Introduction to Regression Modelling

A Brief Overview of Linear Regression

Casey Crary and Tyler McCord

2025-02-21

What is Regression?

- Regression finds relationships between independent predictors on a continuous numeric scale
- Reveals more complexity than hypothesis testing
- ex. What are the factors that predict SAT test scores?

Our Dataset - SAT

- SAT data assembled for a statistics education journal article on the link between SAT scores and measures of educational expenditures
- contains data from 1994-1995

```
head(SAT)
```

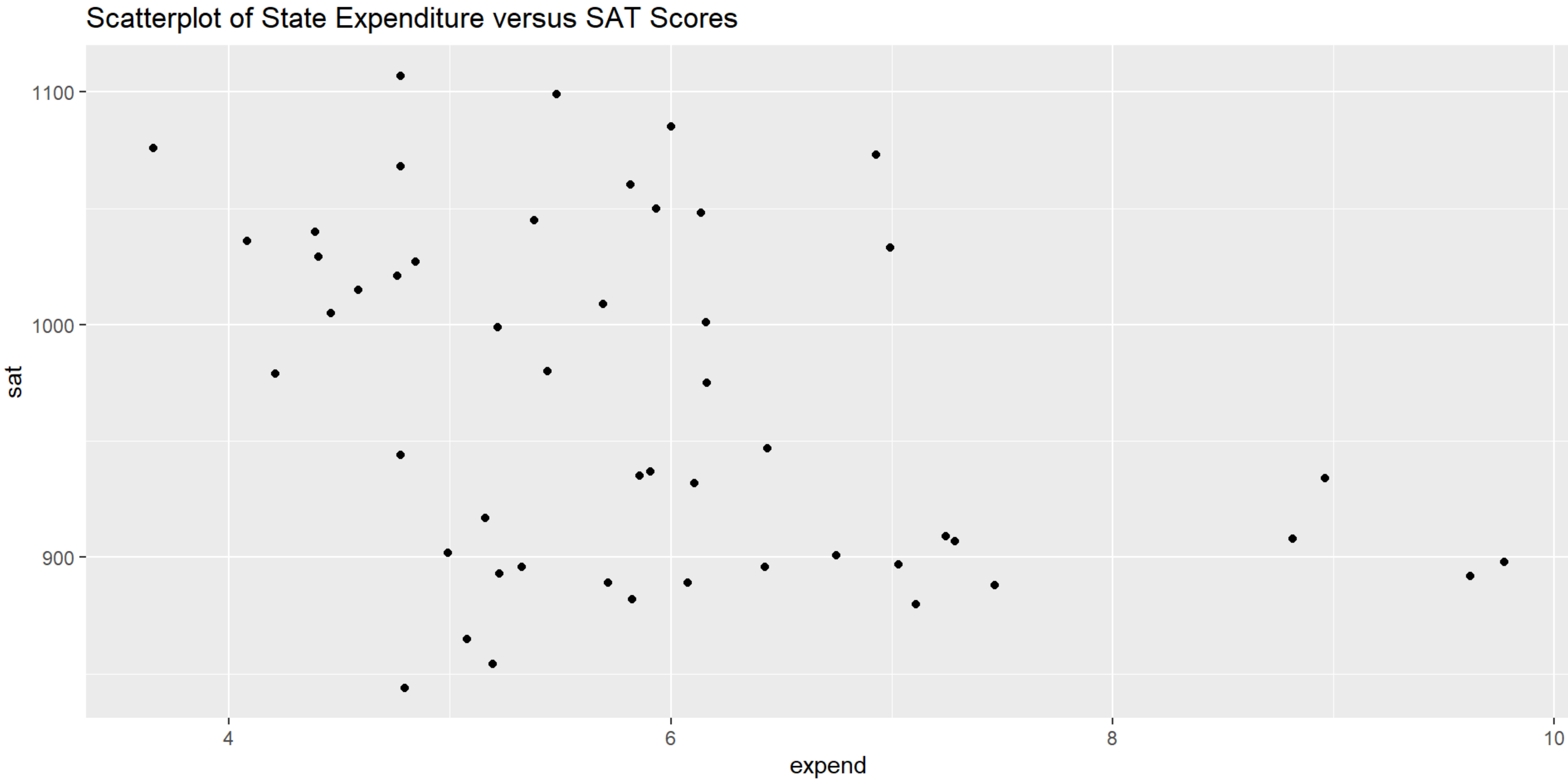
```
       state expend ratio salary frac verbal math  sat
1    Alabama  4.405  17.2 31.144    8    491  538 1029
2     Alaska  8.963  17.6 47.951   47    445  489  934
3    Arizona  4.778  19.3 32.175   27    448  496  944
4   Arkansas  4.459  17.1 28.934    6    482  523 1005
5 California  4.992  24.0 41.078   45    417  485  902
6   Colorado  5.443  18.4 34.571   29    462  518  980
```

