Please Click https://canvas.ust.hk and SFQ on the left panel To Fill out End-of-Term Course Survey.

Thanks for your attention!

Chapter 5. Qualitative Variables as Predictors

Outline

5.1 Introduction and the Salary Survey Data

5.2 Interaction Variables

5.3 Systems of Regression Equations

5.4 Other Applications of Indicative Variables

5.5 Seasonality

5.6 An Example using R

5.1. Introduction and the Salary Survey Data

5.1 Introduction and the Salary Survey Data

Introduction

Qualitative *or* ***categorical*** *variables can be very useful as* ***predictor*** *variables in regression analysis. Qualitative variables such as gender, marital status, or political affiliation can be represented by* ***indicator*** *or* ***dummy*** *variables. These variables take on only* ***two*** *values, usually 0 and 1. The two values signify that the observation belongs to one of two possible categories. The numerical values of indicator variables are* ***not*** *intended to reflect a* ***quantitative ordering*** *of the categories, but only serve to* ***identify*** *category or class membership.*

For example, an analysis of salaries earned by computer programmers may include variables such as ***education****, years of* ***experience****, and* ***gender*** *as predictor variables. The gender variable could be quantified, say, as 1 for female and 0 for male.* ***Indicator*** *variables can also be used in a regression equation to distinguish among three or more groups as well as among classifications across various types of groups. For example, the regression described above may also include an* ***indicator*** *variable to distinguish whether the observation was for a systems or applications programmer. The* ***four*** *conditions determined by* ***gender*** *and* ***type*** *of programming can be represented by combining the two variables.*

5.1 Introduction and the Salary Survey Data

Introduction

Indicator *variables can be used in a variety of ways and may be considered whenever there are qualitative variables affecting a* ***relationship****. We shall illustrate some of the applications with examples and suggest some additional applications. It is hoped that we will recognize the general* ***applicability*** *of the technique from the examples.*

In the ***first*** *example, we look at data on a salary survey, such as the one mentioned above, and use indicator variables to adjust for various categorical variables that affect the regression relationship. The* ***second*** *example uses indicator variables for analyzing and testing for equality of regression relationships in various subsets of a population.*

We continue to assume that the ***response*** *variable is a* ***quantitative*** *continuous variable, but the* ***predictor*** *variables can be* ***quantitative*** *and/or categorical. The case where the* ***response*** *variable is an* ***indicator*** *variable is dealt with in later chapters.*

5.1 Introduction and the Salary Survey Data

Salary Survey Data

The Salary Survey data set was developed from a salary survey of computer professionals in a large corporation. The objective of the survey was to identify and quantify those variables that determine ***salary*** *differentials. In addition, the data be used to determine if the corporation's salary administration guidelines were being followed. The data appear in Table 5.1. The response variable is* ***salary*** *(S) and the predictors are: (1)* ***experience*** *(X), measured in years; (2)* ***education*** *(E), coded as 1 for completion of a high school (H.S.) diploma, 2 for completion of a bachelor degree (B.S.), and 3 for the completion of an advanced degree; and (3)* ***management*** *(M), which is coded as 1 for a person with management responsibility and 0 otherwise. We shall try to measure the effects of these three variables on salary using regression analysis.*

A ***linear*** *relationship will be used for salary and experience. We shall assume that each additional year of experience is worth a* ***fixed*** *salary increment. Education may also be treated in a linear fashion. If the education variable is used in the regression equation in* ***raw*** *form, we would be assuming that each step up in education is worth a* ***fixed increment*** *in salary. That is, with all other variables held constant, the relationship between salary and education is linear. That interpretation is possible but may be too* ***restrictive****. Instead, we shall view education as a* ***categorical*** *variable and define* ***two*** *indicator variables to represent the* ***three*** *categories. These two variables allow us to pick up the effect of education on salary whether or not it is linear. The management variable is also an* ***indicator*** *variable designating the two categories, 1 for management positions and 0 for regular staff positions.*

Table 5.1

5.1 Introduction and the Salary Survey Data

Salary Survey Data

Table 5.1 Salary Survey Data

Row	S	X	E	M	Row	S	X	E	M
1	13876	1	1	1	24	22884	6	2	1
2	11608	1	3	0	25	16978	7	1	1
3	18701	1	3	1	26	14803	8	2	0
4	11283	1	2	0	27	17404	8	1	1
5	11767	1	3	0	28	22184	8	3	1
6	20872	2	2	1	29	13548	8	1	0
7	11772	2	2	0	30	14467	10	1	0
8	10535	2	1	0	31	15942	10	2	0
9	12195	2	3	0	32	23174	10	3	1
10	12313	3	2	0	33	23780	10	2	1
11	14975	3	1	1	34	25410	11	2	1
12	21371	3	2	1	35	14861	11	1	0
13	19800	3	3	1	36	16882	12	2	0
14	11417	4	1	0	37	24170	12	3	1
15	20263	4	3	1	38	15990	13	1	0
16	13231	4	3	0	39	26330	13	2	1
17	12884	4	2	0	40	17949	14	2	0
18	13245	5	2	0	41	25685	15	3	1
19	13677	5	3	0	42	27837	16	2	1
20	15965	5	1	1	43	18838	16	2	0
21	12336	6	1	0	44	17483	16	1	0
22	21352	6	3	1	45	19207	17	2	0
23	13839	6	2	0	46	19346	20	1	0

5.1 Introduction and the Salary Survey Data

Salary Survey Data

Note that when using indicator variables to represent a set of categories, the number of these variables required is one less than the number of categories. For example, in the case of the education categories above, we create two indicator variables E_1 and E_2, where

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

and

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

As stated above, these two variables taken together uniquely represent the three groups. For H.S., $E_1 = 1, E_2 = 0$; for B.S., $E_1 = 0, E_2 = 1$; and for advanced degree, $E_1 = 0, E_2 = 0$. Furthermore, if there were a third variable, E_{i3}, defined to be 1 or 0 depending on whether or not the ith person is in the advanced degree category, then for each person we have $E_1 + E_2 + E_3 = 1$. Then $E_3 = 1 - E_1 - E_2$, showing clearly that one of the variables is superfluous.[2] Similarly, there is only one indicator variable required to distinguish the two management categories. The category that is not represented by an indicator variable is referred to as the *base category* or the *control group* because the regression coefficients of the indicator variables are interpreted relative to the control group.

In terms of the indicator variables described above, the regression model is

$$S = \beta_0 + \beta_1 X + \gamma_1 E_1 + \gamma_2 E_2 + \delta_1 M + \varepsilon. \qquad (5.1)$$

5.1 Introduction and the Salary Survey Data

Salary Survey Data with indicator variable

```
############## create indicator variables for salary data
#############
Salary<-read.table('data/P130.txt',header=TRUE)   ## read the data
n<-dim(Salary)[1]
p<-dim(Salary)[2]-1
Salary$E1<-rep(0,n)
Salary$E1[which(Salary$E==1)]<-1
Salary$E2<-rep(0,n)
Salary$E2[which(Salary$E==2)]<-1
Salary_new<-subset(Salary,select=c("S","X","E1","E2","M"))
```

```
> Salary_new
       S  X E1 E2 M
1  13876  1  1  0 1
2  11608  1  0  0 0
3  18701  1  0  0 1
4  11283  1  0  1 0
5  11767  1  0  0 0
6  20872  2  0  1 1
7  11772  2  0  1 0
8  10535  2  1  0 0
9  12195  2  0  0 0
10 12313  3  0  1 0
11 14975  3  1  0 1
12 21371  3  0  1 1
13 19800  3  0  0 1
14 11417  4  1  0 0
15 20263  4  0  0 1
16 13231  4  0  0 0
17 12884  4  0  1 0
18 13245  5  0  1 0
19 13677  5  0  0 0
20 15965  5  1  0 1
21 12336  6  1  0 0
22 21352  6  0  0 1
23 13839  6  0  1 0
24 22884  6  0  1 1
25 16978  7  1  0 1
26 14803  8  0  1 0
27 17404  8  1  0 1
28 22184  8  0  0 1
29 13548  8  1  0 0
30 14467 10  1  0 0
31 15942 10  0  1 0
32 23174 10  0  0 1
33 23780 10  0  1 1
34 25410 11  0  1 1
35 14861 11  1  0 0
36 16882 12  0  1 0
37 24170 12  0  0 1
38 15990 13  1  0 0
```

5.1 Introduction and the Salary Survey Data

Salary Survey Data

By evaluating (5.1) for the different values of the indicator variables, it follows that there is a different regression equation for each of the six (three education and two management) categories as shown in Table 5.2. According to the proposed model, we may say that the indicator variables help to determine the base salary level as a function of education and management status after adjustment for years of experience.

The results of the regression computations for the model given in (5.1) appear in Table 5.3. The proportion of salary variation accounted for by the model is quite high ($R^2 = 0.957$). At this point in the analysis we should investigate the pattern of residuals to check on model specification. We shall postpone that investigation for now and assume that the model is satisfactory so that we can discuss the interpretation of the regression results. Later we shall return to analyze the residuals and find that the model must be altered.

We see that the coefficient of X is 546.16. That is, each additional year of experience is estimated to be worth an annual salary increment of \$546. The other coefficients may be interpreted by looking into Table 5.2. The coefficient of the management indicator variable, δ_1, is estimated to be 6883.50. From Table 5.2 we interpret this amount to be the average incremental value in annual salary associated with a management position. For the education variables, γ_1 measures the salary differential for the H.S. category relative to the advanced degree category and γ_2 measures the differential for the B.S. category relative to the advanced degree category. The difference, $\gamma_2 - \gamma_1$, measures the differential salary for the H.S. category relative to the B.S. category. From the regression results, in terms of salary for computer professionals, we see that an advanced degree is worth \$2996 more than a high school diploma, a B.S. is worth \$148 more than an advanced degree (this differential is not statistically significant, $t = 0.38$), and a B.S. is worth about \$3144 more than a high school diploma. These salary differentials hold for every fixed level of experience.

Table 5.2, 5.3

5.1 Introduction and the Salary Survey Data

Salary Survey Data

Table 5.2 Regression Equations for the Six Categories of Education and Management

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

Table 5.3 Regression Analysis of Salary Survey Data

Variable	Coefficient	s.e.	t-Test	p-value
Constant	11031.800	383.2	28.80	< 0.0001
X	546.184	30.5	17.90	< 0.0001
E_1	-2996.210	411.8	-7.28	< 0.0001
E_2	147.825	387.7	0.38	0.7049
M	6883.530	313.9	21.90	< 0.0001
$n = 46$	$R^2 = 0.957$	$R_a^2 = 0.953$	$\hat{\sigma} = 1027$	df = 41

Hypothesis test, confidence interval and *prediction interval* can be performed as was done in usual multiple linear regression like Chapter 3.

5.1 Introduction and the Salary Survey Data

Use R for last example

```
> ### linear regression on Salary_new
> qua_mod<-lm(S~.,data=Salary_new)
> summary(qua_mod)

Call:
lm(formula = S ~ ., data = Salary_new)

Residuals:
     Min       1Q   Median       3Q      Max
-1884.60  -653.60    22.23   844.85  1716.47

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 11031.81     383.22  28.787  < 2e-16 ***
X             546.18      30.52  17.896  < 2e-16 ***
E1          -2996.21     411.75  -7.277 6.72e-09 ***
E2            147.82     387.66   0.381    0.705
M            6883.53     313.92  21.928  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1027 on 41 degrees of freedom
Multiple R-squared:  0.9568,    Adjusted R-squared:  0.9525
F-statistic: 226.8 on 4 and 41 DF,  p-value: < 2.2e-16
```

5.1 Introduction and the Salary Survey Data

Use R for last example

Actually, R can automatically perform linear regression without introducing the additional indicator variables. R treats a categorical variable as a *factor* and perform the linear regression.

```
> ### linear regression directly on qualitative predictors
> Salary<-read.table('data/P130.txt',header=TRUE)   ## read the raw data
> Salary$E<-as.factor(Salary$E)   ## change it to factor variable
> Salary<-within(Salary,E<-relevel(E,ref=3))   ## treat category 3 as the reference category
> Salary$M<-as.factor(Salary$M)   ## change it to factor variable
> Salary<-within(Salary,M<-relevel(M,ref=1))   ## treat the category 1 as the reference category
> qua_mod1<-lm(S~.,data=Salary)
> summary(qua_mod1)

Call:
lm(formula = S ~ ., data = Salary)

Residuals:
     Min       1Q   Median       3Q      Max
-1884.60  -653.60    22.23   844.85  1716.47

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 11031.81     383.22  28.787  < 2e-16 ***
X             546.18      30.52  17.896  < 2e-16 ***
E1          -2996.21     411.75  -7.277 6.72e-09 ***
E2            147.82     387.66   0.381    0.705
M1           6883.53     313.92  21.928  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1027 on 41 degrees of freedom
Multiple R-squared:  0.9568,    Adjusted R-squared:  0.9525
F-statistic: 226.8 on 4 and 41 DF,  p-value: < 2.2e-16
```

5.2. Interaction Variables

5.2 Interaction Variables

Returning now to the question of model specification, consider Figure 5.1, where the residuals are plotted against X. The plot suggests that there may be three or more specific levels of residuals. Possibly the indicator variables that have been defined are not adequate for explaining the effects of education and management status. Actually, each residual is identified with one of the six education-management combinations. To see this we plot the residuals against Category (a new categorical variable that takes a separate value for each of the six combinations). This graph is, in effect, a plot of residuals versus a potential predictor variable that has not yet been used in the equation. The graph is given in Figure 5.2. It can be seen from the graph that the residuals cluster by size according to their education-management category. The combinations of education and management have not been satisfactorily treated in the model. Within each of the six groups, the residuals are either almost totally positive or totally negative. This behavior implies that the model given in (5.1) does not adequately explain the relationship between salary and experience, education, and management variables. The graph points to some hidden structure in the data that has not been explored.

