Multiple Linear Regression

Nate Wells

Math 141, 4/28/21

Outline

In this lecture, we will...

- Discuss framework for multiple linear regression and compare to simple linear regression
- Use the `moderndive` packages to create multiple regression models.
- Quantify variance in a linear model using the correlation coefficient

Section 1

Multiple Linear Regression

Many Simple Linear Regression Models

We are often presented situations where several explanatory variables could be used to predict values of a single response variable.

Many Simple Linear Regression Models

We are often presented situations where several explanatory variables could be used to predict values of a single response variable.

- **Response**: Reed thesis page count
- **Potential Explanatory**: year, division, number of check-outs

Many Simple Linear Regression Models

We are often presented situations where several explanatory variables could be used to predict values of a single response variable.

- **Response**: Reed thesis page count
- **Potential Explanatory**: year, division, number of check-outs

- **Response**: Home prices
- **Potential Explanatory**: square feet, number of bedrooms, number of bathrooms

Many Simple Linear Regression Models

We are often presented situations where several explanatory variables could be used to predict values of a single response variable.

- **Response**: Reed thesis page count
- **Potential Explanatory**: year, division, number of check-outs

- **Response**: Home prices
- **Potential Explanatory**: square feet, number of bedrooms, number of bathrooms

- **Response**: Household income
- **Potential Explanatory**: household size, years of education, state of residency

Many Simple Linear Regression Models

We are often presented situations where several explanatory variables could be used to predict values of a single response variable.

- **Response**: Reed thesis page count
- **Potential Explanatory**: year, division, number of check-outs

- **Response**: Home prices
- **Potential Explanatory**: square feet, number of bedrooms, number of bathrooms

- **Response**: Household income
- **Potential Explanatory**: household size, years of education, state of residency

In each case, we could create simple linear regression models for each explanatory variable.

Many Simple Linear Regression Models

We are often presented situations where several explanatory variables could be used to predict values of a single response variable.

- **Response**: Reed thesis page count
- **Potential Explanatory**: year, division, number of check-outs

- **Response**: Home prices
- **Potential Explanatory**: square feet, number of bedrooms, number of bathrooms

- **Response**: Household income
- **Potential Explanatory**: household size, years of education, state of residency

In each case, we could create simple linear regression models for each explanatory variable.

- But the results may be misleading. Several explanatory variables may be highly correlated.

Many Simple Linear Regression Models

We are often presented situations where several explanatory variables could be used to predict values of a single response variable.

- **Response**: Reed thesis page count
- **Potential Explanatory**: year, division, number of check-outs

- **Response**: Home prices
- **Potential Explanatory**: square feet, number of bedrooms, number of bathrooms

- **Response**: Household income
- **Potential Explanatory**: household size, years of education, state of residency

In each case, we could create simple linear regression models for each explanatory variable.

- But the results may be misleading. Several explanatory variables may be highly correlated.

Could we get better predictive power by including all explanatory variables in the *same* model?

Multiple Regression Model

In a **simple linear regression model** (SLR), we express the response variable Y as a linear function of one explanatory variable X:

$$\hat{Y} = \beta_0 + \beta_1 \cdot X$$

Multiple Regression Model

In a **simple linear regression model** (SLR), we express the response variable Y as a linear function of one explanatory variable X:

$$\hat{Y} = \beta_0 + \beta_1 \cdot X$$

In a **multiple linear regression model** (MLR), we express the response variable Y as a linear combination of k explanatory variables $X_1, X_2, \ldots, X_k$:

$$\hat{Y} = \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \cdots + \beta_k \cdot X_k$$

Multiple Regression Model

In a **simple linear regression model** (SLR), we express the response variable Y as a linear function of one explanatory variable X:

$$\hat{Y} = \beta_0 + \beta_1 \cdot X$$

In a **multiple linear regression model** (MLR), we express the response variable Y as a linear combination of k explanatory variables $X_1, X_2, \ldots, X_k$:

$$\hat{Y} = \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \cdots + \beta_k \cdot X_k$$

- In the MLR model, we allow the explanatory variables to either be quantitative or binary categorical (i.e taking values 0 or 1 corresponding to failure or success)

Multiple Regression Model

In a **simple linear regression model** (SLR), we express the response variable Y as a linear function of one explanatory variable X:

$$\hat{Y} = \beta_0 + \beta_1 \cdot X$$

In a **multiple linear regression model** (MLR), we express the response variable Y as a linear combination of k explanatory variables $X_1, X_2, \dots, X_k$:

$$\hat{Y} = \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \cdots + \beta_k \cdot X_k$$

- In the MLR model, we allow the explanatory variables to either be quantitative or binary categorical (i.e taking values 0 or 1 corresponding to failure or success)
- While we lose a nice 2D graphical representation (although higher dimensional graphics are possible), statistical software allows us to estimate coefficients of the model.

