Week ON10

Elmer V Villanueva

27 April 2020

SET YOUR WORKING DIRECTORY!

```
setwd("C:/Users/1/Dropbox/00 - Working Folder/Teaching/DPH112/2019-2020/Week ON10")
```

Announcements

- The two coursework assessments and the final paper have been released. The deadlines are

Assessment	Due Date	Days Till Deadline
Coursework 1	16 May	19
coursework 2	30 May	33
Final Paper	17 June	51

- All the assessments are to be completed INDIVIDUALLY. No collusion is allowed. This means that you cannot discuss this assignment with other students, nor can you share your approach or code.
- The running count for students forwarding errors is as follows:

Student	Items Identified
Yijia Jiang	6
Jing Wang	2
Xinwen Hu	1
Yuxuan Wu	1

Reading

Read and understand Vittinghoff et al., Chapter 4.

Review of previous learning

We have now completed our extension of our understanding of simple linear regression models

$Y = \beta_0 + \beta_1 X_1$

onto multiple linear regression models

$Y = \beta_0 + \beta_1 X_1 + beta_2 X_2 + \cdots + \beta_p X_p = \beta_0 + \sum \beta_p X_p$

The process of development of models under both cases is the same – you build the model and then you test it. Before we move on to different sort of model kinds of models, I want to discuss three very important

topics that are easily demonstrated using our current learning.

Example 1A

In this lesson, we will use data from a random sample of 100 births from the North Carolina Birth Registry. The birth weight (BW) in grams of 100 babies were recorded, as was the length of the pregnancy in weeks (WK). For the purposes of our example, BW is the dependent variable.

Table 1. Data from a simple random sample of 100 births from the North Carolina Birth Registry.

ID	BW	WK	ID	BW	WK
1	3147	40	51	3232	38
2	2977	41	52	3317	40
3	3119	38	53	2863	37
4	3487	38	54	3175	37
5	4111	39	55	3317	40
6	3572	41	56	3714	34
7	3487	40	57	2240	36
8	3147	41	58	3345	39
9	3345	38	59	3119	39
10	2665	34	60	2920	37
11	1559	34	61	3430	41
12	3799	38	62	3232	35
13	2750	38	63	3430	38
14	3487	40	64	4139	39
15	3317	38	65	3714	39
16	3544	43	66	1446	28
17	3459	45	67	3147	39
18	2807	37	68	2580	31
19	3856	40	69	3374	37
20	3260	40	70	3941	40
21	2183	42	71	2070	37
22	3204	38	72	3345	40
23	3005	36	73	3600	40
24	3090	40	74	3232	41
25	3430	39	75	3657	38
26	3119	40	76	3487	39
27	3912	39	77	2948	38
28	3572	40	78	2722	40
29	3884	41	79	3771	40
30	3090	38	80	3799	45
31	2977	42	81	1871	33
32	3799	37	82	3260	39
33	4054	40	83	3969	38
34	3430	38	84	3771	40
35	3459	41	85	3600	40
36	3827	39	86	2693	35
37	3147	44	87	3062	45
38	3289	38	88	2693	36
39	3629	36	89	3033	41
40	3657	36	90	3856	42
41	3175	41	91	4111	40
42	3232	43	92	3799	39

ID	BW	WK	ID	BW	WK
43	3175	36	93	3147	38
44	3657	40	94	2920	36
45	3600	39	95	4054	40
46	3572	40	96	2296	36
47	709	25	97	3402	38
48	624	25	98	1871	33
49	2778	36	99	4167	41
50	3572	35	100	3402	37

```
NCREG <- data.frame(ID <- c(1:100),
                    BW <- c(3147, 2977, 3119, 3487, 4111,
                            3572, 3487, 3147, 3345, 2665,
                            1559, 3799, 2750, 3487, 3317,
                            3544, 3459, 2807, 3856, 3260,
                            2183, 3204, 3005, 3090, 3430,
                            3119, 3912, 3572, 3884, 3090,
                            2977, 3799, 4054, 3430, 3459,
                            3827, 3147, 3289, 3629, 3657,
                            3175, 3232, 3175, 3657, 3600,
                            3572, 709, 624, 2778, 3572,
                            3232, 3317, 2863, 3175, 3317,
                            3714, 2240, 3345, 3119, 2920,
                            3430, 3232, 3430, 4139, 3714,
                            1446, 3147, 2580, 3374, 3941,
                            2070, 3345, 3600, 3232, 3657,
                            3487, 2948, 2722, 3771, 3799,
                            1871, 3260, 3969, 3771, 3600,
                            2693, 3062, 2693, 3033, 3856,
                            4111, 3799, 3147, 2920, 4054,
                            2296, 3402, 1871, 4167, 3402),
                    WK <- c(40, 41, 38, 38, 39, 41, 40, 41, 38, 34,
                            34, 38, 38, 40, 38, 43, 45, 37, 40, 40,
                            42, 38, 36, 40, 39, 40, 39, 40, 41, 38,
                            42, 37, 40, 38, 41, 39, 44, 38, 36, 36,
                            41, 43, 36, 40, 39, 40, 25, 25, 36, 35,
                            38, 40, 37, 37, 40, 34, 36, 39, 39, 37,
                            41, 35, 38, 39, 39, 28, 39, 31, 37, 40,
                            37, 40, 40, 41, 38, 39, 38, 40, 40, 45,
                            33, 39, 38, 40, 40, 35, 45, 36, 41, 42,
                            40, 39, 38, 36, 40, 36, 38, 33, 41, 37))
str(NCREG)
```

