Lec05: Multiple Regression

Stat021
Prof Amanda Luby
Reference: IMS Ch8.1-8.3, Ch25.1-25.2

Multiple regression extends single predictor regression to the case that still has one response but many predictors (which we'll call $x_1, x_2, x_3, \ldots$) that we think are simultaneously connected to the response. By considering how different predictor variables interact, we can uncover complicated relationships between the predictor variable and the response variable.

The plan for today is to cover the basics of multiple regression and extend key parts of what we learned in simple linear regression. Our focus today is *not* on the R code but on building up our intuition and seeing some of the foundations of multiple regression. On Thursday, we'll walk through how to do these things in R.

Agenda:

- Multiple Regression Equation
- Sum of Squares
- R^2 and Adjusted R^2
- ANOVA for regression
- T-tests for regression
- CI's and PI's

1. Data

We will consider data about loans from the peer-to-peer lender, Lending Club. The loan data includes terms of the loan as well as information about the borrower. The outcome variable we would like to better understand is the interest rate assigned to the loan. For instance, all other characteristics held constant, does it matter how many credit cards somebody has? How much does their income matter? Multiple regression will help us answer these and other questions.

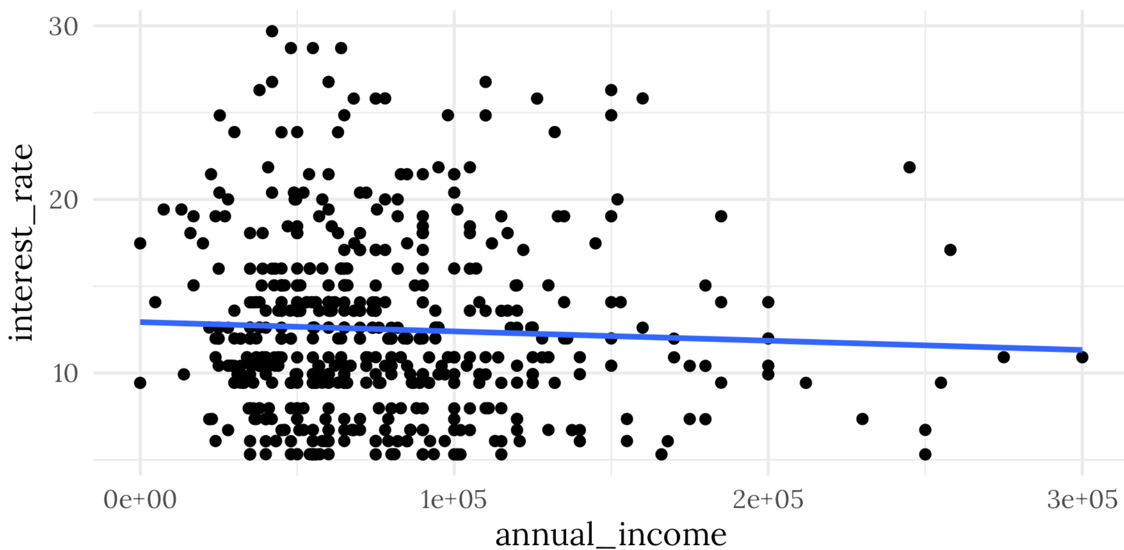
The full dataset includes results from 10,000 loans, and we'll be looking at a random sample of 500 loans and subset of the available variables:

```
## # A tibble: 6 x 7
##   state homeownership annual_income debt_to_income num_total_cc_accounts
##   <fct> <fct>                 <dbl>          <dbl>                 <int>
```

```
## 1 ME    MORTGAGE              60000         27.2                        12
## 2 NY    MORTGAGE              25000         19.8                        23
## 3 CA    RENT                 140000         10.4                        14
## 4 CA    RENT                 160000          9.26                        6
## 5 CT    RENT                  82000          6.2                        15
## 6 NJ    RENT                  98000          8.94                        6
## # ... with 2 more variables: interest_rate <dbl>, loan_amount <int>
```

2. Simple Linear Regression

There are a total of 55 variables in this dataset, so we are only scratching the surface today. We'll return to this dataset throughout the next few weeks! To start with, let's look at the relationship between `interest_rate` and `annual_income`:

