Statistical Modeling

CH.4 - Qualitative Variables

SS 2022 || Prof. Dr. Buchwitz

Outline

Course Contents

Session	Topic
1	Simple Linear Regression
2	Multiple Linear Regression
3	Regression Diagnostics
4	Qualitative Variables as Predictors
5	Transformation of Variables
6	Weighted Least Squares
7	Correlated Errors
8	Analysis of Collinear Data
9	Working with Collinear Data
10	Variable Selection Procedures
11	Logistic Regression
12	Further Topics

Outline

Introduction

- Qualitative or categorical variables (such as gender, marital status, etc.) are useful predictors and are usually called **indicator** or **dummy variables**.
- Those variables usually only take two values, 0 and 1, which signify that the observation belongs to one of two possible categories.
- The numerical values of indicator variables **do not reflect quantitative ordering**.
- **Example Variable:** Gender, coded as 1 for *female* and 0 for *male*.
- Indicator variables can also be used in a regression equation to distinguish between three or more groups.
- The response variable is stil a quantiative continuous in all discussed cases.

Example: Salary Survey Data

```
##        S  X E M
## 1  13876  1 1 1
## 2  11608  1 3 0
## 3  18701  1 3 1
## 4  11283  1 2 0
## 5  11767  1 3 0
## 6  20872  2 2 1
## 7  11772  2 2 0
## 8  10535  2 1 0
## 9  12195  2 3 0
## 10 12313  3 2 0
## 11 14975  3 1 1
## 12 21371  3 2 1
## 13 19800  3 3 1
## 14 11417  4 1 0
## 15 20263  4 3 1
## 16 13231  4 3 0
## 17 12884  4 2 0
## 18 13245  5 2 0
## 19 13677  5 3 0
## 20 15965  5 1 1
## 21 12336  6 1 0
## 22 21352  6 3 1
## 23 13839  6 2 0
## 24 22884  6 2 1
## 25 16978  7 1 1
## 26 14803  8 2 0
## 27 17404  8 1 1
## 28 22184  8 3 1
## 29 13548  8 1 0
## 30 14467 10 1 0
## 31 15942 10 2 0
```

Your turn

Salary survey of computer professionals with objective to identify and quantify variables that determine salary differentials.

S Salary (Response)

X Experience, measured in years

E Education, 1 (High School/HS), 2 (Bachelor/BS), 3 (Advanced Degree/AD)

M Management 1 (is Manager), 0 (no Management Responsibility)

Example: Salary Survey Data

- **Experience:** We assume linearity, which means that each additional year is worth a fixed salary increment.
- **Education:** Can be used in a linear or categorial form.
  - ▶ Using the variable in its raw form would assume that each step up in education is worth a fixed increment in salary. This may be too restrictive.
  - ▶ Using education as categorical variable can be done by defining **two indicator variables**. This allows to pick up the effect of education wether it is linear or not.
- **Management:** Is also an indicator variable, that allows to distinguish between management (1) an regular staff positions (0).

Indicator Variables

When using indicator variables to represent a set of categroies, the number of these variables required is **one less than the number of categories**. For *education* we can create two indicators variables:

$$E_{i1} = \begin{cases} 1, \text{ if the i-th person is in the HS category} \\ 0, \text{otherwise.} \end{cases}$$

$$E_{i2} = \begin{cases} 1, & \text{if the i-th person is in the BS category} \\ 0, & \text{otherwise.} \end{cases}$$

These two variables allow representing the three groups (HS, BS, AD).

HS: $E_1 = 1, E_2 = 0$, BS: $E_1 = 0, E_2 = 1$, AD: $E_1 = 0, E_2 = 0$

■ The regression equation from the Salary Survey Data is:

$$S = \beta_0 + \beta_1 X + \gamma_1 E_1 + \gamma_2 E_2 + \delta_1 M + \epsilon$$

Indicator Variables

- The regression equation from the Salary Survey Data is:

$$S = \beta_0 + \beta_1 X + \gamma_1 E_1 + \gamma_2 E_2 + \delta_1 M + \epsilon$$

- There is a different valid regression equation for each of the six (three education and two managment) categories.

