Multiple Linear Regression

Rebecca C. Steorts, Duke University

STA 325, Chapter 3 ISL

Agenda

- How to extend beyond a SLR
- Multiple Linear Regression (MLR)
- Relationship Between the Response and Predictors
- Model Selection: Forward Selection
- Model Fit
- Predictions
- Application
- Qualitative Predictors with More than Two Levels
- Additivity, Interactions, Polynomial Regression
- Other Issues

Multiple Linear Regression (MLR)

- SLR is a useful approach for predicting a response on the basis of a single predictor variable.
- However, in practice we often have more than one predictor.

Advertising data set

- In the Advertising data, we have examined the relationship between sales and TV advertising.
- We also have data for the amount of money spent advertising on the radio and in newspapers.
- How can we extend our analysis of the advertising data in order to accommodate these two additional predictors?

Run many SLR's!

We could run three SLRs. Not a good idea! Why?

1. It is unclear how to make a single prediction of sales given levels of the three advertising media budgets, since each of the budgets is associated with a separate regression equation.
2. Each of the three regression equations ignores the other two media in forming estimates for the regression coefficients.

MLR

Suppose we have p preditors. Then the MLR takes the form

$$Y = \beta_o + \beta_1 X_1 + \cdots \beta_p X_p + \epsilon, \tag{1}$$

where

- X_j represents the jth predictor
- β_j quantifies the association between that predictor and the response
- We interpret β_j as the average effect on Y of a one unit increase in X_j, holding all other predictors fixed.

Advertising

In the advertising example, the MLR becomes

$$\textit{sales} = \beta_o + \beta_1 \times \textit{TV} + \beta_2 \times \textit{radio} + \beta_3 \times \textit{newspaper} + \epsilon$$

Estimating the Regression Coefficients

As was the case in SLR, $\beta_o, \beta_1, \ldots, \beta_p$ are unknown and must be estimated by $\hat{\beta}_o, \hat{\beta}_1, \ldots, \hat{\beta}_p$.

Given the estimated coefficients (found by minimizing the RSS), we can make predictions using

$$\hat{y} = \hat{\beta}_o + \hat{\beta}_1 X_1 + \cdots + \hat{\beta}_p X_p.$$

Solving for the LSE is beyond the scope of this class given the calculations are quite tedious and require matrix algebra.

Important Questions to Keep in Mind

1. Is at least one of the predictors (X) useful in predicting the response?
2. Do all the predictors help to explain Y, or is only a subset of the predictors useful?
3. How well does the model fit the data?
4. Given a set of predictor values, what response value should we predict, and how accurate is our prediction?

We address these now, point by point.

Relationship Between the Response and Predictors

- In the SLR setting, to determine whether there is a relationship between the response and the predictor we simply check whether $\beta_1 = 0$.
- Are all of the regression coefficients zero?

$$H_o : \beta_o = \beta_1 = \cdots = \beta_p = 0 \tag{2}$$

$$H_a : \text{at least one } \beta_j \text{ is non-zero.} \tag{3}$$

The hypothesis test if performed by computing the F-statistic:

$$F = \frac{(TSS - RSS)/p}{RSS/(n-p-1)}.$$

Look at p-value and then make a conclusion.

Remark: Suppose $p > n$, there are more coefficients β_j to estimate than observations from which to estimate them. One simple approach to overcome this issue is **forward selection.** This high dimensional setting is discussed more in Chapter 6.

Deciding on Important Variables

The first step in a multiple regression analysis is to compute the F-statistic and to examine the associated p-value.

If we conclude on the basis of that p-value that at least one of the predictors is related to the response, then it is natural to wonder which are the guilty ones!

The task of determining which predictors are associated with the response, in order to fit a single model involving only those predictors, is referred to as **variable selection.** (See Chapter 6 for a more in depth discussion).

Model selection

It is infeasible to look at all possible models since there are 2^p models that contain subsets of p features.

For example, if p $= 30$, then we must consider $2^{30} = 1{,}073{,}741{,}824$ models!

We will focus on **forward selection** since it can be used in high dimensional settings.

Forward selection

1. We begin with the null model. This is the that contains an intercept but no predictors.
2. We then fit p simple linear regressions and add to the null model the variable that results in the lowest RSS.
3. We then add to that model the variable that results in the lowest RSS for the new two-variable model.
4. This approach is continued until some stopping rule is satisfied.

Remark: **Forward selection** is a greedy approach, and might include variables early that later become redundant.

Model Fit

Two of the most common numerical measures of model fit are the RSE and R^2, the fraction of variance explained.

These quantities are computed and interpreted in the same fashion as for simple linear regression.

Remark: It turns out that R^2 **will always increase** when more variables are added to the model, even if those variables are only weakly associated with the response.

- This is due to the fact that adding another variable to the least squares equations must allow us to fit the training data (though not necessarily the testing data) more accurately.

Predictions

Predictions are simple to make. But what is the uncertainty about our predictions?

1. The inaccuracy in the coefficient estimates is related to the reducible error from Chapter 2. We can computer a confidence interval in order to see how close $\hat{y}$ is to $f(X)$
2. Assume that $f(X)$ is a linear model is almost never correct, so there is an additional source of reducible error here, called model bias.
3. We can calculate prediction intervals (recall these are always wider than confidence intervals).

