STA 314: Statistical Methods for Machine Learning I

Lecture 3 - Linear models and model selection

Xin Bing

Department of Statistical Sciences
University of Toronto

Review

- We have learned the bias-variance-tradeoff:
  - As the complexity of the fitted model increases, its bias decreases while its variance increases.
  - The variance of any fitted model is roughly proportional to

$$\frac{\text{complexity of } \hat{f}}{n}.$$

  - When the sample size (n) is limited, a fitted model with high complexity performs poorly due to large variance.
  - When n is large enough, a more complex fitted model tends to peform better as they have smaller bias than simpler models.

Review

- We have learned the OLS approach:

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\alpha}}{\operatorname{argmin}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\alpha}\|_2^2.$$

- We have learned the statistical properties of $\hat{\boldsymbol{\beta}}$ under the linear model

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

  - Unbiasedness
  - Estimation error (ℓ_2)
  - Inference (confidence intervals, hypothesis testing).

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.

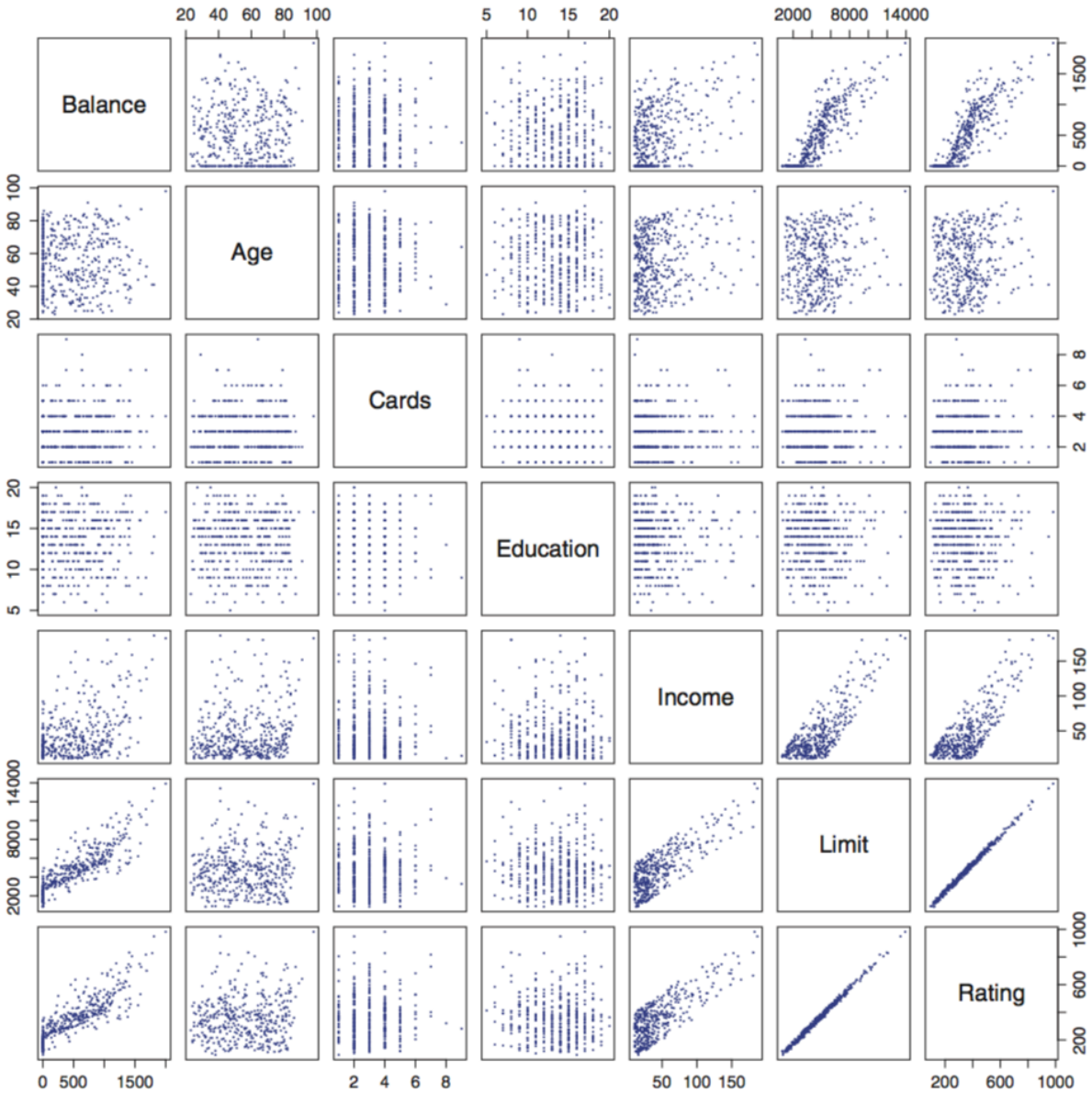
$$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS}.$$