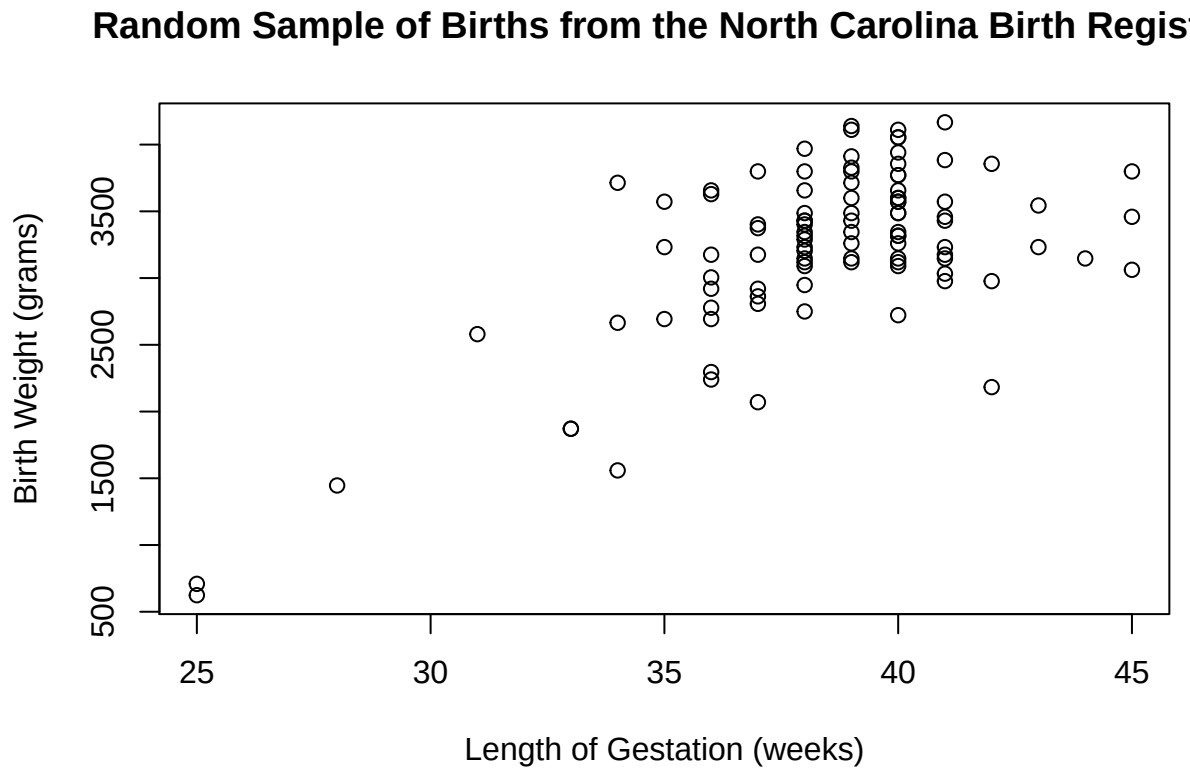
```
## 'data.frame':    100 obs. of  3 variables:
##  $ ID....c.1.100.                                                : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ BW....c.3147..2977..3119..3487..4111..3572..3487..3147..3345..  : num  3147 2977 3119 3487 4111 .
##  $ WK....c.40..41..38..38..39..41..40..41..38..34..34..38..38..40..: num  40 41 38 38 39 41 40 41 38
```

```
head(NCREG)
```

```
##   ID....c.1.100. BW....c.3147..2977..3119..3487..4111..3572..3487..3147..3345..
## 1              1                                                          3147
## 2              2                                                          2977
## 3              3                                                          3119
```

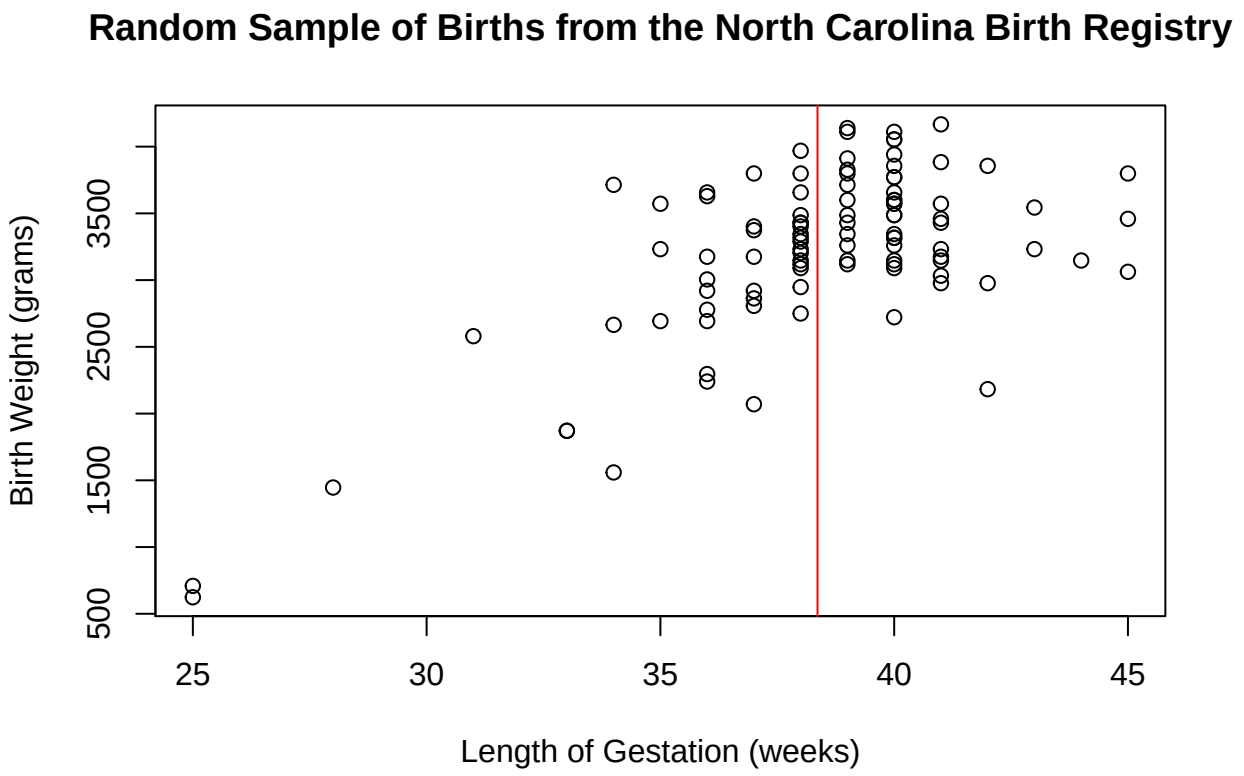
```
## 4              4                                                                3487
## 5              5                                                                4111
## 6              6                                                                3572
##   WK....c.40..41..38..38..39..41..40..41..38..34..34..38..38..40..
## 1                                                                 40
## 2                                                                 41
## 3                                                                 38
## 4                                                                 38
## 5                                                                 39
## 6                                                                 41
```

```
plot(y = NCREG$BW, x = NCREG$WK,
     ylab = "Birth Weight (grams)",
     xlab = "Length of Gestation (weeks)",
     main = "Random Sample of Births from the North Carolina Birth Registry")
```



