Multiple Linear Regression

Types of Predictors

Prof. Maria Tackett

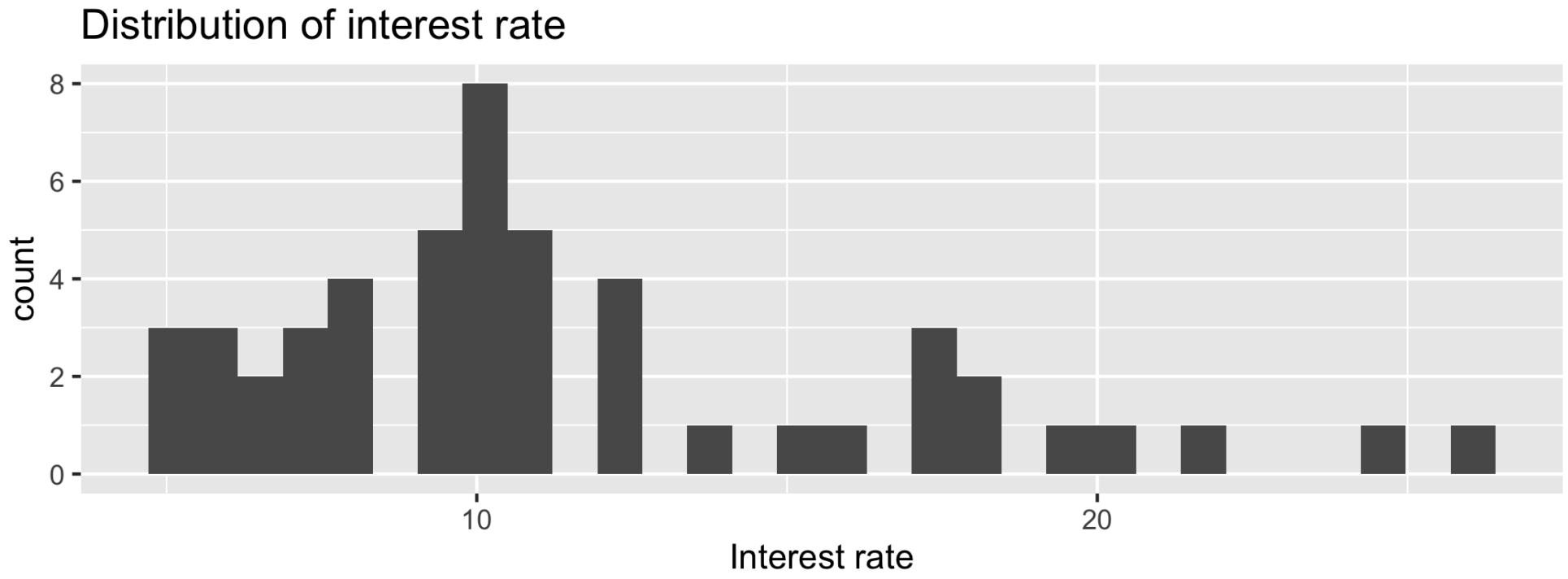
Click here for PDF of slides

Topics

- Mean-centering quantitative predictors
- Using indicator variables for categorical predictors
- Using interaction terms

Peer-to-peer lender

Today's data is a sample of 50 loans made through a peer-to-peer lending club. The data is in the **`loan50`** data frame in the openintro R package.



