Multiple Linear Regression
ANCOVA and PC Regression

Laura Barnes & Julianne Quinn

University of Virginia
Charlottesville, VA

Agenda

Model Diagnostics

- Graphical Diagnostics:
  - Residuals vs. fitted
  - Scale-Location plot of square root of absolute standardized residuals vs. fitted
  - QQ plot of standardized residuals
  - Residual-Leverage plot
- Analytical Diagnostics:
  - We can also diagnose problems in the regression by analyzing the results: coefficient values, performance tests.
  - Multicollinearity is apparent in the flipped signs for coefficients and for large changes in coefficient values with small changes to values in the observations.
  - Diagnosing Simpson's Paradox with coefficient values: coefficients show contradictory patterns; adding variables or replacing variables causes major changes to coefficient values.

When the Tests Fail

- Transform the response: Box-Cox plot
- Transform the predictors: second order model, interaction model, complete second order model
- Add new variables (i.e., new models) for Simpson's paradox and lack of fit;
- Remove or combine variables for multicollinearity;
- Principal components regression or ridge regression for multicollinearity; and
- Use robust methods for regression.

Qualitative Regression Models

- Qualitative predictors require coding to use in regression models.
- The different values of the qualitative variable are called levels.
- Treatment contrasts or dummy variables:
  - If the qualitative variable has m levels then it can be encoded with $m - 1$ dummy variables.
  - Each dummy variable represents a level of the qualitative variable.
  - Each dummy variable takes on one of two values: 0 or 1

Example Qualitative Regression Model

- Consider MPG of different cars. Suppose we have cars in three brands: Toyota, Ford and Chevrolet.
- Code the brand variable with 2 dummy variables as follows:

$$X_1 = \begin{cases} 1 & \text{if Ford} \\ 0 & \text{else} \end{cases} \quad X_2 = \begin{cases} 1 & \text{if Toyota} \\ 0 & \text{else} \end{cases}$$

- Another categorical variable *Cylinder*: Cylinder has two levels, more than 5 cylinders or not.

$$X_3 = \begin{cases} 1 & \text{if more than 5 cylinders} \\ 0 & \text{else} \end{cases}$$

Tests of Understanding

- Write a linear, main effects model with MPG (Y) as a function of brand and cylinder.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon$$

- Write a main effects plus interaction model with MPG (Y) as a function of brand and cylinder.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_1 x_3 + \beta_5 x_2 x_3 + \epsilon$$

Regression Results with Train Data

- Model: ACCDMG $\sim$ Cause (5 levels: E, H, M, S, T)
- Results:

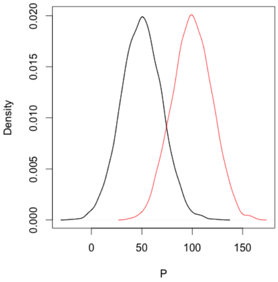
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	169976	24368	6.975	3.82e-12 ***
CauseH	-100517	27384	-3.671	0.000247 ***
CauseM	-44111	29130	-1.514	0.130073
CauseS	-130979	60260	-2.174	0.029825 *
CauseT	11273	28265	0.399	0.690050

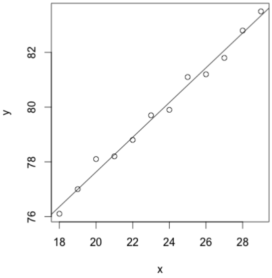
- What is the base case?
- Interpret the above result, what's your conclusion?

Analysis of Variance

- ANOVA provides a method for multiple comparisons of means.
- A one-way ANOVA considers one predictor variable at multiple levels.
- ANOVA treats all predictors as qualitative variables or factors. So it converts quantitative variables to qualitative variables.
- This means it does not require the linear independence assumption but it does reduce the interpretability of the results. Both regression and ANOVA produce identical results for qualitative variables.
- Since regression gives us more information we will use it with both observational and experimental data.

ANOVA vs. Regression



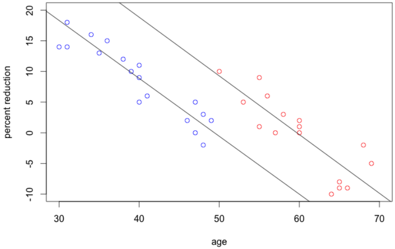


Analysis of Covariance

- Recall that regression allows for association tests while controlling for the values of other variables in the equation.
- ANCOVA combines qualitative and quantitative predictors or explanatory variables.
- Why would we want to use ANCOVA? A fake drug testing example.

