Lecture 20: Qualitative variables as predictors and Interactions II

Pratheepa Jeganathan

11/06/2019

Recap

- What is a regression model?
- Descriptive statistics – graphical
- Descriptive statistics – numerical
- Inference about a population mean
- Difference between two population means
- Some tips on R
- Simple linear regression (covariance, correlation, estimation, geometry of least squares)
  - Inference on simple linear regression model
  - Goodness of fit of regression: analysis of variance.
  - F-statistics.
  - Residuals.
  - Diagnostic plots for simple linear regression (graphical methods).

Recap

- Multiple linear regression
  - Specifying the model.
  - Fitting the model: least squares.
  - Interpretation of the coefficients.
  - Matrix formulation of multiple linear regression
  - Inference for multiple linear regression
    - T-statistics revisited.
    - More F statistics.
    - Tests involving more than one β.
- Diagnostics – more on graphical methods and numerical methods
  - Different types of residuals
  - Influence
  - Outlier detection
  - Multiple comparison (Bonferroni correction)
  - Residual plots:
    - partial regression (added variable) plot,
    - partial residual (residual plus component) plot.

Recap

- Adding qualitative predictors
  - Qualitative variables as predictors to the regression model.
  - Adding interactions to the linear regression model.

Qualitative variables and Interactions

Outline

- Analyzing and testing for equality of regression relationship in various subsets of a population.

Note

- Additive model - no interaction
- Multiplicative model - with interaction
- Start with a simple model and proceed sequentially to more complex model (try to retain the simplest model that has an acceptable residual structure)
- There are situation that we need to fit regression sepeartly for subsets.
  - analyzing and testing for equality of regression relationship in various subsets of a population.

Jobtest employment data (**CH** Page 138)

- We look at an example of a dataset concerning equal opportunity in employment.
  - Suppose there is an aptitude test to screen job applicants.
  - The test measures the applicant's aptitude for the job and shouldn't discriminate by race.
  - We considered a variable that indicates the race, either "White" or "Minority."
  - Let's use the dataset to analyze the implication of the hypothesis for discrimination in hiring.

Variable	Description
TEST	Job aptitude test score
MINORITY	1 if applicant could be considered minority, 0 otherwise
PERF	Job performance evaluation

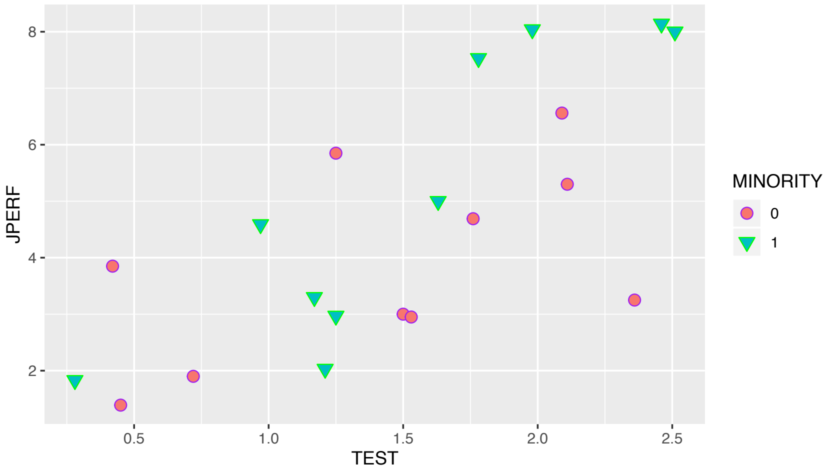
Jobtest employment data

```r
url = 'http://stats191.stanford.edu/data/jobtest.table'
jobtest.table = read.table(url, header=T)
jobtest.table$MINORITY = factor(jobtest.table$MINORITY)
```

