STA 314: Statistical Methods for Machine Learning I

Lecture 2 - Linear regression

Xin Bing

Department of Statistical Sciences
University of Toronto

Review

- Supervised learning is about to estimate (learn) f under the generating mechanism
$$Y = f(X) + \epsilon$$

- For a given estimate estimate $\hat{f}$ of f, we have learned that
  - it follows the bias-variance tradeoff
  - and how to evaluate it

- Different $\hat{f}$'s are different algorithms / methods:
  - parametric methods
  - non-parametric methods

Linear regression

Let $Y \in \mathbb{R}$ be the outcome and $X \in \mathbb{R}^p$ be the (random) vector of p features.

The linear model assumes

$$\begin{aligned} Y &= f(x) + \epsilon \\ &= \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon \qquad \text{(linearity of } f\text{)} \end{aligned}$$

where:

- $\beta_0, \beta_1, \cdots, \beta_p$ are unknown constants.
  - β_0 is called the **intercept**
  - β_j, for $1 \le j \le p$, are the **coefficients** or **parameters** of the p features
- ϵ is the error term satisfying $\mathbb{E}[\epsilon|X] = 0$.

Linear predictor under the linear regression model

Given some estimates $\hat{\beta}_j$ of β_j for $0 \leq j \leq p$, we predict the response at any $X = x$ by the linear predictor

$$\hat{y} := \hat{f}(x) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_p x_p.$$

Question: how to choose $\hat{\beta}_0, \ldots, \hat{\beta}_p$?

Ordinary Least Squares approach (OLS)

Recall that we want to find a function g by

$$\min_g \mathbb{E}\left[(Y - g(X))^2\right].$$

Under linear model, it suffices to find $\alpha_0, \ldots, \alpha_p$ by

$$\min_{\alpha_0,\ldots,\alpha_p} \mathbb{E}\left[\left(Y - \alpha_0 - \alpha_1 X_1 - \cdots - \alpha_p X_p\right)^2\right]$$

In the model fitting step, we use $\mathcal{D}^{train}$ to approximate the above expectation (w.r.t. X and Y).

Specifically, given the training data $\mathcal{D}^{train}$: $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$, we choose $\hat{\beta}_0, \cdots, \hat{\beta}_p$ by minimizing the training MSE.

$$(\hat{\beta}_0, \cdots, \hat{\beta}_p) = \underset{\alpha_0,\cdots,\alpha_p}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \alpha_0 - \alpha_1 x_{i1} - \cdots - \alpha_p x_{ip}\right)^2.$$

Ordinary Least Squares approach (OLS)

Using the matrix notation,

- $\hat{\boldsymbol{\beta}} = [\hat{\beta}_0, \cdots, \hat{\beta}_p]^\top \in \mathbb{R}^{p+1}$, $\boldsymbol{\beta} = [\beta_0, \cdots, \beta_p]^\top \in \mathbb{R}^{p+1}$, $\boldsymbol{\alpha} = [\alpha_0, \cdots, \alpha_p]^\top \in \mathbb{R}^{p+1}$,
- $\mathbf{y} = [y_1, \ldots, y_n]^\top \in \mathbb{R}^n$, $\mathbf{X} = [1, x_1, \ldots, x_n]^\top \in \mathbb{R}^{n \times (p+1)}$,

the OLS estimator of $\boldsymbol{\beta}$ is defined as

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= \underset{\boldsymbol{\alpha} \in \mathbb{R}^{p+1}}{\operatorname{argmin}} \ \frac{1}{n} \sum_{i=1}^{n} \left(y_i - x_i^\top \boldsymbol{\alpha}\right)^2 \\ &= \underset{\boldsymbol{\alpha} \in \mathbb{R}^{p+1}}{\operatorname{argmin}} \ \frac{1}{n} \|\mathbf{y} - \mathbf{X}\boldsymbol{\alpha}\|_2^2.\end{aligned}$$

Comments

- The idea of estimating f by minimizing the training MSE can be applied to (almost) all supervised problems. Specifically,

$$\hat{f} = \underset{g \in \mathcal{F}}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} L(y_i, g(x_i)) \tag{1}$$

