Multiple Linear Regression

Special Predictors & Assumptions

Prof. Maria Tackett

02.12.20

STA 210

[Click for PDF of slides]

Announcements

- HW 02 **due TODAY at 11:59p**
- HW 03 will be assigned Monday and due **Feb 24**
- Analysis of variance questions

Today's agenda

- Special predictors
- Checking assumptions

Peer-to-peer lender

Today's data is a sample of about 9900 applications to a peer-to-peer lending club. The full data is in the `loans_full_schema` dataframe in the `openintro` package.

```
# loan50 dataset from the openintro package
loans <- read_csv("data/loans.csv") %>%
  mutate(bankruptcy = as.factor(bankruptcy))
glimpse(loans)
```

```
## Observations: 9,974
## Variables: 9
## $ verified_income  <chr> "Verified", "Not Verified", "Source Verified", .
## $ debt_to_income   <dbl> 18.01, 5.04, 21.15, 10.16, 57.96, 6.46, 23.66, .
## $ bankruptcy       <fct> 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 1, 0,.
## $ term             <dbl> 60, 36, 36, 36, 36, 36, 60, 60, 36, 36, 60, 60,.
## $ credit_util      <dbl> 0.54759517, 0.15003472, 0.66134832, 0.19673228,.
## $ interest_rate    <dbl> 14.07, 12.61, 17.09, 6.72, 14.07, 6.72, 13.59, .
## $ debt_inc_cent    <dbl> -1.3019882, -14.2719882, 1.8380118, -9.1519882,.
## $ term_cent        <dbl> 16.725887, -7.274113, -7.274113, -7.274113, -7..
## $ credit_util_cent <dbl> 0.14448914, -0.25307131, 0.25824229, -0.2063737.
```