Category	E	M	Regression Equation
1	1	0	$S = (\beta_0 + \gamma_1) + \beta_1 X + \epsilon$
2	1	1	$S = (\beta_0 + \gamma_1 + \delta_1) + \beta_1 X + \epsilon$
3	2	0	$S = (\beta_0 + \gamma_2) + \beta_1 X + \epsilon$
4	2	1	$S = (\beta_0 + \gamma_2 + \delta_1) + \beta_1 X + \epsilon$
5	3	0	$S = \beta_0 + \beta_1 X + \epsilon$
6	3	1	$S = (\beta_0 + \delta_1) + \beta_1 X + \epsilon$

Indicator Variables

```
d <- P130
d$E1 <- as.numeric(d$E == 1)
d$E2 <- as.numeric(d$E == 2)
mod <- lm(S ~ 1 + X + E1 + E2 + M, data=d)
summary(mod)
```

```
##
## Call:
## lm(formula = S ~ 1 + X + E1 + E2 + M, data = d)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1884.6  -653.6    22.2   844.9  1716.5
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   11031.8      383.2   28.79  < 2e-16 ***
## X               546.2       30.5   17.90  < 2e-16 ***
## E1            -2996.2      411.8   -7.28  6.7e-09 ***
## E2              147.8      387.7    0.38      0.7
## M              6883.5      313.9   21.93  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1030 on 41 degrees of freedom
## Multiple R-squared:  0.957,  Adjusted R-squared:  0.953
## F-statistic:  227 on 4 and 41 DF,  p-value: <2e-16
```

Your turn

Interpret the regression coefficients. Assume that the residual patterns are satisfactory.

Table 3

	Dependent variable:	
	S	
	(1)	(2)
X	546.200*** (30.520)	570.100*** (38.560)
E1	−2,996.000*** (411.800)	
E2	147.800 (387.700)	
E		1,579.000*** (262.300)
M	6,884.000*** (313.900)	6,688.000*** (398.300)
Constant	11,032.000*** (383.200)	6,963.000*** (665.700)
Observations	46	46
R^2	0.957	0.928
Adjusted R^2	0.953	0.923
Residual Std. Error	1,027.000 (df = 41)	1,313.000 (df = 42)
F Statistic	226.800*** (df = 4; 41)	179.600*** (df = 3; 42)

Note: *p<0.1; **p<0.05; ***p<0.01

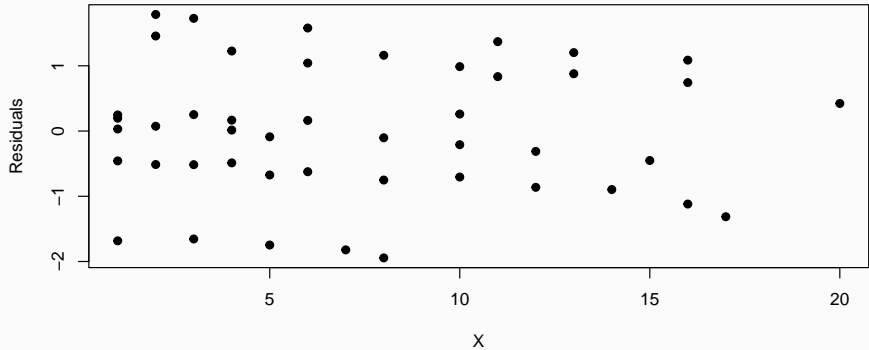**Before we continue we check the residuals**

