Multiple Regression

Instructor: Cheng-Ying Chou

10/12/2020

Regression

- Purposes of Regression
  1. Describe association between Y and X's
  2. Make predictions:
     - Interpolation: making prediction within a range of X's
     - Extrapolation: making prediction outside a range of X's
  3. To "adjust" or "control for" confounding variables
- What is "Y"?
  - an outcome variable
  - 'dependent' variable
  - 'response'
- Type of regression depends on type of Y
  - continuous (linear regression)
  - binary (logistic regression)
  - time-to-event (Cox regression)
  - rare event or rate (Poisson regression)

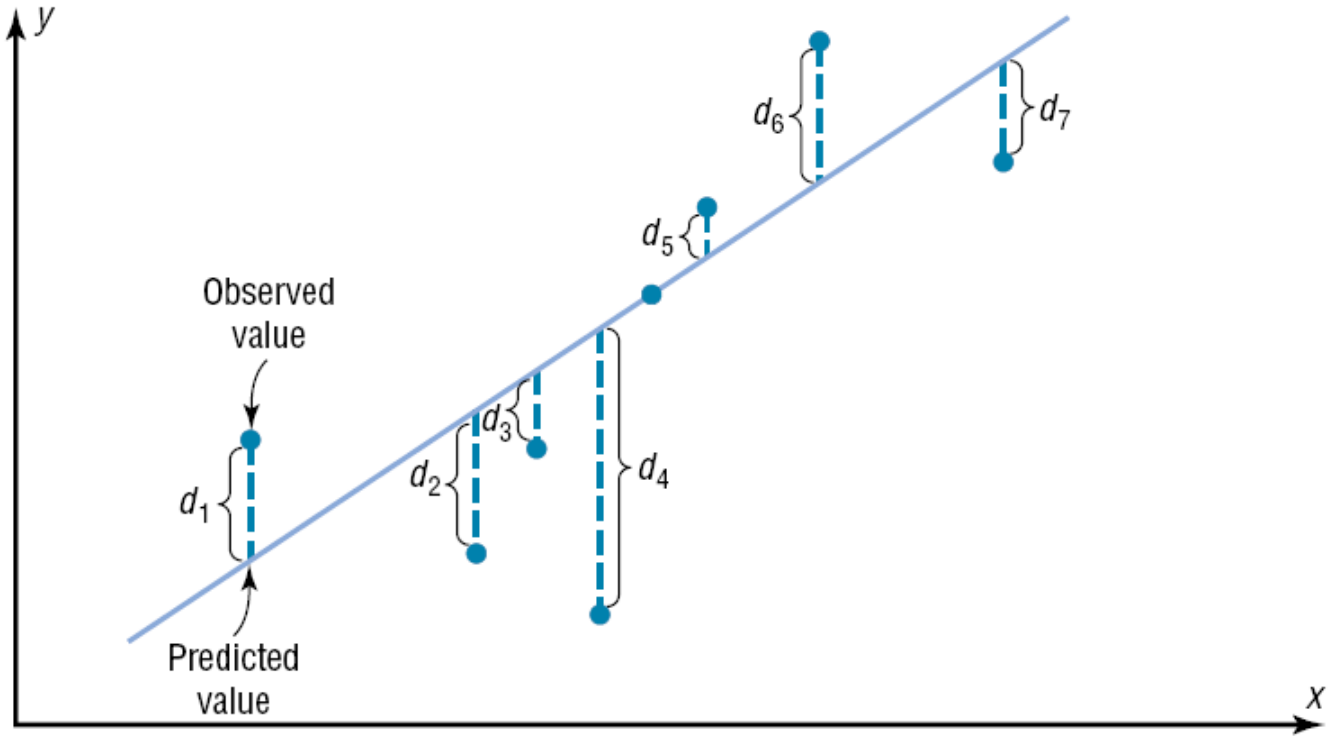
Simple linear regression model

$$y = \beta_0 + \beta_1 x$$

- Where β_1 can be interpreted as the change in the mean of *Y* for a unit change in *x*.
- Also, the variability of Y at a particular value of *x* is determined by the error variance, σ^2.
- This implies there is a distribution of *Y*-values at each *x* and that the variance of this distribution is the same at each *x*.

