Chapter 10

Categorical Predictors

Chanwoo Yoo, Division of Advanced Engineering,
Korea National Open University

Contents

1. Example on Birth Weight and Smoking
2. Two Separate Advantages
3. Coding Qualitative Variables
4. Additive Effects
5. Interaction Effects

1. Example on Birth Weight and Smoking

1. Binary Predictor

- A variable that takes on only two possible values
  - Gender (male, female)
  - Smoking status (smoker, nonsmoker)
  - Treatment (yes, no)
  - Health status (diseased, healthy)
  - Company status (private, public)

2. Data: Birth and Smokers

- Data: Birth and Smokers data
  - y_i (Wgt): birth weight of baby i
  - x_{i1} (Gest): length of gestation of baby i
  - x_{i2} (Smoke): binary variable coded as a 1, if the baby's mother smoked during pregnancy and 0, if she did not

3. Data Load

```
> birthsmokers <- read.table("birthsmokers.txt", header=T)
> attach(birthsmokers)
> head(birthsmokers)
   Wgt Gest Smoke
1 2940   38     1
2 3130   38     0
3 2420   36     1
4 2450   34     0
5 2760   39     1
6 2440   35     1
```

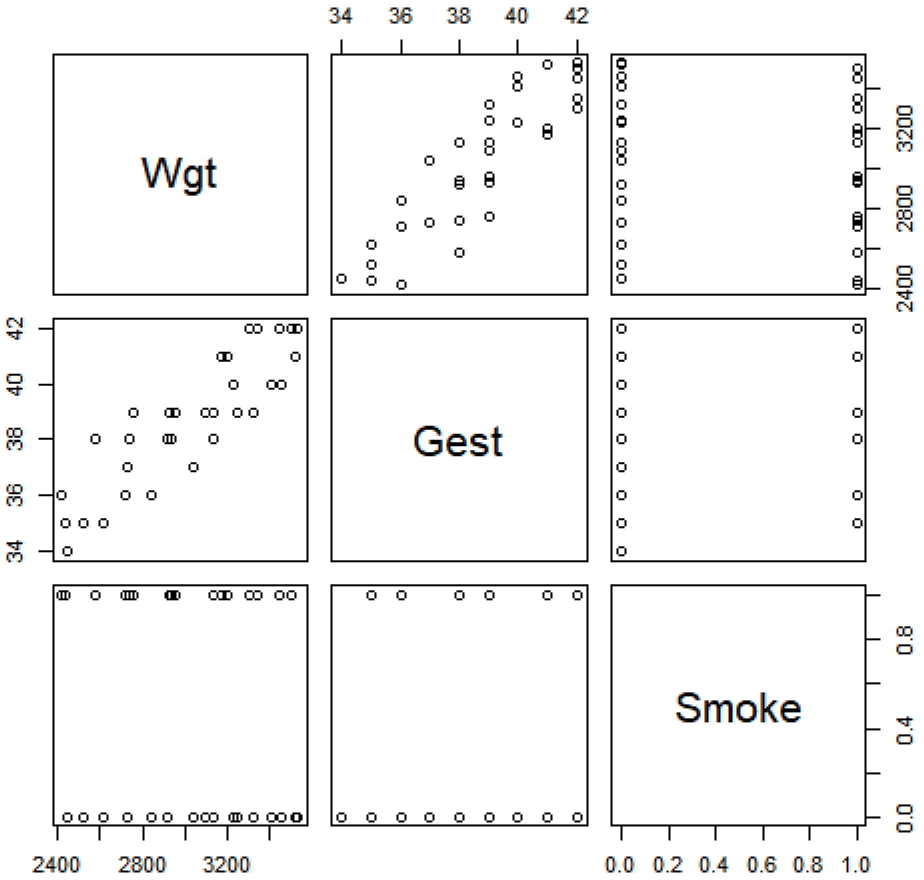
4. Research Question

- Is there a significant difference in mean birth weights for the two groups, after taking into account length of gestation?
  - $y_i = (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}) + \epsilon_i$
  - $H_0: \beta_2 = 0$
  - $H_A: \beta_2 \neq 0$

5. Scatter Plot Matrix

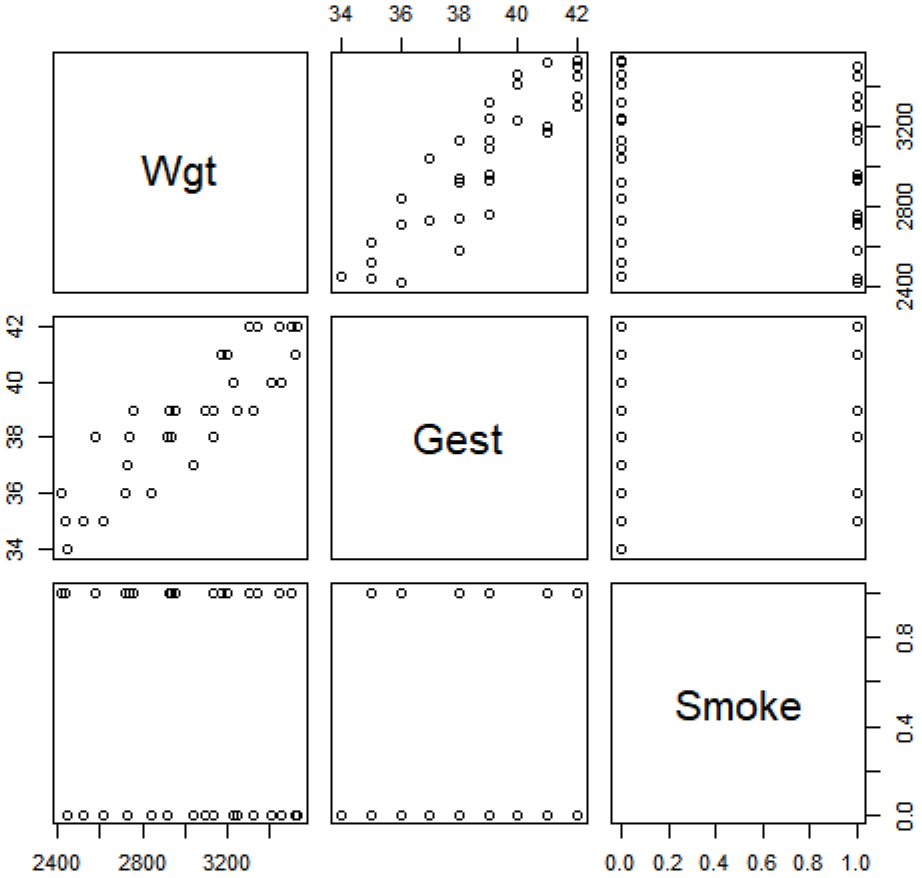
```
pairs(cbind(Wgt, Gest, Smoke))
```

5. Scatter Plot Matrix



- There is a positive linear relationship between length of gestation and birth weight.The variation of the residuals appears to be roughly constant.