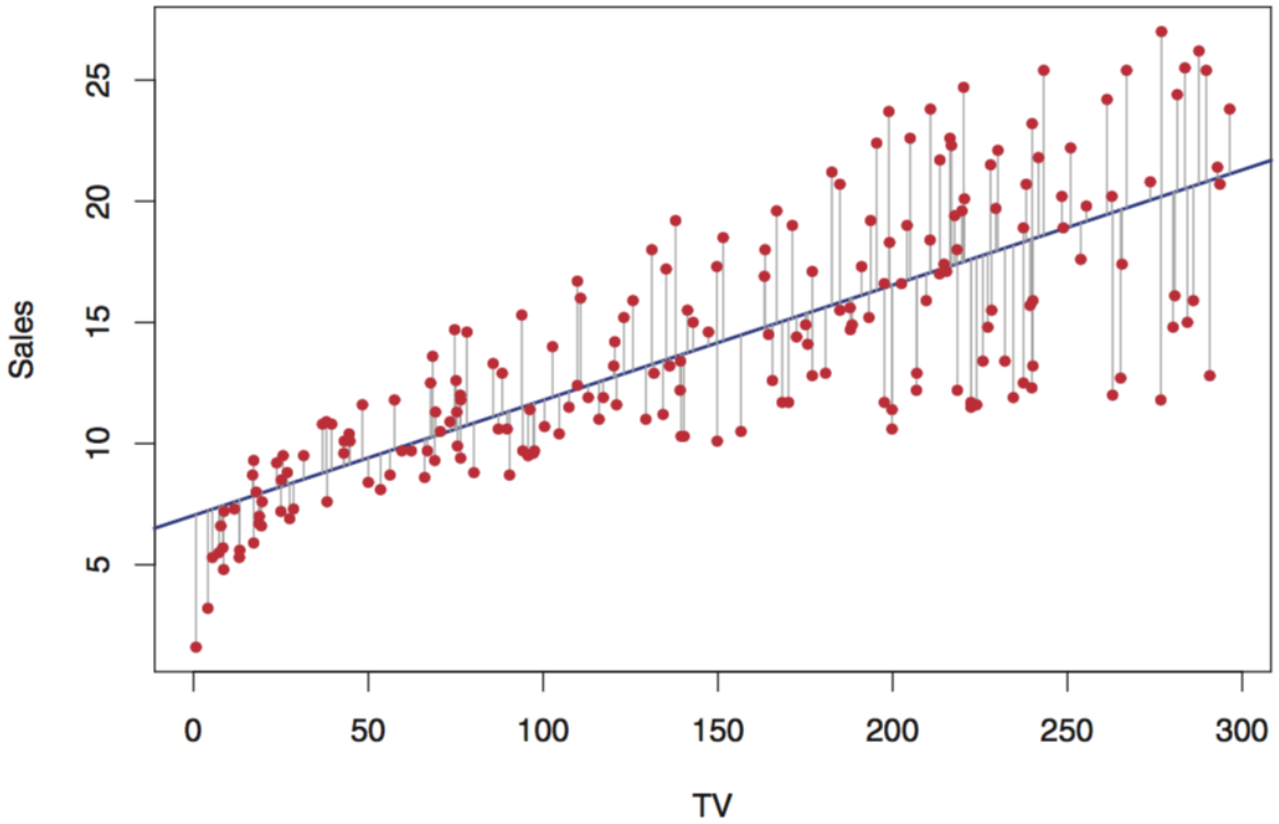
  where $L(\cdot, \cdot)$ is a loss function and $\mathcal{F}$ is a class of choices of g.
- The OLS approach corresponds to $L(a, b) = (a - b)^2$ and $\mathcal{F}(x) = \{x^\top \beta : \beta \in \mathbb{R}^{p+1}\}$.
- In general, the difficulty of solving (1) varies across problems. But, the OLS approach has a unique, closed-form solution:

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

  whenever $\mathbf{X} \in \mathbb{R}^{n \times (p+1)}$ has full column rank.

An example for $p = 1$: Advertising Data

Y : **Sales**, X : **TV** budget, $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i = 7.03 + 0.05 x_i$.



- Each grey segment represents an error. The fitted model compromises by averaging the squared errors.
- A linear fit captures the essence of the relationship, although it is somewhat deficient in the left of the plot.

Some important considerations

- Estimation of β:
  - How close is the point estimation $\hat{\beta}$ to β?
  - (Inference) Can we provide confidence interval / conduct hypothesis testing of β?
- Prediction of Y at $X = x$:
  - How accurate is the point prediction $\hat{y} = x^\top \hat{\beta}$?
  - (Inference) Can we further provide confidence interval of Y?
- Variable (Model) selection:
  - Do all the predictors help to explain Y, or is there only a subset of the predictors useful?
  - Later in Lecture 3

Property of $\hat{\beta} = (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \mathbf{y}$

Take the *design matrix* $\mathbf{X}$ to be deterministic with full column rank. Assume $\epsilon_1, \ldots, \epsilon_n$ are independent with $\mathbb{E}[\epsilon_i] = 0$ and $\mathrm{Var}(\epsilon_i) = \sigma^2$.

- Unbiasedness: $\mathbb{E}[\hat{\beta}] = \beta$
- The covariance matrix of $\hat{\beta}$ is:

$$\mathrm{Cov}(\hat{\beta}) = \sigma^2 (\mathbf{X}^\top \mathbf{X})^{-1}$$

- The above two properties imply the ℓ_2 estimation error

$$\mathbb{E}\left[\|\hat{\beta} - \beta\|_2^2 \right] = \sigma^2 \,\mathrm{Tr}\left[(\mathbf{X}^\top \mathbf{X})^{-1} \right]$$

When $\mathbf{X}^\top \mathbf{X} = n\, \mathbf{I}_{p+1}$,

$$\mathbb{E}\left[\|\hat{\beta} - \beta\|_2^2 \right] = \frac{\sigma^2 (p+1)}{n}.$$

The MSE of estimating β increases as p gets larger.

Inference on β

- The *unknown* variance σ^2 may be estimated by

$$\hat{\sigma}^2 = \frac{1}{n-p-1}\sum_{i=1}^{n}\left(y_i - x_i^\top \hat{\beta}\right)^2.$$

- A 95% confidence interval of β_j has the form of

$$\left[\hat{\beta}_j - 1.96 \cdot SE(\hat{\beta}_j),\quad \hat{\beta}_j + 1.96 \cdot SE(\hat{\beta}_j)\right],$$

where

$$SE(\hat{\beta}_j) = \sqrt{\hat{\sigma}^2[(\mathbf{X}^\top\mathbf{X})^{-1}]_{jj}}$$

- Hypothesis testing

$H_0 : \beta_j = 0$ (There is no linear relationship between Y and X_j)

vs

$H_1 : \beta_j \neq 0$ (There is linear relationship between Y and X_j)

Inference on β

We base on the t-statistic

$$t = \frac{\hat{\beta}_j}{SE(\hat{\beta}_j)}$$

- This has a t-distribution with $n - p - 1$ degrees of freedom, when $\beta_j = 0$.
- Using statistical software, it is easy to compute the probability of observing any value equal to $|t|$ or larger. We call this probability the ***p*-value**.
- In most applications, we reject the null hypothesis if the p-value ≤ 0.05.
- Can be generalized to

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

via F-statistics. (c.f. pp 75-78 of the textbook.)

Results for Advertising Data

Y : **Sales**, X : **TV** budget, $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i = 7.0325 + 0.0475 x_i$.

	Coefficient	Std. Error	t-statistic	p-value
`Intercept`	7.0325	0.4578	15.36	< 0.0001
`TV`	0.0475	0.0027	17.67	< 0.0001

- The p-value for **TV** is smaller than 0.05, so that we reject the null hypothesis $\beta_1 = 0$.
- This indicates that **TV** is significant for predicting **Sales**.

