Multiple Linear Regression
(ISLR 3.2, 3.3)

Yingbo Li

Southern Methodist University

STAT 4399

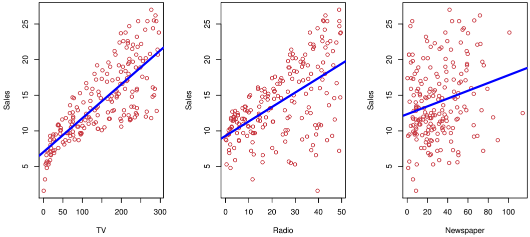
Outline

Multiple linear regression

- One (continuous) response Y, vs. p predictor $X_1, X_2, \ldots, X_p$.
- The multiple linear regression model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon, \text{ where } \epsilon \sim \mathrm{N}(0, \sigma^2)$$

In the context of the `Advertising` data



$$\texttt{Sales} = \beta_0 + \beta_1 \times \texttt{TV} + \beta_2 \times \texttt{Radio} + \beta_3 \times \texttt{Newspaper} + \epsilon$$

Estimating the coefficients: minimize RSS

OLS estimators of $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^T$:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}(\mathbf{X}^T\mathbf{Y}),$$

where

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & X_{1,1} & \cdots & X_{1,p} \\ 1 & X_{2,1} & \cdots & X_{2,p} \\ \vdots & \vdots & & \vdots \\ 1 & X_{n,1} & \cdots & X_{n,p} \end{pmatrix}, \quad \hat{\sigma}^2 = \frac{RSS}{n-p-1}.$$

(Co)variance of the estimator

$$Cov(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$$

Regression plane

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \cdots + \hat{\beta}_p X_p$$

```
> lm2 = lm(Sales ~ ., data = Advertising);
> summary(lm2);

Call:
lm(formula = Sales ~ ., data = Advertising)

Residuals:
    Min      1Q  Median      3Q     Max
-8.8277 -0.8908  0.2418  1.1893  2.8292

Coefficients:  hat(β)                           p-values
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.938889   0.311908   9.422   <2e-16 ***
TV           0.045765   0.001395  32.809   <2e-16 ***
Radio        0.188530   0.008611  21.893   <2e-16 ***
Newspaper   -0.001037   0.005871  -0.177     0.86
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
                                          hat(σ )
Residual standard error: 1.686 on 196 degrees of freedom
Multiple R-squared:  0.8972,    Adjusted R-squared:  0.8956
F-statistic: 570.3 on 3 and 196 DF,  p-value: < 2.2e-16
```

Interpreting the coefficients

$$\hat{\texttt{Sales}} = 2.939 + 0.046 \times \texttt{TV} + 0.189 \times \texttt{Radio} - 0.001 \times \texttt{Newspaper}$$

We interpret β_j as the average effect on Y of a one unit increase in X_j, *holding all other predictors fixed*.

- For a given amount of TV and newspaper advertising, spending an additional $1000 on radio advertising associates with an increase in sales by 189 units on average.

Hypothesis testing on one predictor

Is a specific predictor important? Let's look at X_3 `Newspaper`.

- Simple linear regression
  `Sales` vs. `Newspaper`

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	12.35141	0.62142	19.88	< 2e-16 ***
Newspaper	0.05469	0.01658	3.30	$\boxed{0.00115 \text{ **}}$

- Multiple linear regression
  `Sales` vs. `TV` + `Radio` + `Newspaper`

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2.938889	0.311908	9.422	<2e-16 ***
TV	0.045765	0.001395	32.809	<2e-16 ***
Radio	0.188530	0.008611	21.893	<2e-16 ***
Newspaper	-0.001037	0.005871	-0.177	$\boxed{0.86}$

t-test

The hypotheses for testing for significance of a predictor also takes into account all other predictors.

$$H_0 : \beta_3 = 0 \text{ when other predictors are included in the model.}$$
$$H_1 : \beta_3 \neq 0 \text{ when other predictors are included in the model.}$$

Degrees of freedom for all the t-statistics are $n - p - 1$.

The p-value for `Newspaper` is 0.86. What does this indicate?
If we keep all other predictors in the model, then there is not a significant relationship between newspaper advertising and sales.

`Newspaper` is significant is SLR but not so in MLR. Why?