5. Scatter Plot Matrix

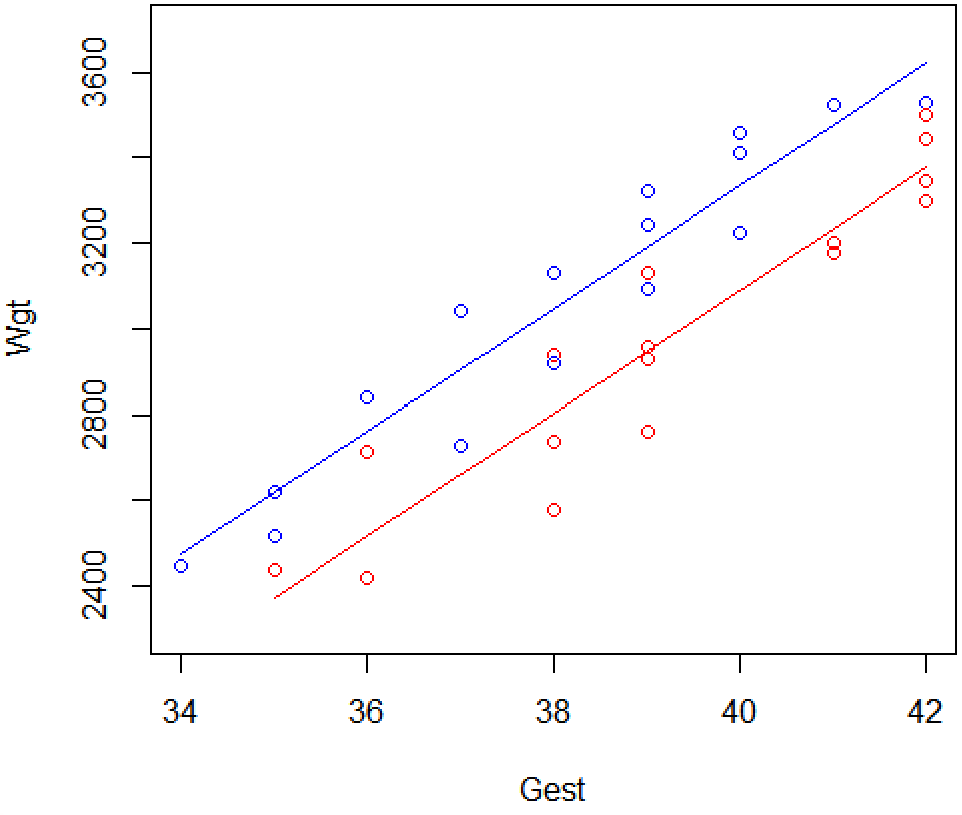


- It is hard to see if any kind of relationship exists between birth weight and smoking status, or between length of gestation and smoking status.

6. Regression Plot

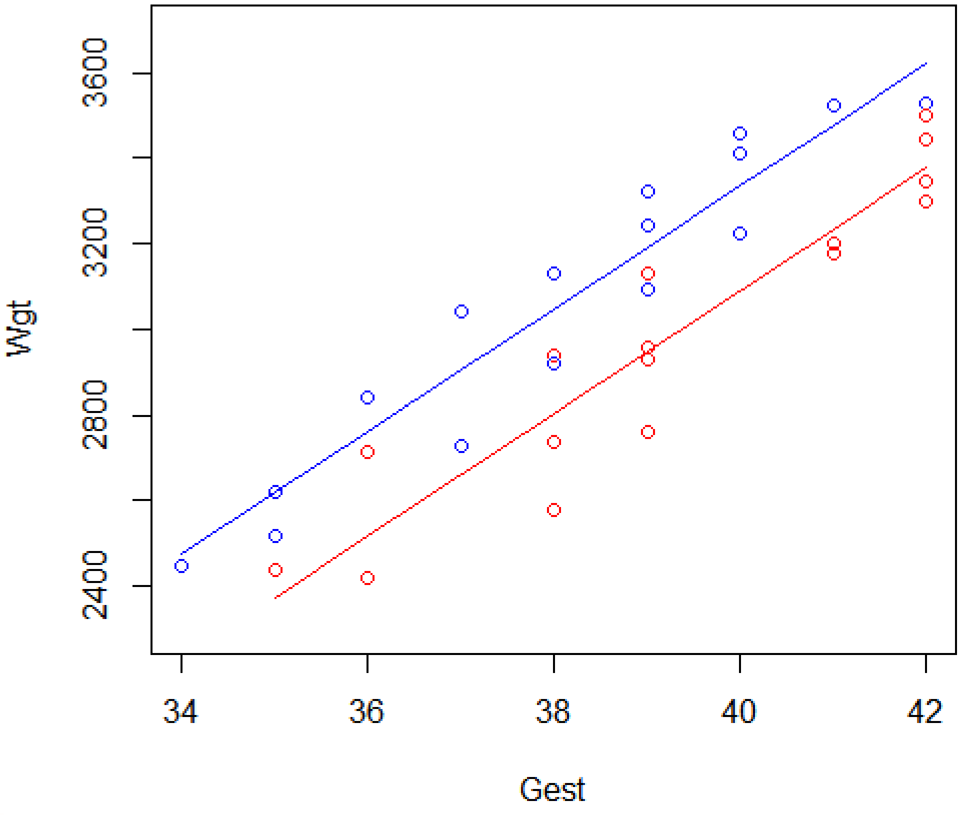
```
model <- lm(Wgt ~ Gest + Smoke)

plot(x=Gest, y=Wgt, ylim=c(2300, 3700),
     col=ifelse(Smoke==1, "red", "blue"),
     panel.last =
       c(lines(sort(Gest[Smoke==0]),
               fitted(model)[Smoke==0][order(Gest[Smoke==0])],
               col="blue"),
         lines(sort(Gest[Smoke==1]),
               fitted(model)[Smoke==1][order(Gest[Smoke==1])],
               col="red")))
```

