Choosing the Best Multiple Regression Model

- Which combination of explanatory variables will best predict the "***Calories***" of a chicken sandwich? Which variables contribute nothing to our understanding of what makes a sandwich more or less healthy?

- It is best to keep the regression model as simple as possible. This translates to limiting the number of x variables to the best one or two or three.

- Regression models should make sense. Choose predictors that are easy to understand. Avoid obscure variables. Remember, adding more predictors will always increase R-sq, but that doesn't mean we should always do it!

Adjusted R^2

- When we add another predictor to the model, R^2 can't go down. Adjusted R^2 is a rough attempt to adjust for this fact, and it incorporates somewhat of a penalty for adding more x-variables to the model. Below are the formulas:

$$R^2 = \frac{SS_{REGRESSION}}{SS_{REG} + SS_{RESIDUAL}}$$

$$= 1 - \frac{SS_{RESIDUAL}}{SS_{TOTAL}}$$

$$R^2_{ADJ} = 1 - \frac{MS_{RESIDUAL}}{MS_{TOTAL}}$$

Rationale: Mean Squares are just Sums of Squares divided by degrees of freedom – therefore, each new predictor won't necessarily increase adjusted R^2. Be careful interpreting it though, as adjusted R^2 does not represent the percentage of variation in y accounted for by the model (it can exceed 100% and dip below 0%).

Root Mean Squared Error (RMSE)

- RMSE is the square root of the average squared distance of a data point from the fitted line.
- If the multiple regression equation fits the data well, RMSE will be Small.
- RMSE will be in the same units as your y-variable.

Evaluating a Model (The Basics)

1. When selecting from a pool of potential x-variables to predict a y-variable, the final model should have a limited number predictors. Avoid overfitting your model.
2. Adjusted R2 should be high. StatCrunch ranks potential models based on this statistic.
3. RMSE should be low.

Example: Using what we know at this point, determine a reasonable model to predict "Calories" in the Chicken Sandwiches dataset.

With $j = 6$ possible x-variables, there are

$2^j - 1 = 2^6 - 1 = 64 - 1 = 63$ possible models.

With Fat in, common sense dictates Trans Fat & Saturated Fat out.

3rd Highest R^2_{ADJ} has Serving Size, Fat, Carbs, Sodium.

All 4 x's look linear against y = Calories.

Run Mult-Reg: Save Res & Predicted
y x

Check for Conditions. No Pattern ✓
Equal Spread ✓
No Outliers ✓

Test Main Hypothesis: $H_0: \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$
(Model is useful) H_A: At least one not zero

$F = 617.25$ P-Value ≤ 0.0001 Reject H_0,
Conclude Model is Useful.

Check if Residuals look normal enough
- skewed histogram, QQ not straight
But $n = 45$ data points so Proceed
with inference if desired

All 4 X-variables are statistically significant

Model: $\widehat{\text{Calories}} = -0.69 + 0.49(\text{Serving Size g}) + 7.89(\text{Fat g}) + 2.88(\text{Carb g}) + 0.06(\text{Sodium mg})$

Model has $R^2 = 98.41\%$
$R^2_{ADJ} = 98.25\%$
RMSE = 20.04 Calories
(On average, data points are about 20 calories away from predicted line)

Leverage, Outliers, and Influential Points

- A data point can be extraordinary in the y-direction, in the x-direction, or by having an unusual combination of values in the x-variables.
- Deviations in the y-direction show up in the residuals.
- Deviations in the x-direction show up as leverage/influence points

Leverage and Influence

- The leverage of a data point is its ability to move the regression model (slopes or intercept) all by itself just by moving in the y direction.
- We cannot see these points in our scatterplots because we are working in 3 or more dimensions.
- An influential point is one that substantially changes the regression model ***for your purposes***.
- To identify potential leverage and influential points, checkbox the "Cook's Distances" in the Multiple Regression menus.
- Cook's Distance Formula:

$$D_i = \frac{\sum_{j=1}^{n} \left(\hat{Y}_j - \hat{Y}_{j(i)} \right)^2}{p \cdot MSE}$$ with