Property of the prediction $\hat{y} = x^\top \hat{\beta}$ at $X = x$

The property of $\hat{\beta}$ can be used to analyze the prediction $\hat{y}$ of y.

- Expectation

$$\mathbb{E}[\hat{y} \mid X = x] = x^\top \mathbb{E}[\hat{\beta}] = x^\top \beta$$

- Variance

$$\text{Var}[\hat{y} \mid X = x] = x^\top \text{Cov}(\hat{\beta}) x = \sigma^2 x^\top (\mathbf{X}^\top \mathbf{X})^{-1} x.$$

- MSE

$$\mathbb{E}[(y - \hat{y})^2 \mid X = x] = \sigma^2 + \sigma^2 x^\top (\mathbf{X}^\top \mathbf{X})^{-1} x.$$

Other considerations in linear regression models

- The coefficient of determination: R^2
- Qualitative Predictors
- Extend to non-linearity
  - Adding interaction terms
  - Adding transformed predictors
- Model diagnosis

The coefficient of determination: R-squared (R^2)

Meaning of R-squared

R^2 is the proportion of the variation in the outcome (Y) that can be explained from the predictors (X).

Recall that for each training point (x_i, y_i), its fitted value is

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_p x_{ip}.$$

Its residual is defined as

$$e_i = y_i - \hat{y}_i.$$

The residual sum of squares is

$$RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2,$$

nothing but the training MSE at $(\hat{\beta}_0, \ldots, \hat{\beta}_p)$.

The coefficient of determination: R^2

- The total sum of squares is

$$TSS = \sum_{i=1}^{n} (y_i - \bar{y})^2, \quad \text{with} \quad \bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i,$$

  quantifying the total variance of Y in the sample $(y_1, \ldots, y_n)$.

- R^2 measures the proportion of variability in Y that can be explained by regressing Y onto X.

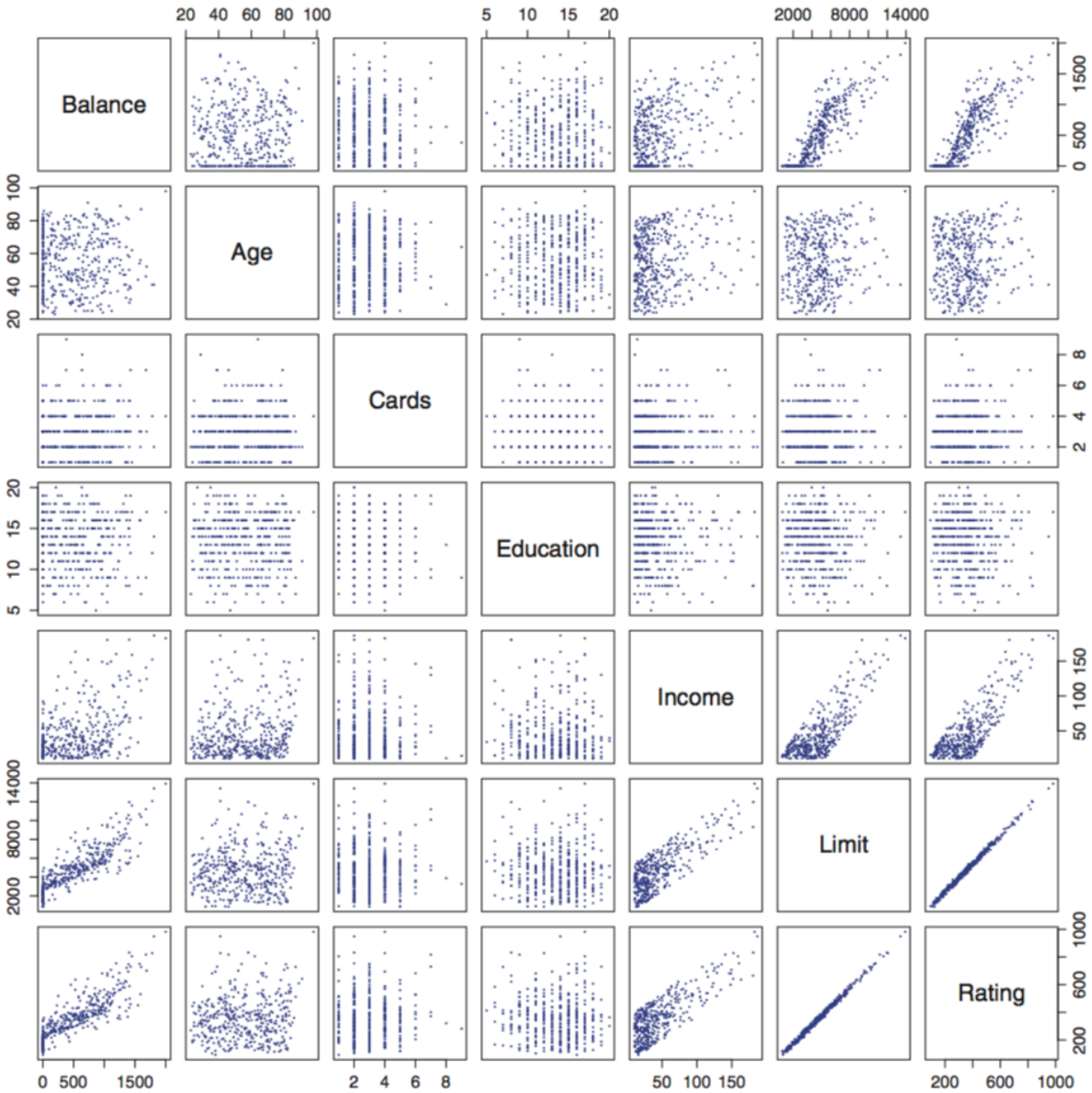
$$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS}.$$

- $0 \leq R^2 \leq 1$. R^2 *close to* 1 indicates a large proportion of the variability in the response that is explained by the predictors.
- However, a large value of R^2 does **NOT** imply that the model fits the data well. It always favors more flexible models, which may overfit the data! (Adjusted R^2 later.)