- $0 \leq R^2 \leq 1$. 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

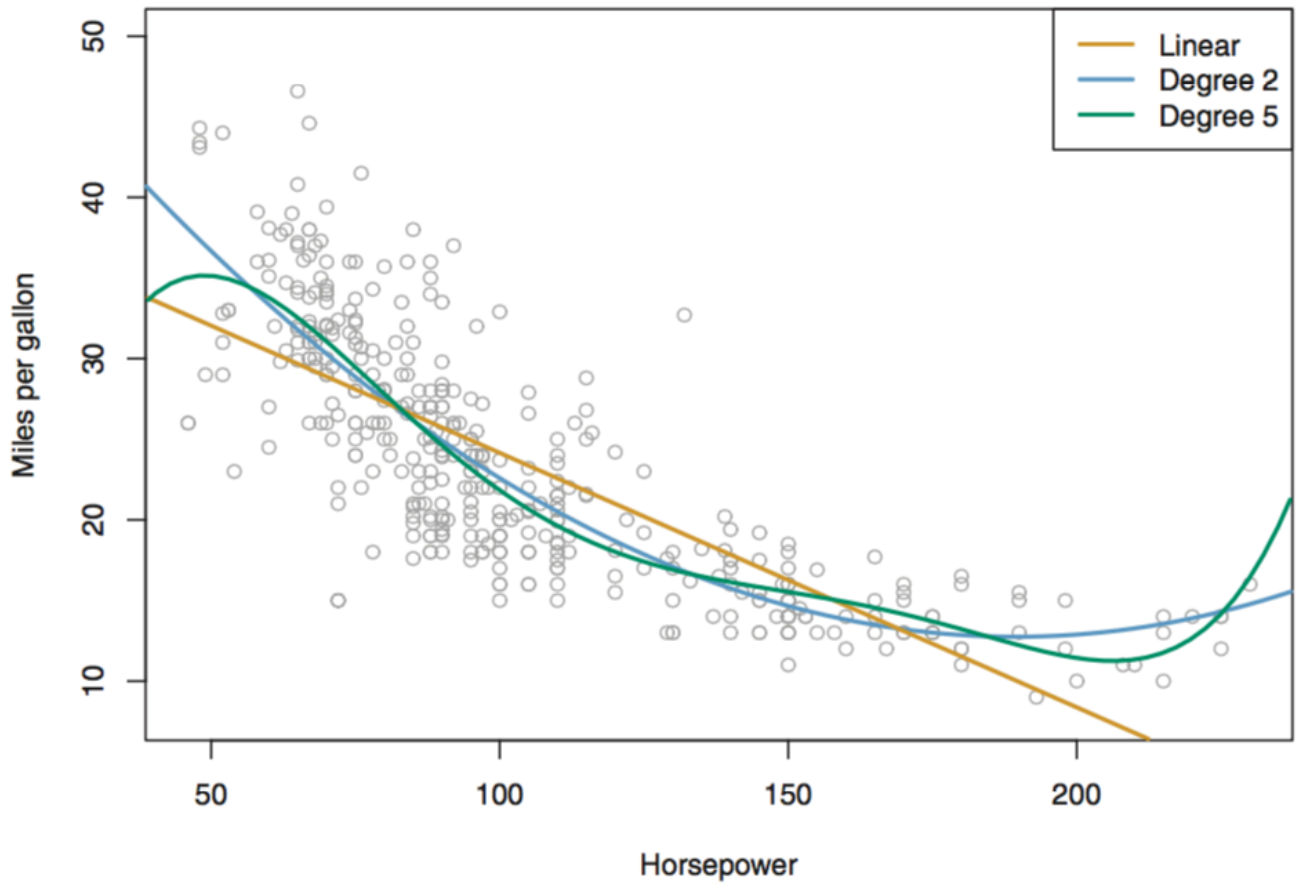
- Sometimes it is the case that an interaction term has a very small p-value whereas the associated **main effects** (in the Advertising Data, TV and radio) have large p-values.