6. Regression Plot



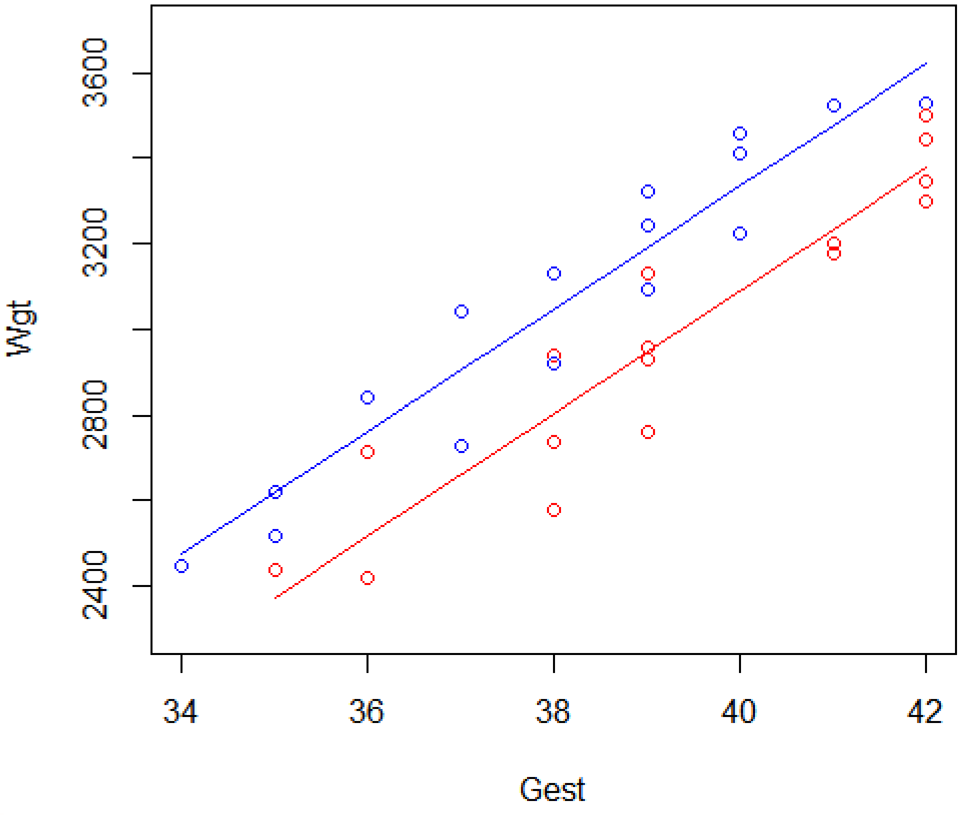
- The blue circles represent the data on non-smoking mothers, while the red circles represent the data on smoking mothers.

6. Regression Plot



- The blue line represents the estimated linear relationship between length of gestation and birth weight for non-smoking mothers, while the red line represents the estimated linear relationship for smoking mothers.

6. Regression Plot



It appears as if the birth weights for non-smoking mothers is higher than that for smoking mothers, regardless of the length of gestation. A hypothesis test or confidence interval would allow us to see if this result extends to the larger population.

7. Model Summary

```
> summary(model)
...
Coefficients:
             Estimate Std. Error t value Pr(>¦t¦)
(Intercept) -2389.573    349.206  -6.843 1.63e-07 ***
Gest          143.100      9.128  15.677 1.07e-15 ***
Smoke        -244.544     41.982  -5.825 2.58e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

7. Model Summary

```
> summary(model)
...
Residual standard error: 115.5 on 29 degrees of freedom
Multiple R-squared:  0.8964,  Adjusted R-squared:  0.8892
F-statistic: 125.4 on 2 and 29 DF,  p-value: 5.289e-15
```

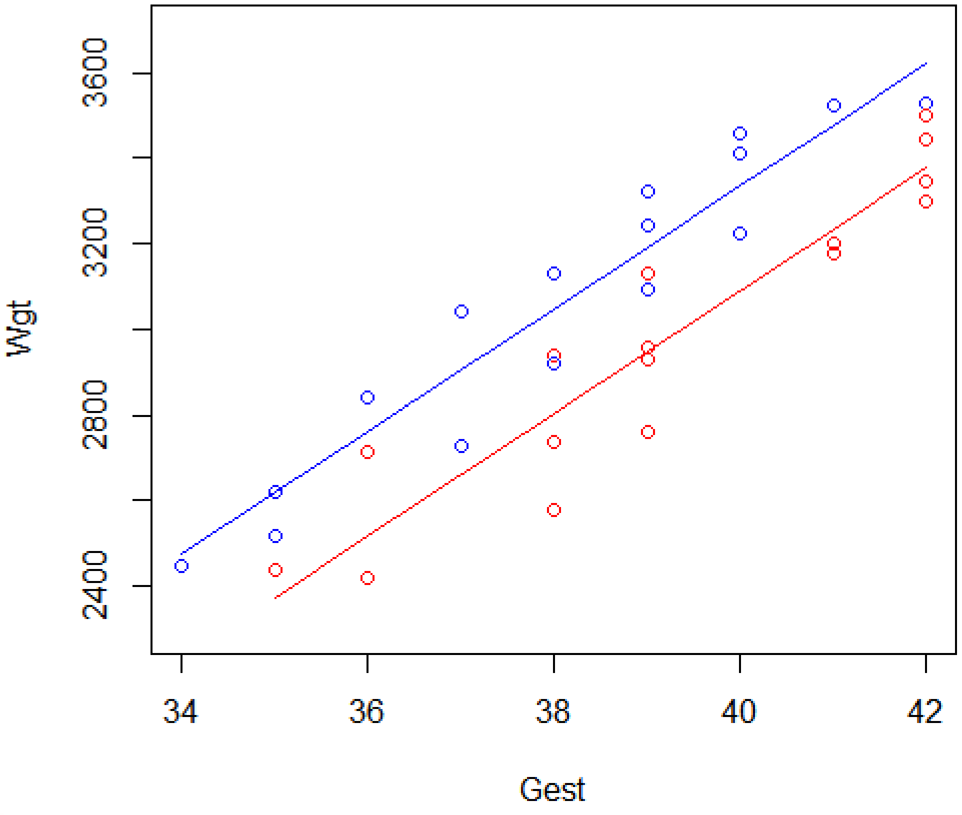
8. ANOVA

