STAT 462

Applied Regression Analysis

8.1 - Example on Birth Weight and Smoking

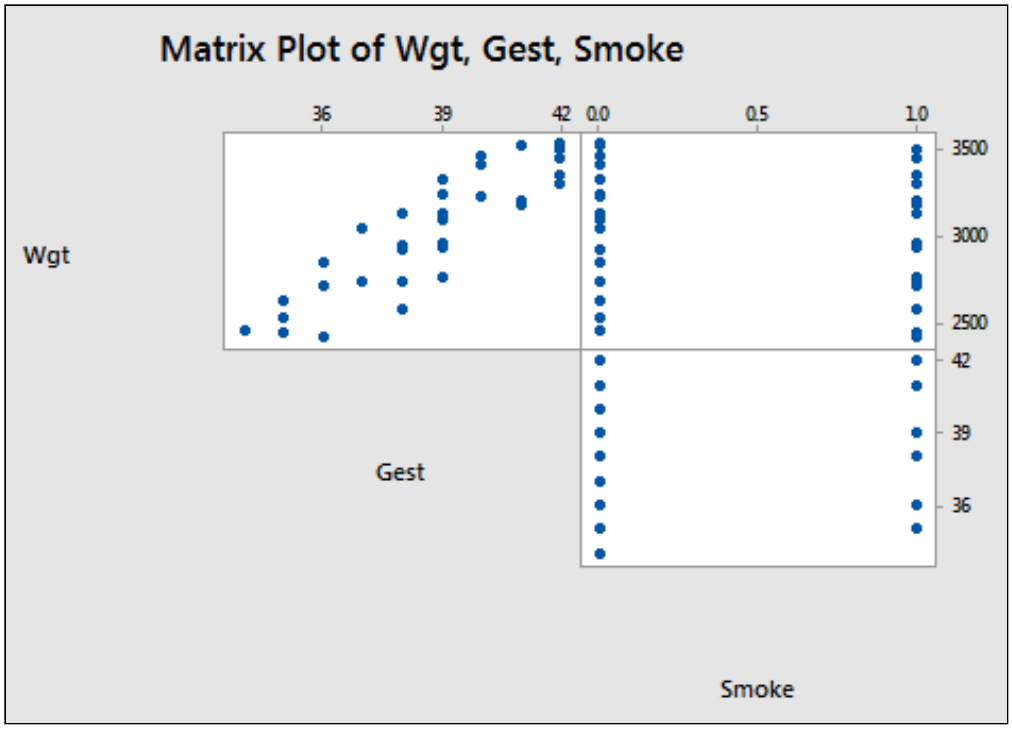
Example: Is a baby's birth weight related to the mother's smoking during pregnancy?

Researchers (Daniel, 1999) interested in answering the above research question collected the following data (birthsmokers.txt (/stat462/sites/onlinecourses.science.psu.edu.stat462/files/data/birthsmokers.txt)) on a random sample of $n = 32$ births:

- Response (y): birth weight (**Weight**) in grams of baby
- Potential predictor (x_1): **Smoking** status of mother (yes or no)
- Potential predictor (x_2): length of gestation (**Gest**) in weeks

The distinguishing feature of this data set is that one of the predictor variables — **Smoking** — is a qualitative predictor. To be more precise, smoking is a "**binary variable**" with only two possible values (yes or no). The other predictor variable (**Gest**) is, of course, quantitative.

The scatter plot matrix:



suggests, not surprisingly, that there is a positive linear relationship between length of gestation and birth weight. That is, as the length of gestation increases, the birth weight of babies tends to increase. It is hard to see if any kind of (marginal) relationship exists between birth weight and smoking status, or between length of gestation and smoking status.

The important question remains — after taking into account length of gestation, is there a significant difference in the average birth weights of babies born to smoking and non-smoking mothers? A **first-order model** with **one binary predictor** and **one quantitative predictor** that helps us answer the question is:

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

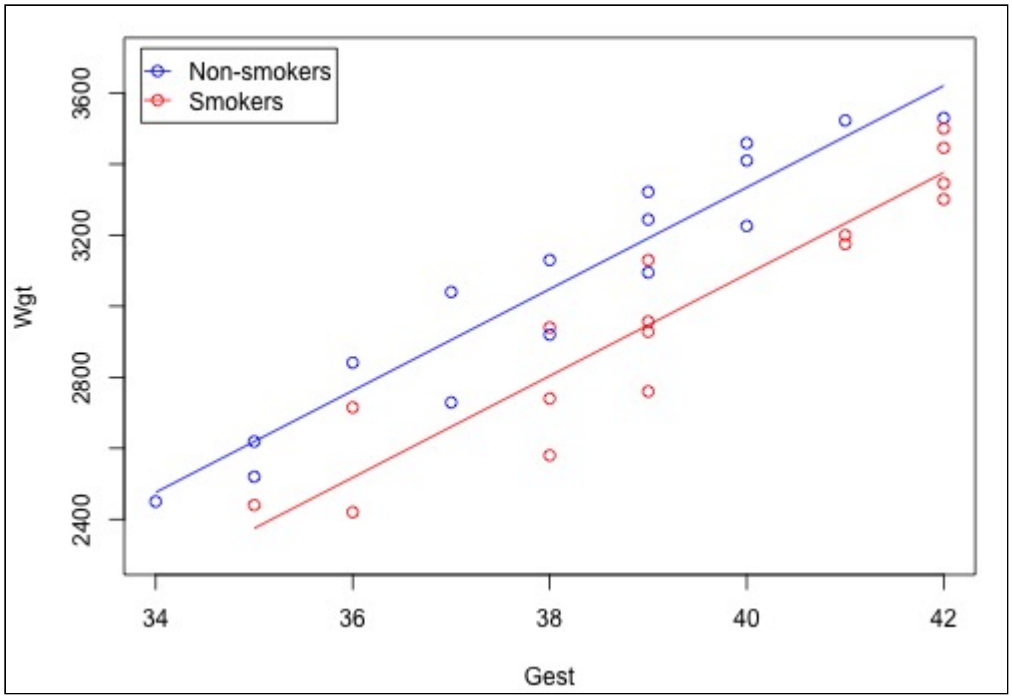
where:

- y_i is the birth weight of baby i
- x_{i1} is length of gestation of baby i
- x_{i2} is a binary variable coded as a 1, if the baby's mother smoked during pregnancy and 0, if she did not

and the **independent** error terms ε_i follow a **normal** distribution with mean 0 and **equal variance** σ^2.

Notice that in order to include a qualitative variable in a regression model, we have to "**code**" the variable, that is, assign a unique number to each of the possible categories. We'll learn more about coding in the remainder of this lesson.

Using the sample data on $n = 32$ births, the plot of the estimated regression function looks like:



The blue circles represent the data on non-smoking mothers (x_2=0), while the red circles represent the data on smoking mothers (x_2=1). And, 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.

At least in this sample of data, 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.

Did you expect the plot of the estimated regression equation to appear as two distinct lines? Let's consider this question. Statistical software tells us that the estimated regression function is:

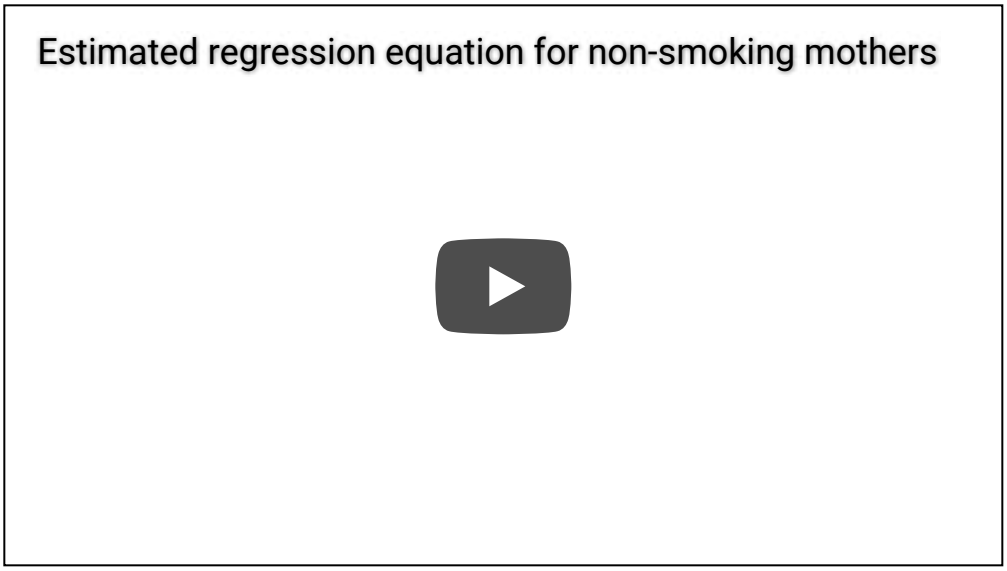
```
Regression Equation

Wgt = -2390 + 143.10 Gest - 244.5 Smoke
```

Therefore, as illustrated in this screencast below, the estimated regression equation for non-smoking mothers (smoking = 0) is:

Weight = - 2390 + 143 Gest

and the estimated regression equation for smoking mothers (when smoking = 1) is:

Weight = - 2635 + 143 Gest



That is, we obtain two different parallel estimated lines (they are parallel because they have the same slope, 143). The difference between the two lines, –245, represents the difference in the average birth weights for a fixed gestation length for smoking and non-smoking mothers in the sample.

How would we answer the following set of research questions? (Do the procedures that appear in parentheses seem appropriate in answering the research question?)

- Is baby's birth weight related to smoking during pregnancy, after taking into account length of gestation? (Conduct a hypothesis test for testing whether the slope parameter for smoking is 0.)
- How is birth weight related to gestation, after taking into account a mother's smoking status? (Calculate and interpret a confidence interval for the slope parameter for gestation.)

Upon analyzing the data, the software output:

STAT 462

Applied Regression Analysis

8.2 - The Basics of Indicator Variables

A "**binary predictor**" is a variable that takes on only two possible values. Here are a few common examples of binary predictor variables that you are likely to encounter in your own research:

- Gender (male, female)
- Smoking status (smoker, nonsmoker)
- Treatment (yes, no)
- Health status (diseased, healthy)
- Company status (private, public)

Example: On average, do smoking mothers have babies with lower birth weight?