- **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.

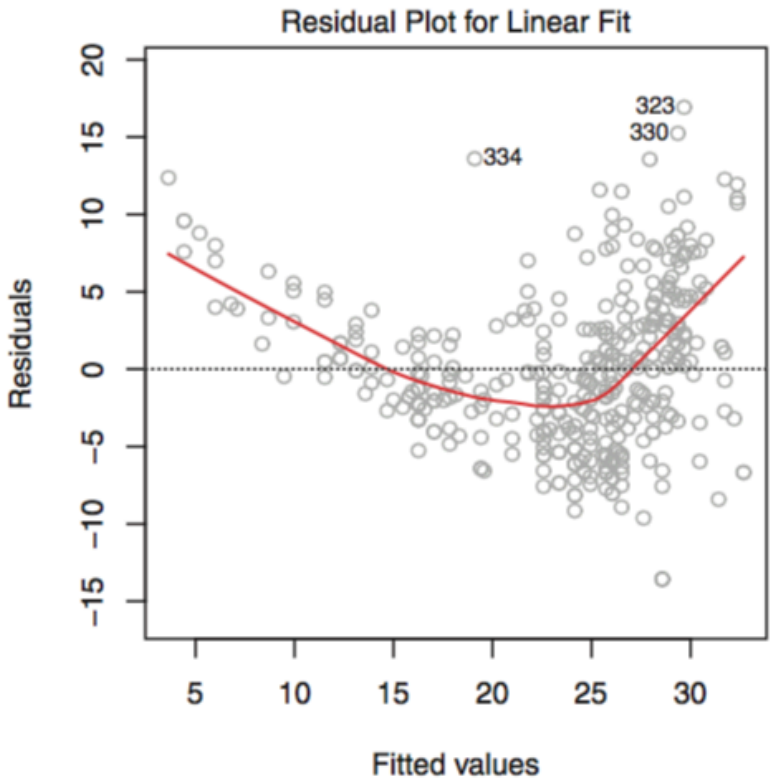
Check Non-linearity

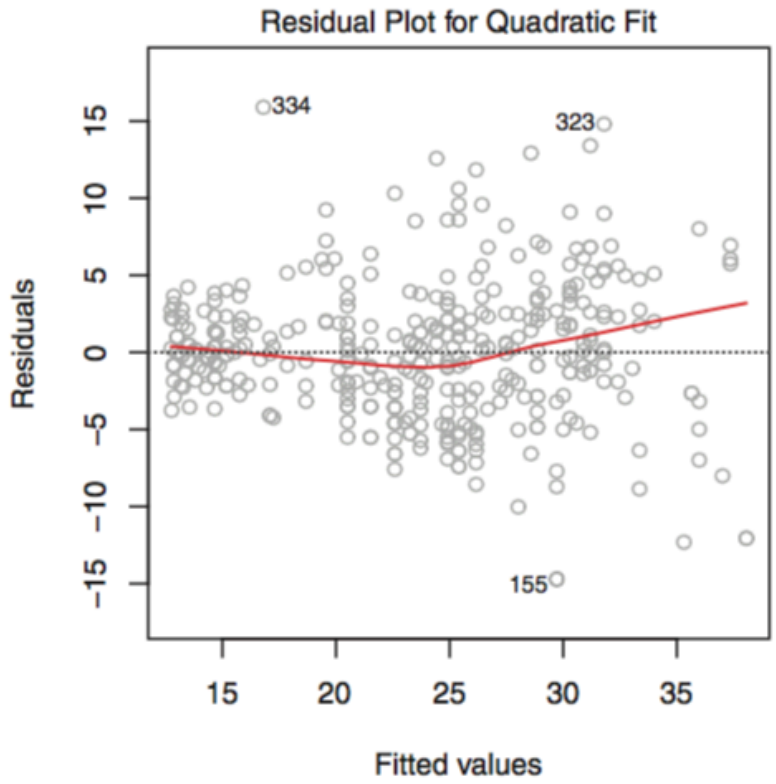
Recall

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

- 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



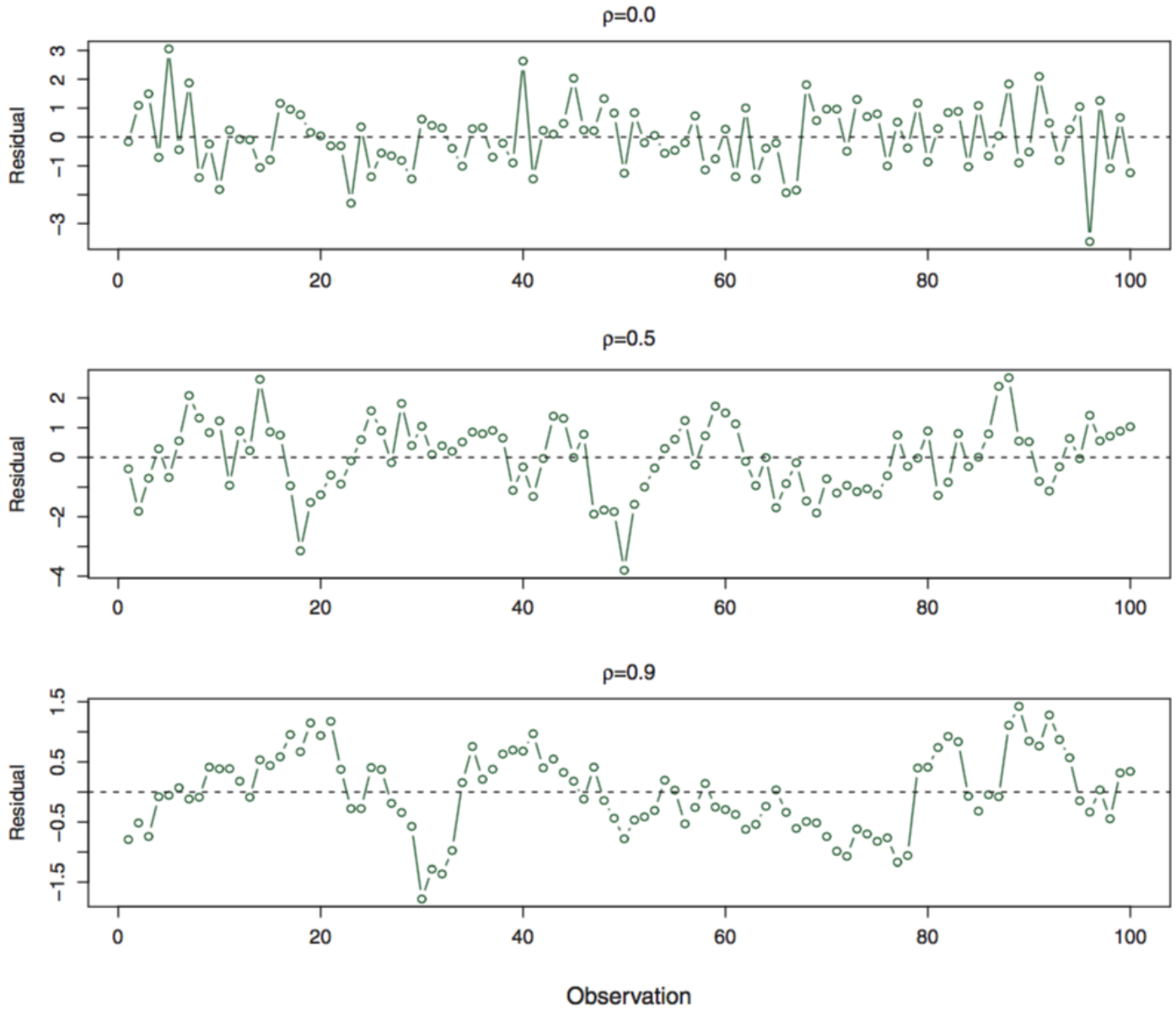