I will mark out the mean value of the independent variable using a red vertical line.

```
plot(y = NCREG$BW, x = NCREG$WK,
     ylab = "Birth Weight (grams)",
     xlab = "Length of Gestation (weeks)",
     main = "Random Sample of Births from the North Carolina Birth Registry")
abline(v = 38.36, col = "red")
```



On regressing BW on WK, we derive the following results:

```
NCREG.LM1 <- lm(BW ~ WK, data = NCREG)
summary(NCREG.LM1)
```

```
##
## Call:
## lm(formula = BW ~ WK, data = NCREG)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1507.23  -280.16    36.14   324.48  1056.60
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1732.12     569.28  -3.043  0.00301 **
## WK            129.10      14.78   8.732 6.79e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 493.7 on 98 degrees of freedom
## Multiple R-squared:  0.4376, Adjusted R-squared:  0.4319
## F-statistic: 76.25 on 1 and 98 DF,  p-value: 6.785e-14
```

The estimated equation is $BW = -1732.1 + 129.1WK$.

Centring Data

We learned that the interpretation of the intercept parameter is often quite problematic depending on the scope of the independent variable. This is the case here. The formal interpretation of $b_0 = -1732.1$ is that if the duration of pregnancy is zero weeks, we would expect the birthweight to be -1,732.1 grams. A negative weight is, of course, impossible. However, even if the estimate was positive, it would be very hard for us to believe this because the value of the independent variable is out of scope.

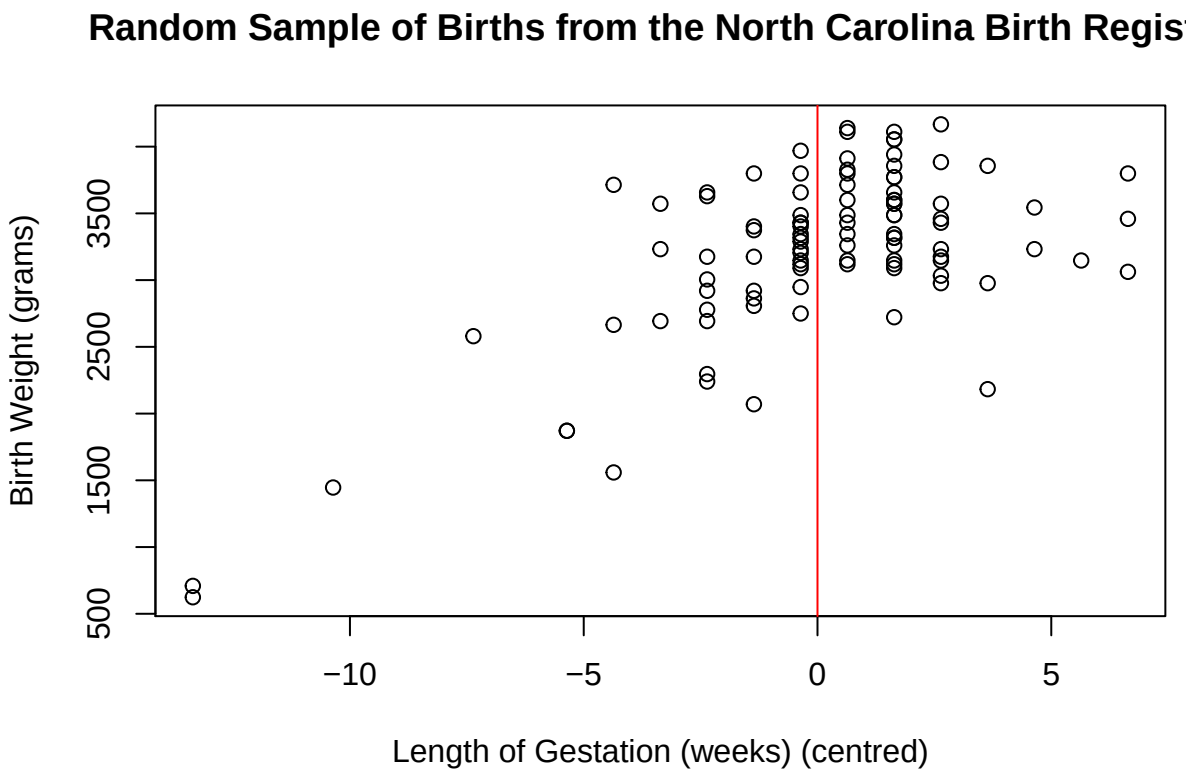
There is an easy soluton to this problem. We can adjust the independent variable by subtracting the mean of all observation from each observation: $x_{ci} = x_i - \bar{x}$. This new variable is centred with a mean of zero, but with the same standard deviation as the original data. If we use this centred variable in the regression, we get the following result:

```
NCREG$WK.CEN <- NCREG$WK - mean(NCREG$WK)
NCREG.LM2 <- lm(BW ~ WK.CEN, data = NCREG)
summary(NCREG.LM2)
```

```
##
## Call:
## lm(formula = BW ~ WK.CEN, data = NCREG)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1507.23  -280.16    36.14   324.48  1056.60
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3220.29      49.37  65.229  < 2e-16 ***
## WK.CEN        129.10      14.78   8.732 6.79e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 493.7 on 98 degrees of freedom
## Multiple R-squared:  0.4376, Adjusted R-squared:  0.4319
## F-statistic: 76.25 on 1 and 98 DF,  p-value: 6.785e-14
```

