
<td>14</td>
<td>52.77</td>
<td>47.39931</td>
<td>5.370692</td>
</tr>
<tr>
<td>8</td>
<td>Team8</td>
<td>Walmart</td>
<td>1</td>
<td>16</td>
<td>36.09</td>
<td>38.54876</td>
<td>-2.458759</td>
</tr>
<tr>
<td>9</td>
<td>Team9</td>
<td>Walmart</td>
<td>2</td>
<td>17</td>
<td>38.96</td>
<td>40.61608</td>
<td>-1.656078</td>
</tr>
<tr>
<td>10</td>
<td>Team10</td>
<td>Walmart</td>
<td>5</td>
<td>8</td>
<td>43.86</td>
<td>47.98058</td>
<td>-4.120581</td>
</tr>
</tbody>
</table>
Residuals

• What threshold should we use?
• Now we can drop observations with too extreme a standardized residual
  • `problemsolving %>%
    filter(abs(StdResidual) <= 3)
    -> problemsolving.NoOutliers`
  • Keep only the observations with a standardized residual within $\pm 3$ standard deviations

• How many data points did we lose?
  • `nrow(problemsolving) -
    nrow(problemsolving.NoOutliers)`

• Then, rerun your model
How Should Outliers Change Interpretation?

- Effect reliable with and without outliers?
  - Hooray!

- Effect only seen if outliers removed?
  - Effect characterizes *most* of the data, but a few exceptions

- Effect only seen with outliers included?
  - Suggests it’s driven by a few observations

- No effect either way?
  - Weep softly at your desk
Week 3.2: Fixed Effect Interactions

- Residuals
- Predicted Values & Residuals
- Using Residuals to Detect Outliers
- Interactions
- In the Model Formula
  - Interpretation
- Lab
A common question of interest is how two variables may *interact* or *moderate* each other.

Does the effect of YearsExperience depend on how many Newcomers there are, or vice versa?

Let’s add an interaction to the model:

```r
model2 <- lmer(MinutesTaken ~ 1 + Newcomers + YearsExperience + Newcomers:YearsExperience + (1|Company), data=problemsolving)
```

: means **interaction**
A common question of interest is how two variables may *interact* or *moderate* each other.

Does the effect of YearsExperience depend on how many Newcomers there are, or vice versa?

In this case, *no significant interaction*
Model Formulae: Interactions

- A shortcut!
- \(1 + \text{Newcomers} \times \text{YearsExperience}\)
  - A * means the interaction plus all of the individual effects
  - For factorial experiments (where we include every combination of independent variables), usually what you want

- Scales up to even more variables:
  \(\text{Newcomers} \times \text{YearsExperience} \times \text{PositiveAffect}\)
  - 3-way interaction, plus all 2-way interactions, plus all 1-way effects
Model Formulae: Interactions

- The interaction of A and B is defined as the special effect of combining A & B … over and above their individual effects
- Thus, it does not make sense to discuss an interaction without including the individual effects

```r
bad.model <- lmer(MinutesTaken ~ 1 + Newcomers:YearsExperience + (1|Company), data=problemsolving)
```
Week 3.2: Fixed Effect Interactions

- Residuals
  - Predicted Values & Residuals
  - Using Residuals to Detect Outliers
- Interactions
  - In the Model Formula
- Interpretation
- Lab
Interpreting Interactions

• In many cases, looking at a figure of the descriptive statistics will be most helpful for understanding.
  • Good to do whenever you’re uncertain.

• Fixed effect estimates:
  • Test the statistical significance of interaction.
  • Provide a numerical description.
Interpreting Interactions

\[ y = 38 + 2.17 \times \text{Newcomers} - 0.10 \times \text{Experience} + (-0.001 \times \text{Newcomers} \times \text{Experience}) \]

- Reduced by the Newcomers effect (larger number = longer solve time) if the team is more experienced.
- Increased by the YearsExperience effect (larger number = longer solve time) if there are more newcomers.

When would this decrease the DV the most? (most negative number):
- When Newcomers is large
- When YearsExperience is large
Interpreting Interactions

• Numerical interaction term tells us how the interaction works:
  • Strengthens individual effects with the same sign as the interaction
  • Weakens individual effects with a different sign as the interaction

• Or, again, just look at the graph 😊
Week 3.2: Fixed Effect Interactions

- Residuals
  - Predicted Values & Residuals
  - Using Residuals to Detect Outliers
- Interactions
  - In the Model Formula
  - Interpretation
- Lab
Lab

- Canvas: Modules → Week 3.2
- stroop.csv