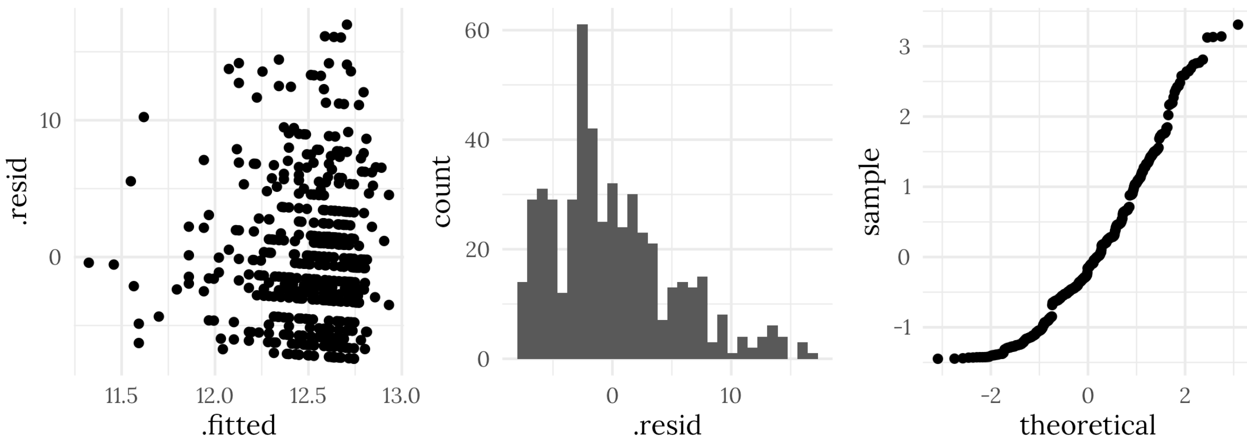



```
income_lm = lm(interest_rate ~ annual_income, data = loans)
tidy(income_lm)
```

```
## # A tibble: 2 x 5
##   term             estimate  std.error statistic   p.value
##   <chr>               <dbl>      <dbl>     <dbl>     <dbl>
## 1 (Intercept)   12.9        0.466          27.8  3.84e-103
## 2 annual_income -0.00000536 0.00000522     -1.03 3.05e-  1
```

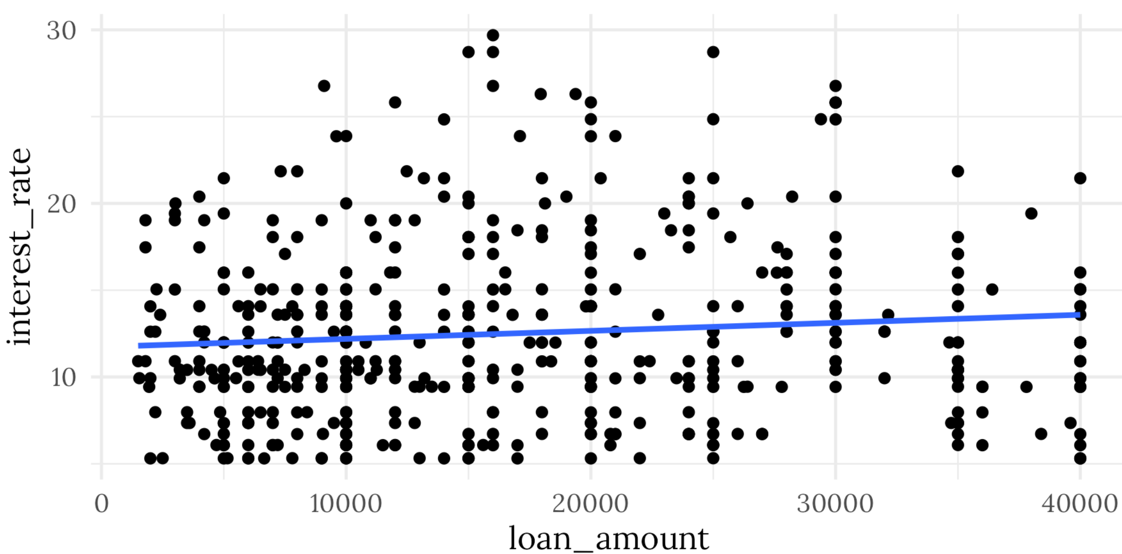
Model:

Assumption Checking:



Hypothesis test for β_1:

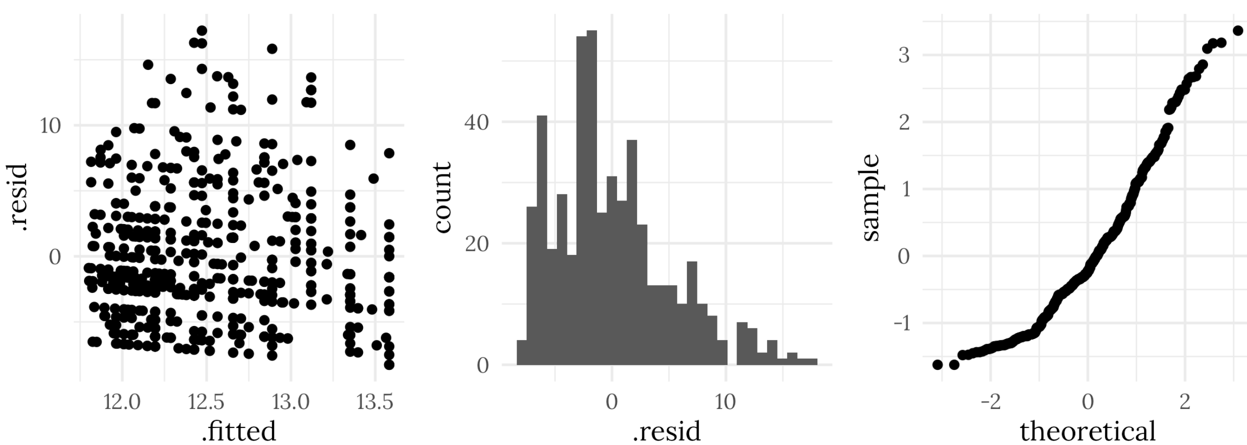
Even if `annual_income` is related to `interest_rate`, there are likely other variables at play. Now let's look at the relationship between `interest_rate` and `loan_amount`:



```r
amount_lm = lm(interest_rate ~ loan_amount, data = loans)
tidy(amount_lm)
```

```
## # A tibble: 2 x 5
##   term          estimate std.error statistic  p.value
##   <chr>            <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept) 11.7       0.438         26.8  1.36e-98
## 2 loan_amount  0.0000463 0.0000220      2.10 3.63e- 2
```

Model:



Hypothesis test for β_1:

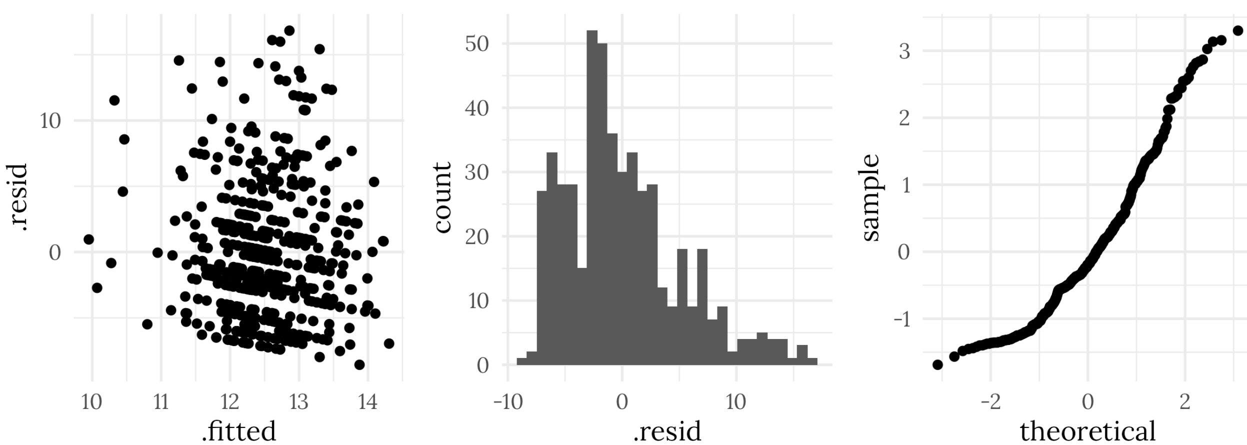
Summary

3. Multiple Regression Model

```
amount_income_lm = lm(interest_rate ~ annual_income + loan_amount, data = loans)
tidy(amount_income_lm)
```

```
## # A tibble: 3 x 5
##   term            estimate  std.error statistic  p.value
##   <chr>              <dbl>      <dbl>     <dbl>    <dbl>
## 1 (Intercept)   12.3       0.518           23.7  9.06e-84
## 2 annual_income -0.0000114 0.00000564      -2.02 4.41e- 2
## 3 loan_amount    0.0000651 0.0000239        2.73 6.60e- 3
```

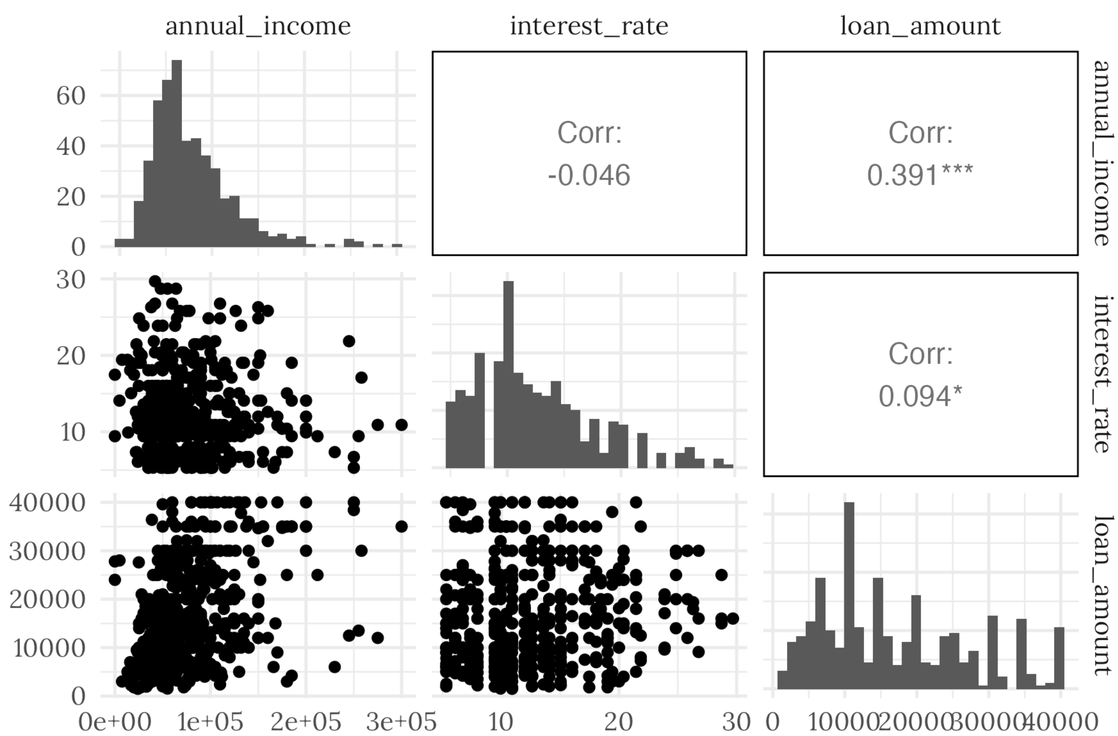
Model:



4. Interpretation of β Coefficients

5. Multicollinearity

Sometimes a set of predictor variables can impact the model in unusual ways, often due to the predictor variables themselves being correlated.

