PROC ANOVA, PROC REG, and PROC GLM

Shannon Pileggi

STAT 330

OUTLINE

Overview

ANOVA

Regression

Methods overview

ANOVA (analysis of variance)

- Dependent variable = quantitative

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

  $H_0 : \mu_1 = \mu_2 = \mu_3 = \ldots \mu_g$ 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: $\beta_1 \neq 0$

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
- repeated measures analysis of variance

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...

Some quirks

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

The Data

Collected from Kelly Blue Book for 2005 used GM cars

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

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

Review

On your own: Match the appropriate statistical method for each research question.

1. one-sample t-test
2. two sample t-test
3. paired t-test
4. correlation
5. simple linear regression
6. multipe linear regression
7. ANOVA

____ Does the population average of price differ by number of doors (2,4) the car comes with?

____ Does the population average of price differ by number of cylinders (4,6,8) the car has?

____ Is there a relationship between price and mileage?

____ Is there a relationship between price and mileage after adjusting for number of doors?

Overview

ANOVA

Regression

Review

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

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

Syntax

SAS Code

```
PROC ANOVA DATA = mydata ;
    CLASS catvar;
    MODEL quantvar = 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

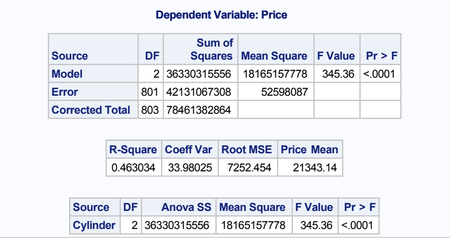
PROC ANOVA Example

SAS Code

```
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?

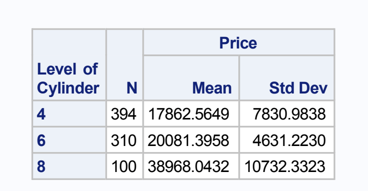


Dependent Variable: Price

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	36330315556	18165157778	345.36	<.0001
Error	801	42131067308	52598087		
Corrected Total	803	78461382864			

R-Square	Coeff Var	Root MSE	Price Mean
0.463034	33.98025	7252.454	21343.14

Source	DF	Anova SS	Mean Square	F Value	Pr > F
Cylinder	2	36330315556	18165157778	345.36	<.0001



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

Conditions for one-way ANOVA

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

On your own: How would you check these conditions?

MEANS options

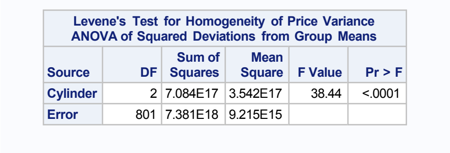
- Many multiple comparison methods available: `TUKEY`, `SCHEFFE`, `DUNCAN`, `BON` (for Bonferroni).
- `ALPHA=` controls overall error rate
- To test the assumption of equal variance, use the `HOVTEST` option (**<u>H</u>**omogeneity **<u>O</u>**f **<u>V</u>**ariance test). For this test, the null hypothesis is H_0: Variances are equal. Smaller p-values lend stronger evidence against this statement.

PROC ANOVA Example, continued

SAS Code

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

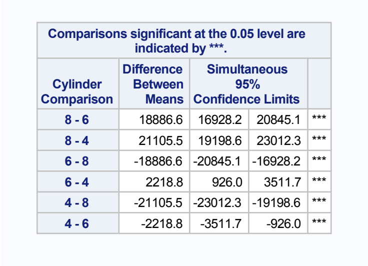
SAS Code



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

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Cylinder	2	7.084E17	3.542E17	38.44	<.0001
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?



Comparisons significant at the 0.05 level are indicated by ***.

Cylinder Comparison	Difference Between Means	Simultaneous 95% Confidence Limits		
8 - 6	18886.6	16928.2	20845.1	***
8 - 4	21105.5	19198.6	23012.3	***
6 - 8	-18886.6	-20845.1	-16928.2	***
6 - 4	2218.8	926.0	3511.7	***
4 - 8	-21105.5	-23012.3	-19198.6	***
4 - 6	-2218.8	-3511.7	-926.0	***

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

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

Overview

ANOVA

Regression

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

Relationship between price and mileage

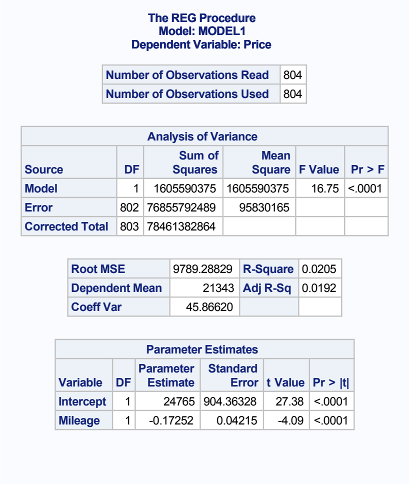
Both PROC REG and PROC GLM can be used for simple linear regression with a quantitative independent variable.

SAS Code

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

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

SAS Code



The REG Procedure
Model: MODEL1
Dependent Variable: Price

Number of Observations Read	804
Number of Observations Used	804

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	1605590375	1605590375	16.75	<.0001
Error	802	76855792489	95830165		
Corrected Total	803	78461382864			

Root MSE	9789.28829	R-Square	0.0205
Dependent Mean	21343	Adj R-Sq	0.0192
Coeff Var	45.86620		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	24765	904.36328	27.38	<.0001
Mileage	1	-0.17252	0.04215	-4.09	<.0001

Other features of PROC REG - multiple model statements

SAS Code

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

SAS Code

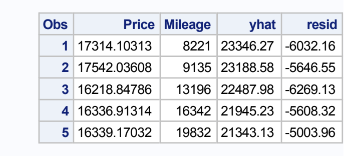
Other features of PROC REG - create data set with residuals and other diagnostic measures

SAS Code

```
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

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

Relationship between price and mileage, adjusting for sound

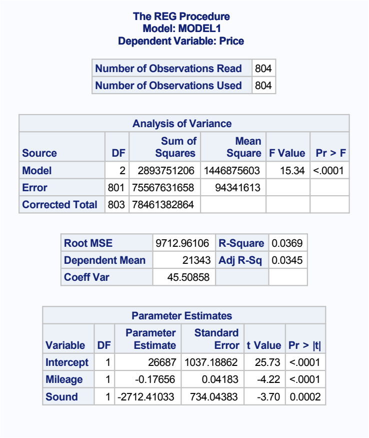
This model can be executed in either PROC REG or PROC GLM because sound is coded as 0/1 (an indicator variable).

SAS Code

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

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

SAS Code



The REG Procedure
Model: MODEL1
Dependent Variable: Price

Number of Observations Read	804
Number of Observations Used	804

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	2893751206	1446875603	15.34	<.0001
Error	801	75567631658	94341613		
Corrected Total	803	78461382864			

Root MSE	9712.96106	R-Square	0.0369
Dependent Mean	21343	Adj R-Sq	0.0345
Coeff Var	45.50858		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	26687	1037.18862	25.73	<.0001
Mileage	1	-0.17656	0.04183	-4.22	<.0001
Sound	1	-2712.41033	734.04383	-3.70	0.0002

Relationship between price and mileage, adjusting for cruise

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

SAS Code

```
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

Exploring a quadratic relationship with mileage

This model can be easily executed in PROC GLM.

SAS Code

```
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)

PROC REG model selection

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.

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?