

Lecture 6: Linear Regression (continued)

Reading: Sections 3.1-3

STATS 202: Data mining and analysis

Sergio Bacallado

October 7, 2013

Multiple linear regression

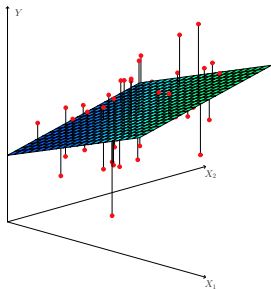


Figure 3.4

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon$$

$$\varepsilon \sim \mathcal{N}(0, \sigma) \quad \text{i.i.d.}$$

or, in matrix notation:

$$\mathbf{y} = \mathbf{X}\beta,$$

where $\mathbf{y} = (y_1, \dots, y_n)^T$,
 $\beta = (\beta_0, \dots, \beta_p)^T$ and \mathbf{X} is our
usual data matrix with an extra
column of zeroes on the left to
account for the intercept.

The estimates $\hat{\beta}$

Our goal again is to minimize the RSS:

$$\begin{aligned}\text{RSS} &= \sum_{i=1}^n (y_i - \hat{y}_i)^2 \\ &= \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_{i,1} - \cdots - \beta_p x_{i,p})^2.\end{aligned}$$

One can show that this is minimized by the vector $\hat{\beta}$:

$$\hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}.$$

Which variables are important?

- ▶ F-test:

$$H_0 : \beta_{p-q+1} = \beta_{p-q+2} = \cdots = \beta_p = 0.$$

RSS_0 is the residual sum of squares for the model in H_0 .

$$F = \frac{(RSS_0 - RSS)/q}{RSS/(n - p - 1)}.$$

- ▶ Special case: $q = p$. Test whether any of the predictors are important.
- ▶ Special case: exclude a single variable. Test whether this variable is important $\sim t$ -tests in R output. **Must be careful with multiple testing.**

How many variables are important?

When choosing a subset of the predictors, we have 2^p choices.

- ▶ **Forward selection:** Starting from a *null model*, include variables one at a time, minimizing the RSS at each step.
- ▶ **Backward selection:** Starting from the *full model*, eliminate variables one at a time, choosing the one with the largest p-value at each step.
- ▶ **Mixed selection:** Starting from a *null model*, include variables one at a time, minimizing the RSS at each step. If the p-value for some variable goes beyond a threshold, eliminate that variable.

How many variables are important?

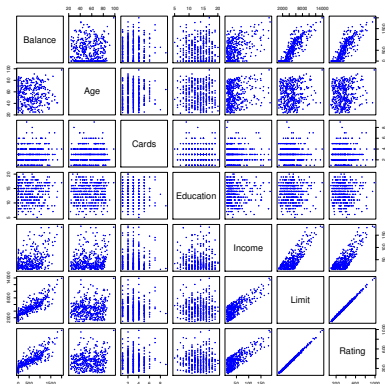
The output of a stepwise selection method is a range of models:

- ▶ $\{\}$
- ▶ $\{\text{tv}\}$
- ▶ $\{\text{tv}, \text{newspaper}\}$
- ▶ $\{\text{tv}, \text{newspaper}, \text{radio}\}$
- ▶ $\{\text{tv}, \text{newspaper}, \text{radio}, \text{facebook}\}$
- ▶ $\{\text{tv}, \text{newspaper}, \text{radio}, \text{facebook}, \text{twitter}\}$

6 choices are better than $2^6 = 64$. We use different *tuning methods* to decide which model to use; e.g. cross-validation, AIC, BIC.

Dealing with categorical or qualitative predictors

Example: Credit dataset



In addition, there are 4 qualitative variables:

- ▶ gender: male, female.
- ▶ student: student or not.
- ▶ status: married, single, divorced.
- ▶ ethnicity: African American, Asian, Caucasian.

Dealing with categorical or qualitative predictors

For each qualitative predictor, e.g. ethnicity:

- ▶ Choose a baseline category, e.g. African American
- ▶ For every other category, define a new predictor:
 - ▶ X_{Asian} is 1 if the person is Asian and 0 otherwise.
 - ▶ $X_{\text{Caucasian}}$ is 1 if the person is Caucasian and 0 otherwise.

The model will be:

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_7 X_7 + \beta_{\text{Asian}} X_{\text{Asian}} + \beta_{\text{Caucasian}} X_{\text{Caucasian}} + \varepsilon.$$

β_{Asian} is the relative effect on balance for being Asian compared to the baseline category.