MRL for Advertising data

- In order to fit a MLR using least squares we again use the lm() function.
- The syntax $\text{lm}(y \sim x1 + x2 + x3)$ is used to fit a model with three predictors.
- The summary()function outputs the regression coefficients for the predictors.

MRL for Advertising data

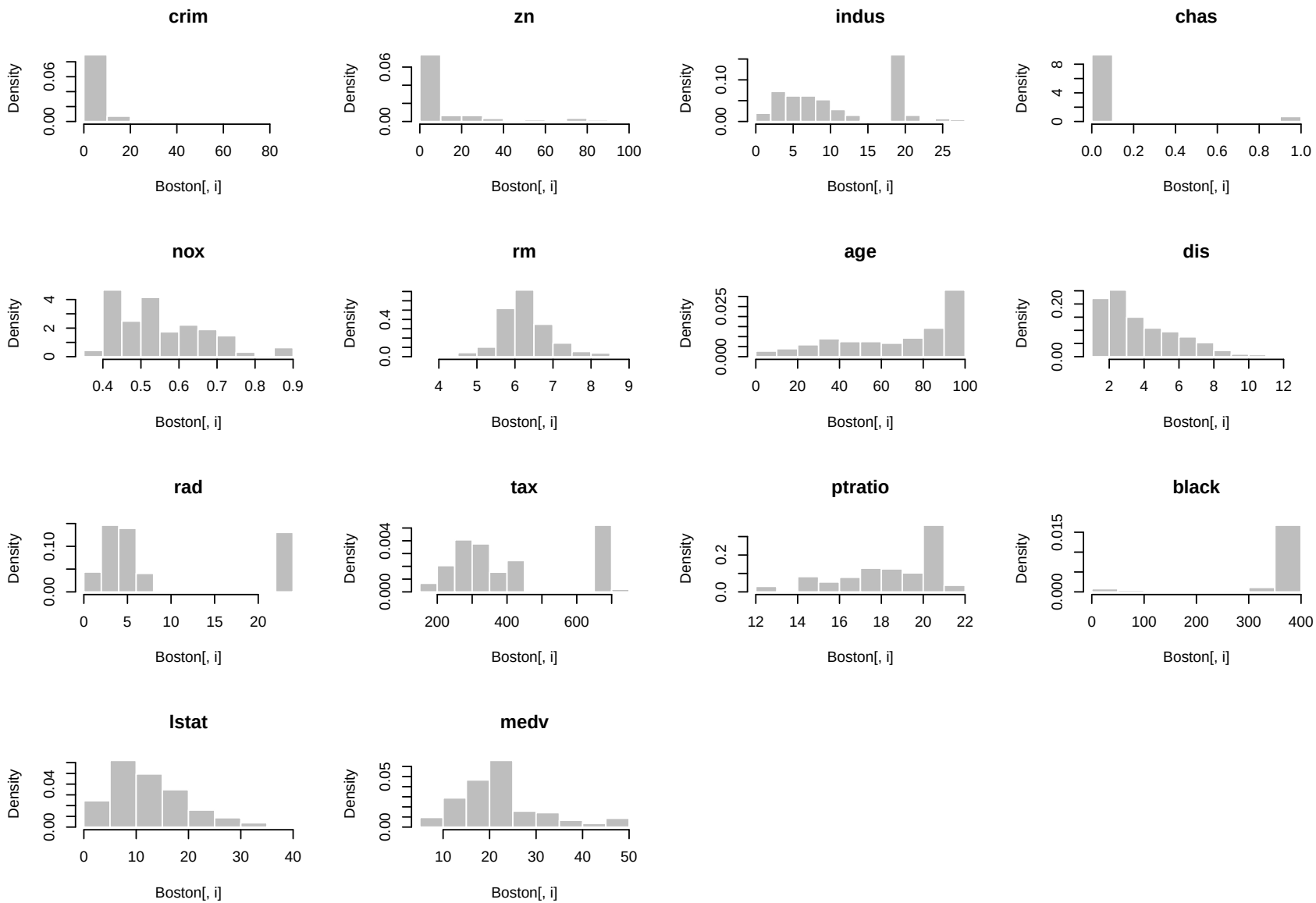
```
attach(Boston)
names(Boston)
```

```
##  [1] "crim"    "zn"      "indus"   "chas"    "nox"     "
##  [8] "dis"     "rad"     "tax"     "ptratio" "black"   "
```

```
dim(Boston)
```

```
## [1] 506  14
```