Finding Parameters

To create an SLR model, we found the equation of a line that minimizes the sum of squared residuals, where

$$\text{Residual} = \text{Observed} - \text{Predicted} \qquad e_i = \hat{y}_i - y_i$$

Finding Parameters

To create an SLR model, we found the equation of a line that minimizes the sum of squared residuals, where

$$\text{Residual} = \text{Observed} - \text{Predicted} \qquad e_i = \hat{y}_i - y_i$$

To create an MLR model, we do the exact same thing!

Finding Parameters

To create an SLR model, we found the equation of a line that minimizes the sum of squared residuals, where

$$\text{Residual} = \text{Observed} - \text{Predicted} \qquad e_i = \hat{y}_i - y_i$$

To create an MLR model, we do the exact same thing!

- The only difference is that instead of the equation describing a line, the equation describes a plane in multidimensional space.
  - If we have 2 explanatory variables, the equation describes a plane in 3D space.

Finding Parameters

To create an SLR model, we found the equation of a line that minimizes the sum of squared residuals, where

$$\text{Residual} = \text{Observed} - \text{Predicted} \qquad e_i = \hat{y}_i - y_i$$

To create an MLR model, we do the exact same thing!

- The only difference is that instead of the equation describing a line, the equation describes a plane in multidimensional space.
  - If we have 2 explanatory variables, the equation describes a plane in 3D space.

We even use the exact same R code to fit the linear model:

```
mod<-lm(Y ~ X1 + X2 + ... + Xk, data = my_data)
```

Credit Card Debt

The `Credit` dataset in the `ISLR` package contains (fabricated) credit card debt and other financial and demographic information for 400 individuals.

Credit Card Debt

The `Credit` dataset in the `ISLR` package contains (fabricated) credit card debt and other financial and demographic information for 400 individuals.

Goal: Build a model that allows us to predict credit debt given financial and demographic information

Credit Card Debt

The `Credit` dataset in the `ISLR` package contains (fabricated) credit card debt and other financial and demographic information for 400 individuals.

Goal: Build a model that allows us to predict credit debt given financial and demographic information

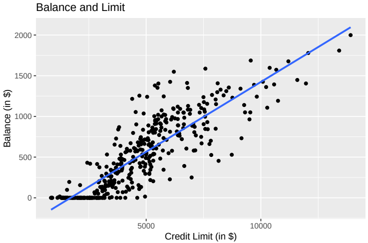
We first consider `Balance` as a function of `Limit` and `Income`

Credit Card Debt

The `Credit` dataset in the `ISLR` package contains (fabricated) credit card debt and other financial and demographic information for 400 individuals.

Goal: Build a model that allows us to predict credit debt given financial and demographic information

We first consider `Balance` as a function of `Limit` and `Income`



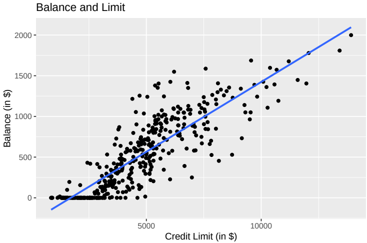
$R = 0.86 \qquad \hat{\text{Balance}} = -292.8 + 0.17 \cdot \text{Limit}$

Credit Card Debt

The `Credit` dataset in the `ISLR` package contains (fabricated) credit card debt and other financial and demographic information for 400 individuals.

Goal: Build a model that allows us to predict credit debt given financial and demographic information

We first consider `Balance` as a function of `Limit` and `Income`



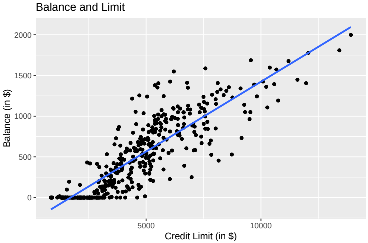
$R = 0.86$ $\hat{\text{Balance}} = -292.8 + 0.17 \cdot \text{Limit}$ $R = 0.46$ $\hat{\text{Balance}} = 246.51 + 6.048 \cdot \text{Income}$

Credit Card Debt

The Credit dataset in the ISLR package contains (fabricated) credit card debt and other financial and demographic information for 400 individuals.

