ANOVA Regression Overview

PROC ANOVA, PROC REG, and PROC GLM

Shannon Pileggi

STAT 330

1 / 29

STAT 330: Lecture 13

ANOVA Overview Regression

OUTLINE

Overview

ANOVA

Regression

STAT 330: Lecture 13 2 / 29

Overview ANOVA Regression •000000

Methods overview

ANOVA (analysis of variance)

- ► Dependent variable = quantitative independent variable = categorical (more than 2 levels)
- ▶ interested in comparing >2 groups

 $H_0: \mu_1 = \mu_2 = \mu_3 = \dots \mu_g$ for g groups

 H_a : at least one mean is different than the others

Linear Regression

- ► Dependent variable = quantitative independent variable(s) = quantitative or categorical
- ▶ interested in examining the relationship between x and y

 H_0 : $\beta_1 = 0$ vs H_a : $\beta_1 \neq 0$

3 / 29

Overview ANOVA Regression 000000

PROCs Overview

All can be used to model a *quantitative* dependent variable.

PROC REG

- ▶ simple linear regression
- polynomial regression
- regression with multiple predictors

PROC ANOVA

- analysis of variance (for balanced data)
- multivariate analysis of variance (MANOVA)
- repeated measures analysis of variance

PROC GLM

- ▶ simple regression
- multiple regression
- ▶ analysis of variance
- analysis of covariance
- response-surface models
- weighted regression
- polynomial regression
- partial correlation
- multivariate analysis of variance

4 / 29

repeated measures analysis of variance

STAT 330: Lecture 13

 Overview
 ANOVA
 Regression

 00●000
 00000000
 000000000

Overview, continued

- ► PROC GLM can do the same type of analyses as PROC REG and PROC ANOVA
- ► PROC REG and PROC ANOVA allow you to do more detailed analysis related specifically to regression and ANOVA, respectively
- ▶ all procedures have their quirks...

4□ > 4₫ > 4 ½ > ½
5 / 20

4□ > 4個 > 4 = > 4 = > = 9 < 0</p>

7 / 29

STAT 330: Lecture 13 5 / 29

ANOVA

Regression 0000000000

The Data

STAT 330: Lecture 13

Overview

0000000

Collected from Kelly Blue Book for 2005 used GM cars

suggested retail price Price number of miles the car has been driven Mileage Make manufacturer of the car Model specific models for each car manufacturer specific type of car model Trim body type Type Cylinder number of cylinders in the engine a more specific measure of engine size Liter number of doors Doors whether the car has cruise control (Y/N)Cruise indicator for upgraded speakers (1 = upgraded)Sound indicator for leather seats (1 = leather)Leather

 Overview
 ANOVA
 Regression

 000 000
 00000000
 000000000

Some quirks

Feature	PROC REG	PROC ANOVA	PROC GLM
Dependent var	quantitative	quantitative	quantitative
Quantitative indepen-	✓	X	✓
dent var			
Categorical indepen-	must be coded as	use CLASS state-	use CLASS state-
dent var	indicator variables	ment	ment
	in the data set (no		
	CLASS statment)		
Higher order terms	must be coded in	can be written in	can be written in
(e.g., squares, inte-	the data set	MODEL statement	MODEL statement
ractions)			
Multiple MODEL state-	✓	X	X
ments			
Parameter estimates	automatic	N/A	may need to use
			SOLUTION option
			if have categorical
			ind var