- Left: A linear regression of **mpg** on **horsepower**. A strong pattern in the residuals indicates non-linearity in the data.
- Right: A linear regression of **mpg** on **horsepower** and **horsepower**^{2}. No clear pattern in the residuals.

Correlation of Error Terms

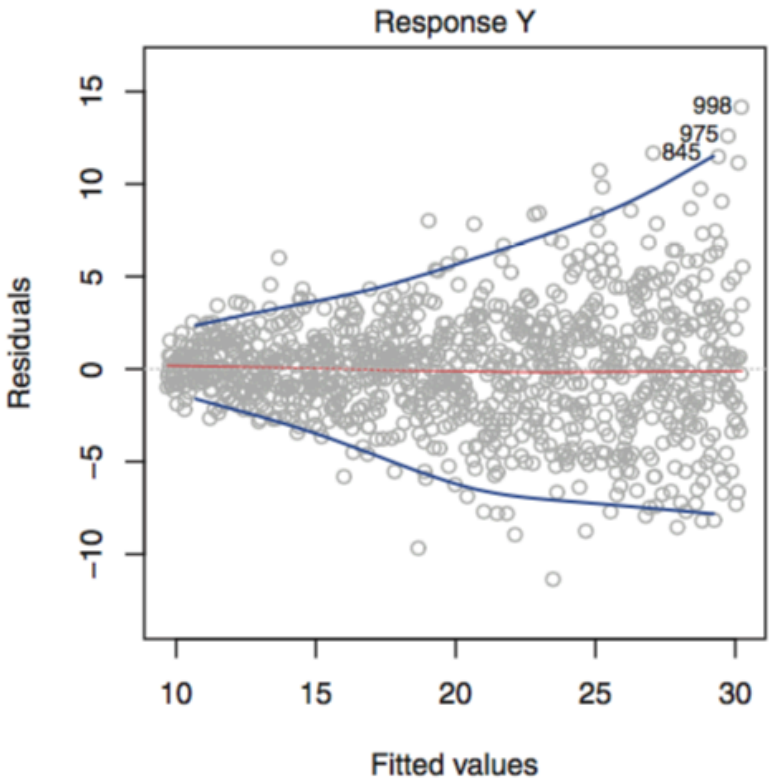
- In the linear regression model, the error terms, $\epsilon_1, \epsilon_2, ..., \epsilon_n$ are uncorrelated.
- If residuals are correlated, the estimated standard error $(\hat{\sigma}^2)$ will not be close to the true standard error (σ^2).
  So, the resulting confidence interval or p-values will not 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.