Jobtest employment data

```r
p = ggplot(data = jobtest.table, aes(x = TEST, y = JPERF,
  shape = MINORITY, col = MINORITY, fill = MINORITY)) +
  geom_point(size = 3) +
  scale_shape_manual(values = c(21,25))+
  scale_color_manual(values = c("purple",
    "green")) +
  xlab("TEST") +
  ylab("JPERF")
```

Jobtest employment data



General model

- In theory, there may be a linear relationship between *JPERF* and *TEST* but it could be different by group.
- Model:

$$JPERF_i = \beta_0 + \beta_1 TEST_i + \beta_2 MINORITY_i + \beta_3 MINORITY_i * TEST_i + \varepsilon_i$$

- Regression functions:

$$Y_i = \begin{cases} \beta_0 + \beta_1 TEST_i + \varepsilon_i & \text{if } MINORITY_i = 0 \\ (\beta_0 + \beta_2) + (\beta_1 + \beta_3) TEST_i + \varepsilon_i & \text{if } MINORITY_i = 1. \end{cases}$$

Our first model: $(\beta_2 = \beta_3 = 0)$

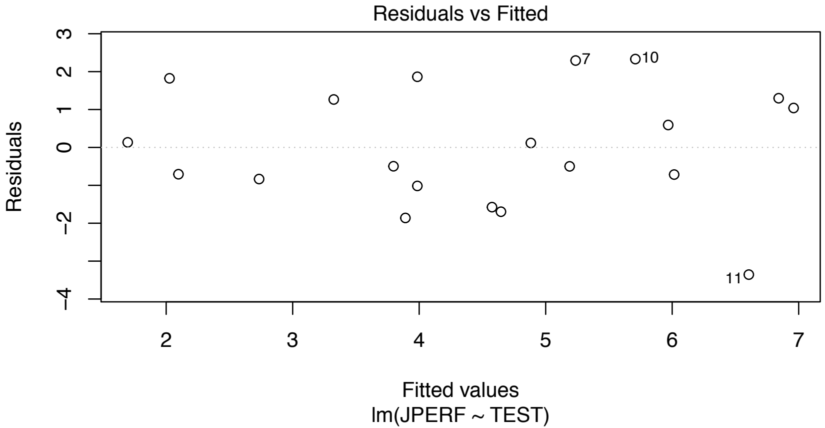
- This has no effect for `MINORITY`.

```
jobtest.lm1 = lm(JPERF ~ TEST, jobtest.table)
#summary(jobtest.lm1)
```

```
Call:
lm(formula = JPERF ~ TEST, data = jobtest.table)

Residuals:
    Min      1Q  Median      3Q     Max
-3.3558 -0.8798 -0.1897  1.2735  2.3312

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.0350     0.8680   1.192 0.248617
TEST          2.3605     0.5381   4.387 0.000356 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.591 on 18 degrees of freedom
Multiple R-squared:  0.5167,    Adjusted R-squared:  0.4899
F-statistic: 19.25 on 1 and 18 DF,  p-value: 0.0003555
```

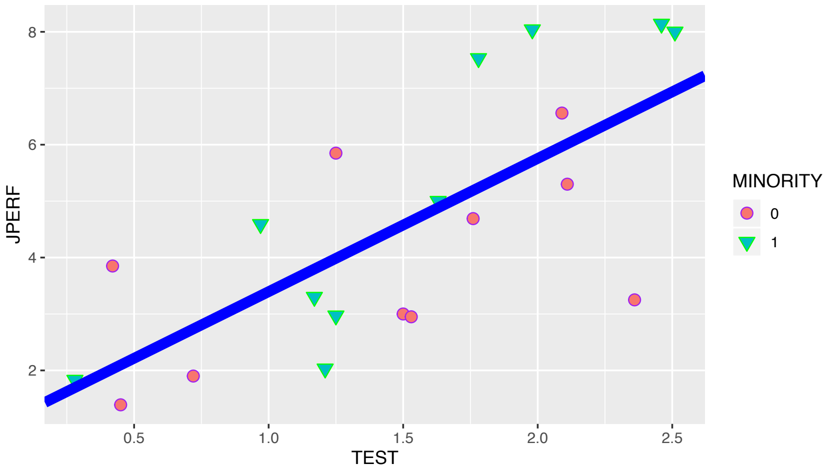
The first model: $(\beta_2 = \beta_3 = 0)$

```
plot(jobtest.lm1, add.smooth = FALSE, which = 1)
```



