Multilinear Regression

Prof Wells

STA 295: Stat Learning

February 13th, 2024

Outline

In today's class, we will. . .

- Generalize the simple regression model to include more than 1 predictor
- Quantify model accuracy for linear regression models (both simple and multiple)
- Implement multiple regression in R

Section 1

Multiple 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**: Home price
- **Predictors**: 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**: Home price
- **Predictors**: square feet, number of bedrooms, number of bathrooms

- **Response**: State high school grad rate
- **Predictors**: poverty rate, fed per capita spending, unemployment rate, percent home ownership

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**: Home price
- **Predictors**: square feet, number of bedrooms, number of bathrooms

- **Response**: State high school grad rate
- **Predictors**: poverty rate, fed per capita spending, unemployment rate, percent home ownership

In each case, we could create simple linear regression models for each predictor 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**: Home price
- **Predictors**: square feet, number of bedrooms, number of bathrooms

- **Response**: State high school grad rate
- **Predictors**: poverty rate, fed per capita spending, unemployment rate, percent home ownership

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

- But its not clear how to combine estimates from multiple models.

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**: Home price
- **Predictors**: square feet, number of bedrooms, number of bathrooms

- **Response**: State high school grad rate
- **Predictors**: poverty rate, fed per capita spending, unemployment rate, percent home ownership

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

- But its not clear how to combine estimates from multiple models.

- 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**: Home price
- **Predictors**: square feet, number of bedrooms, number of bathrooms

- **Response**: State high school grad rate
- **Predictors**: poverty rate, fed per capita spending, unemployment rate, percent home ownership

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

- But its not clear how to combine estimates from multiple models.
- The results may be misleading. Several explanatory variables may be highly correlated.
- And even if none of the predictors have strong association with the response, we are likely to observe a significant predictor just due to chance.

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**: Home price
- **Predictors**: square feet, number of bedrooms, number of bathrooms

- **Response**: State high school grad rate
- **Predictors**: poverty rate, fed per capita spending, unemployment rate, percent home ownership

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

- But its not clear how to combine estimates from multiple models.
- The results may be misleading. Several explanatory variables may be highly correlated.
- And even if none of the predictors have strong association with the response, we are likely to observe a significant predictor just due to chance.

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 f of one predictor variable X:

$$Y = f(X) + \epsilon$$

and estimate f using

$$\hat{Y} = \hat{f}(X) = \hat{\beta}_0 + \hat{\beta}_1 X$$

Multiple Regression Model

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

$$Y = f(X) + \epsilon$$

and estimate f using

$$\hat{Y} = \hat{f}(X) = \hat{\beta}_0 + \hat{\beta}_1 X$$

In a **multiple linear regression model** (MLR), we express the response variable Y as a linear combination of p predictors $X_1, X_2, \ldots, X_p$:

$$Y = f(X_1, \ldots, X_p) + \epsilon$$

and estimate f using

$$\hat{Y} = \hat{f}(X_1, \ldots, X_p) = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \cdots + \hat{\beta}_p X_p$$

Multiple Regression Model

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

$$Y = f(X) + \epsilon$$

and estimate f using

$$\hat{Y} = \hat{f}(X) = \hat{\beta}_0 + \hat{\beta}_1 X$$

In a **multiple linear regression model** (MLR), we express the response variable Y as a linear combination of p predictors $X_1, X_2, \ldots, X_p$:

$$Y = f(X_1, \ldots, X_p) + \epsilon$$

and estimate f using

$$\hat{Y} = \hat{f}(X_1, \ldots, X_p) = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \cdots + \hat{\beta}_p X_p$$

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

Finding Parameters

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

$$\mathrm{RSS} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} \Big(y_i - (\beta_0 + \beta_1 x_1) \Big),$$

which has the solution

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Finding Parameters

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

$$\mathrm{RSS} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} \Big(y_i - (\beta_0 + \beta_1 x_1) \Big),$$

which has the solution

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

And in R, we computed the coefficients using

```
my_mod<-lm(Y ~ X, data = my_data)
summary(my_mod)
```

Finding Parameters

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

$$\mathrm{RSS} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}\Big(y_i - (\beta_0 + \beta_1 x_1)\Big),$$

which has the solution

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

And in R, we computed the coefficients using

```
my_mod<-lm(Y ~ X, data = my_data)
summary(my_mod)
```

To fit an MLR model...