Histograms of Boston data



MRL for Advertising data

```
lm.fit <- lm(medv~., data=Boston)
summary(lm.fit)
```

```
##
## Call:
## lm(formula = medv ~ ., data = Boston)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.595  -2.730  -0.518   1.777  26.199
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.646e+01  5.103e+00   7.144 3.28e-12 ***
## crim        -1.080e-01  3.286e-02  -3.287 0.001087 **
## zn           4.642e-02  1.373e-02   3.382 0.000778 ***
## indus        2.056e-02  6.150e-02   0.334 0.738288
## chas         2.687e+00  8.616e-01   3.118 0.001925 **
## nox         -1.777e+01  3.820e+00  -4.651 4.25e-06 ***
## rm           3.810e+00  4.179e-01   9.116  < 2e-16 ***
## age          6.922e-04  1.321e-02   0.052 0.958229
## dis         -1.476e+00  1.995e-01  -7.398 6.01e-13 ***
## rad          3.060e-01  6.635e-02   4.613 5.07e-06 ***
## tax         -1.233e-02  3.760e-03  -3.280 0.001112 **
## ptratio     -9.527e-01  1.308e-01  -7.283 1.31e-12 ***
## black        9.312e-03  2.686e-03   3.467 0.000573 ***
## lstat       -5.248e-01  5.072e-02 -10.347  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.745 on 492 degrees of freedom
## Multiple R-squared:  0.7406, Adjusted R-squared:  0.7338
## F-statistic: 108.1 on 13 and 492 DF,  p-value: < 2.2e-16
```

MRL for Advertising data

We can access each part of summary by typing ?summary.lm to see what is available

```
#r2
summary(lm.fit)$r.sq
```

```
## [1] 0.7406427
```

```
#RMSE
summary(lm.fit)$sigma
```

```
## [1] 4.745298
```

```
summary(lm.fit)$residuals
```

```
##           1           2           3           4           5
##  -6.003843377 -3.425562379  4.132403281  4.792963511  8.256475767
##           6           7           8           9          10
##   3.443715538 -0.101808268  7.564011571  4.976363147 -0.020262107
##          11          12          13          14          15
##  -3.999496511 -2.686795681  0.793478472  0.847097189 -1.083482050
##          16          17          18          19          20
##   0.602516792  2.572490209  0.588598653  4.021988943 -0.206136033
##          21          22          23          24          25
##   1.076142473  1.928963305 -0.632881292  0.693714654 -0.078338315
##          26          27          28          29          30
##   0.513314391  1.136023454  0.091525719 -1.147372851  0.123571798
##          31          32          33          34          35
##   1.244882410 -3.559232946  4.388942638 -1.182758141 -0.206758913
##          36          37          38          39          40
##  -4.914635265 -2.341937076 -2.108911425  1.784973884 -0.557625688
##          41          42          43          44          45
```

MRL for Advertising data

```
summary(lm.fit)$coefficients
```

```
##                 Estimate  Std. Error      t value     Pr(>|t|)
## (Intercept)  3.645949e+01 5.103458811   7.14407419 3.283438e-12
## crim        -1.080114e-01 0.032864994  -3.28651687 1.086810e-03
## zn           4.642046e-02 0.013727462   3.38157628 7.781097e-04
## indus        2.055863e-02 0.061495689   0.33431004 7.382881e-01
## chas         2.686734e+00 0.861579756   3.11838086 1.925030e-03
## nox         -1.776661e+01 3.819743707  -4.65125741 4.245644e-06
## rm           3.809865e+00 0.417925254   9.11614020 1.979441e-18
## age          6.922246e-04 0.013209782   0.05240243 9.582293e-01
## dis         -1.475567e+00 0.199454735  -7.39800360 6.013491e-13
## rad          3.060495e-01 0.066346440   4.61289977 5.070529e-06
## tax         -1.233459e-02 0.003760536  -3.28000914 1.111637e-03
## ptratio     -9.527472e-01 0.130826756  -7.28251056 1.308835e-12
## black        9.311683e-03 0.002685965   3.46679256 5.728592e-04
## lstat       -5.247584e-01 0.050715278 -10.34714580 7.776912e-23
```

MRL for Advertising data

Note that age has by far a very high p-value.