The first model: $(\beta_2 = \beta_3 = 0)$

```
p + geom_abline(intercept = jobtest.lm1$coef[1],
  slope = jobtest.lm1$coef[2],
  lwd=3, col='blue')
```



Our second model ($\beta_3 = 0$)

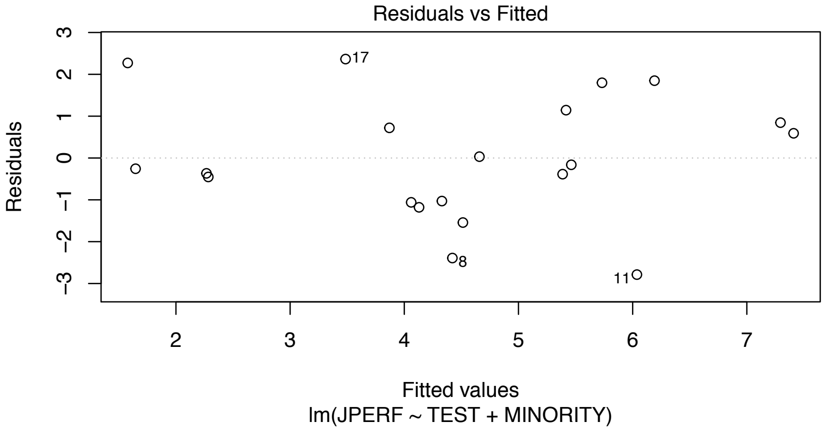
- This model allows for an effect of MINORITY but no interaction between MINORITY and TEST.

```r
jobtest.lm2 = lm(JPERF ~ TEST + MINORITY,
  data = jobtest.table)
#summary(jobtest.lm2)
```

```
Call:
lm(formula = JPERF ~ TEST + MINORITY, data = jobtest.table)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.7872 -1.0370 -0.2095  0.9198  2.3645 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   0.6120     0.8870   0.690 0.499578    
TEST          2.2988     0.5225   4.400 0.000391 ***
MINORITY1     1.0276     0.6909   1.487 0.155246    
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.54 on 17 degrees of freedom
Multiple R-squared:  0.5724,	Adjusted R-squared:  0.5221 
F-statistic: 11.38 on 2 and 17 DF,  p-value: 0.0007312
```

The second model ($\beta_3 = 0$)

```
plot(jobtest.lm2, add.smooth = FALSE, which = 1)
```

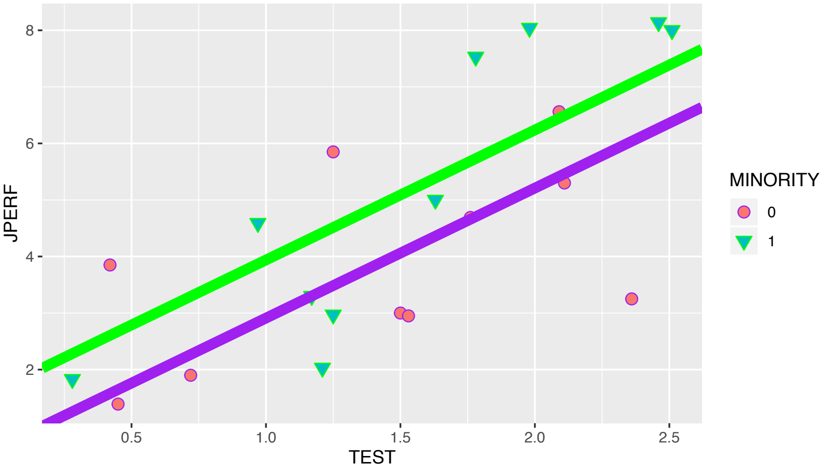


The second model ($\beta_3 = 0$)

```
p2 = p + geom_abline(intercept =
    jobtest.lm2$coef['(Intercept)'],
  slope = jobtest.lm2$coef['TEST'],
  lwd=3, col='purple') +
  geom_abline(intercept =
      (jobtest.lm2$coef['(Intercept)'] +
      jobtest.lm2$coef['MINORITY1']),
  slope = jobtest.lm2$coef['TEST'],
  lwd=3, col='green')
```

