Model Selection and Assumptions

November 15, 2019

Forward Selection

- **Forward selection** is essentially backward selection in reverse.
- We start with the model with no variables.
- We use R^2_{adj} to add one variable at a time.
- We continue to do this until we cannot improve R^2_{adj} any further.

Example: Forward Selection

We start with the intercept-only model and (one at a time) examine the model using

Add ...	`income_ver`	`debt_to_income`	`credit_util`	`bankruptcy`
	$R^2_{adj} = 0.05926$	$R^2_{adj} = 0.01946$	$R^2_{adj} = 0.06452$	$R^2_{adj} = 0.00222$
	`term`	`issued`	`credit_checks`	
	$R^2_{adj} = 0.12855$	$R^2_{adj} = -¡0.00018$	$R^2_{adj} = 0.01711$	

to predict interest rate.

- $R^2_{adj} = 0$ for the intercept-only model.

Example: Forward Selection

- We see the biggest improvement with `term`.
- We then check all of the models with `term` and each other variable.
- Our new baseline R^2_{adj} is 0.12855.

Add `term` and ...			
	`income_ver` $R^2_{adj} = 0.16851$	`debt_to_income` $R^2_{adj} = 0.14368$	`credit_util` $R^2_{adj} = 0.20046$
	`bankruptcy` $R^2_{adj} = 0.13070$	`issued` $R^2_{adj} = 0.12840$	`credit_checks` $R^2_{adj} = 0.14294$

Example: Forward Selection

Moving forward with `term` and `credit_util` (new baseline $R^2_{adj} = 0.20046$)

Add `term`, `credit_util`, and ...	`income_ver` $R^2_{adj} = 0.24183$	`debt_to_income` $R^2_{adj} = 0.20810$	
	`bankruptcy` $R^2_{adj} = 0.20169$	`issued` $R^2_{adj} = 0.20031$	`credit_checks` $R^2_{adj} = 0.21629$

So we will include `income_var`.

Continuing on, we include `debt_to_income`, then `credit_checks`, and `bankruptcy`.

Example: Forward Selection

At this point, we have only `income` left.

- The current R^2_{adj} is 0.25854.
- Including `income`, we find $R^2_{adj} = 0.25843$.

We conclude with the same model we found in the backward elimination.

Model Selection: the P-Value Approach

The p-value may be used instead of R^2_{adj}. For backward elimination

- Build the full model and find the predictor with the largest p-value.
- If the p-value$> \alpha$, remove it and refit the model.
- Repeat with the smaller model.
- When all p-values$< \alpha$, STOP. This is your final model.

Note: it is still important that we remove only one variable at a time!

Model Selection: the P-Value Approach

The p-value may be used instead of R^2_{adj}. For forward selection

- Fit a model for each possible predictor and identify the model with the smallest p-value.
- If that p-value$< \alpha$, add that predictor to the model.
- Repeat, building models with the chosen predictor and each additional potential predictor.
- When none of the remaining predictors have p-value$< \alpha$, STOP. This is the final model.

Note: it is still important that we add only one variable at a time!

Model Selection: R^2_{adj} or P-Value?

- When the primary goal is prediction accuracy, use R^2_{adj}.
  - This is typically the case in machine learning applications.
- When the primary goal is understanding statistical significance, use p-values.

Model Selection: Backward or Forward?

- Both are perfectly valid approaches.
- Statistical software like `R` can automate either process.
- If you have a lot of predictor variables, forward selection may make things easier.
  - Note: we can't fit models where $k \geq n$.
  - In this setting, forward selection may help us choose which variables to include.
- If you have fewer predictor variables, backward elimination may be easier to use.