Simple Linear Regression

- We can see how **one** quantitative variable impacts another quantitative variable
- Find a least squares regression line
  - Minimize the sum of squared residuals
- Let's try predicting `sat` (each state's average SAT score) using `expend` (expenditure per pupil in average daily attendance in public elementary and secondary schools, in thousands of US dollars)

```r
gf_point(data = SAT, sat ~ expend) +
  labs(title = "Scatterplot of State Expenditure versus SAT Scores")
```

```r
gf_point(data = SAT, sat ~ expend) +
  labs(title = "Scatterplot of State Expenditure versus SAT Scores")
```



Fitting the SLR Model - 1

```r
slr <- lm(sat ~ expend, data = SAT)
msummary(slr)
```

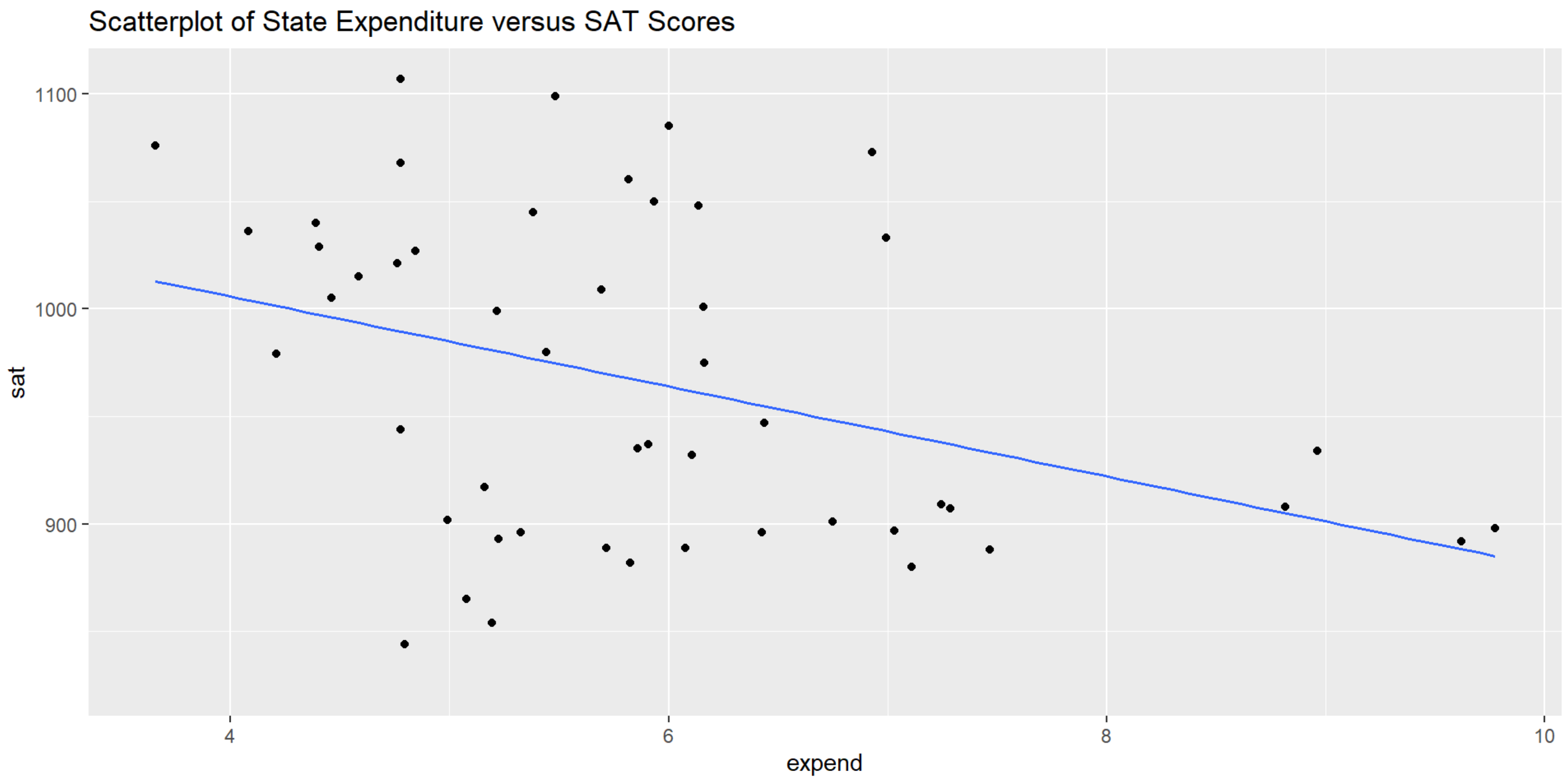
```
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 1089.294     44.390  24.539  < 2e-16 ***
expend       -20.892      7.328  -2.851  0.00641 **

Residual standard error: 69.91 on 48 degrees of freedom
Multiple R-squared:  0.1448,    Adjusted R-squared:  0.127
F-statistic: 8.128 on 1 and 48 DF,  p-value: 0.006408
```