Variables

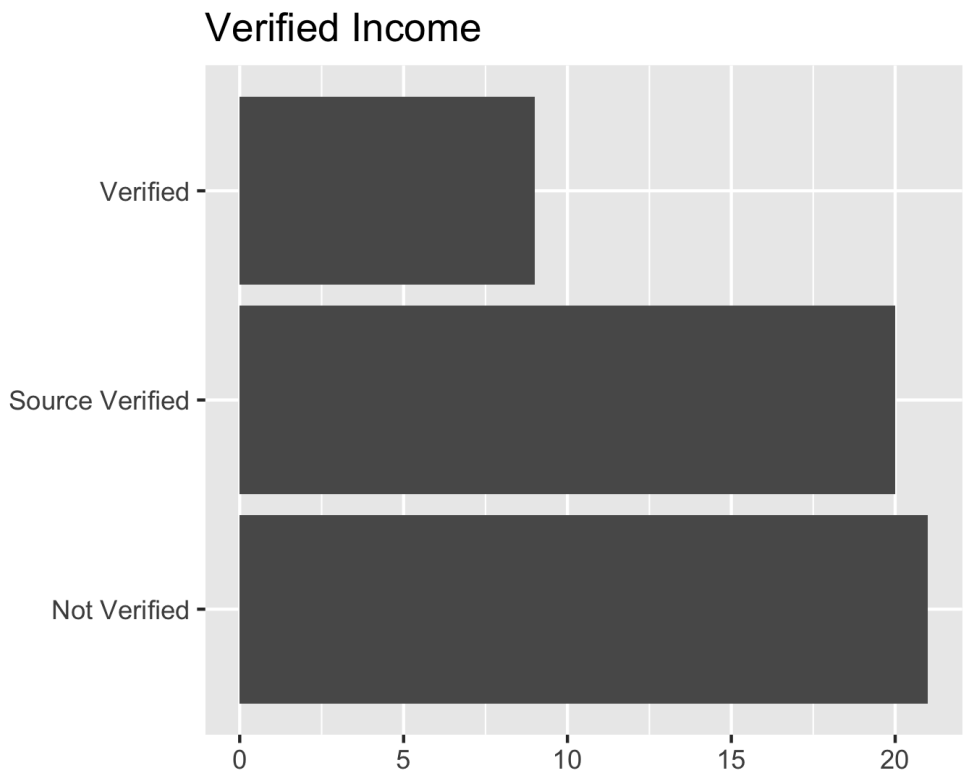
Predictors

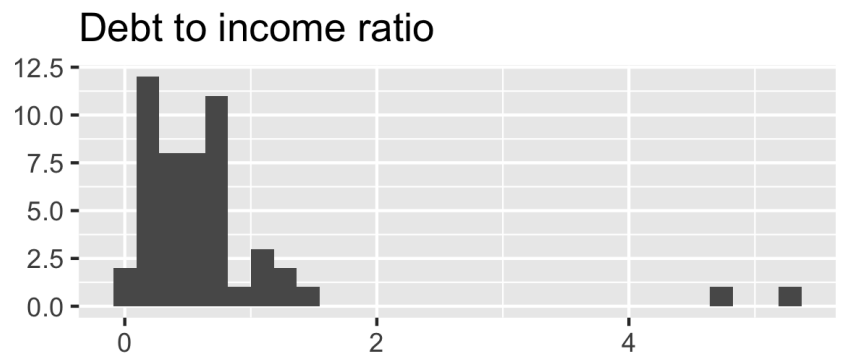
- **`verified_income`**: Whether borrower's income source and amount have been verified (**`Not Verified`**, **`Source Verified`**, **`Verified`**)
- **`debt_to_income`**: Debt-to-income ratio, i.e. the percentage of a borrower's total debt divided by their total income
- **`annual_income`**: Annual income (in $1000s)