Goal: Build a model that allows us to predict credit debt given financial and demographic information

We first consider Balance as a function of Limit and Income



$R = 0.86 \qquad \hat{\text{Balance}} = -292.8 + 0.17 \cdot \text{Limit} \qquad R = 0.46 \qquad \hat{\text{Balance}} = 246.51 + 6.048 \cdot \text{Income}$

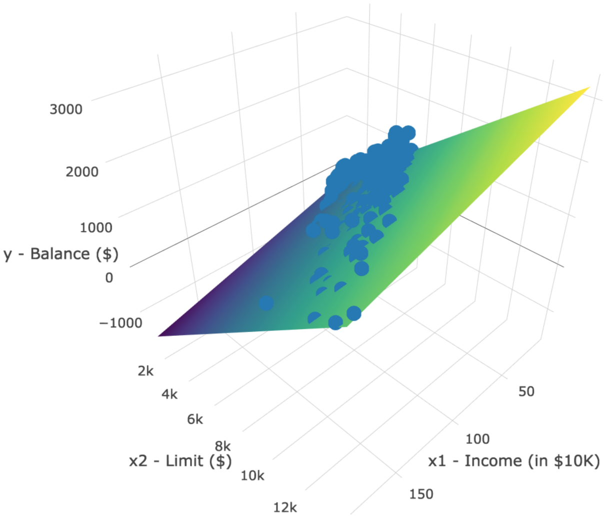
Both variables have some explanatory power for Balance

The Regression Plane

How do Limit and Income *together* explain Balance?

The Regression Plane

How do Limit and Income *together* explain Balance?



Multiple Regression for Debt

Let's find the MLR model

```
mod<-lm(Balance ~ Limit + Income, data = Credit)
```

Multiple Regression for Debt

Let's find the MLR model

```
mod<-lm(Balance ~ Limit + Income, data = Credit)
```

And investigate the regression table

```
get_regression_table(mod)
```

```
## # A tibble: 3 x 7
##   term      estimate std_error statistic p_value lower_ci upper_ci
##   <chr>        <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
## 1 intercept -385.       19.5       -19.8       0 -423.    -347.
## 2 Limit        0.264     0.006      45.0       0    0.253    0.276
## 3 Income      -7.66      0.385     -19.9       0   -8.42    -6.91
```

Multiple Regression for Debt

Let's find the MLR model

```
mod<-lm(Balance ~ Limit + Income, data = Credit)
```

And investigate the regression table

```
get_regression_table(mod)
```

```
## # A tibble: 3 x 7
##   term      estimate std_error statistic p_value lower_ci upper_ci
##   <chr>        <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
## 1 intercept -385.       19.5       -19.8       0 -423.    -347.
## 2 Limit        0.264     0.006      45.0       0    0.253    0.276
## 3 Income      -7.66      0.385     -19.9       0   -8.42    -6.91
```

Which gives us the regression equation:

$$\hat{\text{Balance}} = -385.179 + 0.264 \cdot \text{Limit} - 0.7663 \cdot \text{Income}$$

Multiple Regression for Debt

Let's find the MLR model

```
mod<-lm(Balance ~ Limit + Income, data = Credit)
```

And investigate the regression table

```
get_regression_table(mod)
```

```
## # A tibble: 3 x 7
##   term      estimate std_error statistic p_value lower_ci upper_ci
##   <chr>        <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
## 1 intercept -385.       19.5       -19.8       0 -423.    -347.
## 2 Limit        0.264     0.006      45.0       0    0.253    0.276
## 3 Income      -7.66      0.385     -19.9       0   -8.42    -6.91
```

Which gives us the regression equation:

$$\widehat{\text{Balance}} = -385.179 + 0.264 \cdot \text{Limit} - 0.7663 \cdot \text{Income}$$

- For **fixed** value of Income, increasing Credit Limit by $1 increases Balance by an average of $0.264.

Multiple Regression for Debt

Let's find the MLR model

```
mod<-lm(Balance ~ Limit + Income, data = Credit)
```

And investigate the regression table

```
get_regression_table(mod)
```

```
## # A tibble: 3 x 7
##   term      estimate std_error statistic p_value lower_ci upper_ci
##   <chr>        <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
## 1 intercept -385.       19.5       -19.8       0 -423.    -347.
## 2 Limit        0.264     0.006      45.0       0    0.253    0.276
## 3 Income      -7.66      0.385     -19.9       0   -8.42    -6.91
```

Which gives us the regression equation:

$$\hat{\text{Balance}} = -385.179 + 0.264 \cdot \text{Limit} - 0.7663 \cdot \text{Income}$$

- For **fixed** value of Income, increasing Credit Limit by $1 increases Balance by an average of $0.264.
- While for **fixed** value of Limit, increasing Income by $1000 decreases Balance by an average of $7.66.