```
> anova(model)
Analysis of Variance Table

Response: Wgt
          Df  Sum Sq Mean Sq F value    Pr(>F)    
Gest       1 2895838 2895838 216.962 5.365e-15 ***
Smoke      1  452881  452881  33.931 2.577e-06 ***
Residuals 29  387070   13347                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

9. Confidence Interval

```
> confint(model)
                 2.5 %     97.5 %
(Intercept) -3103.7795 -1675.3663
Gest          124.4312   161.7694
Smoke        -330.4064  -158.6817
```

10. Interpretation of Predictor Variable



- $y_i = (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}) + \epsilon_i$
- β_2 represents how much higher (or lower) the mean response function of the second group is than that of the first group for any value of x_{i1}.

2. Two Separate Advantages

1. Advantages

- Why not just fit two separate regression functions — one for the smokers and one for the non-smokers? Are there advantages to including both the binary and quantitative predictor variables within one multiple regression model?

2. Model Creation

```
model.0 <- lm(Wgt ~ Gest, subset=Smoke==0)
summary(model.0)
```

3. Model Summary

```
> summary(model.0)
...
Coefficients:
            Estimate Std. Error t value Pr(>¦t¦)
(Intercept) -2546.14     457.29  -5.568 6.93e-05 ***
Gest          147.21      11.97  12.294 6.85e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

4. Model Creation

```
model.1 <- lm(Wgt ~ Gest, subset=Smoke==1)
summary(model.1)
```

5. Model Summary

```
> summary(model.1)
...
Coefficients:
            Estimate Std. Error t value Pr(>¦t¦)
(Intercept) -2474.56     553.97  -4.467 0.000532 ***
Gest          139.03      14.11   9.851 1.12e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

6. The First Advantage

- The standard error of the Gest coefficient — SE(Gest) — is smallest for the estimated model based on all 32 data points. Therefore, confidence intervals for the Gest coefficient will be narrower if calculated using the analysis based on all 32 data points.

Model estimated using	SE(Gest)
all 32 data points	9.128
16 nonsmokers	11.97
16 smokers	14.11

6. First Advantage

- There appears to be an advantage in "pooling" and analyzing the data all at once rather than breaking it apart and conducting different analyses for each group. Our regression model assumes that the slope for the two groups are equal. It also assumes that the variances of the error terms are equal. Therefore, it makes sense to use as much data as possible to estimate these quantities.

7. The Second Advantage

- How could you use the results of model.0 and model.1 to determine if the mean birth weight of babies differs between smoking and non-smoking mothers, after taking into account length of gestation?

8. Model Summary: model.0

```
> summary(model.0)
...
Coefficients:
            Estimate Std. Error t value Pr(>¦t¦)
(Intercept) -2546.14     457.29  -5.568 6.93e-05 ***
Gest          147.21      11.97  12.294 6.85e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

9. Model Summary: model.1

```
> summary(model.1)
...
Coefficients:
            Estimate Std. Error t value Pr(>¦t¦)
(Intercept) -2474.56     553.97  -4.467 0.000532 ***
Gest          139.03      14.11   9.851 1.12e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

10. Model Summary: model

```
> summary(model)
...
Coefficients:
             Estimate Std. Error t value Pr(>¦t¦)
(Intercept) -2389.573    349.206  -6.843 1.63e-07 ***
Gest          143.100      9.128  15.677 1.07e-15 ***
Smoke        -244.544     41.982  -5.825 2.58e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

11. The Second Advantage

- There is sufficient evidence to conclude that there is a statistically significant difference in the mean birth weight of all babies of smoking mothers and the mean birth weight of all babies of non-smoking mothers, after taking into account length of gestation.

12. Summary

- "Pooling" your data and fitting one combined regression function allows you to easily and efficiently answer research questions concerning the binary predictor variable.

3. Coding Qualitative Variables

1. Coding Qualitative Variables

- If your qualitative variable defines 2 groups, then you need 1 indicator variable.
- If your qualitative variable defines 3 groups, then you need 2 indicator variables.
- If your qualitative variable defines 4 groups, then you need 3 indicator variables. And, so on.

2. Model Creation

```
Smoke2 <- ifelse(Smoke==1, 0, 1)
model.2 <- lm(Wgt ~ Gest + Smoke + Smoke2)
summary(model.2)
```

3. Model Summary

```
> summary(model.2)
...
Coefficients: (1 not defined because of singularities)
             Estimate Std. Error t value Pr(>¦t¦)
(Intercept) -2389.573    349.206  -6.843 1.63e-07 ***
Gest          143.100      9.128  15.677 1.07e-15 ***
Smoke        -244.544     41.982  -5.825 2.58e-06 ***
Smoke2             NA         NA      NA       NA
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

4. Problematic Coding

- The indicator variable "Smoke2" is "highly correlated" with the indicator variable "Smoke". In fact, "Smoke" and "Smoke2" are perfectly correlated with one another — when "Smoke" is 1, "Smoke2" is always 0 and when "Smoke" is 0, "Smoke2" is always 1. (Described more technically, the columns of the X matrix are linearly dependent.)

5. Coding Rules for Qualitative Variables

- If your qualitative variable defines c groups,
- Choose one group or category to be the "reference" group.
- Observations in this group will have the value zero for all the indicator variables used to code this qualitative variable.
- Each of the remaining c – 1 groups will be represented by one and only one of the c – 1 indicator variables.

