where n_h is the number of points selected from stratum h, $W_h = \frac{N_h}{N}$ is the proportion of stratum h in the population, and σ_h is the standard deviation of the response variate (in this case, weight) in stratum h. In this case, we do not know the true value, so we must use an estimate for the standard deviation. Since there are multiple categorical variables, optimal allocation has to be used for each categorical variable. Suppose that the desired size of the training set is n, and that the number of categorical variables is r. Optimal allocation was used to sample $\frac{n}{r}$ for each categorical variable $x_1, \ldots, x_r$ (yielding a total training set size of n as desired).

Selection of Candidate Models (Automatic and Manual)

Model 1 (automatic selection)

In the previous stage (pre-fitting diagnostics), we were able to reduce the set of covariates to consider when building the model. The set which we are now considering is: gestation, parity, time, number, smoke, mother's age, mother's weight, mother's height, body-mass index, gestation-parity interaction, gestation-mage interaction, income, and mother's ethnicity. Using this set of variables, we will run forward, backwards, and step-wise selection to obtain 3 different models.

We compare these three models using their respective MSPEs (calculated after training the model on the train set).

```
## lm(formula = wt ~ gestation + smoke + mht + meth + I(gestation *
##     parity) + I(mwt * 703/(mht)^2) + parity + mwt, data = birth.data,
##     subset = train.ind)
## [1] 115024.5
```

```
## lm(formula = wt ~ mage + I(mwt * 703/(mht)^2) + smoke + mht +
##     meth + I(gestation * mage) + I(gestation * parity), data = birth.data,
##     subset = train.ind)
## [1] 109848.6
```

```
## lm(formula = wt ~ parity + time + mage + mht + I(mwt * 703/(mht)^2) +
##     I(gestation * parity) + I(gestation * mage) + meth, data = birth.data,
##     subset = train.ind)
## [1] 115157.1
```

The model obtained from stepwise selection has the lowest MSPE so we will select it as one of our candidate models. We will first begin by analyzing the significance of each covariate in the presence of other covariates. For brevity's sake, only the non-categorical variables are shown below.

```
##                            Estimate   Std. Error     t value     Pr(>|t|)
## (Intercept)             2.748489809 17.237324926   0.1594499 8.733635e-01
## mage                   -4.509967490  0.416268998 -10.8342622 2.821334e-25
## I(mwt * 703/(mht)^2)    0.509400476  0.194914428   2.6134570 9.168672e-03
## mht                     1.525102565  0.250294054   6.0932433 1.886433e-09
## I(gestation * mage)     0.016491892  0.001442051  11.4364112 9.598508e-28
## I(gestation * parity)   0.003020828  0.001428050   2.1153519 3.477629e-02
```

For the non-categorical variables, namely mother's age, mother's height, BMI (the body-mass index which is $\left(\frac{703 \times mwt_i}{mht_i^2}\right)$) the gestation-mage interaction variable and the gestation-parity interaction variable, a t-test with the null hypothesis $H_0 : \beta_i = 0$ can be used. The corresponding p-values are all strictly less than 0.05, so at the 0.05 significance level, the null hypothesis can be rejected. Hence, each of mother's age, mother's height, BMI, the gestation-mage interaction variable and the gestation-parity interaction variable are significant in the presence of the other covariates.

For the categorical variables, namely smoke and mother's ethnicity, we must use an F-test to determine if the categorical variables are significant in the presence of the other covariates.

```
## Analysis of Variance Table
##
## Model 1: wt ~ mage + I(mwt * 703/(mht)^2) + mht + meth + I(gestation *
##     mage) + I(gestation * parity)
## Model 2: wt ~ mage + I(mwt * 703/(mht)^2) + smoke + mht + meth + I(gestation *
##     mage) + I(gestation * parity)
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    659 170275
## 2    656 162084  3    8191.2 11.051 4.382e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p value is 4.382×10^{-7}, so at the 0.05 significance level, we can conclude that the categorical variable smoke is significant in the presence of the other covariates.

```
## Analysis of Variance Table
##
## Model 1: wt ~ mage + I(mwt * 703/(mht)^2) + mht + smoke + I(gestation *
##     mage) + I(gestation * parity)
## Model 2: wt ~ mage + I(mwt * 703/(mht)^2) + smoke + mht + meth + I(gestation *
##     mage) + I(gestation * parity)
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
## 1    661 172043
## 2    656 162084  5    9958.7 8.0611 2.16e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p value is 2.16×10^{-7}, so at the 0.05 significance level, we can conclude that the categorical variable mother's ethnicity is significant in the presence of the other covariates. Thus, in this particular model, all covariates are significant in the presence of each other.

