- Rearranging the terms we can see (3.33) that the relationship between sales and TV is dependent on radio:

rearranging terms we can see (3.33) that the relationship between sales and TV is dependent on radio

$$Y = \beta_0 + \underbrace{(\beta_1 + \beta_3 X_2)}_{\tilde{\beta}_1} X_1 + \beta_2 X_2 + \varepsilon$$

Interpretation of beta_3

- Beta_3 is the increase in effectiveness of TV spending associated with a one-unit increase in radio spending (or vice-versa).

β_3 is the slope multiplied with the interaction term

Interpretation

```
Call:
lm(formula = sales ~ TV * radio, data = Advertising)

Residuals:
    Min      1Q  Median      3Q     Max 
-6.3366 -0.4028  0.1831  0.5948  1.5246 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 6.750e+00  2.479e-01  27.233   <2e-16 ***
TV          1.910e-02  1.504e-03  12.699   <2e-16 ***
radio       2.886e-02  8.905e-03   3.241   0.0014 ** 
TV:radio    1.086e-03  5.242e-05  20.727   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9435 on 196 degrees of freedom
Multiple R-squared:  0.9678,    Adjusted R-squared:  0.9673 
F-statistic:  1963 on 3 and 196 DF,  p-value: < 2.2e-16
```

H_0 given intercept TV, radio $\beta_3 = 0$

H_1: given intercept TV, radio $\beta_3 \neq 0$

we should include term b/c p-value is very small

Hierarchical principle

- If we include an interaction in a model, we should also include the main effects, even if the p-values associated with their coefficients are not significant.
- The rationale (as on textbook):
- If X1*X2 is related to the response, then whether or not the coefficients of X1 or X2 are exactly zero is of little interest.
- Also X1*X2 is typically correlated with X1 and X2, and so leaving them out tends to alter the meaning of the interaction.

example cant take out radio if we are
interacting with radio.

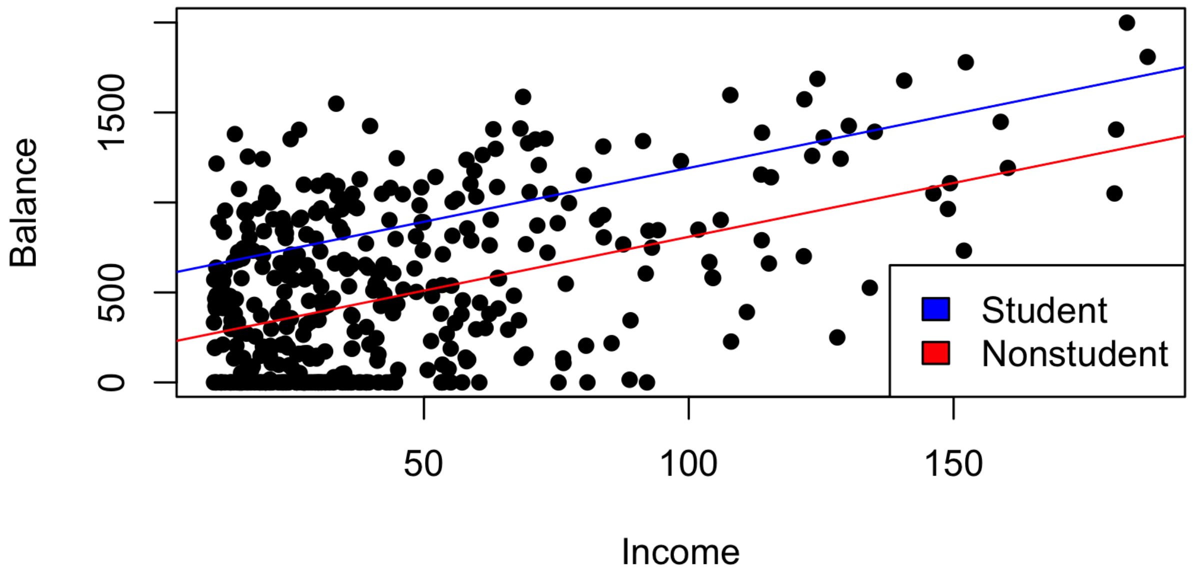
Qualitative and quantitative variable interaction