Fitting the SLR Model - 2

- Use `lm()` to fit the model
- `msummary()` to find necessary information
  - Estimate - Coefficients for each variable, this looks like y=mx + b
  - Pr (>|t|) indicates the p-value for each predictor, minimize
  - Residual standard error - every data point is, on average, this far away from the line
  - Multiple R-Squared - the amount of variation in y that is explained by x, maximize
  - F-statistic p-value - significance of the entire model,

SLR Plotted

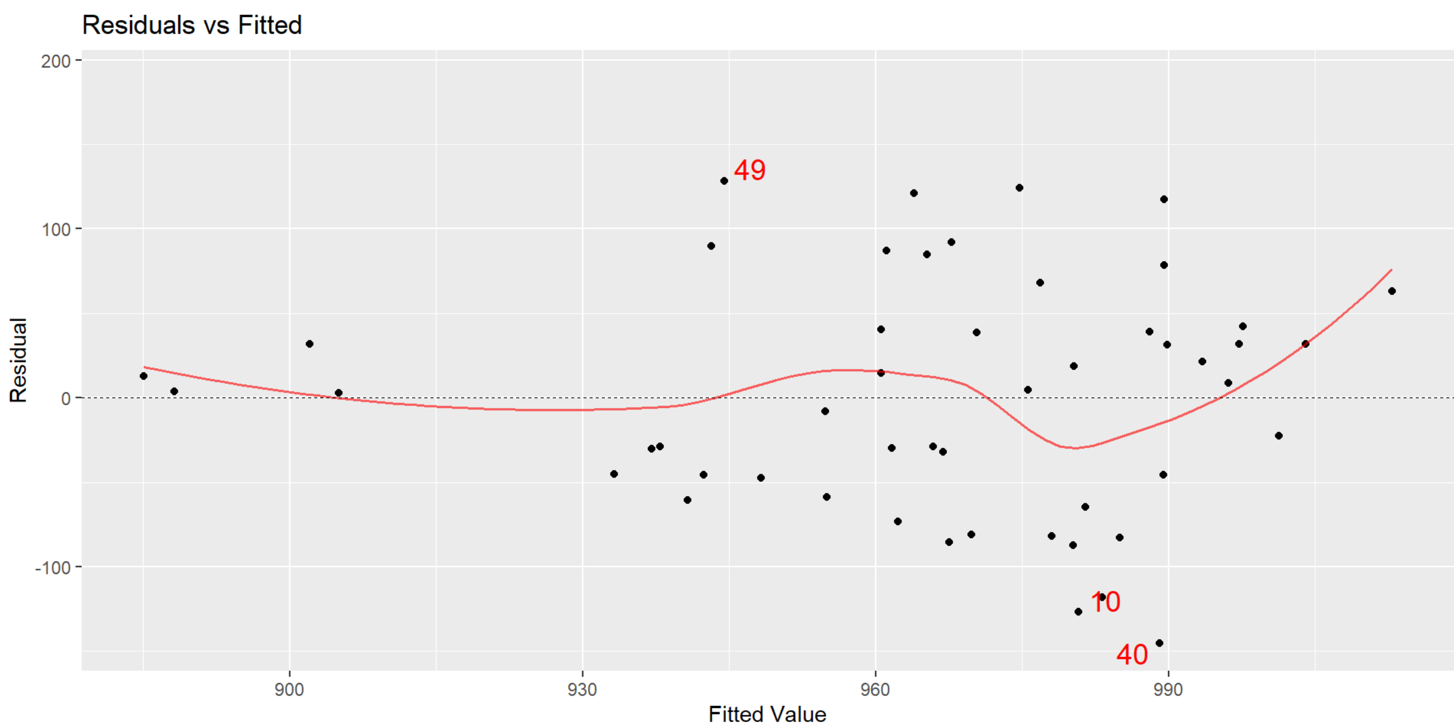
```