Finding Parameters

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

$$\mathrm{RSS} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}\Big(y_i - (\beta_0 + \beta_1 x_1)\Big),$$

which has the solution

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$$

And in R, we computed the coefficients using

```
my_mod<-lm(Y ~ X, data = my_data)
summary(my_mod)
```

To fit an MLR model. . .

we do the exact same thing!

Finding Parameters MLR

To create a MLR model, we find the values of $(\beta_0, \beta_1, \ldots, \beta_p)$ that minimizes RSS, where

$$\mathrm{RSS} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}\Big(y_i - (\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p)\Big)^2,$$

which has the solution

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

Finding Parameters MLR

To create a MLR model, we find the values of $(\beta_0, \beta_1, \ldots, \beta_p)$ that minimizes RSS, where

$$\mathrm{RSS} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} \Big(y_i - (\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p) \Big)^2,$$

which has the solution

$$\hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$$

- $\hat{\beta}$ is the $(p+1)$-vector of coefficient estimates $(\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p)$
- $\mathbf{y}$ is the n-vector of observed responses
- $\mathbf{X}$ is the matrix (or dataframe) consisting of n rows of observations on p predictors (along with a column of 1's to create an *intercept*).

Finding Parameters MLR

To create a MLR model, we find the values of $(\beta_0, \beta_1, \ldots, \beta_p)$ that minimizes RSS, where

$$\mathrm{RSS} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} \Big(y_i - (\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p) \Big)^2,$$

which has the solution

$$\hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$$

- $\hat{\beta}$ is the $(p+1)$-vector of coefficient estimates $(\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p)$
- **y** is the *n*-vector of observed responses
- **X** is the matrix (or dataframe) consisting of n rows of observations on p predictors (along with a column of 1's to create an *intercept*).

- In the SLR case with 1 predictor, the estimates $\hat{\beta}_0, \hat{\beta}_1$ describe a line in 2D space:

$$y = \hat{\beta}_0 + \hat{\beta}_1 x$$

Finding Parameters MLR

To create a MLR model, we find the values of $(\beta_0, \beta_1, \ldots, \beta_p)$ that minimizes RSS, where

$$\mathrm{RSS} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}\Big(y_i - (\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p)\Big)^2,$$

which has the solution

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

- $\hat{\beta}$ is the $(p+1)$-vector of coefficient estimates $(\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p)$
- $\mathbf{y}$ is the n-vector of observed responses
- $\mathbf{X}$ is the matrix (or dataframe) consisting of n rows of observations on p predictors (along with a column of 1's to create an *intercept*).

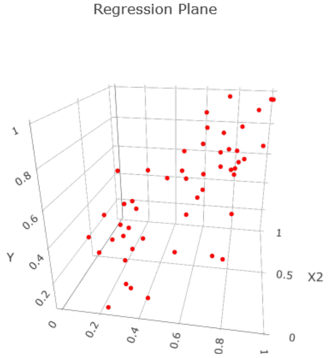
- In the SLR case with 1 predictor, the estimates $\hat{\beta}_0, \hat{\beta}_1$ describe a line in 2D space:

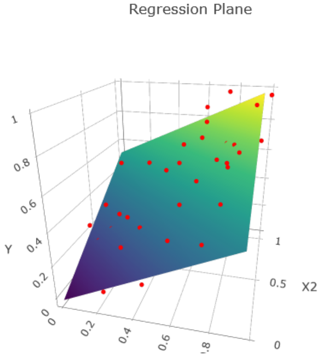
$$y = \hat{\beta}_0 + \hat{\beta}_1 x$$

- But if we have 2 predictors, the estimates $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$ describe a plane in 3D space

$$y = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$$

The Plane of Best Fit





An interactive graphic available under topics list for Tuesday 2-13 on course website

Example: Credit Card Balance

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

Example: Credit Card Balance

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

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

Example: Credit Card Balance

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

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

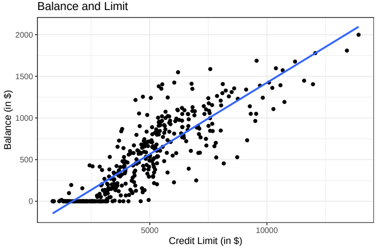
We first consider `Balance` as a function of credit `Limit` and `Income` (in $1000s)

Example: Credit Card Balance

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

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

We first consider Balance as a function of credit Limit and Income (in $1000s)