Comparing MLR and SLR

Wait. . .

Comparing MLR and SLR

Wait. . .

- The SLR for Balance and Income was

$$\hat{\text{Balance}} = 246.51 + 6.048 \cdot \text{Income}$$

Comparing MLR and SLR

Wait. . .

- The SLR for Balance and Income was

$$\hat{\text{Balance}} = 246.51 + 6.048 \cdot \text{Income}$$

- That is, increasing Income by $1000 **INCREASED** Balance by $6.05.

Comparing MLR and SLR

Wait. . .

- The SLR for Balance and Income was

$$\hat{\text{Balance}} = 246.51 + 6.048 \cdot \text{Income}$$

- That is, increasing Income by $1000 **INCREASED** Balance by $6.05.
- But the MLR is

$$\hat{\text{Balance}} = -385.179 + 0.264 \cdot \text{Limit} - 0.7663 \cdot \text{Income}$$

Comparing MLR and SLR

Wait. . .

- The SLR for Balance and Income was

$$\hat{\text{Balance}} = 246.51 + 6.048 \cdot \text{Income}$$

- That is, increasing Income by \$1000 **INCREASED** Balance by \$6.05.
- But the MLR is

$$\hat{\text{Balance}} = -385.179 + 0.264 \cdot \text{Limit} - 0.7663 \cdot \text{Income}$$

- Not only has MLR given us a new rate of change, but it's completely switched the direction!

Comparing MLR and SLR

Wait. . .

- The SLR for Balance and Income was

$$\hat{\text{Balance}} = 246.51 + 6.048 \cdot \text{Income}$$

- That is, increasing Income by $1000 **INCREASED** Balance by $6.05.
- But the MLR is

$$\hat{\text{Balance}} = -385.179 + 0.264 \cdot \text{Limit} - 0.7663 \cdot \text{Income}$$

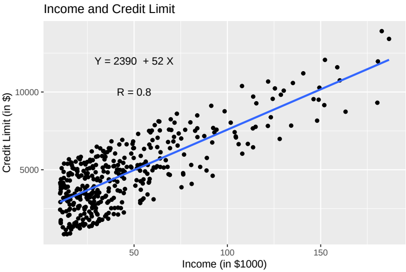
- Not only has MLR given us a new rate of change, but it's completely switched the direction!
- How is this possible?

Income and Credit Limit

Let's consider the relationship between income and credit limit

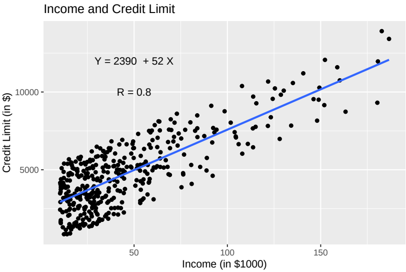
Income and Credit Limit

Let's consider the relationship between income and credit limit



Income and Credit Limit

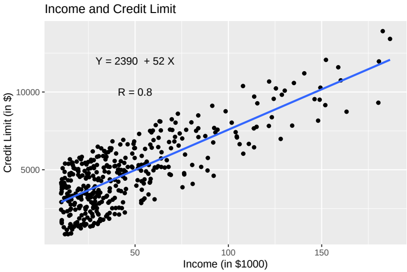
Let's consider the relationship between income and credit limit



In a vacuum, as income increases, so too does credit limit.

Income and Credit Limit

Let's consider the relationship between income and credit limit

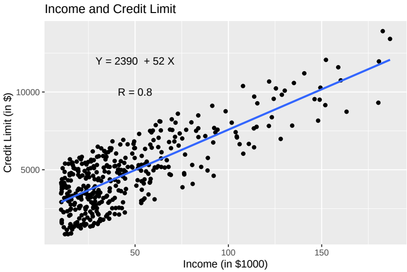


In a vacuum, as income increases, so too does credit limit.

- So in the SLR model, when we assess the change in Debt due to increase in Income, we are implicitly also increasing Credit Limit

Income and Credit Limit

Let's consider the relationship between income and credit limit



In a vacuum, as income increases, so too does credit limit.