The estimated equation now becomes $BW = 3220.3 + 129.1 WK.CEN$. While we have retained the same estimate b_1 across both models, the intercepts have changed substantially. How, then, do we interpret this information. Perhaps the centred scatterplot will give us a clue.

```
plot(y = NCREG$BW, x = NCREG$WK.CEN,
     ylab = "Birth Weight (grams)",
     xlab = "Length of Gestation (weeks) (centred)",
     main = "Random Sample of Births from the North Carolina Birth Registry")
abline(v = 0, col = "red")
```



When we centre an independent variable, the zero point is now the mean of the distribution. This is seen clearly in the scatterplot above. While the formal interpretation of the intercept does not change – it is the value of the dependent variable when the independent variable is zero – the meaning of the zero point is now quite meaningful. Thus, for the results shown above, the estimated birthweight is 3220.3 grams during the mean length of gestation of 38.4 weeks. The interpretation of the slope parameter has not changed: for every week that the pregnancy progresses, the birthweight increases by 129.1 grams.

Scaling Data

Centring only brings us so far. Note what happens when we change the unit of the independent variable. For example, instead of weeks, we are interested in the daily change in birthweight.

```
NCREG$DAY.CEN <- NCREG$WK.CEN * 7
NCREG.LM3 <- lm(BW ~ DAY.CEN, data = NCREG)
summary(NCREG.LM3)
```

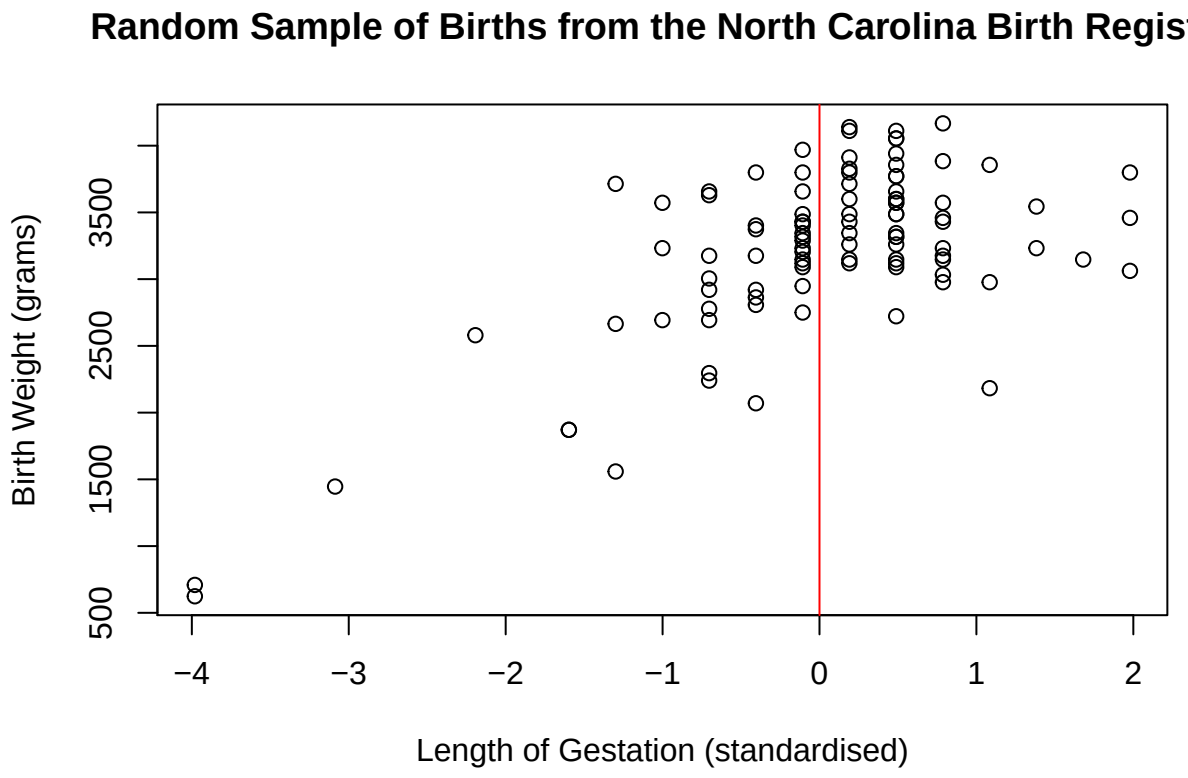
```
##
## Call:
## lm(formula = BW ~ DAY.CEN, data = NCREG)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1507.23  -280.16    36.14   324.48  1056.60
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept) 3220.290     49.369  65.229  < 2e-16 ***
## DAY.CEN       18.443      2.112   8.732 6.79e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 493.7 on 98 degrees of freedom
## Multiple R-squared:  0.4376, Adjusted R-squared:  0.4319
## F-statistic: 76.25 on 1 and 98 DF,  p-value: 6.785e-14
```

The slope estimate has changed because we changed the unit of the independent variable. We encountered this problem when we discussed the covariance. The solution we used was to standardise the data so that they become unitless. In the case of covariance, the resulting standardised measure was called the correlation.