Regression

- **Best fit** means that the sum of the squares of the vertical distance from each point to the line is at a minimum.



Variation

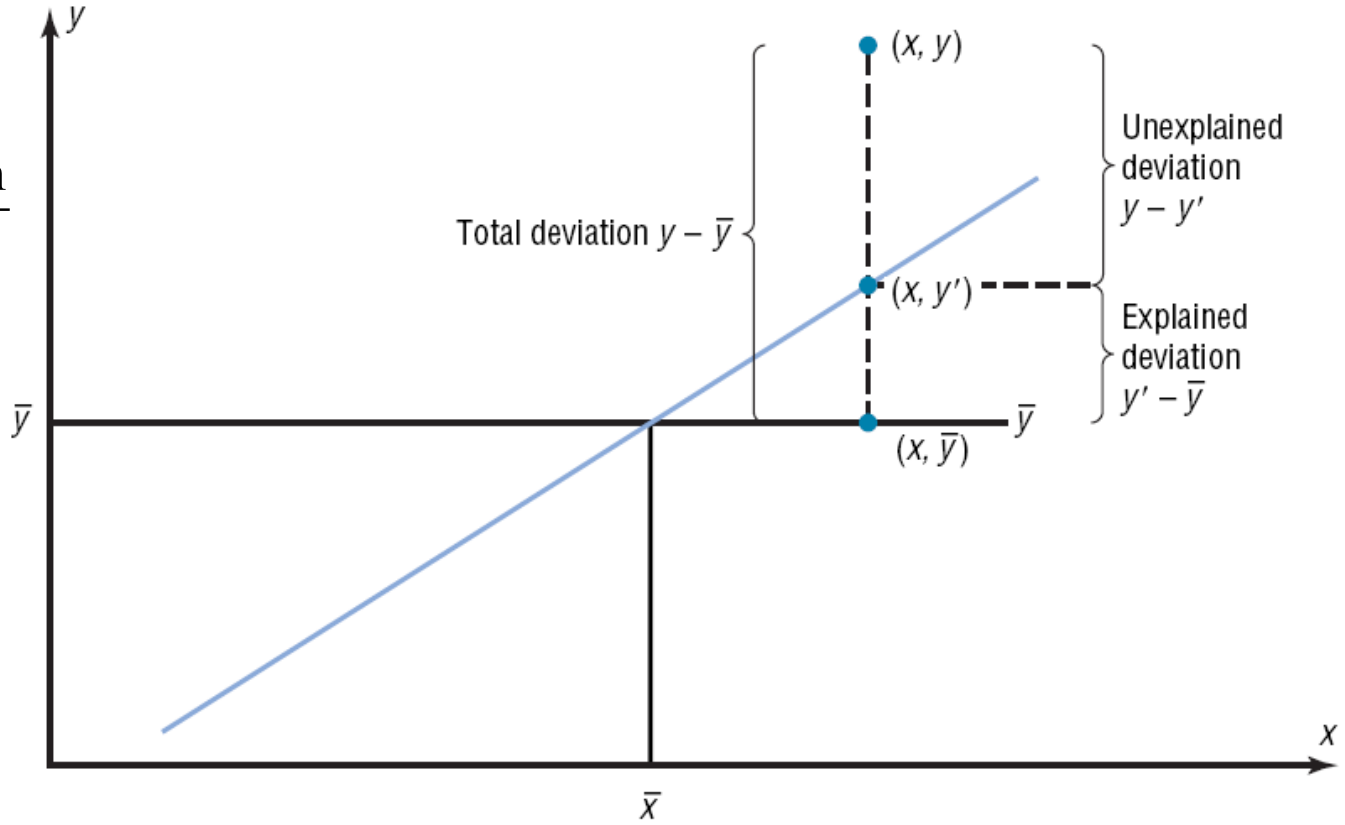
Due to relationship

- **Total variation = explained variation + unexplained variation**

$$\sum(y_i - \bar{y})^2 = \sum(\hat{y}_i - \bar{y})^2 + \sum(y_i - \hat{y}_i)^2$$