In the previous section, we briefly investigated data (birthsmokers.txt (https://onlinecourses.science.psu.edu/stat462/sites/onlinecourses.science.psu.edu.stat462/files/data/birthsmokers.txt)) on a random sample of $n = 32$ births that allow researchers (Daniel, 1999) to answer the above research question. The researchers collected the following data:

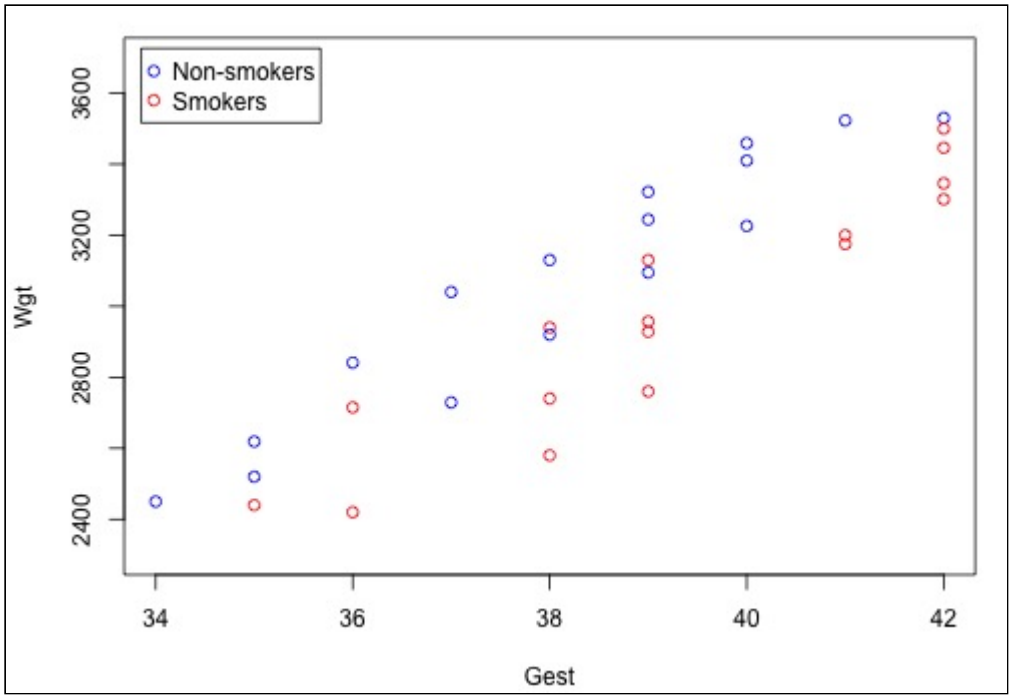
- Response (y): birth weight (**Weight**) in grams of baby
- Potential predictor (x_1): length of gestation (**Gest**) in weeks
- Potential predictor (x_2): **Smoking** status of mother (smoker or non-smoker)

In order to include a qualitative variable in a regression model, we have to "**code**" the variable, that is, assign a unique number to each of the possible categories. A common coding scheme is to use what's called a "**zero-one indicator variable**." Using such a variable here, we code the binary predictor **Smoking** as:

- $x_{i2} = 1$, if mother i smokes
- $x_{i2} = 0$, if mother i does not smoke

In doing so, we use the tradition of assigning the value of 1 to those having the characteristic of interest and 0 to those not having the characteristic. Tradition is less important, though, than making sure you keep track of your coding scheme so that you can properly draw conclusions. Incidentally, other terms sometimes used instead of "**zero-one indicator variable**" are "**dummy variable**" or "**binary variable**".

A scatter plot of the data in which blue circles represent the data on non-smoking mothers ($x_2=0$) and red circles represent the data on smoking mothers ($x_2=1$):



suggests that there might be two distinct linear trends in the data — one for smoking mothers and one for non-smoking mothers. Therefore, a **first order model** with **one binary and one quantitative predictor** appears to be a natural model to formulate for these data. That is:

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

where:

- y_i is birth weight of baby *i* in *grams*
- x_{i1} is the length of gestation of baby *i* in *weeks*
- x_{i2} = 1, if the mother smoked during pregnancy, and x_{i2} = 0, if she did not

and the **independent** error terms ε_i follow a **normal** distribution with mean 0 and **equal variance** σ^2.

How does a model containing a (0,1) indicator variable for two groups yield two distinct response functions? In short, this screencast below, illustrates how the mean response function:

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

yields one regression function for non-smoking mothers (x_{i2} = 0):

$$\mu_Y = \beta_0 + \beta_1 x_{i1}$$

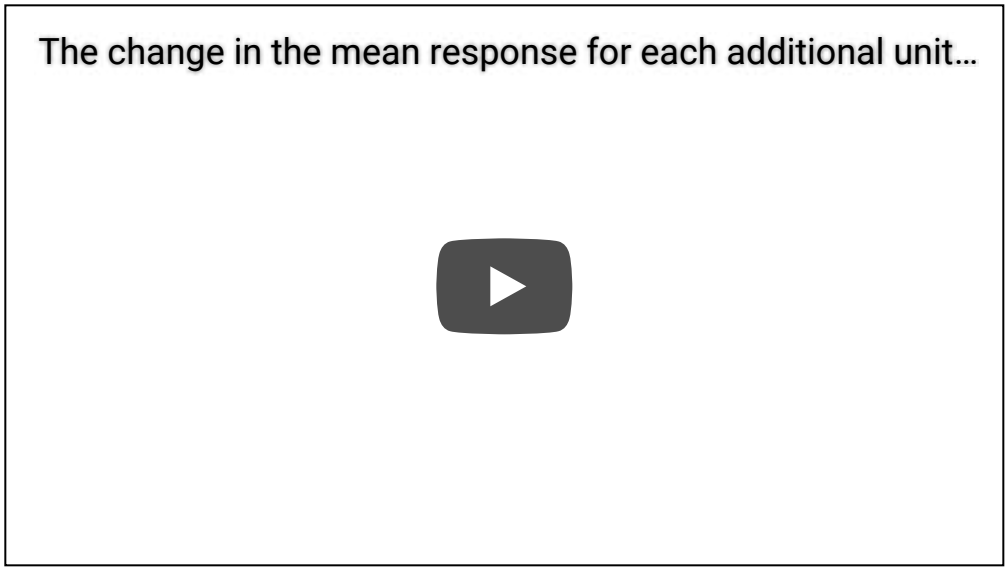
and one regression function for smoking mothers (x_{i2} = 1):