We can do the same here. In effect, we wish to produce an independent variable that is *unitless*. To do this, we can scale a centred variable by dividing by the estimate of the standard deviation s_x.

```
NCREG$WK.STD <- NCREG$WK.CEN/sd(NCREG$WK.CEN)
plot(y = NCREG$BW, x = NCREG$WK.STD,
     ylab = "Birth Weight (grams)",
     xlab = "Length of Gestation (standardised)",
     main = "Random Sample of Births from the North Carolina Birth Registry")
abline(v = 0, col = "red")
```



IMPORTANT: Remember that a centred and scaled variable is called a standardised variable.

The estimated linear model is

```
NCREG.LM4 <- lm(BW ~ WK.STD, data = NCREG)
summary(NCREG.LM4)
```

```
##
## Call:
## lm(formula = BW ~ WK.STD, data = NCREG)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1507.23  -280.16    36.14   324.48  1056.60
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3220.29      49.37  65.229  < 2e-16 ***
## WK.STD        433.28      49.62   8.732 6.79e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 493.7 on 98 degrees of freedom
## Multiple R-squared:  0.4376, Adjusted R-squared:  0.4319
## F-statistic: 76.25 on 1 and 98 DF,  p-value: 6.785e-14
```

How do we interpret the slope parameter for standardised independent variables? Very easy. Applying the usual interpretation, we state that for every unit increase in the length of gestation, the birthweight increases by 433.3 grams. There is no difference in our formal of interpretation, although the meaning changes. Remember that a standardised variable has mean of zero and standard deviation of one. Therefore, a unit increase means that the independent variable increases by 1 standard deviation. Another way to interpret this is to say that an increase of one standard deviation in the length of gestation increases birthweight by 433.3 grams.

What is the value of a single standard deviation?

```
sd(NCREG$WK.CEN)
```

```
## [1] 3.356044
```

Centring, scaling and standardising can be performed on the dependent variable, too. Some famous examples of standardised variables in health and medicine include bone minreal density, anthropometric data and IQ. Sometimes, standardised vaiables are called *z-scores*.

Qualitative Independent Variables

So far, we have considered only independent variables that have been *continuous scaled* that have yielded numerical values that were measurements in the usual sense of the word. Frequently, however, it is desirable to use one or more qualitative variables as independent variables in the regression model. Qualitative variables, it will be recalled, are those variables whose "values" are categories and that convey the concept of attribute rather than amount or quantity. Another name for qualitative variables are nominal or categorical varaibles. The variable marital status, for example, is a qualitative variable whose categories are "single," "married," "widowed," and "divorced." Other examples of qualitative variables include sex (male or female), diagnosis, race, occupation, and immunity status to some disease.

IMPORTANT: In R, qualitative variables are *factor* variables.

In order to incorporate a qualitative independent variable ina regression model, it must be quantified in some manner. This may be accomplished through the use of what are known as *dummy variables*. A dummy variable is a variable that assumes only a finite number of values (such as 0 or 1) for the purpose of identifying the different categories of a qualitative variable.

The term "dummy" is used to indicate the fact that the numerical values (such as 0 and 1) assumed by the variable have no quantitative meaning but are used merely to identify different categories of the qualitative

variable under consideration. Qualitative variables are sometimes called indicator variables, and when there are only two categories, they are sometimes called *dichotomous* or *binary* variables.

IMPORTANT: Technically, any numerical values can be used. For example, to indicate the two categories of sex, you can set "males" to be 56 and "females" to be $sin(\sqrt{30} - \pi)$. Doing this comes at the cost of complicating the interpretation of the results. For the remainder of this module, we will specify codes that are zero and one. For example, "males" is one and "females" is zero.

Where the qualitative variable has k categories, $k - 1$ dummy variables must be defined for all the categories to be properly coded. The variable sex, with two categories, can be quantified by the use of only one dummy variable.

Table 2. Dummy coding for sex.

Level of sex	Variable X1
Male	1
Female	0

Note how all levels of the variable are defined by a single dummy variable. To specify that an observation is male, then $x_1 = 1$. To specify that an observation is female, then $x_1 = 0$

If your variable has four levels, then you require three dummy variables. For example, take the variable `location of house` which may have four categories – Suzhou, Nanjing, Wuxi, Other Jiangsu.

Table 3. Dummy coding for location of house

Level of variable	Variable X1	Variable X2	Variable X3
Suzhou	1	0	0
Nanjing	0	1	0
Wuxi	0	0	1
Other Jiangsu	0	0	0

To specify that someone's house is in Suzhou, will mean that $x_1 = 1; x_2 = 0; x_3 = 0$. Someone with a house in Nanjing will have this pattern: $x_1 = 0; x_2 = 1; x_3 = 0$. Someone with a houses in Wujiang will have this pattern: $x_1 = 0; x_2 = 0; x_3 = 0$.

Let us use an example.

Example 1B

The data I gave in Example 1A was incomplete. In fact, we have access to a second independent variable – the smoking status of the mother (SM). The smoking status is coded 1 if the mother is a smoker and 0 if the mother is a non-smoker.

Table 4. Data from a simple random sample of 100 births from the North Carolina Birth Registry.

