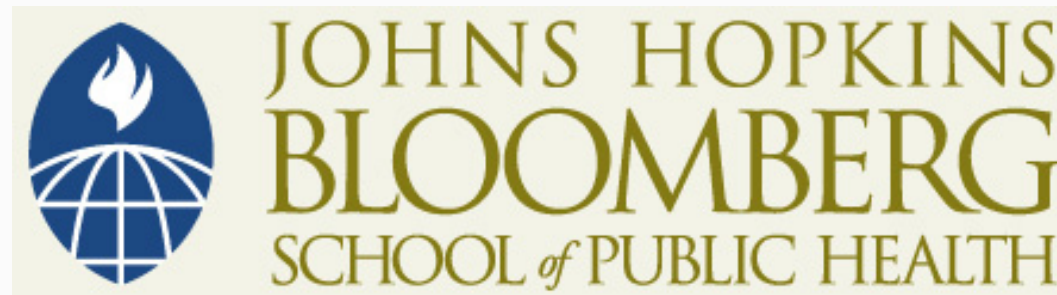


This work is licensed under a [Creative Commons Attribution-NonCommercial-ShareAlike License](https://creativecommons.org/licenses/by-nc-sa/4.0/). Your use of this material constitutes acceptance of that license and the conditions of use of materials on this site.



Copyright 2009, The Johns Hopkins University and John McGready. All rights reserved. Use of these materials permitted only in accordance with license rights granted. Materials provided "AS IS"; no representations or warranties provided. User assumes all responsibility for use, and all liability related thereto, and must independently review all materials for accuracy and efficacy. May contain materials owned by others. User is responsible for obtaining permissions for use from third parties as needed.



JOHNS HOPKINS
BLOOMBERG
SCHOOL *of* PUBLIC HEALTH

Section D

Handling Multiple Categorical Predictors in Multiple Linear Regression: ANOVA as a Regression Model

The Situation

- Sometimes regression scenarios include predictors that are not continuous, not binary, but multi-categorical
- Examples
 - Subject's race (White, African-American, Hispanic, Asian, Other)
 - City of residence (Baltimore, Chicago, Tokyo, Madrid)

The Situation

- How can this type of situation be handled in a regression framework?
- We'll explore with an example using a data set containing information about average SAT scores in 51 U.S. states (treating D.C. as a state)—the averages were based on random samples of students taken within each of the 51 states

SAT Scores Example

- The SAT (Scholastic Aptitude Test) is taken by many U.S. high school students to fulfill requirements for entry into most colleges or universities
- The test is made up of two components: verbal and quantitative (math)
- This analysis will use the quantitative score, which ranges from 200-800 (we will refer to these simply as SAT scores for simplicity)
- This data comes from the book *Statistics with Stata 8*, by Lawrence Hamilton

SAT Scores Example

- Data consists of 51 observations: the cumulative average SAT quantitative section scores for the 51 U.S. states for students taking the test in 1990
- Additional information on each observation includes geographical region of the state (West, Northeast, South, Midwest) and per-pupil education expenditures in each state in 1990

SAT Scores Example

- A key question
 - Do average SAT scores differ across the four regions of the country and, if so, what is the magnitude of these differences?

Snippet of the Data

- Here is a snippet of the data in Stata (msat is mean math SAT score for the state, region is a “labeled” numerical variable)

```
      +-----+
      | msat    region |
      |-----|
1.   |  515     South  |
2.   |  481     West   |
3.   |  490     West   |
4.   |  523     South  |
5.   |  482     West   |
      |-----|
6.   |  506     West   |
7.   |  468     N. East |
```

ANOVA

- Analysis of variance testing for differences between the four states yields a p-value of less than 0.001
- So there are at least some statistical differences in SAT quantitative scores across the four regions—but in order to find out which regions are statistically different and to figure out by how large (and in what direction) these differences occur would require a lot of t-tests

ANOVA as a Regression Model

- Could this analysis be done by a linear regression relating SAT scores to region?
- How can we handle a predictor that takes on four categories?

ANOVA as a Regression Model

- Arbitrarily give each region a numerical value ($x_1 = 1$ for Western region states, 2 for Northeastern states, 3 for Southern states, and 4 for mid-Western states for example) and fit SLR of

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1$$

- Where \hat{y} is estimated mean SAT score, and x_1 is region as defined above

ANOVA as a Regression Model

- This is not a good idea!!!
- Coding is arbitrary, could have assigned $x_1 = 1$ for Midwest, etc. . . .
 - Estimated coefficient of region will depend on arbitrary coding
- Coding “assumes” mean SAT score differences between regions is “incremental”
 - Example—difference in average SAT scores between Southern states ($x_1 = 3$) and Western States ($x_1 = 1$) is twice the difference between Northeastern States ($x_1 = 2$) and Western States ($x_1 = 1$)