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

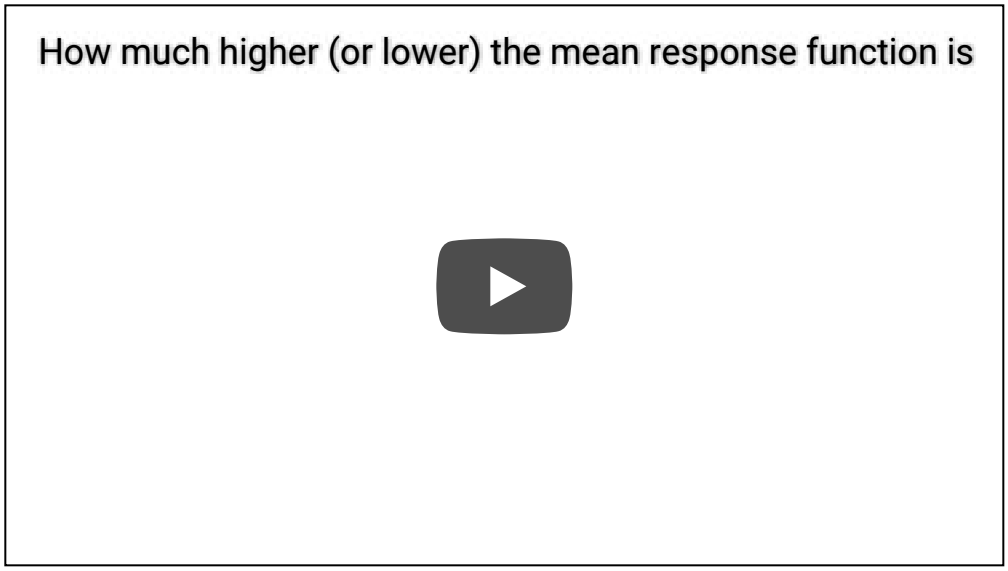
Note that the two formulated regression functions have the same slope (β_1) but different intercepts (β_0 and $\beta_0 + \beta_2$) — mathematical characteristics that, based on the above scatter plot, appear to summarize the trend in the data well.

Now, given that we generally use regression models to answer research questions, we need to figure out how each of the parameters in our model enlightens us about our research problem! The *fundamental principle* is that you can determine the meaning of any regression coefficient by seeing what effect changing the value of the predictor has on the mean response $\boldsymbol{\mu_Y}$. Here's the interpretation of the regression coefficients in a regression model with one (0, 1) binary indicator variable and one quantitative predictor:

- β_1 represents the change in the mean response μ_Y for each additional unit increase in the quantitative predictor x_1 ... for both groups.



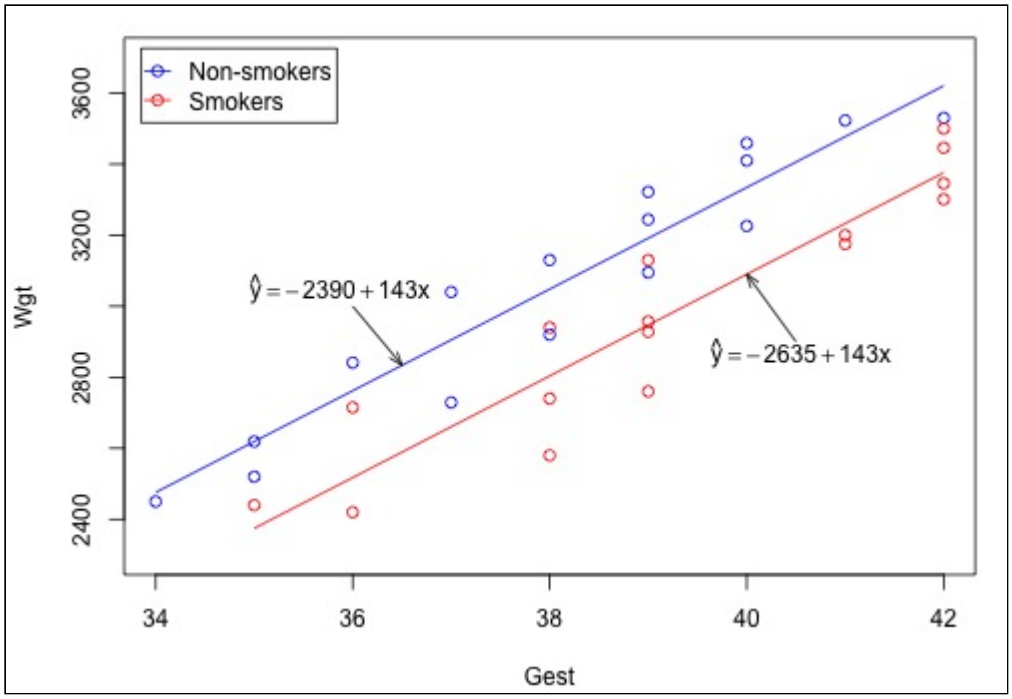
- β_2 represents how much higher (or lower) the mean response function of the second group is than that of the first group... for any value of x_1.