Qualitative Predictors

- Some predictors are not quantitative but are qualitative, taking a discrete set of values.
- These are also called **categorical predictors** or **factor variables**.
- See for example the scatterplot matrix of the credit card data.

Credit Card Data



Credit Card Data

In addition to the 7 quantitative variables, there are four qualitative variables:

- gender
- student (student status)
- status (marital status)
- ethnicity (Caucasian, African American (AA) or Asian).

Qualitative predictors with two levels

Example (study the difference in credit card balance between males and females, ignoring the other variables)

We create a new **dummay variable** of the predictor (gender):

$$x_i = \begin{cases} 1 & \text{if } i\text{th person is female} \\ 0 & \text{if } i\text{th person is male} \end{cases}$$

Resulting model:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if } i\text{th person is female} \\ \beta_0 + \epsilon_i & \text{if } i\text{th person is male.} \end{cases}$$

Qualitative predictors with more than two levels

With more than two levels, we create additional dummy variables. For example, for the ethnicity variable we create two dummy variables. The first could be

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

and the second could be

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

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

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if } i\text{th person is Asian} \\ \beta_0 + \beta_2 + \epsilon_i & \text{if } i\text{th person is Caucasian} \\ \beta_0 + \epsilon_i & \text{if } i\text{th person is AA.} \end{cases}$$

- There are always one fewer dummy variables than the number of levels.
- The level when all dummy variables are 0 – African American in this example – is known as the baseline.

Credit Card Data

	Coefficient	Std. error	t-statistic	p-value
`Intercept`	531.00	46.32	11.464	< 0.0001
`ethnicity[Asian]`	−18.69	65.02	−0.287	0.7740
`ethnicity[Caucasian]`	−12.50	56.68	−0.221	0.8260

Interpretation: The Asian category tends to have 18.69 less debt than the AA category, and that the Caucasian category tends to have 12.50 less debt than the AA category.

Extension to non-linearity: adding **interaction** terms

- Consider the model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon.$$

  Regardless of the value of X_2, one-unit increase in X_1 will lead to β_1-unit increase in Y.

- Consider the model with **interaction** terms

$$\begin{aligned} Y &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \epsilon \\ &= \beta_0 + (\underbrace{\beta_1 + \beta_3 X_2}_{\tilde{\beta}_1}) X_1 + \beta_2 X_2 + \epsilon. \end{aligned}$$

  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.

- β_1 and β_2 are the coefficients of the **main effects** while β_3 is that of the **interaction**.

Interaction

Example (Gender + Education)

Consider

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

where

$$x_{i1} = \begin{cases} 1 & \text{if the } i\text{the person is female} \\ 0 & \text{if the } i\text{the person is male} \end{cases}$$

x_{i2} = the number of years of education.

Interpretation of β_2: one more year education leads to β_2-unit change in credit card balance with gender held fixed.

Example (Gender + Education)

Now consider

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1} x_{i2} + \epsilon_i$$
$$= \begin{cases} \beta_0 + \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i2} + \epsilon_i, & \text{if the } i\text{the person is female} \\ \beta_0 + \beta_2 x_{i2} + \epsilon_i, & \text{if the } i\text{the person is male} \end{cases}$$

where

$$x_{i1} = \begin{cases} 1 & \text{if female} \\ 0 & \text{if male} \end{cases}, \quad x_{i2} = \text{the number of years of education.}$$

- **Interpretation of β_2**: one more year education leads to β_2-unit change in credit card balance with for male.
- **Interpretation of β_3**: for one more year education, the difference in credit card balance between female and male is β_3.
- **How about $\beta_3 + \beta_2$?**

Interactions

Read pages 89-90 of the textbook for more examples.

Hierarchy at the presence of interactions

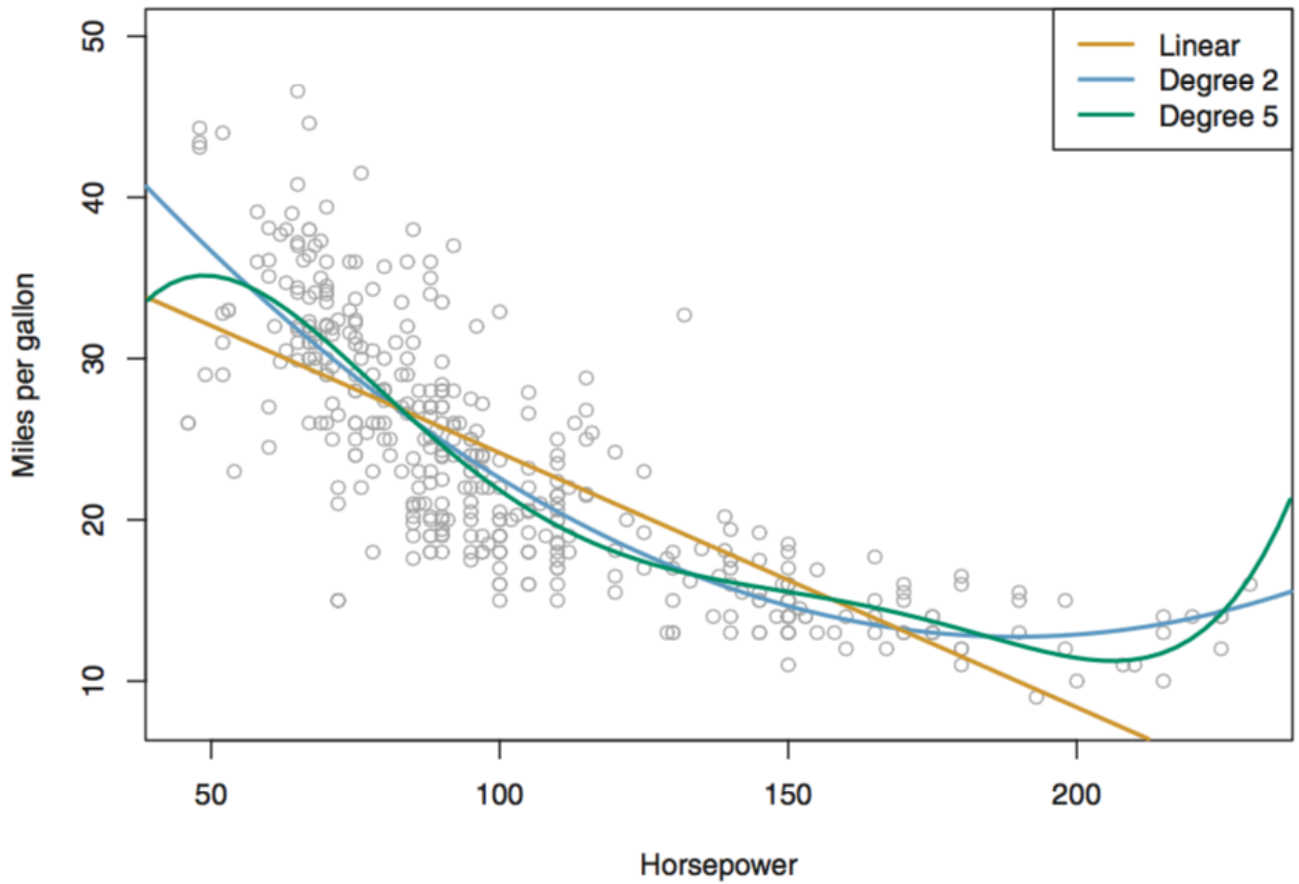
- Sometimes it is the case that an interaction term has a very small p-value whereas the associated **main effects** (in the Advertising Data, TV and radio) have large p-values.