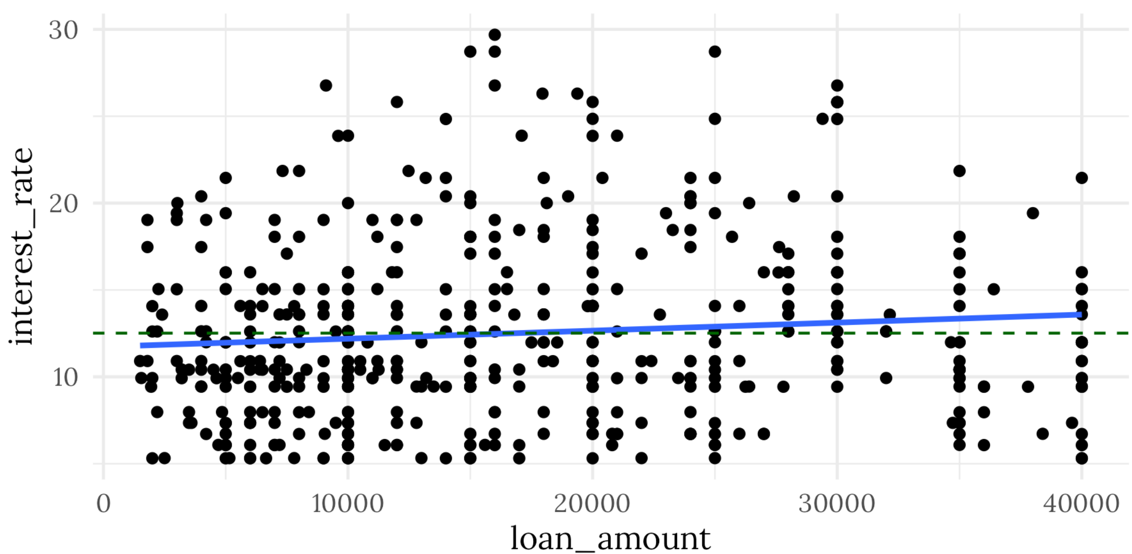


Recall our interpretation of the multiple regression coefficient: if other x-variables remain the same, we expect the y-variable to be $\hat{\beta}_i$ lower/higher, on average. If there is a strong degree of multicollinearity, we *can't* assume that all other x-variables can remain the same if x_i changes. This makes it quite difficult to interpret the coefficient or evaluate the p-value. In practice, there will almost always be some amount of collinearity among the predictor variables.

Even if the variables are very collinear, we likely have unbiased predictions of the response variable and so such models can still be useful for prediction purposes (rather than inference).

6. Inference

7. Sum of Squares



8. R^2 and adjusted R^2

```
summary(amount_income_lm)
```

```
##
## Call:
## lm(formula = interest_rate ~ annual_income + loan_amount, data = loans)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -8.57  -3.69  -0.98   2.49  16.83
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    1.23e+01   5.18e-01   23.73   <2e-16 ***
## annual_income -1.14e-05   5.64e-06   -2.02   0.0441 *
## loan_amount    6.51e-05   2.39e-05    2.73   0.0066 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.11 on 497 degrees of freedom
## Multiple R-squared:  0.0168, Adjusted R-squared:  0.0129
## F-statistic: 4.25 on 2 and 497 DF,  p-value: 0.0147
```

- R^2 for income-only model: 0.00211
- R^2 for amount-only model: 0.00877

```r
anova(amount_income_lm)
```

```
## Analysis of Variance Table
##
## Response: interest_rate
##                Df Sum Sq Mean Sq F value Pr(>F)
## annual_income   1     28    27.8    1.07 0.3022
## loan_amount     1    194   194.2    7.44 0.0066 **
## Residuals     497  12975    26.1
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

9. ANOVA for Regression

10. CI's and PI's

Just like in simple linear regression, we need to be careful when making confidence intervals for predictions. The math for computing standard errors gets even more tricky in the multiple regression setting, so we'll rely on software to compute it for us.

```
new_loan = tibble(
  annual_income = 87000,
  loan_amount = 40000
)
augment(amount_income_lm, newdata = new_loan, interval = "confidence")
```

```
## # A tibble: 1 x 5
##   annual_income loan_amount .fitted .lower .upper
##           <dbl>       <dbl>   <dbl>  <dbl>  <dbl>
## 1         87000       40000    13.9   12.8   15.1
```

```
augment(amount_income_lm, newdata = new_loan, interval = "prediction")
```

```
## # A tibble: 1 x 5
##   annual_income loan_amount .fitted .lower .upper
##           <dbl>       <dbl>   <dbl>  <dbl>  <dbl>
## 1         87000       40000    13.9   3.81   24.0
```