Collinearity

- Two predictor variables are said to be collinear when they are correlated, and this *collinearity* complicates model estimation.

```
> round(cor(Advertising[, 1:3]), 3);
             TV Radio Newspaper
TV        1.000 0.055     0.057
Radio     0.055 1.000     0.354
Newspaper 0.057 0.354     1.000
```

- We don't like adding highly correlated predictors to the model
  - Adding X_j to the model, if its highly correlated with a predictor X_k that is already in, brings almost no additional explanation ability.

$$R^2(\texttt{Sales} \sim \texttt{Radio}) = 0.3320$$

$$R^2(\texttt{Sales} \sim \texttt{Radio} + \texttt{Newspaper}) = 0.3327$$

  - More dangerously, highly correlated predictors make model unstable. An extreme case: $X_2 = X_1$. Then no unique solution for (β_1, β_2).
- Therefore, we prefer the simplest best model, i.e. *parsimonious* model.

Hypothesis testing: F-test

Is at least one of the predictors $X_1, \ldots, X_p$ useful in predicting Y?

$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0.$

$H_1 :$ at least one $\beta_j \neq 0.$

Test statistic:

$$F = \frac{SS_{reg}/p}{RSS/(n-p-1)} \overset{\text{under} H_0}{\sim} F_{p,n-p-1}$$

ANOVA (analysis of variance) table

ANOVA:	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Regression	3	4860.3	1620.1	570.27	0.00
Residuals	196	556.8	2.8		
Total	199	5417.1			

Comparing two models

- Model 1: contains $X_1, \ldots, X_q$ (the smaller model, suppose $q < p$)
- Model 2: contains $X_1, \ldots, X_p$ (the larger model)

F-test:

$H_0 : \beta_{q+1} = \beta_{q+2} = \cdots = \beta_p = 0.$

$H_1 :$ at least one of the above β_j is nonzero.

Test statistic:

$$F = \frac{(SS_{reg,2} - SS_{reg,1})/(p-q)}{RSS_2/(n-p-1)} \overset{\text{under} H_0}{\sim} F_{p-q,n-p-1}$$

Model 2 has more predictors, so it explains the variability of Y at least as well as Model 1.

- $R_2^2 \geq R_1^2$.
- $SS_{reg,2} \geq SS_{reg,1}$, and thus $RSS_2 \leq RSS_1$.

Categorical predictors

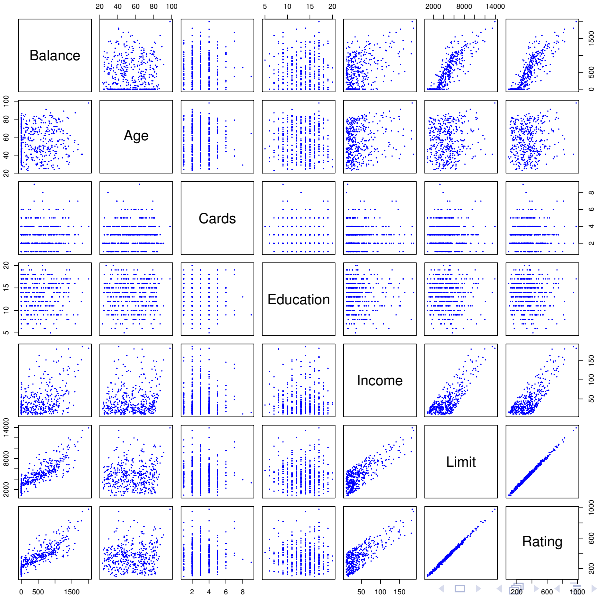
In normal linear regression

- Y has to be continuous (or it violates the $\epsilon \sim \mathrm{N}(0, \sigma^2)$ assumption)
- X_j can be either continuous or discrete.

For example, let's look at the `Credit` data:

- Response: `Balance` (average credit card debt)
- Numerical predictors: `Age`, `Cards` (number of cards), `Education`, `Income`, `Limit` (credit limit), `Rating` (credit rating)
- Categorical predictors: `Gender`, `Student` (Yes/No), `Married`, `Ethnicity` (African American, Asian, Caucasian)

Scatterplots among continuous variables. What do you find?



Categorical predictors with two levels

Formulate the categorical predictor `Gender` as a 0-1 dummy variable:

$$X_{i,1} = \begin{cases} 1 & \text{if the } i\text{th person is female} \\ 0 & \text{if the } i\text{th person male} \end{cases}$$

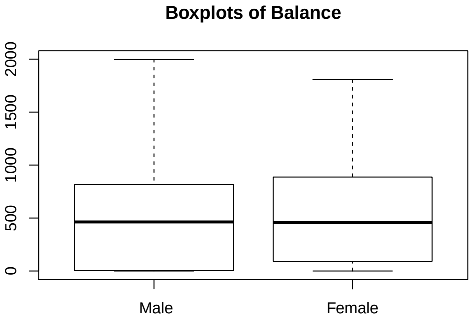
The level assigned with 0 is called the baseline. Here, male is the baseline.

