Predictive Comparisons

Cole Brokamp

November 17th, 2016

Table of Contents

Sources

This presentation is based on:

- Gelman A, Pardoe I. Average predictive comparisons for models with nonlinearity, interactions, and variance components. Sociological Methodology. 2007 Dec 1;37(1):23-51
- White paper by David Chudzicki (davidchudzicki.com/predcomps)

What is a Predictive Comparison?

How to estimate expected change in outcome associated with a unit difference in one of the inputs?

- Simple case of linear regression without interactions or nonlinearity: regression coefficient
- Not valid for models with transformed outcome variables, interactions, polynomials, and other nonlinearities and nonadditivities
- Depends on the values of the predictors

Predictive Comparison should

- be defined for each input rather than each linear predictor
- allow for comparison of different models

Notation

Model: $p(y|x, \theta)$

- $x = (u, v)$
- u is the input of interest
- v is all other inputs

Predictive Comparison focuses on expected change in y corresponding to a specified change in $u(u^{(1)} \to u^{(2)})$

$$\delta_u(u^{(1)} \to u^{(2)}, v, \theta) = \frac{\mathbb{E}(y \mid u^{(2)}, v, \theta) - \mathbb{E}(y \mid u^{(1)}, v, \theta)}{u^{(2)} - u^{(1)}}$$

Disclaimers

- Assume $E(y|x, \theta)$ is a known function with an estimation of inferential uncertainty about θ
- Predictive comparison means summarizing predictive models, *NOT* estimating casual effects

Existing Methods

Direct Examination of Regression Coefficients

- Useful when directly interpretable, like coefficients of linear regression model without interactions
- Interactions prevent coefficients being interpreted as individual outputs

Define Predictive Comparisons at a Central Value

- Evaluate function at central value of x and perturb one input at a time
- Problems arise when input space is spread out such that no single central value is representative or inputs are binary/bimodal

Existing Methods

Transformed Coefficients

- Can sometimes clarify model interpretations (ex: exponentiated coefficients of a log regression model are multiplicative effects)
- Method here leads to interpretation on the original scale of the response variable (ex: probability scale rather than odds for logistic regression)
- Probabilities are more familiar and intuitive to work with than odds, although predictive comparisons can be used on the transformed scale

Standardized Coefficients

- Depend on sample variation in the inputs
- No meaning for categorical inputs
- Difficult to deal with input transformations and interactions

Generalizing Regression Coefficients

$$\hat{y} = f(u, v) = \mathbb{E}(y \mid u, v, \theta)$$

Given f and a choice of $u^{(1)}$, $u^{(2)}$, and v, the outcome change due to a one unit change in u while holding v constant is

$$\delta_{u^{(1)} \to u^{(2)}} = \frac{f(u^{(2)}, v) - f(u^{(1)}, v)}{u^{(2)} - u^{(1)}}$$

If f is a linear model with no interactions, the above does not depend on $u^{(1)}$, $u^{(2)}$, or v

Table of Contents

Average Over All Transitions

Average Predictive Comparison Δ_u is the mean value of δ_u over $u^{(1)}$, $u^{(2)}$, v, and θ

$$\delta_u = \frac{\mathbb{E}(y \mid u^{(2)}, v, \theta) - \mathbb{E}(y \mid u^{(1)}, v, \theta)}{u^{(2)} - u^{(1)}}$$

Average in both numerator and denominator for all increasing transitions of u:

$$\frac{\int \int_{u^{(1)} < u^{(2)}} du^{(1)} du^{(2)} \int dv \int d\theta (\mathbb{E}(y \mid u^{(2)}, v, \theta) - \mathbb{E}(y \mid u^{(1)}, v, \theta)) p(u^{(1)} \mid v) p(u^{(2)} \mid v) p(v) p(\theta)}{\int \int_{u^{(1)} < u^{(2)}} du^{(1)} du^{(2)} \int dv \int d\theta (u^{(2)} - u^{(1)}) p(u^{(2)} \mid v) p(v) p(\theta)}$$

Separate Numerator and Denominator

Compute the numerator and denominator seperately instead of taking $\mathbb{E}(\delta_{u^{(1)} \to u^{(2)}}, v)$

$$APC = \frac{\mathbb{E}(\Delta_f)}{\mathbb{E}(\Delta_u)}$$

where

- $\Delta_f = f(u^{(2)}, v) - f(u^{(1)}, v)$
- $\Delta_u = u^{(2)} - u^{(1)}$
- $\mathbb{E}$ is the expectation under the following process:
  1. sample v from (marginal) distributions of inputs
  2. sample $u^{(1)}$ and $u^{(2)}$ independently from distribution of u conditional on v