- So in the SLR model, when we assess the change in Debt due to increase in Income, we are implicitly also increasing Credit Limit
  - We could say Credit Limit is a confounding variable in the SLR model.

The Regression Plane Revisited

In the MLR model, we may freely change both Income and Credit Limit

The Regression Plane Revisited

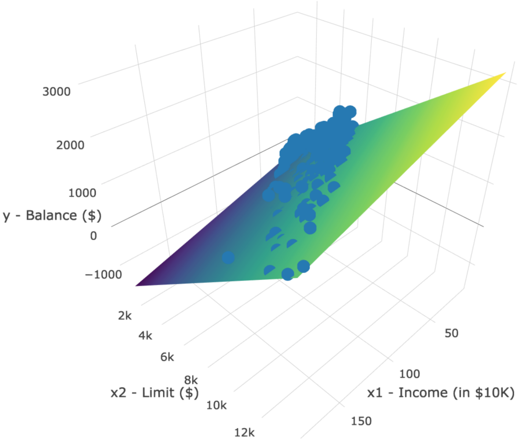
In the MLR model, we may freely change both Income and Credit Limit

- This corresponds to the fact that there is a unique Balance point on the regression plane for each pair of Income / Credit Limit values.

The Regression Plane Revisited

In the MLR model, we may freely change both Income and Credit Limit

- This corresponds to the fact that there is a unique Balance point on the regression plane for each pair of Income / Credit Limit values.



Debt vs. Income Revisited

We can lump Credit Limits into 4 brackets (low, med-low, med-high, high) to create a categorical variable and analyze the SLR for Balance and Income for each level of Credit Limit

Debt vs. Income Revisited

We can lump Credit Limits into 4 brackets (low, med-low, med-high, high) to create a categorical variable and analyze the SLR for Balance and Income for each level of Credit Limit

```
Credit_bracket<-Credit %>%
  mutate(credit_bracket = case_when(
    Limit < quantile(Limit, .25) ~ "low",
    Limit > quantile(Limit, .25) & Limit < median(Limit) ~ "med-low",
    Limit > median(Limit) & Limit < quantile(Limit, .75) ~ "med-high",
    Limit > quantile(Limit, .75) ~ "high")) %>%
  mutate(credit_bracket = fct_relevel(
    credit_bracket, "high", "med-high", "med-low", "low"))
```

```
##   Income Limit Balance credit_bracket
## 1     15  3606     333        med-low
## 2    106  6645     903           high
## 3    105  7075     580           high
## 4    149  9504     964           high
## 5     56  4897     331       med-high
## 6     80  8047    1151           high
```

Debt vs. Income Revisited

Debt vs. Income Revisited

- Note that within each credit bracket, increasing income corresponds to either decreasing or relatively flat change in Balance

Debt vs. Income Revisited

- Note that within each credit bracket, increasing income corresponds to either decreasing or relatively flat change in Balance
  - This is an example of **Simpson's Paradox**: a trend present in the aggregate data can reverse itself when data is considered by group.

How Strong is a Linear Model?

For SLR, we used the correlation coefficient R to assess model strength.

How Strong is a Linear Model?

For SLR, we used the correlation coefficient R to assess model strength.

- The value R^2 has utility too! It represents the percentage of variability in values of the response variable just due to variability in explanatory variable.

How Strong is a Linear Model?

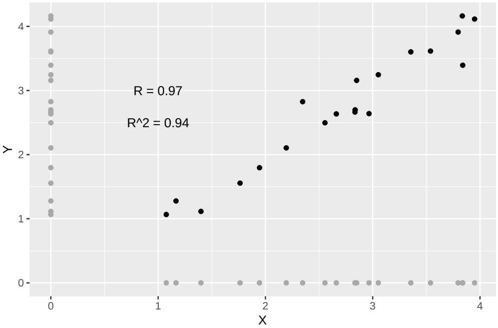
For SLR, we used the correlation coefficient R to assess model strength.

- The value R^2 has utility too! It represents the percentage of variability in values of the response variable just due to variability in explanatory variable.
  - If $R \approx \pm 1$, then $R^2 \approx 1$: nearly all the variability in response is due to variability in the explanatory variable.

How Strong is a Linear Model?

For SLR, we used the correlation coefficient R to assess model strength.

- The value R^2 has utility too! It represents the percentage of variability in values of the response variable just due to variability in explanatory variable.
  - If $R \approx \pm 1$, then $R^2 \approx 1$: nearly all the variability in response is due to variability in the explanatory variable.