ID	BW	WK	SM	ID	BW	WK	SM
1	3147	40	0	51	3232	38	0
2	2977	41	0	52	3317	40	0
3	3119	38	0	53	2863	37	0
4	3487	38	0	54	3175	37	0
5	4111	39	0	55	3317	40	0
6	3572	41	0	56	3714	34	0

ID	BW	WK	SM	ID	BW	WK	SM
7	3487	40	0	57	2240	36	0
8	3147	41	0	58	3345	39	0
9	3345	38	1	59	3119	39	0
10	2665	34	0	60	2920	37	0
11	1559	34	0	61	3430	41	0
12	3799	38	0	62	3232	35	0
13	2750	38	0	63	3430	38	0
14	3487	40	0	64	4139	39	0
15	3317	38	0	65	3714	39	0
16	3544	43	1	66	1446	28	1
17	3459	45	0	67	3147	39	1
18	2807	37	0	68	2580	31	0
19	3856	40	0	69	3374	37	0
20	3260	40	0	70	3941	40	0
21	2183	42	1	71	2070	37	0
22	3204	38	0	72	3345	40	0
23	3005	36	0	73	3600	40	0
24	3090	40	1	74	3232	41	0
25	3430	39	0	75	3657	38	1
26	3119	40	0	76	3487	39	0
27	3912	39	0	77	2948	38	0
28	3572	40	0	78	2722	40	0
29	3884	41	0	79	3771	40	0
30	3090	38	0	80	3799	45	0
31	2977	42	0	81	1871	33	0
32	3799	37	0	82	3260	39	0
33	4054	40	0	83	3969	38	0
34	3430	38	1	84	3771	40	0
35	3459	41	0	85	3600	40	0
36	3827	39	0	86	2693	35	1
37	3147	44	1	87	3062	45	0
38	3289	38	0	88	2693	36	0
39	3629	36	0	89	3033	41	0
40	3657	36	0	90	3856	42	0
41	3175	41	1	91	4111	40	0
42	3232	43	1	92	3799	39	0
43	3175	36	0	93	3147	38	0
44	3657	40	1	94	2920	36	0
45	3600	39	0	95	4054	40	0
46	3572	40	0	96	2296	36	0
47	709	25	0	97	3402	38	0
48	624	25	0	98	1871	33	1
49	2778	36	0	99	4167	41	0
50	3572	35	0	100	3402	37	1

```
NCREG2 <- data.frame(ID <- c(1:100),
                     BW <- c(3147, 2977, 3119, 3487, 4111,
                             3572, 3487, 3147, 3345, 2665,
                             1559, 3799, 2750, 3487, 3317,
                             3544, 3459, 2807, 3856, 3260,
                             2183, 3204, 3005, 3090, 3430,
```

```
                        3119, 3912, 3572, 3884, 3090,
                        2977, 3799, 4054, 3430, 3459,
                        3827, 3147, 3289, 3629, 3657,
                        3175, 3232, 3175, 3657, 3600,
                        3572, 709, 624, 2778, 3572,
                        3232, 3317, 2863, 3175, 3317,
                        3714, 2240, 3345, 3119, 2920,
                        3430, 3232, 3430, 4139, 3714,
                        1446, 3147, 2580, 3374, 3941,
                        2070, 3345, 3600, 3232, 3657,
                        3487, 2948, 2722, 3771, 3799,
                        1871, 3260, 3969, 3771, 3600,
                        2693, 3062, 2693, 3033, 3856,
                        4111, 3799, 3147, 2920, 4054,
                        2296, 3402, 1871, 4167, 3402),
                    WK <- c(40, 41, 38, 38, 39, 41, 40, 41, 38, 34,
                        34, 38, 38, 40, 38, 43, 45, 37, 40, 40,
                        42, 38, 36, 40, 39, 40, 39, 40, 41, 38,
                        42, 37, 40, 38, 41, 39, 44, 38, 36, 36,
                        41, 43, 36, 40, 39, 40, 25, 25, 36, 35,
                        38, 40, 37, 37, 40, 34, 36, 39, 39, 37,
                        41, 35, 38, 39, 39, 28, 39, 31, 37, 40,
                        37, 40, 40, 41, 38, 39, 38, 40, 40, 45,
                        33, 39, 38, 40, 40, 35, 45, 36, 41, 42,
                        40, 39, 38, 36, 40, 36, 38, 33, 41, 37),
                    SM <- c(0, 0, 0, 0, 0, 0, 0, 0, 1, 0,
                        0, 0, 0, 0, 0, 1, 0, 0, 0, 0,
                        1, 0, 0, 1, 0, 0, 0, 0, 0, 0,
                        0, 0, 0, 1, 0, 0, 1, 0, 0, 0,
                        1, 1, 0, 1, 0, 0, 0, 0, 0, 0,
                        0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
                        0, 0, 0, 0, 0, 1, 1, 0, 0, 0,
                        0, 0, 0, 0, 1, 0, 0, 0, 0, 0,
                        0, 0, 0, 0, 0, 1, 0, 0, 0, 0,
                        0, 0, 0, 0, 0, 0, 0, 1, 0, 1))
str(NCREG2)
```

```
## 'data.frame':    100 obs. of  4 variables:
##  $ ID....c.1.100.                                                : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ BW....c.3147..2977..3119..3487..4111..3572..3487..3147..3345..  : num  3147 2977 3119 3487 4111 .
##  $ WK....c.40..41..38..38..39..41..40..41..38..34..34..38..38..40..: num  40 41 38 38 39 41 40 41 38
##  $ SM....c.0..0..0..0..0..0..0..0..1..0..0..0..0..0..0..1..0..0..  : num  0 0 0 0 0 0 0 0 1 0 ...
```

```
head(NCREG2)
```

```
##   ID....c.1.100. BW....c.3147..2977..3119..3487..4111..3572..3487..3147..3345..
## 1              1                                                           3147
## 2              2                                                           2977
## 3              3                                                           3119
## 4              4                                                           3487
## 5              5                                                           4111
## 6              6                                                           3572
##   WK....c.40..41..38..38..39..41..40..41..38..34..34..38..38..40..
## 1                                                              40
## 2                                                              41
```

```
## 3                                                                    38
## 4                                                                    38
## 5                                                                    39
## 6                                                                    41
##   SM....c.0..0..0..0..0..0..0..0..1..0..0..0..0..0..0..0..1..0..0..
## 1                                                                   0
## 2                                                                   0
## 3                                                                   0
## 4                                                                   0
## 5                                                                   0
## 6                                                                   0
```