Equivalent to taking a weighted average of all δ_u as defined in the APC with weights $(u^{(2)} - u^{(1)})$

Estimate

$$\hat{\Delta}_u = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{s=1}^{S} w_{ij}[E(y \mid u_j, v_i, \theta_s) - E(y \mid u_i, v_i, \theta_s)]sign(u_j - u_i)}{\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{s=1}^{S} w_{ij}(u_j - u_i)sign(u_j - u_i)}$$

- Consider only increasing transitions of u in δ_u to avoid expected changes cancelling each other out
- Possible for predictive comparisons to be negative for some values of v and positive for other values of v such that it cancels out, so use root square or absolute values
- Use random subsets when n and s are large

Choosing Inputs

- We want to use representative values of v and transitions in u that are representative of what actually occurs at those values of v
- Observations represent samples from the joint distribution u, v i.e. each row is a sample from v followed by a sample $u^{(1)}$ conditional on v
- Difficult part is drawing another sample $u^{(2)}$ conditional on v
- To approximate, use weights $w_{ij} = w(v_i, v_j)$

Weights

- In theory, transitions are from points $u^{(1)}$ to $u^{(2)}$ with a common v
- This is unlikely to occur in reality, to approximate use weights $w_{ij} = w(v_i, v_j)$
- Reflects how likely it is for u to transition from u_i to u_j when $v = v_i$
- Goal is to approximate the distribution of $p(u^{(2)} \mid v)$ by giving higher weight to pairs of points with more similar v

Weighting Function

Gelman suggests using the following based on Mahalanobis distances:

$$w(v_i, v_j) = \left[1 + (v_i - v_j)^T \Sigma_v^{-1} (v_i - v_j)\right]^{-1}$$

- Mahalanobis distances is problematic for unordered categorical variables with more than one level
- Other proximity based weighting functions could be considered

Proposed Differences

Chudzicki:

- Use absolute APCs instead of RMS APCs (and always show signed and absolute versions alongside one another)
- Don't divide by difference in inputs
  - APC could be high but variation in input is so small that it doesn't make a difference
  - Allows for direct comparisons of APCs across different variables with different units
- Weights: use a subset of nearest observations for calculating weights, which can differ for start and end transition points

Table of Contents

Software

- R Package `predcomps`
- Collaboration with David Chudzicki
- Currently under active development, API likely to change

```
> library(predcomps)
> set.seed(6655)
> library(tidyverse)
```

Simple linear model

Simulate data

```
> n <- 200
> d <-
+   data_frame(x1 = runif(n,0,1),
+              x2 = runif(n,0,1),
+              x3 = runif(n,0,10)) %>%
+   mutate(y = 2 * x1 + (-2) * x2 + 1 * x3 +
+            rnorm(n,0,0.1))
```