- Consider the Credit data set
- We wish to predict balance using the income (quantitative) and student (qualitative) variables.
- In the absence of an interaction term (like before), the model takes the form (3.34)

$$\text{balance}_i = \beta_1 \text{income}_i + \begin{cases} \beta_0 + \beta_2 + \varepsilon, & i\text{-th person is student} \\ \beta_0 + \varepsilon, & i\text{-th person is non-student} \end{cases}$$

- The lines for students and non-students have different intercepts, beta_0+beta_2 versus beta_0, but have the same slope, beta_1
- What does this imply?

The unit increase in balance vs income is the same for students and nonstudents



- This limitation can be addressed by adding an interaction variable
- We can multiply income with the indicator variable student.
- Our model is now (3.35):

$$\text{balance}_i = \beta_0 + \beta_1 \cdot \text{income}_i + \beta_2 \cdot \text{student}_i + \beta_3 \cdot \text{income}_i \cdot \text{student}$$

$$= \begin{cases} (\beta_0 + \beta_2) + (\beta_1 + \beta_3) \cdot \text{income}_i + \varepsilon_i & \text{, if student} \\ \beta_0 + \beta_1 \cdot \text{income}_i + \varepsilon_i & \text{, non student} \end{cases}$$

Interpretation of the 4 coefficients

```
Call:
lm(formula = Balance ~ Income + Student + Income:Student, data = Credit)

Residuals:
    Min      1Q  Median      3Q     Max
-773.39 -325.70  -41.13  321.65  814.04

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       200.6232    33.6984   5.953 5.79e-09 ***
Income              6.2182     0.5921  10.502  < 2e-16 ***
StudentYes        476.6758   104.3512   4.568 6.59e-06 ***
Income:StudentYes  -1.9992     1.7313  -1.155    0.249
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 391.6 on 396 degrees of freedom
Multiple R-squared:  0.2799,    Adjusted R-squared:  0.2744
F-statistic:  51.3 on 3 and 396 DF,  p-value: < 2.2e-16
```

Interaction

that's high?