ANOVA as a Regression Model

- Alternative approach—designate one region as “reference” region, say Western region, and make binary indicators for each of the three other regions
 - $x_1 = 1$ if Northeastern state, 0 otherwise
 - $x_2 = 1$ if Southern state, 0 otherwise
 - $x_3 = 1$ if mid-Western state, 0 otherwise

ANOVA as a Regression Model

- Here is a table showing the x values for each region

Region	x_1	x_2	x_3
West	0	0	0
Northeast	1	0	0
South	0	1	0
Mid-west	0	0	1

ANOVA as a Regression Model

- Fit the regression model

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3$$

- Here, each coefficient estimates mean SAT score difference between a region that has a corresponding x value of 1 and the reference region (Western states)
- Notice, the intercept has meaning here—it's the estimated mean when all x s are 0, the estimated mean SAT score for Western states!

ANOVA as a Regression Model

- Example

- For Northeastern states ($x_1 = 1, x_2 = 0, x_3 = 0$) model predicts

$$\begin{aligned}\hat{y} &= \hat{\beta}_0 + \hat{\beta}_1 * 1 + \hat{\beta}_2 * 0 + \hat{\beta}_3 * 0 \\ &= \hat{\beta}_0 + \hat{\beta}_1\end{aligned}$$

- For Western states ($x_1 = 0, x_2 = 0, x_3 = 0$) model predicts

$$\begin{aligned}\hat{y} &= \hat{\beta}_0 + \hat{\beta}_1 * 0 + \hat{\beta}_2 * 0 + \hat{\beta}_3 * 0 \\ &= \hat{\beta}_0\end{aligned}$$

ANOVA as a Regression Model

- Example

- So $\hat{y}_{NE} - \hat{y}_W = \hat{\beta}_0 + \hat{\beta}_1 - \hat{\beta}_0 = \hat{\beta}_1$

- Similar results can be shown for other coefficients

ANOVA as a Regression Model

- Stata results
- Notice, data in the following format . . .

```
+-----+
      msat  region
-----
1.    515      3
2.    481      1
3.    490      1
4.    523      3
5.    482      1
-----
6.    506      1
7.    468      2
8.    464      3
```

ANOVA as a Regression Model

- “xi” option before regression command will automatically create binary indicators for a multi-categorical variable
- Syntax
 - xi: regress msat i.region

ANOVA as a Regression Model

■ Stata results

```
. xi: regress msat i.region
i.region          _Iregion_1-4          (naturally coded; _Iregion_1 omitted)

-----+-----
Source |           SS          df           MS                Number of obs =      50
-----+-----
Model | 26433.6728           3    8811.22428                F( 3, 46) = 12.26
Residual | 33050.8072          46    718.495808                Prob > F      = 0.0000
-----+-----
Total | 59484.48            49   1213.96898                R-squared     = 0.4444
                                           Adj R-squared = 0.4081
                                           Root MSE    = 26.805

-----+-----
msat |           Coef.    Std. Err.      t    P>|t|     [95% Conf. Interval]
-----+-----
_Iregion_2 | -32.01709    11.62333     -2.75   0.008   -55.41364   -8.620547
_Iregion_3 | -11.64904    10.00874     -1.16   0.250   -31.79559    8.497512
_Iregion_4 | 35.45513     10.7305      3.30   0.002    13.85576    57.0545
   _cons | 498.4615     7.434306    67.05   0.000    483.4971   513.426
-----+-----
```

Stata Results

- Resulting regression equation

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3$$

$$\hat{y} = 498.5 - 32.0x_1 - 11.6x_2 + 35.5x_3$$

ANOVA as a Regression Model

■ Overall F-test

```
. xi: regress msat i.region
i.region          _Iregion_1-4      (naturally coded; _Iregion_1 omitted)
```

Source	SS	df	MS	Number of obs =	50
Model	26433.6728	3	8811.22428	F(3, 46) =	12.26
Residual	33050.8072	46	718.495808	Prob > F =	0.0000
Total	59484.48	49	1213.96898	R-squared =	0.4444

Adj R-squared = 0.4081
Root MSE = 26.805

msat	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
_Iregion_2	-32.01709	11.62333	-2.75	0.008	-55.41364 -8.620547
_Iregion_3	-11.64904	10.00874	-1.16	0.250	-31.79559 8.497512
_Iregion_4	35.45513	10.7305	3.30	0.002	13.85576 57.0545
_cons	498.4615	7.434306	67.05	0.000	483.4971 513.426

ANOVA as a Regression Model

- This is the overall test for . . .
 - $H_0: \beta_1 = \beta_2 = \beta_3$: no differences in mean SAT scores across the four regions
 - H_a : at least one region has different mean SAT scores than the others
 - This is the same exact test that we did with the traditional ANOVA approach