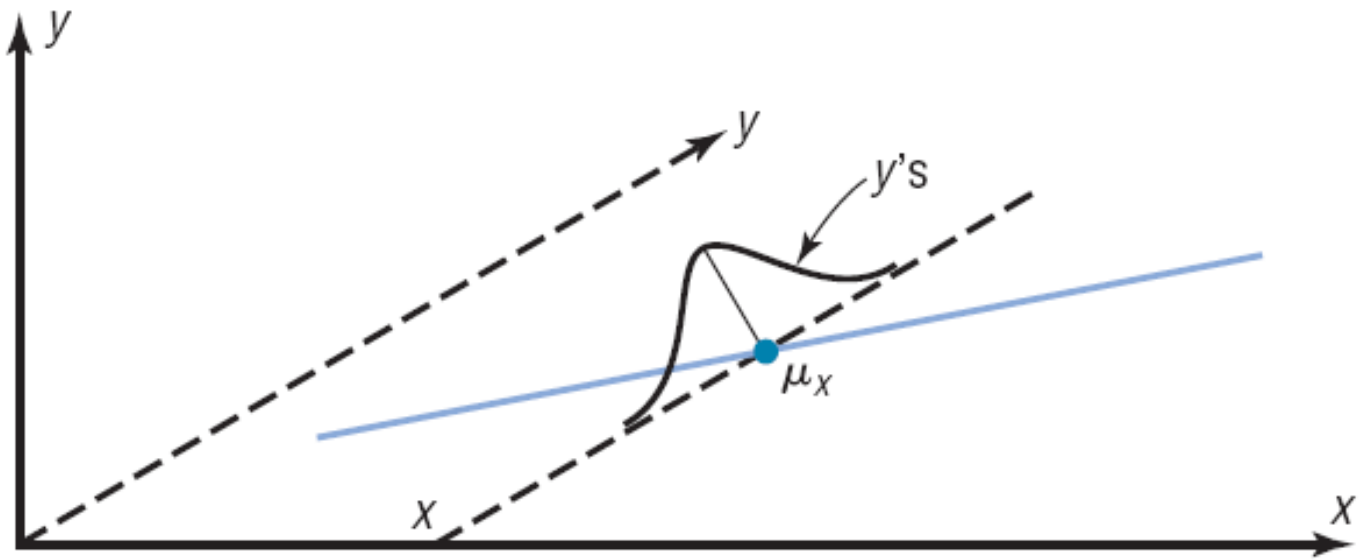
SST = SSR + SSE

$$r^2 = \frac{\text{explained variation}}{\text{total variation}}$$



Assumptions for Valid Predictions

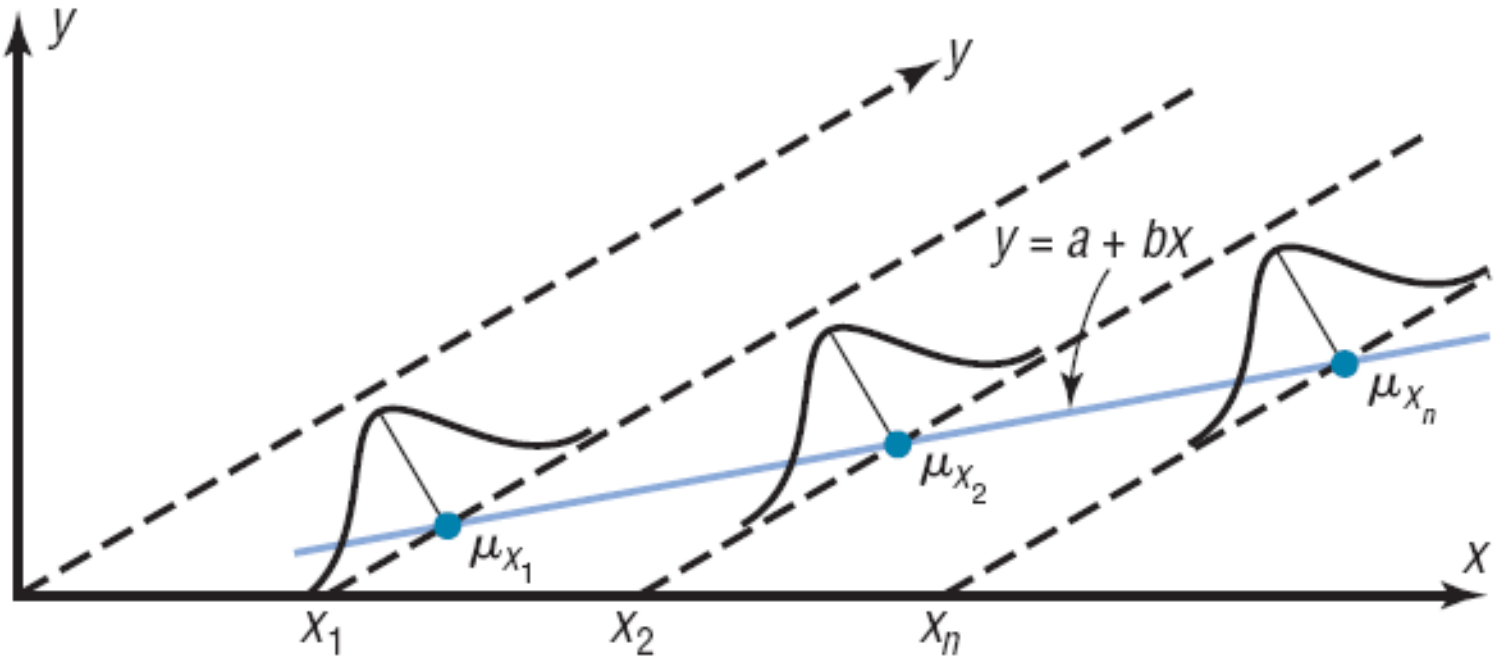
1. The sample is a random sample.
2. For any specific value of the independent variable x, the value of the dependent variable y must be normally distributed about the regression line.



(a) Dependent variable *y* normally distributed

Assumptions for Valid Predictions

3. The standard deviation of each of the dependent variables must be the same for each value of the independent variable.



(b) $\sigma_1 = \sigma_2 = \cdots = \sigma_n$