The second model ($\beta_3 = 0$)

```
p2
```



Our third model ($\beta_2 = 0$)

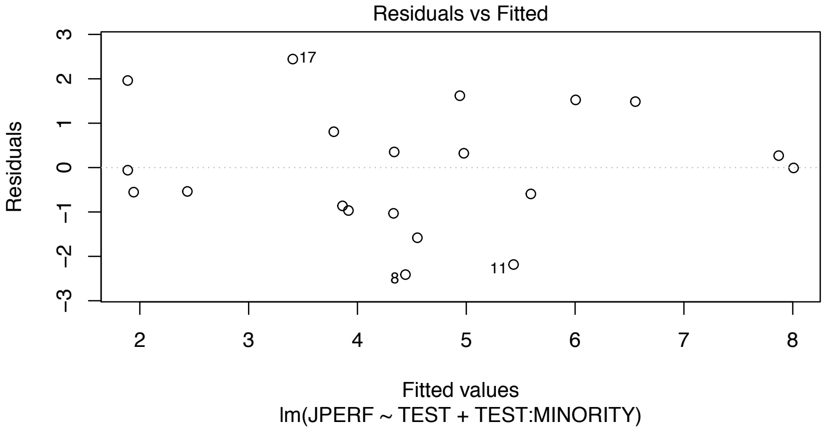
- This model includes an interaction between `TEST` and `MINORITY`.
- These lines have the same intercept but possibly different slopes within the `MINORITY` groups.

```
jobtest.lm3 = lm(JPERF ~ TEST + TEST:MINORITY,
  data = jobtest.table)
#summary(jobtest.lm3)
```

```
Call:
lm(formula = JPERF ~ TEST + TEST:MINORITY, data = jobtest.table)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.41100 -0.88871 -0.03359  0.97720  2.44440 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)   
(Intercept)     1.1211     0.7804   1.437  0.16900   
TEST            1.8276     0.5356   3.412  0.00332 **
TEST:MINORITY1  0.9161     0.3972   2.306  0.03395 * 
---
Signif. codes:  
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.429 on 17 degrees of freedom
Multiple R-squared:  0.6319,	Adjusted R-squared:  0.5886 
F-statistic: 14.59 on 2 and 17 DF,  p-value: 0.0002045
```

The third model ($\beta_2 = 0$)

```
plot(jobtest.lm3, add.smooth = FALSE, which = 1)
```