4□▶ 4□▶ 4 = ▶ 4 = ▶ 9 < 0</p>

8 / 29

STAT 330: Lecture 13

Overview

0000000

ANOVA Regression

Get started

```
SAS Code

LIBNAME flash "&path";

PROC CONTENTS DATA = flash.cars VARNUM;

RUN;

PROC MEANS DATA = flash.cars;

VAR price mileage liter;

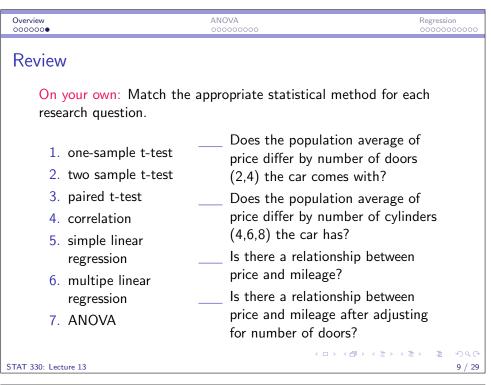
RUN;

PROC FREQ DATA = flash.cars;

TABLES make type cylinder doors cruise sound leather;

RUN;

SAS Code
```





Which figure would you produce to examine the relationship between price and number of cylinders (4,6,8)?

←□ → ←□ → ←□ → ←□ → □

11 / 29

- 1. histogram
- 2. single boxplot
- 3. side by side boxplot
- 4. scatter plot

```
Regression
                                  00000000
Syntax
                                  SAS Code _
        PROC ANOVA DATA = mydata ;
             CLASS catvar;
             MODEL quantuar = catvar ;
             MEANS catvar / options;
        QUIT ;
                                  _ SAS Code _
       ► CLASS - specify categorical independent variable (treatment)
       ► MODEL - specify relationship
       ▶ MEANS - estimates means for all levels of CLASS variable
                                                   ◆ロト ◆個 ト ◆ 恵 ト ◆ 恵 ・ 夕 へ ○
STAT 330: Lecture 13
                                                                          12 / 29
```

 Overview 000000
 ANOVA 000000
 Regression 0000000000

PROC ANOVA Example

```
PROC ANOVA DATA = flash.cars;
CLASS cylinder;
MODEL price = cylinder;
MEANS cylinder;
QUIT;
SAS Code
```

On your own: What is the next step in this analysis?



		Pri	ice
Level of Cylinder	N	Mean	Std Dev
4	394	17862.5649	7830.9838
6	310	20081.3958	4631.2230
8	100	38968.0432	10732.3323

13 / 29

 Overview
 ANOVA
 Regression

 0000000
 0000●0000
 0000000000

Conditions for one-way ANOVA

- 1. observations are independent (in each of the g groups)
- 2. normal underlying population distribution OR $n \ge 30$ in each group
- 3. each group has (about) the same variability (equal variance)

On your own: How would you check these conditions?

 ✓ □ ➤ ✓ ⓓ ➤ ✓ ໙ ▷ ✓ ໙ ▷
 E
 ✓ ℚ へ

 STAT 330: Lecture 13
 14 / 29

 Overview
 ANOVA
 Regression

 000000
 00000●000
 000000000

MEANS options

STAT 330: Lecture 13

- ► Many multiple comparison methods available: TUKEY, SCHEFFE, DUNCAN, BON (for Bonferroni).
- ► ALPHA= controls overall error rate
- ► To test the assumption of equal variance, use the HOVTEST option (Homogeneity Of Variance test). For this test, the null hypothesis is H₀: Variances are equal. Smaller p-values lend stronger evidence against this statement.

 Overview
 ANOVA
 Regression

 000000
 000000●00
 000000000

PROC ANOVA Example, continued

PROC ANOVA DATA = flash.cars;
CLASS cylinder;
MODEL price = cylinder;
MEANS cylinder
/ TUKEY HOVTEST;
QUIT;
SAS Code

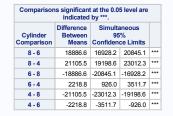
 Levene's Test for Homogeneity of Price Variance ANOVA of Squared Deviations from Group Means

 Source
 DF
 Squares
 Square
 F Value
 Pr > F

 Cylinder
 2
 7.084E17
 3.542E17
 38.44
 <.0001</td>

 Error
 801
 7.381E18
 9.215E15

On your own: Is the equality of variance condition satisfied? For which cylinder comparisons do we have evidence of a difference in population mean price?