$\hat{\beta}_0 = 200.6232$

Averge balance for non students

$\hat{\beta}_1 = 6.2182$

Averarge increase in balance for an increase $1 in income non students

$\hat{\beta}_2 = 476.6758$

Averge difference in balance between studens & no students when no income

$\hat{\beta}_3 = -1.9992$

Averaye difference in slope (balance v income) between studends & non students

Interpretation of the 4 coefficients

H0: $\beta_3 = 0$

H1 $\beta_3 \neq 0$

pval = .244

β3 is

the coefficient

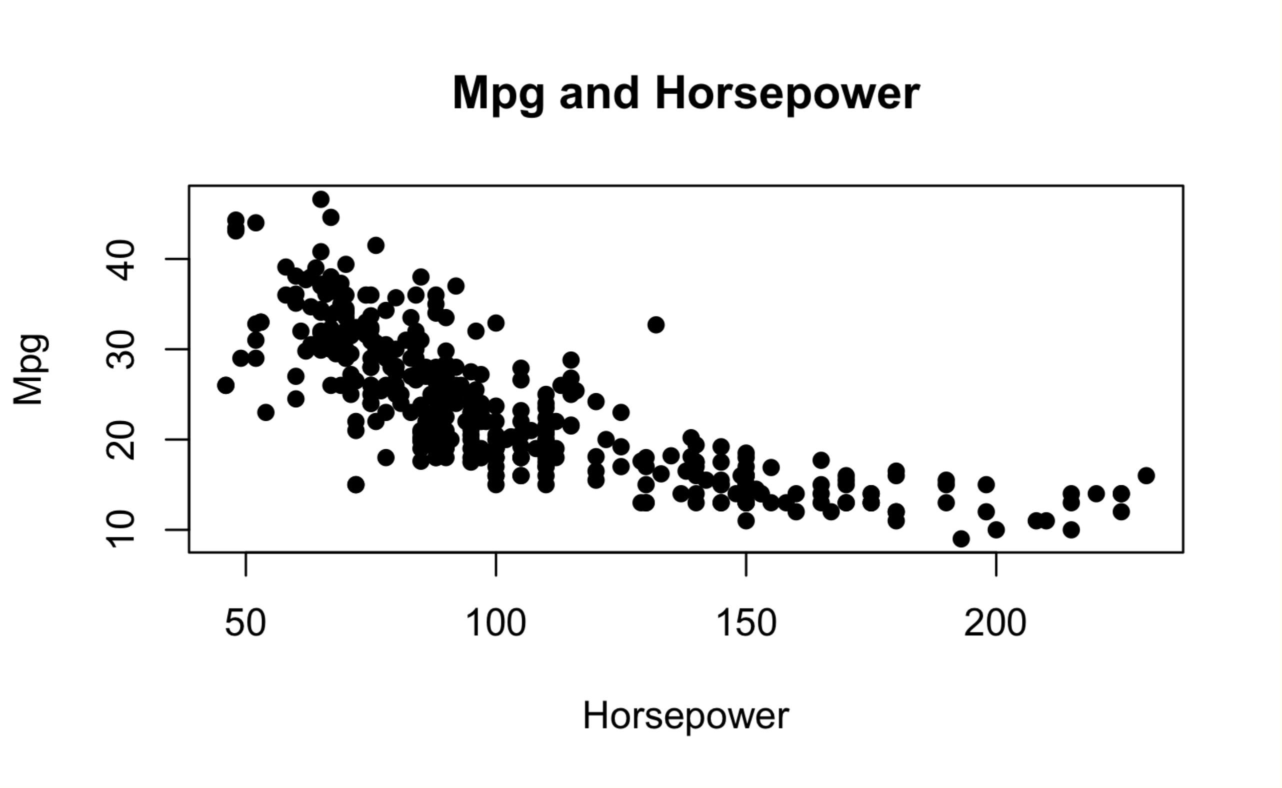
infront of interaction

term

NON LINEAR RELATIONSHIPS

Auto dataset

- Scatterplot



Curved relationship

- Here, the relationship is quite clear, but also somewhat curved
- So we consider using a quadratic(polynomial) regression
- Let Y denote mpg and let X denote horsepower, then the relationship could be (3.36)

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \varepsilon$$

Linear model

- The polynomial regression is still a linear model, and can be solved using the same idea as least squares minimization
- Let X1 = X, and let X2 = X^2
- The underlying model becomes

First entries : 130, 165, 150, 150, 140, 148

Matrix form of regression

$$X = \begin{bmatrix} 1 & x_1 & x_1^2 & \cdots & x_n^p \\ 1 & x_2 & x_2^2 & \cdots & x_n^p \\ 1 & x_3 & x_3^2 & \cdots & x_n^p \\ \vdots & & \vdots & \vdots & \\ 1 & x_n & x_n^2 & \cdots & x_n^p \end{bmatrix}$$

here for one x variable

for $Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \ldots + \beta_p x^p + \varepsilon$

Looks like

$$
X = \begin{bmatrix} 1 & 130 & 169000 \\ 1 & 165 & 27225 \\ 1 & 150 & 22500 \\ 1 & 150 & 22500 \\ 1 & 140 & 14600 \\ 1 & 198 & 39204 \end{bmatrix}
$$

Click to add title

- Coefficients:
-

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	56.9000997	1.8004268	31.60	<2e-16	***
horsepower	-0.4661896	0.0311246	-14.98	<2e-16	***
I(horsepower^2)	0.0012305	0.0001221	10.08	<2e-16	***

H_0: model includes intercept, horsepower

H_1: model includes intercept, horsepower and horsepower squared

Degree

- Discussion:
- If including X^2 leads to an improvement in the model (with intercept and X), why don't we include X^3, X^4, X^5 etc?