Figure 5.1, 5.2

5.2 Interaction Variables

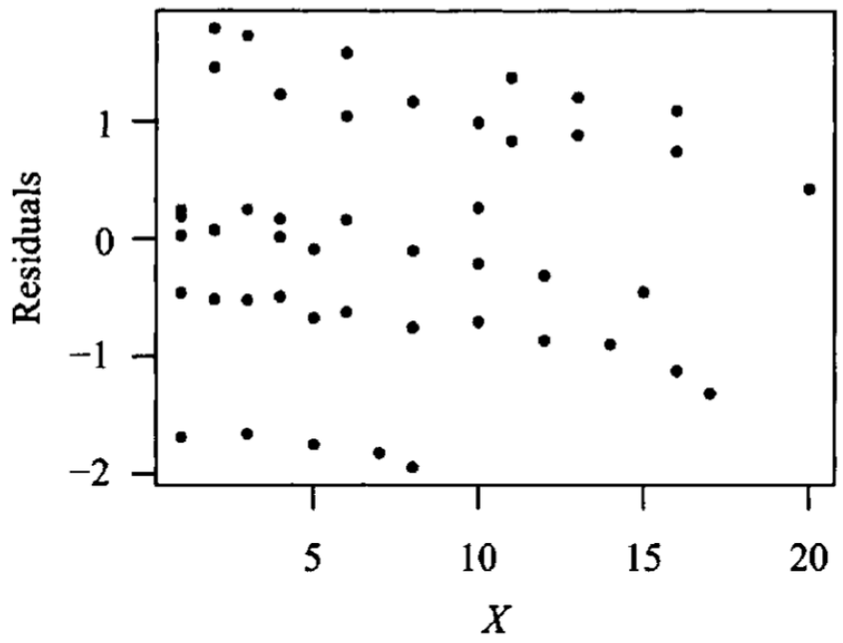


Figure 5.1 Standardized residuals versus years of experience (X).

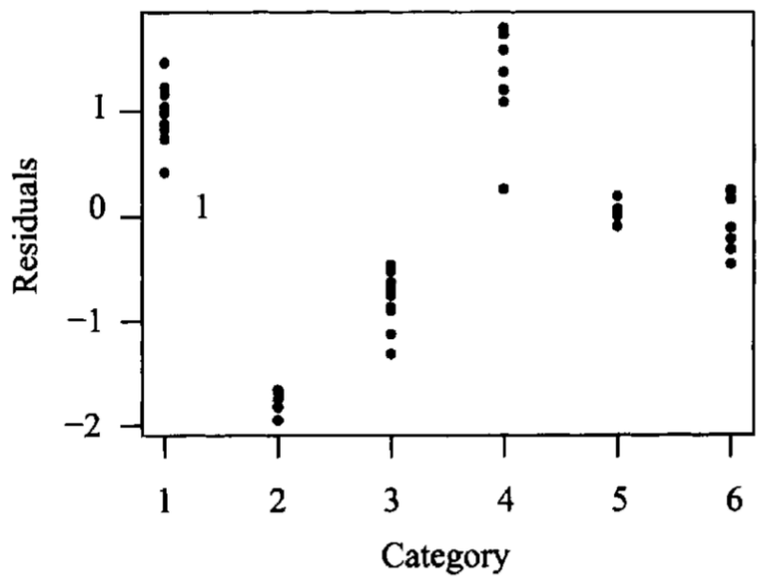


Figure 5.2 Standardized residuals versus education-management categorical variable.

5.2 Interaction Variables

The graphs strongly suggest that the effects of education and management status on salary determination are not additive. Note that in the model in (5.1) and its further exposition in Table 5.2, the incremental effects of both variables are determined by additive constants. For example, the effect of a management position is measured as δ_1, independently of the level of educational attainment. The nonadditive effects of these variables can be evaluated by constructing additional variables that are used to measure what may be referred to as *multiplicative* or *interaction effects*. Interaction variables are defined as products of the existing indicator variables $(E_1 \cdot M)$ and $(E_2 \cdot M)$. The inclusion of these two variables on the right-hand side of (5.1) leads to a model that is no longer additive in education and management, but recognizes the multiplicative effect of these two variables.

The expanded model is

$$\begin{aligned} S &= \beta_0 + \beta_1 X + \gamma_1 E_1 + \gamma_2 E_2 + \delta_1 M \\ &\quad + \alpha_1 (E_1 \cdot M) + \alpha_2 (E_2 \cdot M) + \varepsilon. \end{aligned} \tag{5.2}$$

The regression results are given in Table 5.4. The residuals from the regression of the expanded model are plotted against X in Figure 5.3. Note that observation 33 is an outlier. Salary is overpredicted by the model. Checking this observation in the listing of the raw data, it appears that this particular person seems to have fallen behind by a couple of hundred dollars in annual salary as compared to other persons with similar characteristics. To be sure that this single observation is not overly affecting the regression estimates, it has been deleted and the regression rerun. The new results are given in Table 5.5.

Table 5.4, 5.5
Figure 5.3

5.2 Interaction Variables

Table 5.4 Regression Analysis of Salary Data: Expanded Model

Variable	Coefficient	s.e.	t-Test	p-value
Constant	11203.40	79.07	141.7	< 0.0001
X	496.99	5.57	89.3	< 0.0001
E_1	-1730.75	105.30	-16.4	< 0.0001
E_2	-349.08	97.57	-3.6	0.0009
M	7047.41	102.60	68.7	< 0.0001
$E_1 \cdot M$	-3066.04	149.30	-20.5	< 0.0001
$E_2 \cdot M$	1836.49	131.20	14.0	< 0.0001
$n = 46$	$R^2 = 0.999$	$R_a^2 = 0.999$	$\hat{\sigma} = 173.8$	df = 39

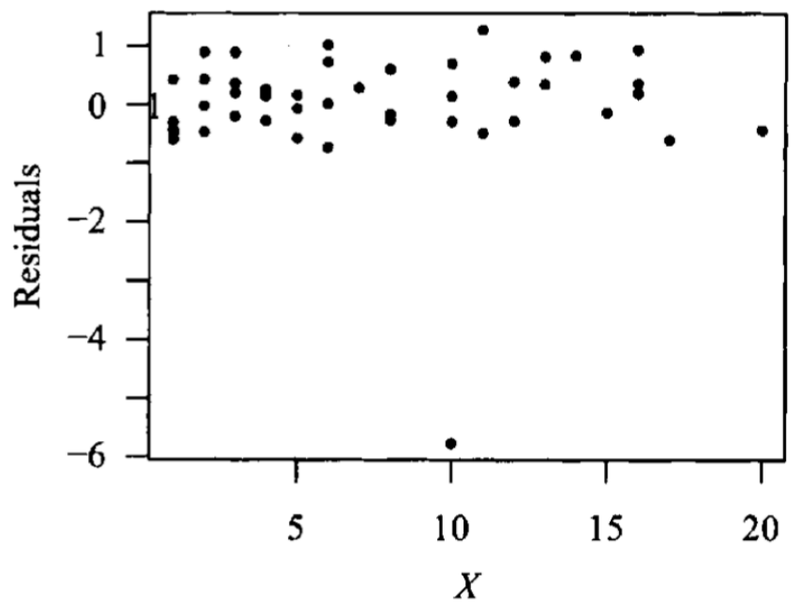


Figure 5.3 Standardized residuals versus years of experience: Expanded model.

Table 5.5

5.2 Interaction Variables

Table 5.5 Regression Analysis of Salary Data: Expanded Model, Observation 33 Deleted.

Variable	Coefficient	s.e.	t-Test	p-value
Constant	11199.70	30.54	367.0	< 0.0001
X	498.41	2.15	232.0	< 0.0001
E_1	−1741.28	40.69	−42.8	< 0.0001
E_2	−357.00	37.69	−9.5	< 0.0001
M	7040.49	39.63	178.0	< 0.0001
$E_1 \cdot M$	−3051.72	57.68	−52.9	< 0.0001
$E_2 \cdot M$	1997.62	51.79	38.6	< 0.0001
$n = 45$	$R^2 = 1.0$	$R_a^2 = 1.0$	$\hat{\sigma} = 67.13$	df = 38

*The regression coefficients are basically **unchanged**. However, the standard deviation of the **residuals** has been reduced to $67.28 and the proportion of **explained** variation has reached 0.9998. The plot of residuals versus X (Figure 5.4) appears **satisfactory** compared with the similar residual plot for the **additive** model. In addition, the plot of residuals for each education-management category (Figure 5.5) shows that each of these groups has residuals that appear to be **symmetrically** distributed about zero. Therefore the introduction of the interaction terms has produced an accurate representation of salary variations. The relationships between salary and experience, education, and management status appear to be adequately described by the model given in (5.2).*

Figure 5.4, 5.5

5.2 Interaction Variables

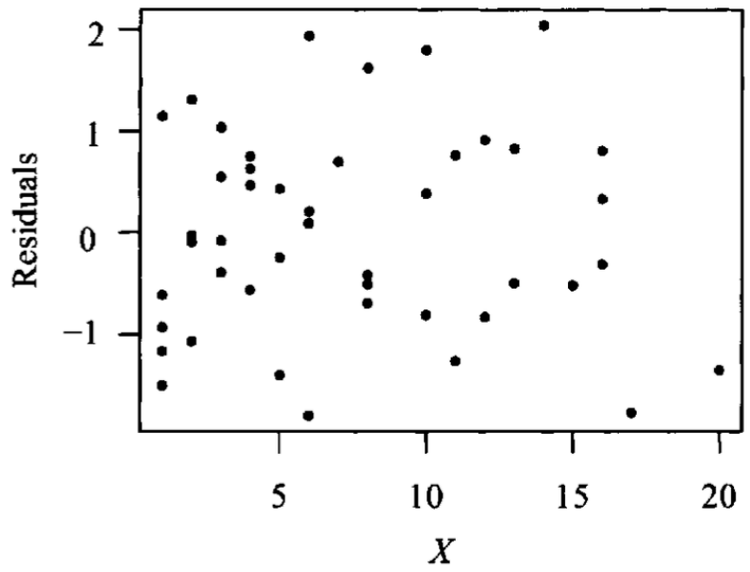


Figure 5.4 Standardized residuals versus years of experience: Expanded model, observation 33 deleted.

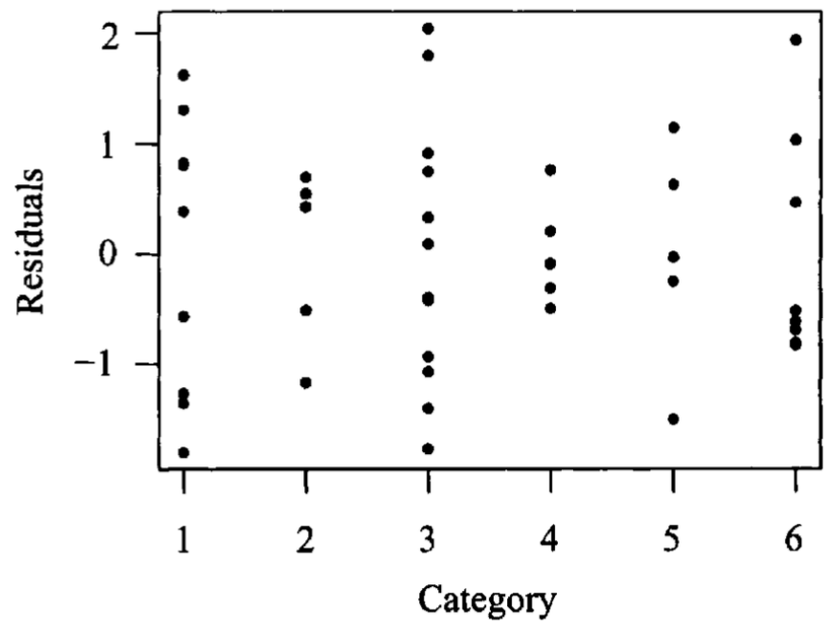


Figure 5.5 Standardized residuals versus education-management categorical variable: Expanded model, observation 33 deleted.