Variables

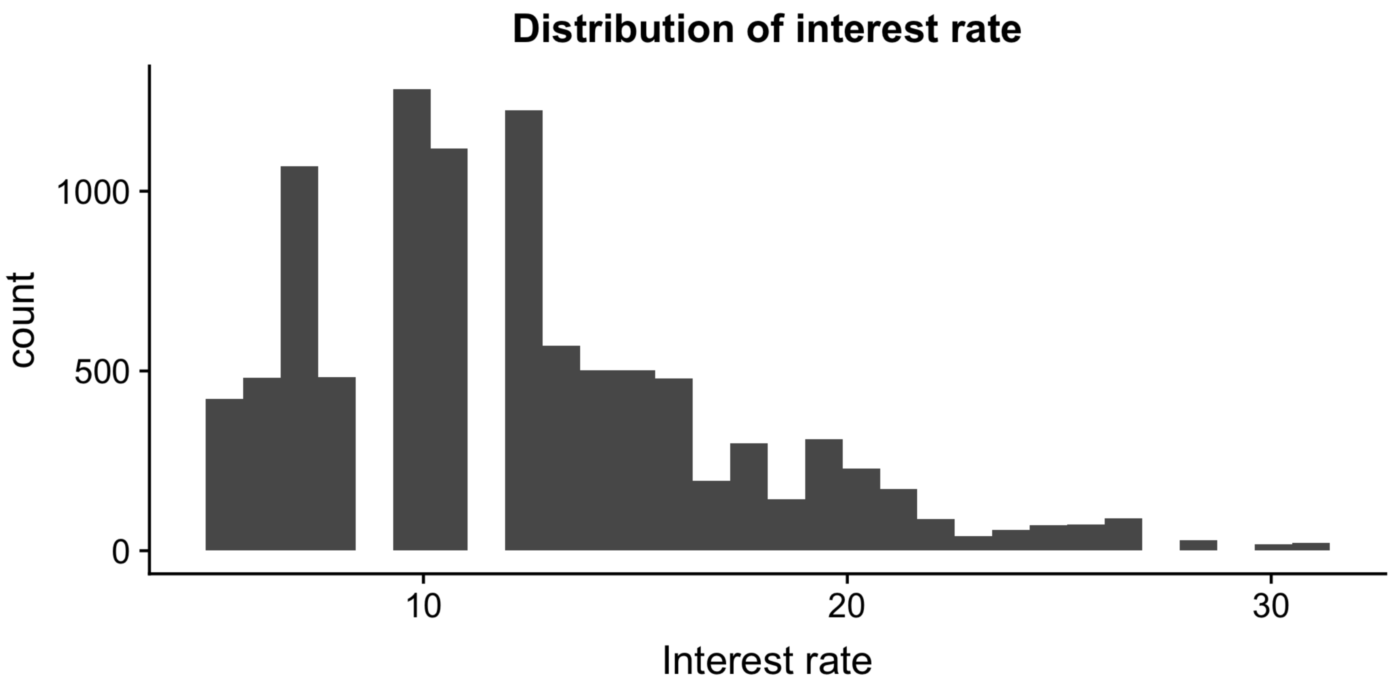
Predictors

- `verified_income`: Whether borrower's income source and amount have been verified (`Not Verified`, `Source Verified`, `Verified`)
- `debt_to_income`: Debt-to-income ratio, i.e. the percentage of a borrower's total debt divided by their total income
- `bankruptcy`: Indicator of whether borrower has had a bankruptcy in the past (`0`: No, `1`: Yes)
- `term`: Length of the loan in months
- `credit_util`: What fraction of total credit a borrower is utilizing, i.e. total credit utilizied divided by total credit limit

