WEEK 05

INSTRUCTOR: YANAN WU

TA: KHADIJA NISAR

SPRING 2025

5.1 REGRESSION

MULTIPLE REGRESSION

- Simple linear regression: Bivariate - two variables: y and x
- Multiple linear regression: Multiple variables: y and $x_1, x_2, x_3, \cdots$

LOANS DATA

variable	description
`interest_rate`	Interest rate for the loan.
`income_ver`	Categorical variable describing whether the borrower's income source and amount have been verified, with levels `verified`, `source_only`, and `not`.
`debt_to_income`	Debt-to-income ratio, which is the percentage of total debt of the borrower divided by their total income.
`credit_util`	Of all the credit available to the borrower, what fraction are they utilizing. For example, the credit utilization on a credit card would be the card's balance divided by the card's credit limit.
`bankruptcy`	An indicator variable for whether the borrower has a past bankruptcy in her record. This variable takes a value of `1` if the answer is "yes" and `0` if the answer is "no".
`term`	The length of the loan, in months.
`issued`	The month and year the loan was issued, which for these loans is always during the first quarter of 2018.
`credit_checks`	Number of credit checks in the last 12 months. For example, when filing an application for a credit card, it is common for the company receiving the application to run a credit check.

Figure 9.2: Variables and their descriptions for the `loans` data set.

CATEGORICAL VARIABLE AS PREDICTORS

```
       X          interest_rate              income_ver   debt_to_income     credit_util     bankruptcy       term        issue
 Min.   :    1   Min.   : 5.31   Not Verified   :3573   Min.   :  0.00   Min.   :0.0000   no :8759   Min.   :36.00   1:3389
 1st Qu.: 2502   1st Qu.: 9.43   Source Verified:4112   1st Qu.: 11.06   1st Qu.:0.1690   yes:1213   1st Qu.:36.00   2:2979
 Median : 5000   Median :11.98   Verified       :2287   Median : 17.57   Median :0.3600              Median :36.00   3:3604
 Mean   : 5001   Mean   :12.42                          Mean   : 19.31   Mean   :0.4031              Mean   :43.27
 3rd Qu.: 7502   3rd Qu.:15.05                          3rd Qu.: 25.00   3rd Qu.:0.6070              3rd Qu.:60.00
 Max.   :10000   Max.   :30.94                          Max.   :469.09   Max.   :1.8350              Max.   :60.00
 credit_checks
 Min.   : 0.000
 1st Qu.: 0.000
 Median : 1.000
 Mean   : 1.958
 3rd Qu.: 3.000
 Max.   :29.000
```

Which variables are categorical variable？

2-LEVEL CATEGORICAL VARIABLE AS PREDICTORS

```
Residuals:
    Min      1Q  Median      3Q     Max
-7.7573 -3.6373 -0.4473  2.7181 18.6081

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   12.33189    0.05334 231.204  < 2e-16 ***
bankruptcyyes  0.73538    0.15293   4.809 1.54e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.992 on 9970 degrees of freedom
Multiple R-squared:  0.002314,  Adjusted R-squared:  0.002214
F-statistic: 23.12 on 1 and 9970 DF,  p-value: 1.542e-06
```

Reference level: $backruptcy$ = no

$$\widehat{rate} = 12.332 + 0.735 * backruptcy_{yes}$$

REGRESSION MODEL

$$\widehat{rate} = 12.332 + 0.735 * backruptcy_{yes}$$

1. Borrower with bankruptcy record
2. Borrower without bankruptcy record

A. $\widehat{rate} = 12.332 + 0.735 * 0$

B. $\widehat{rate} = 12.332 + 0.735 * 1$

Which value (with bankruptcy or without bankruptcy) has the higher interest rate?