5.2 Interaction Variables

With the standard error of the residuals estimated to be \$67.28, we can believe that we have uncovered the actual and very carefully administered salary formula. Using 95% confidence intervals, each year of experience is estimated to be worth between \$494.08 and \$502.72. These increments of approximately \$500 are added to a starting salary that is specified for each of the six education-management groups. Since the final regression model is not additive, it is rather difficult to directly interpret the coefficients of the indicator variables. To see how the qualitative variables affect salary differentials, we use the coefficients to form estimates of the base salary for each of the six categories. These results are presented in Table 5.6 along with standard errors and confidence intervals. The standard errors are computed using the formulas in Chapter 3.

Using a regression model with indicator variables and interaction terms, it has been possible to account for almost all the variation in salaries of computer professionals selected for this survey. The level of accuracy with which the model explains the data is very rare! We can only conjecture that the methods of salary administration in this company are precisely defined and strictly applied.

Table 5.6

5.2 Interaction Variables

Table 5.6 Estimates of Base Salary Using the Nonadditive Model in (5.2)

Category	E	M	Coefficients	Estimate of Base Salary[a]	s.e.[a]	95% Confidence Interval
1	1	0	$\beta_0 + \gamma_1$	9459	31	(9398, 9520)
2	1	1	$\beta_0 + \gamma_1 + \delta + \alpha_1$	13448	32	(13385, 13511)
3	2	0	$\beta_0 + \gamma_2$	10843	26	(10792, 10894)
4	2	1	$\beta_0 + \gamma_2 + \delta + \alpha_2$	19880	33	(19815, 19945)
5	3	0	β_0	11200	31	(11139, 11261)
6	3	1	$\beta_0 + \delta$	18240	29	(18183, 18297)

[a] Recorded to the nearest dollar.

5.2 Interaction Variables

Remarks

In retrospect, we see that an equivalent model may be obtained with a different set of indicator variables and regression parameters. One could define five variables, each taking on the values of 1 or 0, corresponding to five of the six education-management categories. The numerical estimates of base salary and the standard errors of Table 5.6 would be the same. The advantage to proceeding as we have is that it allows us to separate the effects of the three sets of predictor variables, (1) education, (2) management, and (3) education-management interaction. Recall that interaction terms were included only after we found that an additive model did not satisfactorily explain salary variations. In general, we start with simple models and proceed sequentially to more complex models if necessary. We shall always hope to retain the simplest model that has an acceptable residual structure.

5.2 Interaction Variables

Use R on the expanded model

```
> ################   Interaction Model on Salary Data
> Salary_expd<-Salary_new
> Salary_expd$E1M <- Salary_expd$E1 * Salary_expd$M      ## create the new interaction variable E1*M
> Salary_expd$E2M <- Salary_expd$E2 * Salary_expd$M      ## create the new interaction variable E2*M
> expd_mod<-lm(S~.,data=Salary_expd)
> summary(expd_mod)

Call:
lm(formula = S ~ ., data = Salary_expd)

Residuals:
    Min      1Q  Median      3Q     Max
-928.13  -46.21   24.33   65.88  204.89

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 11203.434     79.065 141.698  < 2e-16 ***
X             496.987      5.566  89.283  < 2e-16 ***
E1          -1730.748    105.334 -16.431  < 2e-16 ***
E2           -349.078     97.568  -3.578 0.000945 ***
M            7047.412    102.589  68.695  < 2e-16 ***
E1M         -3066.035    149.330 -20.532  < 2e-16 ***
E2M          1836.488    131.167  14.001  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 173.8 on 39 degrees of freedom
Multiple R-squared:  0.9988,    Adjusted R-squared:  0.9986
F-statistic:  5517 on 6 and 39 DF,  p-value: < 2.2e-16
```

5.2 Interaction Variables

Use R on the expanded model

```
> ## delete the outlier observation
> expd_mod1<-lm(S~.,data=Salary_expd[-33,])
> summary(expd_mod1)

Call:
lm(formula = S ~ ., data = Salary_expd[-33, ])

Residuals:
     Min       1Q   Median       3Q      Max
-112.884  -43.636   -5.036   46.622  128.480

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 11199.714     30.533 366.802  < 2e-16 ***
X             498.418      2.152 231.640  < 2e-16 ***
E1          -1741.336     40.683 -42.803  < 2e-16 ***
E2           -357.042     37.681  -9.475 1.49e-11 ***
M            7040.580     39.619 177.707  < 2e-16 ***
E1M         -3051.763     57.674 -52.914  < 2e-16 ***
E2M          1997.531     51.785  38.574  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 67.12 on 38 degrees of freedom
Multiple R-squared:  0.9998,    Adjusted R-squared:  0.9998
F-statistic: 3.543e+04 on 6 and 38 DF,  p-value: < 2.2e-16
```

5.2 Interaction Variables

Use R on the expanded model

Similarly, R can automatically perform the interaction model without introducing the additional variables. We now use the *original* dataset to fit the interaction model on R automatically.

```
> ################    Interaction Model by R without creating new variables
> expd_mod2<-lm(S~.+E*M, data=Salary)
> summary(expd_mod2)

Call:
lm(formula = S ~ . + E * M, data = Salary)

Residuals:
    Min      1Q  Median      3Q     Max
-928.13  -46.21   24.33   65.88  204.89

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 11203.434     79.065 141.698  < 2e-16 ***
X             496.987      5.566  89.283  < 2e-16 ***
E1          -1730.748    105.334 -16.431  < 2e-16 ***
E2           -349.078     97.568  -3.578 0.000945 ***
M1           7047.412    102.589  68.695  < 2e-16 ***
E1:M1       -3066.035    149.330 -20.532  < 2e-16 ***
E2:M1        1836.488    131.167  14.001  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 173.8 on 39 degrees of freedom
Multiple R-squared:  0.9988,    Adjusted R-squared:  0.9986
F-statistic:  5517 on 6 and 39 DF,  p-value: < 2.2e-16
```

5.2 Interaction Variables

Use R on the expanded model

Similarly, R can automatically perform the interaction model without introducing the additional variables. We now use the *original* dataset to fit the interaction model on R automatically.

```
> ################    Interaction Model by R without creating new variables and deleted outlier
> expd_mod3<-lm(S~.+E*M, data=Salary[-33,])
> summary(expd_mod3)

Call:
lm(formula = S ~ . + E * M, data = Salary[-33, ])

Residuals:
     Min       1Q   Median       3Q      Max
-112.884  -43.636   -5.036   46.622  128.480

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 11199.714     30.533 366.802  < 2e-16 ***
X             498.418      2.152 231.640  < 2e-16 ***
E1          -1741.336     40.683 -42.803  < 2e-16 ***
E2           -357.042     37.681  -9.475 1.49e-11 ***
M1           7040.580     39.619 177.707  < 2e-16 ***
E1:M1       -3051.763     57.674 -52.914  < 2e-16 ***
E2:M1        1997.531     51.785  38.574  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 67.12 on 38 degrees of freedom
Multiple R-squared:  0.9998,    Adjusted R-squared:  0.9998
F-statistic: 3.543e+04 on 6 and 38 DF,  p-value: < 2.2e-16
```

5.3. Systems of Regression Equations

5.3 Systems of Regression Equations

*A **collection** of data may consist of **two** or **more** distinct subsets, each of which may require a **separate** regression equation. Serious bias may be incurred if one regression relationship is used to represent the **pooled** data set. An analysis of this problem can be accomplished using indicator variables. An analysis of separate regression equations for subsets of the data may be applied to **cross-sectional**. The example discussed below treats **cross-sectional**.*

The model for the two groups can be different in all aspects or in only some aspects. In this section we discuss three distinct cases:

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.

We illustrate these cases below when we have only one quantitative predictor variable. These ideas can be extended straightforwardly to the cases where there are more than one quantitative predictor variable.

5.3 Systems of Regression Equations

Case 1: Models with Different Slopes and Different intercepts

*We illustrate this case with an **important** problem concerning equal opportunity in employment. Many large corporations and government agencies administer a **preemployment** test in an attempt to screen job applicants. The test is supposed to measure an applicant's **aptitude** for the job and the results are used as part of the information for making a hiring decision. The federal government has ruled that these tests **(1)** must measure abilities that are directly related to the job under consideration and **(2)** must not discriminate on the basis of race or national origin. Operational definitions of requirements (1) and (2) are rather elusive. We shall not try to resolve these operational problems. We shall take one approach involving race represented as **two groups, white and minority**. The hypothesis that there are separate regressions relating test scores to job performance for the two groups will be examined. The implications of this hypothesis for discrimination in hiring are discussed.*

Let Y represent job performance and let X be the score on the preemployment test. We want to compare

$$\begin{aligned}
&\text{Model 1 (Pooled):} && y_{ij} = \beta_0 + \beta_1 x_{ij} + \varepsilon_{ij}, \quad j = 1, 2; \quad i = 1, 2, \cdots, n_j, \\
&\text{Model 2 (Minority):} && y_{i1} = \beta_{01} + \beta_{11} x_{i1} + \varepsilon_{i1}, \qquad (5.3) \\
&\text{Model 2 (White):} && y_{i2} = \beta_{02} + \beta_{12} x_{i2} + \varepsilon_{i2}.
\end{aligned}$$

5.3 Systems of Regression Equations

Case 1: Models with Different Slopes and Different intercepts

*Figure 5.6 depicts the two models. In **model 1**, race distinction is ignored, the data are **pooled**, and there is **one** regression line. In **model 2** there is a separate regression relationship for the two subgroups, each with **distinct** regression coefficients. We shall assume that the variances of the residual terms are the **same** in **each** subgroup.*

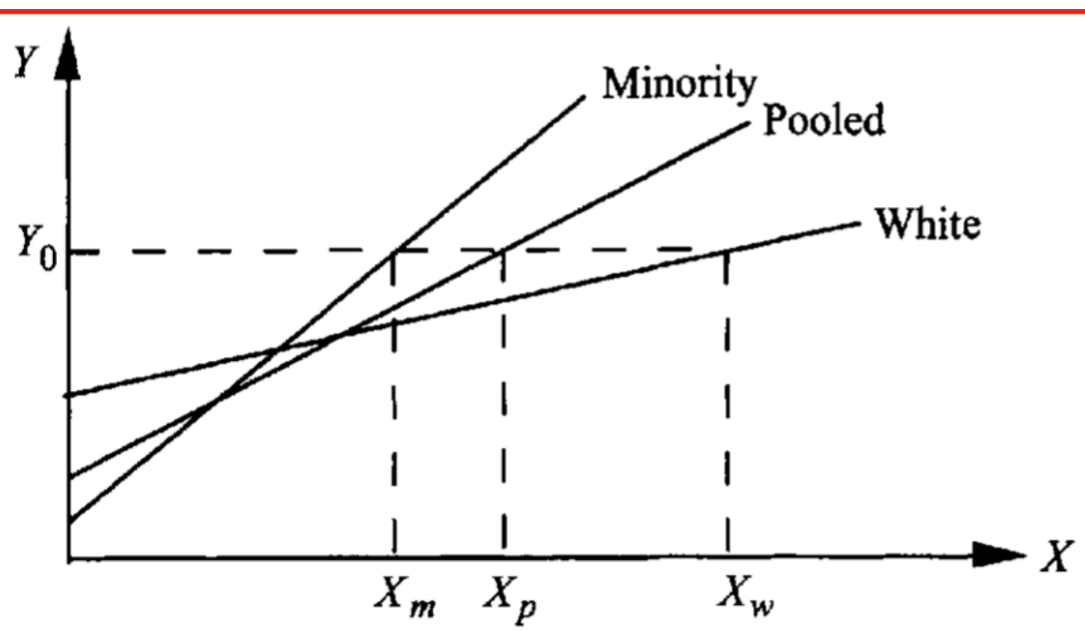


Figure 5.6 Requirements for employment on pretest.

5.3 Systems of Regression Equations

Case 1: Models with Different Slopes and Different intercepts

Before analyzing the data, let us briefly consider the types of errors that could be present in interpreting and applying the results. If Y_0, as seen on the graph, has been set as the minimum required level of performance, then using Model 1, an acceptable score on the test is one that exceeds X_p. However, if Model 2 is in fact correct, the appropriate test score for whites is X_w and for minorities is X_m. Using X_p in place of X_m and X_w represents a relaxation of the pretest requirement for whites and a tightening of that requirement for minorities. Since inequity can result in the selection procedure if the wrong model is used to set cutoff values, it is necessary to examine the data carefully. It must be determined whether there are two distinct relationships or whether the relationship is the same for both groups and a single equation estimated from the pooled data is adequate. Note that whether Model 1 or Model 2 is chosen, the values X_m, X_w, and X_p are estimates subject to sampling errors and should only be used in conjunction with appropriate confidence intervals. (Construction of confidence intervals is discussed in the following paragraphs.)

