Multiple Linear Regression

Kyle M. Lang

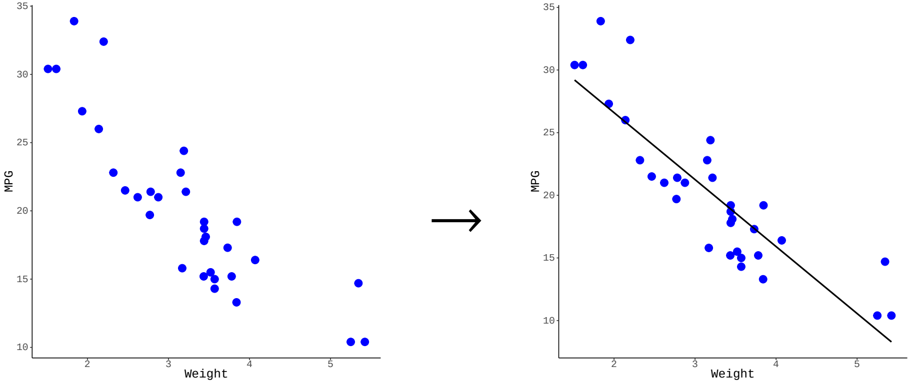
Department of Methodology & Statistics
Utrecht University

Outline

Graphical Representations of Regression Models

A regression of two variables can be represented on a 2D scatterplot.

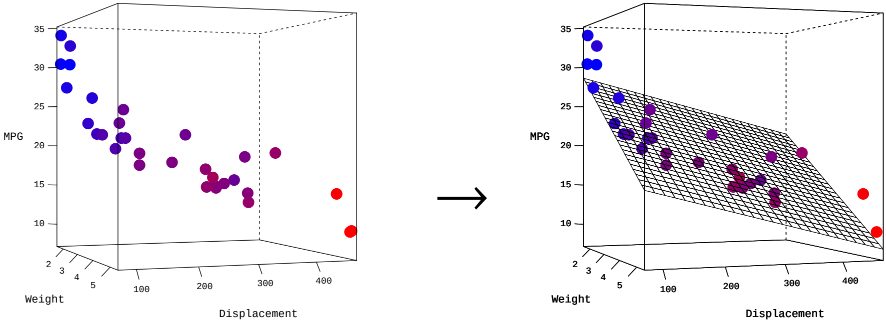
- Simple linear regression implies a 1D line in 2D space.



Graphical Representations of Regression Models

Adding an additional predictor leads to a 3D point cloud.

- A regression model with two IVs implies a 2D plane in 3D space.



Partial Effects

In MLR, we want to examine the *partial effects* of the predictors.

- What is the effect of a predictor after controlling for some other set of variables?

This approach is crucial to controlling confounds and adequately modeling real-world phenomena.

Example

```r
## Read in the 'diabetes' dataset:
dDat <- readRDS("../data/diabetes.rds")

## Simple regression with which we're familiar:
out1 <- lm(bp ~ age, data = dDat)
```

Asking: What is the effect of age on average blood pressure?

Example

```r
partSummary(out1, -1)

## Residuals:
##     Min      1Q  Median      3Q     Max
## -31.188  -8.897  -1.209   8.612  39.952
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 77.47605    2.38132  32.535  < 2e-16
## age          0.35391    0.04739   7.469 4.39e-13
##
## Residual standard error: 13.04 on 440 degrees of freedom
## Multiple R-squared:  0.1125,Adjusted R-squared:  0.1105
## F-statistic: 55.78 on 1 and 440 DF,  p-value: 4.393e-13
```

Example

```r
## Add in another predictor:
out2 <- lm(bp ~ age + bmi, data = dDat)
```

Asking: What is the effect of BMI on average blood pressure, *after controlling for age?*

- We're partialing age out of the effect of BMI on blood pressure.