1. Residuals vs. Years of Experience
2. Residuals vs. Categories from Dummys

Regression Diagnostics

```
plot(x = d$X, y = rstandard(mod), pch=19,
     ylab="Residuals", xlab = "X",
     main = "Standardized Residuals vs. Years of Experience (X)")
```
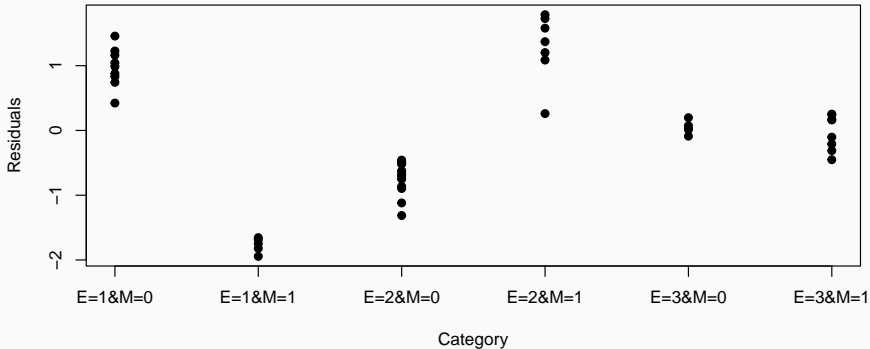

Standardized Residuals vs. Years of Experience (X)

Regression Diagnostics

```
d$cat <- factor((paste0("E=",d$E,"&M=",d$M)))
plot(x = as.numeric(d$cat), y = rstandard(mod), pch=19, xaxt="n",
     ylab="Residuals", xlab = "Category",
     main = "Standardized Residuals vs. Education-Management Category")
axis(1,at=1:6,labels=levels(d$cat))
```


Standardized Residuals vs. Education–Management Category

What is wrong with the residuals:

- Depending on the category the residuals are almost entirely positive or negative.
- The **pattern of the residuals is highly moderated by the associated group** (education-management category). This makes it clear that the combinations of education and management have not been treated sufficiently in the model.
- The residual plots provide evidence that the effects of education and management status on salary determination are **not additive**.

The multiplicative pattern needs to be embedded in the model!

- Interaction effects are *multiplicative* effects that allow capturing nonadditive effects in variables.
- Interaction variables are products of existing indicator variables.
- Using the Salary Survey Data this can be achieved by creating the two interaction effects ($E_1 \cdot M$) and ($E_2 \cdot M$) and **adding** them to the model.
- The interaction effects **do not replace** the indicator variables.

Interaction Effects

```
mod <- lm(S ~ 1 + X + E1 + E2 + M + E1*M + E2*M, data=d)
summary(mod)
```

Your turn

Is that model sufficient?

```
##
## Call:
## lm(formula = S ~ 1 + X + E1 + E2 + M + E1 * M + E2 * M, data = d)
##
## Residuals:
##    Min    1Q Median    3Q    Max
## -928.1  -46.2   24.3   65.9  204.9
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 11203.43      79.07  141.70  < 2e-16 ***
## X             496.99       5.57   89.28  < 2e-16 ***
## E1          -1730.75     105.33  -16.43  < 2e-16 ***
## E2           -349.08      97.57   -3.58  0.00095 ***
## M            7047.41     102.59   68.70  < 2e-16 ***
## E1:M        -3066.04     149.33  -20.53  < 2e-16 ***
## E2:M         1836.49     131.17   14.00  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 174 on 39 degrees of freedom
## Multiple R-squared:  0.999,  Adjusted R-squared:  0.999
## F-statistic: 5.52e+03 on 6 and 39 DF,  p-value: <2e-16
```

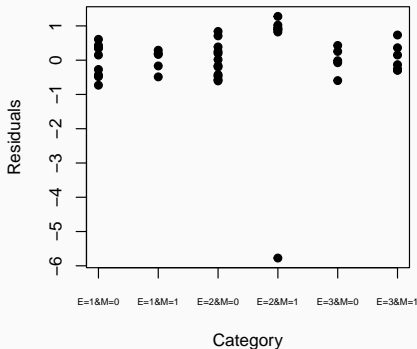# Regression Diagnostics

```
summary(rstandard(mod))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -5.773  -0.286   0.150   0.001   0.418   1.277
```


Standardized Residuals vs. Years of Experience (X)
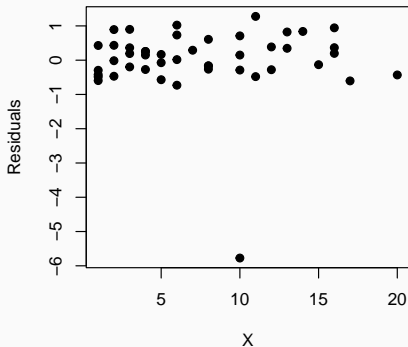
Standardized Residuals vs. Education–Management Category

