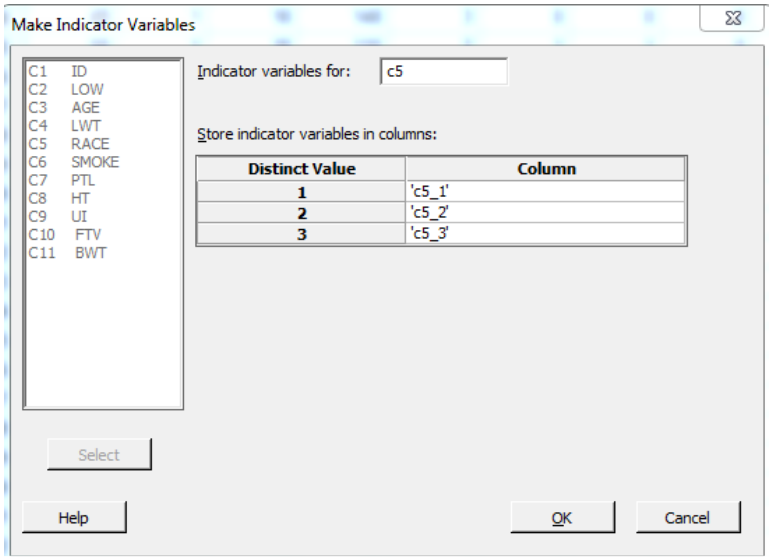
The low birth weight set has information on 189 babies. These are the variables:

ID	Identification number We won't use this here.
LOW	Binary, with 0 = low birth weight (< 2,500 g) 1 = OK birth weight (≥ 2,500 g) We won't use this here.
AGE	Mother's age in years
LWT	Mother's last weight (pounds) before becoming pregnant
RACE	Coded as 1 = white, 2 = African-American, 3 = other
SMOKE	Binary, with 0 = no, 1 = yes
PTL	Number of previous premature labors
HT	Binary, with 0 = no hypertension (high blood pressure), 1 = yes
UI	Binary, with 0 = no uterine irritability, 1 = yes
FTV	Number of physician visits during first trimester
BWT	Birth weight in grams

The objective is to explain BWT in terms of the other variables.

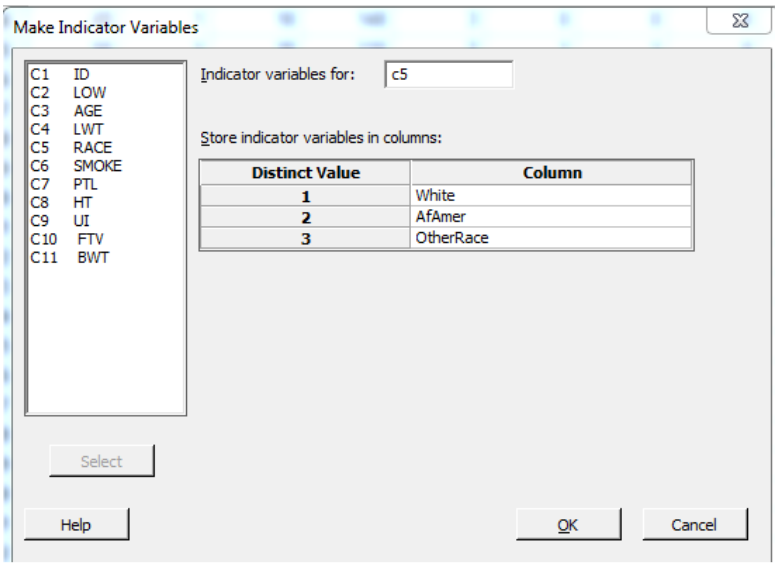
Our first problem is that the 1, 2, 3 coding of RACE is artificial. We will replace this three-level categorical variable with three binary variables. Minitab can do this through **Calc** ⇒ **Make Indicator Variables**. You'll get this panel:



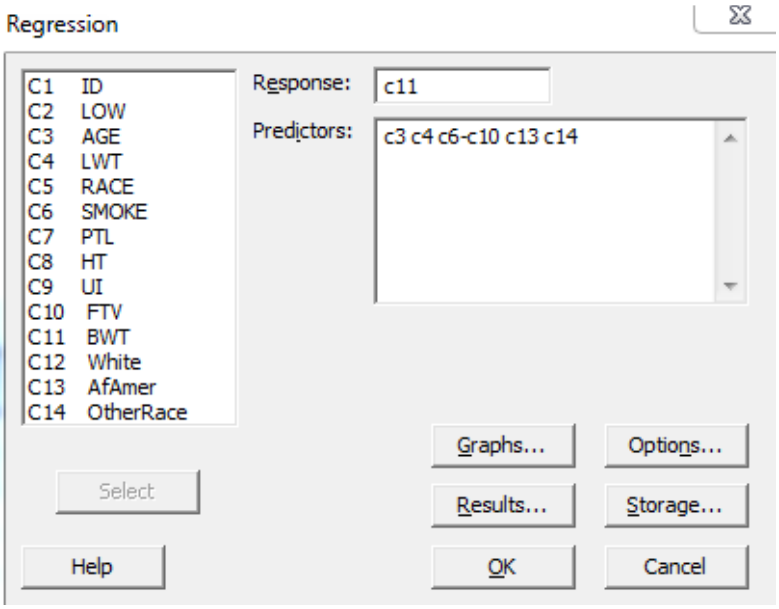
The new binary variables will appear in new columns. Minitab will assign names to these columns in obvious ways. For example, the new column c5_2 will have values