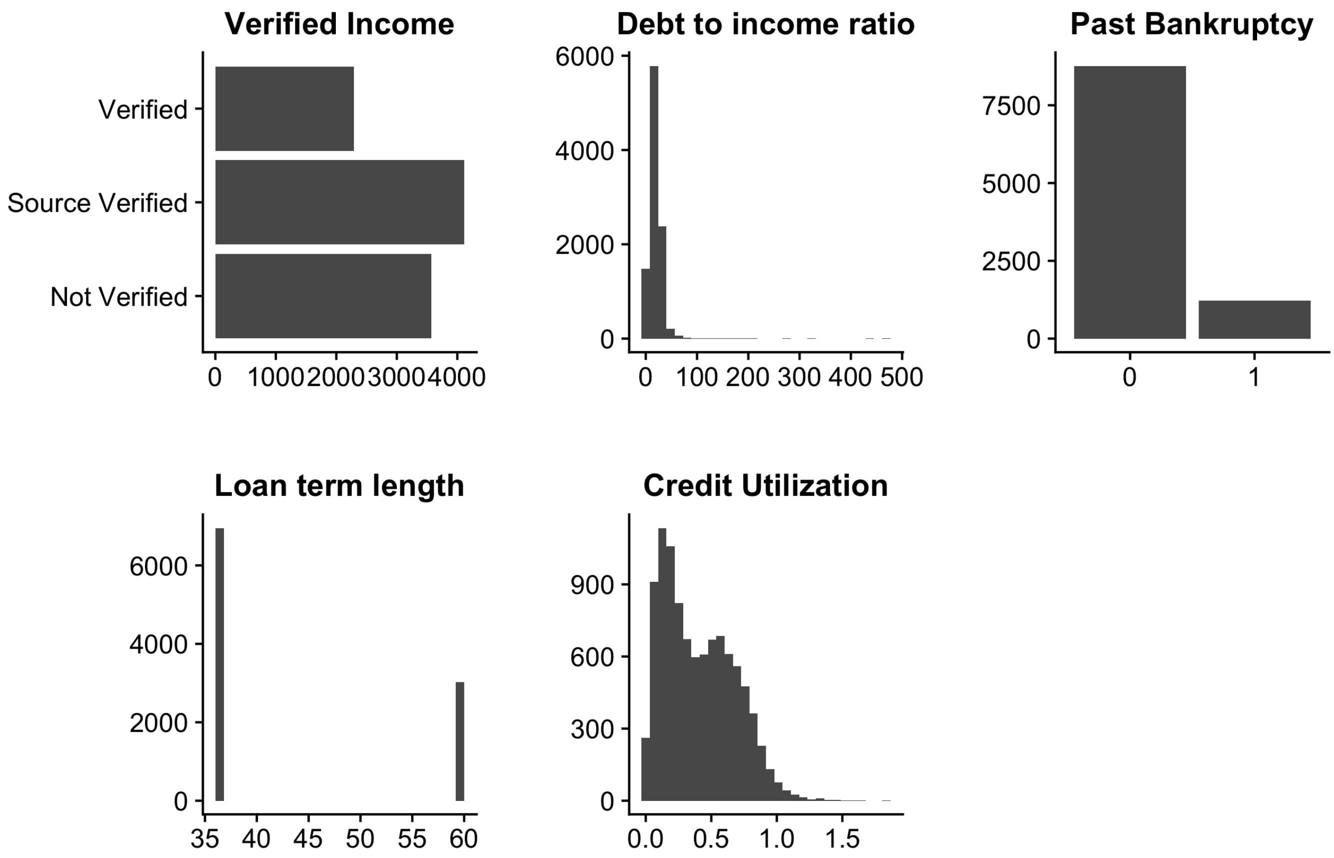
Response

- `interest_rate`: Interest rate for the loan

Response variable, `interest_rate`

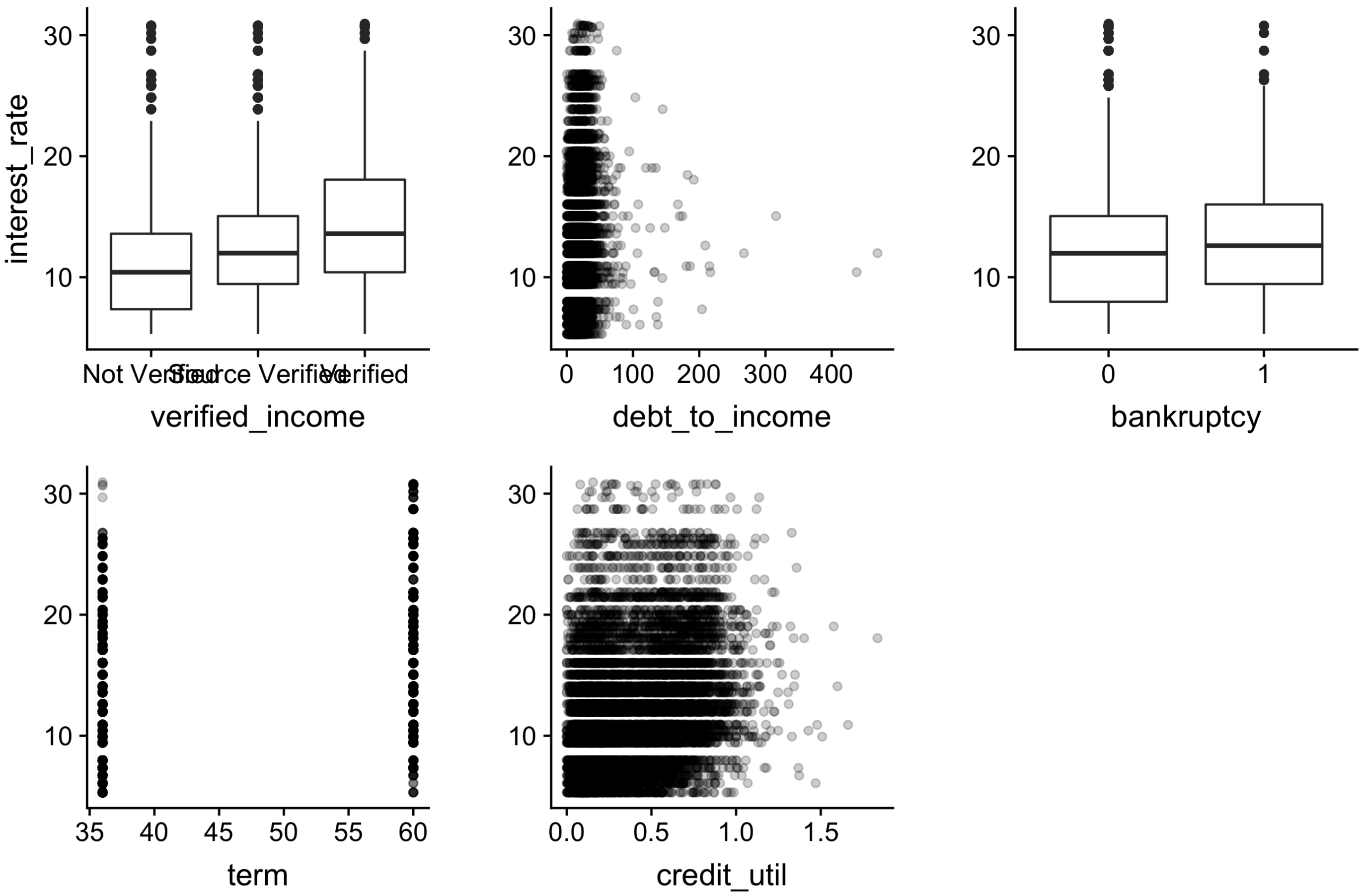


Predictor variables



STA 210

Response vs. Predictors



Regression Model

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	2.233	0.198	11.276	0	1.845	2.621
verified_incomeSource Verified	1.098	0.100	11.028	0	0.903	1.293
verified_incomeVerified	2.665	0.118	22.635	0	2.434	2.896
debt_to_income	0.023	0.003	7.689	0	0.017	0.029
bankruptcy1	0.525	0.133	3.951	0	0.265	0.785
term	0.154	0.004	38.800	0	0.146	0.162
credit_util	4.838	0.163	29.676	0	4.519	5.158

Special Predictors

Interpreting the Intercept

term	estimate	std.error	statistic	p.value
(Intercept)	2.233	0.198	11.276	0
verified_incomeSource Verified	1.098	0.100	11.028	0
verified_incomeVerified	2.665	0.118	22.635	0
debt_to_income	0.023	0.003	7.689	0
bankruptcy1	0.525	0.133	3.951	0
term	0.154	0.004	38.800	0
credit_util	4.838	0.163	29.676	0

- Based on our model, what subset of borrowers do we expect to have an interest rate of 2.233%? In other words, what subset of borrowers are included in the intercept?
- Is this interpretation meaningful? Why or why not?

