

Multivariate Linear Regression

I am examining the number of nicotine dependence symptoms (response variable) as a function of smoking quantity, major depression, gender, and ethnicity.

Usage:

```
my.lm <- lm(QuantitativeResponse ~ Explanatory1 + Explanatory2 + ..., data =
  my.subset)
summary(my.lm)
```

```
> NDcount.lm <- lm(ND.symp.count ~ SmkQuant + MAJORDEPLIFE + SEX + ETHRACE2A,
+ data=nesarc.subset)
> summary(NDcount.lm)
```

Call:

```
lm(formula = ND.symp.count ~ SmkQuant + MAJORDEPLIFE + SEX +
  ETHRACE2A, data = nesarc.subset)
```

Residuals:

Min	1Q	Median	3Q	Max
-14.693	-2.619	-0.640	2.109	12.693

The coefficient is listed here, and can be used to interpret the direction (+/-) and strength (magnitude) of each explanatory variable.

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	3.00532	0.21156	14.205	<2e-16 ***
SmkQuant	0.15344	0.01089	14.088	<2e-16 ***
MAJORDEPLIFE1	3.04502	0.21830	13.949	<2e-16 ***
SEX2	-0.10783	0.18899	-0.571	0.5684
ETHRACE2A2	0.11651	0.28636	0.407	0.6842
ETHRACE2A3	0.28959	0.61273	0.473	0.6365
ETHRACE2A4	1.43270	0.50853	2.817	0.0049 **
ETHRACE2A5	-0.56437	0.24299	-2.323	0.0203 *

SmkQuant is significantly ($p < 2e-16$) and positively (+ coefficient) associated with ND.symp.count. Since SmkQuant is quantitative, the coefficient of 0.15 means that every additional cigarette smoked (one-unit increase in SmkQuant) is associated with a 0.15 increase in the number of ND symptoms.

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.746 on 1630 degrees of freedom
(4200 observations deleted due to missingness)

Multiple R-squared: 0.2194, Adjusted R-squared: 0.2161
F-statistic: 65.46 on 7 and 1630 DF, p-value: < 2.2e-16

The adjusted R^2 means that the combination of smoking quantity, major depression, gender, and ethnicity explains 21.61% of the variance in the number of nicotine dependence symptoms.

ETHRACE2A is a categorical variable, and the "5" after the variable name means that this line of the table refers to ethnicity group 5 compared to the reference group 1 (the lowest level, and the one missing from this table). There is a significant ($p = 0.0203$) and negative (-coefficient) difference in the number of nicotine dependence symptoms for group 5 relative to group 1. The coefficient of -0.56 means that Hispanics (ethnic group 5) on average experiences 0.56 fewer nicotine dependence symptoms compared to Whites (group 1.)

Multivariate Logistic Regression

I am examining the presence/absence of a nicotine dependence diagnosis (response variable) as a function of smoking quantity, major depression, gender, and ethnicity.

Usage:

```
my.glm <- glm(QuantitativeResponse ~ Explanatory1 + Explanatory2 + ..., data
              = my.subset, family="binomial")
summary(my.glm)
exp(my.glm$coefficients)      # For odds ratios
exp(confint(my.glm))         # For 95% confidence intervals of odds ratios
```

```
> ND.glm <- glm(TAB12MDX ~ SmkQuant + MAJORDEPLIFE + SEX + ETHRACE2A,
+               data=nesarc.subset, family="binomial")
> summary(ND.glm)
```

Call:

```
glm(formula = TAB12MDX ~ SmkQuant + MAJORDEPLIFE + SEX + ETHRACE2A,
     family = "binomial", data = nesarc.subset)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-2.7073	-1.0087	-0.7156	1.0716	1.7547

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-1.064749	0.119024	-8.946	< 2e-16 ***
SmkQuant	0.066474	0.006553	10.144	< 2e-16 ***
MAJORDEPLIFE1	0.909948	0.112860	8.063	7.47e-16 ***
SEX2	-0.010700	0.099188	-0.108	0.9141
ETHRACE2A2	0.088517	0.151606	0.584	0.5593
ETHRACE2A3	0.220670	0.314691	0.701	0.4832
ETHRACE2A4	0.469801	0.268380	1.751	0.0800 .
ETHRACE2A5	-0.299834	0.130988	-2.289	0.0221 *

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 2646.6 on 1913 degrees of freedom
 Residual deviance: 2427.7 on 1906 degrees of freedom
 (3924 observations deleted due to missingness)
 AIC: 2443.7

Number of Fisher Scoring iterations: 4

```
> exp(ND.glm$coefficients)
```

	SmkQuant	MAJORDEPLIFE1	SEX2	ETHRACE2A2
(Intercept)	0.3448143	2.4841938	0.9893570	1.0925531
ETHRACE2A3	1.2469122			
ETHRACE2A4	1.5996764			
ETHRACE2A5	0.7409412			

The coefficient is listed here, and can **only be used to interpret the direction** (+/-). The coefficients need to be converted to odds ratios before the strength of the relationship can be interpreted.

Look first at the p-value for to see whether each term is significant before going on to odds ratios. No need to discuss odds ratios if the p-value doesn't show a significant relationship

These are the odds ratios, which are obtained by exponentiating the coefficients from the above table. Odds ratios are easier to interpret

Since this is a categorical variable, first figure out which level it's referring to, and what the reference group is (which is always missing from the results). This odds ratio refers to ethnic group 5 relative to group 1 (the reference group). This indicates that Hispanics (group 5) are only 74% as likely as Whites (group 1) to have nicotine dependence.

Since the above table showed that smoking quantity is significantly associated with an increase in the odds of having nicotine dependence, this odds ratio means that for every additional cigarette smoked (a 1-unit increase in this quantitative variable), one's odds of nicotine dependence are 1.07 times greater. In other words, the odds of nicotine dependence increase by 7% with every additional cigarette smoked.

```
> exp(confint(ND.glm))
```

waiting for profiling to be done...

		2.5 %	97.5 %
(Intercept)	0.2724703	0.4345419	
SmkQuant	1.0552795	1.0827431	
MAJORDEPLIFE1	1.9934072	3.1032123	
SEX2	0.8144679	1.2016454	
ETHRACE2A2	0.8110362	1.4701558	
ETHRACE2A3	0.6752027	2.3345977	
ETHRACE2A4	0.9454660	2.7183982	
ETHRACE2A5	0.5724646	0.9569247	

These are the 95% confidence intervals for the odds ratios. That is, when you're making inferences about the larger (unobserved) population, they give the range that is likely to contain the true population odds ratio.

Even though the odds ratio for **your sample** is 1.07, the true population odds ratio might be slightly different, due to random variation in sampling. The odds ratio indicates that there's a 95% certainty that the **true population odds ratio** falls somewhere between 1.055 and 1.082.

The odds ratio for Hispanics (group 5) vs. Whites (reference group) among your sample is 0.74, but the true population odds ratio falls somewhere between 0.57 and 0.96.