ANCOVA Example

- Patients who received an inflammation reducing medication are shown by the red dots. Did it work? Simpson's paradox strikes again!



Quantitative Variable Added

- Recall the train accident problem. We built a model: ACCDMG $\sim$ CAUSE (five levels: E,H,M, S,T)
- Train speed (TRNSPD) may predict damages and we want to add this quantitative variable to the model. This is the essence of ANCOVA.
- Write the main effects models: ACCDMG $\sim$ CAUSE and ACCDMG $\sim$ CAUSE+TRNSPD and the interaction model.
- How can we compare them?
  - In general we test interactions using the Partial F test.
  - Regardless of the encoding we test a qualitative variable using the Partial F test.

Quantitative Variable Added

- Look at the main effects and the interaction results that follow. Use Partial F tests to make a recommendation.
- Look at the t-tests. Do we ever remove a variable in an interaction term from the main effects part of the model?

Main Effects Results ANCOVA

- Model.cause: $log(ACCDMG + 1) \sim CAUSE$
- Result:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	10.847752	0.073335	147.921	< 2e-16 ***
CauseH	-0.442112	0.082411	-5.365	8.79e-08 ***
CauseM	-0.387382	0.087666	-4.419	1.03e-05 ***
CauseS	-0.620046	0.181352	-3.419	0.000638 ***
CauseT	0.004585	0.085064	0.054	0.957014

Residual standard error: 1.196 on 2715 degrees of freedom
Multiple R-squared:0.03051, Adjusted R-squared: 0.02908
F-statistic: 21.36 on 4 and 2715 DF, p-value: < 2.2e-16

Main Effects Results ANCOVA

- Model.cause+trnspd:
  $log(ACCDMG + 1) \sim CAUSE + TRNSPD$
- Result:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	10.562932	0.075696	139.544	< 2e-16 ***
CauseH	-0.246274	0.082213	-2.996	0.00276 **
CauseM	-0.418358	0.085631	-4.886	1.09e-06 ***
CauseS	-0.409950	0.177980	-2.303	0.02133 *
CauseT	0.088169	0.083361	1.058	0.29029
TRNSPD	0.016968	0.001464	11.592	< 2e-16 ***

Residual standard error: 1.168 on 2714 degrees of freedom
Multiple R-squared: 0.07624, Adjusted R-squared: 0.07454
F-statistic: 44.8 on 5 and 2714 DF, p-value: < 2.2e-16

Interaction Results ANCOVA

- Model.interaction: $log(ACCDMG + 1) \sim CAUSE + TRNSPD + CAUSE : TRNSPD$
- Result:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	10.531683	0.096002	109.703	< 2e-16 ***
CauseH	-0.308234	0.107096	-2.878	0.00403 **
CauseM	-0.180733	0.113306	-1.595	0.11081
CauseS	-0.470313	0.247140	-1.903	0.05714 .
CauseT	-0.040416	0.109184	-0.370	0.71129
TRNSPD	0.018830	0.003862	4.875	1.15e-06 ***
CauseH:TRNSPD	0.015913	0.006995	2.275	0.02299 *
CauseM:TRNSPD	-0.012950	0.004378	-2.958	0.00312 **
CauseS:TRNSPD	0.018941	0.036969	0.512	0.60844
CauseT:TRNSPD	0.011615	0.004673	2.486	0.01299 *

Residual standard error: 1.155 on 2710 degrees of freedom Multiple R-squared: 0.0979, Adjusted R-squared: 0.0949 F-statistic: 32.68 on 9 and 2710 DF, p-value: < 2.2e-16

Partial F Tests ANCOVA

- Model.cause vs. Model.cause+trnspd:
  Analysis of Variance Table
  Model 1: log(ACCDMG + 1)$\sim$ Cause
  Model 2: log(ACCDMG + 1) $\sim$Cause + TRNSPD

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	2715	3883.9				
2	2714	3700.7	1	183.23	134.37	< 2.2e-16 ***

Partial F Tests ANCOVA

- Model.cause+trnspd vs. Model.interaction:
  Analysis of Variance Table
  Model 1: log(ACCDMG + 1)$\sim$ Cause + TRNSPD
  Model 2: log(ACCDMG + 1) $\sim$Cause + TRNSPD+Cause:TRNSPD