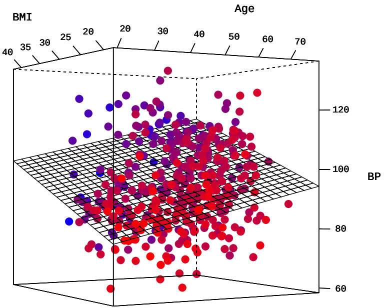
Example

```
partSummary(out2, -1)

## Residuals:
##     Min      1Q  Median      3Q     Max
## -29.287  -8.198  -0.178   8.413  41.026
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 52.24654    3.83168  13.635  < 2e-16
## age          0.28651    0.04504   6.362 5.02e-10
## bmi          1.08053    0.13363   8.086 6.06e-15
##
## Residual standard error: 12.18 on 439 degrees of freedom
## Multiple R-squared:  0.2276,Adjusted R-squared:  0.224
## F-statistic: 64.66 on 2 and 439 DF,  p-value: < 2.2e-16
```

Interpretation

- The expected average blood pressure for an unborn patient with a negligible extent is 52.25.
- For each year older, average blood pressure is expected to increase by 0.29 points, after controlling for BMI.
- For each additional point of BMI, average blood pressure is expected to increase by 1.08 points, after controlling for age.



Multiple R^2

How much variation in blood pressure is explained by the two models?

- Check the R^2 values.

```
## Extract R^2 values:
r2.1 <- summary(out1)$r.squared
r2.2 <- summary(out2)$r.squared

r2.1

## [1] 0.1125117

r2.2

## [1] 0.2275606
```

F-Statistic

How do we know if the R^2 values are significantly greater than zero?

- We use the F-statistic to test $H_0 : R^2 = 0$ vs. $H_1 : R^2 > 0$.

```
f1 <- summary(out1)$fstatistic
f1

##     value     numdf     dendf
##  55.78116   1.00000 440.00000

pf(q = f1[1], df1 = f1[2], df2 = f1[3], lower.tail = FALSE)

##        value
## 4.392569e-13
```

F-Statistic

```
f2 <- summary(out2)$fstatistic
f2

##    value    numdf    dendf
##  64.6647   2.0000 439.0000

pf(f2[1], f2[2], f2[3], lower.tail = FALSE)

##        value
## 2.433518e-25
```

Comparing Models

How do we quantify the additional variation explained by BMI, above and beyond age?

- Compute the ΔR^2

```
## Compute change in R^2:
r2.2 - r2.1

## [1] 0.115049
```

Significance Testing

How do we know if ΔR^2 represents a significantly greater degree of explained variation?

- Use an F-test for $H_0 : \Delta R^2 = 0$ vs. $H_1 : \Delta R^2 > 0$

```
## Is that increase significantly greater than zero?
anova(out1, out2)

## Analysis of Variance Table
##
## Model 1: bp ~ age
## Model 2: bp ~ age + bmi
##   Res.Df   RSS Df Sum of Sq      F    Pr(>F)
## 1    440 74873
## 2    439 65167  1    9706.1 65.386 6.057e-15 ***
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Model Comparison

We can also compare models based on their prediction errors.

- For OLS regression, we usually compare MSE values.

```
mse1 <- MSE(y_pred = predict(out1), y_true = dDat$bp)
mse2 <- MSE(y_pred = predict(out2), y_true = dDat$bp)

mse1

## [1] 169.3963

mse2

## [1] 147.4367
```

In this case, the MSE for the model with BMI included is smaller.

- We should prefer the the larger model.

CATEGORICAL PREDICTORS

Categorical Predictors

Most of the predictors we've considered thus far have been *quantitative*.

- Continuous variables that can take any real value in their range
- Interval or Ratio scaling

We often want to include grouping factors as predictors.

- These variables are *qualitative*.
  - Their values are simply labels.
  - There is no ordering of the categories.
  - Nominal scaling

How to Model Categorical Predictors

We need to be careful when we include categorical predictors into a regression model.

- The variables need to be coded before entering the model

Consider the following indicator of major:
$X_{maj} = \{1 = \textit{Law}, 2 = \textit{Economics}, 3 = \textit{Data Science}\}$

- What would happen if we naïvely used this variable to predict program satisfaction?

How to Model Categorical Predictors

```
mDat <- readRDS("../data/major_data.rds")
mDat[seq(25, 150, 25), ]

##     sat majF majN
## 25  1.9  law    1
## 50  1.4  law    1
## 75  4.3 econ    2
## 100 4.1 econ    2
## 125 5.7   ds    3
## 150 5.1   ds    3

out1 <- lm(sat ~ majN, data = mDat)
```

How to Model Categorical Predictors

```
partSummary(out1, -1)