6. Different Coding Scheme

- What if we had instead used (1, −1) coding? That is, what if we created a (1, −1) indicator variable, say, defined as:
  - $x_{i2} = 1$, if the mother smokes
  - $x_{i2} = -1$, if mother does not smoke
  - $y_i = (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}) + \epsilon_i$
  - β_2 represents how far each group is "offset" from the overall "average".

6. Different Coding Scheme

```
Smoke3 <- ifelse(Smoke==1, 1, -1)
model.3 <- lm(Wgt ~ Gest + Smoke3)
summary(model.3)
```

6. Different Coding Scheme

```
> summary(model.3)
...
Coefficients:
             Estimate Std. Error t value Pr(>¦t¦)
(Intercept) -2511.845    353.449  -7.107 8.07e-08 ***
Gest          143.100      9.128  15.677 1.07e-15 ***
Smoke3       -122.272     20.991  -5.825 2.58e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

6. Different Coding Scheme

```
> summary(model)
...
Coefficients:
             Estimate Std. Error t value Pr(>¦t¦)
(Intercept) -2389.573    349.206  -6.843 1.63e-07 ***
Gest          143.100      9.128  15.677 1.07e-15 ***
Smoke        -244.544     41.982  -5.825 2.58e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

6. Different Coding Scheme

```
> summary(model.3)
...
Residual standard error: 115.5 on 29 degrees of freedom
Multiple R-squared:  0.8964,  Adjusted R-squared:  0.8892
F-statistic: 125.4 on 2 and 29 DF,  p-value: 5.289e-15
```

6. Different Coding Scheme

```
> summary(model)
...
Residual standard error: 115.5 on 29 degrees of freedom
Multiple R-squared:  0.8964,  Adjusted R-squared:  0.8892
F-statistic: 125.4 on 2 and 29 DF,  p-value: 5.289e-15
```

6. Different Coding Scheme

```
> predict(model, interval="confidence",
+         newdata=data.frame(Gest=38, Smoke=1))
       fit      lwr      upr
1 2803.693 2740.599 2866.788

> predict(model.3, interval="confidence",
+         newdata=data.frame(Gest=38, Smoke3=1))
       fit      lwr      upr
1 2803.693 2740.599 2866.788
```

6. Different Coding Scheme

```
> predict(model, interval="confidence",
+           newdata=data.frame(Gest=38, Smoke=0))
       fit     lwr      upr
1 3048.237 2989.12 3107.355

> predict(model.3, interval="confidence",
+           newdata=data.frame(Gest=38, Smoke3=-1))
       fit     lwr      upr
1 3048.237 2989.12 3107.355
```

7. Summary

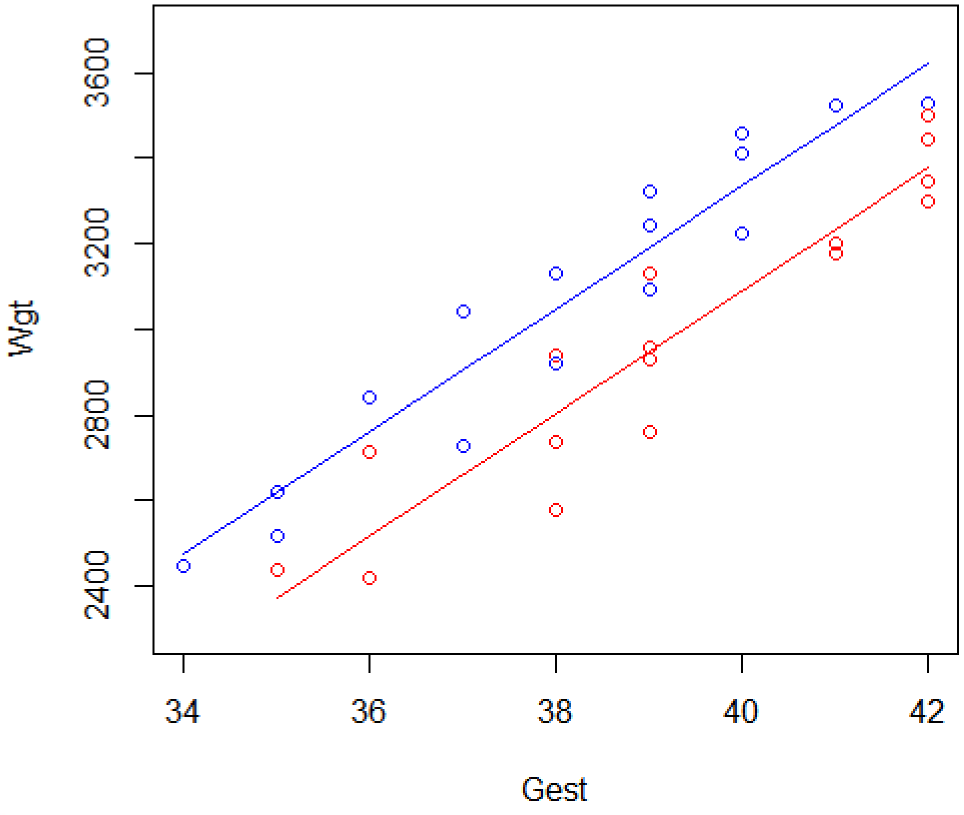
- Regardless of the coding scheme used, we obtain the same two estimated functions and draw the same scientific conclusions.
- It's just how we arrive at those conclusions that differs.
- The meanings of the regression coefficients differ.

4. Additive Effects

1. Additive Effects

- Does the effect of the gestation length on mean birth weight depend on whether or not the mother is a smoker?