Mean-Centered Variables

- To have a meaningful interpretation of the intercept, use **mean-centered** predictor variables in the model (quantitative predictors only)
- A mean-centered variable is calculated by subtracting the mean from each value of the variable, i.e.

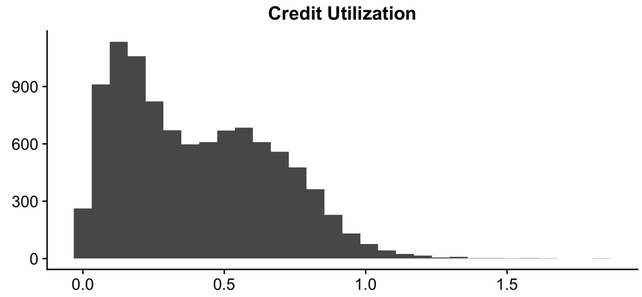
$$x_{ip} - \bar{x}_{.p}$$

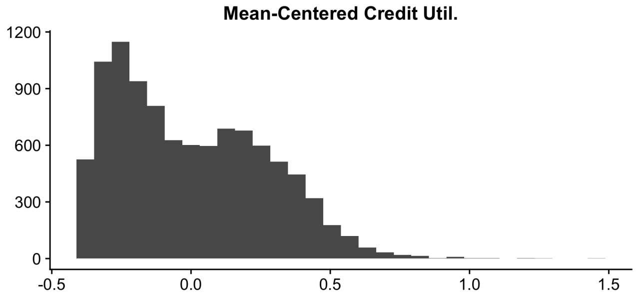
- Now the intercept is interpreted as the expected value of the response at the mean value of all quantitative predictors

STA 210

Loans: mean-centered variables

```
loans <- loans %>%
  mutate(debt_inc_cent = debt_to_income - mean(debt_to_income),
         term_cent = term - mean(term),
         credit_util_cent = credit_util - mean(credit_util))
```





In-class exercise

term	estimate	std.error	statistic	p.value
(Intercept)	2.233	0.198	11.276	0
verified_incomeSource Verified	1.098	0.100	11.028	0
verified_incomeVerified	2.665	0.118	22.635	0
debt_to_income	0.023	0.003	7.689	0
bankruptcy1	0.525	0.133	3.951	0
term	0.154	0.004	38.800	0
credit_util	4.838	0.163	29.676	0

- Go to http://bit.ly/sta210-sp20-mean-center and describe how the model would change if `debt_inc_cent`, `term_cent`, and `credit_util_cent` were used in the model in[illegible] original versions of these variables.

03:00

STA 210

How model changes with mean-centered variables

Indicator (dummy) variables

- Suppose there is a categorical variable with k levels (categories)
- Make k indicator variables (also known as dummy variables)
- Use $k-1$ of the indicator variables in the model
  - Can't uniquely estimate all k variables at once if the intercept is in the model
- Level that doesn't have a variable in the model is called the **baseline**
- Coefficients interpreted as the change in the mean of the response over the baseline

STA 210

Indicator variables: $k = 2$

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	11.293	0.074	151.718	0	11.148	11.439
verified_incomeSource Verified	1.098	0.100	11.028	0	0.903	1.293
verified_incomeVerified	2.665	0.118	22.635	0	2.434	2.896
debt_inc_cent	0.023	0.003	7.689	0	0.017	0.029
bankruptcy1	0.525	0.133	3.951	0	0.265	0.785
term_cent	0.154	0.004	38.800	0	0.146	0.162
credit_util_cent	4.838	0.163	29.676	0	4.519	5.158

Interpreting `bankruptcy` in the model:

Indicator variables: $k > 2$

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	11.293	0.074	151.718	0	11.148	11.439
verified_incomeSource Verified	1.098	0.100	11.028	0	0.903	1.293
verified_incomeVerified	2.665	0.118	22.635	0	2.434	2.896
debt_inc_cent	0.023	0.003	7.689	0	0.017	0.029
bankruptcy1	0.525	0.133	3.951	0	0.265	0.785
term_cent	0.154	0.004	38.800	0	0.146	0.162
credit_util_cent	4.838	0.163	29.676	0	4.519	5.158

Interpreting `verified_income` in the model:

STA 210

Interaction Terms

- **Case**: Relationship of the predictor variable with the response depends on the value of another predictor variable
  - This is an **interaction effect**
- Create a new interaction variable that is one predictor variable times the other in the interaction
- **Good Practice**: When including an interaction term, also *include the associated main effects* (each predictor variable on its own) even if their coefficients are not statistically significant

STA 210

Add interaction term

```
model_w_int <- lm(interest_rate ~ verified_income + debt_inc_cent
                   bankruptcy + term_cent + credit_util_cent +
                   debt_inc_cent * verified_income,
                 data = loans)
```

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	11.298	0.074	151.764	0.000	11.152	11.444
verified_incomeSource Verified	1.094	0.100	10.940	0.000	0.898	1.290
verified_incomeVerified	2.704	0.119	22.730	0.000	2.471	2.937
debt_inc_cent	0.032	0.005	6.527	0.000	0.022	0.041
bankruptcy1	0.525	0.133	3.954	0.000	0.265	0.786
term_cent	0.154	0.004	38.764	0.000	0.146	0.162
credit_util_cent	4.841	0.163	29.689	0.000	4.521	5.160
verified_incomeSource Verified:debt_inc_cent	-0.009	0.007	-1.243	0.214	-0.023	0.005
verified_incomeVerified:debt_inc_cent	-0.019	0.007	-2.699	0.007	-0.033	-0.005

STA 210

Checking model assumptions

Assumptions

Inference on the regression coefficients and predictions are reliable only when the regression assumptions are reasonably satisfied:

1. **Linearity:** Response variable has a linear relationship with the predictor variables in the model
2. **Constant Variance:** The regression variance is the same for all set of predictor variables $(x_1, \ldots, x_p)$
3. **Normality:** For a given set of predictors $(x_1, \ldots, x_p)$, the response, y, follows a Normal distribution around its mean
4. **Independence:** All observations are independent

> We will use plots of the residuals to check these assumptions

STA 210

Checking linearity assumption

- Make the following plots:
  - Plot the residuals vs. the predicted (fitted) values
  - Plot the residuals vs. each predictor variable
- These plots should have no systematic / obvious pattern, i.e there should be no apparent structure
- A systematic pattern may suggestion that interactions or higher-order terms (like quadratic terms) are required.

STA 210

Checking constant variance assumption

- Make a plot of the residuals vs. the predicted (fitted) values
- The height of the cloud of points should be constant as you go from left to right on the plot

Checking normality assumption

- Make the following plots:
  - Histogram of the residuals
  - Normal QQ-Plot of the residuals
- The histogram should be approximately unimodal and symmetric.
- The points on the Normal QQ-Plot should generally follow a straight diagonal line

Checking independence assumption

- In the indepednece assumption, we assume the residuals are not correlated
- If your data were collected over time, plot the residuals in time order
- There should be no pattern in the plot.
  - A cyclical pattern indicates the residuals are correlated, a violation of the assumption.
- Can generally conclude this assumption is resonably met unless there are clear violations

augment

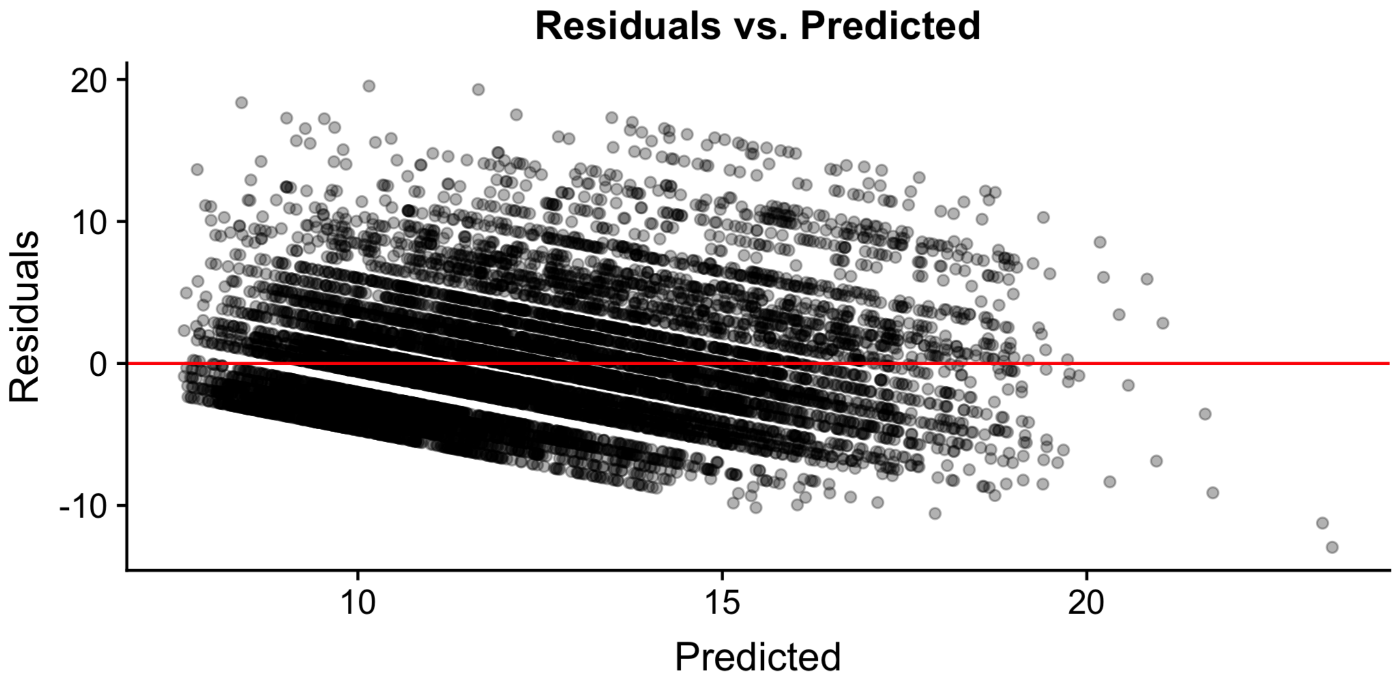
- Use the **augment** function in the **broom** package to calculate residuals, predicted values, and other model diagnostics

```
loans_aug <- augment(model_w_int)
glimpse(loans_aug)
```