How do we re-run the regression, omitting age?

```
lm.fit1 <- lm(medv~. -age, data=Boston)
summary(lm.fit1)
```

```
##
## Call:
## lm(formula = medv ~ . - age, data = Boston)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.6054  -2.7313  -0.5188   1.7601  26.2243
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  36.436927   5.080119   7.172 2.72e-12 ***
## crim         -0.108006   0.032832  -3.290 0.001075 **
## zn            0.046334   0.013613   3.404 0.000719 ***
## indus         0.020562   0.061433   0.335 0.737989
## chas          2.689026   0.859598   3.128 0.001863 **
## nox         -17.713540   3.679308  -4.814 1.97e-06 ***
## rm            3.814394   0.408480   9.338  < 2e-16 ***
## dis          -1.478612   0.190611  -7.757 5.03e-14 ***
## rad           0.305786   0.066089   4.627 4.75e-06 ***
## tax          -0.012329   0.003755  -3.283 0.001099 **
## ptratio      -0.952211   0.130294  -7.308 1.10e-12 ***
## black         0.009321   0.002678   3.481 0.000544 ***
## lstat        -0.523852   0.047625 -10.999  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

Qualitative Predictors

So far, we have only considered quantitative predictors.

What if the predictors are qualitative?

As an example, the Credit data set records balance (average credit card debt for individuals) for

- quantiative predictors: age, cards, education, rating
- qualitative predictors: gender, student (status), status (marital), and ethnicity.

Qualitative Predictors with Two Levels

If the qualitative predictor (factor) has to levels, we can create an indicator or dummy variable.

Suppose we are dealing with gender (male, female)

Then we create a new variable, where

$$x_i = \begin{cases} 1, \text{if the } i\textit{th} \text{ person is female} \\ 0, \text{otherwise} \end{cases}$$

Qualitative Predictors with Two Levels

Using our indicator variable

$$x_i = \begin{cases} 1, \text{if the i}th \text{ person is female} \\ 0, \text{otherwise,} \end{cases}$$

we can then use this in our regression model:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \tag{4}$$

$$= \begin{cases} \beta_0 + \beta_1 + \epsilon_i, \text{if the i}th \text{ person is female} \\ \beta_0 + \epsilon_i, \text{otherwise} \end{cases} \tag{5}$$

- Now β_0 can be interpreted as the overall average credit card balance (ignoring the gender effect), and