Regression Diagnostics

```
d$res     <- residuals(mod)
d$res_std <- rstandard(mod)
tail(d, n=15)
```

```
##          S  X  E M E1 E2     cat     res res_std
## 32 23174 10 3 1  0  0 E=3&M=1  -46.72 -0.2885
## 33 23780 10 2 1  0  1 E=2&M=1 -928.13 -5.7735
## 34 25410 11 2 1  0  1 E=2&M=1  204.89  1.2773
## 35 14861 11 1 0  1  0 E=1&M=0  -78.54 -0.4796
## 36 16882 12 2 0  0  1 E=2&M=0   63.80  0.3866
## 37 24170 12 3 1  0  0 E=3&M=1  -44.69 -0.2784
## 38 15990 13 1 0  1  0 E=1&M=0   56.48  0.3465
## 39 26330 13 2 1  0  1 E=2&M=1  130.91  0.8226
## 40 17949 14 2 0  0  1 E=2&M=0  136.83  0.8383
## 41 25685 15 3 1  0  0 E=3&M=1  -20.65 -0.1316
## 42 27837 16 2 1  0  1 E=2&M=1  146.95  0.9437
## 43 18838 16 2 0  0  1 E=2&M=0   31.85  0.1983
## 44 17483 16 1 0  1  0 E=1&M=0   58.52  0.3648
## 45 19207 17 2 0  0  1 E=2&M=0  -96.14 -0.6047
## 46 19346 20 1 0  1  0 E=1&M=0  -66.43 -0.4310
```

```
d <- d[-33, ] # Remove problematic observation
```

Interaction Effects

```
##                          Model Summary
## -----------------------------------------------------------------
## R                     1.000      RMSE                 67.119
## R-Squared             1.000      Coef. Var             0.392
## Adj. R-Squared        1.000      MSE                4504.951
## Pred R-Squared        1.000      MAE                  51.794
## -----------------------------------------------------------------
##                               ANOVA
## -----------------------------------------------------------------
##                  Sum of
##                  Squares      DF      Mean Square      F         Sig.
## -----------------------------------------------------------------
## Regression   957607113.080     6     159601185.513   35427.955   0.0000
## Residual        171188.120    38         4504.951
## Total        957778301.200    44
## -----------------------------------------------------------------
##                         Parameter Estimates
## -----------------------------------------------------------------
##     model        Beta    Std. Error   Std. Beta      t       Sig        lower        upper
## -----------------------------------------------------------------
## (Intercept)   11199.714     30.533                366.802   0.000    11137.902    11261.525
##           X     498.418      2.152      0.557     231.640   0.000      494.062      502.774
##          E1   -1741.336     40.683     -0.304     -42.803   0.000    -1823.693    -1658.979
##          E2    -357.042     37.681      0.052      -9.475   0.000     -433.324     -280.761
##           M    7040.580     39.619      0.738     177.707   0.000     6960.376     7120.785
##        E1:M   -3051.763     57.674     -0.149     -52.914   0.000    -3168.519    -2935.008
##        E2:M    1997.531     51.785      0.103      38.574   0.000     1892.697     2102.364
## -----------------------------------------------------------------
```

Note: The level accuracy with which the model explains the data is very rare! Usually Goodness of fit indicators are worse.

Interaction Effects

```
mod <- lm(S ~ 1 + X + E1 + E2 + M + E1*M + E2*M, data=d)
equatiomatic::extract_eq(mod, use_coefs=F, intercept="beta", wrap=T)
```

$$S = \beta_0 + \beta_1(X) + \beta_2(E1) + \beta_3(E2) +$$
$$\beta_4(M) + \beta_5(E1 \times M) + \beta_6(E2 \times M) + \epsilon$$

(1)

```
equatiomatic::extract_eq(mod, use_coefs=T, coef_digits=4, wrap=T)
```

$$\widehat{S} = 11199.7138 + 498.4178(X) - 1741.3359(E1) - 357.0423(E2) +$$
$$7040.5801(M) - 3051.7633(E1 \times M) + 1997.5306(E2 \times M)$$

(2)