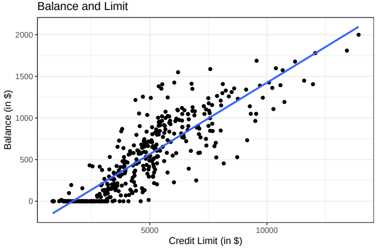
$R = 0.86 \qquad \hat{\text{Balance}} = -292.8 + 0.17 \cdot \text{Limit}$

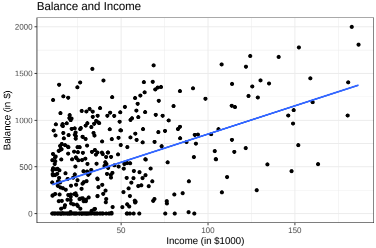
Example: Credit Card Balance

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

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

We first consider `Balance` as a function of credit `Limit` and `Income` (in $1000s)





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

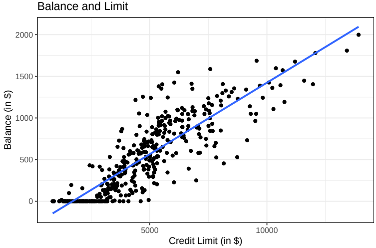
$R = 0.46 \qquad \hat{\text{Balance}} = 246.51 + 6.048 \cdot \text{Income}$

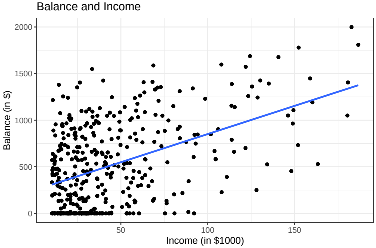
Example: Credit Card Balance

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

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

We first consider `Balance` as a function of credit `Limit` and `Income` (in $1000s)





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

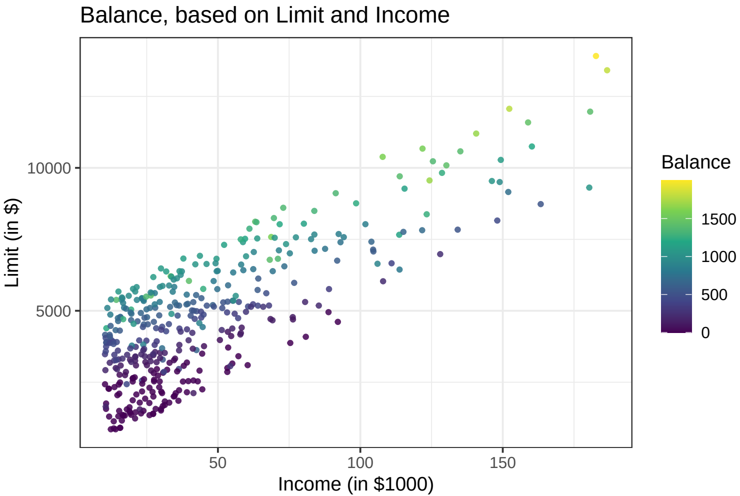
Both variables have some explanatory power for Balance

The Multilinear Visualization

How do Limit and Income *together* explain Balance?

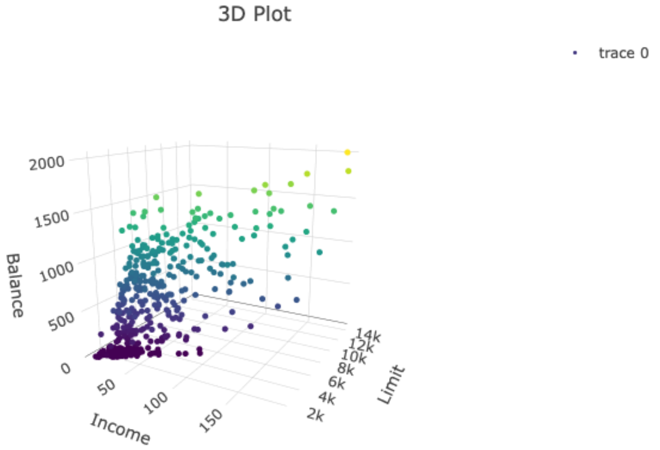
The Multilinear Visualization

How do Limit and Income *together* explain Balance?