Least squares method

- To finding estimated values of intercept and slopes for the model by minimizing SSE

$$SSE = \Sigma(y_i - \hat{y}_i)^2 = \Sigma(y_i - \beta_0 - \beta_1 x_i)^2$$

$$(\hat{\beta}_0, \hat{\beta}_1) = \arg\min_{\beta_0, \beta_1} \Sigma(y_i - \beta_0 - \beta_1 x_i)^2$$

$$\frac{\partial SSE}{\partial \beta_0} = -2\Sigma(y_i - \beta_0 - \beta_1 x_i) = 0$$

$$\frac{\partial SSE}{\partial \beta_1} = -2\Sigma(y_i - \beta_0 - \beta_1 x_i)x_i = 0$$

Pure error

- SST=SSR+SSE

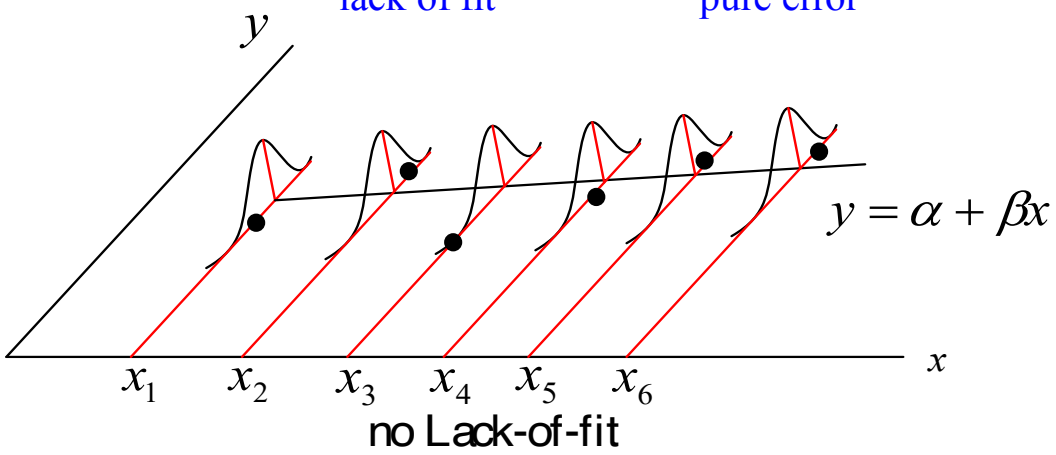
$$\sum(y_i-\bar{y})^2=\sum(\hat{y}_i-\bar{y})^2+\sum(y_i-\hat{y}_i)^2$$