Simple linear model

Fit linear model

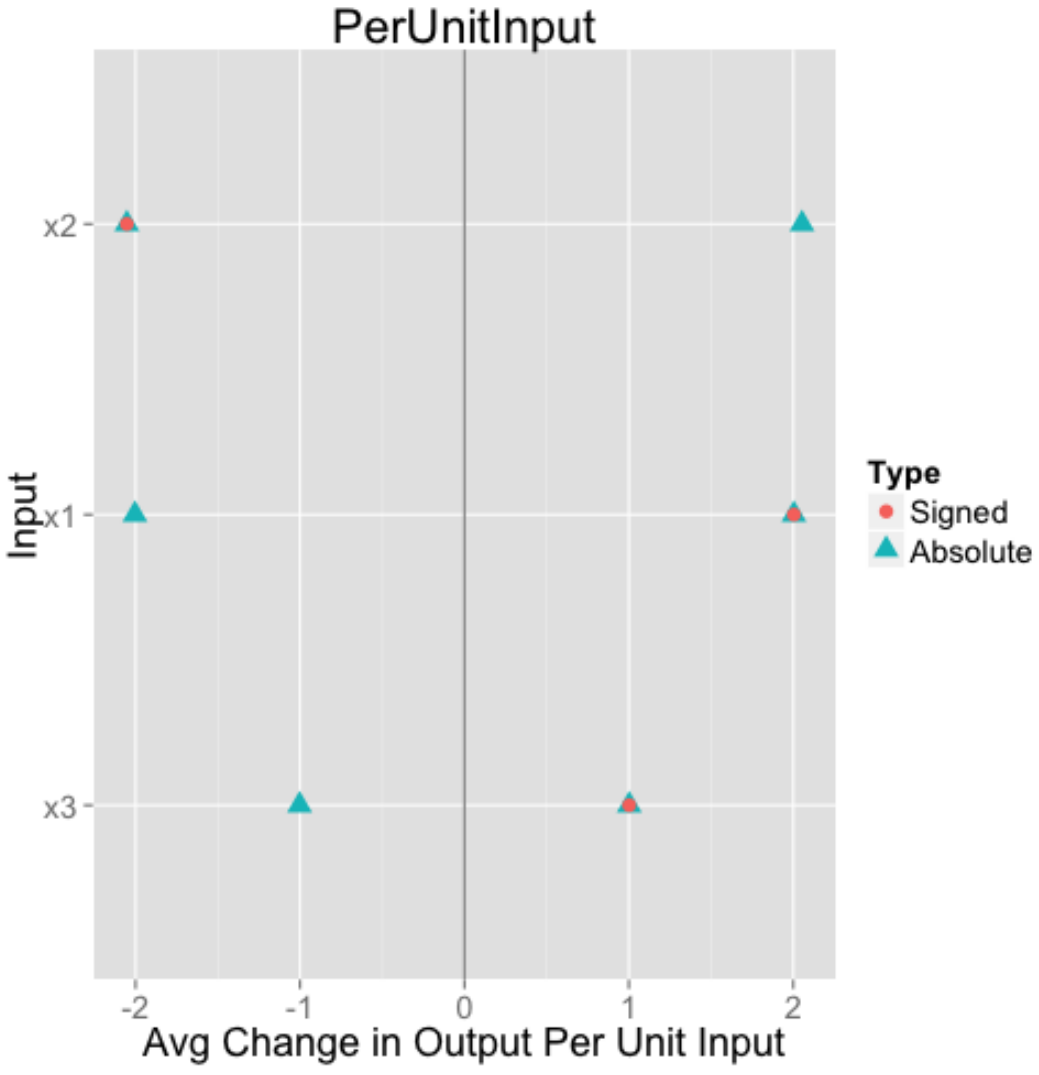
```
> lm_fit <- lm(y ~ ., data=d)
> coefficients(lm_fit) %>%
+   round(2)
```

```
(Intercept)          x1          x2          x3
       0.00        1.99       -1.98        1.00
```

Calculate and plot APCs

```
> GetPredCompsDF(lm_fit,df=d) %>%
+   PlotPredCompsDF(variant='PerUnitInput')
```