Figure 5.6

5.3 Systems of Regression Equations

Case 1: Models with Different Slopes and Different intercepts

Data were collected for this analysis using a special employment program. ***Twenty*** *applicants were hired on a trial basis for six weeks. One week was spent in a training class. The remaining five weeks were spent on the job. The participants were selected from a pool of applicants by a method that was not related to the* ***preemployment*** *test scores.* ***A test*** *was given at the end of the training period and a* ***work performance evaluation*** *was developed at the end of the six-week period. These* ***two*** *scores were combined to form an index of job performance. (Those employees with unsatisfactory performance at the end of the six-week period were dropped.) The data appear in Table 5. We refer to this data set as the* ***Preemployment Testing Data.***

Table 5.7 Data on Preemployment Testing Program

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

5.3 Systems of Regression Equations

Case 1: Models with Different Slopes and Different intercepts

Formally, we want to test the null hypothesis $H_0 : \beta_{11} = \beta_{12}, \beta_{01} = \beta_{02}$ against the alternative that there are substantial differences in these parameters. The test can be performed using indicator variables. Let z_{ij} be defined to take the value 1 if $j = 1$ and to take the value 0 if $j = 2$. That is, Z is a new variable that has the value 1 for a minority applicant and the value 0 for a white applicant. We consider the two models

$$\text{Model 1:} \quad y_{ij} = \beta_0 + \beta_1 x_{ij} + \varepsilon_{ij}$$

$$\text{Model 3:} \quad y_{ij} = \beta_0 + \beta_1 x_{ij} + \gamma z_{ij} + \delta(z_{ij} \cdot x_{ij}) + \varepsilon_{ij}. \qquad (5.4)$$

The variable $(z_{ij} \cdot x_{ij})$ represents the interaction between the group (race) variable Z and the preemployment test X.

5.3 Systems of Regression Equations

Case 1: Models with Different Slopes and Different intercepts

Note that Model 3 is equivalent to Model 2. This can be seen if we observe that for the minority group, $x_{ij} = x_{i1}$ and $z_{ij} = 1$; hence Model 3 becomes

$$\begin{aligned} y_{i1} &= \beta_0 + \beta_1 x_{i1} + \gamma + \delta x_{i1} + \varepsilon_{i1} \\ &= (\beta_0 + \gamma) + (\beta_1 + \delta) x_{i1} + \varepsilon_{i1} \\ &= \beta_{01} + \beta_{11} x_{i1} + \varepsilon_{i1}, \end{aligned}$$

which is the same as Model 2 for minority with $\beta_{01} = \beta_0 + \gamma$ and $\beta_{11} = \beta_1 + \delta$. Similarly, for the white group, we have $x_{ij} = x_{i2}$, $z_{ij} = 0$, and Model 3 becomes

$$y_{i2} = \beta_0 + \beta_1 x_{i2} + \varepsilon_{i2},$$

which is the same as Model 2 for white with $\beta_{02} = \beta_0$ and $\beta_{12} = \beta_1$. Therefore, a comparison between Models 1 and 2 is equivalent to a comparison between Models 1 and 3. Note that Model 3 can be viewed as a full model (FM) and Model 1 as a restricted model (RM) because Model 1 is obtained from Model 3 by setting $\gamma = \delta = 0$. Thus, our null hypothesis H_0 now becomes $H_0 : \gamma = \delta = 0$. The hypothesis is tested by constructing an F-Test for the comparison of two models as described in Chapter 3. In this case, the test statistics is

$p = 3, k = 2, n = 20$ (in the framework of Chapter 3)

$$F = \frac{[\text{SSE(RM)} - \text{SSE(FM)}]/2}{\text{SSE(FM)}/16},$$

which has 2 and 16 degrees of freedom. (Why?) Proceeding with the analysis of the data, the regression results for Model 1 and Model 3 are given in Tables 5.8 and 5.9. The plots of residuals against the predictor variable (Figures 5.7 and 5.8) look acceptable in both cases. The one residual at the lower right in Model 1 may require further investigation.

Table 5.8, 5.9
Figure 5.7, 5.8

Qualitative Predictors

5.3 Systems of Regression Equations

Case 1: Models with Different Slopes and Different intercepts

Table 5.8 Regression Results, Preemployment Testing Data: Model 1

Variable	Coefficient	s.e.	t-Test	p-value
Constant	1.03	0.87	1.19	0.2486
TEST (X)	2.36	0.54	4.39	0.0004
$n = 20$	$R^2 = 0.52$	$R_a^2 = 0.49$	$\hat{\sigma} = 1.59$	df = 18

Table 5.9 Regression Results, Preemployment Testing Data: Model 3

Variable	Coefficient	s.e.	t-Test	p-value
Constant	2.01	1.05	1.91	0.0736
TEST (X)	1.31	0.67	1.96	0.0677
RACE (Z)	−1.91	1.54	−1.24	0.2321
RACE · TEST ($X \cdot Z$)	2.00	0.95	2.09	0.0527
$n = 20$	$R^2 = 0.664$	$R_a^2 = 0.601$	$\hat{\sigma} = 1.41$	df = 16

Figure 5.7, 5.8

5.3 Systems of Regression Equations

Case 1: Models with Different Slopes and Different intercepts

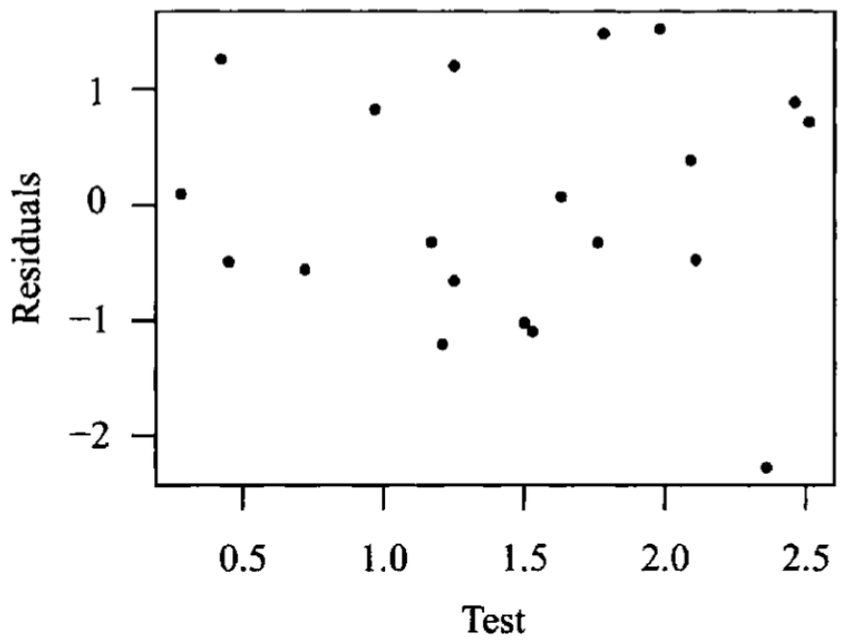


Figure 5.7 Standardized residuals versus test score: Model 1.

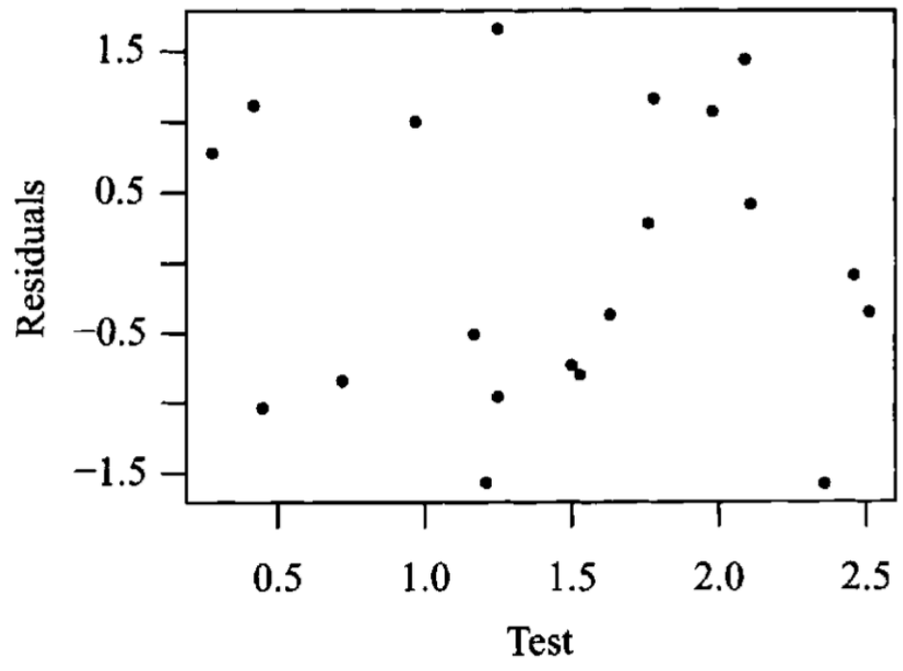


Figure 5.8 Standardized residuals versus test score: Model 3.

5.3 Systems of Regression Equations

Case 1: Models with Different Slopes and Different intercepts

To evaluate the formal hypothesis we compute the F-ratio specified previously, which is equal to

$$F = \frac{(45.51 - 31.81)/2}{31.81/16} = 3.4$$

and is significant at a level slightly above 5%. Therefore, on the basis of this test we would conclude that the relationship is probably different for the two groups. Specifically, for minorities we have

$$Y_1 = 0.10 + 3.31X_1$$

and for whites we have

$$Y_2 = 2.01 + 1.32X_2.$$

The results are very similar to those that were described in Figure 5.5 when the problem of bias was discussed. The straight line representing the relationship for minorities has a larger slope and a smaller intercept than the line for whites. If a pooled model were used, the types of biases discussed in relation to Figure 5.6 would occur.

5.3 Systems of Regression Equations

Case 1: Models with Different Slopes and Different intercepts

Equal Variance Assumption

Although the formal procedure using indicator variables has led to the plausible conclusion that the relationships are different for the two groups, the data for the individual groups have not been looked at carefully. Recall that it was assumed that the variances were identical in the two groups. This assumption was required so that the only distinguishing characteristic between the two samples was the pair of regression coefficients. In Figure 5.9 a plot of residuals versus the indicator variable is presented. There does not appear to be a difference between the two sets of residuals. We shall now look more closely at each group. The regression coefficients for each sample taken separately are presented in Table 5.10. The residuals are shown in Figures 5.10 and 5.11. The regression coefficients are, of course, the values obtained from Model 3. The standard errors of the residuals are 1.29 and 1.51 for the minority and white samples, respectively. The residual plots against the test score are acceptable in both cases. An interesting observation that was not available in the earlier analysis is that the preemployment test accounts for a major portion of the variation in the minority sample, but the test is only marginally useful in the white sample.

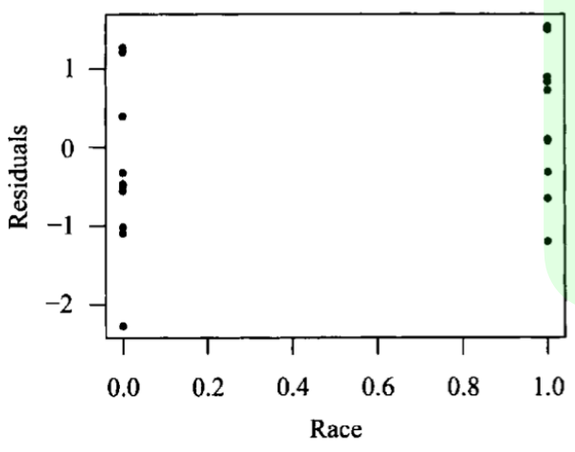


Figure 5.9 Standardized residuals versus race: Model 1.

Figure 5.10, 5.11
Table 5.10

5.3 Systems of Regression Equations

Case 1: Models with Different Slopes and Different intercepts

Equal Variance Assumption

Table 5.10 Separate Regression Results

Sample	$\hat{\beta}_0$	$\hat{\beta}_1$	t_1	R^2	$\hat{\sigma}$	df
Minority	0.10	3.31	5.31	0.78	1.29	8
White	2.01	1.31	1.82	0.29	1.51	8

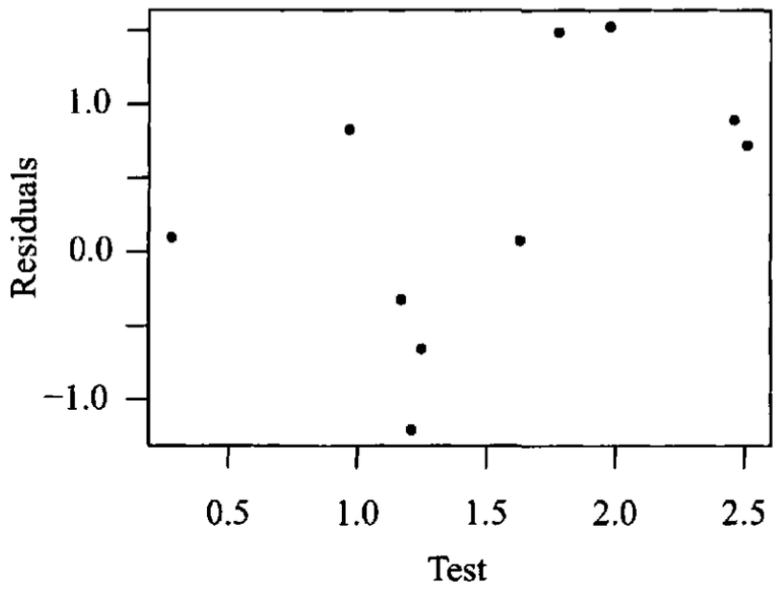


Figure 5.10 Standardized residuals versus test: Model 1, minority only.