- **Hierarchy principle**:

  If we include an interaction term X_1X_2 in the model, we should also include the main effects X_1 and X_2, even if the p-values associated with their coefficients are not significant.

Extention to non-linearity: adding transformed predictors

For a number of cars, their **mpg** and **horsepower** are shown in the figure.



The linear regression (orange); the linear regression fit for a model that includes **horsepower**^{2} (blue); the linear regression fit for a model that includes all polynomials of **horsepower** up to 5th-degree (green).

Non-linearity

The figure suggests that

$$mpg = \beta_0 + \beta_1 \times horsepower + \beta_2 \times horsepower^2 + \epsilon,$$

may provide a better fit.

	Coefficient	Std. Error	t-statistic	p-value
Intercept	56.9001	1.8004	31.6	< 0.0001
horsepower	-0.4662	0.0311	-15.0	< 0.0001
horsepower^2	0.0012	0.0001	10.1	< 0.0001

- A simple approach for incorporating non-linear associations in a linear model is to include **transformed versions of the predictors** in the model.
- By the end of the day, it is still a linear model!
  Can be fitted by least squared with $X_1 = horsepower$, and $X_2 = horsepower^2$.

Diagnosis of Linear Models

- Non-linearity of the response-predictor relationships.
- Correlation of the error terms among training samples.
- Non-constant variance of error terms.
- Outliers.
- Collinearity.

