Mean centering demo

From VoxBoWiki

Jump to: navigation, search

What happens when you mean center a covariate in a GLM, and how does that interact with the inclusion of an intercept? Below you'll find a short answer, a session in the statistical pacakge R that demonstrates the point, and finally, the R script. [R] is freely available for Linux, OSX, and Windows.

Summary

In case you just want to know the answer, here it is. Think of the mean of your signal as part of what your covariates are trying to model. There are four possibilities.

If you include an intercept and your covariates are mean centered (the typical model), then everything will be fine. The intercept will be weighted to explain the mean signal, and your independent variables will be scaled to explain variance in the signal.

If you omit the intercept term and your covariates are mean centered, then your model can't explain the mean at all, your error will be very high, and your t statistics will be low.

If you don't center your covariates, and you include the intercept, then your covariates will end up with the same weights and t values, but they will also end up accounting for some of the mean (unless by chance they end up with zero weights). So the intercept will no longer reflect the mean signal.

If you don't center your covariates and omit the intercept then your covariates will end up with high weights (to model the mean signal) and t-values. This correctly reflects the fact that your covariates do a good job of modeling the mean signal, although that is usually not the model you want to evaluate.

An interactive session in R

Here's a demo, in R, of all this in action.

> # What's the point of mean centering our GLM covariates? Let's first

> # set up some variables. Both the independent variable y and the

> # dependent variable x have non-zero means. cx is just x after

> # mean-centering.

>

> x <- rnorm(50)+4

> y <- rnorm(50)+4

> cx <- x-mean(x)

>

> # Double-check the means

>

> mean(x)

[1] 3.817591

> mean(y)

[1] 4.133391

> mean(cx)

[1] 1.687962e-16

>

> # Now let's try four different models. First with an intercept and

> # and a mean-centered independent variable (x).

>

> fm <- lm(y ~ 1+cx)

> summary(fm)

Call:

lm(formula = y ~ 1 + cx)

Residuals:

Min 1Q Median 3Q Max

-1.95933 -0.70649 -0.01367 0.68533 1.80542

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 4.133391 0.142810 28.943 <2e-16 \*\*\*

cx 0.007507 0.144429 0.052 0.959

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.01 on 48 degrees of freedom

Multiple R-Squared: 5.628e-05, Adjusted R-squared: -0.02078

F-statistic: 0.002702 on 1 and 48 DF, p-value: 0.9588

>

> # With an intercept and a mean centered covariate, the intercept

> # weight is the mean of the dependent variable and the beta/t/p for x

> # are modest, as they should be for randomly generated variables

>

> # Now let's try it without a non-centered x:

>

> fm <- lm(y ~ 1+x)

> summary(fm)

Call:

lm(formula = y ~ 1 + x)

Residuals:

Min 1Q Median 3Q Max

-1.95933 -0.70649 -0.01367 0.68533 1.80542

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 4.104732 0.569565 7.207 3.55e-09 \*\*\*

x 0.007507 0.144429 0.052 0.959

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.01 on 48 degrees of freedom

Multiple R-Squared: 5.628e-05, Adjusted R-squared: -0.02078

F-statistic: 0.002702 on 1 and 48 DF, p-value: 0.9588

>

> # Failing to mean-center the independent variable doesn't affect its

> # beta, t, or p value. But it does affect the intercept, which is no

> # longer the mean of y. This is because the weighting of x now also

> # contributes to the mean. (If you multiply the mean of x by its

> # coefficient, it comes out to the difference in the intercept

> # weight.) If you care about the intercept value, mean center your

> # covariates.

>

> # What happens if we omit the intercept, first with the mean-centered

> # independent variable (x):

>

> fm <- lm(y ~ 0+cx)

> summary(fm)

Call:

lm(formula = y ~ 0 + cx)

Residuals:

Min 1Q Median 3Q Max

2.174 3.427 4.120 4.819 5.939

Coefficients:

Estimate Std. Error t value Pr(>|t|)

cx 0.007507 0.614051 0.012 0.99

Residual standard error: 4.293 on 49 degrees of freedom

Multiple R-Squared: 3.05e-06, Adjusted R-squared: -0.02041

F-statistic: 0.0001495 on 1 and 49 DF, p-value: 0.9903

>

> # Without an intercept, we can use the mean-centered version of x and

> # get the same coefficient as before, because there's no way to scale

> # cx to model the mean signal. But because the error is much higher

> # (we're not modeling the mean at all), we get a very low t-value.

>

> # lastly, with no intercept and the offset independent variable:

>

> fm <- lm(y ~ 0+x)

> summary(fm)

Call:

lm(formula = y ~ 0 + x)

Residuals:

Min 1Q Median 3Q Max

-3.6665 -0.4518 0.3709 1.3366 3.0817

Coefficients:

Estimate Std. Error t value Pr(>|t|)

x 1.01513 0.05172 19.63 <2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.442 on 49 degrees of freedom

Multiple R-Squared: 0.8872, Adjusted R-squared: 0.8849

F-statistic: 385.3 on 1 and 49 DF, p-value: < 2.2e-16

>

> # x now has a mean, so the best least squares fit will give it a high

> # beta weight. It gets all the credit, so we get a ridiculously high

> # t-value. The statistic is correct, but it's usually the wrong model

> # for what we want to test.