←□ → ←□ → ← □ → □ → □ ←
 16 / 29

◆□▶◆□▶◆□▶◆□▶ □ める

15 / 29

STAT 330: Lecture 13

ANOVA Regression Overview 000000000

The same analysis, but with PROC GLM

SAS Code

PROC GLM DATA = flash.cars : CLASS cylinder; MODEL price = cylinder; QUIT ; _ SAS Code _

- ▶ same base output as PROC ANOVA
- requires much more work to get multiple comparisons

4□ > 4回 > 4 = > 4 = > = 9 < ○</p> 17 / 29

ANOVA Overview Regression 00000000

Warning

- ▶ the ANOVA procedure is designed to handle balanced data (groups with equal sample sizes)
 - ▶ for one-way ANOVA, PROC ANOVA still works ok even for unbalanced data
- ▶ if you have unbalanced data and you want to do something more complex than one-way ANOVA, use PROC GLM

Was PROC ANOVA valid for analyzing the relationship between price and cylinder?

- 1. Yes
- 2. No

18 / 29

STAT 330: Lecture 13

Overview ANOVA Regression •0000000000

STAT 330: Lecture 13

Regression

Overview ANOVA Regression 0000000000

Review

Which figure would you produce to examine the relationship between price and mileage?

- 1. histogram
- 2. single boxplot
- 3. side by side boxplot
- 4. scatter plot

◆ロト ◆個ト ◆夏ト ◆夏ト 夏 りへで

19 / 29

◆ロト ◆個 ト ◆ 恵 ト ◆ 恵 ・ 夕 へ ○ 20 / 29 STAT 330: Lecture 13

 Overview
 ANOVA
 Regression

 0000000
 000000000
 0000000000

Relationship between price and mileage

Both PROC REG and PROC GLM can be used for simple linear regression with a quantitative independent variable.

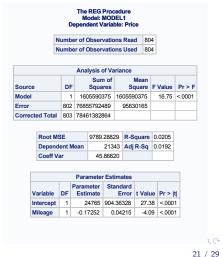
```
PROC REG DATA = flash.cars;
MODEL price = mileage;
QUIT;

PROC GLM DATA = flash.cars;
MODEL price = mileage;
QUIT;

SAS Code
```

Overview

STAT 330: Lecture 13



STAT 330: Lecture 13 21 /

Other features of PROC REG - create data set with

ANOVA

residuals and other diagnostic measures

```
PROC REG DATA = flash.cars ;
   MODEL price = mileage ;
   OUTPUT OUT = reg_results PREDICTED = yhat RESIDUAL = resid ;
QUIT ;

PROC PRINT DATA = reg_results (obs = 5) ;
   VAR price mileage yhat resid ;
RUN ;

SAS Code
```

Obs	Price	Mileage	yhat	resid
1	17314.10313	8221	23346.27	-6032.16
2	17542.03608	9135	23188.58	-5646.55
3	16218.84786	13196	22487.98	-6269.13
4	16336.91314	16342	21945.23	-5608.32
5	16339.17032	19832	21343.13	-5003.96

4 □ → 4 ∰ → 4 ½ → 4 ½ → ½ · √ Q (~ 23 / 29

Regression

00000000000

 Overview
 ANOVA
 Regression

 0000000
 00000000
 000●0000000

Other features of PROC REG - multiple model statements

```
PROC REG DATA = flash.cars ;

MODEL price = mileage ;

MODEL price = liter ;

QUIT ;

SAS Code
```

 ✓□ > ✓□ > ✓□ > ✓ □ >

 Overview
 ANOVA
 Regression

 000000
 00000000
 000000000

PROC REG confidence limits

MODEL quantvar = independent var(s) / options;

Confidence interval options include:

- ▶ CLB confidence limits for parameters
- ▶ CLI confidence limits for an individual predicted value
- ► CLM confidence limits for an average/expected value of dependent variable