Upon fitting our formulated regression model to our data, statistical software output tells us:

```
Regression Equation

Wgt = -2390 + 143.10 Gest - 244.5 Smoke
```

Unfortunately, this output doesn't precede the phrase "regression equation" with the adjective "estimated" in order to emphasize that we've only obtained an estimate of the actual unknown population regression function. But anyway — if we set **Smoking** once equal to 0 and once equal to 1 — we obtain, as hoped, two distinct estimated lines:



Now, let's use our model and analysis to answer the following research question: *Is there a significant difference in mean birth weights for the two groups, after taking into account length of gestation?* As is always the case, the first thing we need to do is to "translate" the research question into an appropriate statistical procedure. We can show that if the slope parameter β_2 is 0, there is no difference in the means of the two groups — for any length of gestation. That is, we can answer our research question by testing the null hypothesis $H_0 : \beta_2 = 0$ against the alternative $H_A : \beta_2 \neq 0$.

Well, that's easy enough! The software output:

```
Model Summary

      S    R-sq  R-sq(adj)  R-sq(pred)
115.530  89.64%     88.92%      87.60%


Coefficients

Term        Coef  SE Coef  T-Value  P-Value   VIF
Constant   -2390      349    -6.84    0.000
Gest      143.10     9.13    15.68    0.000  1.06
Smoke     -244.5     42.0    -5.83    0.000  1.06


Regression Equation

Wgt = -2390 + 143.10 Gest - 244.5 Smoke
```

reports that the P-value is < 0.001. At just about any significance level, we can reject the null hypothesis $H_0 : \beta_2 = 0$ in favor of the alternative hypothesis $H_A : \beta_2 \neq 0$. 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 babies of all non-smoking mothers, after taking into account length of gestation.

A 95% confidence interval for β_2 tells us the magnitude of the difference. A 95% t-multiplier with n-p = 32-3 = 29 degrees of freedom is $t_{(0.025,\ 29)} = 2.0452$. Therefore, a 95% confidence interval for β_2 is:

$$-244.54 \pm 2.0452(41.98) \text{ or } (-330.4, -158.7).$$

We can be 95% confident that the mean birth weight of smoking mothers is between 158.7 and 330.4 grams *less than* the mean birth weight of non-smoking mothers, for a fixed length of gestation. It is up to the researchers to debate whether or not the difference is a meaningful difference.

‹ 8.1 - Example on Birth Weight and Smoking (/stat462/node/160) up (/stat462/node/86) 8.3 - Two Separate Advantages › (/stat462/node/162)

STAT 462

Applied Regression Analysis

8.3 - Two Separate Advantages

Perhaps somewhere along the way in our most recent discussion, you thought "why not just fit two separate regression functions — one for the smokers and one for the non-smokers?" (If you didn't think of it, I thought of it for you!) Are there advantages to including both the binary and quantitative predictor variables within one multiple regression model? The answer is yes! In this section, we explore the two primary advantages.

The first advantage

An easy way of discovering the first advantage is to analyze the data three times — once using the data on all 32 subjects, once using the data on only the 16 non-smokers, and once using the data on only the 16 smokers. Then, we can investigate the effects of the different analyses on important things such as sizes of standard errors of the coefficients and the widths of confidence intervals. Let's try it!

(birthsmokers_02.txt (/stat462/sites/onlinecourses.science.psu.edu.stat462/files/data/birthsmokers_02.txt))

Here's statistical software output for the analysis using a (0,1) indicator variable and the data on all 32 subjects. Let's just run through the output and collect information on various values obtained:

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	-2390	349	-6.84	0.000	
Gest	143.10	9.13	15.68	0.000	1.06
Smoke	-244.5	42.0	-5.83	0.000	1.06

Regression Equation

Wgt = -2390 + 143.10 Gest - 244.5 Smoke

The standard error of the Gest coefficient is 9.13. Recall that this value quantifies how much the estimated *Gest* coefficient would vary from sample to sample. And, the following output:

```
Variable  Setting
Gest           38
Smoke           1


    Fit   SE Fit        95% CI              95% PI
2803.69  30.8496  (2740.60, 2866.79)  (2559.13, 3048.26)


Variable  Setting
Gest           38
Smoke           0


    Fit   SE Fit        95% CI              95% PI
3048.24  28.9051  (2989.12, 3107.36)  (2804.67, 3291.81)
```

tells us that for mothers with a 38-week gestation, the width of the confidence interval for the mean birth weight is 126.2 for smoking mothers and 118.2 for non-smoking mothers.