1. Additive Effects

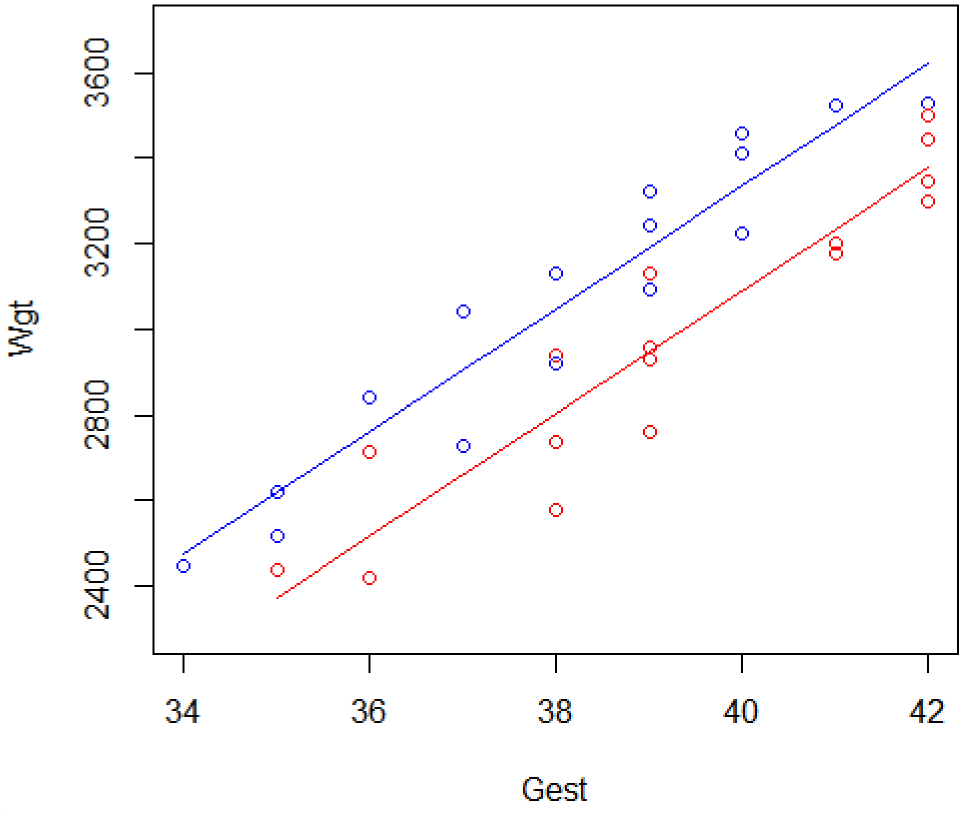


- Answer: No.
- Regardless of whether or not the mother is a smoker, for each additional one-week of gestation, the mean birth weight is predicted to increase by 143 grams.

1. Additive Effects

- Does the effect of smoking on mean birth weight depend on the length of gestation?

1. Additive Effects



- Answer: No.
- For a fixed length of gestation, the mean birth weight of babies born to smoking mothers is predicted to be 245 grams lower than the mean birth weight of babies born to non-smoking mothers.

1. Additive Effects

- When two predictors do not interact, we say that each predictor has an "additive effect" on the response. More formally, a regression model contains additive effects if the response function can be written as a sum of functions of the predictor variables:

$$\mu_Y = f_1(x_1) + f_2(x_2) + \cdots + f_{p-1}(x_{p-1})$$

1. Additive Effects

- For example, our regression model for the birth weights of babies contains additive effects, because the response function can be written as a sum of functions of the predictor variables:

$$\mu_Y = (\beta_0) + (\beta_1 x_{i1}) + (\beta_2 x_{i2})$$

5. Interaction Effects

1. Data: Depression

- Data: Depression
  - y_i (y): measure of the effectiveness of the treatment for individual i
  - x_{i1} (age): age (in years) of individual i
  - x_{i2} (x2): 1 if individual i received treatment A and 0, if not
  - x_{i3} (x3): 1 if individual i received treatment B and 0, if not

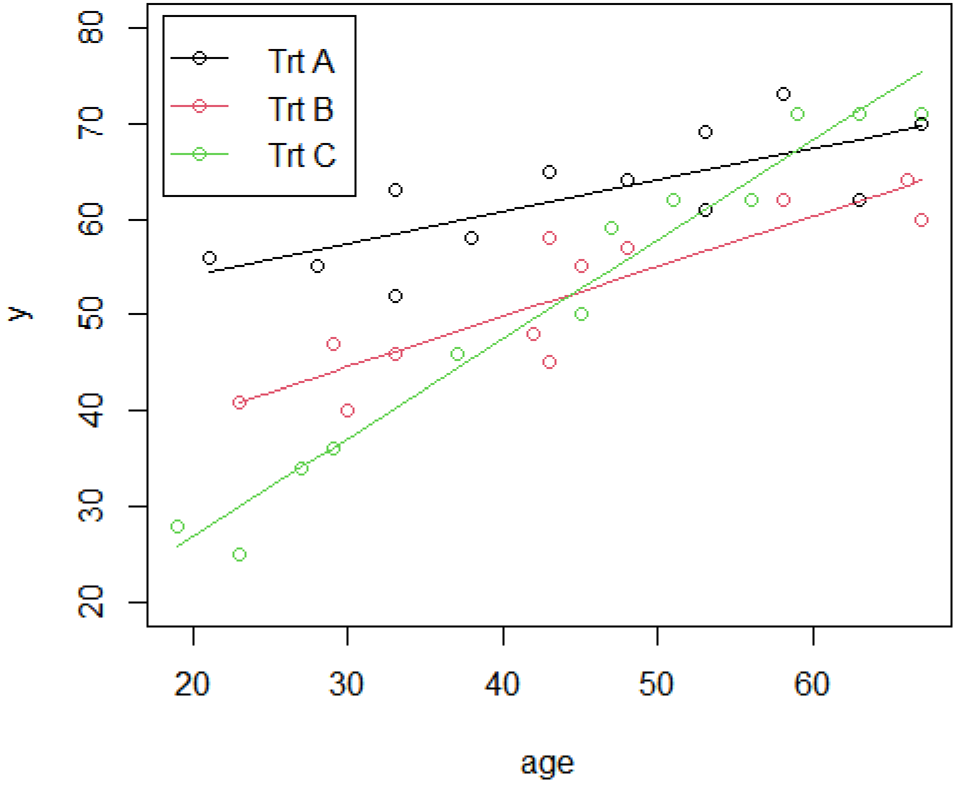
1. Data: Depression

```
> head(depression)
   y age x2 x3 TRT
1 56  21  1  0   A
2 41  23  0  1   B
3 40  30  0  1   B
4 28  19  0  0   C
5 55  28  1  0   A
6 25  23  0  0   C
```

2. Regression Plot

```r
to_num <- function(x) {
  if (x == "A") 1
  else if (x == "B") 2
  else if (x == "C") 3
}
plot(x=age, y=y, col=sapply(TRT, to_num))
legend("topleft", col=1:3, pch=1,
       inset=0.02, x.intersp = 1.5, y.intersp = 1.3,
       legend=c("Trt A", "Trt B", "Trt C"))
```

2. Regression Plot



- The slopes of fitting lines would not be the same.
- In this case, we need to include what are called "interaction terms" in our formulated regression model.