2-LEVEL CATEGORICAL VARIABLE AS PREDICTORS

```
Residuals:
    Min      1Q  Median      3Q     Max
-7.7573 -3.6373 -0.4473  2.7181 18.6081

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  12.33189    0.05334 231.204  < 2e-16 ***
bankruptcyyes  0.73538    0.15293   4.809 1.54e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.992 on 9970 degrees of freedom
Multiple R-squared:  0.002314,  Adjusted R-squared:  0.002214
F-statistic: 23.12 on 1 and 9970 DF,  p-value: 1.542e-06
```

- **Intercept:** The estimated average interest rate for borrower without a bankruptcy record is 12.332

  This is the value we get if we plug in 0 for the explanatory variable

- **Slope:** The estimated average interest rate for borrower with a bankruptcy record is 0.735 higher than borrower without a bankruptcy record

  Then, the estimated average interest rate for borrower with a bankruptcy record is 12.332 + 0.735= 13.067

  This is the value we get if we plug in 1 for the explanatory variable

3-LEVEL CATEGORICAL VARIABLE AS PREDICTORS

```
Residuals:
    Min      1Q  Median      3Q     Max
-9.0466 -3.7249 -0.6549  2.5350 19.7151

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                 11.07486    0.08104   136.7   <2e-16 ***
income_verSource Verified    1.44009    0.11079    13.0   <2e-16 ***
income_verVerified           3.28178    0.12972    25.3   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.844 on 9969 degrees of freedom
Multiple R-squared:  0.06056,   Adjusted R-squared:  0.06037
F-statistic: 321.3 on 2 and 9969 DF,  p-value: < 2.2e-16
```

Which income verification (Source verified, verified, Non verified) is the reference level?

$$\widehat{rate} = 11.075 + 1.44 * income_{sourceVer} + 2.282 * income_{verified}$$

REGRESSION MODEL

$$\widehat{rate} = 11.075 + 1.44 * income_{sourceVer} + 2.282 * income_{verified}$$

1. *Income_ver* take a value of $sourceVeri$
2. *Income_ver* take a value of $Verified$
3. *Income_ver* take a value of $\mathrm{Non}Verified$

A. $\widehat{rate} = 11.075 + 1.44 * 0 + 2.282 * 0$

B. $\widehat{rate} = 11.075 + 1.44 * 1 + 2.282 * 0$

C. $\widehat{rate} = 11.075 + 1.44 * 0 + 2.282 * 1$

Which income verification approach has the highest interest rate?

1. Source verification
2. Verification
3. Not verification
4. Can not tell

ASSESSING MANY x_i IN A MODEL

lm(formula = interest_rate ~ income_ver + debt_to_income + credit_util + bankruptcy + issue + credit_checks, data = loans)

Multiple regression aims to minimize SSE:

$$SSE = e_1^2 + e_2^2 + \cdots + e_n^2 = \sum_{i=1}^{n} e_i^2$$

OUTPUT FOR THE REGRESSION MODEL

Which two variables are not important?

$$credit_{util} = \frac{credit\ banlance}{credit\ limit}$$

Interpretation for β_{credit_util}:

All other variables held constant, if someone's credit utilization increase 1 unit, the change in interest rate is 4.90.

```
Call:
lm(formula = interest_rate ~ income_ver + debt_to_income + credit_util +
    bankruptcy + term + issue + credit_checks, data = loans)

Residuals:
     Min       1Q   Median       3Q      Max
-12.1117  -3.1003  -0.7256   2.3307  18.8157

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)               1.942374   0.206167   9.421  < 2e-16 ***
income_verSource Verified 0.998043   0.099200  10.061  < 2e-16 ***
income_verVerified        2.562094   0.117203  21.860  < 2e-16 ***
debt_to_income            0.021808   0.002937   7.425 1.22e-13 ***
credit_util               4.897255   0.161900  30.249  < 2e-16 ***
bankruptcyyes             0.391221   0.132295   2.957  0.00311 **
term                      0.153340   0.003945  38.871  < 2e-16 ***
issue2                   -0.047150   0.108057  -0.436  0.66259
issue3                   -0.087727   0.102938  -0.852  0.39410
credit_checks             0.228311   0.018244  12.514  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.3 on 9962 degrees of freedom
Multiple R-squared:  0.2602,    Adjusted R-squared:  0.2596
F-statistic: 389.4 on 9 and 9962 DF,  p-value: < 2.2e-16
```