Your Turn

Compare the models `mod1`, `mod2` and `mod3`. Use them to calculate the base salaries (no experience) for each of the six possible education-management categories.

```
# Data Preparation
d <- P130[-33, ]
d$cat <- factor((paste0("E=",d$E,"&M=",d$M)))
d$E.fac <- factor(d$E)

# Model estimation
mod1 <- lm(S ~ 1 + X + E.fac + M + E.fac*M, data=d)
mod2 <- lm(S ~ 1 + X + cat, data=d)
mod3 <- lm(S ~ 1 + X + E.fac*M, data=d)
```

Category	E	M	Estimated Base Salary	95% CI Low	95% CI High
1	1	0	9458	9396	9521
2	2	1	19881	19814	19947
3	3	0	11200	11138	11262
4	1	1	13447	13383	13511
5	2	0	10843	10790	10896
6	3	1	18240	18183	18298

- All models lead to the **same estimates for the base salaries**. This shows that from a technical point using the `cat` variable (instead of the intercation effects) allows to capture the variation in the data.
- It is still **beneficial to use interaction effects** as we did, because this allows to separate the effects of the three sets of predictor variables education, management and education-management interaction.

Systems of Regression Equations

A data set may consist of **two or more distinct subsets**, which may require individual regression equations to avoid bias. Subsets may occur cross-sectional or over time and need to be treated differently:

- Cross-Sectional Data
    1. Each group has a separate regression model.
    2. The models have the same intercept but different slopes.
    3. the models have the same slope but different intercepts.
- Time Series Data
    1. Calendar Effects, e.g. Seasonality
    2. Stability of regression parameters over time

Example: Preemployment Test

```
##    TEST RACE JPERF
## 1  0.28    1  1.83
## 2  0.97    1  4.59
## 3  1.25    1  2.97
## 4  2.46    1  8.14
## 5  2.51    1  8.00
## 6  1.17    1  3.30
## 7  1.78    1  7.53
## 8  1.21    1  2.03
## 9  1.63    1  5.00
## 10 1.98    1  8.04
## 11 2.36    0  3.25
## 12 2.11    0  5.30
## 13 0.45    0  1.39
## 14 1.76    0  4.69
## 15 2.09    0  6.56
## 16 1.50    0  3.00
## 17 1.25    0  5.85
## 18 0.72    0  1.90
## 19 0.42    0  3.85
## 20 1.53    0  2.95
```

Your turn

TEST Score on the preemployment test.

RACE Dummy to indicate if individual is part of a minority (1) or not (0).

JPERF Job Performance Ranking after 6 weeks on the job.
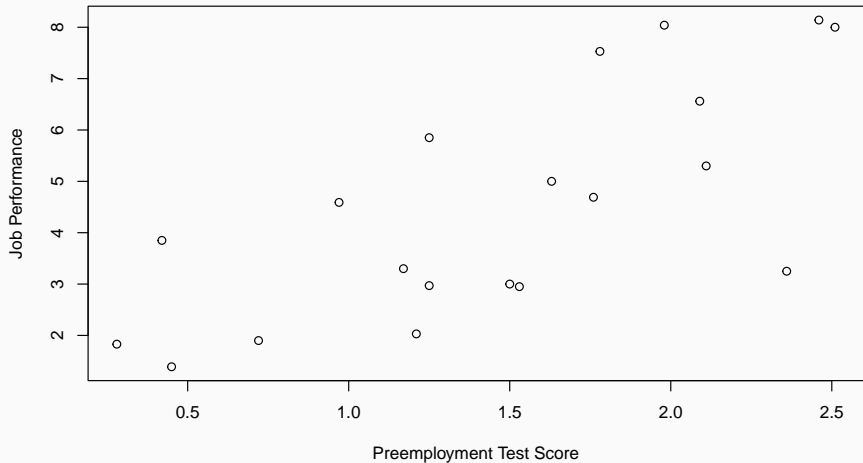
Example: Preemployment Test

For simplicity and generality we refer to the job performance as *Y* and the score on the preemployment test as *X*. We want to compare the following two models:

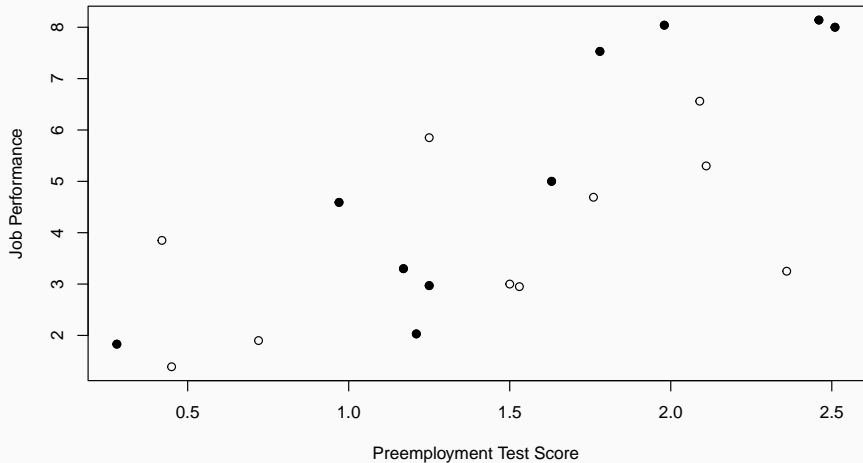Model 1 (Pooled): $\qquad y_{ij} = \beta_0 + \beta_1 x_{ij} + \epsilon_{ij}$

Model 2 (Minority): $\qquad y_{i1} = \beta_{01} + \beta_{11} x_{i1} + \epsilon_{i1}$

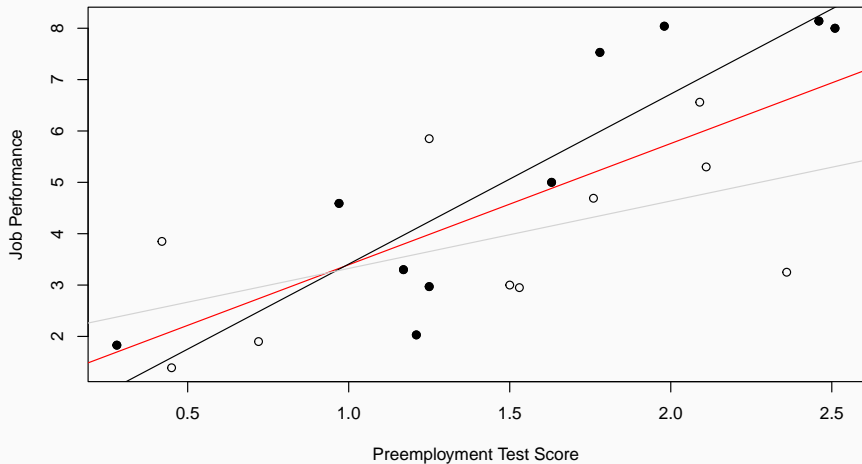Model 2 (non Minority): $\quad y_{i2} = \beta_{02} + \beta_{12} x_{i2} + \epsilon_{i2}$

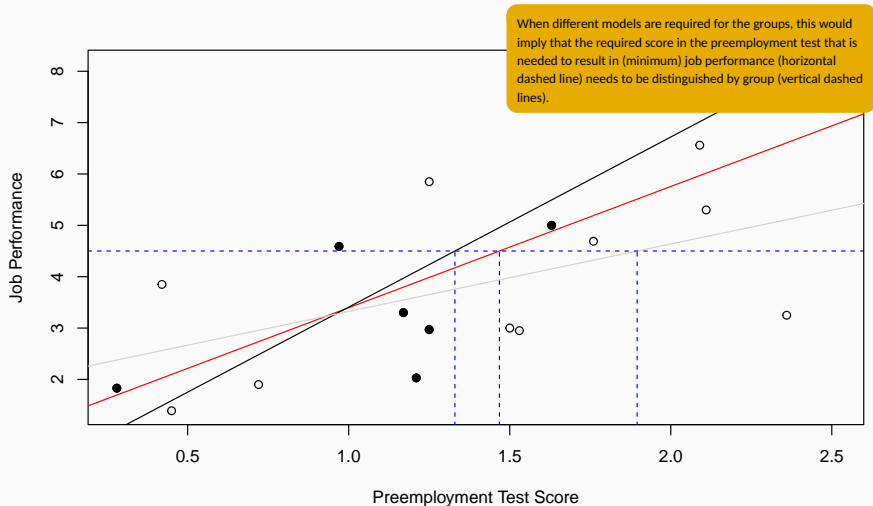When different models are required for the groups, this would imply that the required score in the preemployment test that is needed to result in (minimum) job performance (horizontal dashed line) needs to be distinguished by group (vertical dashed lines).

