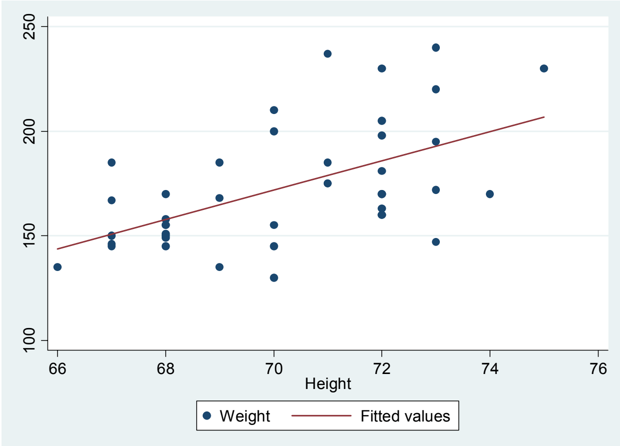
Lack of Fit Test Example Using Male Weight and Height Data

Data represent a sample of n = 43 college males measured at c = 10 different heights. There are multiple weight observations at most of the heights, which are measured to the nearest inch. Here is a plot of the data with the regression line shown, and the R Commands used to generate the test for lack of fit:



R Commands; Dataset is called HtWt and the two variables are named Height and Weight:

```
> Reduced <- lm(Weight~Height, data=HtWt)
> Full <- lm(Weight ~ 0 + as.factor(Height),
data = HtWt)
> anova(Reduced, Full)
```

> #The next command is not needed, but used to show the connection with the usual SSE and anova table:

```
> anova(Reduced)
```

R Commands and Results, with lack of fit test results in bold:

```
> Reduced <- lm(Weight ~ Height, data = HtWt)
> Full <- lm(Weight ~ 0 + as.factor(Height), data = HtWt)
> anova (Reduced, Full)
Analysis of Variance Table

Model 1: Weight ~ Height                [This is the reduced model]
Model 2: Weight ~ 0 + as.factor(Height) [This is the full model]

  Res.Df      RSS   Df    Sum of Sq        F     Pr(>F)
1     41    23617
2     33    20712    8       2905.1   0.5786     0.7879

> #The following is not needed, but shows the connection with regression.
> anova (Reduced) [This shows the anova table for the usual regression]
Analysis of Variance Table
Response: Weight
          Df Sum Sq Mean Sq F value    Pr(>F)
Height     1  11277   11277  19.578 6.978e-05 ***
Residuals 41  23617     576
```

Notes:

- Model 1 is the usual linear regression model, which is the *reduced* model in this case; SSE(R) = 23617
- Model 2 is telling R to consider Height as a "Factor" instead of a continuous variable, thus treating it as categorical and fitting the mean at each height. SSE(F) = 20712 = SSE(PE) [F = Full, PE = pure error]
- The Lack of Fit SSE is SSE(LF) = SSE(R) – SSE(F) = 23617 – 20712 = 2905
- $F^* = \dfrac{2905.1/8}{20712/33} = 0.5786$. Do not reject H_0 because the *p*-value = .7879.
- Conclude that using the linear regression is almost as good as using separate means at the 10 heights. Advantage of linear regression is that we can predict even for heights in between those measured.