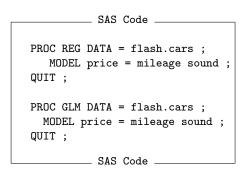
 (□)
 (□)
 (□)
 (□)
 (□)
 (□)
 (□)
 (□)
 (□)
 (□)
 (□)
 (□)
 (□)
 (□)
 (□)
 (□)
 (□)
 (□)
 (□)
 (□)
 (□)
 (□)
 (□)
 (□)
 (□)
 (□)
 (□)
 (□)
 (□)
 (□)
 (□)
 (□)
 (□)
 (□)
 (□)
 (□)
 (□)
 (□)
 (□)
 (□)
 (□)
 (□)
 (□)
 (□)
 (□)
 (□)
 (□)
 (□)
 (□)
 (□)
 (□)
 (□)
 (□)
 (□)
 (□)
 (□)
 (□)
 (□)
 (□)
 (□)
 (□)
 (□)
 (□)
 (□)
 (□)
 (□)
 (□)
 (□)
 (□)
 (□)
 (□)
 (□)
 (□)
 (□)
 (□)
 (□)
 (□)
 (□)
 (□)
 (□)
 (□)
 (□)
 (□)
 (□)
 (□)
 (□)
 (□)
 (□)
 (□)
 (□)
 (□)
 (□)
 (□)
 (□)
 (□)
 (□)
 (□)
 (□)
 (□)
 (□)
 (□)
 (□)
 (□)
 (□)
 (□)
 (□)
 (□)
 (□)
 (□)
 (□)
 (□)
 (□)
 (□)
 (□)
 (□)
 (□)
 (□)
 (□)
 (□)
 (□)</

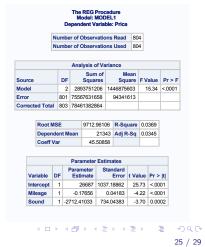
 Overview
 ANOVA
 Regression

 0000000
 00000000
 00000000

Relationship between price and mileage, adjusting for sound

This model can be executed in either PROC REG or PROC GLM because sound is coded as 0/1 (an indicator variable).





 Overview
 ANOVA
 Regression

 000000
 00000000
 000000●000

Relationship between price and mileage, adjusting for cruise

This model can only be executed in PROC GLM because cruise is coded as Y/N.

```
PROC GLM DATA = flash.cars ;

CLASS cruise ;

MODEL price = mileage cruise / SOLUTION ;

QUIT ;

SAS Code
```

- ▶ Use the CLASS statement for the categorical variable
- ▶ Use SOLUTION option to obtain parameter estimates
- ► You could do this with PROC REG if you coded cruise as 0/1 in the data set



 Overview
 ANOVA
 Regression

 000000
 00000000
 0000000●00

Exploring a quadratic relationship with mileage

This model can be easily executed in PROC GLM.

```
PROC GLM DATA = flash.cars;

MODEL price = mileage mileage*mileage;

QUIT;

SAS Code
```

- You could do this with PROC REG if you coded mileage_squared in the data set
- ► The same idea applies to interaction terms (PROC GLM can handle them in the model statement, PROC REG needs the variables to be coded in the data set)

 Overview
 ANOVA
 Regression

 0000000
 000000000
 000000000

PROC REG model selection

STAT 330: Lecture 13

STAT 330: Lecture 13

Another *option* for the MODEL statement in PROC REG allows you to do automated model selection. There are 9 model selection methods available.

```
SAS Code

PROC REG DATA = flash.cars ;

MODEL price = mileage liter sound leather /

SELECTION = RSQUARE ;

QUIT ;

SAS Code
```

This example uses the R-squared method to examine all possible models based on the 4 independent variables.

□ ► < □ ► < □ ► < □ ► < □ ► < □ ► < □ ► < □ ►
 28 / 29

 Overview
 ANOVA
 Regression

 000000
 00000000
 00000000●

Conditions for regression

- 1. observations are independent
- 2. linear relationship between x and y
- 3. normally distributed errors about the regression line
- 4. constant variability in *y* about the regression line (constant variance)

On your own: How would you check these conditions?