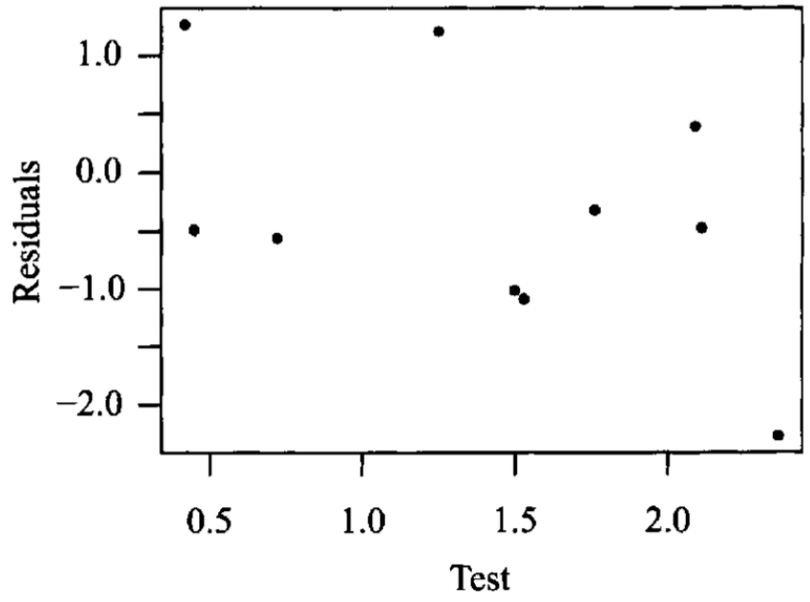


Figure 5.11 Standardized residuals versus test: Model 1, white only.

5.3 Systems of Regression Equations

Case 1: Models with Different Slopes and Different intercepts

Equal Variance Assumption

Our previous conclusion is still valid. The two regression equations are different. Not only are the regression coefficients different, but the residual mean squares also show slight differences. Of more importance, the values of R^2 are greatly different. For the white sample, $R^2 = 0.29$ is so small ($t = 1.82; 2.306$ is required for significance) that the preemployment test score is not deemed an adequate predictor of job success. This finding has bearing on our original objective since it should be a prerequisite for comparing regressions in two samples that the relationships be valid in each of the samples when taken alone. Concerning the validity of the preemployment test, we conclude that if applied as the law prescribes, with indifference to race, it will give biased results for both racial groups. Moreover, based on these findings, we may be justified in saying that the test is of no value for screening white applicants.

5.3 Systems of Regression Equations

Case 1: Models with Different Slopes and Different intercepts

Use R to obtain the regression results

```
> ############################   Preemployment Testing Data -- Case 1
> Preemp<-read.table('data/P140.txt',header=TRUE)   ## read the data
> Preemp$RACE<- as.factor(Preemp$RACE)      ## as this variable to a factor variable
> model1<-lm(JPERF~TEST,data=Preemp)        ## Mode 1: y = beta0 + beta1 x + epsilon
> summary(model1)

Call:
lm(formula = JPERF ~ TEST, data = Preemp)

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
```

```
> Preemp
   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
```

```
> model2<-lm(JPERF~TEST+RACE+TEST*RACE,data=Preemp)    ## Model 3: y = beta0 +beta1 x+ gamma z + delta (z*x) + epsilon
> summary(model2)

Call:
lm(formula = JPERF ~ TEST + RACE + TEST * RACE, data = Preemp)

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 .
RACE1        -1.9132     1.5403  -1.242   0.2321
TEST:RACE1    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
```

5.3 Systems of Regression Equations

Case 1: Models with Different Slopes and Different intercepts

Use R to obtain the regression results

```
> ######### obtain F value by direct calculation
> n<-20
> p<-3
> k<-2
> SSEfm<-sum((summary(model2)$residual)^2)
> SSErm<-sum((summary(model1)$residual)^2)
> Fval<-(SSErm-SSEfm)/SSEfm*16/2
> Fval
[1] 3.516061
```

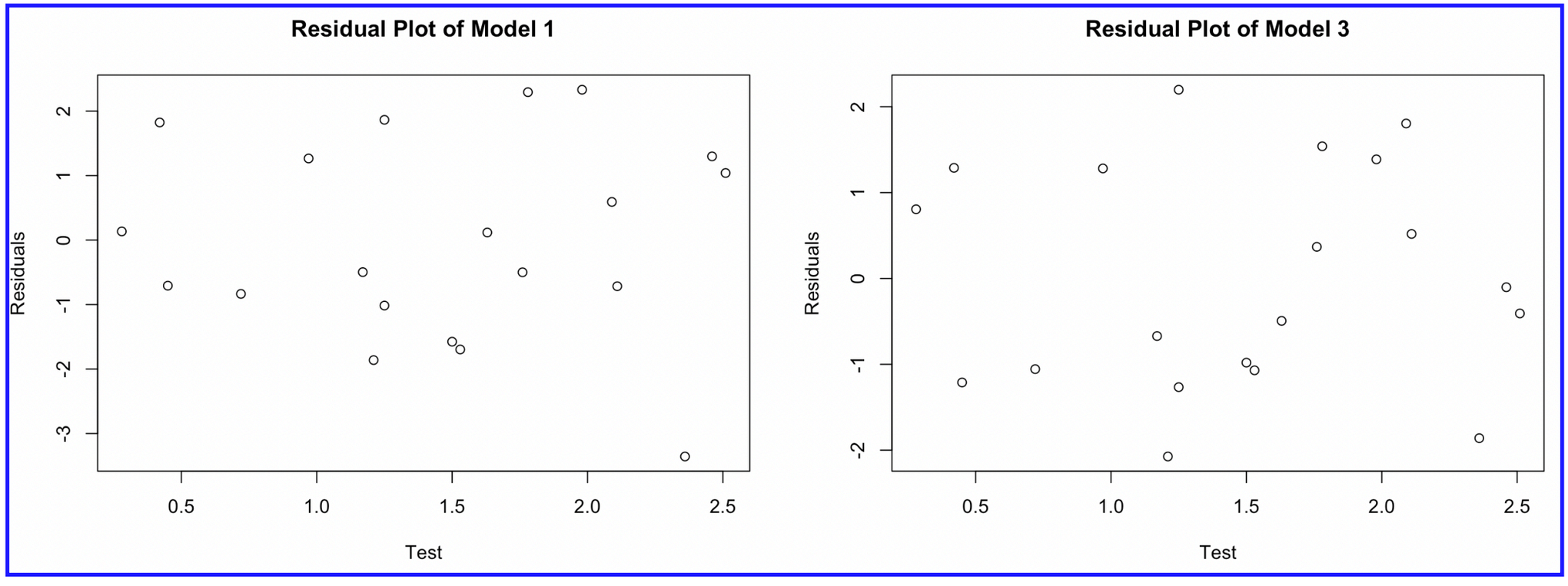
```
> ######### obtain F value and model testing by anova
> anova(model1,model2)
Analysis of Variance Table

Model 1: JPERF ~ TEST
Model 2: JPERF ~ TEST + RACE + TEST * RACE
  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 ‘ ’ 1
```

5.3 Systems of Regression Equations

Case 1: Models with Different Slopes and Different intercepts

Use R to obtain the regression results

```
######## Residual plot
par(mfrow=c(1,2))
plot(Preemp$TEST,summary(model1)$residual,xlab="Test",ylab="Residuals",main="Residual Plot of Model 1")
plot(Preemp$TEST,summary(model2)$residual,xlab="Test",ylab="Residuals",main="Residual Plot of Model 3")
```



5.3 Systems of Regression Equations

Case 1: Models with Different Slopes and Different intercepts

Regression on Separate dataset by Race

```
> ########## Regression on Minority
> model_mnt<-lm(JPERF~TEST,data=Preemp[1:10,])
> summary(model_mnt)

Call:
lm(formula = JPERF ~ TEST, data = Preemp[1:10, ])

Residuals:
    Min      1Q  Median      3Q     Max
-2.0734 -0.6267 -0.2548  1.1624  1.5394

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.09712    1.03519   0.094 0.927564
TEST         3.31095    0.62411   5.305 0.000724 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.292 on 8 degrees of freedom
Multiple R-squared:  0.7787,    Adjusted R-squared:  0.751
F-statistic: 28.14 on 1 and 8 DF,  p-value: 0.0007239
```

```
> ########## Regression on White Race
> model_white<-lm(JPERF~TEST,data=Preemp[11:20,])
> summary(model_white)

Call:
lm(formula = JPERF ~ TEST, data = Preemp[11:20, ])

Residuals:
    Min      1Q  Median      3Q     Max
-1.8599 -1.0663 -0.3061  1.0957  2.1980

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.0103     1.1291   1.780    0.113
TEST          1.3134     0.7208   1.822    0.106

Residual standard error: 1.512 on 8 degrees of freedom
Multiple R-squared:  0.2933,    Adjusted R-squared:  0.205
F-statistic:  3.32 on 1 and 8 DF,  p-value: 0.1059
```

5.3 Systems of Regression Equations

Case 2: Models with Same Slope and Different intercepts

In the previous subsection we dealt with the case where the two groups have distinct models with different sets of coefficients as given by Models 1 and 2 in (5.3) and as depicted in Figure 5.6. Suppose now that there is a reason to believe that the two groups have the same slope, β_1, and we wish to test the hypothesis that the two groups also have the same intercept, that is, $H_0 : \beta_{01} = \beta_{02}$. In this case we compare

$$\begin{aligned} &\text{Model 1 (Pooled):} && y_{ij} = \beta_0 + \beta_1 x_{ij} + \varepsilon_{ij}, \quad j = 1, 2; \quad i = 1, 2, \cdots, n_j, \\ &\text{Model 2 (Minority):} && y_{i1} = \beta_{01} + \beta_1 x_{i1} + \varepsilon_{i1}, \\ &\text{Model 2 (White):} && y_{i2} = \beta_{02} + \beta_1 x_{i2} + \varepsilon_{i2}. \end{aligned} \tag{5.5}$$

Notice that the two models have the same value of the slope β_1 but different values of the intercepts β_{01} and β_{02}. Using the indicator variable Z defined earlier, we can write Model 2 as

$$\text{Model 3:} \quad y_{ij} = \beta_0 + \beta_1 x_{ij} + \gamma z_{ij} + \varepsilon_{ij}. \tag{5.6}$$

Note the absence of the interaction variable $(z_{ij} \cdot x_{ij})$ from Model 3 in (5.6). If it is present, as it is in (5.4), the two groups would have two models with different slopes and different intercepts.