3. Model with Interaction Terms

- $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_{12} x_{i1} x_{i2} + \beta_{13} x_{i1} x_{i3} + \epsilon_i$

Treatment	Formulated regression function
If $x_{i2} = 1\ and\ x_{i3} = 0$	$\mu_Y = (\beta_0 + \beta_2) + (\beta_1 + \beta_{12}) x_{i1}$
If $x_{i2} = 0\ and\ x_{i3} = 1$	$\mu_Y = (\beta_0 + \beta_3) + (\beta_1 + \beta_{13}) x_{i1}$
If $x_{i2} = 0\ and\ x_{i3} = 0$	$\mu_Y = \beta_0 + \beta_1 x_{i1}$

3. Model with Interaction Terms

Treatment	Formulated regression function
If $x_{i2} = 1\ and\ x_{i3} = 0$	$\mu_Y = (\beta_0 + \beta_2) + (\beta_1 + \beta_{12})x_{i1}$
If $x_{i2} = 0\ and\ x_{i3} = 1$	$\mu_Y = (\beta_0 + \beta_3) + (\beta_1 + \beta_{13})x_{i1}$
If $x_{i2} = 0\ and\ x_{i3} = 0$	$\mu_Y = \beta_0 + \beta_1 x_{i1}$

- The effect of the individual's age on the treatment's mean effectiveness depends on the treatment(x_{i2} and x_{i3}).

3. Model with Interaction Terms

Treatment	Formulated regression function
If $x_{i2} = 1\ and\ x_{i3} = 0$	$\mu_Y = (\beta_0 + \beta_2) + (\beta_1 + \beta_{12})x_{i1}$
If $x_{i2} = 0\ and\ x_{i3} = 1$	$\mu_Y = (\beta_0 + \beta_3) + (\beta_1 + \beta_{13})x_{i1}$
If $x_{i2} = 0\ and\ x_{i3} = 0$	$\mu_Y = \beta_0 + \beta_1 x_{i1}$

- The effect of treatment on the treatment's mean effectiveness depends on the individual's age.

4. Interaction between Predictors

- Two predictors interact if the effect on the response variable of one predictor depends on the value of the other.
- A slope parameter can no longer be interpreted as the change in the mean response for each unit increase in the predictor, while the other predictors are held constant.

5. Interaction Effect

- A regression model contains interaction effects if

$$\mu_Y \neq f_1(x_1) + f_2(x_2) + \cdots + f_{p-1}(x_{p-1})$$

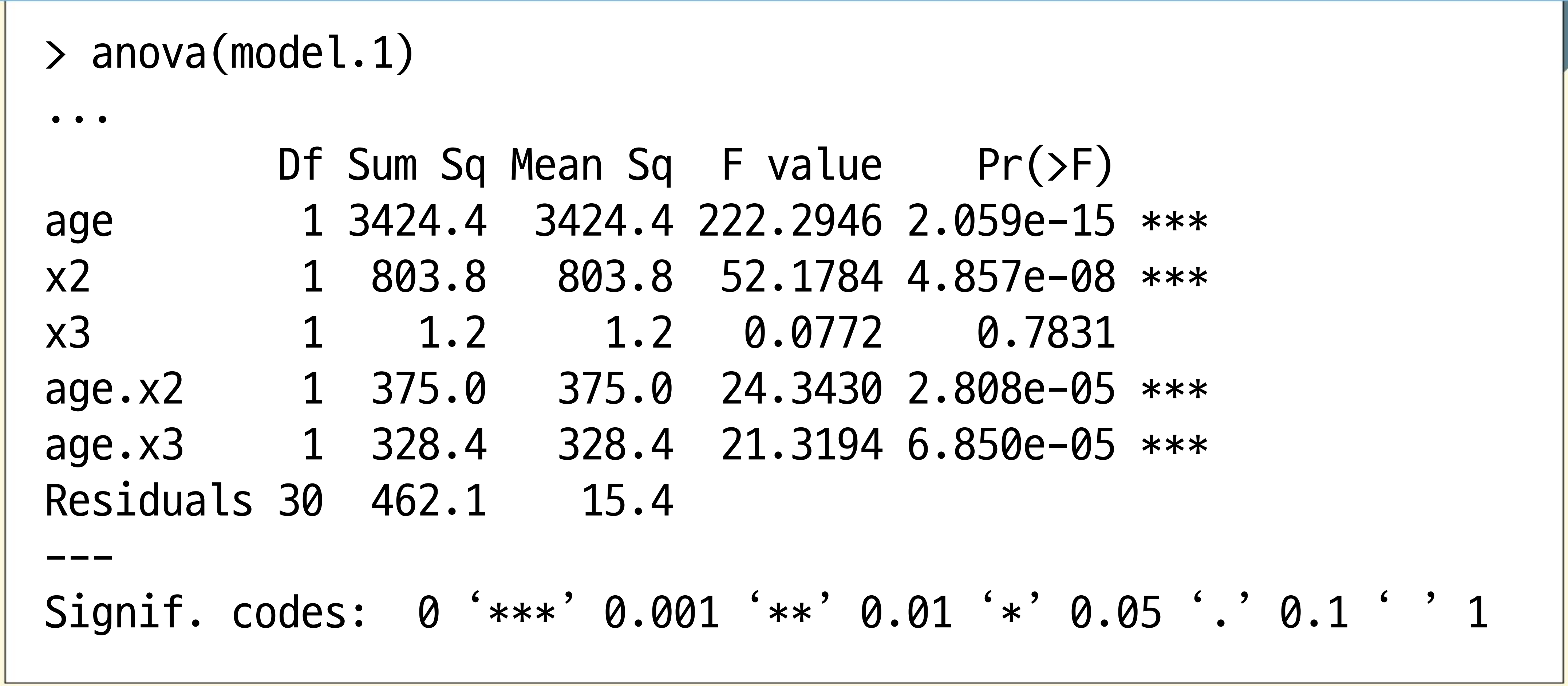
- For our example concerning treatment for depression, the mean response:

$$\mu_Y = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_{12} x_{i1} x_{i2} + \beta_{13} x_{i1} x_{i3}$$