Let's do that again, but this time for the output on just the 16 non-smoking mothers:

```
Coefficients

Term       Coef  SE Coef  T-Value  P-Value   VIF
Constant  -2546      457    -5.57    0.000
Gest_0    147.2     12.0    12.29    0.000  1.00


Regression Equation

Wgt_0 = -2546 + 147.2 Gest_0
```

The standard error of the Gest coefficient is 12.0. And:

```
Variable  Setting
Gest_0         38


    Fit   SE Fit        95% CI              95% PI
3047.72  26.7748  (2990.30, 3105.15)  (2811.30, 3284.15)
```

for non-smoking mothers with a 38-week gestation, the width of the confidence interval for the mean birth weight is 114.9.

And, let's do the same thing one more time for the output on just the 16 smoking mothers:

```
Coefficients

Term       Coef  SE Coef  T-Value  P-Value   VIF
Constant  -2475      554    -4.47    0.001
Gest_1    139.0     14.1     9.85    0.000  1.00


Regression Equation

Wgt_1 = -2475 + 139.0 Gest_1
```

The standard error of the Gest coefficient is 14.1. And:

```
Variable  Setting
Gest_1         38


    Fit   SE Fit        95% CI             95% PI
2808.53  35.8088  (2731.73, 2885.33)  (2526.39, 3090.67)
```

for smoking mothers with a 38-week gestation, the length of the confidence interval is 153.6.

Here's a summary of what we've gleaned from the three pieces of output:

Model estimated using…	**SE(Gest)**	**Width of CI for μ_Y**
all 32 data points	9.13	(NS) 118.2 (S) 126.2
16 nonsmokers	12.0	114.9
16 smokers	14.1	153.6

Let's see what we learn from this investigation:

- 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. (This is a good thing!)
- The width of the confidence interval for the mean weight of babies born to smoking mothers is narrower for the estimated model based on all 32 data points (126.2 compared to 153.6), and not substantially different for non-smoking mothers (118.2 compared to 114.9). (Another good thing!)

In short, 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.

The second advantage

An easy way of discovering the second advantage of fitting one "combined" regression function using all of the data is to consider how you'd answer the research question if you broke apart the data and conducted two separate analyses obtaining:

Nonsmokers

```
Coefficients

Term       Coef  SE Coef  T-Value  P-Value   VIF
Constant  -2546      457    -5.57    0.000
Gest_0    147.2     12.0    12.29    0.000  1.00


Regression Equation

Wgt_0 = -2546 + 147.2 Gest_0
```

Smokers

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	-2475	554	-4.47	0.001	
Gest_1	139.0	14.1	9.85	0.000	1.00

Regression Equation

Wgt_1 = -2475 + 139.0 Gest_1

How could you use these results to determine if the mean birth weight of babies differs between smoking and non-smoking mothers, after taking into account length of gestation? Not completely obvious, is it?! It actually could be done with much more (complicated) work than would be necessary if you analyze the data as a whole and fit one combined regression function:

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	-2390	349	-6.84	0.000	
Gest	143.10	9.13	15.68	0.000	1.06
Smoke	-244.5	42.0	-5.83	0.000	1.06

Regression Equation

Wgt = -2390 + 143.10 Gest - 244.5 Smoke

As we previously discussed, answering the research question merely involves testing the null hypothesis $H_0 : \beta_2 = 0$ against the alternative $H_A : \beta_2 \neq 0$. The P-value is < 0.001. 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.

In summary, "pooling" your data and fitting one combined regression function allows you to easily and efficiently answer research questions concerning the binary predictor variable.

‹ 8.2 - The Basics of Indicator Variables (/stat462/node/161) up (/stat462/node/86) 8.4 - Coding Qualitative Variables › (/stat462/node/163)

STAT 462

Applied Regression Analysis

8.4 - Coding Qualitative Variables

In this section, we focus on issues concerning the coding of qualitative variables. In particular, we:

- learn a general rule for the number of indicator variables that are necessary in coding a qualitative variable
- investigate the impact of using a different coding scheme, such as (1, -1) coding, on the interpretation of the regression coefficients

A general rule for coding a qualitative variable

In the birth weight example, we coded the qualitative variable **Smoking** by creating a (0, 1) indicator variable that took on the value 1 for smoking mothers and 0 for non-smoking mothers. What if we had instead tried to use **two** indicator variables? That is, what if we created one (0, 1) indicator variable, x_{i2} say, for smoking mothers defined as:

- $x_{i2} = 1$, if mother smokes
- $x_{i2} = 0$, if mother does not smoke

and one (0, 1) indicator variable, x_{i3} say, for non-smoking mothers defined as:

- $x_{i3} = 1$, if mother does not smoke
- $x_{i3} = 0$, if mother smokes

In this case, our modified regression function with **two binary predictors** and **one quantiative predictor** would be:

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

where:

- μ_Y is the mean birth weight for given predictors
- x_{i1} is the length of gestation of baby i
- $x_{i2} = 1$, if mother smokes and $x_{i2} = 0$, if mother does not
- $x_{i3} = 1$, if mother does not smoke and $x_{i3} = 0$, if mother smokes

Do you see where this is going? Let's see what the implication of such a coding scheme is on the data analysis:

```
Regression Analysis: Weight versus Gest, x2*, x3*

* x3* is highly correlated with other X variables
* x3* has been removed from the equation

The regression equation is
Weight = - 2390 + 143 Gest - 245 x2*

Predictor        Coef     SE Coef          T        P
Constant      -2389.6       349.2      -6.84    0.000
Gest          143.100       9.128      15.68    0.000
x2*           -244.54       41.98      -5.83    0.000

S = 115.5       R-Sq = 89.6%     R-Sq(adj) = 88.9%
```