The Multilinear Visualization II

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

```
summary(mod)$coefficients
```

```
##                 Estimate   Std. Error   t value      Pr(>|t|)
## (Intercept) -385.1792604 19.464801525 -19.78850  3.878764e-61
## Limit          0.2643216  0.005879729  44.95471 7.717386e-158
## Income        -7.6633230  0.385072058 -19.90101  1.260933e-61
```

Multiple Regression for Debt

Let's find the MLR model

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

And investigate the regression table

```
summary(mod)$coefficients
```

```
##                Estimate   Std. Error   t value      Pr(>|t|)
## (Intercept) -385.1792604 19.464801525 -19.78850  3.878764e-61
## Limit          0.2643216  0.005879729  44.95471 7.717386e-158
## Income        -7.6633230  0.385072058 -19.90101  1.260933e-61
```

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

```
summary(mod)$coefficients
```

```
##                 Estimate   Std. Error   t value      Pr(>|t|)
## (Intercept) -385.1792604 19.464801525 -19.78850  3.878764e-61
## Limit          0.2643216  0.005879729  44.95471 7.717386e-158
## Income        -7.6633230  0.385072058 -19.90101  1.260933e-61
```

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.

Multiple Regression for Debt

Let's find the MLR model

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

And investigate the regression table

```
summary(mod)$coefficients
```

```
##                Estimate   Std. Error   t value      Pr(>|t|)
## (Intercept) -385.1792604 19.464801525 -19.78850  3.878764e-61
## Limit          0.2643216  0.005879729  44.95471 7.717386e-158
## Income        -7.6633230  0.385072058 -19.90101  1.260933e-61
```

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

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

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

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

- 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

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

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

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

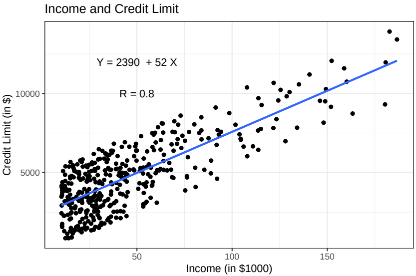
- Not only has MLR given us a new rate of change, but it's completely switched the direction!
- How is this possible?

Income and Credit Limit

Let's consider the relationship between income and credit limit

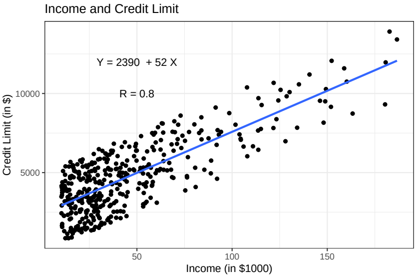
Income and Credit Limit

Let's consider the relationship between income and credit limit



Income and Credit Limit

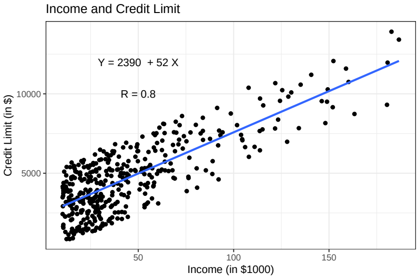
Let's consider the relationship between income and credit limit



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

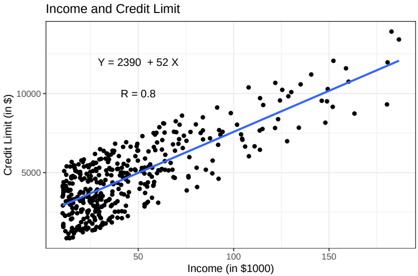


In a vacuum, as income increases, so too does credit limit.

- So in the SLR model, when we assess the change in Balance 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

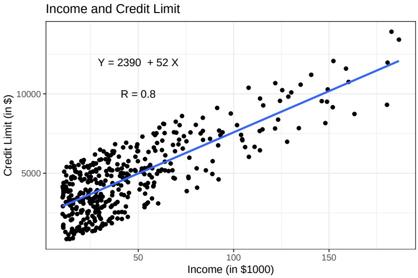


In a vacuum, as income increases, so too does credit limit.

- So in the SLR model, when we assess the change in Balance 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.