gf_point(sat ~ expend, data = SAT) %>%
  gf_lm() +
  labs(title = "Scatterplot of State Expenditure versus SAT Scores")
```



Conditions (LINE) - 1

- Linearity: predictor(s) and response should have a linear relationship

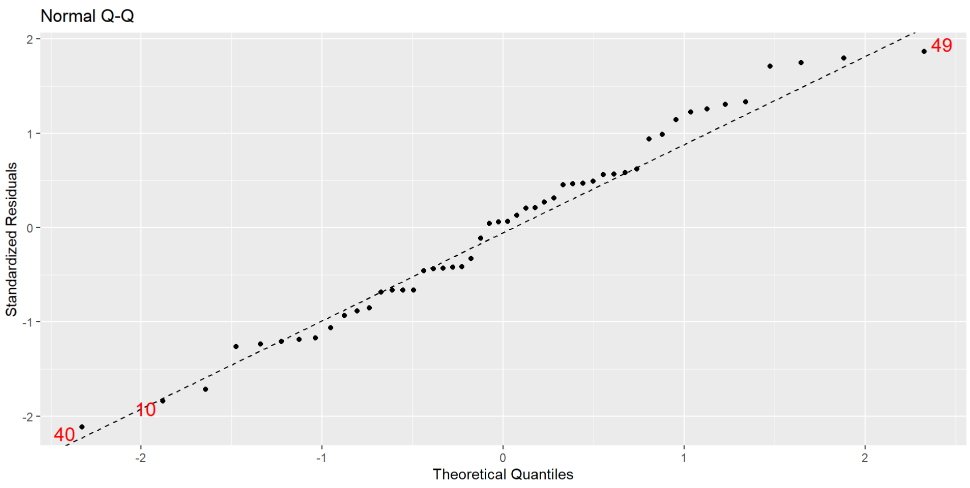
```
1 mplot(slr, which = 1)
```



Conditions (LINE) - 2

- Independence: observations should be independent of each other, this has to do with experimental design
- Normality: We want the residuals to be normally distributed. Use a qqplot or check distribution of residuals.