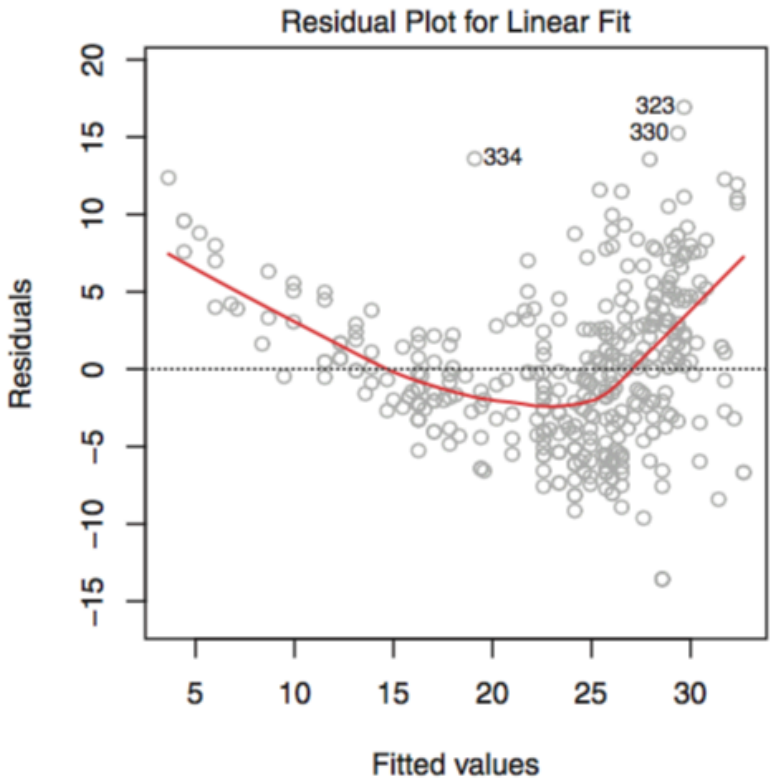
Check Non-linearity

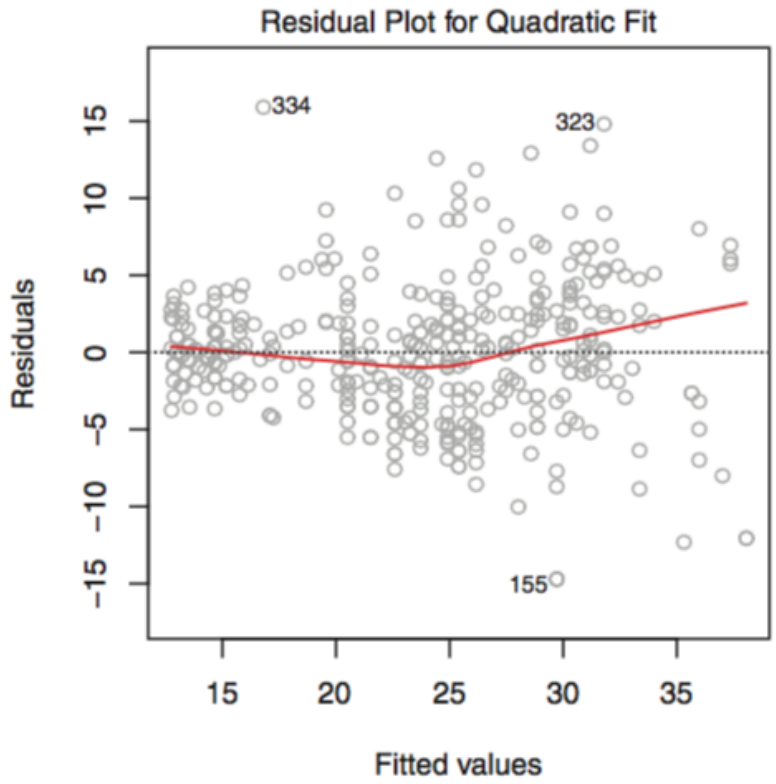
Recall

$$\hat{y}_i = x_i^\top \hat{\boldsymbol{\beta}}, \qquad \forall\, i \in \{1, \ldots, n\}.$$

- For a fitted linear regression model, we plot its residuals, $e_i = y_i - \hat{y}_i$ versus the fitted value $\hat{y}_i$.
- If the linear model works well, there should be no apparent patterns in the plot (points randomly centered around 0).
- If there seems to be certain pattern, we could consider adding non-linear transformation of the predictors to the model. (E.g. X^2 or $\log X$ for the univariate predictor).

Car Data



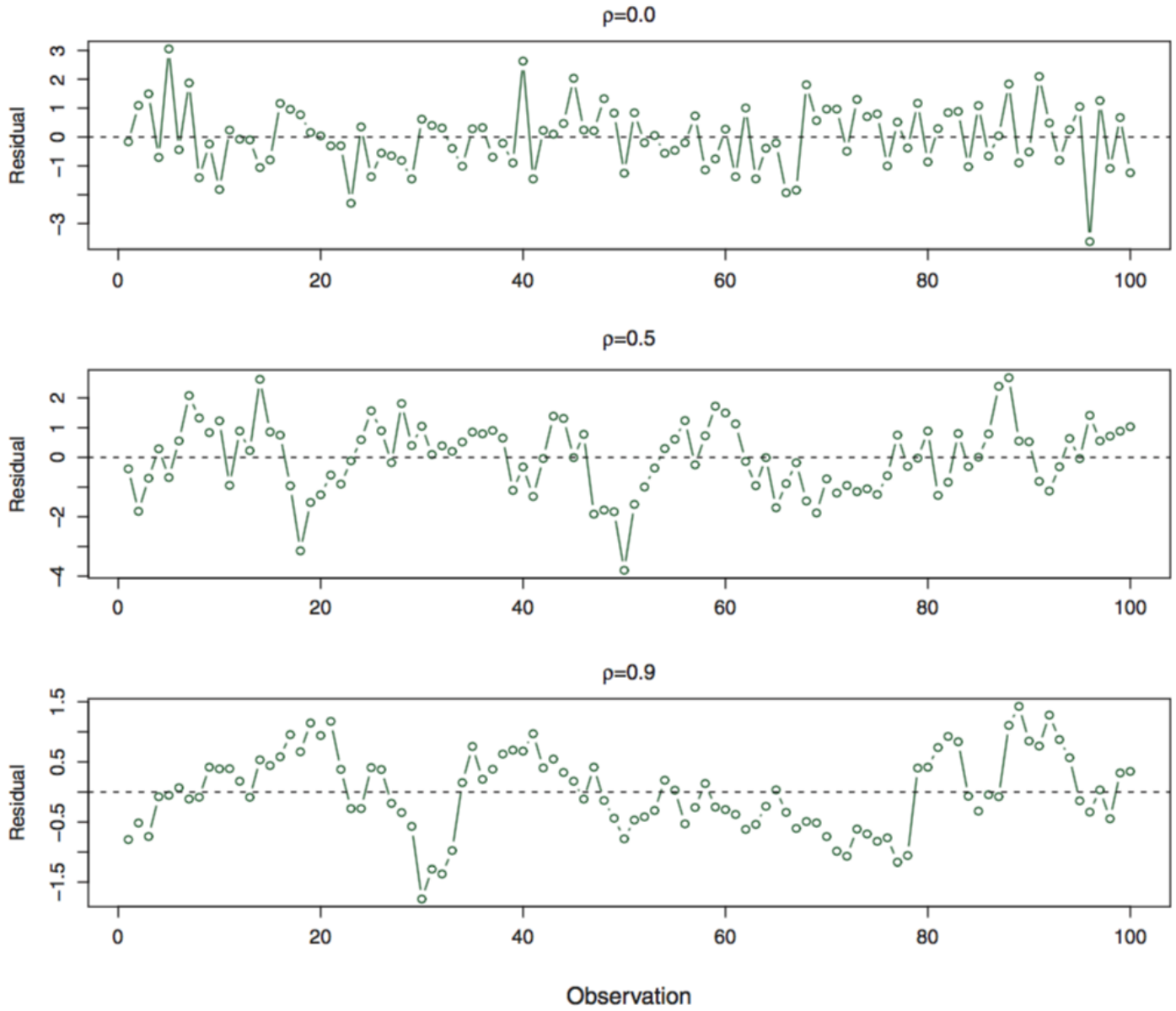


- Left: A linear regression of **mpg** on **horsepower**. A strong pattern in the residuals indicates non-linearity in the data.
- Right: A linear regression of **mpg** on **horsepower** and **horsepower**^{2}. No clear pattern in the residuals.