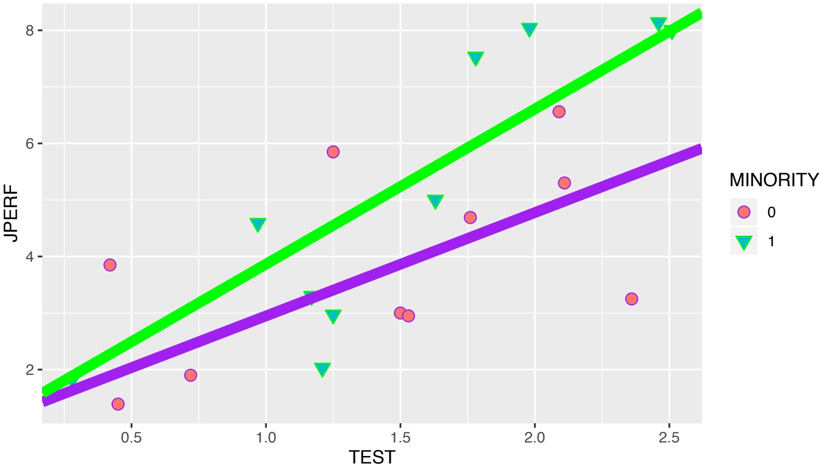


The third model ($\beta_2 = 0$)

```
p3 = p + geom_abline(intercept =
    jobtest.lm3$coef['(Intercept)'],
  slope = jobtest.lm3$coef['TEST'],
  lwd=3, col='purple') +
  geom_abline(intercept =
      jobtest.lm3$coef['(Intercept)'],
  slope =
      (jobtest.lm3$coef['TEST'] +
          jobtest.lm3$coef['TEST:MINORITY1']),
  lwd=3, col='green')
```

The third model ($\beta_2 = 0$)

```
p3
```



Our final model: no constraints

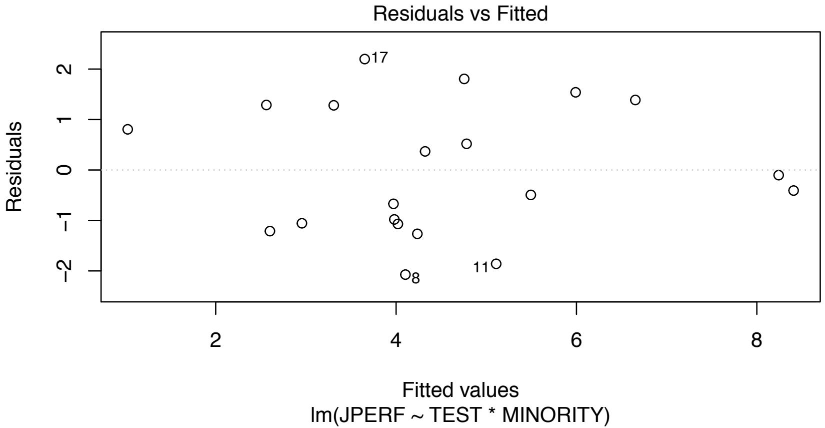
- This model allows for different intercepts and different slopes.
- The expression TEST*MINORITY is shorthand for TEST + MINORITY + TEST:MINORITY.

```
jobtest.lm4 = lm(JPERF ~ TEST * MINORITY,
  data = jobtest.table)
#summary(jobtest.lm4)
```

```
Call:
lm(formula = JPERF ~ TEST * MINORITY, data = jobtest.table)

Residuals:
    Min      1Q  Median      3Q     Max
-2.0734 -1.0594 -0.2548  1.2830  2.1980

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)      2.0103     1.0501   1.914   0.0736 .
TEST             1.3134     0.6704   1.959   0.0677 .
MINORITY1       -1.9132     1.5403  -1.242   0.2321
TEST:MINORITY1   1.9975     0.9544   2.093   0.0527 .
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.407 on 16 degrees of freedom
Multiple R-squared:  0.6643,    Adjusted R-squared:  0.6013
F-statistic: 10.55 on 3 and 16 DF,  p-value: 0.0004511
```