Models with different Slopes and different Intercepts

- What we want to test the Preemployment Test data for are differences in intercept and slope using the following Null.

$$H_0 : \beta_{11} = \beta_{12}, \beta_{01} = \beta_{02}$$

- This test can be performed using an **interaction term** by using a variable z_{ij} that takes the value 1 if an individual is part of a minority group and 0 otherwise. This leads to two relevant models:

Model 1 (Pooled): $y_{ij} = \beta_0 + \beta_1 x_{ij} + \epsilon_{ij}$

Model 3 (Interaction): $y_{ij} = \beta_0 + \beta_1 x_{ij} + \gamma z_{ij} + \delta(z_{ij} \cdot x_{ij}) + \epsilon_{ij}$

- This model is **equivalent** to the previously discussed Model 2.

Models with different Slopes and different Intercepts

|	Model 1	Model 2	Model 2	Model 3
	Pooled	Minority	White	Interaction
(Intercept)	1.03	0.10	2.01	2.01
	(0.87)	(1.04)	(1.13)	(1.05)
TEST	2.36^{***}	3.31^{***}	1.31	1.31
	(0.54)	(0.62)	(0.72)	(0.67)
RACE				-1.91
				(1.54)
TEST:RACE				2.00
				(0.95)
R^2	0.52	0.78	0.29	0.66
Adj. R^2	0.49	0.75	0.20	0.60
Num. obs.	20	10	10	20

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table 5

- Model 1 can be seen as a restricted version (RM) of model 3, the full model (FM), with $\gamma = \delta = 0$.

Models with different Slopes and different Intercepts

```
df <- cbind(P140,res = rstandard(mod3))
plot(x=df$TEST, y=df$res,
     ylab="Residuals", xlab="Preemployment Test Score")
points(df$TEST[df$RACE == T], df$res[df$RACE == T],pch=19)
```
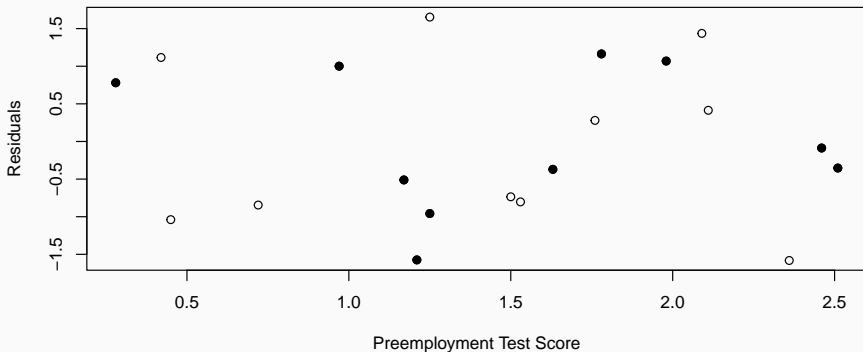
Models with different Slopes and different Intercepts

- The framework using the models as FM and [RM] for comparison.

$$F = \frac{[SSE(RM) - SSE(\ldots)]}{SSE(FM)/1\ldots}$$

```
(SSE_RM <- sum(residuals(mod1)^2))
```

```
## [1] 45.57
```

```
(SSE_FM <- sum(residuals(mod3)^2))
```

```
## [1] 31.66
```

```
(F_stat <- ((SSE_RM - SSE_FM)/2)/(SSE_FM/16))
```

```
## [1] 3.516
```

```
pf(F_stat, df1=2, df2=16, lower.tail=FALSE)
```

```
## [1] 0.05424
```

34

Models with same Slope and different Intercepts

- Assuming we have a reason to believe that only the intercepts for the two groups are different can be achieved using the indicator variable (and omitting the interaction term).