sum of square = sum of square + sum of square
about the mean due to regression about regression

- Error = lack-of-fit + pure error

$$\sum(y_i-\hat{y}_i)^2=\underbrace{\sum(\bar{y}_i-\hat{y}_i)^2}_{\text{lack of fit}}+\underbrace{\sum(y_i-\bar{y}_i)^2}_{\text{pure error}}$$

From:
Walpole著2002
probability
& statistics
for engineering
& scientists



統計軟體將Error SS分解為 {Lack of fit / Pure error} 兩項

X	Y
150	77.4
150	76.7
150	78.2
200	84.1
200	84.5
200	83.7
250	88.9
250	89.2
250	89.7
300	94.8
300	94.7
300	95.9

變方分析 Analysis of variance (ANOVA)

source	df	SS	MS	F	P
Regression	1	509.1050	509.10	1531.60	< 0.0001 (顯著)
Error	10	4.0117	0.401		
Lack-of-fit	2	1.3517	0.6758	2.03	0.19 (不顯著)
Pure error	8	2.6600	0.3325		
Total	11	513.11			

Statistical Inference for Linear Regression

$$H_0 : \beta = 0$$
$$H_1 : \beta \neq 0$$

Practice with exAM.csv

1. Does there exist a linear relationship between y and x?

2. Does lack of fit exist?

Lack-of-fit and pure error

```
 ## with data from Greene (1993):
test<-read.csv("exAM.csv",header=T,sep=",");
fit<-lm(y~x, data=test)
summary(fit)
anova(fit)
plot(test$x,test$y)

#install.packages("alr3")
library(alr3)
pureErrorAnova(fit)

## Another way to show lack-of-fit
fit2<-lm(y~factor(x), data=test)
anova(fit,fit2)
```

Multiple Regression Analysis (MRA)

- Review Simple Regression Analysis
- Multiple Regression Analysis
  - Design requirements
  - Multiple regression model
  - R2
  - Testing R2 and b's
  - Comparing models
  - Comparing standardized regression coefficients

Multiple Regression Analysis (MRA)

- Method for studying the relationship between a dependent variable and two or more independent variables.
- Purposes:
  - Prediction
  - Explanation
  - Theory building

Design Requirements

- One dependent variable (criterion)
- Two or more independent variables (predictor variables).
- Sample size: >= 50 (at least 10 times as many cases as independent variables)

Assumptions for Multiple Regression

1. ***Normality assumption***—for any specific value of the independent variable, the values of the y variable are normally distributed.
2. ***Equal-variance assumption***—the variances (or standard deviations) for the y variables are the same for each value of the independent variable.
3. ***Linearity assumption***—there is a linear relationship between the dependent variable and the independent variables.
4. ***Non-multicollinearity assumption***—the independent variables are not correlated.
5. ***Independence assumption***—the values for the y variables are independent.