ANOVA as a Regression Model

- Some of the estimated regional differences

```
. xi: regress msat i.region
i.region      _Iregion_1-4      (naturally coded; _Iregion_1 omitted)

      Source |           SS          df           MS              Number of obs =      50
-----+-----+-----+-----+-----+-----+-----+-----
      Model |    26433.6728         3    8811.22428          F( 3, 46) =     12.26
      Residual |   33050.8072        46    718.495808          Prob > F      =  0.0000
-----+-----+-----+-----+-----+-----+-----
      Total |    59484.48         49   1213.96898          R-squared      =  0.4444
                                          Adj R-squared  =  0.4081
                                          Root MSE     =  26.805
```

msat	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
_Iregion_2	-32.01709	11.62333	-2.75	0.008	-55.41364	-8.620547
_Iregion_3	-11.64904	10.00874	-1.16	0.250	-31.79559	8.497512
Iregion_4	35.45513	10.7305	3.30	0.002	13.85576	57.0545
_cons	498.4615	7.434306	67.05	0.000	483.4971	513.426

Results

- A statistically significant relationship was found between mean SAT scores and student's region of the country ($p < .0001$ by F-test)
- Students from northeastern states had SAT scores of 32 points lower on average than students from western states (95% CI 8.6 to 55.4 points lower)

Results

- Students from southern states had SAT scores of 11.6 points lower on average than students from western states (95% CI 31.6 points lower to 8.5 points higher)
- Students from mid-western states had SAT scores of 35.5 points higher on average than students from western states (95% CI 13.9 points to 57.0 points higher)
- Regional differences account for 44% of the variation in SAT scores

Results

- What about other comparisons—for example, SAT scores for Northeastern states to mid-western states?
 - One approach—recode indicators for region making “mid-west” the reference group—more work!
 - Another option—use existing coefficients

Results

- Recall $\hat{\beta}_1 = \hat{y}_{NE} - \hat{y}_W = -32.0$ estimates the average difference in SAT scores for northeastern states minus (compared to) western states
- Recall $\hat{\beta}_3 = \hat{y}_{MW} - \hat{y}_W = 35.5$ estimates the average difference in SAT scores for mid-western states minus (compared to) western states

Results

- So :

$$\begin{aligned}\hat{\beta}_1 - \hat{\beta}_3 &= \hat{y}_{NE} - \hat{y}_W - (\hat{y}_{MW} - \hat{y}_W) \\ &= \hat{y}_{NE} - \hat{y}_W - \hat{y}_{MW} + \hat{y}_W \\ &= \hat{y}_{NE} - \hat{y}_{MW}\end{aligned}$$

- So the estimated mean difference in SAT scores between northeastern states and mid-western states is given by $(-32.0-35.5)$
= -67.5 points

Results

- We can employ Stata to do this and get a 95% CI (just FYI)
- The “lincom” command can be run after any regression to give estimates for differences in coefficients

ANOVA as a Regression Model

- We need to use names Stata gives to coefficients in command

```
. xi: regress msat i.region
i.region      _Iregion_1-4      (naturally coded; _Iregion_1 omitted)

-----+-----
Source |           SS          df           MS              Number of obs =      50
-----+-----
Model  |    26433.6728         3    8811.22428          F( 3, 46) =      12.26
Residual |    33050.8072        46    718.495808          Prob > F      =    0.0000
-----+-----
Total  |     59484.48         49   1213.96898          R-squared     =    0.4444
                                           Adj R-squared =    0.4081
                                           Root MSE     =    26.805

-----+-----
msat |           Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-----+-----
_Iregion_2 |   -32.01709    11.62333     -2.75   0.008   -55.41364   -8.620547
_Iregion_3 |   -11.64904    10.00874     -1.16   0.250   -31.79559    8.497512
_Iregion_4 |    35.45513    10.7305     3.30   0.002    13.85576    57.0545
   _cons |    498.4615     7.434306    67.05   0.000    483.4971    513.426
-----+-----
```

ANOVA as a Regression Model

- Syntax
 - `lincom _Iregion_2 - _Iregion_4`

```
. lincom _Iregion_2- _Iregion_4
```

```
(1) _Iregion_2 - _Iregion_4 = 0
```

msat	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
(1)	-67.47222	11.81979	-5.71	0.000	-91.26423	-43.68021

Recap

- ANOVA is just a specific form of linear regression
- In general, if we have a categorical predictor with k categories, we designate one category as the reference group and create $k-1$ binary indicators x_1, x_2, x_{k-1} for all other levels of the predictor
- Coefficients are interpretable as mean difference in the outcome between each of the $k-1$ categories and the reference group

Advantages

- Not only do we get an overall test for any mean outcome differences between the groups being compared, we also get estimates and 95% CIs for some of the differences
- This approach also gives a R^2 value
- We can also expand the regression model to include more predictors (example—SAT scores predicted by both region and per-pupil state expenditures)