Note that SM appears in R as a numeric variable. We need to convert it into a factor variable.

```
NCREG2$SM.F <- factor(NCREG2$SM,
                      levels = c(0,1),
                      labels = c("Non-smoker", "Smoker"))
str(NCREG2)
```

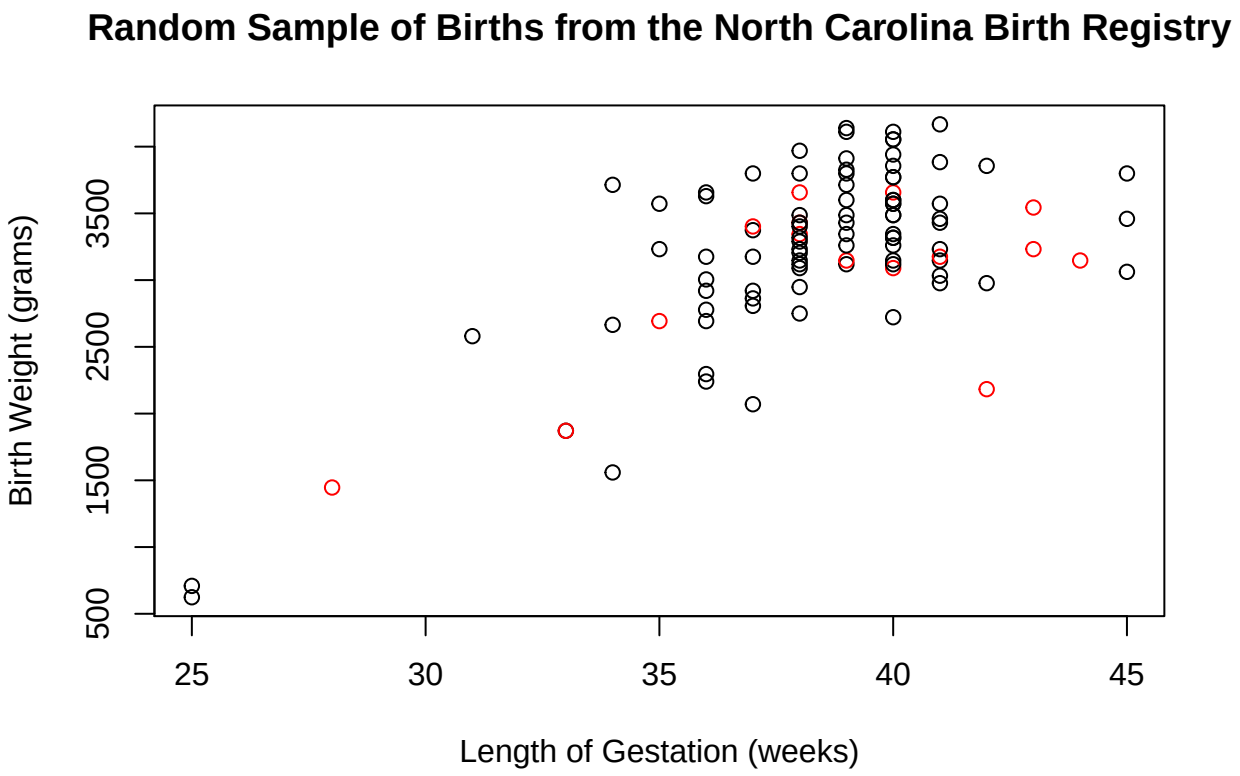
```
## 'data.frame':    100 obs. of  5 variables:
##  $ ID....c.1.100.                                                  : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ BW....c.3147..2977..3119..3487..4111..3572..3487..3147..3345..  : num  3147 2977 3119 3487 4111 .
##  $ WK....c.40..41..38..38..39..41..40..41..38..34..34..38..38..40..: num  40 41 38 38 39 41 40 41 38
##  $ SM....c.0..0..0..0..0..0..0..0..1..0..0..0..0..0..0..1..0..0..  : num  0 0 0 0 0 0 0 0 1 0 ...
##  $ SM.F                                                            : Factor w/ 2 levels "Non-smoker"
```

```
head(NCREG2)
```

```
##   ID....c.1.100. BW....c.3147..2977..3119..3487..4111..3572..3487..3147..3345..
## 1              1                                                           3147
## 2              2                                                           2977
## 3              3                                                           3119
## 4              4                                                           3487
## 5              5                                                           4111
## 6              6                                                           3572
##   WK....c.40..41..38..38..39..41..40..41..38..34..34..38..38..40..
## 1                                                              40
## 2                                                              41
## 3                                                              38
## 4                                                              38
## 5                                                              39
## 6                                                              41
##   SM....c.0..0..0..0..0..0..0..0..1..0..0..0..0..0..0..1..0..0..       SM.F
## 1                                                             0 Non-smoker
## 2                                                             0 Non-smoker
## 3                                                             0 Non-smoker
## 4                                                             0 Non-smoker
## 5                                                             0 Non-smoker
## 6                                                             0 Non-smoker
```

With this extra information, our scatterplot will change considerably.

```
with(NCREG2, plot(WK, BW, col = SM.F,
                  ylab = "Birth Weight (grams)",
                  xlab = "Length of Gestation (weeks)",
                  main = "Random Sample of Births from the North Carolina Birth Registry"))
```



The black circles are mothers who are non-smokers and the red ones are mothers who are smokers.