- β_1 is the amount that females are above the average and males are below the average.

Qualitative Predictors with More than Two Levels

When a qualitative predictor has more than two levels, a single dummy variable cannot represent all possible values. In this situation, we can create additional dummy variables.

For example, for the ethnicity variable we create two dummy variables.

$$x_{i1} = \begin{cases} 1, \text{if the } i\textit{th} \text{ person is Asian} \\ 0, \text{if the } i\textit{th} \text{ person is not Asian} \end{cases}$$

and

$$x_{i2} = \begin{cases} 1, \text{if the } i\textit{th} \text{ person is Caucasian} \\ 0, \text{if the } i\textit{th} \text{ person is not Caucasian} \end{cases}$$

Qualitative Predictors with More than Two Levels

Then both of these variables can be used in the regression equation, in order to obtain the model

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i \tag{6}$$

$$= \begin{cases} \beta_0 + \beta_1 + \epsilon_i, \text{if the i}th \text{ person is Asian} \\ \beta_0 + \beta_2 + \epsilon_i, \text{if the i}th \text{ person is Caucasian} \\ \beta_0 + \epsilon_i, \text{if the i}th \text{ person is African American} \end{cases} \tag{7}$$

Qualitative Predictors with More than Two Levels

- Now β_0 can be interpreted as the average credit card balance for African Americans,
- β_1 can be interpreted as the difference in the average balance between the Asian and African American categories, and
- β_2 can be interpreted as the difference in the average balance between the Caucasian and African American categories.
- There will always be one fewer dummy variable than the number of levels.
- The level with no dummy variable — African American in this example — is known as the **baseline**.

Extensions of the Linear Model

We discuss removing two of the most important assumptions of the linear model: additivity and linearity.

Additivity

The **additive assumption** means that the effect of changes in a predictor X_j on the response Y is independent of the values of the other predictors.

Removing the Additive Effect

Suppose that spending money on radio advertising actually increases the effectiveness of TV ad- vertising, so that the slope term for TV should increase as radio increases.