Dealing with categorical or qualitative predictors

- ▶ The model fit and predictions are independent of the choice of the baseline category.
- ▶ However, hypothesis tests derived from these variables are affected by the choice.
 - ▶ **Solution:** To check whether ethnicity is important, use an F -test for the hypothesis $\beta_{\text{Asian}} = \beta_{\text{Caucasian}} = 0$. This does not depend on the coding.
- ▶ Other ways to encode qualitative predictors produce the same fit \hat{f} , but the coefficients have different interpretations; e.g. *contrasts*.

How good are the predictions?

The function `predict` in R output predictions from a linear model;
eg. $x_0 = (5, 10, 15)$:

```
> predict(lm.fit, data.frame(lstat=(c(5,10,15))),  
          interval="confidence")  
      fit   lwr   upr  
1 29.80 29.01 30.60  
2 25.05 24.47 25.63  
3 20.30 19.73 20.87
```

“Confidence intervals” reflect the uncertainty on $\hat{\beta}$; ie. confidence interval for $f(x_0)$.

```
> predict(lm.fit, data.frame(lstat=(c(5,10,15))),  
          interval="prediction")  
      fit   lwr   upr  
1 29.80 17.566 42.04  
2 25.05 12.828 37.28  
3 20.30  8.078 32.53
```

“Prediction intervals” reflect uncertainty on $\hat{\beta}$ and the irreducible error ε as well; i.e. confidence interval for y_0 .

Recap

So far, we have:

- ▶ Defined Multiple Linear Regression
- ▶ Discussed how to test the importance of variables.
- ▶ Described one approach to choose a subset of variables.
- ▶ Explained how to code qualitative variables.
- ▶ Now, how do we evaluate model fit? Is the linear model any good? What can go wrong?

How good is the fit?

To assess the fit, we focus on the residuals.

- ▶ $R^2 = \text{Corr}(Y, \hat{Y})$, always increases as we add more variables.
- ▶ The residual standard error (RSE) does not always improve with more predictors:

$$\text{RSE} = \sqrt{\frac{1}{n - p - 1} \text{RSS}}.$$

- ▶ **Visualizing the residuals** can reveal phenomena that are not accounted for by the model.

Potential issues in linear regression

1. Interactions between predictors
2. Non-linear relationships
3. Correlation of error terms
4. Non-constant variance of error (heteroskedasticity).
5. Outliers
6. High leverage points
7. Colinearity

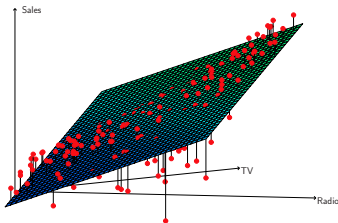
Interactions between predictors

Linear regression has an *additive* assumption:

$$\text{sales} = \beta_0 + \beta_1 \times \text{tv} + \beta_2 \times \text{radio} + \varepsilon$$

i.e. An increase of \$100 dollars in TV ads causes a fixed increase in sales, regardless of how much you spend on radio ads.

If we visualize the residuals, it is clear that this is false:



Interactions between predictors

One way to deal with this is to include multiplicative variables in the model:

$$\text{sales} = \beta_0 + \beta_1 \times \text{tv} + \beta_2 \times \text{radio} + \beta_3 \times (\text{tv} \cdot \text{radio}) + \varepsilon$$

The **interaction variable** is high when both tv and radio are high.