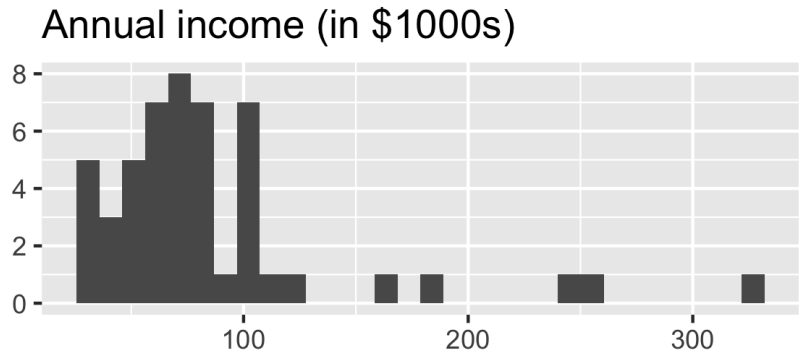
Response

- **`interest_rate`**: Interest rate for the loan

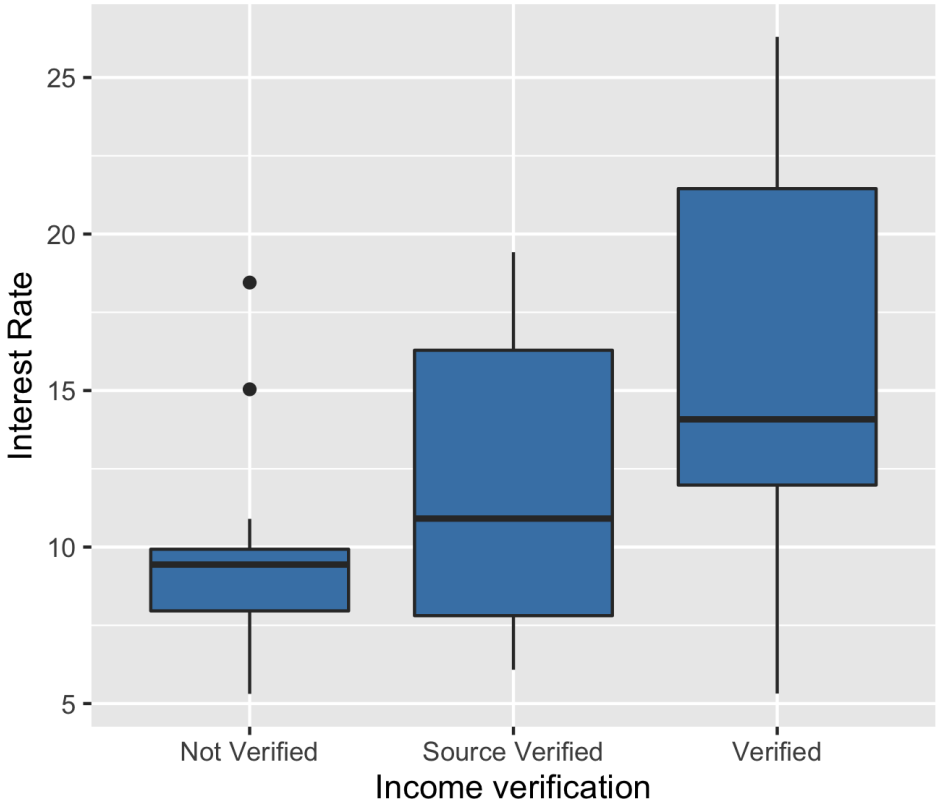
Predictor variables

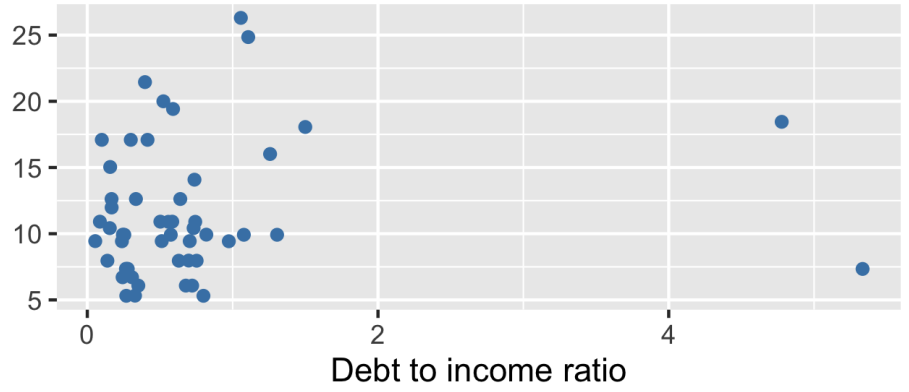


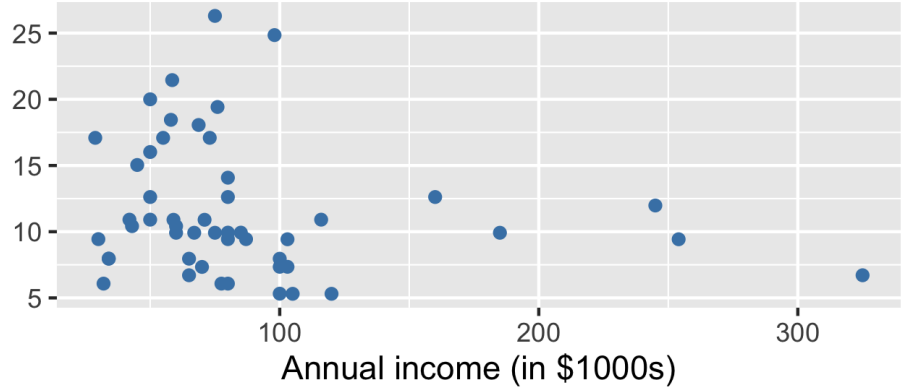




Response vs. Predictors







Regression Model

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	10.726	1.507	7.116	0.000	7.690	13.762
debt_to_income	0.671	0.676	0.993	0.326	-0.690	2.033
verified_incomeSource Verified	2.211	1.399	1.581	0.121	-0.606	5.028
verified_incomeVerified	6.880	1.801	3.820	0.000	3.253	10.508