## Residuals:
##    Min     1Q Median     3Q    Max
## -1.303 -0.313 -0.113  0.342  1.342
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.33200    0.12060  -2.753  0.00664
## majN         2.04500    0.05582  36.632  < 2e-16
##
## Residual standard error: 0.5582 on 148 degrees of freedom
## Multiple R-squared:  0.9007,Adjusted R-squared:    0.9
## F-statistic:  1342 on 1 and 148 DF,  p-value: < 2.2e-16
```

Dummy Coding

The most common way to code categorical predictors is *dummy coding*.

- A G-level factor must be converted into a set of $G - 1$ dummy codes.
- Each code is a variable on the dataset that equals 1 for observations corresponding to the code's group and equals 0, otherwise.
- The group without a code is called the *reference group*.

Example Dummy Code

Let's look at the simple example of coding biological sex:

	sex	male
1	female	0
2	male	1
3	male	1
4	female	0
5	male	1
6	female	0
7	female	0
8	male	1
9	female	0
10	female	0

Example Dummy Codes

Now, a slightly more complex example:

	drink	juice	tea
1	juice	1	0
2	coffee	0	0
3	tea	0	1
4	tea	0	1
5	tea	0	1
6	tea	0	1
7	juice	1	0
8	tea	0	1
9	coffee	0	0
10	juice	1	0

Using Dummy Codes

To use the dummy codes, we simply include the $G - 1$ codes as $G - 1$ predictor variables in our regression model.

$$Y = \beta_0 + \beta_1 X_{male} + \varepsilon$$
$$Y = \beta_0 + \beta_1 X_{juice} + \beta_2 X_{tea} + \varepsilon$$

- The intercept corresponds to the mean of Y for the reference group.
- Each slope represents the difference between the mean of Y in the coded group and the mean of Y in the reference group.

Example

First, an example with a single, binary dummy code:

```r
## Read in some data:
cDat <- readRDS("../data/cars_data.rds")

## Fit and summarize the model:
out2 <- lm(price ~ mtOpt, data = cDat)
```

Example

```
partSummary(out2, -1)

## Residuals:
##     Min      1Q  Median      3Q     Max
## -10.341  -6.338  -3.141   2.662  38.059
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   23.841      1.623  14.691   <2e-16
## mtOpt         -6.603      2.004  -3.295   0.0014
##
## Residual standard error: 9.18 on 91 degrees of freedom
## Multiple R-squared:  0.1066,Adjusted R-squared:  0.09679
## F-statistic: 10.86 on 1 and 91 DF,  p-value: 0.001403
```

Interpretations

- The average price of a car without the option for a manual transmission is $\hat{\beta}_0 = 23.84$ thousand dollars.
- The average difference in price between cars that have manual transmissions as an option and those that do not is $\hat{\beta}_1 = -6.6$ thousand dollars.

Example

Fit a more complex model:

```
out3 <- lm(price ~ front + rear, data = cDat)
partSummary(out3, -1)

## Residuals:
##     Min      1Q  Median      3Q     Max
## -14.050  -6.250  -1.236   3.264  32.950
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 17.63000    2.76119   6.385 7.33e-09
## front       -0.09418    2.96008  -0.032  0.97469
## rear        11.32000    3.51984   3.216  0.00181
##
## Residual standard error: 8.732 on 90 degrees of freedom
## Multiple R-squared:  0.2006,Adjusted R-squared:  0.1829
## F-statistic: 11.29 on 2 and 90 DF,  p-value: 4.202e-05
```

Interpretations

- The average price of a four-wheel-drive car is $\hat{\beta}_0 = 17.63$ thousand dollars.
- The average difference in price between front-wheel-drive cars and four-wheel-drive cars is $\hat{\beta}_1 = -0.09$ thousand dollars.
- The average difference in price between rear-wheel-drive cars and four-wheel-drive cars is $\hat{\beta}_2 = 11.32$ thousand dollars.

Example

Include two sets of dummy codes:

```
out4 <- lm(price ~ mtOpt + front + rear, data = cDat)
partSummary(out4, -c(1, 2))

## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  21.7187     2.9222   7.432 6.25e-11
## mtOpt        -5.8410     1.8223  -3.205  0.00187
## front        -0.2598     2.8189  -0.092  0.92677
## rear         10.5169     3.3608   3.129  0.00237
##
## Residual standard error: 8.314 on 89 degrees of freedom
## Multiple R-squared:  0.2834,Adjusted R-squared:  0.2592
## F-statistic: 11.73 on 3 and 89 DF,  p-value: 1.51e-06
```

Interpretations

- The average price of a four-wheel-drive car that does not have a manual transmission option is $\hat{\beta}_0 = 21.72$ thousand dollars.
- After controlling for drive type, the average difference in price between cars that have manual transmissions as an option and those that do not is $\hat{\beta}_1 = -5.84$ thousand dollars.
- After controlling for transmission options, the average difference in price between front-wheel-drive cars and four-wheel-drive cars is $\hat{\beta}_2 = -0.26$ thousand dollars.
- After controlling for transmission options, the average difference in price between rear-wheel-drive cars and four-wheel-drive cars is $\hat{\beta}_3 = 10.52$ thousand dollars.

Significance Testing

For variables with only two levels, we can test the overall factor's significance by evaluating the significance of a single dummy code.

```
partSummary(out2, -c(1, 2))

## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   23.841      1.623  14.691   <2e-16
## mtOpt         -6.603      2.004  -3.295   0.0014
##
## Residual standard error: 9.18 on 91 degrees of freedom
## Multiple R-squared:  0.1066,Adjusted R-squared:  0.09679
## F-statistic: 10.86 on 1 and 91 DF,  p-value: 0.001403
```

Significance Testing

For variables with more than two levels, we need to simultaneously evaluate the significance of each of the variable's dummy codes.

```
partSummary(out4, -c(1, 2))

## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   21.7187     2.9222   7.432 6.25e-11
## mtOpt         -5.8410     1.8223  -3.205  0.00187
## front         -0.2598     2.8189  -0.092  0.92677
## rear          10.5169     3.3608   3.129  0.00237
##
## Residual standard error: 8.314 on 89 degrees of freedom
## Multiple R-squared:  0.2834,Adjusted R-squared:  0.2592
## F-statistic: 11.73 on 3 and 89 DF,  p-value: 1.51e-06
```

Significance Testing

```
summary(out4)$r.squared - summary(out2)$r.squared

## [1] 0.1767569

anova(out2, out4)

## Analysis of Variance Table
##
## Model 1: price ~ mtOpt
## Model 2: price ~ mtOpt + front + rear
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     91 7668.9
## 2     89 6151.6  2    1517.3 10.976 5.488e-05 ***
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Significance Testing

What about models where a single nominal factor is the only predictor?

```
partSummary(out3, -c(1, 2))

## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 17.63000    2.76119   6.385 7.33e-09
## front       -0.09418    2.96008  -0.032  0.97469
## rear        11.32000    3.51984   3.216  0.00181
##
## Residual standard error: 8.732 on 90 degrees of freedom
## Multiple R-squared:  0.2006,Adjusted R-squared:  0.1829
## F-statistic: 11.29 on 2 and 90 DF,  p-value: 4.202e-05
```

Significance Testing

We can compare back to an "intercept-only" model.

```
out0 <- lm(price ~ 1, data = cDat)
partSummary(out0, -1)

## Residuals:
##     Min      1Q  Median      3Q     Max
## -12.11   -7.31   -1.81    3.79   42.39
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   19.510      1.002   19.48   <2e-16
##
## Residual standard error: 9.659 on 92 degrees of freedom
```

Significance Testing

```
r2Diff <- summary(out3)$r.squared - summary(out0)$r.squared
r2Diff

## [1] 0.2006386

anova(out0, out3)

## Analysis of Variance Table
##
## Model 1: price ~ 1
## Model 2: price ~ front + rear
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     92 8584.0
## 2     90 6861.7  2    1722.3 11.295 4.202e-05 ***
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Significance Testing

We don't actually need to do the explicit model comparison, though.

```
r2Diff

## [1] 0.2006386

summary(out3)$r.squared

## [1] 0.2006386

anova(out0, out3)[2, "F"]

## [1] 11.29494

summary(out3)$fstatistic[1]

##    value
## 11.29494
```

Conclusion

- Each variable in a regression model corresponds to a dimension in the data-space.
  - A regression model with P predictors implies a P-dimensional (hyper)-plane in (P + 1)-dimensional space.
- The coefficients in MLR are partial coefficients.
  - Each effect is interpreted as holding other predictors constant.
- Categorical predictors must be coded before they can be used in our models.
  - The regression coefficients represent group mean differences.