Let us fit the regression model.

```
NCREG2.LM1 <- lm(BW ~ WK + SM.F, data = NCREG2)
summary(NCREG2.LM1)
```

```
##
## Call:
## lm(formula = BW ~ WK + SM.F, data = NCREG2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1260.45  -265.80     9.26   311.53  1016.59
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1724.42     558.84  -3.086  0.00265 **
## WK            130.05      14.52   8.957 2.39e-14 ***
## SM.FSmoker   -294.40     135.78  -2.168  0.03260 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 484.6 on 97 degrees of freedom
## Multiple R-squared:  0.4636, Adjusted R-squared:  0.4525
## F-statistic: 41.92 on 2 and 97 DF,  p-value: 7.594e-14
```

The estimated equation is $BW = -1724.4 + 130.1WK - 294.4SM.FSmoker$.

Let us try to estimate the birthweight of a baby born after 38 weeks of pregnancy to a mother who smokes. In this case $WK = 38$ and $SM.F = 1$ and the equation becomes $BW = -1724.4 + 130.1 \times 38 - 294.4 \times 1 = 2925$.

In contrast, let us estimate the birthweight of a baby born after 38 weeks of pregnancy, but this time,t he mother is a non-smoker. In this case $WK = 38$ and $SM.F = 0$. The resulting equation is $BW = -1724.4 + 130.1 \times 38 - 294.4 \times 0 = 3219.4$,.

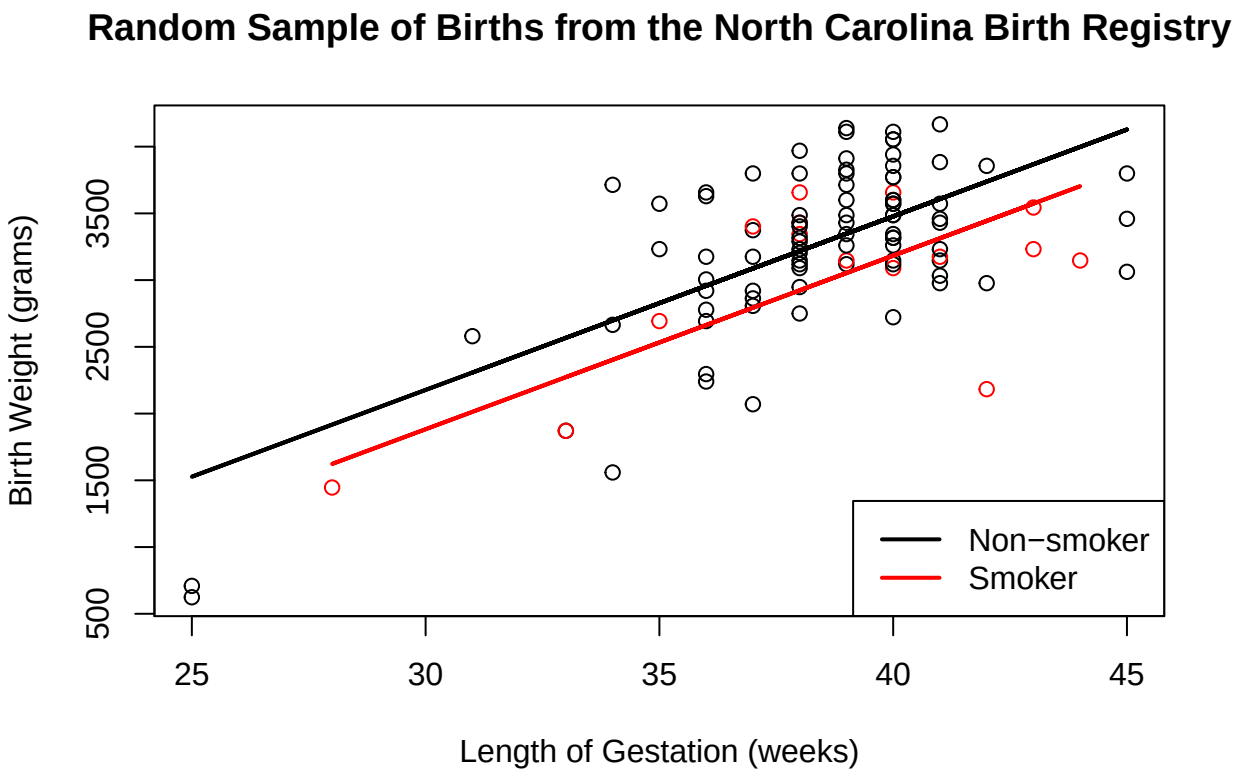
If we hold the length of pregnancy constant, what is the difference between the birthweight of a baby born to a mother who smokes versus a mother who does not smoke? $2925 - 3219.4 = -294.4$. This is the estimate of the slope.

Thus, the interpretation of the slope estimate for the variable `SM.F` proceeds by two stages. First we have to ask what the *reference category* is. In R, the reference category is the level of the dummy variable that is omitted from the estimates. In the present case, R provides estimates for smokers, but omits non-smokers. Thus, the reference category is "non-smokers".

Once we know the reference category, then we can interpret the estimate. The above results show that women who smoke have babies that weigh 294.4 grams less than women who are non-smokers, after adjusting for the length of gestation.

We can visualise the relationship between the two smoking categories as follows

```
with(NCREG2, plot(WK, BW, col = SM.F,
                  ylab = "Birth Weight (grams)",
                  xlab = "Length of Gestation (weeks)",
                  main = "Random Sample of Births from the North Carolina Birth Registry"))
legend("bottomright", legend=c("Non-smoker","Smoker"),
       lty=c(1,1,1), lwd=2, cex=1, col=c("black","red"))
lines(NCREG2$WK[NCREG2$SM.F=="Non-smoker"],
      fitted(NCREG2.LM1)[NCREG2$SM.F=="Non-smoker"],
      lwd=2, col="black")
lines(NCREG2$WK[NCREG2$SM.F=="Smoker"],
      fitted(NCREG2.LM1)[NCREG2$SM.F=="Smoker"],
      lwd=2, col="red")
```



IMPORTANT: There are other types of coding schemes we can use instead of dummy variable coding. These are beyond the scope of this discussion, however.

THE END