Model Strength for MLR

We can also compute R^2 for MLR. In particular,

$$R^2 = 1 - \frac{\text{variability in residuals}}{\text{variability in outcomes}} = 1 - \frac{\mathrm{Var}(e_i)}{\mathrm{Var}(y_i)}$$

Model Strength for MLR

We can also compute R^2 for MLR. In particular,

$$R^2 = 1 - \frac{\text{variability in residuals}}{\text{variability in outcomes}} = 1 - \frac{\mathrm{Var}(e_i)}{\mathrm{Var}(y_i)}$$

- Usually, we use software to compute

Model Strength for MLR

We can also compute R^2 for MLR. In particular,

$$R^2 = 1 - \frac{\text{variability in residuals}}{\text{variability in outcomes}} = 1 - \frac{\text{Var}(e_i)}{\text{Var}(y_i)}$$

- Usually, we use software to compute

```
mod_credit<-lm(Balance ~ Income + Limit , data = Credit)
get_regression_summaries(mod_credit)
```

```
## # A tibble: 1 x 9
##   r_squared adj_r_squared    mse  rmse sigma statistic p_value    df  nobs
##       <dbl>         <dbl>  <dbl> <dbl> <dbl>     <dbl>   <dbl> <dbl> <dbl>
## 1     0.871          0.87 27177.  165.  165.     1342.       0     2   400
```

Model Strength for MLR

We can also compute R^2 for MLR. In particular,

$$R^2 = 1 - \frac{\text{variability in residuals}}{\text{variability in outcomes}} = 1 - \frac{\text{Var}(e_i)}{\text{Var}(y_i)}$$

- Usually, we use software to compute

```
mod_credit<-lm(Balance ~ Income + Limit , data = Credit)
get_regression_summaries(mod_credit)
```

```
## # A tibble: 1 x 9
##   r_squared adj_r_squared    mse  rmse sigma statistic p_value    df  nobs
##       <dbl>         <dbl>  <dbl> <dbl> <dbl>     <dbl>   <dbl> <dbl> <dbl>
## 1     0.871          0.87 27177.  165.  165.     1342.       0     2   400
```

- But it turns out this formula gives a **biased** estimate of the variability in the *population* explained by the model.

Model Strength for MLR

We can also compute R^2 for MLR. In particular,

$$R^2 = 1 - \frac{\text{variability in residuals}}{\text{variability in outcomes}} = 1 - \frac{\text{Var}(e_i)}{\text{Var}(y_i)}$$

- Usually, we use software to compute

```
mod_credit<-lm(Balance ~ Income + Limit , data = Credit)
get_regression_summaries(mod_credit)
```

```
## # A tibble: 1 x 9
##   r_squared adj_r_squared    mse  rmse sigma statistic p_value    df  nobs
##       <dbl>         <dbl>  <dbl> <dbl> <dbl>     <dbl>   <dbl> <dbl> <dbl>
## 1     0.871          0.87 27177.  165.  165.     1342.       0     2   400
```

- But it turns out this formula gives a **biased** estimate of the variability in the *population* explained by the model.
- Instead, we use the adjusted R:

$$R^2 = 1 - \frac{\text{Var}(e_i)}{\text{Var}(y_i)} \cdot \frac{n-1}{n-k-1}$$

Model Strength for MLR

We can also compute R^2 for MLR. In particular,

$$R^2 = 1 - \frac{\text{variability in residuals}}{\text{variability in outcomes}} = 1 - \frac{\mathrm{Var}(e_i)}{\mathrm{Var}(y_i)}$$

- Usually, we use software to compute

```
mod_credit<-lm(Balance ~ Income + Limit , data = Credit)
get_regression_summaries(mod_credit)
```

```
## # A tibble: 1 x 9
##   r_squared adj_r_squared    mse  rmse sigma statistic p_value    df  nobs
##       <dbl>         <dbl>  <dbl> <dbl> <dbl>     <dbl>   <dbl> <dbl> <dbl>
## 1     0.871          0.87 27177.  165.  165.     1342.       0     2   400
```

- But it turns out this formula gives a **biased** estimate of the variability in the *population* explained by the model.
- Instead, we use the adjusted R:

$$R^2 = 1 - \frac{\mathrm{Var}(e_i)}{\mathrm{Var}(y_i)} \cdot \frac{n-1}{n-k-1}$$

- This adjusted R^2 is usually a bit smaller than R^2, and the difference decreases as n gets large.