Given a fixed budget of $100,000, spending half on radio and half on TV may increase sales more than allocating the entire amount to either TV or to radio. This is referred to as an **interaction effect**.

Interaction Effect

One way of extending this model to allow for interaction effects is to include a third predictor, called an **interaction term**, which is constructed by computing the product of X_1 and X_2, which results in

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \epsilon. \tag{8}$$

How does inclusion of this interaction term relax the additive assumption?

Equation 8 can be rewritten as

$$\begin{aligned} Y &= \beta_0 + (\beta_1 + \beta_3 X_2) X_1 + \beta_2 X_2 + \epsilon && (9) \\ &= \beta_0 + \tilde{\beta}_1 X_1 + \beta_2 X_2 + \epsilon && (10) \\ & && (11) \end{aligned}$$

where $\tilde{\beta}_1 = \beta_1 + \beta_3 X_2$.

Interaction Effect

Since $\tilde{\beta}_1$ changes with X_2, the effect of X_1 on Y is no longer constant: adjusting X_2 will change the impact of X_1 on Y.

Linear Assumption

The linear assumption means that the change in the response Y due to a one-unit change in X_j is constant, regardless of the value of X_j.

In our previous analysis of the Advertising data, we concluded that both TV and radio seem to be associated with sales.

For example, the linear model states that the average effect on sales of a one-unit increase in TV is always β_1, regardless of the amount spent on radio.

However, the simple model may be incorrect.

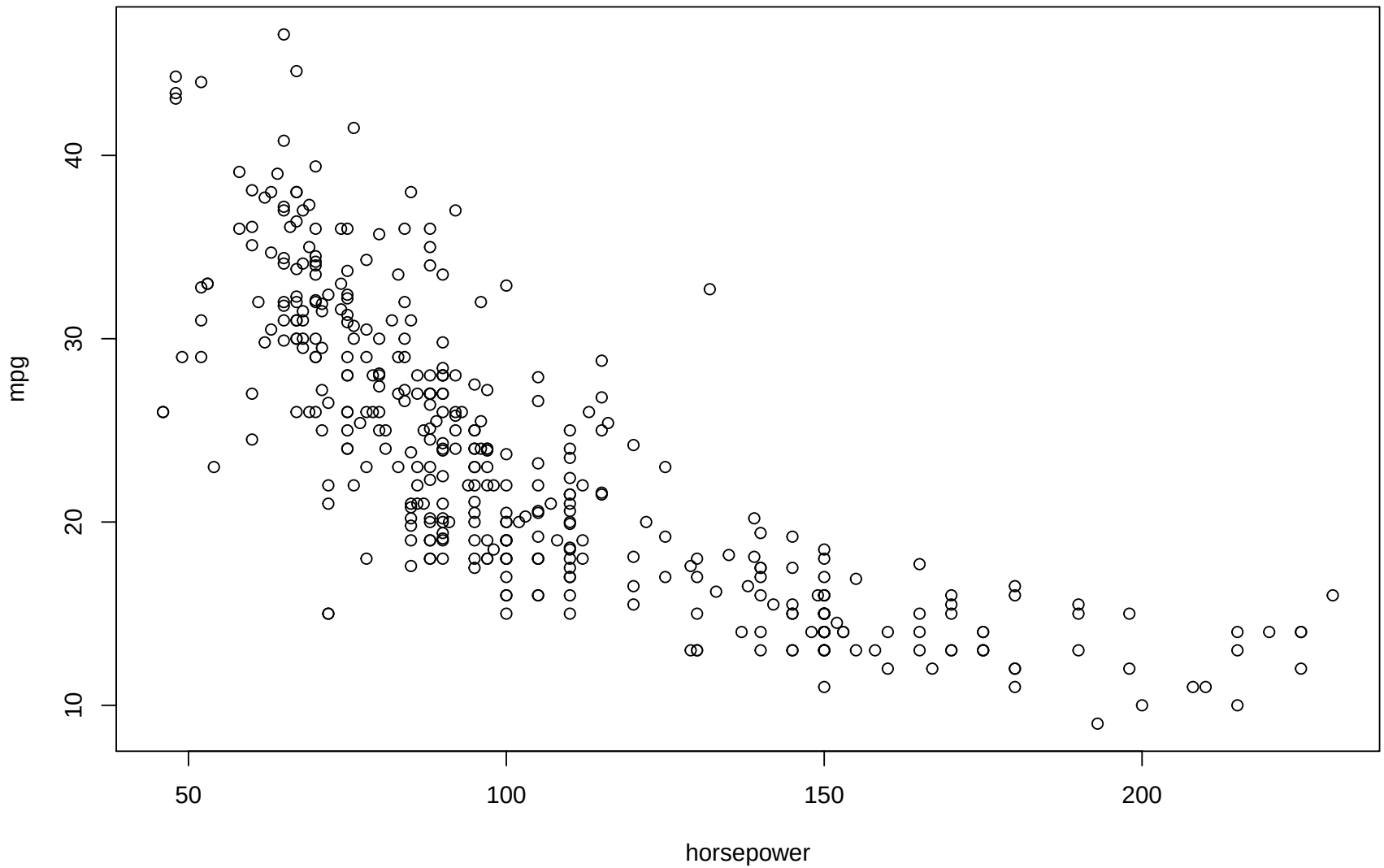
Polynomial regression

We present a very simple way to directly extend the linear model to accommodate non-linear relationships, using **polynomial regression**

Polynomial regression

Let us consider a motivating example from the Auto data set.

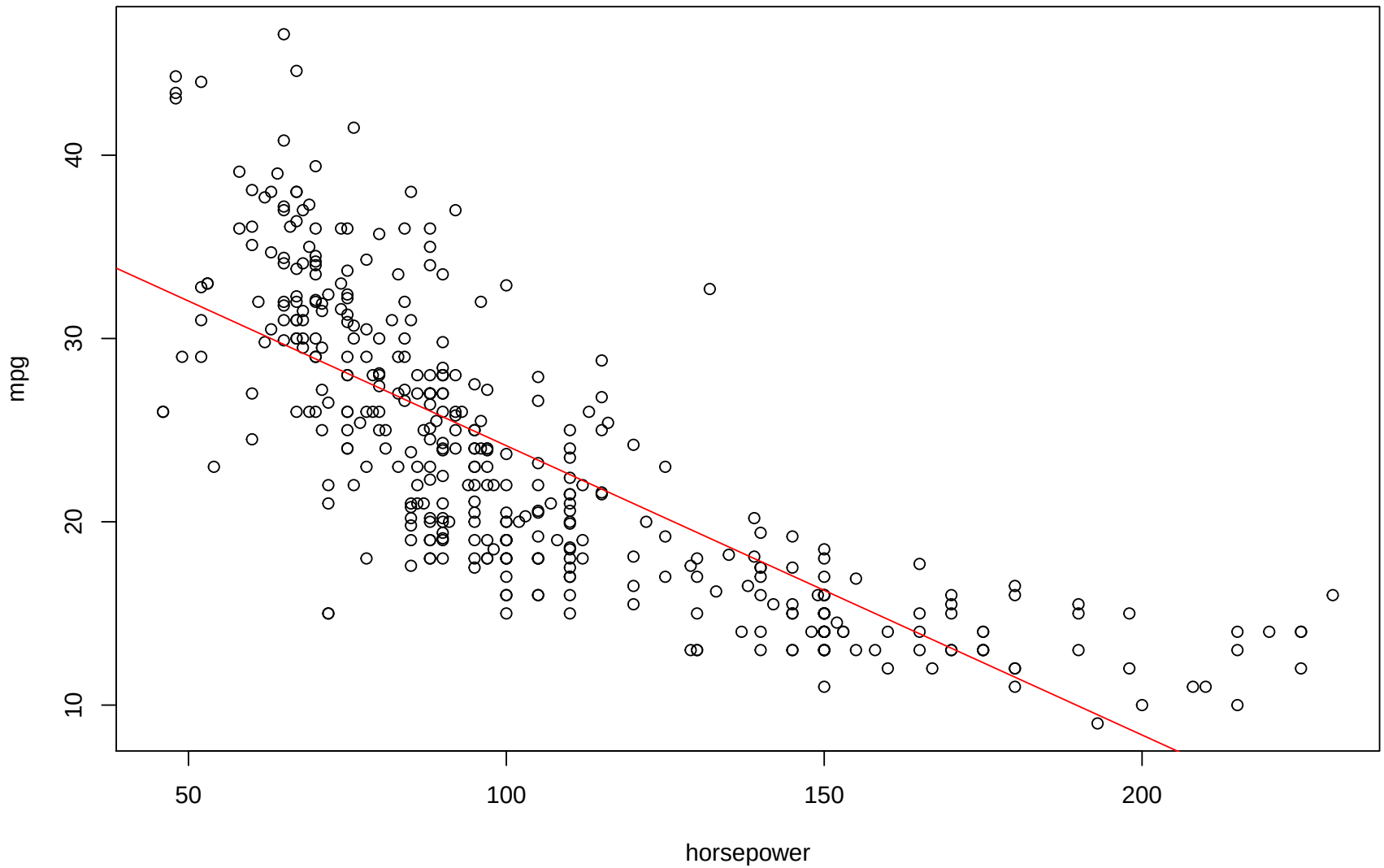
```
attach(Auto)
plot(horsepower, mpg)
```



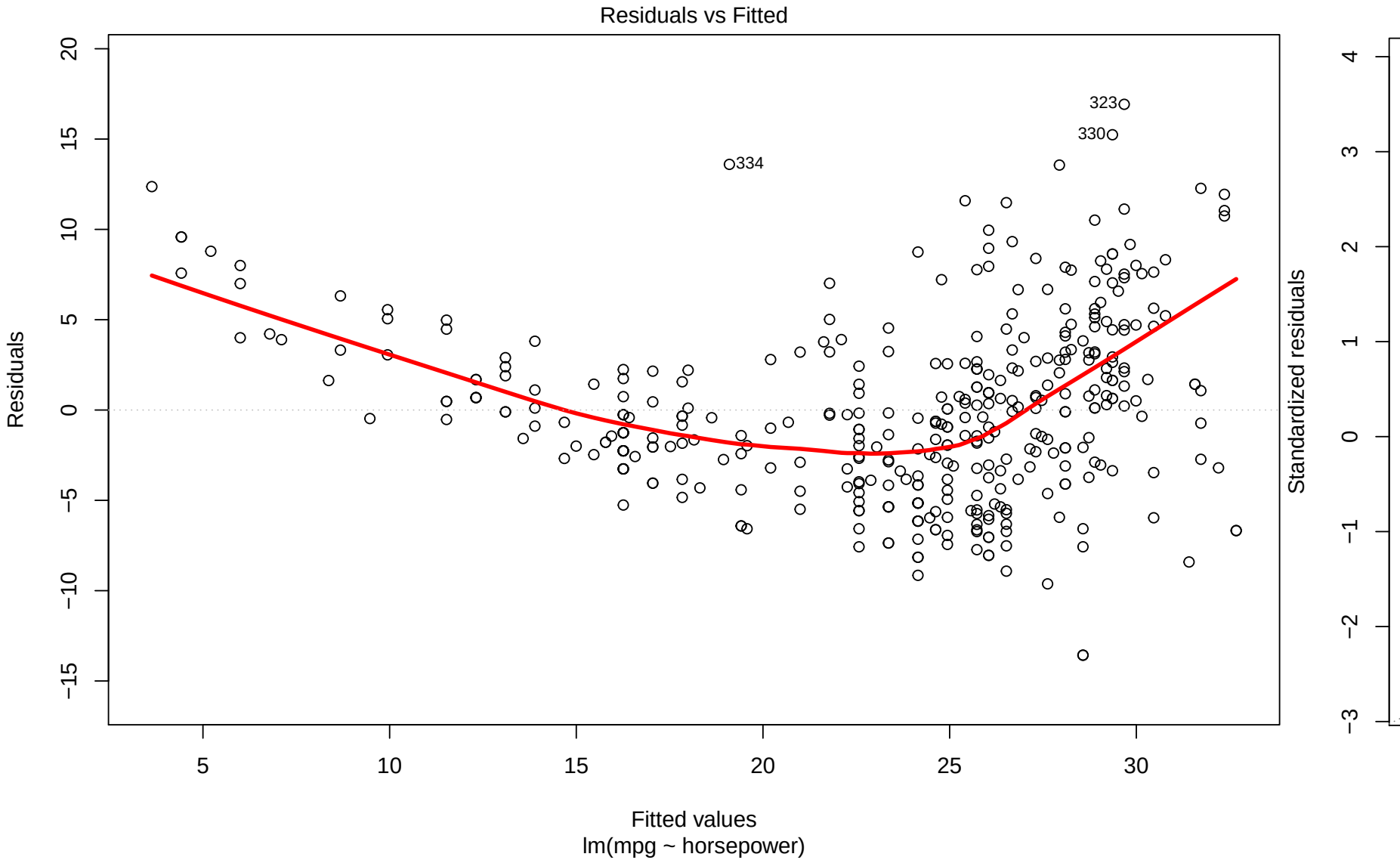
Polynomial regression

ATTN: I can't get the polynomail onto the regression line!

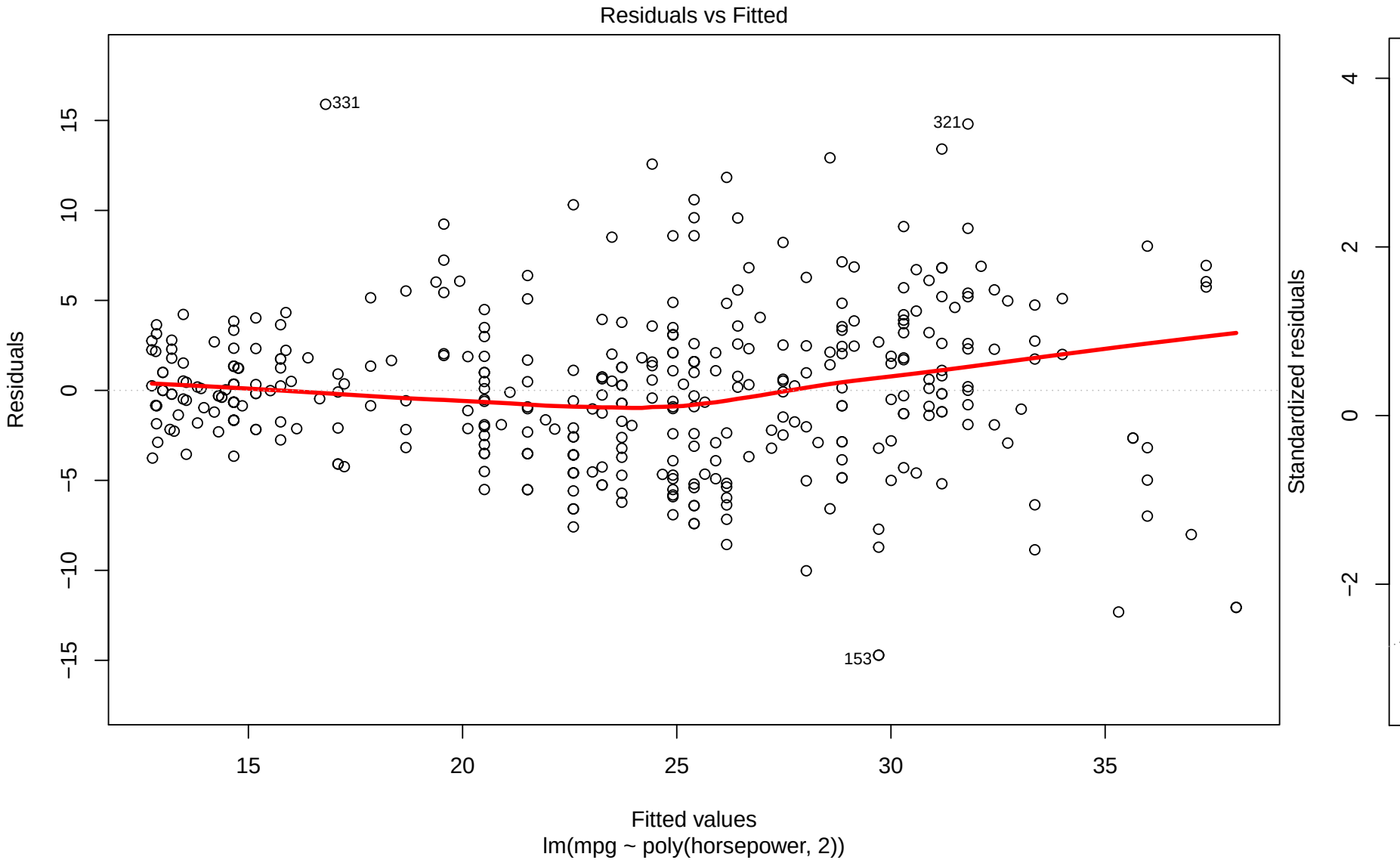
```
## Warning in abline(lm.fit.poly2, lwd = 3, col = "brown"): only using the
## first two of 3 regression coefficients
```



Linear regression



Polynomial regression



Potential Problems

There are many issues that can arise in linear regression since it is a simple method.

Non-linearity of the response-predictor relationships