annual_income	-0.021	0.011	-1.804	0.078	-0.043	0.002

- Describe the subset of borrowers who are expected to get an interest rate of 10.726% based on our model
- Is this interpretation meaningful? Why or why not?

Mean-centered variables

Mean-Centered Variables

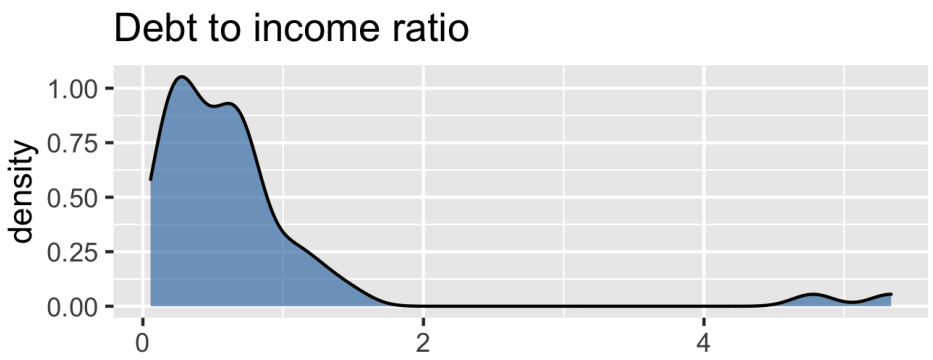
If we are interested in interpreting the intercept, we can **mean-center** the quantitative predictors in the model.

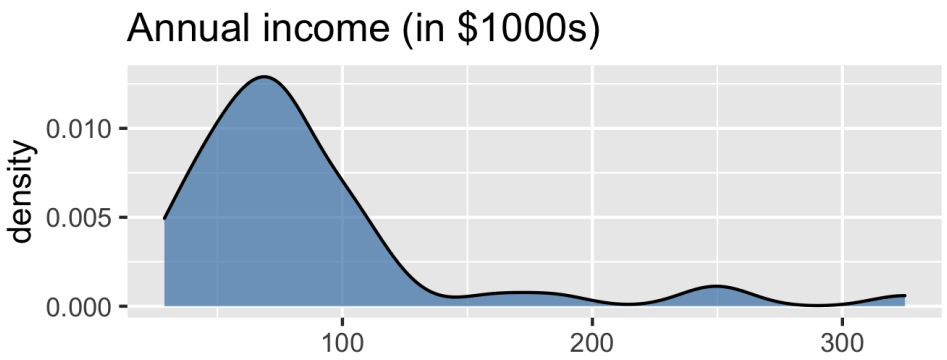
We can mean-center a quantitative predictor X_j using the following:

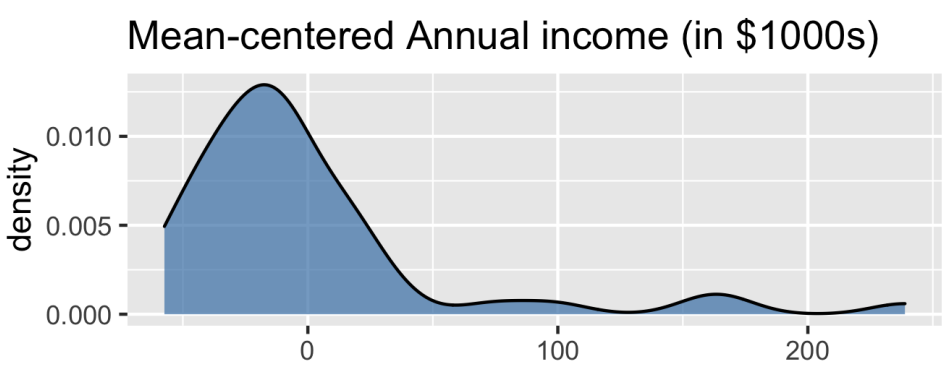
$$X_{j_{Cent}} = X_j - \bar{X}_j$$

If we mean-center all quantitative variables, then the intercept is interpreted as the expected value of the response variable when all quantitative variables are at their mean value.

Loans data: mean-center variables







Using mean-centere variables in the model

How do you expect the model to change if we use the **`debt_inc_cent`** and **`annual_income_cent`** in the model?

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	9.444	0.977	9.663	0.000	7.476	11.413
debt_inc_cent	0.671	0.676	0.993	0.326	-0.690	2.033
verified_incomeSource Verified	2.211	1.399	1.581	0.121	-0.606	5.028
verified_incomeVerified	6.880	1.801	3.820	0.000	3.253	10.508
annual_income_cent	-0.021	0.011	-1.804	0.078	-0.043	0.002

Original vs. mean-centered model

term	estimate
(Intercept)	10.726
debt_to_income	0.671
verified_incomeSource Verified	2.211
verified_incomeVerified	6.880
annual_income	-0.021

term	estimate
(Intercept)	9.444
debt_inc_cent	0.671
verified_incomeSource Verified	2.211
verified_incomeVerified	6.880
annual_income_cent	-0.021

Indicator variables

Indicator variables

- Suppose there is a categorical variable with K categories (levels)
- We can make K indicator variables - one indicator for each category
- An **indicator variable** takes values 1 or 0
  - 1 if the observation belongs to that category
  - 0 if the observation does not belong to that category

Indicator variable for `verified_income`

```
loan50 <- loan50 %>%
  mutate(not_verified =
           if_else(verified_income == "Not Verified", 1, 0),
         source_verified =
           if_else(verified_income == "Source Verified", 1, 0),
         verified =
           if_else(verified_income == "Verified", 1, 0)
  )
```

```
## # A tibble: 3 × 4
##   verified_income not_verified source_verified verified
##   <fct>                  <dbl>           <dbl>    <dbl>
## 1 Not Verified               1               0        0
## 2 Verified                   0               0        1
## 3 Source Verified            0               1        0
```

Indicators in the model

We will use $K - 1$ of the indicator variables in the model

The **baseline** is the category that doesn't have a term in the model.

The coefficients of the indicator variables in the model are interpreted as the expected change in the response compared to the baseline, holding all other variables constant.

Interpreting `verified_income`

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	9.444	0.977	9.663	0.000	7.476	11.413
debt_inc_cent	0.671	0.676	0.993	0.326	-0.690	2.033
verified_incomeSource Verified	2.211	1.399	1.581	0.121	-0.606	5.028
verified_incomeVerified	6.880	1.801	3.820	0.000	3.253	10.508
annual_income_cent	-0.021	0.011	-1.804	0.078	-0.043	0.002

The baseline category is "Not verified".

Interpreting `verified_income`

A person with source verified income is expected to take a loan with an interest rate that is 2.211% higher than the rate on loans to those whose income is not verified, holding all else constant.

A person with verified income is expected to take a loan with an interest rate that is 6.880% higher than the rate on loans to those whose income is not verified, holding all else constant.