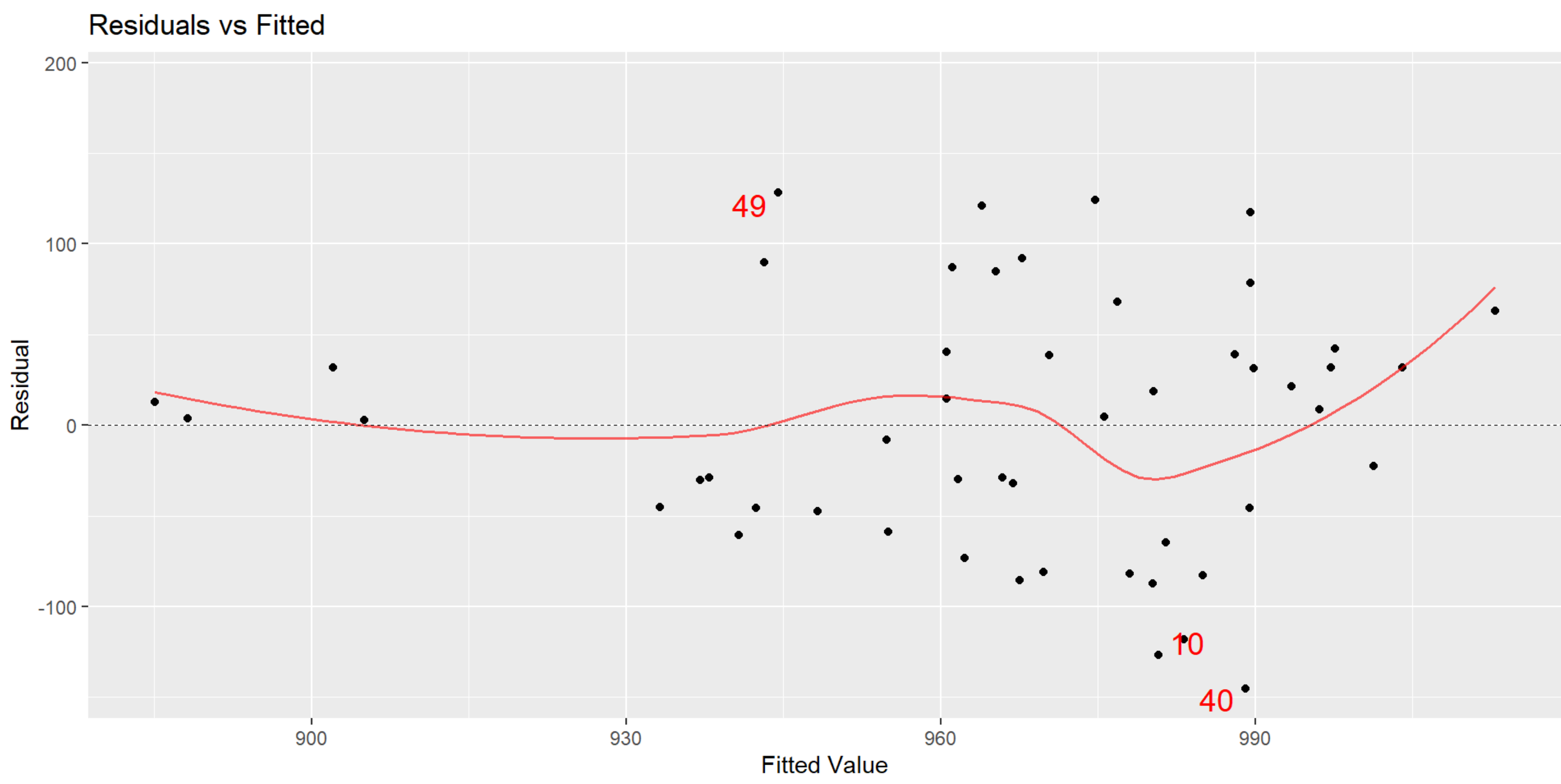
```
mplot(slr, which = 2)
```



Conditions (LINE) - 3

- Equal Variance: residuals should have consistent variation

```
mplot(slr, which = 1)
```