6. Model Creation

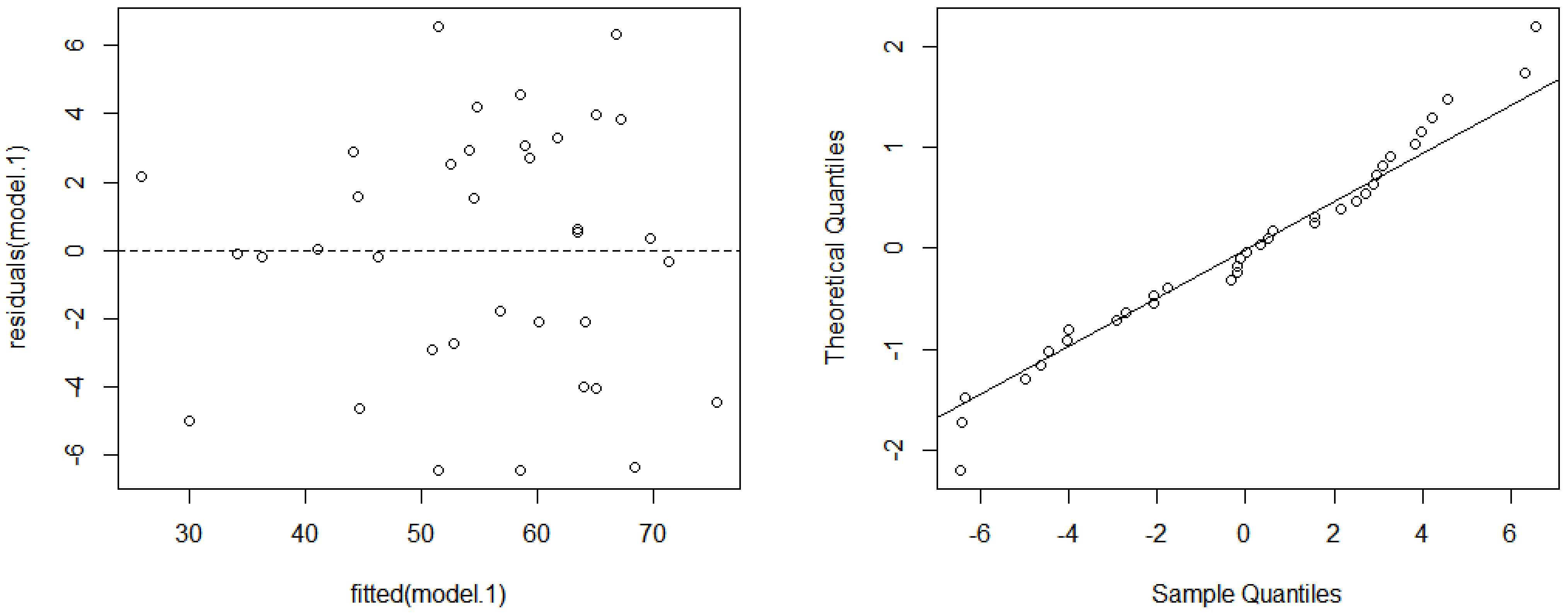
```
age.x2 <- age*x2
age.x3 <- age*x3
model.1 <- lm(y ~ age + x2 + x3 + age.x2 + age.x3)
anova(model.1)

plot(x=fitted(model.1), y=residuals(model.1),
     panel.last = abline(h=0, lty=2))

qqnorm(residuals(model.1), main="", datax=TRUE)
qqline(residuals(model.1), datax=TRUE)
```

7. ANOVA



```
> anova(model.1)
...
          Df Sum Sq Mean Sq  F value    Pr(>F)    
age        1 3424.4  3424.4 222.2946 2.059e-15 ***
x2         1  803.8   803.8  52.1784 4.857e-08 ***
x3         1    1.2     1.2   0.0772    0.7831    
age.x2     1  375.0   375.0  24.3430 2.808e-05 ***
age.x3     1  328.4   328.4  21.3194 6.850e-05 ***
Residuals 30  462.1    15.4                       
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

8. Plots


residuals(model.1)
6
4
2
0
-2
-4
-6
30
40
50
60
70
fitted(model.1)
Theoretical Quantiles
2
1
0
-1
-2
-6
-4
-2
0
2
4
6
Sample Quantiles

9. Question 1

- For every age, is there a difference in the mean effectiveness for the three treatments?
  - $H_0: \beta_2 = \beta_3 = \beta_{12} = \beta_{13} = 0$
  - H_A: At least one of these slope parameters is not 0.

9. Question 1

- $F = \frac{(803.8+1.2+375.0+328.4)/4}{15.4} = 24.49$

```
> pf(24.49, 4, 30, lower.tail=FALSE)
[1] 4.436433e-09
```

- We can reject the null hypothesis. There is sufficient evidence at the $\alpha = 0.05$ level to conclude that there is a significant difference in the mean effectiveness for the three treatments.

10. Question 2

- Does the effect of age on the treatment's effectiveness depend on treatment?
  - $H_0: \beta_{12} = \beta_{13} = 0$
  - H_A: At least one of these interaction parameters is not 0.

10. Question 2

- $F = \frac{(375.0+328.4)/2}{15.4} = 22.84$

```
> pf(22.84, 2, 30, lower.tail=FALSE)
[1] 9.377822e-07
```

- We can reject the null hypothesis. There is sufficient evidence at the $\alpha = 0.05$ level to conclude that the effect of age on the treatment's effectiveness depends on the treatment.

Next

Chapter 11
Data Transformations