Interaction terms

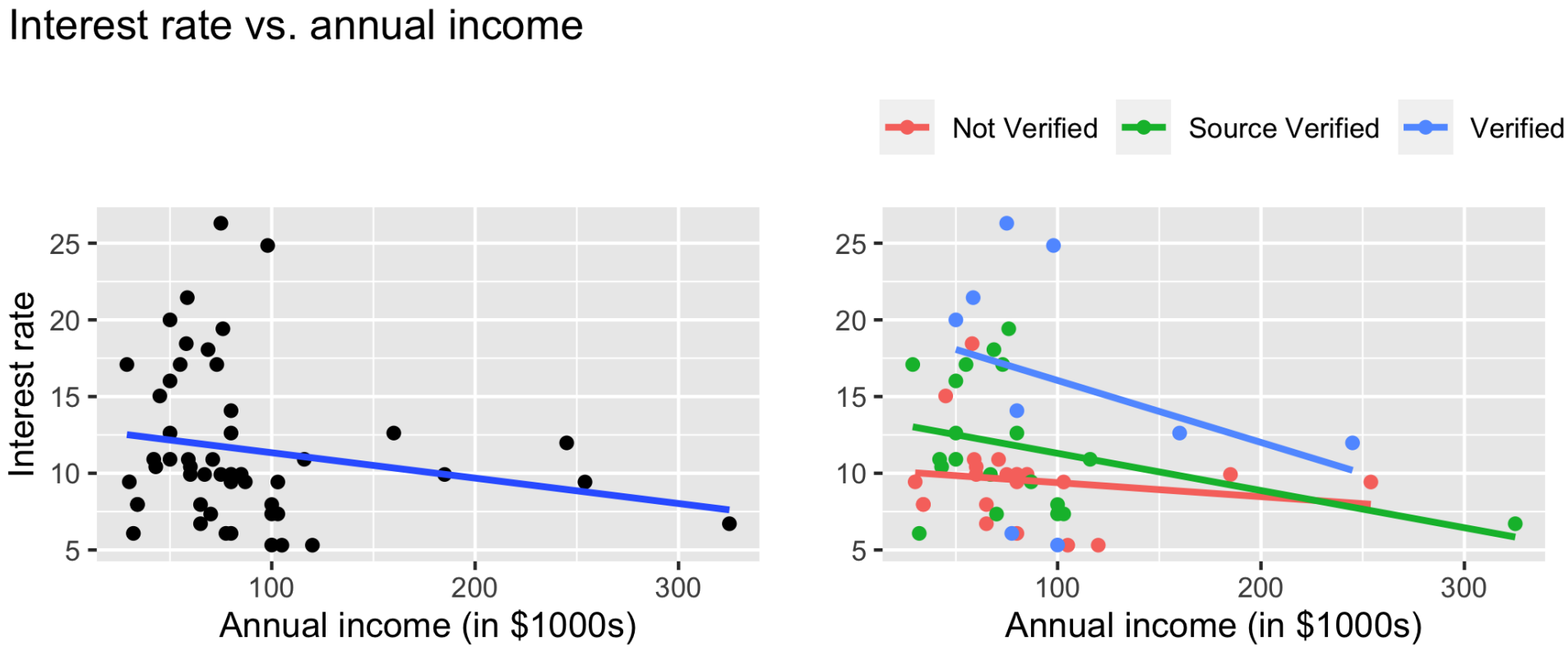
Interaction Terms

Sometimes the relationship between a predictor variable and the response depends on the value of another predictor variable

This is an **interaction effect**

To account for this, we can include **interaction terms** in the model.

Interest rate vs. annual income



The lines are **not parallel** indicating there is an **interaction effect**. The slope of annual income differs based on the income verification.

Interaction term in model

term	estimate	std.error	statistic	p.value
(Intercept)	9.484	0.989	9.586	0.000
debt_inc_cent	0.691	0.685	1.009	0.319
annual_income_cent	-0.007	0.020	-0.341	0.735
verified_incomeSource Verified	2.157	1.418	1.522	0.135
verified_incomeVerified	7.181	1.870	3.840	0.000
annual_income_cent:verified_incomeSource Verified	-0.016	0.026	-0.643	0.523
annual_income_cent:verified_incomeVerified	-0.032	0.033	-0.979	0.333

Interpreting interaction terms

What the interaction means:

The effect of annual income on the interest rate differs by -0.016 when the income is source verified compared to when it is not verified, holding all else constant.

Interpreting `annual_income` for source verified:

If the income is source verified, we expect the interest rate to decrease by 0.023% (-0.007 + -0.016) for each additional thousand dollars in annual income, holding all else constant.

Recap

- Mean-centering quantitative predictors
- Using indicator variables for categorical predictors
- Using interaction terms