Correlation of Error Terms

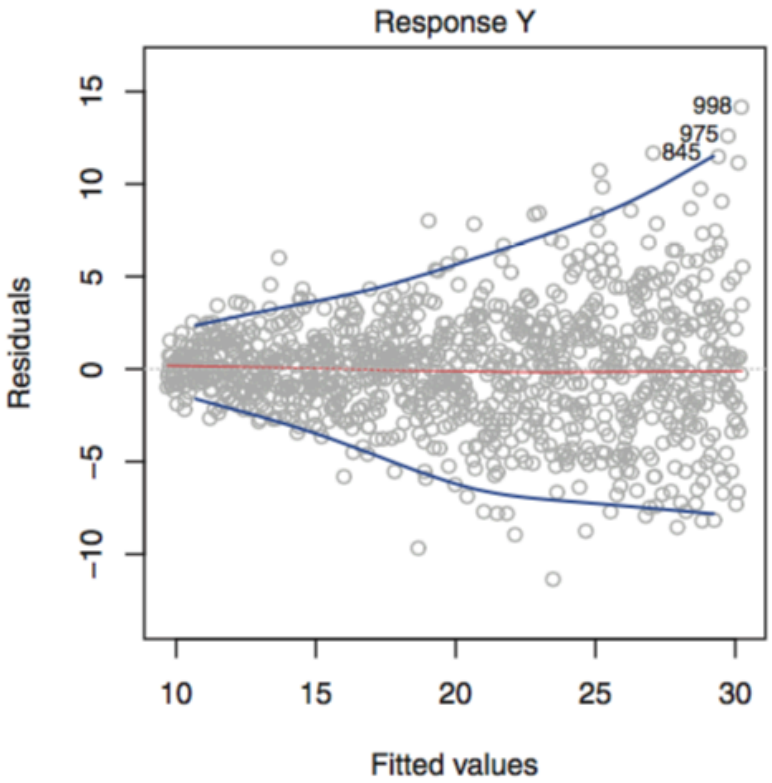
- In the linear regression model, the error terms, $\epsilon_1, \epsilon_2, ..., \epsilon_n$ are uncorrelated.
- If residuals are correlated, the estimated standard error $(\hat{\sigma}^2)$ will not be close to the true standard error (σ^2).
  So, the resulting confidence interval or p-values will not be accurate.
- If the samples are independent, then the uncorrelateness can be justified.
- In practice, we could still check this via plotting all the residuals.

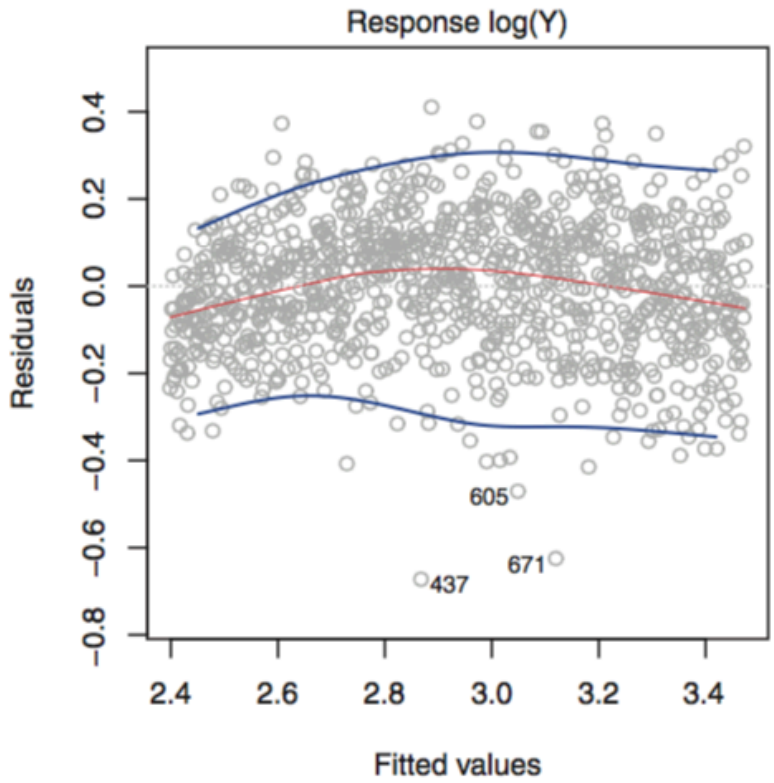
Example


ρ=0.0
Residual
−3 −1 0 1 2 3
0 20 40 60 80 100
ρ=0.5
Residual
−4 −2 0 1 2
0 20 40 60 80 100
ρ=0.9
Residual
−1.5 −0.5 0.5 1.5
0 20 40 60 80 100
Observation