Then a simple linear regression that regresses `Balance` on `Gender` is

$$Y_i = \beta_0 + \beta_1 X_{i,1} + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if the } i\text{th person is female} \\ \beta_0 + \epsilon_i & \text{if the } i\text{th person male} \end{cases}$$

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   509.80      33.13  15.389   <2e-16 ***
GenderFemale   19.73      46.05   0.429    0.669
```

- The average credit card debt is estimated to be \$509.80 for males and \$509.80 + \$19.73 = \$529.53 for females.
- p-value > 0.05, so the difference between genders is not significant.



Categorical predictors with more than two levels

For a categorical predictor with l levels, we create $l - 1$ dummy variables.

For variable `Ethnicity`:

- By alphabetical order, let “African American” be the baseline level.
- We create 2 dummy variables

$$X_{i,1} = \begin{cases} 1 & \text{if the } i\text{th person is Asian} \\ 0 & \text{if the } i\text{th person is not Asian} \end{cases}$$

$$X_{i,2} = \begin{cases} 1 & \text{if the } i\text{th person is Caucasian} \\ 0 & \text{if the } i\text{th person is not Caucasian} \end{cases}$$

Regress `Balance` on `Ethnicity`:

$$Y_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \epsilon_i$$

```
Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)          531.00      46.32  11.464   <2e-16 ***
EthnicityAsian       -18.69      65.02  -0.287    0.774
EthnicityCaucasian   -12.50      56.68  -0.221    0.826
```

All dummy variables are insignificant. So we do not need to include variable `Ethnicity` in the model.

How about some of the dummy variables are significant?

Adding interaction terms

- In our previous analysis of the `Advertising` data, we assumed that the effect on sales of increasing one advertising medium is independent of the amount spent on the other media.
- But suppose that spending money on radio advertising actually increases the effectiveness of TV advertising, so that the slope term for TV should increase as radio increases.
- Let's consider the model with `TV`, `Radio`, and their interaction term.

$$\begin{aligned}\texttt{Sales} &= \beta_0 + \beta_1 \times \texttt{TV} + \beta_2 \times \texttt{Radio} + \beta_3 \times \texttt{TV} \times \texttt{Radio} + \epsilon \\ &= \beta_0 + (\beta_1 + \beta_3 \times \texttt{Radio}) \times \texttt{TV} + \beta_2 \times \texttt{Radio} + \epsilon\end{aligned}$$

Interpretation

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 6.750e+00  2.479e-01  27.233   <2e-16 ***
TV          1.910e-02  1.504e-03  12.699   <2e-16 ***
Radio       2.886e-02  8.905e-03   3.241   0.0014 **
TV:Radio    1.086e-03  5.242e-05  20.727   <2e-16 ***
```

- R^2 increase to 0.9678 from 0.8972.
  This means that $(0.9678 - 0.8972)/(1 - 0.8972) = 69\%$ of the variability in sales that remains after fitting the additive model has been explained by the interaction term.
- An increase in TV advertising of $1000 is associated with increased sales of $(\beta_1 + \beta_3 \times \texttt{Radio}) \times 1000 = 19.10 + 1.086 \times \texttt{Radio}$ units.
- An increase in radio advertising of $1000 is associated with increased sales of $(\beta_2 + \beta_3 \times \texttt{TV}) \times 1000 = 28.86 + 1.086 \times \texttt{Radio}$ units.

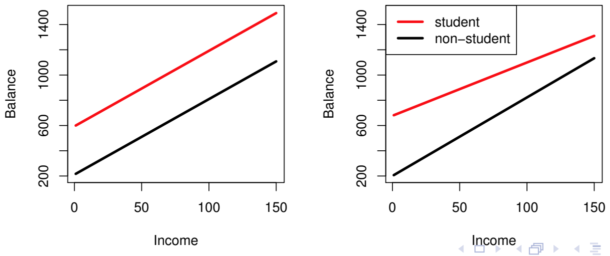
If the interaction term is significant, but the associated main effects are not. Should we include the main effects?

Interactions involving dummy variables

In the `Credit` data, we include X_1 `Income` and a dummy variable X_2 for `Student` (1 – yes; 0 – no).