INTERPRETATION

All else being equal, borrower with bankruptcy record

a) are estimated to have interest rate 0.39 lower

b) are estimated to have interest rate 0.39 higher

than borrower without bankruptcy record

```
Call:
lm(formula = interest_rate ~ income_ver + debt_to_income + credit_util +
    bankruptcy + term + issue + credit_checks, data = loans)

Residuals:
     Min       1Q   Median       3Q      Max
-12.1117  -3.1003  -0.7256   2.3307  18.8157

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                 1.942374   0.206167   9.421  < 2e-16 ***
income_verSource Verified   0.998043   0.099200  10.061  < 2e-16 ***
income_verVerified          2.562094   0.117203  21.860  < 2e-16 ***
debt_to_income              0.021808   0.002937   7.425 1.22e-13 ***
credit_util                 4.897255   0.161900  30.249  < 2e-16 ***
bankruptcyyes               0.391221   0.132295   2.957  0.00311 **
term                        0.153340   0.003945  38.871  < 2e-16 ***
issue2                     -0.047150   0.108057  -0.436  0.66259
issue3                     -0.087727   0.102938  -0.852  0.39410
credit_checks               0.228311   0.018244  12.514  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.3 on 9962 degrees of freedom
Multiple R-squared:  0.2602,    Adjusted R-squared:  0.2596
F-statistic: 389.4 on 9 and 9962 DF,  p-value: < 2.2e-16
```

R^2 VS. ADJUSTED R^2

$$R^2 = \frac{explained\ variability}{total\ variability}$$

$$R^2_{adj} = 1 - \left(\frac{SS_{Error}}{SS_{Total}} \times \frac{n-1}{n-p-1} \right)$$

where n is the number of cases and p is the number of predictors (explanatory variables) in the model.

- Because p is never negative, R^2_{adj} will always be smaller than R^2.
- R^2_{adj} applies a penalty for the number of predictors included in the model.
- Therefore, we choose models with higher R^2_{adj} over others.

R^2 VS. ADJUSTED R^2

```
Call:
lm(formula = interest_rate ~ income_ver + debt_to_income + credit_util +
    bankruptcy + term + issue + credit_checks, data = loans)

Residuals:
     Min       1Q   Median       3Q      Max
-12.1117  -3.1003  -0.7256   2.3307  18.8157

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)               1.942374   0.206167   9.421  < 2e-16 ***
income_verSource Verified 0.998043   0.099200  10.061  < 2e-16 ***
income_verVerified        2.562094   0.117203  21.860  < 2e-16 ***
debt_to_income            0.021808   0.002937   7.425 1.22e-13 ***
credit_util               4.897255   0.161900  30.249  < 2e-16 ***
bankruptcyyes             0.391221   0.132295   2.957  0.00311 **
term                      0.153340   0.003945  38.871  < 2e-16 ***
issue2                   -0.047150   0.108057  -0.436  0.66259
issue3                   -0.087727   0.102938  -0.852  0.39410
credit_checks             0.228311   0.018244  12.514  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.3 on 9962 degrees of freedom
Multiple R-squared:  0.2602,    Adjusted R-squared:  0.2596
F-statistic: 389.4 on 9 and 9962 DF,  p-value: < 2.2e-16
```

COLLINEARITY BETWEEN INDEPENDENT VARIABLES

- Two independent variables are said to be collinear when they are correlated, and this collinearity complicates model estimation.
- We don't like adding independent that are associated with each other to the model, because often times the addition of such variable brings nothing to the table. Instead, we prefer the simplest best model, i.e. **parsimonious** model.
- While it's impossible to avoid collinearity from arising in observational data, experiments are usually designed to prevent correlation among predictors.

COLLINEARITY DIAGNOSTICS

VIF: variance-inflation factor

- Variance inflation factor measure the inflation in the variances of the parameter estimates due to collinearities that exist among the predictors.

$$VIF_i = \frac{1}{1 - R_i^2} = \frac{1}{Tolerance}$$

Step 1: $x_1 = \alpha_0 + \alpha_2 x_2 + \alpha_3 x_3 + \cdots \alpha_i x_i + \varepsilon$

where R_1^2 is the coefficient of determination of the regression equation in step one

COLLINEARITY DIAGNOSTICS

- $VIF > 5$: A cutoff of 5 is also commonly used
- VIF > 10: indicating multicollinearity is high

	GVIF	Df	GVIF^(1/(2*Df))
income_ver	1.053346	2	1.013078
debt_to_income	1.047341	1	1.023397
credit_util	1.025391	1	1.012616
bankruptcy	1.008434	1	1.004208
term	1.020978	1	1.010434
issue	1.002292	2	1.000573
credit_checks	1.017132	1	1.008530

WEEK 03

CODE DEMO SESSION

Instructor: Yanan Wu
TA: Khadija Nisar

Spring 2025