Non-constant Variance of the Error Terms

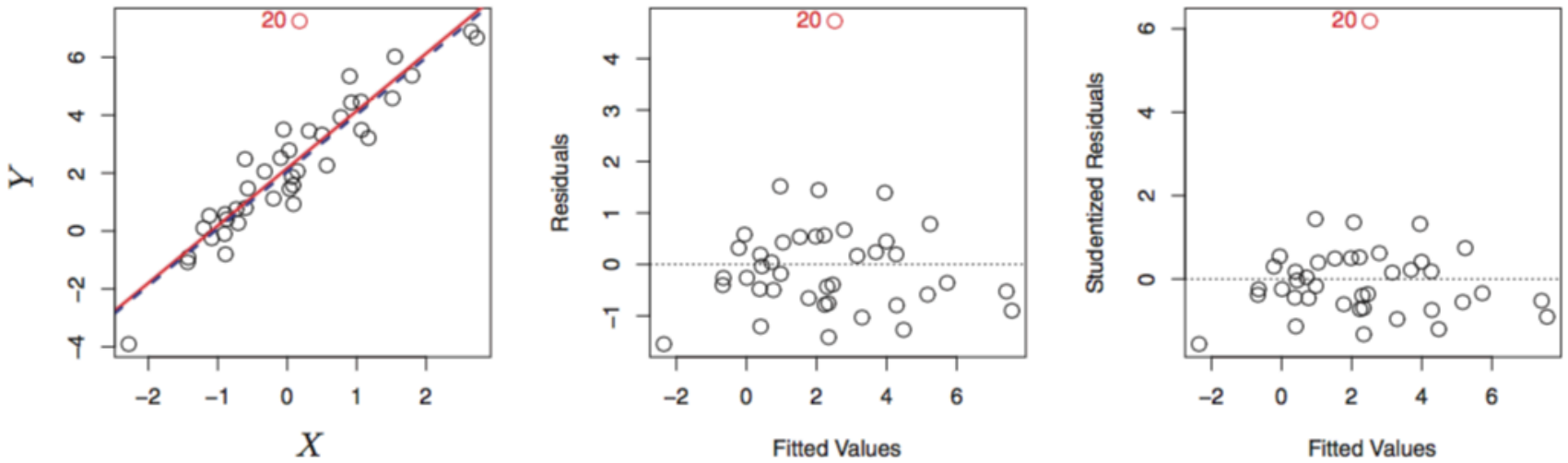
- We assume $\text{Var}(\epsilon_i) = \sigma^2$ for all $1 \leq i \leq n$.
- In the residual plot, if we see the variances of residual change with the fitted value $\hat{y}_i$. This phenomenon is known as **heteroscedasticity**.
- What can we do? We transform Y (e.g., $\log Y$).





Detection of Outliers

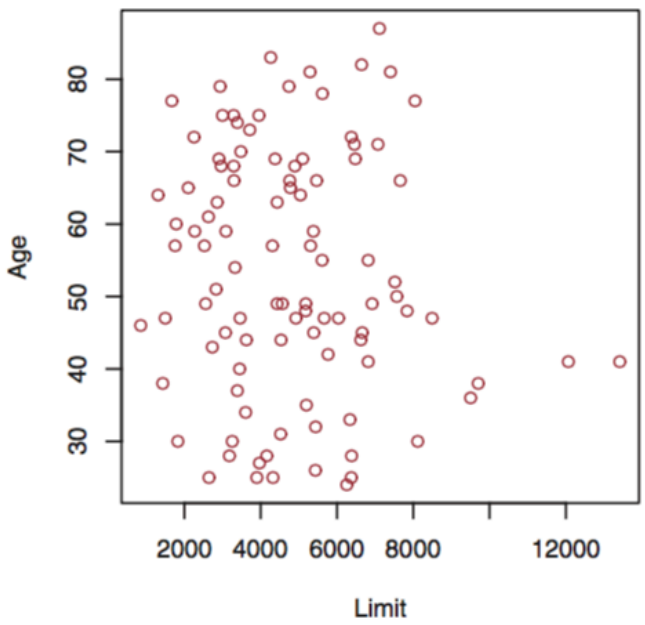
- An outlier is a point for which y_i is far from the value predicted by the model.
- How to find outliers? Calculate the studentized residuals, computed by $e_i/\hat{\sigma}$, where $\hat{\sigma}^2$ is the estimate of the error variance σ^2.
- Usually, observations whose studentized residuals are greater than 3 in absolute value are possible outliers.

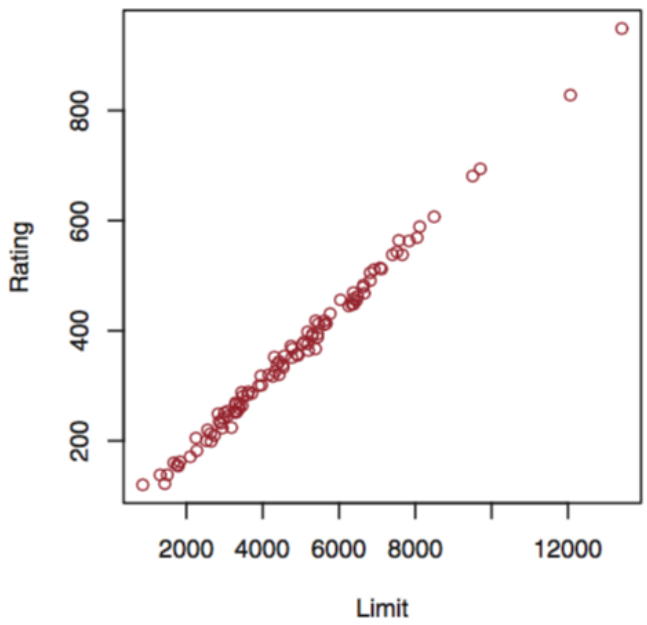


Collinearity

- Collinearity refers to the situation in which two or more predictors are highly correlated.
- If two predictors tend to increase or decrease together, it is difficult to determine how each of them is associated with the response. Furthermore, the variance of the OLS estimator gets larger.
- How to detect collinearity?
  - examine the correlation matrix of $X_1, ..., X_p$.
  - use variance inflation factor, VIF (more sophisticated, we will not discuss it further).
- How to handle collinearity? If X_1 and X_2 are collinearity,
  - drop one of X_1 and X_2 in the regression.
  - combine X_1 and X_2 (e.g., take the average, but can be difficult to interpret).
  - ridge regularization (later in Lecture 4).

Credit Card Data





		Coefficient	Std. error	t-statistic	p-value
Model 1	`Intercept`	−173.411	43.828	−3.957	< 0.0001
	`age`	−2.292	0.672	−3.407	0.0007
	`limit`	0.173	0.005	34.496	< 0.0001
Model 2	`Intercept`	−377.537	45.254	−8.343	< 0.0001
	`rating`	2.202	0.952	2.312	0.0213
	`limit`	0.025	0.064	0.384	0.7012

So far we have covered many aspects of the OLS estimator under the linear regression model

$$Y = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p + \epsilon.$$

Question: is it always a good choice?