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \epsilon$$

	No interaction	With interaction
No student	$Y = \beta_0 + \beta_1 X_1 + \epsilon$	
Student	$Y = (\beta_0 + \beta_2) + \beta_1 X_1 + \epsilon$	$Y = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) X_1 + \epsilon$

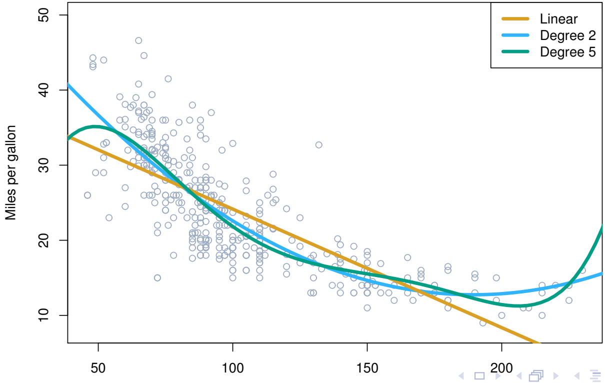


Other extensions

- Adding quadric terms

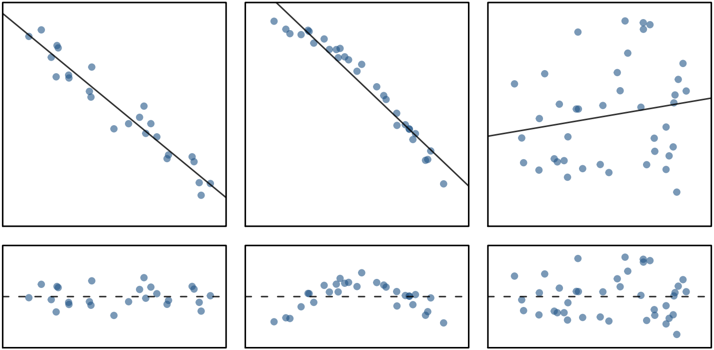
$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \epsilon$$

- Adding more polynomials (we will revisit this in Ch 7).



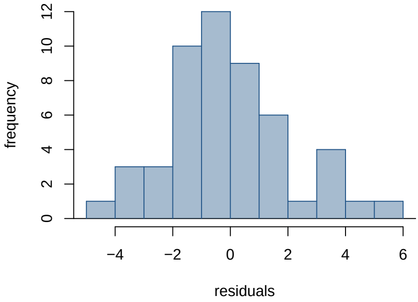
Model diagnostics: (1) linearity

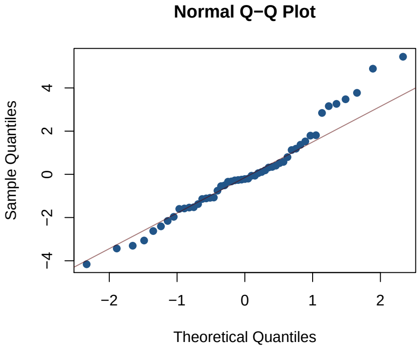
- The relationship between the explanatory and the response variable should be linear.
- Check using a scatterplot of the data, or a *residuals plot*.



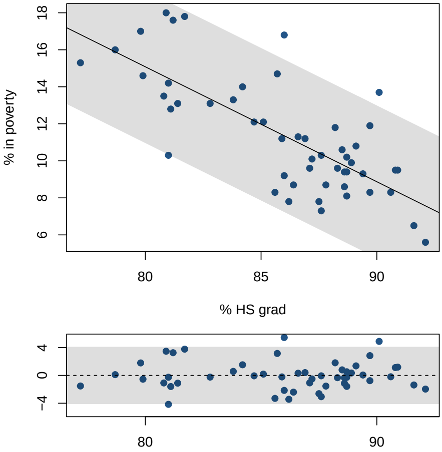
Model diagnostics: (2) nearly normal residuals

- The residuals should be nearly normal.
- This condition may not be satisfied when there are unusual observations that don't follow the trend of the rest of the data.
- Check using a histogram or normal probability plot of residuals.