Simple linear model

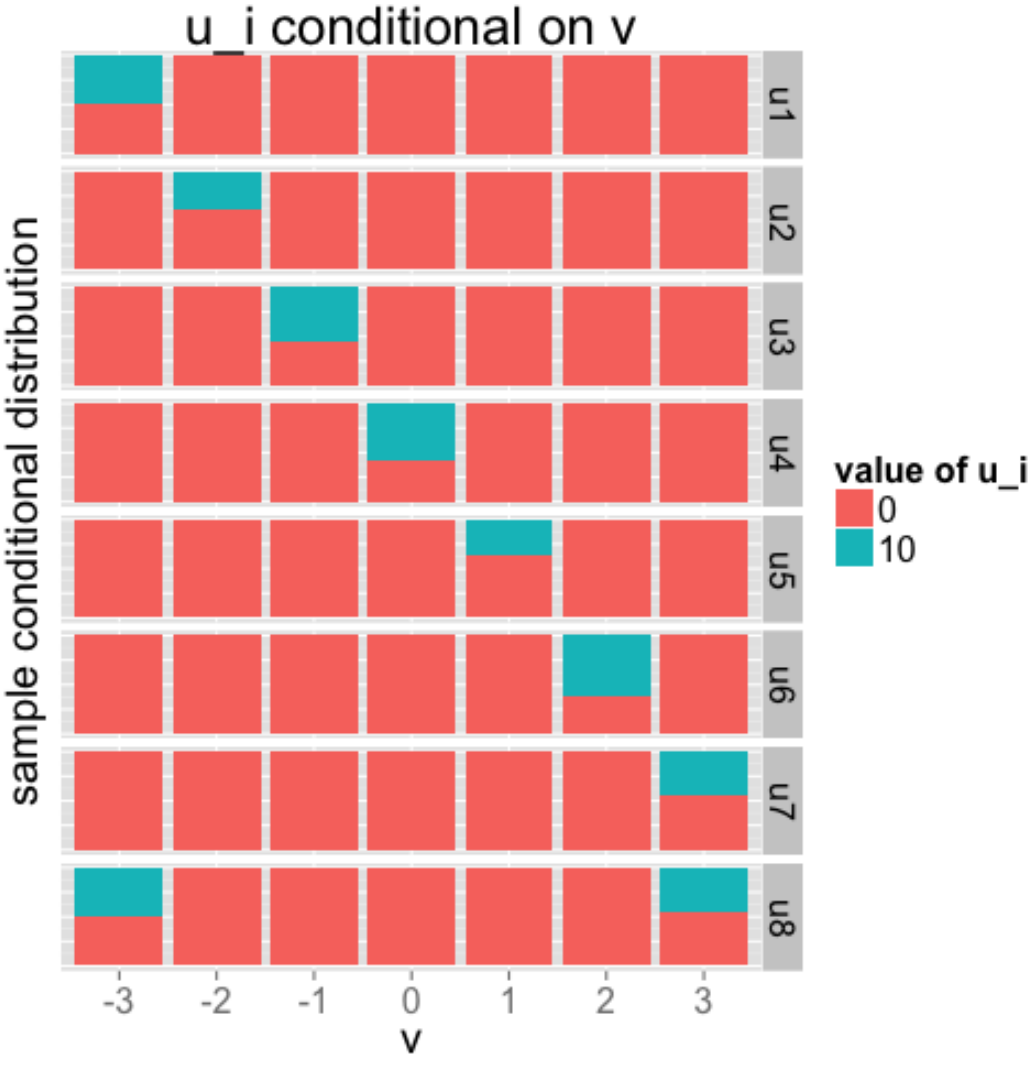


Linear model with interactions

Simulate data

- 9 inputs: v and $u_1, u_2, ..., u_8$
- v distributed uniformly among $-3, -2, ..., 2, 3$
- each u mostly constant at 0 but at one value of v ($v = -3$ for u_1, $v = -2$ for $u2$, etc), u can be either 0 or 10
- *except* input u_8: at either $v = -3$ or $v = 3$, u_8 can be 0 or 10; otherwise u_8 is 0

Linear model with interactions



Linear model with interactions

Simulate output; each u has the same role

$$\mathbb{E}[y] = vu_1 + vu_2 + vu_3 + vu_4 + vu_5 + vu_6 + vu_7 + vu_8$$

Differences in APCs will arise from input correlation structure, *not* from output function alone