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 Balance 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

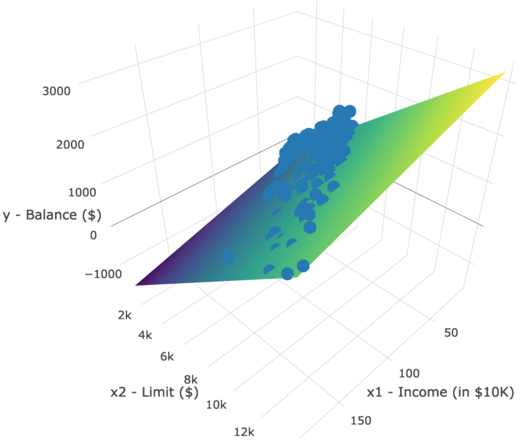
In the MLR model, we may freely change both Income and Credit Limit

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



Balance 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

Balance 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

Section 2

Assessing Model Accuracy

How Strong is a Linear Model?

- In an linear model model, $Y = f(X) + \epsilon$. Even if we could perfectly predict f using $\hat{f}$, our model would still have non-zero MSE.

How Strong is a Linear Model?

- In an linear model model, $Y = f(X) + \epsilon$. Even if we could perfectly predict f using $\hat{f}$, our model would still have non-zero MSE.
- Recall **Residual Standard Error** (RSE) measures the average size of deviations of the response from the linear regression line. It is given by

$$\mathrm{RSE} = \sqrt{\frac{1}{n-1-p}\mathrm{RSS}} = \sqrt{\frac{1}{n-1-p}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

How Strong is a Linear Model?

- In an linear model model, $Y = f(X) + \epsilon$. Even if we could perfectly predict f using $\hat{f}$, our model would still have non-zero MSE.
- Recall **Residual Standard Error** (RSE) measures the average size of deviations of the response from the linear regression line. It is given by

$$\mathrm{RSE} = \sqrt{\frac{1}{n-1-p}\mathrm{RSS}} = \sqrt{\frac{1}{n-1-p}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

- It has the property that

$$\mathrm{Avg}(\mathrm{RSE}^2) = \mathrm{Var}(\epsilon)$$

- Which means that $\mathrm{Abvg}(\mathrm{RSE}) \approx \mathrm{sd}(\epsilon)$

Five Flavors of Error

Which of the following are most likely to **decrease** as more and more predictors are added to a linear model (select all that apply)?

a. test MSE
b. training MSE
c. RSS
d. RSE
e. $\text{Var}(\epsilon)$

The R^2 statistic

Large RSE indicates poor model fit, while small RSE indicates good fit. But how do we determine how small is **small**?

The R^2 statistic

Large RSE indicates poor model fit, while small RSE indicates good fit. But how do we determine how small is **small**?

- The answer depends on the units of Y

The R^2 statistic

Large RSE indicates poor model fit, while small RSE indicates good fit. But how do we determine how small is **small**?

- The answer depends on the units of Y

An alternative, standardized measure of goodness of fit is the coefficient of variation, R^2:

$$R^2 = 1 - \frac{\text{RSS}}{\text{TSS}} \qquad \text{where TSS} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

The R^2 statistic

Large RSE indicates poor model fit, while small RSE indicates good fit. But how do we determine how small is **small**?

- The answer depends on the units of Y

An alternative, standardized measure of goodness of fit is the coefficient of variation, R^2:

$$R^2 = 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}} \qquad \text{where TSS} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

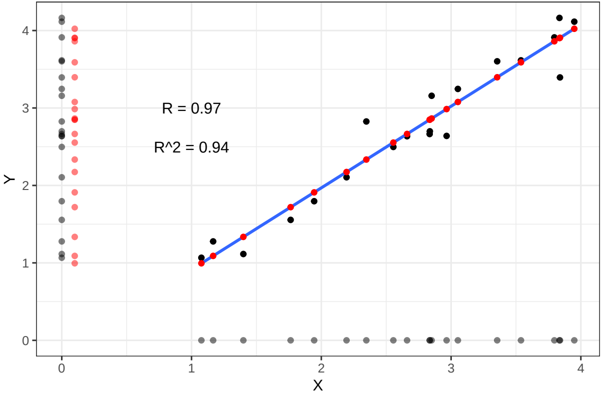
- The value of R^2 is always between 0 and 1, and represents the percentage of variability in values of the response just due to variability in the predictors.

Values of R^2

If $R^2 \approx 1$: nearly all the variability in response is due to variability in the predictor variable.

Values of R^2

If $R^2 \approx 1$: nearly all the variability in response is due to variability in the predictor variable.