This can be checked using residual plots to identify for non-linearity.

That is plot the residuals e_i versus x_i.

In the MLR setting, plot the residual versus the predicted (fitted) values $\hat{y}_i$.

An ideal plot will have no pattern!

Correlation of Error Terms

An important assumption of the linear model is that the error terms are uncorrelated.

If there is correlation in the error terms, then the estimation standard errors will tend to understimate the true standard errors.

As a result, confidence and prediction intervals will be more narrow than they should be.

Why might we have correlated errors? We could have time series data, where such data points that appear are correlated by the time structure.

Non-constant Variance of Error Terms

Another important assumption is that the linear regresion model is that the error terms have constant variance (σ^2).

Often, the variance of the error terms are non-constant.

One can identify non-constant errors (or heterscedarcity) from the presence of a **funnel shape** in the residual plot.

One easy solution is transforming the response using a concave function such as log or sqrt.

Then recheck the residual plot.

Outliers

An outlier is a point for which y_i is far from the value predicted by the model.

Outliers can arise for a variety of reasons, such as incorrect recording of an observation during data collection.

Residual (studentized) plots can be used to identify outliers.

High Leverage Points

We just saw that outliers are observations for which the response y_i is unusual given the predictor x_i.

In contrast, observations with high leverage high leverage have an unusual value for x_i.

High leverage observations tend to have a sizable impact on the estimated regression line.

For this reason, it is important to identify high leverage observations.

In order to quantify an observation's leverage, we compute the **leverage statistic**. A large value of this statistic indicates an observation with high leverage leverage.

For SLR,

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})}{\sum_{j=1}^{n}(x_j - \bar{x}}.$$

Collinearity

Collinearity refers to the situation in which two or more predictor variables are closely related to one another.

Collinearity

Let's consider the Credit data set.