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	2714	3700.7				
2	2710	3613.9	4	86.751	16.263	3.6e-13 ***

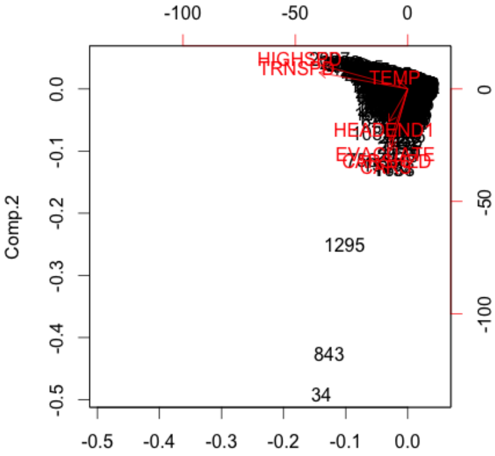
Principal Components Regression

- We talked about how to select variables.
- Remove or combine variables for multicollinearity.
- Rather than select variables for the model, principal components puts weights on the variables. How are those weights chosen?
- By definition the principal components are orthogonal. Hence, the combinations of principal components eliminates multicollinearity problems.
- The loadings on the principal components can reveal latent variables or "factors". These are higher order variables that represent possible contributors to the response.

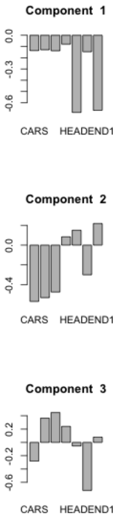
PC Regression Steps

1. Find the principal components for the quantitative variables in the data set after removing the response variable.
2. Determine how many principal components explain a sufficient amount of the variance (e.g., 90%). Select these components as the predictors.
3. Compute the principal components scores or values for each observation. This is a matrix product of the principal component with each observation matrix. Call this the principal component data matrix.
4. Regress the response variable against the principal component data matrix.

PC Regression Example: Rail Accident PC Biplot



PC Regression Example: Rail Accident PC Loadings



PC Regression Example: Rail Accident Results

	# of PC	R^2	BIC
50% Variance	2	0.8748	77674.11
75% Variance	3	0.09097	77671.59
90% Variance	5	0.09138	77686.19

- What else methods can we use to compare the above three models?
- What about comparing PC regression models with the main effect model?

Multiple Linear Regression

- Concept of multiple linear regression
- Least square estimate
- Model assumptions
- Variable selections
- Model diagnostics
- Nonlinear variables
- Qualitative variables
- ANOVA and ANCOVA

How to Build Multiple Linear Regression Models?

- Building multiple linear regression model is not as simple as typing $lm(r \sim ., data = acts11)$ in R!
- Begin with graphical analysis;
- Select variables (both quantitative and qualitative variables) for multiple linear regression models;
- Measure the performance of models: F tests, R^2, AIC, BIC, etc.;
- Diagnose models: graphical and analytical;
- Adjust models: transformations, higher order models, variable selection, PC regression, etc.;
- Repeat the above steps as necessary;
- Get several alternative models, select the best model(s) for recommendation.

Model Selections

- Model selection means choosing the model or models to use as the basis for our analysis and ultimately our recommendations.
- Model selection consists of choosing variables, transformations, and combinations among the variables and levels in qualitative variables.
- Model selection is important, but why?

Reason for Model Selections

- Okham's (Occam's) Razor.
- Extra terms can add noise to the predictions. More data is not necessarily better.
- Multicollinearity.
- Leaving out variables causes inaccurate understanding and predictions. This is Simpson's paradox.
- It can cost more to get data on more variables.
- We have to make recommendations. If models give competing answers, we need to pick from among these.

Approaches to Model Selection

- t-tests. This is not a good approach because of multiplicity.
- Partial F test results. This is a good approach, but not for non-nested models.
- Criterion based also good but with limits.
- Automated selection, forward, backward, and stepwise. Quick and dirty.
- Principal components can provides variable extraction versus selection.

Approaches to Model Selection

- Test sets when we have enough data.
- Cross-validation when we don't have enough data.
- Choose models based on diagnostics.
- Bootstrapping when nonparametric methods are needed.
- Model selection is hard! Focus on the problem (not the mechanics). This is why good systems engineers are in such demand.