Introduction to Multiple Regression

Math 430, Winter 2017

The data

- In an effort to understand the important aspects of a satisfactory supervisor, clerical employees at a large financial organization company were asked to rate their immediate supervisor.
- The survey questions were designed to measure overall performance of the supervisor, as well as six additional characteristics.
- Employees were asked to rate the following statements on a scale from 0 to 100 (0 meaning "completely disagree" to 100 meaning "completely agree")

The data

Variable	Description
`rating`	Overall rating of supervisor performance
`complaints`	Score for “Your supervisor handles employee complaints appropriately.”
`privileges`	Score for “Your supervisor allows special privileges.”
`learn`	Score for “Your supervisor provides opportunities to learn new things.”
`raises`	Score for “Your supervisor bases raises on performance.”
`critical`	Score for “Your supervisor is too critical of poor performance.”
`advance`	Score for “I am not satisfied with the rate I am advancing in the company?”

The data

```
supervisor <- read.table("https://github.com/math430-lu/data/raw/master/supervisor.txt",
                         header = TRUE)
head(supervisor)
```

```
##   overall complaints privileges learn raises critical advance
## 1      43         51         30    39     61       92      45
## 2      63         64         51    54     63       73      47
## 3      71         70         68    69     76       86      48
## 4      61         63         45    47     54       84      35
## 5      81         78         56    66     71       83      47
## 6      43         55         49    44     54       49      34
```

Problem overview

Primary research question

- What makes a good (or bad) supervisor?

Analysis

- What is the response variable?
- What should we use for the predictor?

Multiple Regression

The model

- Model

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ip} + e_i, \quad e_i \overset{iid}{\sim} \mathcal{N}(0, \sigma_e^2)$$

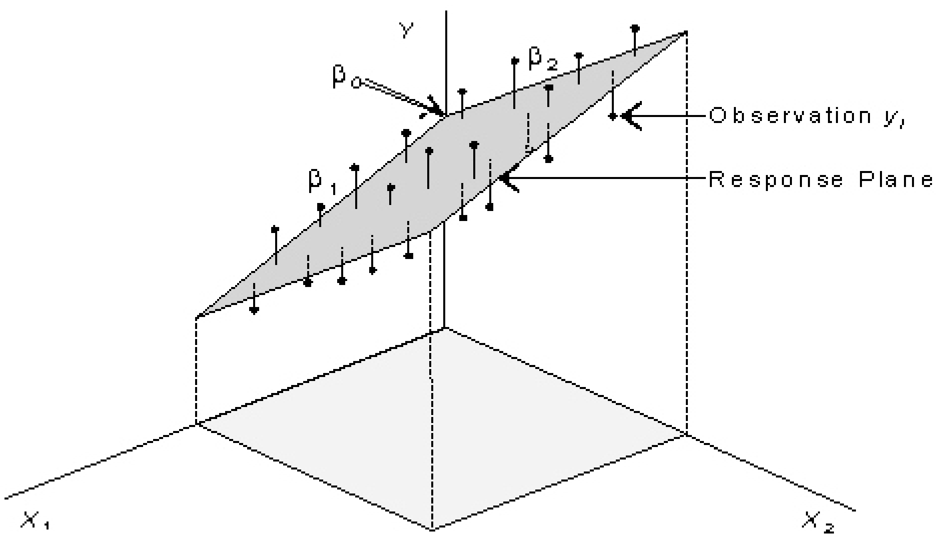
- In matrix form

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$$

- Assumptions – same as in SLR

The geometry

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ip} + e_i$$



Interpreting β_0

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ip} + e_i, \quad e_i \overset{iid}{\sim} \mathcal{N}(0, \sigma_e^2)$$

Uncentered model:

Centered model:

Interpreting β_k

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ip} + e_i, \quad e_i \overset{iid}{\sim} \mathcal{N}(0, \sigma_e^2)$$