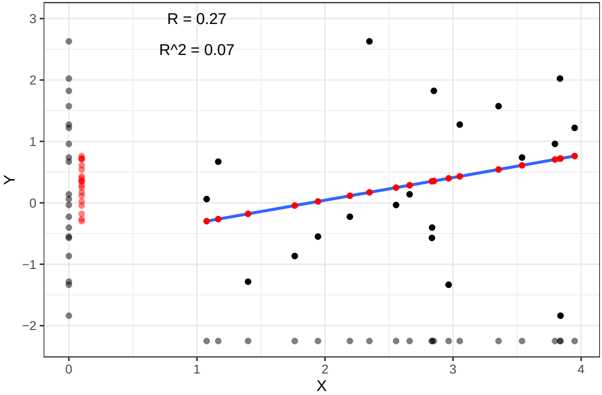


Values of R^2

If $R^2 \approx 0$: almost none of the variability in response is due to variability in the predictor variable.

Values of R^2

If $R^2 \approx 0$: almost none of the variability in response is due to variability in the predictor variable.



Formulas for R^2 in terms of correlation

For SLR,

$$R^2 = [\mathrm{Cor}(X, Y)]^2 = \left[\frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}} \right]^2 = \left[\frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} \right]^2$$

Formulas for R^2 in terms of correlation

For SLR,

$$R^2 = [\mathrm{Cor}(X, Y)]^2 = \left[\frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}}\right]^2 = \left[\frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}\sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}\right]^2$$

For MLR,

$$R^2 = \left[\mathrm{Cor}(Y, \hat{Y})\right]^2$$

Formulas for R^2 in terms of correlation

For SLR,

$$R^2 = [\mathrm{Cor}(X, Y)]^2 = \left[\frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}}\right]^2 = \left[\frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}\right]^2$$

For MLR,

$$R^2 = \left[\mathrm{Cor}(Y, \hat{Y})\right]^2$$

We will usually use software to compute R^2.

Model Accuracy in R

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

summary(mod_credit)
```

```
##
## Call:
## lm(formula = Balance ~ Income + Limit, data = Credit)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -232.79 -115.45  -48.20   53.36  549.77
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept) -385.17926   19.46480  -19.79   <2e-16 ***
## Income        -7.66332    0.38507  -19.90   <2e-16 ***
## Limit          0.26432    0.00588   44.95   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 165.5 on 397 degrees of freedom
## Multiple R-squared:  0.8711, Adjusted R-squared:  0.8705
## F-statistic:  1342 on 2 and 397 DF,  p-value: < 2.2e-16
```

Model Accuracy in R

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

summary(mod_credit)
```

```
##
## Call:
## lm(formula = Balance ~ Income + Limit, data = Credit)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -232.79 -115.45  -48.20   53.36  549.77
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept) -385.17926   19.46480  -19.79   <2e-16 ***
## Income        -7.66332    0.38507  -19.90   <2e-16 ***
## Limit          0.26432    0.00588   44.95   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 165.5 on 397 degrees of freedom
## Multiple R-squared:  0.8711, Adjusted R-squared:  0.8705
## F-statistic:  1342 on 2 and 397 DF,  p-value: < 2.2e-16
```

We can use `summary(mod)$r.sq` or `summary(mod)$sigma` to access R^2 and RSE directly.

Adjusted R^2

- It turns out that the samples's R^2 gives a **biased** estimate of the variability in the *population* explained by the model.

Adjusted R^2

- It turns out that the samples's R^2 gives a **biased** estimate of the variability in the *population* explained by the model.
- Instead, we use the adjusted R:

$$R^2_{\text{adjusted}} = 1 - \frac{\text{RSS}}{\text{TSS}} \frac{n-1}{n-p-1}$$

Adjusted R^2

- It turns out that the samples's R^2 gives a **biased** estimate of the variability in the *population* explained by the model.
- Instead, we use the adjusted R:

$$R^2_{\text{adjusted}} = 1 - \frac{\text{RSS}}{\text{TSS}} \frac{n-1}{n-p-1}$$

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

Adjusted R^2

- It turns out that the samples's R^2 gives a **biased** estimate of the variability in the *population* explained by the model.
- Instead, we use the adjusted R:

$$R^2_{\text{adjusted}} = 1 - \frac{\text{RSS}}{\text{TSS}} \frac{n-1}{n-p-1}$$

- This adjusted R^2 is usually a bit smaller than R^2, and the difference decreases as n gets large.
- While *adjusted* R^2 is better metric for evaluating models than R^2, it still gives an overly optimistic estimate of error. Why?