$\hat{Y}_j$ = prediction from full model for case j

$\hat{Y}_{j(i)}$ = prediction for case j with case i removed

MSE = Mean Square Error

p = number of fitted parameters

- Some have suggested that any Cook's Distances that exceed $4/n$ should warrant further investigation (accuracy check, closer examination, collect more data near the point).

  n = number of cases

Example: Back to the chicken sandwiches analysis, our finalized model included serving size, fat, carbs, and sodium to predict calories. Run the analysis and analyze the Cook's Distances for potential influential points.

$n = 45$ sandwiches so any $D_i > \frac{4}{45} = 0.089$ should be looked at carefully.

Histo / Boxplot look for outliers

Dairy Queen Crispy : $D = 0.167$
Subway Oven Roasted : $D = 0.150$
Dairy Queen Grilled : $D = 0.149$
HARDEE'S Low Carb : $D = 0.139$
Carl's Jr Spicy : $D = 0.113$

High Leverage, Potential Influence.

Original : $R^2 = 98.41\%$, RMSE = 20.37

No DQ Crispy: $R^2 = 98.55\%$, RMSE = 19.31

⋮

No DQ, Sub, DQ : $R^2 = 98.94\%$, RMSE = 16.56

⋮

Remove all 5 : $R^2 = 99.16\%$, RMSE = 15.03

What's the right thing to do?

Residuals / Studentized Residuals

- Until now, we've looked at residuals "Straight Up", but then, what's a big residual?
- Take every residual, divide it by an estimate of its standard deviation, and you have what are known as Studentized Residuals.
- Tough one: Studentized residuals follow a Student's t model.
- Any studentized residual that stands out deserves attention.
- Checkbox the "Studentized Residuals" box on StatCrunch. Any value exceeding ±2 can generally be accepted as "far off the line".

Example: Continuing with Chicken Sandwiches, examine the studentized residuals.

Dairy Queen Grilled has $e_{student} = -2.75$

KFC Oven Toasted has $e_{student} = -2.02$

CARL'S JR Spicy has $e_{student} = -2.26$

All 3 sandwiches are much lower in calories than would be expected for their fat, serving size, carbs, and sodium.

Reasons?

Indicator Variables / Dummy Variables

- We can add categorical variables to a linear regression model that, until now, has only been able to handle quantitative variables.

- **Technique:** Suppose we had an additional variable in the Chicken Sandwiches dataset called "***Cheese***". It's a "yes" or "no" valued variable. Does cheese add calories? Probably. What about to a model that already includes serving size, fat grams, carbs, and sodium? We can code sandwiches with cheese as 1 and sandwiches without as 0. Then we run the regression analysis as before, treating this indicator variable as any other variable.

- We will get to an example with traditional indicator variables, but they are useful for another task.

- Code any potential outliers and influential points as indicator variables. Create a column in StatCrunch for each, putting 0s in for every row except the sandwich in question. Run the model with these variables included. If the *P*-values are small, it indicates that these sandwiches really don't fit with the rest of the bunch:

(1) Add 6 columns of 50 zeros to dataset
Data \ Sequence

(2) Run mult-regression with our 4 X variables and these 6 new dummy variables too.

(3) This model removes the effects of these sandwich's influence and the effects of outliers while making it clear these sandwiches are different

Final Model: For all sandwiches except the 6 special ones:

$$\widehat{\text{Calories}} = -0.827 + 0.536(\text{Serving Size}) + 7.538(\text{Fat Grams}) + 2.967(\text{Carb grams}) + 0.061(\text{Sodium mg})$$

Hardees could go either way.

The rest, add or subtract calories based on the coefficients

$R^2 = 99.41\%$

Root MSE = 13.23 calories.