Interpreting e_i

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ip} + e_i, \quad e_i \overset{iid}{\sim} \mathcal{N}(0, \sigma_e^2)$$

- Same as before: the distance an observation falls from the "line"
- Represents random error (that we can't model)

Interpreting σ_e^2

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ip} + e_i, \quad e_i \overset{iid}{\sim} \mathcal{N}(0, \sigma_e^2)$$

- Same as before: the typical (average) distance an observation falls from the "line"

Fitting the model

Target:

- Prediction equation

$$\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_{i1} + \widehat{\beta}_2 x_{i2} + \cdots + \widehat{\beta}_k x_{ip}$$

i.e.

$$\widehat{\mathbf{Y}} = \mathbf{X}\widehat{\boldsymbol{\beta}}$$

- Standard error for MLR model: $\widehat{\sigma}_e$

Fitting the model

Procedure:

- Least squares estimation: choose the coefficients to minimize

$$\text{SSE} = \sum_i \left(Y_i - \widehat{Y}_i\right)^2$$

- $\widehat{\boldsymbol{\beta}} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{Y}$
- Use $s^2 = \sqrt{\dfrac{\text{RSS}}{n-p-1}}$

Using R

```
super.lm <- lm(overall ~ complaints + privileges + learn + raises + critical + advance,
               data = supervisor)
summary(super.lm)
```

```
##
## Call:
## lm(formula = overall ~ complaints + privileges + learn + raises +
##     critical + advance, data = supervisor)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.9418  -4.3555   0.3158   5.5425  11.5990
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 10.78708   11.58926   0.931 0.361634
## complaints   0.61319    0.16098   3.809 0.000903 ***
## privileges  -0.07305    0.13572  -0.538 0.595594
## learn        0.32033    0.16852   1.901 0.069925 .
## raises       0.08173    0.22148   0.369 0.715480
## critical     0.03838    0.14700   0.261 0.796334
## advance     -0.21706    0.17821  -1.218 0.235577
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.068 on 23 degrees of freedom
## Multiple R-squared:  0.7326, Adjusted R-squared:  0.6628
## F-statistic:  10.5 on 6 and 23 DF,  p-value: 1.24e-05
```

Using R

```
head(model.matrix(super.lm))
```

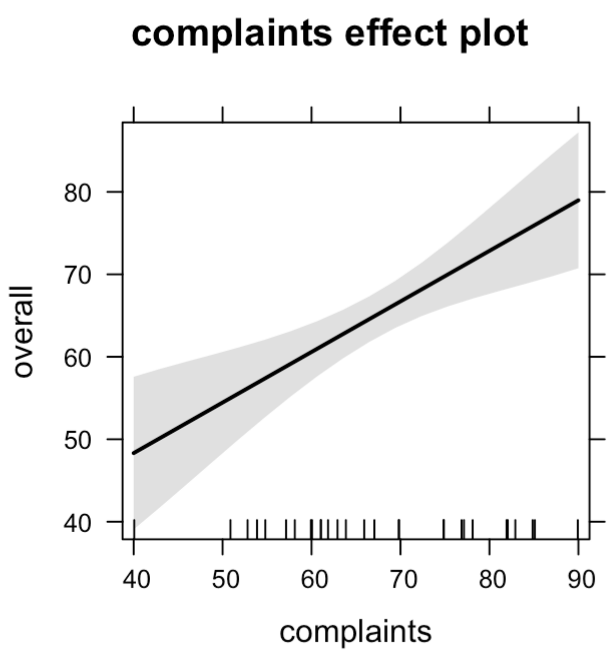
```
##   (Intercept) complaints privileges learn raises critical advance
## 1           1         51         30    39     61       92      45
## 2           1         64         51    54     63       73      47
## 3           1         70         68    69     76       86      48
## 4           1         63         45    47     54       84      35
## 5           1         78         56    66     71       83      47
## 6           1         55         49    44     54       49      34
```