Simple vs. Multiple Regression

- One dependent variable Y predicted from one independent variable X
- One regression coefficient
- **r^2**: proportion of variation in dependent variable Y predictable from X

- One dependent variable Y predicted from a set of independent variables (X1, X2 ….Xk)
- One regression coefficient for each independent variable
- **R^2**: proportion of variation in dependent variable Y predictable by set of independent variables (X's)

Multiple regression in matrix form

$$y_i = \beta_0 1 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i = \mathbf{x}_i^{\mathsf{T}} \beta + \varepsilon_i, \qquad i = 1, \ldots, n,$$

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \qquad \mathbf{X} = \begin{pmatrix} \mathbf{x}_1^{\mathsf{T}} \\ \mathbf{x}_2^{\mathsf{T}} \\ \vdots \\ \mathbf{x}_n^{\mathsf{T}} \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix} \qquad \epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

- Estimated parameters

$$\mathbf{y} = X\beta + \varepsilon$$

$$E(\mathbf{y}) = \mathbf{X}\beta$$

$$\operatorname{cov}(\mathbf{y}) = \sigma^2 \mathbf{I}$$

$$\mathbf{b} = (\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{y}$$

$$E(\epsilon) = 0$$

$$\operatorname{cov}(\epsilon) = \sigma^2 \mathbf{I}$$

Fitting the model

- LS provides estimates of the unknown model parameters, $\beta_0, \beta_1, \ldots, \beta_k$ which minimizes Q

$$Q = SSE = \sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik})]^2$$

$$\frac{\partial SSE}{\partial \beta_0} = -2 \Sigma [y_i - (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik})] = 0$$

$$\frac{\partial SSE}{\partial \beta_j} = -2 \Sigma [y_i - (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik})] x_{ij} = 0$$

Statistical Inference for Multiple Regression

- Determine which predictor variables have statistically significant effects.
- We test the hypotheses:

$$H_{0j}: \beta_j = 0 \quad vs. \quad H_{1j}: \beta_j \neq 0$$

- If we can't reject H_{0j}, then x_j is not a significant predictor of y.

The main model-fitting commands

- lm (linear models for fixed effects)
- lme and lmer (linear models for mixed effects)
- glm (generalized linear models for fixed effects)
- nls (nonlinear least squares)
- gam (generalized additive models)

Basic syntax for a regression analysis in R

$\mathrm{lm}(Y \sim model)$ $\mathrm{glm}(Y \sim model)$

Syntax	Model	Comments
Y ~ A	$Y = \beta_o + \beta_1 A$	Straight-line with an implicit y-intercept
Y ~ -1 + A	$Y = \beta_1 A$	Straight-line with no y-intercept; that is, a fit forced through (0,0)
Y ~ A + I(A^2)	$Y = \beta_o + \beta_1 A + \beta_2 A^2$	Polynomial model; note that the identity function **I( )** allows terms in the model to include normal mathematical symbols.
Y ~ A + B	$Y = \beta_o + \beta_1 A + \beta_2 B$	A first-order model in A and B without interaction terms.
Y ~ A:B	$Y = \beta_o + \beta_1 AB$	A model containing only first-order interactions between A and B.
Y ~ A*B	$Y = \beta_o + \beta_1 A + \beta_2 B + \beta_3 AB$	A full first-order model with a term; an equivalent code is Y ~ A + B + A:B.
Y ~ (A + B + C)^2	$Y = \beta_o + \beta_1 A + \beta_2 B + \beta_3 C + \beta_4 AB + \beta_5 AC + \beta_6 AC$	A model including all first-order effects and interactions up to the n^{th} order, where n is given by **( )^n**. An equivalent code in this case is Y ~ A*B*C – A:B:C.

Sample formulas

For a model with response y and predictors a, b and c

Model	Interpretation
Y ~ 1	Just the intercept
Y ~ a	One main effect
Y ~ -1+a	No intercept
Y ~ a+b	Two main effects
Y ~ a+b+c+a:b	Three main effects and an interaction between a and b
Y ~ a*b	All main effects and interactions (a+b+a:b)
Y ~ factor(a)	Create dummy variables for a (if not already a factor)
Y ~ (a+b+c)^2	All main effects and second-order interations (a+b+c+a:b+b:c+a:c)
Y ~ I(a^2)	Transform a to a^2
Log(Y) ~ a	Log transform Y
Y ~ a/b/c	Factor c nested within factor b within factor a
Y ~ .	Main effect for each column in the dataframe