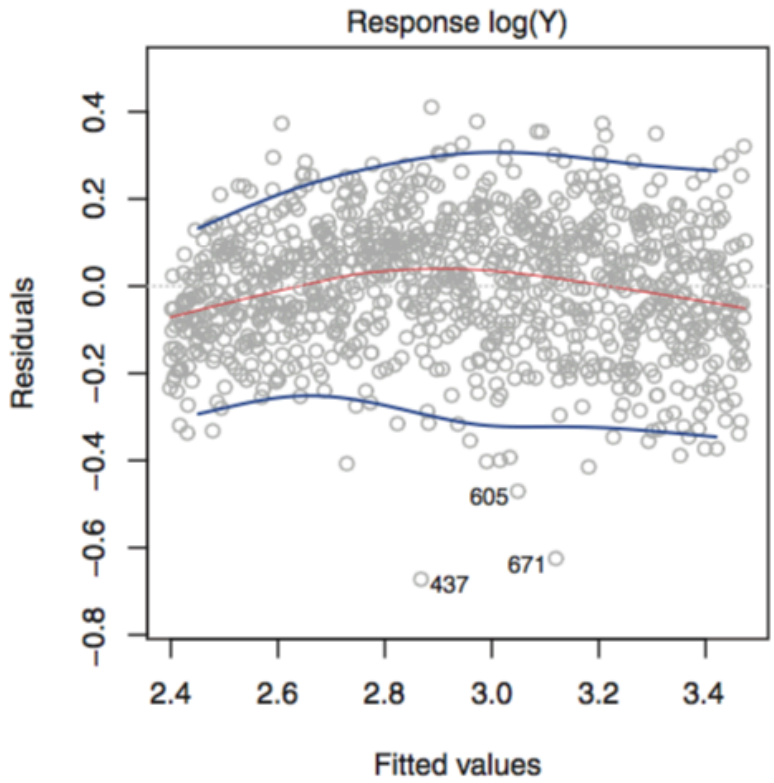
Example


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

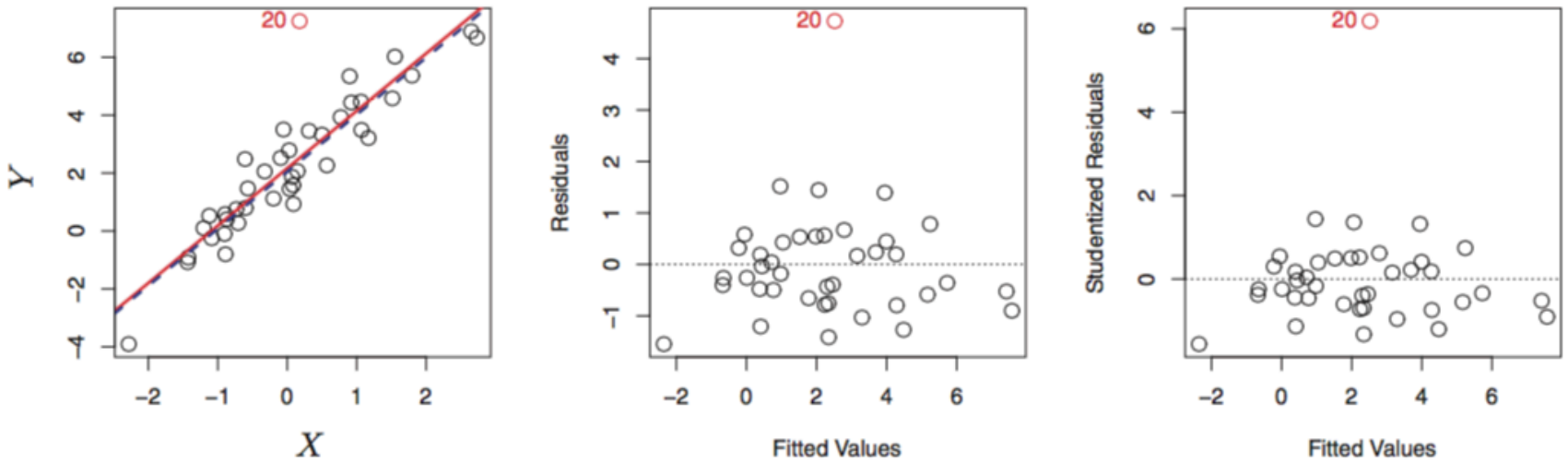
- We assume $\text{Var}(\epsilon_i) = \sigma^2$ for all $1 \leq i \leq n$.
- In the residual plot, if we see the variances of residual change with the fitted value $\hat{y}_i$. This phenomenon is known as **heteroscedasticity**.
- What can we do? We transform Y (e.g., $\log Y$).