The final model

```
plot(jobtest.lm4, add.smooth = FALSE, which = 1)
```

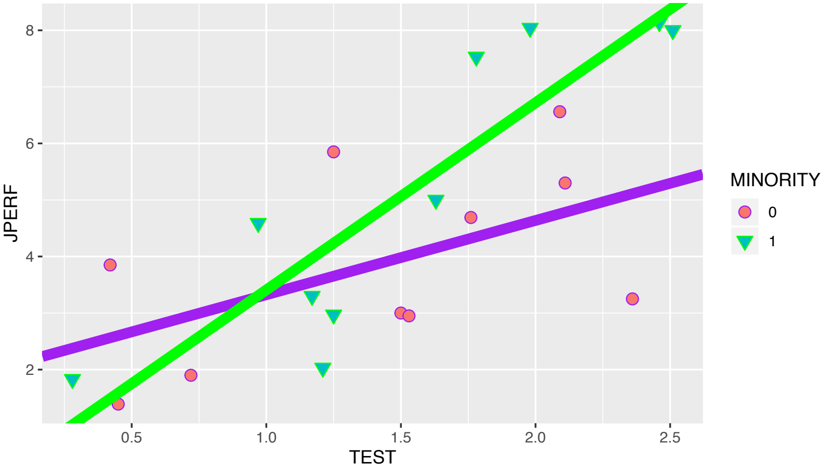


The final model

```
p4 = p + geom_abline(intercept =
    jobtest.lm4$coef['(Intercept)'],
  slope = jobtest.lm4$coef['TEST'],
  lwd=3, col='purple') +
  geom_abline(intercept =
      (jobtest.lm4$coef['(Intercept)'] +
         jobtest.lm4$coef['MINORITY1']),
  slope =
      (jobtest.lm4$coef['TEST'] +
         jobtest.lm4$coef['TEST:MINORITY1']),
  lwd=3, col='green')
```

The final model

```
p4
```



Comparing models

- We can use F test statistic.
- Is there any effect of MINORITY on slope or intercept?

```
# ~ TEST vs. ~ TEST * MINORITY
anova(jobtest.lm1, jobtest.lm4)
```

```
## Analysis of Variance Table
##
## Model 1: JPERF ~ TEST
## Model 2: JPERF ~ TEST * MINORITY
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1     18 45.568
## 2     16 31.655  2    13.913 3.5161 0.05424 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
```

Comparing models

- Is there any effect of MINORITY on intercept? (Assuming we have accepted the hypothesis that the slope is the same within each group).

```
# ~ TEST vs. ~ TEST + MINORITY
anova(jobtest.lm1, jobtest.lm2)
```

```
## Analysis of Variance Table
##
## Model 1: JPERF ~ TEST
## Model 2: JPERF ~ TEST + MINORITY
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     18 45.568
## 2     17 40.322  1    5.2468 2.2121 0.1552
```

Comparing models

- We could also have allowed for the possiblity that the slope is different within each group and still check for a different intercept.

```
# ~ TEST + TEST:MINORITY vs.
# ~ TEST * MINORITY
anova(jobtest.lm3, jobtest.lm4)
```

```
## Analysis of Variance Table
##
## Model 1: JPERF ~ TEST + TEST:MINORITY
## Model 2: JPERF ~ TEST * MINORITY
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     17 34.708
## 2     16 31.655  1    3.0522 1.5427 0.2321
```

Comparing models

- Is there any effect of MINORITY on slope? (Assuming we have accepted the hypothesis that the intercept is the same within each group).

```
# ~ TEST vs. ~ TEST + TEST:MINORITY
anova(jobtest.lm1, jobtest.lm3)
```

```
## Analysis of Variance Table
##
## Model 1: JPERF ~ TEST
## Model 2: JPERF ~ TEST + TEST:MINORITY
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1     18 45.568
## 2     17 34.708  1    10.861 5.3196 0.03395 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
```

Comparing models

- Again, we could have allowed for the possibility that the intercept is different within each group.

```
# ~ TEST + MINORITY vs.
# # ~ TEST * MINORITY
anova(jobtest.lm2, jobtest.lm4)
```

```
## Analysis of Variance Table
##
## Model 1: JPERF ~ TEST + MINORITY
## Model 2: JPERF ~ TEST * MINORITY
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1     17 40.322
## 2     16 31.655  1    8.6661 4.3802 0.05265 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
```

Comparing models

- In summary, taking the several tests into account here, there does seem to be some evidence that the slope is different within the two groups.

Model selection

- Already with this simple dataset (simpler than the IT salary data) we have 4 competing models.
- How are we going to arrive at a final model?
- This highlights the need for *model selection*. (**CH** Chapter 11)

Reference

- **CH**: Chapter 5.4-5.7.
- Lecture notes of Jonathan Taylor .