Interactions between predictors

R makes it easy to include interaction variables in the model:

```
> lm.fit=lm(Sales~.+Income:Advertising+Price:Age,data=Carseats)
> summary(lm.fit)

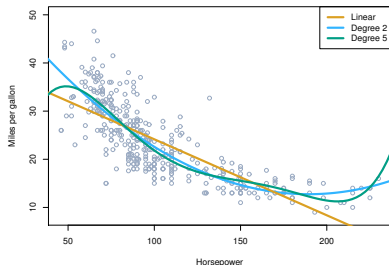
Call:
lm(formula = Sales ~ . + Income:Advertising + Price:Age, data =
    Carseats)

Residuals:
    Min       1Q   Median       3Q      Max
-2.921  -0.750   0.018   0.675   3.341

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    6.575565    1.008747    6.52 2.2e-10 ***
CompPrice      0.092937    0.004118   22.57 < 2e-16 ***
Income         0.010894    0.002604    4.18 3.6e-05 ***
Advertising     0.070246    0.022609    3.11 0.00203 **
Population     0.000159    0.000368    0.43 0.66533
Price        -0.100806    0.007440   -13.55 < 2e-16 ***
ShelveLocGood  4.848676    0.152838   31.72 < 2e-16 ***
ShelveLocMedium 1.953262    0.125768   15.53 < 2e-16 ***
Age           -0.057947    0.015951   -3.63 0.00032 ***
Education     -0.020852    0.019613   -1.06 0.28836
UrbanYes       0.140160    0.112402    1.25 0.21317
USYes         -0.157557    0.148923   -1.06 0.29073
Income:Advertising 0.000751    0.000278    2.70 0.00729 **
Price:Age       0.000107    0.000133    0.80 0.42381
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```


Non-linearities

Example: Auto dataset.



A scatterplot between a predictor and the response may reveal a non-linear relationship.

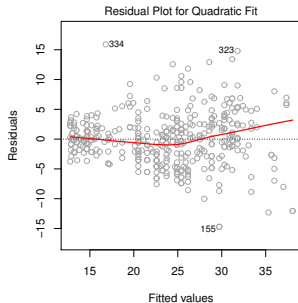
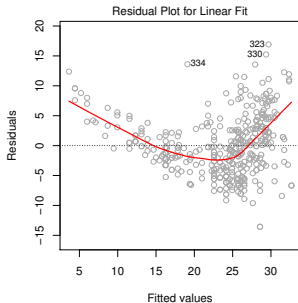
Solution: include polynomial terms in the model.

$$\begin{aligned} \text{MPG} = & \beta_0 + \beta_1 \times \text{horsepower} \\ & + \beta_2 \times \text{horsepower}^2 \\ & + \beta_3 \times \text{horsepower}^3 \\ & + \dots + \varepsilon \end{aligned}$$

Non-linearities

In 2 or 3 dimensions, this is easy to visualize. What do we do when we have too many predictors?

Plot the residuals against the *response* and look for a pattern:



Correlation of error terms

We assumed that the errors for each sample are independent:

$$y_i = f(x_i) + \varepsilon_i \quad ; \quad \varepsilon_i \sim \mathcal{N}(0, \sigma) \text{ i.i.d.}$$

What if this breaks down?

The main effect is that this invalidates any assertions about Standard Errors, confidence intervals, and hypothesis tests:

Example: Suppose that by accident, we double the data (we use each sample twice). Then, the standard errors would be artificially smaller by a factor of $\sqrt{2}$.

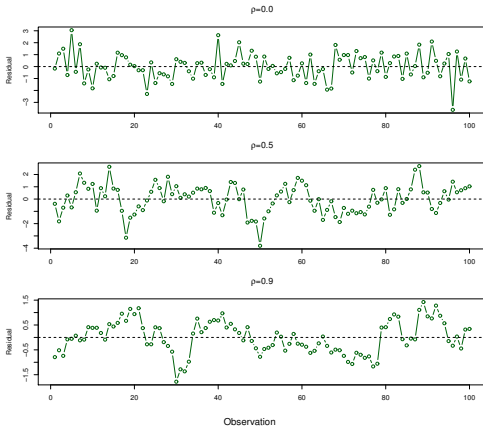
Correlation of error terms

When could this happen in real life:

- ▶ **Time series:** Each sample corresponds to a different point in time. The errors for samples that are close in time are correlated.
- ▶ **Spatial data:** Each sample corresponds to a different location in space.
- ▶ Study on predicting height from weight at birth. Suppose some of the subjects in the study are in the same family, their shared environment could make them deviate from $f(x)$ in similar ways.

Correlation of error terms

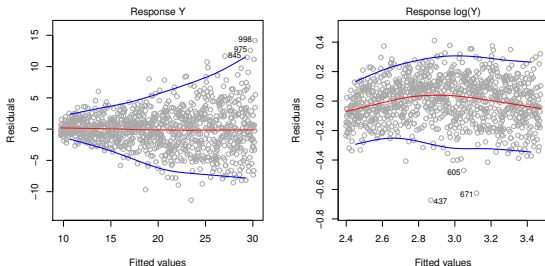
Simulations of time series with increasing correlations between ε_i .



Non-constant variance of error (heteroskedasticity)

The variance of the error depends on the input.

To diagnose this, we can plot residuals vs. fitted values:



Solution: If the trend in variance is relatively simple, we can transform the response using a logarithm, for example.