Interpreting β_k

Effects plots

- Idea: visualize the effect of a predictor by fixing the other predictors at their sample mean (i.e. $\bar{x}_k$ values)

```
library(effects)
plot(Effect("complaints", super.lm))
```



Inference

Review: The ANOVA identity

$$\text{Total sum of squares (SST)} = \sum_{i=1}^{n} \left(y_i - \overline{y}\right)^2$$

$$\text{Sum of squares error (RSS)} = \sum_{i=1}^{n} \left(y_i - \widehat{y}_i\right)^2$$

$$\text{Sum of squares due to model (SSreg)} = \sum_{i=1}^{n} \left(\widehat{y}_i - \overline{y}\right)^2$$

ANOVA tables

Review: R^2

Coefficient of **Multiple** Determination

$$R^2 = \frac{\text{SSModel}}{\text{SSTotal}} = 1 - \frac{\text{SSE}}{\text{SSTotal}}$$

Inference for all coefficients

Testing $\mathbf{H_0} : \boldsymbol{\beta_1} = \boldsymbol{\beta_2} = \cdots = \boldsymbol{\beta_k} = \mathbf{0}$

Inference for all coefficients

```
summary(super.lm)
```

```
##
## Call:
## lm(formula = overall ~ complaints + privileges + learn + raises +
##     critical + advance, data = supervisor)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.9418  -4.3555   0.3158   5.5425  11.5990
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 10.78708   11.58926   0.931 0.361634
## complaints   0.61319    0.16098   3.809 0.000903 ***
## privileges  -0.07305    0.13572  -0.538 0.595594
## learn        0.32033    0.16852   1.901 0.069925 .
## raises       0.08173    0.22148   0.369 0.715480
## critical     0.03838    0.14700   0.261 0.796334
## advance     -0.21706    0.17821  -1.218 0.235577
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.068 on 23 degrees of freedom
## Multiple R-squared:  0.7326, Adjusted R-squared:  0.6628
## F-statistic:  10.5 on 6 and 23 DF,  p-value: 1.24e-05
```

Inference for all coefficients

```
null.lm <- lm(overall ~ 1, data = supervisor)
anova(null.lm, super.lm)
```

```
## Analysis of Variance Table
##
## Model 1: overall ~ 1
## Model 2: overall ~ complaints + privileges + learn + raises + critical +
##     advance
##   Res.Df  RSS Df Sum of Sq      F   Pr(>F)
## 1     29 4297
## 2     23 1149  6      3148 10.502 1.24e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
1 - pf(10.502, df1 = 23, df2 = 29)
```

```
## [1] 1.113971e-08
```

Inference for a single coefficient

Testing $\mathbf{H_0}$: $\beta_j = \beta_j^0$

Inference for a single coefficient

```
library(broom)
tidy(super.lm)
```

```
##          term    estimate  std.error  statistic      p.value
## 1 (Intercept) 10.78707639 11.5892572  0.9307824 0.3616337210
## 2  complaints  0.61318761  0.1609831  3.8090182 0.0009028679
## 3  privileges -0.07305014  0.1357247 -0.5382229 0.5955939205
## 4       learn  0.32033212  0.1685203  1.9008516 0.0699253459
## 5      raises  0.08173213  0.2214777  0.3690310 0.7154800884
## 6    critical  0.03838145  0.1469954  0.2611064 0.7963342642
## 7     advance -0.21705668  0.1782095 -1.2179862 0.2355770486
```

Inference for a single coefficient

Confidence interval for β_j

Inference for a single coefficient

```
confint(super.lm, level = 0.9)
```

```
##                      5 %        95 %
## (Intercept) -9.07542163 30.64957440
## complaints   0.33728323  0.88909198
## privileges  -0.30566483  0.15956454
## learn        0.03150994  0.60915429
## raises      -0.29785215  0.46131642
## critical    -0.21354986  0.29031275
## advance     -0.52248482  0.08837146
```