Example: Backward Selection Using P-Values

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.9251	0.2102	9.16	<0.0001
income_ver: *source_only*	0.9750	0.0991	9.83	<0.0001
income_ver: *verified*	2.5374	0.1172	21.65	<0.0001
debt_to_income	0.0211	0.0029	7.18	<0.0001
credit_util	4.8959	0.1619	30.24	<0.0001
bankruptcy	0.3864	0.1324	2.92	0.0035
term	0.1537	0.0039	38.96	<0.0001
issued: *Jan2018*	0.0276	0.1081	0.26	0.7981
issued: *Mar2018*	-0.0397	0.1065	-0.37	0.7093
credit_checks	0.2282	0.0182	12.51	<0.0001

Example: Backward Selection Using P-Values

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.9213	0.1982	9.69	<0.0001
income_ver: *source_only*	0.9740	0.0991	9.83	<0.0001
income_ver: *verified*	2.5355	0.1172	21.64	<0.0001
debt_to_income	0.0211	0.0029	7.19	<0.0001
credit_util	4.8958	0.1619	30.25	<0.0001
bankruptcy	0.3869	0.1324	2.92	0.0035
term	0.1537	0.0039	38.97	<0.0001
credit_checks	0.2283	0.0182	12.51	<0.0001

Model Conditions

Multiple regression models

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \epsilon$$

depend on the following conditions:

1. Nearly normal residuals.
2. Constant variability of residuals.
3. Independence.
4. Each variable linearly related to the outcome.

Diagnostic Plots

We will consider our final model for the loan data:

$$\begin{aligned}\hat{rate} =& 1.921 + 0.974 \times \texttt{income\_ver}_{\texttt{source}} + 2.535 \times \texttt{income\_ver}_{\texttt{verified}} \\ &+ 0.021 \times \texttt{debt\_income} + 4.896 \times \texttt{credit\_util} + 0.387 \times \texttt{bankruptcy} \\ &+ 0.154 \times \texttt{term} + 0.228 \times \texttt{credit\_check}\end{aligned}$$

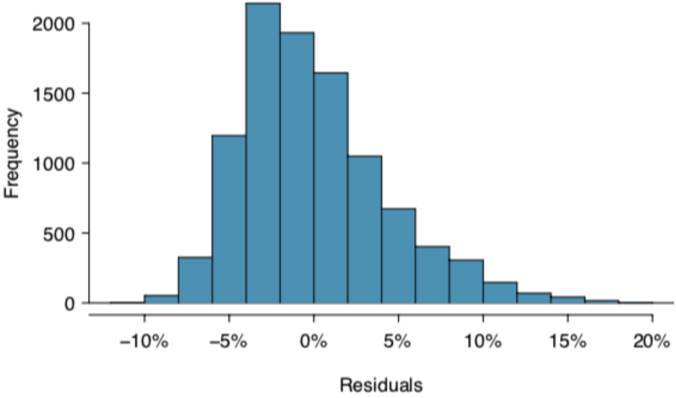
and will examine it for any issues with the model conditions.

Check for Normality

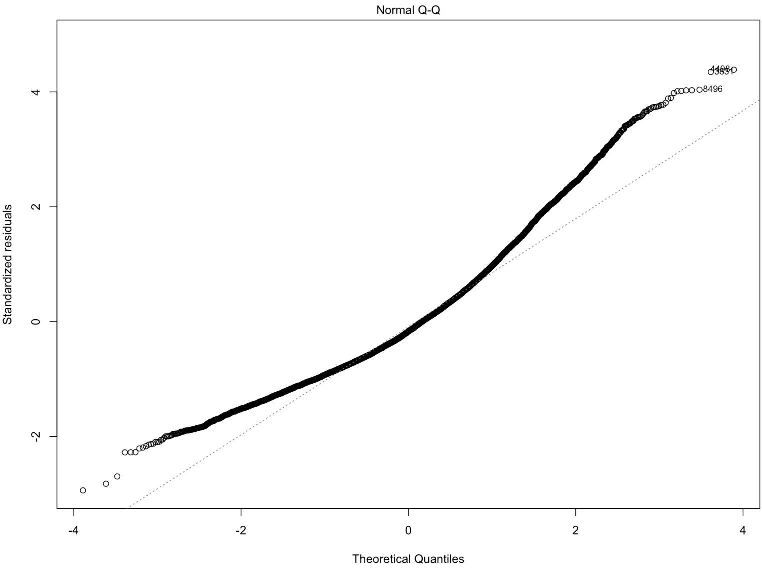
As with simple linear regression, there are two ways to check for normality:

1. Histograms
2. Q-Q Plots

Check for Normality: Histogram



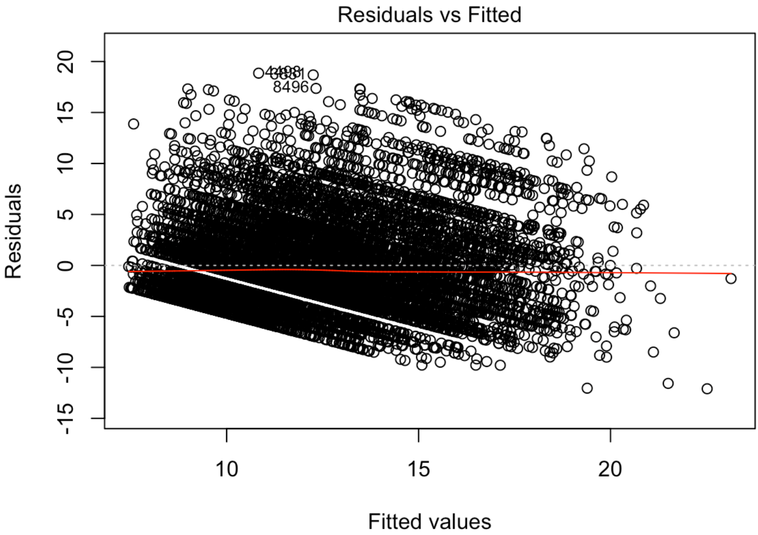
Check for Normality: Q-Q Plots



The Normality Assumption

- Since this is such a large dataset (10000 observations), we can relax this assumption some.
- *However*, our prediction intervals may not be valid if we do.

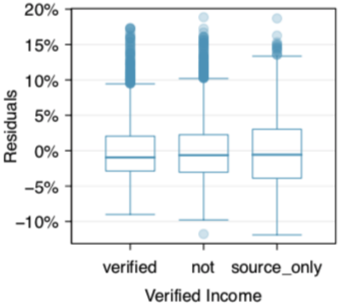
Constant Variance

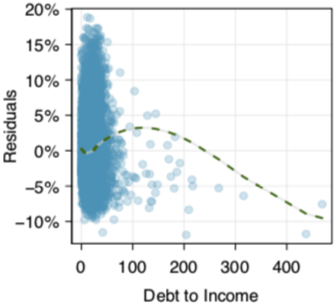


Other Useful Diagnostic Plots

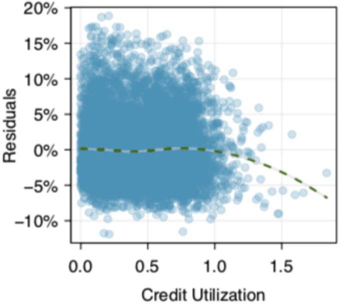
- For data taken in sequence, we might plot *residuals in order of data collection.*
  - This can help identify correlation between cases.
  - If we find connections, we may want to look into methods for **time series**.
- We may also want to look at the residuals plotted against each predictor variable.
  - Look for change in variability and patterns in the data.

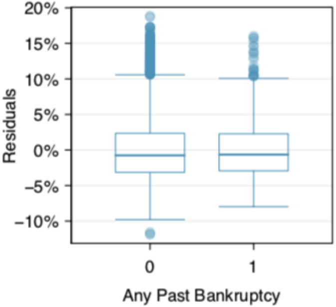
Residuals Versus Specific Predictor Variables



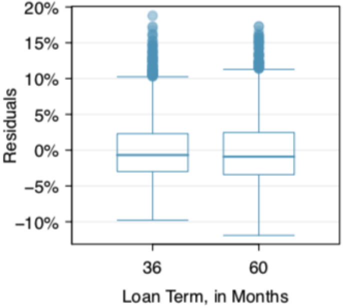


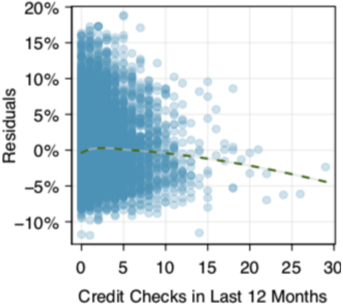
Residuals Versus Specific Predictor Variables





Residuals Versus Specific Predictor Variables





Now What?