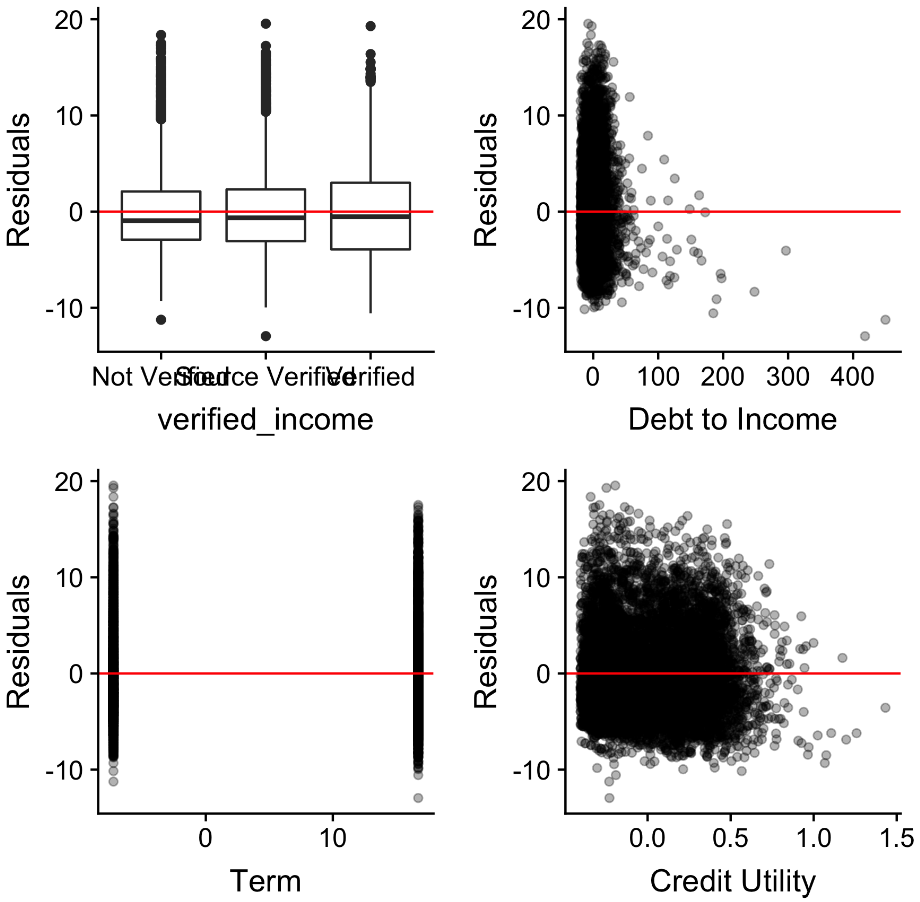
```
## Observations: 9,974
## Variables: 13
## $ interest_rate    <dbl> 14.07, 12.61, 17.09, 6.72, 14.07, 6.72, 13.59, .
## $ verified_income  <chr> "Verified", "Not Verified", "Source Verified", .
## $ debt_inc_cent    <dbl> -1.3019882, -14.2719882, 1.8380118, -9.1519882,.
## $ bankruptcy       <fct> 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 1, 0,.
## $ term_cent        <dbl> 16.725887, -7.274113, -7.274113, -7.274113, -7..
## $ credit_util_cent <dbl> 0.14448914, -0.25307131, 0.25824229, -0.2063737.
## $ .fitted          <dbl> 17.261500, 9.028191, 12.563387, 8.890419, 15.05.
## $ .se.fit          <dbl> 0.11707102, 0.16010940, 0.08736166, 0.09266451,.
## $ .resid           <dbl> -3.1914996, 3.5818088, 4.5266127, -2.1704190, -.
## $ .hat             <dbl> 0.0007303468, 0.0013660418, 0.0004066981, 0.000.
## $ .sigma           <dbl> 4.332063, 4.332032, 4.331944, 4.332127, 4.33217.
## $ .cooksd          <dbl> 4.411042e-05, 1.040504e-04, 4.938101e-05, 1.277.
## $ .std.resid       <dbl> -0.73700191, 0.82739787, 1.04514571, -0.5011389.
```