  Model 1 (Pooled): $\quad y_{ij} = \beta_0 + \beta_1 x_{ij} + \epsilon_{ij}$

  Model 4 (Indicator only): $\quad y_{ij} = \beta_0 + \beta_1 x_{ij} + \gamma z_{ij} + \underline{\delta(z_{ij} \cdot x_{ij})} + \epsilon_{ij}$

- In the case where $z_{ij} = 1$ (which indicates the non-minority group) the coefficient γ can be added to the intercept β_0 to obtain the effective intercept for that respecitve group.

- The resulting models represent **two parallel lines** (same slopes) with intercepts β_0 and $\beta_0 + \gamma$.

Models with same Slope and different Intercepts

```
mod4  <- lm(JPERF ~ 1 + TEST + RACE, data=P140)
```

- Significance can be tested using the *F*-Test. As the FM and RM differ by one parameter, results are eqauivalent to the *t*-Test.

	Model 1	Model 2	Model 2	Model 3	Model 4
	Pooled	Minority	White	Interaction	Indicator
(Intercept)	1.03	0.10	2.01	2.01	0.61
	(0.87)	(1.04)	(1.13)	(1.05)	(0.89)
TEST	2.36***	3.31***	1.31	1.31	2.30***
	(0.54)	(0.62)	(0.72)	(0.67)	(0.52)
RACE				-1.91	1.03
				(1.54)	(0.69)
TEST:RACE				2.00	
				(0.95)	
R^2	0.52	0.78	0.29	0.66	0.57
Adj. R^2	0.49	0.75	0.20	0.60	0.52
Num. obs.	20	10	10	20	20

$^{***}p < 0.001; ^{**}p < 0.01; ^{*}p < 0.05$

Table 6

Models with different Slopes and same Intercept

■ Finally we can hypothesize that the two groups have the same intercept β_0 but different slopes, which can be done by including only the interaction.

Model 1 (Pooled): $y_{ij} = \beta_0 + \beta_1 x_{ij} + \epsilon_{ij}$

Model 5 (Interaction only): $y_{ij} = \beta_0 + \beta_1 x_{ij} + \gamma z_{ij} + \delta(z_{ij} \cdot x_{ij}) + \epsilon_{ij}$

```
mod5  <- lm(JPERF ~ 1 + TEST + RACE:TEST, data=P140)
```

■ Inference for the δ can be carrioud out using the *F*-Test or the *t*-Test. The FM and RM again only differ by one parameter.

Systems of Regeression Equations

■ The final results for all discussed cases for the preemployment test data look like follows.

	Model 1	Model 2	Model 2	Model 3	Model 4	Model 5
	Pooled	Minority	White	Full Interaction	Indicator	Interaction
(Intercept)	1.03	0.10	2.01	2.01	0.61	1.12
	(0.87)	(1.04)	(1.13)	(1.05)	(0.89)	(0.78)
TEST	2.36^{***}	3.31^{***}	1.31	1.31	2.30^{***}	1.83^{**}
	(0.54)	(0.62)	(0.72)	(0.67)	(0.52)	(0.54)
RACE				-1.91	1.03	
				(1.54)	(0.69)	
TEST:RACE				2.00		0.92^{*}
				(0.95)		(0.40)
R^2	0.52	0.78	0.29	0.66	0.57	0.63
Adj. R^2	0.49	0.75	0.20	0.60	0.52	0.59
Num. obs.	20	10	10	20	20	20

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table 7

Time Series Data

- Another interesting field of study is temporal structure in the data, which could fill a whole course by itself. Therefore we only briefly look at two ideas.

1) Calendar Effects, e.g. Seasonality

- Can be modeled by including time as regressor, e.g. in the form of (mulitple) indicators for e.g. Week/Month/Quarter/Year
- The number of indicator variables is $m - 1$ where m is the frequency of the time effects (e.g. $m = 4$ for Quarters).

2) Stability of Parameters over Time

- By combining inidcator and interaction terms one can model intertemporal and interspatial relationships. Insignifcance of the interactions with all indicators then provices evidence stability over time.