```
Credit <- read.csv("data/credit.csv",header=TRUE, sep=",")
pdf(file = "examples/collinear.pdf")
par(mfrow=c(1, 2))
plot(Credit$Limit, Credit$Age, xlab="Limit", ylab="Age")
plot(Credit$Limit, Credit$Rating, xlab="Limit",
     ylab="Rating")
dev.off()
```

```
## pdf
##   2
```

Collinearity

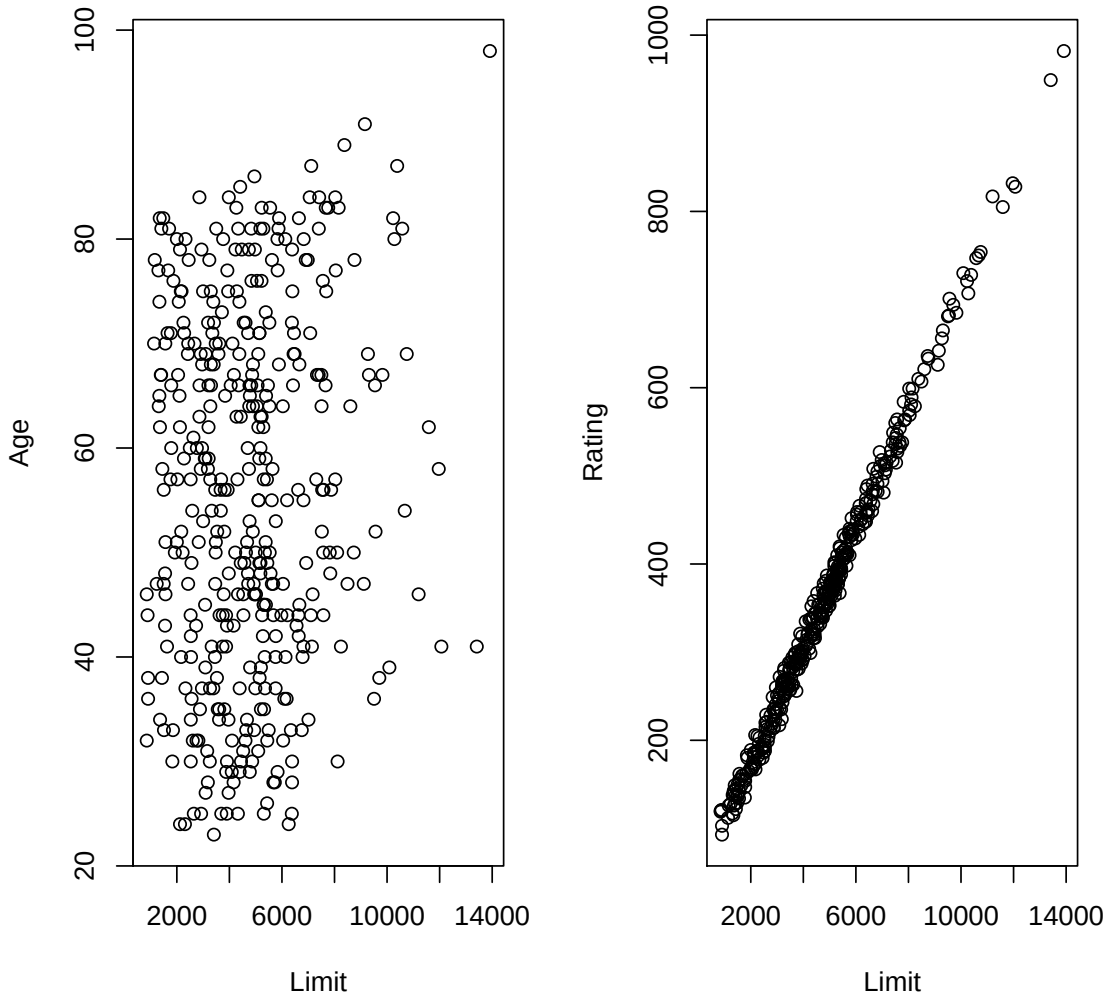


Figure 1: Scatterplots of the observations from the Credit data set. Left: Plot of age versus limit. Right: A plot of age versus rating, with a high

The Marketing Plan

Exercise: Now that we have walked through this, see Section 3.4 and answer these questions on your own.