- If we choose this as our final model, *we must report the observed abnormalities*!
- The second option is to look for ways to continue to improve the model.

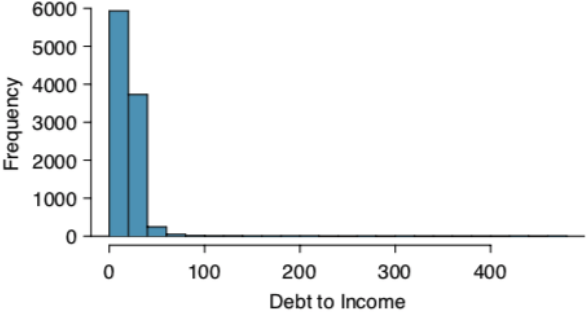
Transformations

One way to improve model fit is to *transform* one or more predictor variables.

- If a variable has a lot of skew and large values have a lot of leverage, we might try
  - Log transformation ($\log x$)
  - Square root transformation ($\sqrt{x}$)
  - Inverse transformation ($1/x$)

There are many valid transformations!

Example: Debt to Income



- We want to deal with this extreme skew.
- There are some cases where `debt_to_income` $= 0$.
- This will make log and inverse transformations infeasible.

Example: Debt to Income

First we will try a square root transformation

- We create a new variable, sqrt_debt_to_income

$$\texttt{sqrt\_debt\_to\_income} = \sqrt{\texttt{debt\_to\_income}}$$

We then refit the model with sqrt_debt_to_income.

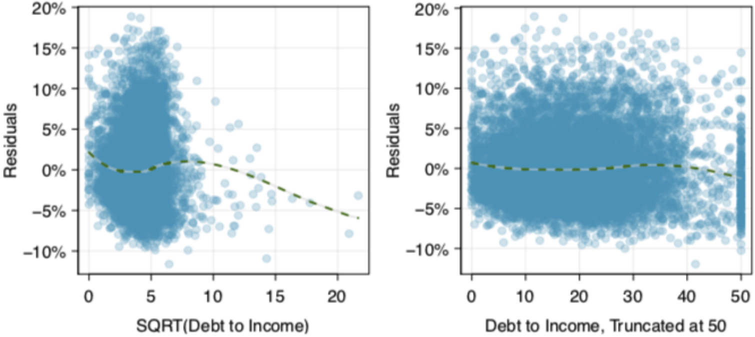
Example: Debt to Income

We will also try a truncation at 50.

- We create a new variable, `debt_to_income_50`.
    - Any values > 50 are shrunk to 50.

We then refit the model with `debt_to_income_50`.

Example: Debt to Income



The truncation does a good job fixing the constant variance assumption for this variable.

Example: Debt to Income

- With the debt to income issue fixed, we should recheck our model assumptions.
- We will find the same issues with the other variables.
- If we decide that this is our final model, we would need to acknowledge these issues.

Example: Debt to Income

The new model is

$$\begin{aligned}\hat{rate} =& 1.562 + 1.002 \times \texttt{income\_ver}_{\texttt{source}} + 2.436 \times \texttt{income\_ver}_{\texttt{verified}} \\ & + 0.048 \times \texttt{debt\_income} + 4.698 \times \texttt{credit\_util} + 0.394 \times \texttt{bankruptcy} \\ & + 0.153 \times \texttt{term} + 0.223 \times \texttt{credit\_check}\end{aligned}$$

Notice that the coefficient for `debt_income` doubled when we dealt with those high leverage outliers.

Reporting Results

- While we may report models that with conditions that are slightly violated,
  - ...as long as we acknowledge the violations in our reporting.
- we shouldn't report results when conditions are grossly violated.
- If familiar methods won't cut it, reach out to an expert.