Detection of Outliers

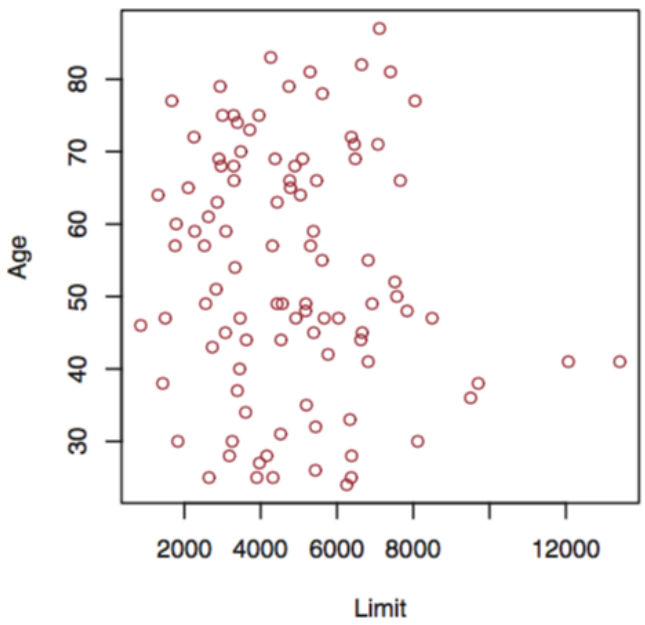
- An outlier is a point for which y_i is far from the value predicted by the model.
- How to find outliers? Calculate the studentized residuals, computed by $e_i/\hat{\sigma}$, where $\hat{\sigma}^2$ is the estimate of the error variance σ^2.
- Usually, observations whose studentized residuals are greater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

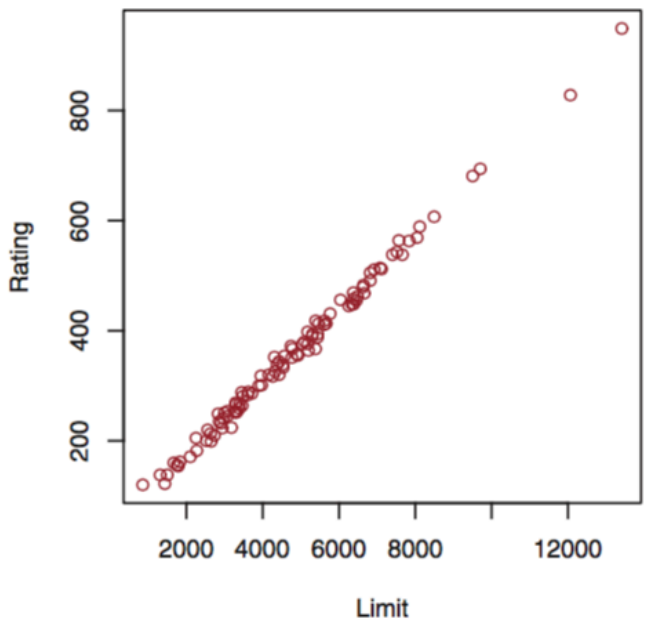


Collinearity

- 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?

Why consider alternatives to the OLS estimator?

Alternative fitting procedures to OLS could yield **better prediction accuracy** and **model interpretability**.

- Prediction / Estimation: the OLS estimator

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

  has large variance when p is large. Especially, if $p > n$, then OLS estimator is not unique and its variance is infinite.

- Interpretability: By removing irrelevant features – that is, by setting some coefficient estimates to zero – we can obtain a model that is more parsimonious hence more interpretable.

What are the alternatives?

- **Subset Selection**. We identify a subset of the p predictors that we believe to be related to the response. We then fit a model using the OLS approach on the identified set of predictors.
- **Shrinkage**. We fit a model involving all p predictors, but the estimated coefficients are shrunken towards zero relative to the OLS estimator. This shrinkage (also known as regularization) has the effect of reducing variance. Some could also perform variable selection.
- **Dimension Reduction**. We project the p predictors into a M-dimensional subspace, where $M < p$. This is achieved by computing M different linear combinations, or projections, of the original predictors. Then the resulting M projections are used as new predictors to fit a linear regression model by OLS.

How to choose the optimal one among a set of models?

Example

$$\text{Model 1:} \quad Y = \alpha_0 + \alpha_1 X_1 + \epsilon$$
$$\text{Model 2:} \quad Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

lead to different predictors at $X = x$

$$\hat{y}_1 = \hat{\alpha}_0 + \hat{\alpha}_1 x_1 \qquad \text{v.s.} \qquad \hat{y}_2 = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2.$$

- Ideally, we choose the one that has a **smaller expected MSE**.
- The expected MSE can be approximated when we have $\mathcal{D}_{test}$.

$$MSE_T = \frac{1}{m} \sum_{i=1}^{m} \left(y_{Ti} - \hat{f}(x_{Ti}) \right)^2$$

- What if we don't have $\mathcal{D}_{test}$?