Check linearity: Residuals vs. Predicted

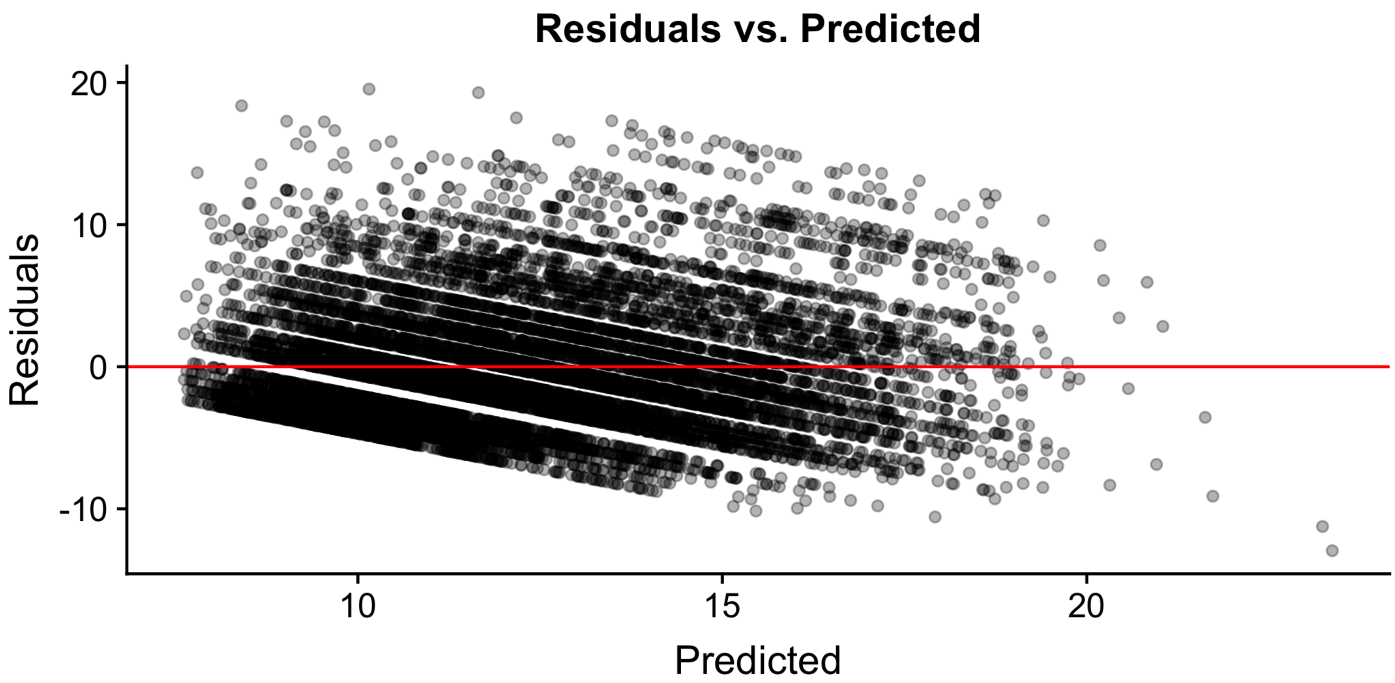
```
ggplot(data = loans_aug, aes(x = .fitted, y = .resid)) +
  geom_point(alpha = 0.3) +
  geom_hline(yintercept = 0, color = "red") +
  labs(x = "Predicted", y = "Residuals",
       title = "Residuals vs. Predicted")
```