SLR Interpretation

- Fitted equation: $\hat{sat} = -20.89 * expend + 1089.29$
  - For every 1000 dollar increase in expenditure per pupil in public elementary and secondary schools, we predict SAT scores to decrease by 20.89 points
  - It doesn't make sense to interpret the intercept here.
- `expend` has a low p-value (0.0064) for `sat`
  - This p-value is less than alpha=0.05, so therefore `expend` is a significant predictor for `sat`.
  - Significance of a predictor means that it has a non-zero relationship with the response.
- Multiple R-Squared = 0.145
  - This means that 14.5% of the variability in SAT scores can

Multiple Linear Regression

- Extension of SLR, with more predictors
- Let’s fit a multiple linear regression model with 3 predictors:
  - `expend` - expenditure per pupil in average daily attendance in public elementary and secondary schools, 1994-95 (in thousands of US dollars)
  - `ratio` - average pupil/teacher ratio in public elementary and secondary schools, Fall 1994
  - `salary` - estimated average annual salary of teachers in public elementary and secondary schools, 1994-95 (in thousands of US dollars)
- We will again use `sat` as our response variable

Multicollinearity

- Sometimes, our predictors are correlated with each other, but this makes isolating each predictor's effect harder
- Let's look at each predictor's VIF score
  - We want VIF less than 5

```r
library(car) #load this package to use vif()
mlr1 <- lm(sat ~ expend + ratio + salary, data = SAT)
vif(mlr1)
```

```
  expend    ratio   salary 
9.387552 2.285359 8.095274 
```

```r
msummary(mlr1)
```

```
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 1069.234    110.925   9.639 1.29e-12 ***
expend        16.469     22.050   0.747   0.4589    
ratio          6.330      6.542   0.968   0.3383    
salary        -8.823      4.697  -1.878   0.0667 .  
```

```
Residual standard error: 68.65 on 46 degrees of freedom
Multiple R-squared:  0.2096,    Adjusted R-squared:  0.1581
F-statistic: 4.066 on 3 and 46 DF,  p-value: 0.01209
```

MLR Model 2

- Fit a model with only ratio and salary

```r
mlr2 <- lm(sat ~ ratio + salary, data = SAT)
msummary(mlr2)
```

```
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 1113.877     93.003  11.977 6.95e-16 ***
ratio          2.666      4.307   0.619  0.53894    
salary        -5.538      1.643  -3.371  0.00151 ** 

Residual standard error: 68.33 on 47 degrees of freedom
Multiple R-squared:    0.2, Adjusted R-squared:  0.166 
F-statistic: 5.876 on 2 and 47 DF,  p-value: 0.005277
```

Model 3

- Fit a model with only salary (technically, this is a SLR model)

```
mlr3 <- lm(sat ~ salary, data = SAT)
msummary(mlr3)
```

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 1158.859     57.659  20.098  < 2e-16 ***
salary        -5.540      1.632  -3.394  0.00139 **

Residual standard error: 67.89 on 48 degrees of freedom
Multiple R-squared:  0.1935,    Adjusted R-squared:  0.1767
F-statistic: 11.52 on 1 and 48 DF,  p-value: 0.001391
```

Advantages of Linear Regression

- Easily interperatable
- Computational time is low
- Easy to understand

Disadvantages of Linear Regression

- Conditions
- Strongly affected by outliers
- Real-world relationships are often not 1-to-1

Different Types of Regression

- Logistic Regression for binary outcomes
- Polynomial Regression for non-linear relationships between variables
- Lasso Regression is linear regression with a penalty element

Sources

- Thanks to Chelsea Wang for help with the original slides.
- https://hbr.org/2015/11/a-refresher-on-regression-analysis
- https://www.geeksforgeeks.org/add-regression-line-to-ggplot2-plot-in-r/
- https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/assumptions-of-linear-regression/
- https://online.stat.psu.edu/stat462/node/91/
- https://medium.com/@satyavishnumolakala/linear-regression-pros-cons-62085314aef0
- https://www.geeksforgeeks.org/types-of-regression-techniques/