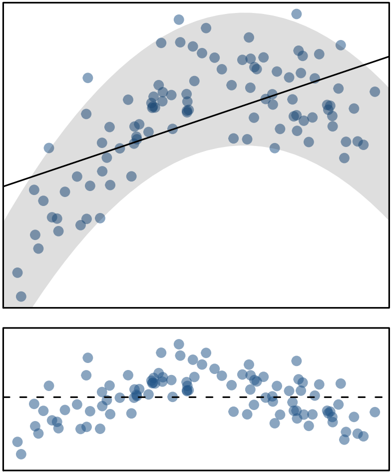
Model diagnostics: (3) constant variability



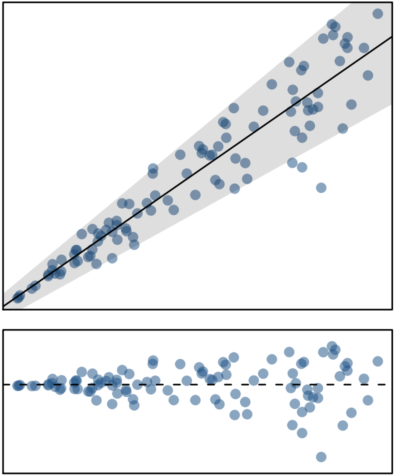
- The variability of points around the least squares line should be roughly constant.
- This implies that the variability of residuals around the 0 line should be roughly constant as well.
- Also called *homoscedasticity*.
- Check using a histogram or normal probability plot of residuals.

What conditions are violated?

Linear model 1



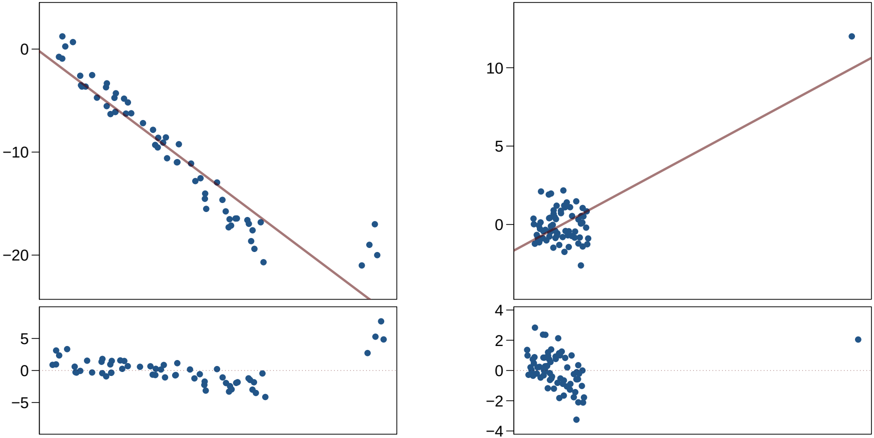
Linear model 2



Model diagnostics: (4) outliers and leverage points

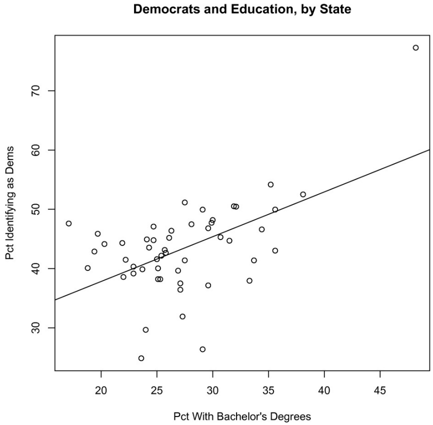
How do the outliers influence the least squares line?

Think of where the regression line would be with and without the outliers.



Some terminology

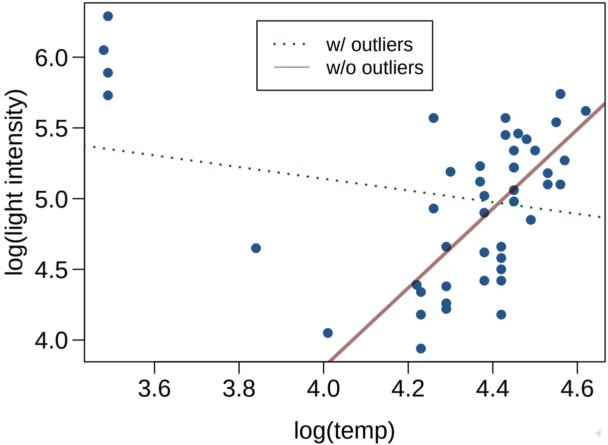
- *Outliers* are points that fall away from the cloud of points.
- Outliers that fall horizontally away from the center of the cloud are called *leverage* points.
- High leverage points that actually influence the slope of the regression line are called *influential* points.



Note: *http://www.progressivepolicy.org/blog/more-college-graduates-more-democratic-voters/*

Influential points

Data are available on the log of the surface temperature and the log of the light intensity of 47 stars in the star cluster CYG OB1.



Types of outliers