Adjusted R^2

- It turns out that the samples's R^2 gives a **biased** estimate of the variability in the *population* explained by the model.
- Instead, we use the adjusted R:

$$R^2_{\text{adjusted}} = 1 - \frac{\text{RSS}}{\text{TSS}} \frac{n-1}{n-p-1}$$

- This adjusted R^2 is usually a bit smaller than R^2, and the difference decreases as n gets large.
- While *adjusted* R^2 is better metric for evaluating models than R^2, it still gives an overly optimistic estimate of error. Why?
  - The formula for adjusted R^2 depends on RSS, which is computed on **training** data.

The Hypothesis Test (optional)

- **Goal**: test whether any predictors have linear relationship with response.

The Hypothesis Test (optional)

- **Goal**: test whether any predictors have linear relationship with response.
  - It would be fallacious to conduct p many significant tests comparing each predictor to the response. (Why?)

The Hypothesis Test (optional)

- **Goal**: test whether any predictors have linear relationship with response.
  - It would be fallacious to conduct p many significant tests comparing each predictor to the response. (Why?)

Hypotheses:

$$H_0 : \beta_1 = \cdots = \beta_p = 0 \qquad H_a : \text{ at least one of } \beta_i \neq 0$$

The Hypothesis Test (optional)

- **Goal**: test whether any predictors have linear relationship with response.
    - It would be fallacious to conduct p many significant tests comparing each predictor to the response. (Why?)

Hypotheses:

$$H_0 : \beta_1 = \cdots = \beta_p = 0 \qquad H_a : \text{ at least one of } \beta_i \neq 0$$

Test statistic:

$$F = \frac{(\text{TSS} - \text{RSS})/p}{\text{RSS}/(n - p - 1)}$$

The Hypothesis Test (optional)

- **Goal**: test whether any predictors have linear relationship with response.
  - It would be fallacious to conduct p many significant tests comparing each predictor to the response. (Why?)

Hypotheses:

$$H_0 : \beta_1 = \cdots = \beta_p = 0 \qquad H_a : \text{ at least one of } \beta_i \neq 0$$

Test statistic:

$$F = \frac{(\mathrm{TSS} - \mathrm{RSS})/p}{\mathrm{RSS}/(n - p - 1)}$$

Under H_0, the stat F has F-distributed with $(p, n - p - 1)$ degrees of freedom

The Hypothesis Test (optional)

- **Goal**: test whether any predictors have linear relationship with response.
  - It would be fallacious to conduct p many significant tests comparing each predictor to the response. (Why?)

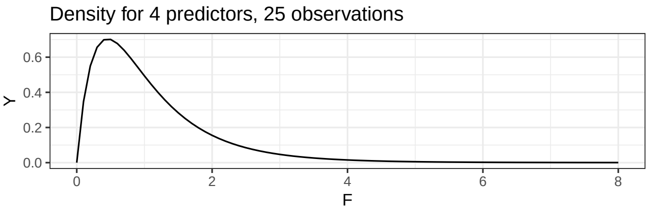
Hypotheses:

$$H_0 : \beta_1 = \cdots = \beta_p = 0 \qquad H_a : \text{ at least one of } \beta_i \neq 0$$

Test statistic:

$$F = \frac{(\text{TSS} - \text{RSS})/p}{\text{RSS}/(n - p - 1)}$$

Under H_0, the stat F has F-distributed with $(p, n - p - 1)$ degrees of freedom



Typical Values of the F statistic (optional)

Provided conditions for linear regression are met,

$$\text{Avg}\left[\frac{\text{RSS}}{n-p-1}\right] = \sigma^2 = \text{Var}(\epsilon)$$

Typical Values of the F statistic (optional)

Provided conditions for linear regression are met,

$$\text{Avg}\left[\frac{\text{RSS}}{n-p-1}\right] = \sigma^2 = \text{Var}(\epsilon)$$

And if H_0 is also true, then

$$\text{Avg}\left[\frac{\text{TSS}-\text{RSS}}{p}\right] = \sigma^2 = \text{Var}(\epsilon)$$

Typical Values of the F statistic (optional)

Provided conditions for linear regression are met,

$$\text{Avg}\left[\frac{\text{RSS}}{n-p-1}\right] = \sigma^2 = \text{Var}(\epsilon)$$

And if H_0 is also true, then

$$\text{Avg}\left[\frac{\text{TSS}-\text{RSS}}{p}\right] = \sigma^2 = \text{Var}(\epsilon)$$

Hence, if there is truly no relationship between any of the predictors and the response, the the average value of

$$F = \frac{(\text{TSS}-\text{RSS})/p}{\text{RSS}/(n-p-1)}$$

is 1.