As you can see **in blue**, the statistical software has problems with fitting the model. This is not a problem unique to the software used here (Minitab) — any statistical software would have problems. At issue is that the indicator variable x_3 is "highly correlated" with the indicator variable x_2. In fact, x_2 and x_3 are perfectly correlated with one another — when x_2 is 1, x_3 is always 0 and when x_2 is 0, x_3 is always 1. (Described more technically, the columns of the $\boldsymbol{X}$ matrix are linearly dependent — if you add the x_2 and x_3 columns you get the column of 1's for the intercept term.) As you can see ("x3 has been removed from the equation"), the software attempts to fix the problem for us by dropping from the model the last predictor variable listed.

How do we prevent such problems from occurring when coding a qualitative variable? The short answer is to always create one fewer indicator variable than the number of groups defined by the qualitative variable. That is, in general, a qualitative variable defining c groups should be represented by c - 1 indicator variables, each taking on values 0 and 1. For example:

- 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.

Then, choose one group or category to be the "reference" group (often it will be clear from the application which group should be the reference group, such as a control group in a medical experiment, but, if not, then the group with the most observations is often a good choice; if all groups are the same size and there is no obvious reference group then simply select the most convenient 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. For examples of how this works in practice, see the three-group examples in Section 6.1, where "no cooling" is the reference group, and Section 8.6, where treatment C is the reference group.

The impact of using a different coding scheme

In the birth weight example, we coded the qualitative variable **Smoking** by creating a (0, 1) indicator variable that took on the value 1 for smoking mothers and 0 for non-smoking mothers. What if we had instead used (1, -1) coding? That is, what if we created a (1, -1) indicator variable, x_{i2} say, defined as:

- x_{i2} = 1, if the mother smokes
- x_{i2} = -1, if mother does not smoke

In this case, our modified regression function using a (1, -1) coding scheme would be:

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

where:

- y_i is the birth weight of baby i
- x_{i1} is the length of gestation of baby i
- x_{i2} = 1, if the mother smokes and x_{i2} = -1, if the mother does not smoke

The mean response function:

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

yields two different response functions. If the mother is a non-smoker, then x_{i2} = -1 and the mean response function looks like:

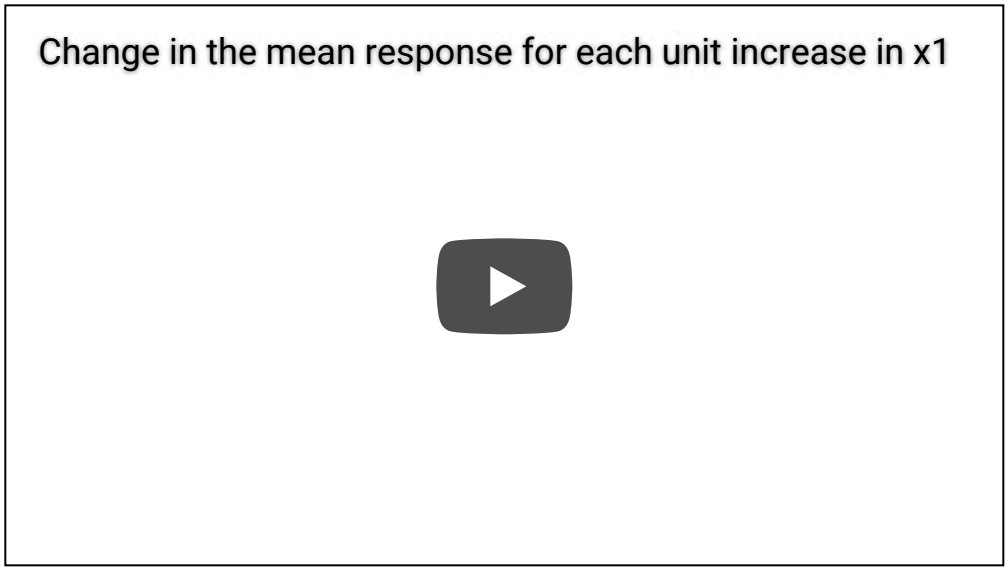
$$\mu_Y = (\beta_0 - \beta_2) + \beta_1 x_{i1}$$

while if the mother is a smoker, then x_{i2} = 1 and the mean response function looks like:

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

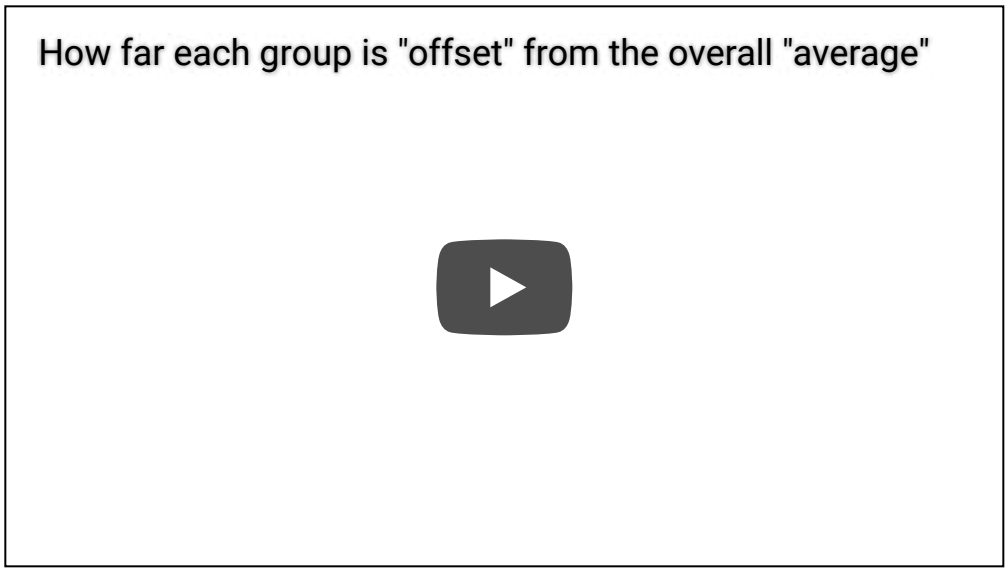
Recall that the *fundamental principle* is that you can determine the meaning of any regression coefficient by seeing what effect changing the value of the predictor has on the mean response μ_Y. Here's the interpretation of the regression coefficients in a regression model with one (1, -1) binary indicator variable and one quantitative predictor, as well as an illustration of how the meaning was determined:

- β_1 represents the change in the mean response μ_Y for each additional unit increase in the quantitative predictor x_1 for both groups. Note that the interpretation of this slope parameter is no different than the interpretation when using (0, 1) coding.



- β_0 represents the overall "average" intercept ignoring group.

- β_2 represents how far each group is "offset" from the overall "average"



Using the (1, −1) coding scheme, the software tells us:

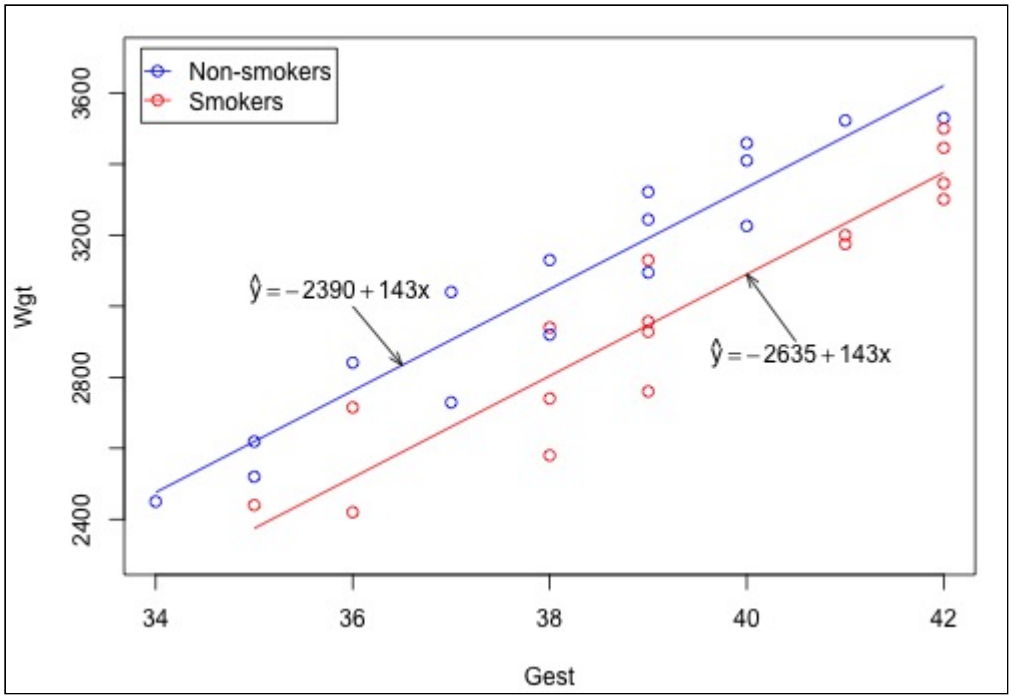
```
The regression equation is
Weight = - 2512 + 143 Gest - 122 Smoking2
```

The estimate of the smoking parameter, -122, tells us that each group is "offset" from the overall "average" by 122 grams. And, if each group is offset from the overall average by 122 grams, then the estimate of the difference in the mean weights of the two groups must be 122+122 or 244 grams. Recall that the estimated smoking coefficient obtained when using the (0, 1) coding scheme was -245:

```
The regression equation is
Weight = - 2390 + 143 Gest - 245 Smoking
```

telling us that the difference in the mean weights of the two groups is 245 grams. Makes sense? (The fact that the estimated coefficient is -245 rather than -244 is just due to rounding.)

If we set **Smoking2** once equal to -1 and once equal to 1, we obtain the same two distinct estimated lines:



In short, 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. That's why it is fundamentally critical that you not only keep track of how you code your qualitative variables, but also can figure out how your coding scheme impacts the interpretation of the regression coefficients. Furthermore, when reporting your results, you should make sure you explain the coding scheme you used. And, when interpreting others' results, you should make sure you know what coding scheme they used!

‹ 8.3 - Two Separate Advantages (/stat462/node/162) up (/stat462/node/86) 8.5 - Additive Effects › (/stat462/node/164)