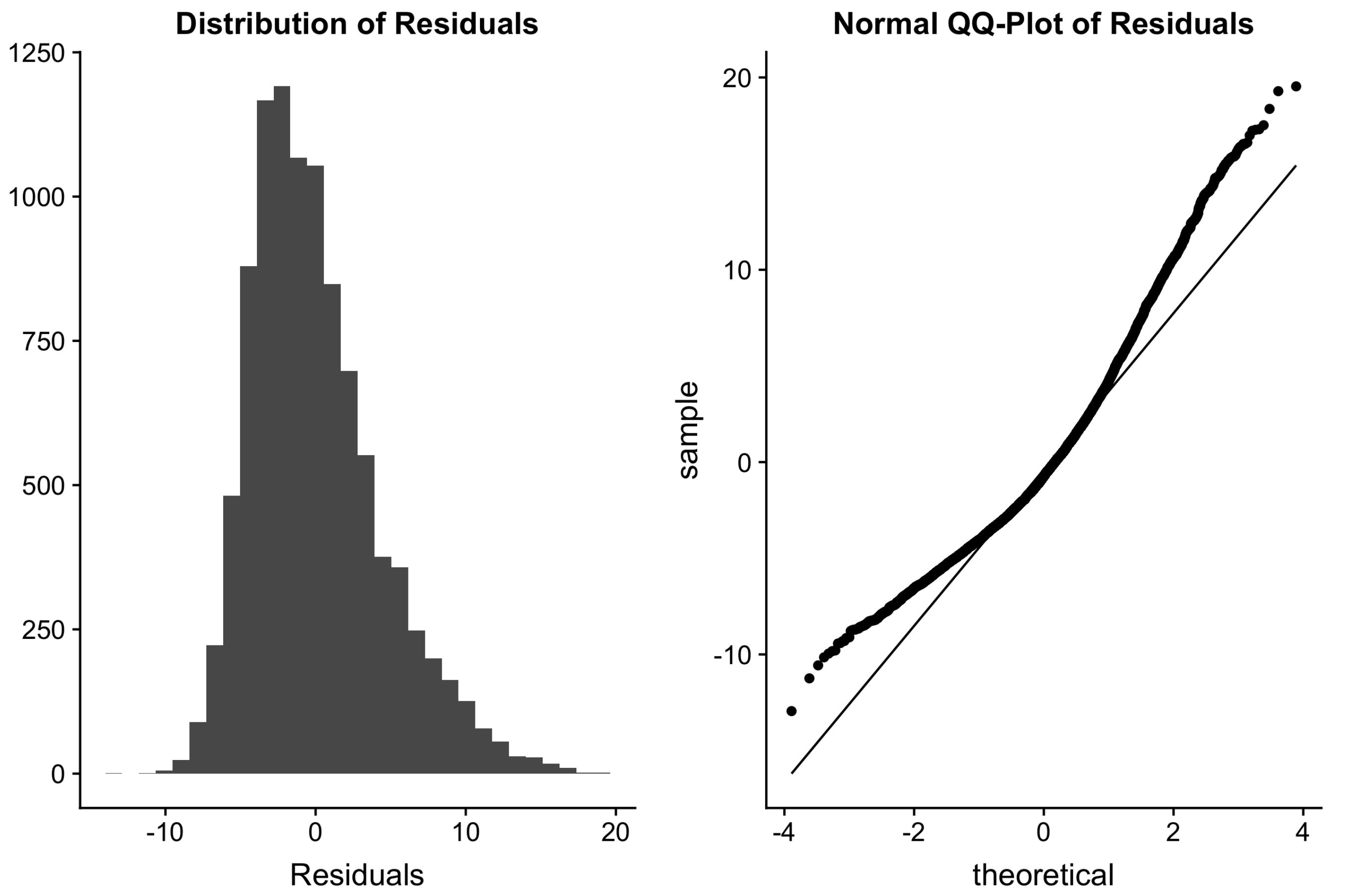
Check linearity: Residuals vs. Predictors



Check constant variance



Check Normality



STA 210

Checking independence

Can check residuals versus observation number if you think there is some structure / order to the dataset. Below is the code for this dataset:

```
loans_aug <- loans_aug %>%
  mutate(obs_num = 1:nrow(loans_aug))
```

```
ggplot(data = loans_aug, aes(x = obs_num, y = .resid)) +
  geom_point(alpha = 0.3) +
  labs(x = "Observation Number", y = "Residuals",
       title = "Residuals vs. Observation Number")
```

Use EDA but don't solely rely on it

- Look at a scatterplot of the response variable vs. each of the predictor variables in the exploratory data analysis before calculating the regression model
- This is a good way to check for obvious departures from linearity or constant variance
- This is **not** definitive, but it can give you an indication early on if you might need to use interactions, higher-order terms, or do a transformation (more on that next week)

STA 210