HW4

Anya Conti

February 14, 2017

```
## Loading required package: car
```

```
summary(M1)
```

```
##
## Call:
## lm(formula = Time ~ T1 + T2, data = transact)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4652.4  -601.3     2.4   455.7  5607.4
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 144.36944  170.54410   0.847    0.398
## T1            5.46206    0.43327  12.607   <2e-16 ***
## T2            2.03455    0.09434  21.567   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1143 on 258 degrees of freedom
## Multiple R-squared:  0.9091, Adjusted R-squared:  0.9083
## F-statistic:  1289 on 2 and 258 DF,  p-value: < 2.2e-16
```

```
summary(M2)
```

```
##
## Call:
## lm(formula = Time ~ a + d, data = transact)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4652.4  -601.3     2.4   455.7  5607.4
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 144.3694   170.5441   0.847    0.398
## a             7.4966     0.3654  20.514  < 2e-16 ***
## d             1.7138     0.2548   6.726 1.12e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1143 on 258 degrees of freedom
## Multiple R-squared:  0.9091, Adjusted R-squared:  0.9083
## F-statistic:  1289 on 2 and 258 DF,  p-value: < 2.2e-16
```

```
summary(M3)
```

```
##
## Call:
```

```
## lm(formula = Time ~ T2 + d, data = transact)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4652.4  -601.3     2.4   455.7  5607.4
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 144.3694   170.5441   0.847    0.398
## T2            7.4966     0.3654  20.514   <2e-16 ***
## d             5.4621     0.4333  12.607   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1143 on 258 degrees of freedom
## Multiple R-squared:  0.9091, Adjusted R-squared:  0.9083
## F-statistic:  1289 on 2 and 258 DF,  p-value: < 2.2e-16
```

```
summary(M4)
```

```
##
## Call:
## lm(formula = Time ~ T1 + T2 + a + d, data = transact)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4652.4  -601.3     2.4   455.7  5607.4
##
## Coefficients: (2 not defined because of singularities)
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 144.36944  170.54410   0.847    0.398
## T1            5.46206    0.43327  12.607   <2e-16 ***
## T2            2.03455    0.09434  21.567   <2e-16 ***
## a                  NA         NA      NA       NA
## d                  NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1143 on 258 degrees of freedom
## Multiple R-squared:  0.9091, Adjusted R-squared:  0.9083
## F-statistic:  1289 on 2 and 258 DF,  p-value: < 2.2e-16
```

4.2.1. There exists perfect multicollinearity: an exact linear relationship among the independent variables (because of $a = \frac{t1+t2}{2}$ and $d = t1 - t2$ which are both linear combinations of t1 and t2). Calculating a regression thus becomes impossible (matrix would not be invertible) and so R ommits the last two variables to avoid this problem.

4.2.2. The intercept of every single model, along with its statistics (standard error, t-value, p-value) are all the same between every model. In addition, the R^2 values are the same. The first and last models are exactly the same. The coefficients in the others are all linear combinations of those. The standard error, t-value, and p-values are altered for these new parameter estimates.

4.2.3. It may appear different at first glance but it is actually the same, because t2 is also included in the d variable.

$$\hat{time} = \hat{\beta}_{03} + \hat{\beta}_{23} t2 + \hat{\beta}_{43} d$$

$$\hat{time} = \hat{\beta}_{03} + \hat{\beta}_{23}t2 + \hat{\beta}_{43}(t1 - t2)$$
$$\hat{time} = \hat{\beta}_{03} + \hat{\beta}_{23}t2 + \hat{\beta}_{43}t1 - \hat{\beta}_{43}t2$$
$$\hat{time} = \hat{\beta}_{03} + \hat{\beta}_{23}t2 - \hat{\beta}_{43}t2 + \hat{\beta}_{43}t1$$
$$\hat{time} = \hat{\beta}_{03} + (\hat{\beta}_{23} - \hat{\beta}_{43})t2 + \hat{\beta}_{43}t1$$

So these models are the same given the following:

$$\hat{\beta}_{23} - \hat{\beta}_{43} = \hat{\beta}_{21}$$

And

$$\hat{\beta}_{43} = \hat{\beta}_{11}$$

And these are true in the regressions.

$$\hat{\beta}_{23} - \hat{\beta}_{43} = \hat{\beta}_{21} = 2.035$$

$$\hat{\beta}_{43} = \hat{\beta}_{11} = 5.462$$

Note: something similar can be shown for all of these models because of a and d being linearly dependent with T1 and T2, so all these models are essentially the same