Fit a linear model to data

- z <- lm(response ~ explanatory, data = mydata)
- z <- lm(response ~ explanatory, data = mydata, na.action = na.exclude)
  - The argument *na.action = na.exclude* is optional --- it tells R to keep track of cases having missing values, in which case the residuals and predicted values will have NA's inserted for those cases.
- The most useful functions to extract results
  - summary(z) # parameter estimates and overall model fit
  - plot(z) # plots of residuals, normal quantiles, leverage
  - coef(z) # model coefficients (means, slopes, intercepts)
  - confint(z) # confidence intervals for parameters
  - resid(z) # residuals
  - predict(z) # predicted values
  - predict(z, newdata = mynewdata, interval="confidence") # the model to predict values for new observations contained in the data frame mynewdata.
  - fitted(z) # predicted values
  - anova(z1, z2) # compare fits of 2 models, "full" vs "reduced"
  - anova(z) # ANOVA table (** terms tested sequentially **)

Tire tread wear vs. mileage

- The table gives the measurements on the groove of one tire after every 4000 miles.
- Our Goal: to build a model to find the relation between the mileage and groove depth of the tire.

Mileage (in 1000 miles)	Groove Depth (in mils)
0	394.33
4	329.50
8	291.00
12	255.17
16	229.33
20	204.83
24	179.00
28	163.83
32	150.33

```
# Linear Regression
mile<-c(0,4,8,12,16,20,24,28,32);
depth<-c(394.33, 329.5, 291, 255.17, 229.33, 204.83, 179, 163.83,
150.33);


mod1<-lm(depth~mile + I(mile^2))
summary(mod1)

Call:
lm(formula = depth ~ mile + I(mile^2))

Residuals:
    Min      1Q  Median      3Q     Max
-8.4212 -2.9548  0.4207  3.4892  8.0652

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 386.26485    4.79996   80.47 2.48e-10 ***
mile        -12.77238    0.69948  -18.26 1.74e-06 ***
I(mile^2)     0.17162    0.02103    8.16 0.000182 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.906 on 6 degrees of freedom
Multiple R-squared:  0.9961,    Adjusted R-squared:  0.9948
F-statistic: 762.8 on 2 and 6 DF,  p-value: 6.011e-08
```

Depth=386.26-12.77mile+0.172mile^2

Testing Significance of b's

- H0: $\beta = 0$
- $t_{observed} = \dfrac{b - \beta}{\text{standard error of } b}$

- with N-(k+1) df

Example: t-test of b

- $t_{observed} = (-.44 - 0)/14.24$
- $t_{observed} = -.03$
- $t_{critical(.05,2,100)} = 1.97$

- Decision: cannot reject the null hypothesis.
- Conclusion: the population β for GSC is not significantly different from 0.

Multiple regression code

```
# Multiple Linear Regression Example
fit <- lm(y ~ x1 + x2 + x3, data=mydata)
summary(fit) # show results

# Other useful functions
coefficients(fit) # model coefficients
confint(fit, level=0.95) # CIs for model parameters
fitted(fit) # predicted values
residuals(fit) # residuals
anova(fit) # anova table
vcov(fit) # covariance matrix for model parameters
influence(fit) # regression diagnostics
```

```
# diagnostic plots
layout(matrix(c(1,2,3,4),2,2)) # optional 4 graphs/page
plot(fit)

# compare models
fit1 <- lm(y ~ x1 + x2 + x3 + x4, data=mydata)
fit2 <- lm(y ~ x1 + x2)
anova(fit1, fit2)
```