Linear model with interactions

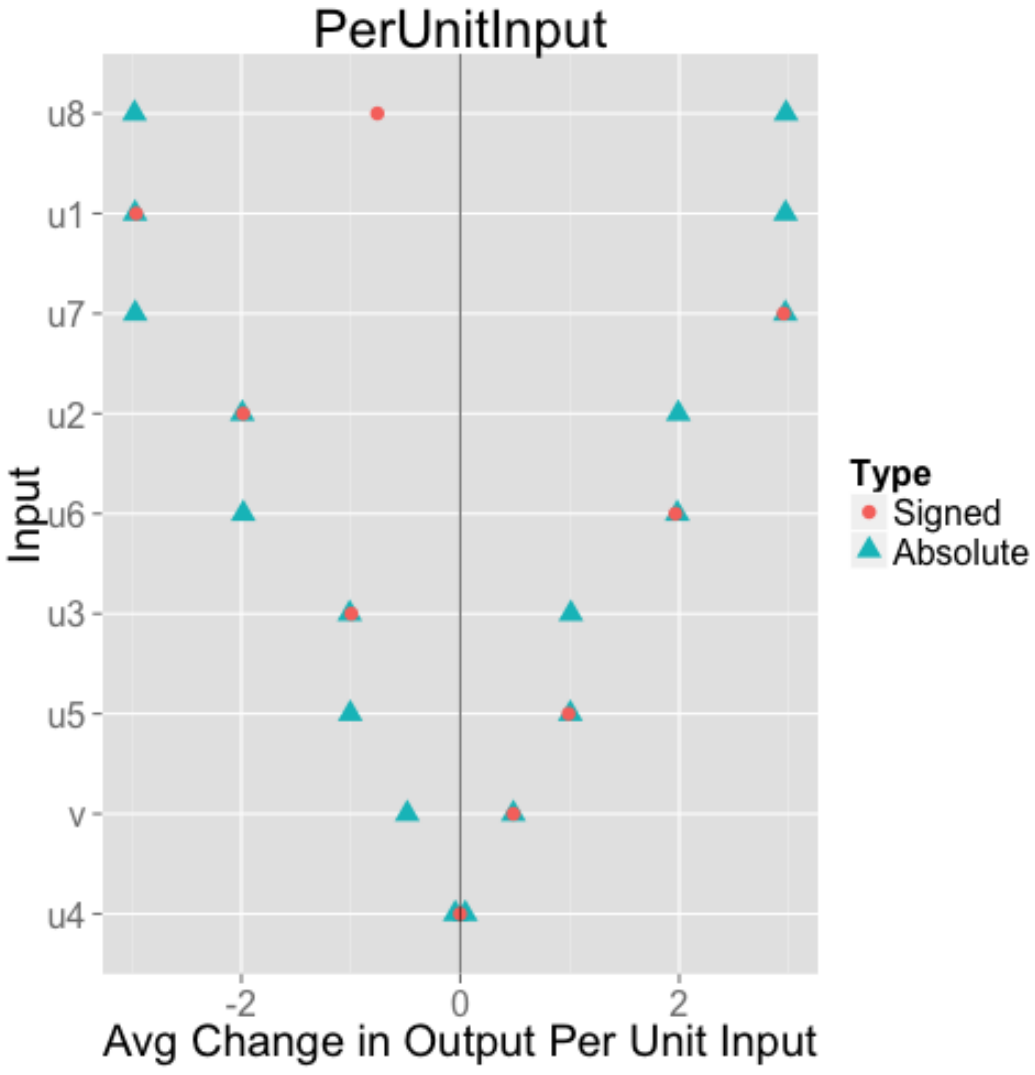


Table of Contents

Finding the best weighting function

- Gelman suggests using Mahalanobis distance, possible to include a constant term to increase weights for closer pairs
- Other possibilities could include random forest proximity, . . .
- Simulation experiments could compare the effect of using different weighting functions on accurately estimating the true predictive comparison value

APC for inference in random forest

- `RFinfer`: R package for random forest prediction variance
- Use to calculate prediction confidence intervals
- Quantifying uncertaintity in estimating θ allow for application of APC to random forest
- Future research
  - using APC for model inference
  - create new plots to interpret the effect of predictor
  - how does APC correspond with variable importance? out of bag error?

Random Forest Implementation

- Use proximity and model from same fitted forest?
- Use known asymptotic prediction behavior instead of sampling for θ_s
- Similar to work on Causal Forests by others:
  - Susan Athey and Guido Imbens. Machine learning methods for estimating heterogeneous causal effects. arXiv preprint arXiv:1504.01132, 2015.
  - Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. arXiv preprint arXiv:1510.04342, 2015.

Standard Errors

Variation in data is different from uncertainty in θ:

- Variation in x leads to variation in δ_u and average over this in estimating δ_u
- In contrast, uncertainty in θ should propogate to uncertainty in APCs
- Thus, treat u, v as fixed and θ as random
- Similar to inference for regression, where standard errors dervied from distribution of outcomes, conditional on inputs

Standard Errors

Compute using standard methods:

If $\hat{\delta}_u = \frac{1}{S}\sum_{s=1}^{S}\hat{\Delta}_u^s$, where

$$\hat{\Delta}_u^s = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}[E(y \mid u_j, v_i, \theta_s) - E(y \mid u_j, v_j, \theta_s)]sign(u_j - u_i)}{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(u_j - u_i)sign(u_j - u_i)}$$

then,

$$S.E.(\hat{\Delta}_u) = \left[\frac{1}{S-1}\sum_{s=1}^{S}(\hat{\Delta}_u^s - \hat{\Delta}_u)^2\right]^{\frac{1}{2}}$$

Thank You

cole.brokamp@cchmc.org

- Gelman A, Pardoe I. Average predictive comparisons for models with nonlinearity, interactions, and variance components. Sociological Methodology. 2007 Dec 1;37(1):23-51
- White paper by David Chudzicki (davidchudzicki.com/predcomps)
- https://github.com/dchudz/predcomps