Typical Values of the F statistic (optional)

Provided conditions for linear regression are met,

$$\mathrm{Avg}\left[\frac{\mathrm{RSS}}{n-p-1}\right] = \sigma^2 = \mathrm{Var}(\epsilon)$$

And if H_0 is also true, then

$$\mathrm{Avg}\left[\frac{\mathrm{TSS}-\mathrm{RSS}}{p}\right] = \sigma^2 = \mathrm{Var}(\epsilon)$$

Hence, if there is truly no relationship between any of the predictors and the response, the the average value of

$$F = \frac{(\mathrm{TSS}-\mathrm{RSS})/p}{\mathrm{RSS}/(n-p-1)}$$

is 1.

Moreover, it is unlikely that F is drastically larger than 1.

Poll: TSS and RSS (optional)

Suppose we have a linear model with 25 observations and 4 predictors. Which of the following provides the best evidence of a relationship between the response and at least 1 of the predictors?

a. $\text{TSS} = 64$, $\text{RSS} = 4$
b. $\text{TSS} = 4$, $\text{RSS} = 16$
c. $\text{TSS} = 48$, $\text{RSS} = 8$
d. $\text{TSS} = 4$, $\text{RSS} = 4$

Improving Model Accuracy

What do we do when the full model is inaccurate on test data (i.e. high test MSE)?

Improving Model Accuracy

What do we do when the full model is inaccurate on test data (i.e. high test MSE)?

- Remove redundant variables strongly correlated with each other
- Remove spurious variables that aren't linearly related with response

Improving Model Accuracy

What do we do when the full model is inaccurate on test data (i.e. high test MSE)?

- Remove redundant variables strongly correlated with each other
- Remove spurious variables that aren't linearly related with response

Two common methods of identifying extraneous predictors:

1. **Backwards Elimination**:
    - Start with the full model, remove the variable with highest p-value, and refit. Continue to do so until all p-values are small, or accuracy ceases to improve.

Improving Model Accuracy

What do we do when the full model is inaccurate on test data (i.e. high test MSE)?

- Remove redundant variables strongly correlated with each other
- Remove spurious variables that aren't linearly related with response

Two common methods of identifying extraneous predictors:

1. **Backwards Elimination**:
   - Start with the full model, remove the variable with highest p-value, and refit. Continue to do so until all p-values are small, or accuracy ceases to improve.
2. **Forward Selection**:
   - Start with the null model, create p many SLR models (one for each predictor), and select the one with best accuracy. Repeat with this new model, creating $p-1$ two predictor models (one for each remaining predictor). Continue until accuracy ceases to improve.

Improving Model Accuracy

What do we do when the full model is inaccurate on test data (i.e. high test MSE)?

- Remove redundant variables strongly correlated with each other
- Remove spurious variables that aren't linearly related with response

Two common methods of identifying extraneous predictors:

1. **Backwards Elimination**:
   - Start with the full model, remove the variable with highest p-value, and refit. Continue to do so until all p-values are small, or accuracy ceases to improve.
2. **Forward Selection**:
   - Start with the null model, create p many SLR models (one for each predictor), and select the one with best accuracy. Repeat with this new model, creating $p - 1$ two predictor models (one for each remaining predictor). Continue until accuracy ceases to improve.

- Is it possible that none of these models will have the best possible accuracy among all subsets of predictors?

Improving Model Accuracy

What do we do when the full model is inaccurate on test data (i.e. high test MSE)?

- Remove redundant variables strongly correlated with each other
- Remove spurious variables that aren't linearly related with response

Two common methods of identifying extraneous predictors:

1. **Backwards Elimination**:
    - Start with the full model, remove the variable with highest p-value, and refit. Continue to do so until all p-values are small, or accuracy ceases to improve.
2. **Forward Selection**:
    - Start with the null model, create p many SLR models (one for each predictor), and select the one with best accuracy. Repeat with this new model, creating $p - 1$ two predictor models (one for each remaining predictor). Continue until accuracy ceases to improve.

- Is it possible that none of these models will have the best possible accuracy among all subsets of predictors?
    - Yes. But we'll cover detailed model selection in Chapter 6.