BELOW IS PROBLEM REDONE WITH $a = t1 + t2$ instead of $a = \frac{t1+t2}{2}$

Analysis is all the same, just numbers are slightly different

```
summary(M1)
```

```
##
## Call:
## lm(formula = Time ~ T1 + T2, data = transact)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4652.4  -601.3     2.4   455.7  5607.4
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 144.36944  170.54410   0.847    0.398
## T1            5.46206    0.43327  12.607   <2e-16 ***
## T2            2.03455    0.09434  21.567   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1143 on 258 degrees of freedom
## Multiple R-squared:  0.9091, Adjusted R-squared:  0.9083
## F-statistic:  1289 on 2 and 258 DF,  p-value: < 2.2e-16
```

```
summary(M2)
```

```
##
## Call:
## lm(formula = Time ~ a + d, data = transact)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -4652.4  -601.3     2.4   455.7  5607.4
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 144.3694   170.5441   0.847    0.398
## a             3.7483     0.1827  20.514  < 2e-16 ***
## d             1.7138     0.2548   6.726 1.12e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1143 on 258 degrees of freedom
## Multiple R-squared:  0.9091, Adjusted R-squared:  0.9083
## F-statistic:  1289 on 2 and 258 DF,  p-value: < 2.2e-16
```

```
summary(M3)
```

```
##
## Call:
## lm(formula = Time ~ T2 + d, data = transact)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4652.4  -601.3     2.4   455.7  5607.4
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 144.3694   170.5441   0.847    0.398
## T2            7.4966     0.3654  20.514   <2e-16 ***
## d             5.4621     0.4333  12.607   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1143 on 258 degrees of freedom
## Multiple R-squared:  0.9091, Adjusted R-squared:  0.9083
## F-statistic:  1289 on 2 and 258 DF,  p-value: < 2.2e-16
```

```
summary(M4)
```

```
##
## Call:
## lm(formula = Time ~ T1 + T2 + a + d, data = transact)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4652.4  -601.3     2.4   455.7  5607.4
##
## Coefficients: (2 not defined because of singularities)
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 144.36944  170.54410   0.847    0.398
## T1            5.46206    0.43327  12.607   <2e-16 ***
## T2            2.03455    0.09434  21.567   <2e-16 ***
## a                  NA         NA      NA       NA
## d                  NA         NA      NA       NA
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1143 on 258 degrees of freedom
## Multiple R-squared:  0.9091, Adjusted R-squared:  0.9083
## F-statistic:  1289 on 2 and 258 DF,  p-value: < 2.2e-16
```

4.2.1. There exists perfect multicollinearity: an exact linear relationship among the independent variables (because of $a = t1 + t2$ and $d = t1 - t2$ which are both linear combinations of t1 and t2). Calculating a regression thus becomes impossible (matrix would not be invertible) and so R ommits the last two variables to avoid this problem.

4.2.2. The intercept of every single model, along with its statistics (standard error, t-value, p-value) are all the same between every model. In addition, the R^2 values are the same. The first and last models are exactly the same. The coefficients in the others are all linear combinations of those. The standard error, t-value, and p-values are altered for these new parameter estimates.

4.2.3. It may appear different at first glance but it is actually the same, because t2 is also included in the d variable.

$$\hat{time} = \hat{\beta}_{03} + \hat{\beta}_{23}t2 + \hat{\beta}_{43}d$$
$$\hat{time} = \hat{\beta}_{03} + \hat{\beta}_{23}t2 + \hat{\beta}_{43}(t1 - t2)$$
$$\hat{time} = \hat{\beta}_{03} + \hat{\beta}_{23}t2 + \hat{\beta}_{43}t1 - \hat{\beta}_{43}t2$$
$$\hat{time} = \hat{\beta}_{03} + \hat{\beta}_{23}t2 - \hat{\beta}_{43}t2 + \hat{\beta}_{43}t1$$
$$\hat{time} = \hat{\beta}_{03} + (\hat{\beta}_{23} - \hat{\beta}_{43})t2 + \hat{\beta}_{43}t1$$

So these models are the same given the following:

$$\hat{\beta}_{23} - \hat{\beta}_{43} = \hat{\beta}_{21}$$

And

$$\hat{\beta}_{43} = \hat{\beta}_{11}$$

And these are true in the regressions.

$$\hat{\beta}_{23} - \hat{\beta}_{43} = \hat{\beta}_{21} = 2.035$$

$$\hat{\beta}_{43} = \hat{\beta}_{11} = 5.462$$

Note: something similar can be shown for all of these models because of a and d being linearly dependent with T1 and T2, so all these models are essentially the same

4.6. For every 1% increase in the percent of the population living in Urban areas, there is an estimated 1% decrease in fertility.

Extra

Below is shown sepred, and then the 95% interval of the prediction is right below that.

```
## [1] 2.267491
```

```
##        fit      lwr      upr
## 1 64.58925 60.14112 69.03737
```

Below is shown sefit, and then the 95% confidence interval of the fit is shown below that.

```
## [1] 0.07313503
```

```
##        fit      lwr      upr
## 1 64.58925 64.44578 64.73271
```