Model 2 (Manual)

The following model was created manually using the evaluation done in the previous stage and intuition. It combines income, smoke and number (which measures the magnitude of the mother's smoking habit) with other biological factors such as gestation, parity, and the mother's height (which was found to be selected frequently by automatic model selection even when different combinations of variables were used). For brevity's sake, only the non-categorical variables are shown below.

```
##                          Estimate  Std. Error   t value     Pr(>|t|)
## (Intercept)           -79.0281462 21.39450487 -3.693853 2.395507e-04
## gestation               0.3921665  0.05445977  7.201031 1.665261e-12
## mht                     1.4121580  0.25418283  5.555678 4.042965e-08
## I(gestation * parity)   0.0360373  0.01623026  2.220376 2.673891e-02
## parity                 -9.2947283  4.47561147 -2.076750 3.821870e-02
```

For the non-categorical variables, namely gestation, mother's height, parity and the gestation-parity interaction variable, a t-test with the null hypothesis $H_0 : \beta_i = 0$ can be used. The corresponding p-values are all strictly less than 0.05, so at the 0.05 significance level, the null hypothesis can be rejected. Hence, each of gestation, mother's height, parity, and the gestation-parity interaction covariate are significant in the presence of the other covariates.

```
## Analysis of Variance Table
##
## Model 1: wt ~ gestation + mht + I(gestation * parity) + parity + number +
##     income
## Model 2: wt ~ gestation + mht + smoke + I(gestation * parity) + parity + 
##     number + income
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)   
## 1    649 172235                                
## 2    647 169747  2    2488.4 4.7424 0.009023 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The *p* value is 0.009, so at the 0.05 significance level, the null hypothesis that smoke is insignificant in the presence of other covariates can be rejected. Hence, the categorical variable smoke is significant in the presence of the other covariates.

```
## Analysis of Variance Table
## 
## Model 1: wt ~ gestation + mht + smoke + I(gestation * parity) + parity + 
##     income
## Model 2: wt ~ gestation + mht + smoke + I(gestation * parity) + parity + 
##     number + income
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    654 172028                           
## 2    647 169747  7    2281.3 1.2422 0.2772
```

The *p* value is 0.28, so at the 0.05 significance level, the null hypothesis that number is insignificant in the presence of other covariates cannot be rejected. So we accept the null hypothesis that number is insignificant in the presence of the other covariates

```
## Analysis of Variance Table
## 
## Model 1: wt ~ gestation + mht + smoke + I(gestation * parity) + parity + 
##     number
## Model 2: wt ~ gestation + mht + smoke + I(gestation * parity) + parity + 
##     number + income
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    655 171436                           
## 2    647 169747  8    1689.2 0.8048 0.5985
```

The *p* value is 0.60, so at the 0.05 significance level, the null hypothesis that income is insignificant in the presence of other covariates cannot be rejected. So we accept the null hypothesis that number is insignificant in the presence of the other covariates. So we further refine the model by removing income and number.

Refined Model 2

```
##                          Estimate  Std. Error   t value     Pr(>|t|)
## (Intercept)           -74.2265183 21.12198303 -3.514183 4.711590e-04
## gestation               0.3852541  0.05363568  7.182794 1.843752e-12
## mht                     1.3660556  0.24823798  5.503008 5.341820e-08
## I(gestation * parity)   0.0382404  0.01599552  2.390694 1.709499e-02
## parity                 -9.8405138  4.41234785 -2.230222 2.606718e-02
```

Since the p-values corresponding to the non-categorical variables are all strictly less than 0.05, at the 0.05 significance level, the null hypothesis can be rejected. Hence, each of gestation, mother's height, parity, and the gestation-parity interaction covariate are significant in the presence of the other covariates in the refined model.

```
## Analysis of Variance Table
## 
## Model 1: wt ~ gestation + mht + I(gestation * parity) + parity
## Model 2: wt ~ gestation + mht + smoke + I(gestation * parity) + parity
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    665 182741
## 2    662 173703  3      9038 11.482 2.402e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p value is 2.402×10^{-7}, so at the 0.05 significance level, the null hypothesis that smoke is insignificant in the presence of other covariates can be rejected. Hence, the categorical variable smoke is significant in the presence of the other covariates.
To summarize, the two models we will proceed to compare are:

Model 1:

```
##           (Intercept)                  mage  I(mwt * 703/(mht)^2)
##           2.748489809          -4.509967490           0.509400476
##              smokenow  smokeuntil pregnancy           smokeused to
##          -7.323507616           0.802088003          -1.041515927
##                   mht             methAsian        methCaucasian
##           1.525102565           3.739051306           7.925141013
##           methMexican             methMixed             methOther
##          19.735897417           7.997043034           7.052603812
##   I(gestation * mage) I(gestation * parity)
##           0.016491892           0.003020828
```

Model 2:

```
##           (Intercept)             gestation                   mht
##           -74.2265183             0.3852541             1.3660556
##              smokenow  smokeuntil pregnancy          smokeused to
##            -7.6680553             0.3658084            -0.8063291
## I(gestation * parity)                parity
##             0.0382404            -9.8405138
```