Model selection

There are two common approaches for model selection when we don't have $\mathcal{D}_{test}$:

- We can avoid estimating the expected MSE by making an adjustment to the training error to account for the model complexity:
  - Mallow's C_p
  - AIC
  - BIC
  - adjusted R^2

- We can directly estimate the expected MSE by manually creating a "test set" using data-splitting techniques:
  - validation set approach
  - cross-validation approach

Avoid estimating the expected MSE: C_p, AIC, BIC, and adjusted R^2

- These techniques adjust the training error for the model complexity.
- They are only used to select among a set of parametric models with different numbers of predictors.

For any given fitted model $\hat{f}$, let $\hat{f}(x_i)$ be the fitted value for the i the observation. For instance, for a fitted linear model with p predictors,

$$\hat{f}(x_i) = x_i^\top \hat{\boldsymbol{\beta}} = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_p x_{ip}.$$

Recall that

$$e_i = y_i - \hat{f}(x_i)$$

is the ith residual. The residual sum of squares (RSS) is defined as

$$RSS = RSS(\hat{f}) = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left(y_i - \hat{f}(x_i)\right)^2.$$

Mallow's C_p

Let p be the total # of parameters in the model and $\hat{\sigma}^2$ is an estimate of $\sigma^2 = \text{Var}(\epsilon)$.

$$C_p = \frac{1}{n}(RSS + 2p\hat{\sigma}^2).$$

- Essentially, the C_p adds a penalty $2d\hat{\sigma}^2$ to the training MSE to adjust for the fact that the training error tends to underestimate the test error.
- C_p tends to take on a small value for models with a low test error, so when determining which of a set of models is best, we choose the model with the lowest C_p value.
- C_p is mainly for (linear) fitted models (such as via OLS) in regression problems

AIC and BIC

Let $\hat{f}$ be the fitted model obtained from the MLE approach.
Let $L := L(\hat{f})$ be the maximized value of the likelihood function for $\hat{f}$.

- **AIC**:

$$AIC = -2 \log L + 2p,$$

In the linear model with $\epsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$,

$$AIC = \frac{C_p}{\hat{\sigma}^2}.$$

In this case, AIC is proportional to C_p, selecting the same model.

- **BIC**:

$$BIC = -2 \log L + (\log n)p,$$

BIC places a heavier penalty $(\log n)p$ on models with many predictors, and hence results selecting smaller-size models than AIC and C_p.

AIC and BIC

- For both AIC and BIC, we select the best model that has the lowest value.
- To compute AIC and BIC, we need to specify the likelihood, i.e. the distribution of $Y \mid X$, and to compute the maximum likelihood estimator.
- AIC and BIC can also be used for selecting parametric models in classification problems.

Adjusted R^2

Recall that the total sum of squares (TSS) is defined as

$$TSS = \sum_{i=1}^{n}(y_i - \bar{y}), \qquad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i.$$

The adjusted R^2 is defined as

$$Adjusted\ R^2 = 1 - \frac{RSS/(n-p-1)}{TSS/(n-1)}.$$

- Unlike C_p, AIC, and BIC, for which a **small** value indicates a model with low test error, a **large** value of adjusted R^2 indicates a model with small test error.
- Maximizing the adjusted R^2 is equivalent to minimizing $RSS/(n-p-1)$. While RSS always decreases as the number of variables in the model increases, $RSS/(n-p-1)$ may increase or decrease, due to the presence of p in the denominator.

Adjusted R^2 vs R^2

Recall that

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

By contrast,

$$Adjusted\ R^2 = 1 - \frac{RSS/(n-p-1)}{TSS/(n-1)}.$$

Remark. Unlike the R^2 statistic, the adjusted R^2 statistic pays a price for the inclusion of unnecessary variables in the model.