1 at those positions in which RACE = 2
0 at those positions in which RACE ≠ 2

You can rename these columns after they are created. You can also rename them here. Revise the panel as this:



Then ask for this regression:

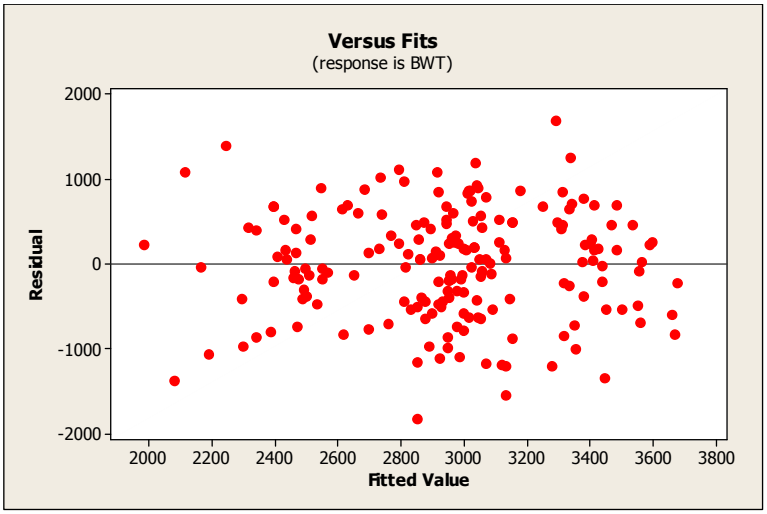


Notice that C5 is not used. For the race indicators, we can use any two of the three. If we specify all three, Minitab will throw out the last one named.

> In general, if a categorical variable has J levels, we only need $J - 1$ binary variables to make a unique identification.

For convenience, we omitted the White indicator; this was completely arbitrary.

Here's the residual versus fitted plot. There are no problems with this.



Here is the regression detail:

Regression Analysis: BWT versus AGE, LWT, ...

```
The regression equation is
BWT = 2928 - 3.66 AGE + 4.37 LWT - 351 SMOKE - 49 PTL - 594 HT - 515 UI
      - 14.1 FTV - 489 AfAmer - 357 OtherRace

Predictor    Coef  SE Coef      T      P
Constant   2927.7    312.8   9.36  0.000
AGE        -3.657    9.616  -0.38  0.704
LWT         4.373    1.734   2.52  0.013
SMOKE      -350.7    106.4  -3.29  0.001
PTL         -48.6    101.9  -0.48  0.634
HT         -594.4    202.3  -2.94  0.004
UI         -514.8    138.8  -3.71  0.000
FTV        -14.10    46.45  -0.30  0.762
AfAmer     -489.5    149.9  -3.27  0.001
OtherRace  -356.7    114.7  -3.11  0.002

S = 650.070   R-Sq = 24.3%   R-Sq(adj) = 20.5%

Analysis of Variance
Source           DF        SS       MS     F      P
Regression        9  24273167  2697019  6.38  0.000
Residual Error  179  75643886   422592
Total           188  99917053
```

The Seq SS section is omitted here. This lists contributions to $SS_{\text{Regression}}$ for the variables in the order in which they are named. This is not helpful at this point.

```
Unusual Observations
Obs    AGE     BWT     Fit   SE Fit  Residual  St Resid
 68   28.0  3303.0  3126.5   337.5     176.5      0.32 X
 94   25.0  3637.0  2240.3   284.9    1396.7      2.39RX
130   45.0  4990.0  3286.9   216.3    1703.1      2.78R
131   28.0   709.0  2079.2   176.0   -1370.2     -2.19R
132   29.0  1021.0  2847.1   166.4   -1826.1     -2.91R
133   34.0  1135.0  2186.5   261.2   -1051.5     -1.77 X
136   27.0  1588.0  3128.2   113.8   -1540.2     -2.41R
155   24.0  2100.0  3443.4    99.9   -1343.4     -2.09R
R denotes an observation with a large standardized residual.
X denotes an observation whose X value gives it large leverage.
```

So . . . what do we think about this?

There are six points noted with R. This is perfectly reasonable, and none has St Resid below -3 or above +3.

The three points with X should be examined. There may be interesting stories for those points.

This regression is statistically significant, but it's very disappointing. Why?