5.3 Systems of Regression Equations

Case 2: Models with Same Slope and Different intercepts

The equivalence of Models 2 and 3 can be seen by noting that for the minority group, where $x_{ij} = x_{i1}$ and $z_{ij} = 1$, Model 3 becomes

$$\begin{aligned} y_{i1} &= \beta_0 + \beta_1 x_{i1} + \gamma + \varepsilon_{i1} \\ &= (\beta_0 + \gamma) + \beta_1 x_{i1} + \varepsilon_{i1} \\ &= \beta_{01} + \beta_1 x_{i1} + \varepsilon_{i1}, \end{aligned}$$

which is the same as Model 2 for minority with $\beta_{01} = \beta_0 + \gamma$. Similarly, Model 3 for the white group becomes

$$y_{i2} = \beta_0 + \beta_1 x_{i2} + \varepsilon_{i2}.$$

Thus, Model 2 (or equivalently, Model 3) represents two parallel lines[4] (same slope) with intercepts $\beta_0 + \gamma$ and β_0. Therefore, our null hypothesis implies a restriction on γ in Model 3, namely, $H_0 : \gamma = 0$. To test this hypothesis, we use the F-Test

$$F = \frac{[\text{SSE(RM)} - \text{SSE(FM)}]/1}{\text{SSE(FM)}/17},$$

$p = 2, k = 2, n = 20$ (in the framework of Chapter 3)

which has 1 and 17 degrees of freedom. Equivalently, we can use the t-Test for testing $\gamma = 0$ in Model 3, which is

$$t = \frac{\hat{\gamma}}{\text{s.e.}(\hat{\gamma})},$$

which has 17 degrees of freedom. Again, the validation of the assumptions of Model 3 should be done before any conclusions are drawn from these tests. For the current example, we leave the computations of the above tests and the conclusions based on them, as an exercise

5.3 Systems of Regression Equations

Case 2: Models with Same Slope and Different intercepts

Use R to obtain the regression results

```
> ##############################   Preemployment Testing Data -- Case 2
> Preemp<-read.table('data/P140.txt',header=TRUE)   ## read the data
> Preemp$RACE<- as.factor(Preemp$RACE)      ## as this variable to a factor variable
> model1<-lm(JPERF~TEST,data=Preemp)        ## Mode 1: y = beta0 + beta1 x + epsilon
> summary(model1)

Call:
lm(formula = JPERF ~ TEST, data = Preemp)

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
```

```
> model2<-lm(JPERF~TEST+RACE,data=Preemp)    ## Model 3: y = beta0 +beta1 x+ gamma z + epsilon
> summary(model2)

Call:
lm(formula = JPERF ~ TEST + RACE, data = Preemp)

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 ***
RACE1         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
```

5.3 Systems of Regression Equations

Case 2: Models with Same Slope and Different intercepts

Use R to obtain the regression results

```
> ######### obtain F value by direct calculation
> n<-20
> p<-2
> k<-2
> SSEfm<-sum((summary(model2)$residual)^2)
> SSErm<-sum((summary(model1)$residual)^2)
> Fval<-(SSErm-SSEfm)/SSEfm*(n-p-1)/(p+1-k)
> Fval
[1] 2.212087
```

```
> ######### obtain F value and model testing by anova
> anova(model1,model2)
Analysis of Variance Table

Model 1: JPERF ~ TEST
Model 2: JPERF ~ TEST + RACE
  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
```

5.3 Systems of Regression Equations

Case 3: Models with Same Intercept and Different Slopes

Now we deal with the third case where the two groups have the same intercept, β_0, and we wish to test the hypothesis that the two groups also have the same slope, that is, $H_0 : \beta_{11} = \beta_{12}$. In this case we compare

$$\begin{aligned} &\text{Model 1 (Pooled):} && y_{ij} = \beta_0 + \beta_1\ x_{ij} + \varepsilon_{ij},\quad j = 1, 2;\quad i = 1, 2, \cdots, n_j, \\ &\text{Model 2 (Minority):} && y_{i1} = \beta_0 + \beta_{11} x_{i1} + \varepsilon_{i1}, \qquad (5.7) \\ &\text{Model 2 (White):} && y_{i2} = \beta_0 + \beta_{12} x_{i2} + \varepsilon_{i2}. \end{aligned}$$

Note that the two models have the same value of the intercept β_0 but different values of the slopes β_{11} and β_{12}. Using the indicator variable Z defined earlier, we can write Model 2 as

$$\text{Model 3:} \quad y_{ij} = \beta_0 + \beta_1 x_{ij} + \delta(z_{ij} \cdot x_{ij}) + \varepsilon_{ij}. \qquad (5.8)$$

5.3 Systems of Regression Equations

Case 3: Models with Same Intercept and Different Slopes

Observe the presence of the interaction variable $(z_{ij} \cdot x_{ij})$ but the absence of the individual contribution of the variable Z. The equivalence of Models 2 and 3 can be seen by observing that for the minority group, where $x_{ij} = x_{i1}$ and $z_{ij} = 1$, Model 3 becomes

$$\begin{aligned} y_{i1} &= \beta_0 + \beta_1 x_{i1} + \delta x_{i1} + \varepsilon_{i1} \\ &= \beta_0 + (\beta_1 + \delta) x_{i1} + \varepsilon_{i1} \\ &= \beta_0 + \beta_{11} x_{i1} + \varepsilon_{i1}, \end{aligned}$$

which is the same as Model 2 for minority with $\beta_{11} = \beta_1 + \delta$. Similarly, Model 3 for the white group becomes

$$y_{i2} = \beta_0 + \beta_{12} x_{i2} + \varepsilon_{i2}.$$

Therefore, our null hypothesis implies a restriction on δ in Model 3, namely, $H_0 : \delta = 0$. To test this hypothesis, we use the F-Test

$$F = \frac{[\text{SSE(RM)} - \text{SSE(FM)}]/1}{\text{SSE(FM)}/17},$$

$p = 2, k = 2, n = 20$ (in the framework of Chapter 3)

which has 1 and 17 degrees of freedom. Equivalently, we can use the t-Test for testing $\delta = 0$ in Model 3, which is

$$t = \frac{\hat{\delta}}{\text{s.e.}(\hat{\delta})},$$

which has 17 degrees of freedom. Validation of the assumptions of Model 3, the computations of the above tests, and the conclusions based on them are left as an exercise

5.3 Systems of Regression Equations

Case 3: Models with Same Intercept and Different Slopes

Use R to obtain the regression results

```
> ##############################   Preemployment Testing Data -- Case 3
> Preemp<-read.table('data/P140.txt',header=TRUE)   ## read the data
> Preemp$RACE<- as.factor(Preemp$RACE)      ## as this variable to a factor variable
> model1<-lm(JPERF~TEST,data=Preemp)        ## Mode 1: y = beta0 + beta1 x + epsilon
> summary(model1)

Call:
lm(formula = JPERF ~ TEST, data = Preemp)

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
```

```
> model2<-lm(JPERF~TEST+RACE:TEST,data=Preemp)    ## Model 3: y = beta0 +beta1 x+ delta (z*x) + epsilon
> summary(model2)

Call:
lm(formula = JPERF ~ TEST + RACE:TEST, data = Preemp)

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:RACE1    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
```

5.3 Systems of Regression Equations

Case 3: Models with Same Intercept and Different Slopes

Use R to obtain the regression results

```
> ######### obtain F value by direct calculation
> n<-20
> p<-2
> k<-2
> SSEfm<-sum((summary(model2)$residual)^2)
> SSErm<-sum((summary(model1)$residual)^2)
> Fval<-(SSErm-SSEfm)/SSEfm*(n-p-1)/(p+1-k)
> Fval
[1] 5.319603
```

```
> ######### obtain F value and model testing by anova
> anova(model1,model2)
Analysis of Variance Table

Model 1: JPERF ~ TEST
Model 2: JPERF ~ TEST + RACE:TEST
  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 ‘ ’ 1
```

5.4. Other Applications of Indicative Variables

5.4 Other Applications of Indicative Variables

Applications of indicator variables such as those described in Section 5.4 can be extended to cover a variety of problems

Suppose, for example, that we wish to compare the means of $k \geq 2$ populations or groups. The techniques commonly used here is known as the *analysis of variance* (ANOVA). A random sample of size n_j is taken from the jth population, $j = 1, \cdots, k$. We have a total of $n = n_1 + \cdots + n_k$ observations on the response variable. Let y_{ij} be the ith response in the jth sample. Then y_{ij} can be modeled as

$$y_{ij} = \mu_0 + \mu_1 x_{i1} + \cdots + \mu_p x_{ip} + \varepsilon_{ij}. \tag{5.9}$$

5.4 Other Applications of Indicative Variables

In this model there are $p = k - 1$ indicator predictor variables $x_{i1}, \cdots, x_{ip}$. Each variable x_{ij} is 1 if the corresponding response is from population j, and zero otherwise. The population that is left out is usually known as the *control* group. All indicator variables for the control group are equal to zero. Thus, for the control group, (5.9) becomes

$$y_{ij} = \mu_0 + \varepsilon_{ij}. \tag{5.10}$$

In both (5.9) and (5.10), ε_{ij} are random errors assumed to be independent normal variables with zero means and constant variance σ^2. The constant μ_0 represents the mean of the control group and the regression coefficient μ_j can be interpreted as the difference between the means of the control and jth groups. If $\mu_j = 0$, then the means of the control and jth groups are equal. The null hypothesis $\mathrm{H}_0 : \mu_1 = \cdots = \mu_p = 0$ that all groups have the same mean can be represented by the model in (5.10). The alternate hypothesis that at least one of the μ_j's is different from zero can be represented by the model in (5.9). The models in (5.9) and (5.10) can be viewed as full and reduced models, respectively. Hence H_0 can be tested using the F-Test given in Chapter 3. Thus, the use of indicator variables allowed us to express ANOVA techniques as a special case of regression analysis. Both the number of quantitative predictor variables and the number of distinct groups represented in the data by indicator variables may be increased.

5.4 Other Applications of Indicative Variables

ANOVA by multiple Linear Regression

Data

Group 1: $y_{11}, y_{12}, y_{13}, \cdots, y_{1n_1}$ → sampled from a population with mean μ_1

Group 2: $y_{21}, y_{22}, y_{23}, \cdots, y_{2n_2}$ → sampled from a population with mean μ_2

$\vdots$

Group k: $y_{k1}, y_{k2}, y_{k3}, \cdots, y_{kn_k}$ → sampled from a population with mean μ_k

ANOVA

$$H_0 : \mu_1 = \cdots = \mu_k \quad \text{versus} \quad H_1 : \text{ not all of them are equal}$$

Formulate as Regression Problem

Combine all the data together, and create a categorical variable X

Y	X
y_{11}	1
y_{12}	1
$\vdots$	$\vdots$
y_{1n_1}	1

Y	X
y_{21}	2
y_{22}	2
$\vdots$	$\vdots$
y_{2n_2}	2

$\cdots$

Y	X
y_{k1}	k
y_{k2}	k
$\vdots$	$\vdots$
y_{kn_k}	k

5.4 Other Applications of Indicative Variables

ANOVA by multiple Linear Regression

Since X is categorical with k categories, we create $k-1$ indicator variables $X_1, \cdots, X_{k-1}$ and treat the k-th group as the control group. If the observation comes from j-th group, then $X_j = 1$ and all other X's are zero. The data now reads as follows.

Y	X_1	X_2	$\cdots$	X_{k-1}
y_{11}	1	0	$\cdots$	0
y_{12}	1	0	$\cdots$	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
y_{1n_1}	1	0	$\cdots$	0
y_{21}	0	1	$\cdots$	0
y_{22}	0	1	$\cdots$	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
y_{2n_2}	0	1	$\cdots$	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$y_{(k-1)1}$	0	0	$\cdots$	1
$y_{(k-1)2}$	0	0	$\cdots$	1
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$y_{(k-1)n_{k-1}}$	0	0	$\cdots$	1
y_{k1}	0	0	$\cdots$	0
y_{k2}	0	0	$\cdots$	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
y_{kn_k}	0	0	$\cdots$	0

5.4 Other Applications of Indicative Variables

ANOVA by multiple Linear Regression

Then the ANOVA test is equivalent to test for the linear regression models

$$\text{Full model}: Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_{k-1} X_{k-1} + \varepsilon$$
$$\text{Reduced model}: Y = \beta_0 + \varepsilon$$

Therefore, ANOVA test of the multiple populations can be done as a special case of model testing in multiple linear regression.

5.5 Seasonality

5.5 Seasonality

The data set we use as an example here, referred to as the Ski Sales data, is shown in Table 5.11. The data consist of two variables: the sales, S, in millions for a firm that manufactures skis and related equipment for the year 1964-1973, and personal disposable income, PDI. Each of these variables is measure quarterly.

The model is an equation that relates S to PDI, that is, $S_t = \beta_0 + \beta_1 \text{PDI}_t + \varepsilon_t$, where S_t is sales in millions in the tth period and PDI_t is the corresponding personal disposable income. Our approach here is to assume the existence of a seasonal effect on sales that is determined on a quarterly basis. To measure this effect we may define indicator variables to characterize the seasonality. Since we have four quarters, we define three indicator variables, Z_1, Z_2, and Z_3, where

$$z_{t1} = \begin{cases} 1, & \text{if the } t\text{th period is a first quarter,} \\ 0, & \text{otherwise,} \end{cases}$$

$$z_{t2} = \begin{cases} 1, & \text{if the } t\text{th period is a second quarter,} \\ 0, & \text{otherwise,} \end{cases}$$

$$z_{t3} = \begin{cases} 1, & \text{if the } t\text{th period is a third quarter,} \\ 0, & \text{otherwise.} \end{cases}$$

Table 5.11

5.5 Seasonality

Table 5.11 Disposable Income and Ski Sales for Years 1964–1973

Row	Date	Sales	PDI	Row	Date	Sales	PDI
1	Q1/64	37.0	109	21	Q1/69	44.9	153
2	Q2/64	33.5	115	22	Q2/69	41.6	156
3	Q3/64	30.8	113	23	Q3/69	44.0	160
4	Q4/64	37.9	116	24	Q4/69	48.1	163
5	Q1/65	37.4	118	25	Q1/70	49.7	166
6	Q2/65	31.6	120	26	Q2/70	43.9	171
7	Q3/65	34.0	122	27	Q3/70	41.6	174
8	Q4/65	38.1	124	28	Q4/70	51.0	175
9	Q1/66	40.0	126	29	Q1/71	52.0	180
10	Q2/66	35.0	128	30	Q2/71	46.2	184
11	Q3/66	34.9	130	31	Q3/71	47.1	187
12	Q4/66	40.2	132	32	Q4/71	52.7	189
13	Q1/67	41.9	133	33	Q1/72	52.2	191
14	Q2/67	34.7	135	34	Q2/72	47.0	193
15	Q3/67	38.8	138	35	Q3/72	47.8	194
16	Q4/67	43.7	140	36	Q4/72	52.8	196
17	Q1/68	44.2	143	37	Q1/73	54.1	199
18	Q2/68	40.4	147	38	Q2/73	49.5	201
19	Q3/68	38.4	148	39	Q3/73	49.5	202
20	Q4/68	45.4	151	40	Q4/73	54.3	204

5.5 Seasonality

Reformulate the data with indicator variables

```
################################
################################ Seasonality
Ski_data<-read.table('data/P149.txt',header=TRUE)   ## read the data
n<-dim(Ski_data)[1]
Ski_data$Z1<-rep(0,n)
Ski_data$Z2<-rep(0,n)
Ski_data$Z3<-rep(0,n)
Ski_data$Z1[which(grepl("Q1", Ski_data$Date, fixed=TRUE))]<-1
Ski_data$Z2[which(grepl("Q2", Ski_data$Date, fixed=TRUE))]<-1
Ski_data$Z3[which(grepl("Q3", Ski_data$Date, fixed=TRUE))]<-1
```

```
> Ski_data
    Date Sales PDI Z1 Z2 Z3
1  Q1/64  37.0 109  1  0  0
2  Q2/64  33.5 115  0  1  0
3  Q3/64  30.8 113  0  0  1
4  Q4/64  37.9 116  0  0  0
5  Q1/65  37.4 118  1  0  0
6  Q2/65  31.6 120  0  1  0
7  Q3/65  34.0 122  0  0  1
8  Q4/65  38.1 124  0  0  0
9  Q1/66  40.0 126  1  0  0
10 Q2/66  35.0 128  0  1  0
11 Q3/66  34.9 130  0  0  1
12 Q4/66  40.2 132  0  0  0
13 Q1/67  41.9 133  1  0  0
14 Q2/67  34.7 135  0  1  0
15 Q3/67  38.8 138  0  0  1
16 Q4/67  43.7 140  0  0  0
17 Q1/68  44.2 143  1  0  0
18 Q2/68  40.4 147  0  1  0
19 Q3/68  38.4 148  0  0  1
20 Q4/68  45.4 151  0  0  0
21 Q1/69  44.9 153  1  0  0
22 Q2/69  41.6 156  0  1  0
23 Q3/69  44.0 160  0  0  1
24 Q4/69  48.1 163  0  0  0
25 Q1/70  49.7 166  1  0  0
26 Q2/70  43.9 171  0  1  0
27 Q3/70  41.6 174  0  0  1
28 Q4/70  51.0 175  0  0  0
29 Q1/71  52.0 180  1  0  0
30 Q2/71  46.2 184  0  1  0
31 Q3/71  47.1 187  0  0  1
32 Q4/71  52.7 189  0  0  0
33 Q1/72  52.2 191  1  0  0
34 Q2/72  47.0 193  0  1  0
35 Q3/72  47.8 194  0  0  1
36 Q4/72  52.8 196  0  0  0
37 Q1/73  54.1 199  1  0  0
38 Q2/73  49.5 201  0  1  0
39 Q3/73  49.5 202  0  0  1
40 Q4/73  54.3 204  0  0  0
```

5.5 Seasonality

Regression results on Ski dataset

```
> ############# Fit the regression model
> Ski_data<-Ski_data[,c("Sales","PDI","Z1","Z2","Z3")]
> ski_mod<-lm(Sales~.,data=Ski_data)
> summary(ski_mod)

Call:
lm(formula = Sales ~ ., data = Ski_data)

Residuals:
     Min       1Q   Median       3Q      Max
-2.51356 -0.86028  0.03654  0.67965  2.67306

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 14.77202    1.05058  14.061 5.78e-16 ***
PDI          0.19904    0.00619  32.155  < 2e-16 ***
Z1           0.35312    0.52150   0.677    0.503
Z2          -5.28382    0.52018 -10.158 5.62e-12 ***
Z3          -5.29210    0.51977 -10.182 5.28e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.162 on 35 degrees of freedom
Multiple R-squared:  0.9728,    Adjusted R-squared:  0.9697
F-statistic:   313 on 4 and 35 DF,  p-value: < 2.2e-16
```

Insignificant

So there are actually only two seasons.

5.6 An Example Using R

5.6 An Example Using R

For this chapter, we will briefly use the built in dataset `mtcars` before returning to our `autompg` dataset that we created in the last chapter. The `mtcars` dataset is somewhat smaller, so we'll quickly take a look at the entire dataset.

```
mtcars
```

```
##                    mpg cyl  disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4         21.0   6 160.0 110 3.90 2.620 16.46  0  1    4    4
## Mazda RX4 Wag     21.0   6 160.0 110 3.90 2.875 17.02  0  1    4    4
## Datsun 710        22.8   4 108.0  93 3.85 2.320 18.61  1  1    4    1
## Hornet 4 Drive    21.4   6 258.0 110 3.08 3.215 19.44  1  0    3    1
## Hornet Sportabout 18.7   8 360.0 175 3.15 3.440 17.02  0  0    3    2
## Valiant           18.1   6 225.0 105 2.76 3.460 20.22  1  0    3    1
## Duster 360        14.3   8 360.0 245 3.21 3.570 15.84  0  0    3    4
## Merc 240D         24.4   4 146.7  62 3.69 3.190 20.00  1  0    4    2
## Merc 230          22.8   4 140.8  95 3.92 3.150 22.90  1  0    4    2
## Merc 280          19.2   6 167.6 123 3.92 3.440 18.30  1  0    4    4
## Merc 280C         17.8   6 167.6 123 3.92 3.440 18.90  1  0    4    4
## Merc 450SE        16.4   8 275.8 180 3.07 4.070 17.40  0  0    3    3
## Merc 450SL        17.3   8 275.8 180 3.07 3.730 17.60  0  0    3    3
## Merc 450SLC       15.2   8 275.8 180 3.07 3.780 18.00  0  0    3    3
## Cadillac Fleetwood  10.4   8 472.0 205 2.93 5.250 17.98  0  0    3    4
## Lincoln Continental 10.4   8 460.0 215 3.00 5.424 17.82  0  0    3    4
## Chrysler Imperial   14.7   8 440.0 230 3.23 5.345 17.42  0  0    3    4
## Fiat 128          32.4   4  78.7  66 4.08 2.200 19.47  1  1    4    1
```

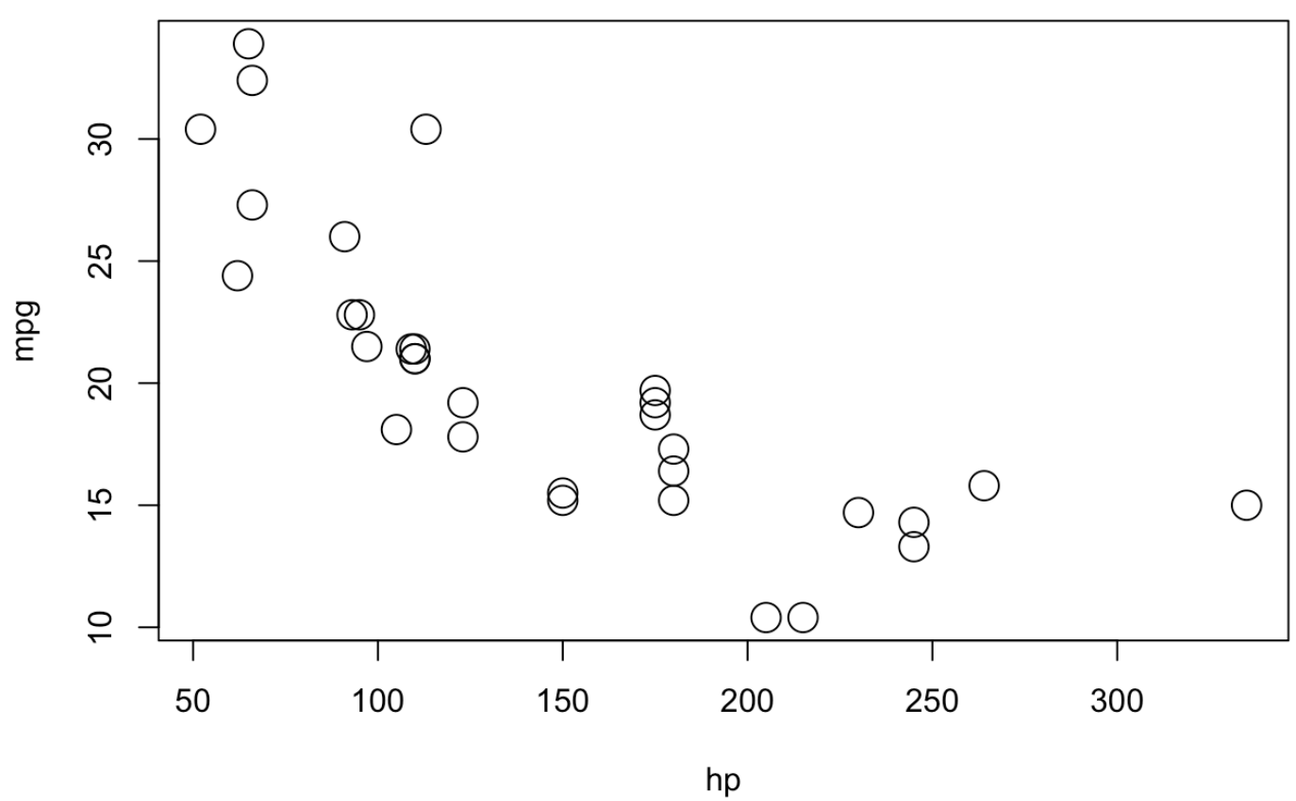
5.6 An Example Using R

We will be interested in three of the variables: `mpg`, `hp`, and `am`.

- `mpg`: fuel efficiency, in miles per gallon.
- `hp`: horsepower, in foot-pounds per second.
- `am`: transmission. Automatic or manual.

As we often do, we will start by plotting the data. We are interested in `mpg` as the response variable, and `hp` as a predictor.

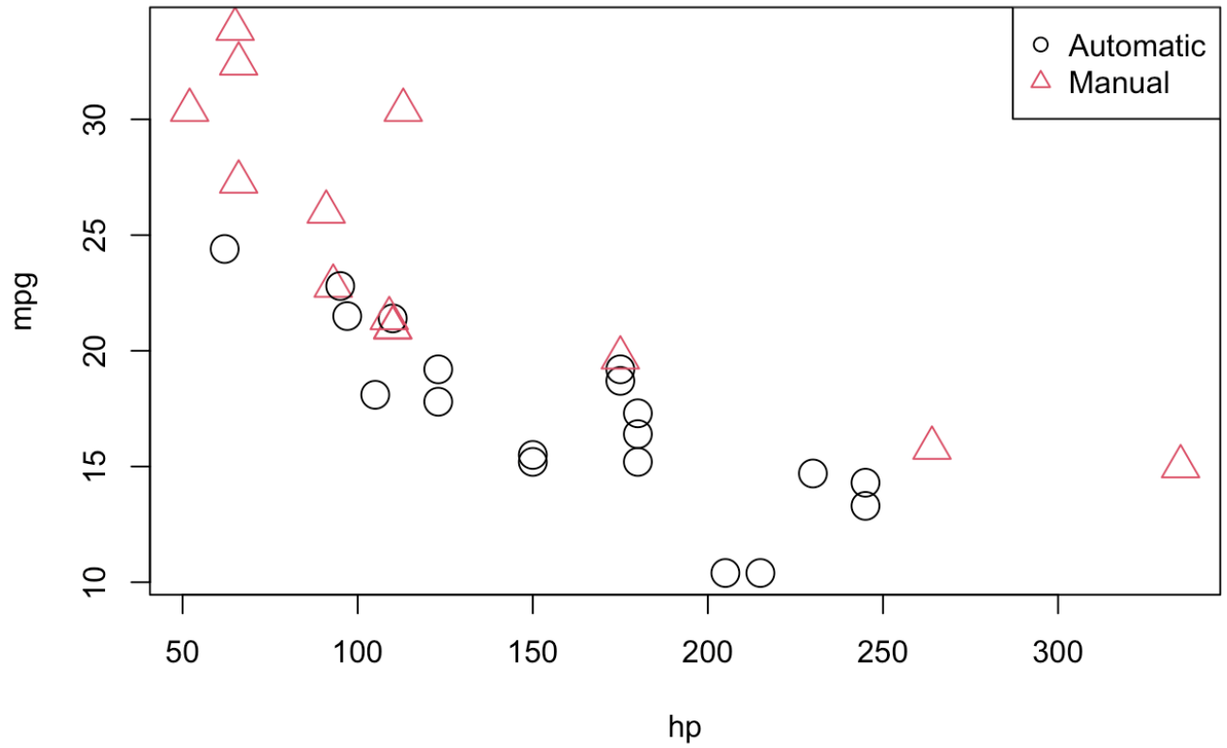
```
plot(mpg ~ hp, data = mtcars, cex = 2)
```



5.6 An Example Using R

Since we are also interested in the transmission type, we could also label the points accordingly.

```
plot(mpg ~ hp, data = mtcars, col = am + 1, pch = am + 1, cex = 2)
legend("topright", c("Automatic", "Manual"), col = c(1, 2), pch = c(1, 2))
```



We used a common `R` "trick" when plotting this data. The `am` variable takes two possible values; `0` for automatic transmission, and `1` for manual transmissions. `R` can use numbers to represent colors, however the color for `0` is white. So we take the `am` vector and add `1` to it. Then observations with automatic transmissions are now represented by `1`, which is black in `R`, and manual transmission are represented by `2`, which is red in `R`. (Note, we are only adding `1` inside the call to `plot()`, we are not actually modifying the values stored in `am`.)

Using only 1 predictor

5.6 An Example Using R

We now fit the SLR model

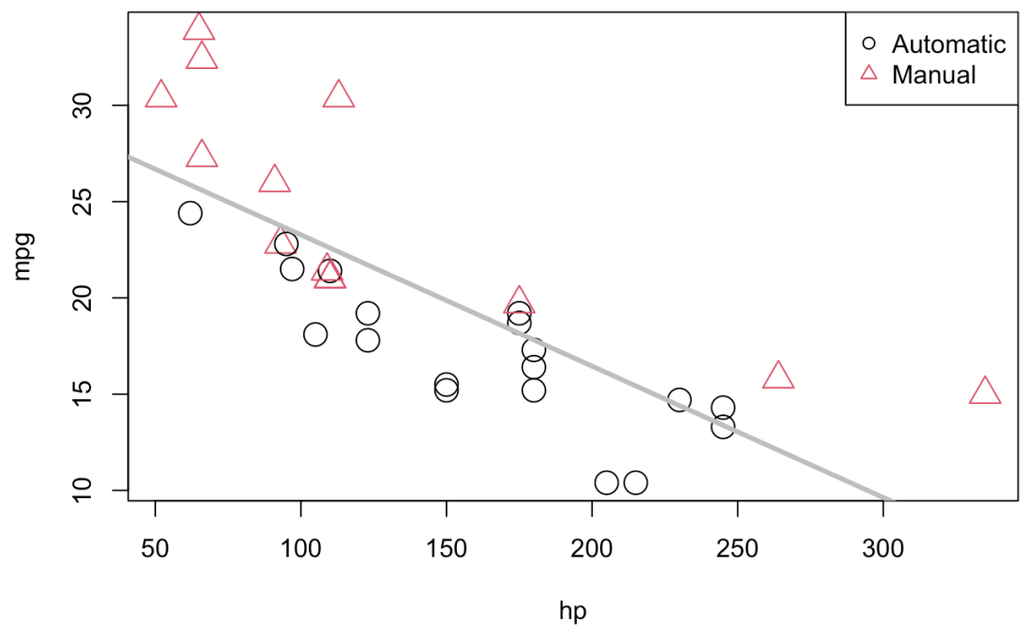
$$Y = \beta_0 + \beta_1 x_1 + \epsilon,$$

where Y is `mpg` and x_1 is `hp`. For notational brevity, we drop the index i for observations.

```
mpg_hp_slr = lm(mpg ~ hp, data = mtcars)
```

We then re-plot the data and add the fitted line to the plot.

```
plot(mpg ~ hp, data = mtcars, col = am + 1, pch = am + 1, cex = 2)
abline(mpg_hp_slr, lwd = 3, col = "grey")
legend("topright", c("Automatic", "Manual"), col = c(1, 2), pch = c(1, 2))
```



We should notice a pattern here. The red, manual observations largely fall above the line, while the black, automatic observations are mostly below the line. This means our model underestimates the fuel efficiency of manual transmissions, and overestimates the fuel efficiency of automatic transmissions. To correct for this, we will add a predictor to our model, namely, `am` as x_2.

Using 2 predictors

5.6 An Example Using R

Our new model is

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon,$$

where x_1 and Y remain the same, but now

$$x_2 = \begin{cases} 1 & \text{manual transmission} \\ 0 & \text{automatic transmission} \end{cases}.$$

In this case, we call x_2 a **dummy variable**. A dummy variable is somewhat unfortunately named, as it is in no way "dumb". In fact, it is actually somewhat clever. A dummy variable is a numerical variable that is used in a regression analysis to "code" for a binary categorical variable. Let's see how this works.

First, note that `am` is already a dummy variable, since it uses the values `0` and `1` to represent automatic and manual transmissions. Often, a variable like `am` would store the character values `auto` and `man` and we would either have to convert these to `0` and `1`, or, as we will see later, `R` will take care of creating dummy variables for us.

So, to fit the above model, we do so like any other multiple regression model we have seen before.

```
mpg_hp_add = lm(mpg ~ hp + am, data = mtcars)
```

Briefly checking the output, we see that `R` has estimated the three β parameters.

```
mpg_hp_add
```

```
##
## Call:
## lm(formula = mpg ~ hp + am, data = mtcars)
##
## Coefficients:
## (Intercept)           hp           am
##    26.58491     -0.05889      5.27709
```

Using 2 predictors

5.6 An Example Using R

Since x_2 can only take values `0` and `1`, we can effectively write two different models, one for manual and one for automatic transmissions.

For automatic transmissions, that is $x_2 = 0$, we have,

$$Y = \beta_0 + \beta_1 x_1 + \epsilon.$$

Then for manual transmissions, that is $x_2 = 1$, we have,

$$Y = (\beta_0 + \beta_2) + \beta_1 x_1 + \epsilon.$$

Notice that these models share the same slope, β_1, but have different intercepts, differing by β_2. So the change in `mpg` is the same for both models, but on average `mpg` differs by β_2 between the two transmission types.

We'll now calculate the estimated slope and intercept of these two models so that we can add them to a plot. Note that:

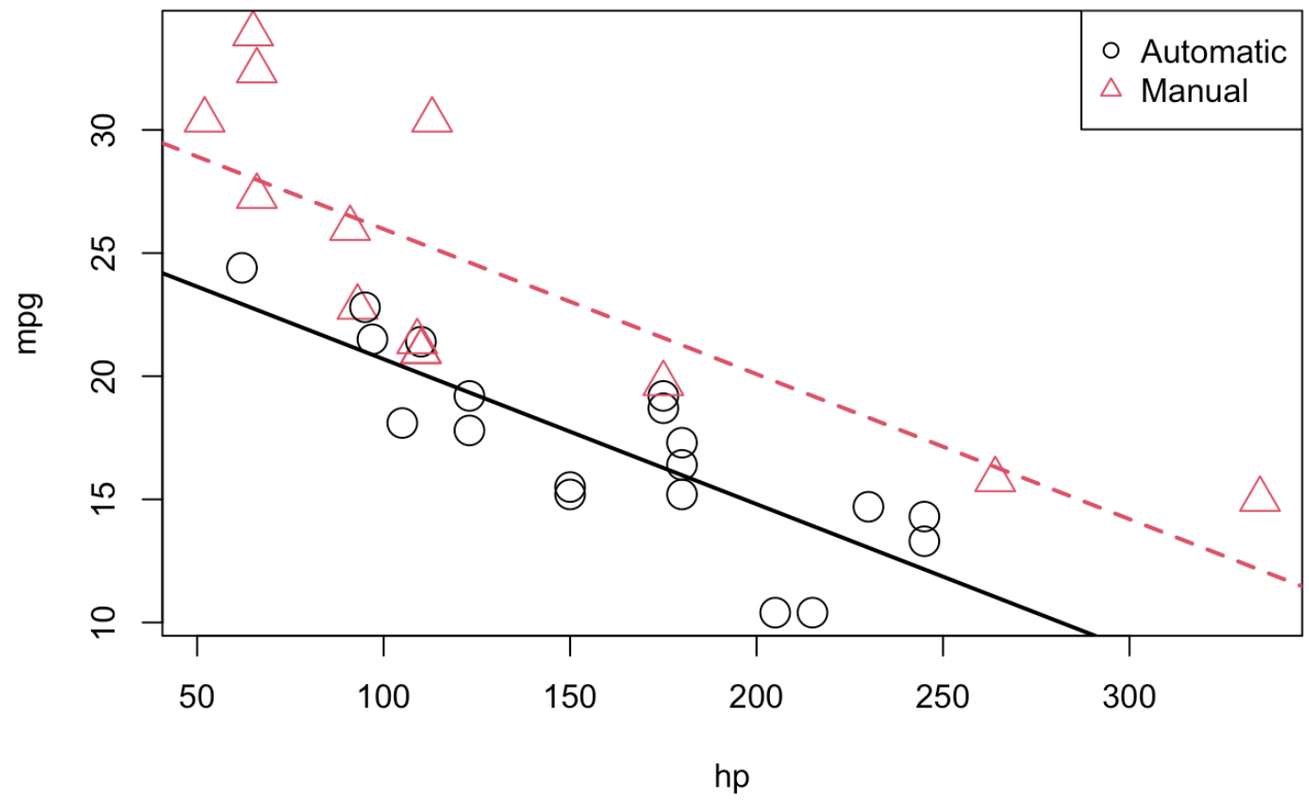
- $\hat{\beta}_0$ = `coef(mpg_hp_add)[1]` = 26.5849137
- $\hat{\beta}_1$ = `coef(mpg_hp_add)[2]` = -0.0588878
- $\hat{\beta}_2$ = `coef(mpg_hp_add)[3]` = 5.2770853

Using 2 predictors

5.6 An Example Using R

Re-plotting the data, we use these slopes and intercepts to add the "two" fitted models to the plot.

```
plot(mpg ~ hp, data = mtcars, col = am + 1, pch = am + 1, cex = 2)
abline(int_auto, slope_auto, col = 1, lty = 1, lwd = 2) # add line for auto
abline(int_manu, slope_manu, col = 2, lty = 2, lwd = 2) # add line for manual
legend("topright", c("Automatic", "Manual"), col = c(1, 2), pch = c(1, 2))
```



We notice right away that the points are no longer systematically incorrect. The red, manual observations vary about the red line in no particular pattern without underestimating the observations as before. The black, automatic points vary about the black line, also without an obvious pattern.

5.6 An Example Using R

They say a picture is worth a thousand words, but as a statistician, sometimes a picture is worth an entire analysis. The above picture makes it plainly obvious that β_2 is significant, but let's verify mathematically. Essentially we would like to test:

$$H_0 : \beta_2 = 0 \quad \text{vs} \quad H_1 : \beta_2 \neq 0.$$

This is nothing new. Again, the math is the same as the multiple regression analyses we have seen before. We could perform either a t or F test here. The only difference is a slight change in interpretation. We could think of this as testing a model with a single line (H_0) against a model that allows two lines (H_1).

To obtain the test statistic and p-value for the t-test, we would use

```
summary(mpg_hp_add)$coefficients["am",]
```

```
##     Estimate   Std. Error      t value     Pr(>|t|)
## 5.277085e+00 1.079541e+00 4.888270e+00 3.460318e-05
```

To do the same for the F test, we would use

```
anova(mpg_hp_slr, mpg_hp_add)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ hp
## Model 2: mpg ~ hp + am
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
## 1     30 447.67
## 2     29 245.44  1    202.24 23.895 3.46e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

5.6 An Example Using R

Notice that these are indeed testing the same thing, as the p-values are exactly equal. (And the F test statistic is the t test statistic squared.)

Recapping some interpretations:

- $\hat{\beta}_0 = 26.5849137$ is the estimated average `mpg` for a car with an automatic transmission and **0** `hp`.
- $\hat{\beta}_0 + \hat{\beta}_2 = 31.8619991$ is the estimated average `mpg` for a car with a manual transmission and **0** `hp`.
- $\hat{\beta}_2 = 5.2770853$ is the estimated **difference** in average `mpg` for cars with manual transmissions as compared to those with automatic transmission, for **any** `hp`.
- $\beta_1 = -0.0588878$ is the estimated change in average `mpg` for an increase in one `hp`, for **either** transmission types.

We should take special notice of those last two. In the model,

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon,$$

we see β_1 is the average change in Y for an increase in x_1, *no matter* the value of x_2. Also, β_2 is always the difference in the average of Y for *any* value of x_1. These are two restrictions we won't always want, so we need a way to specify a more flexible model.

Here we restricted ourselves to a single numerical predictor x_1 and one dummy variable x_2. However, the concept of a dummy variable can be used with larger multiple regression models. We only use a single numerical predictor here for ease of visualization since we can think of the "two lines" interpretation. But in general, we can think of a dummy variable as creating "two models," one for each category of a binary categorical variable.