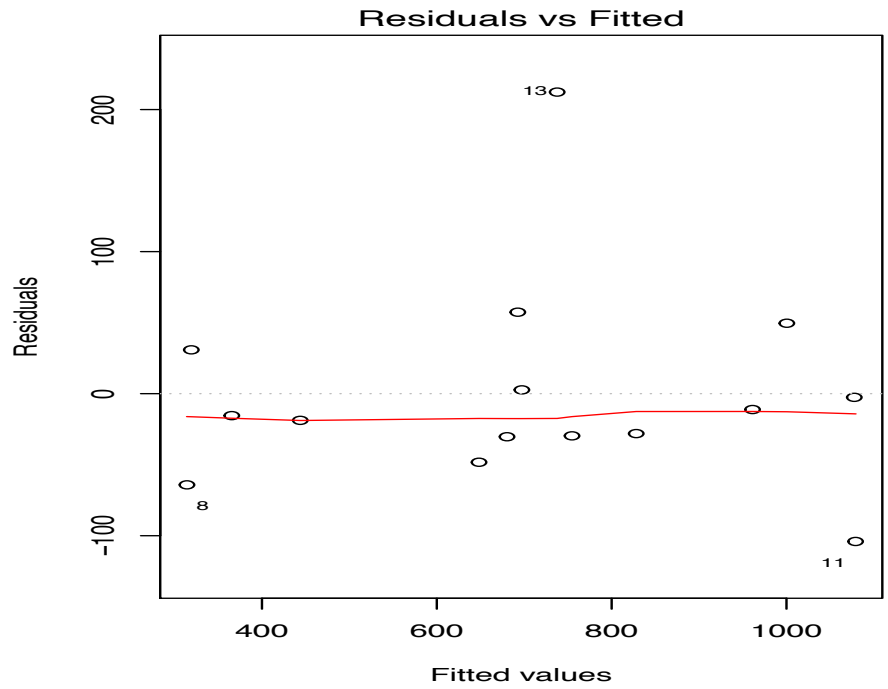
Calculate volumes (volume) and page areas (area) for the books on which information is given in the dataset oddbooks.

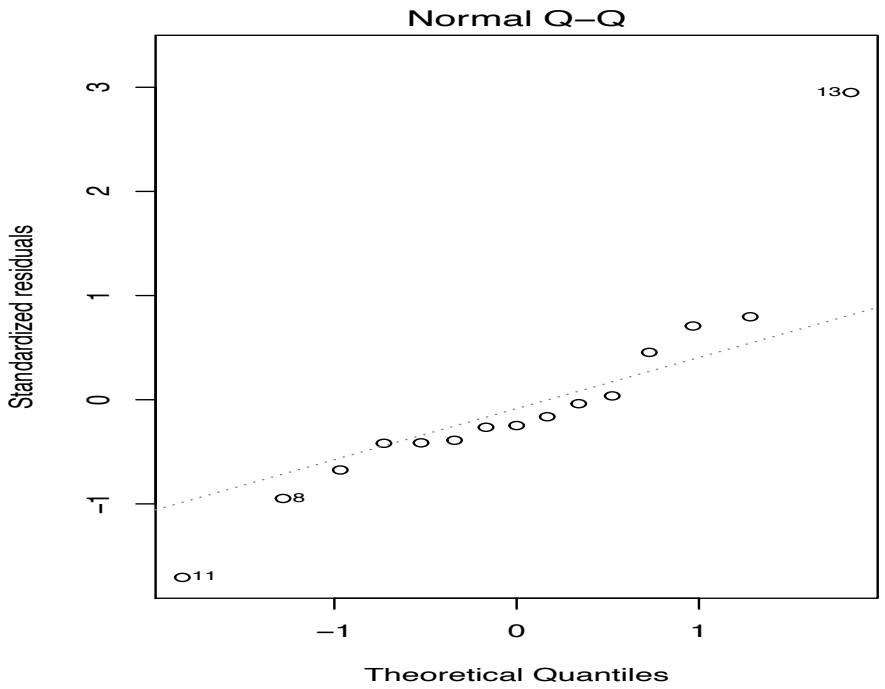
a) Plot log(weight) against log(volume), and fit a regression line.
b) Plot log(weight) against log(area), and again fit a regression line.
c) Which of the lines (a) and (b) gives the better fit?
d) Repeat (a) and (b), now with log(density) in place of log(weight) as the dependent variable. Comment on how results from these regressions may help explain the results obtained in (a) and (b).

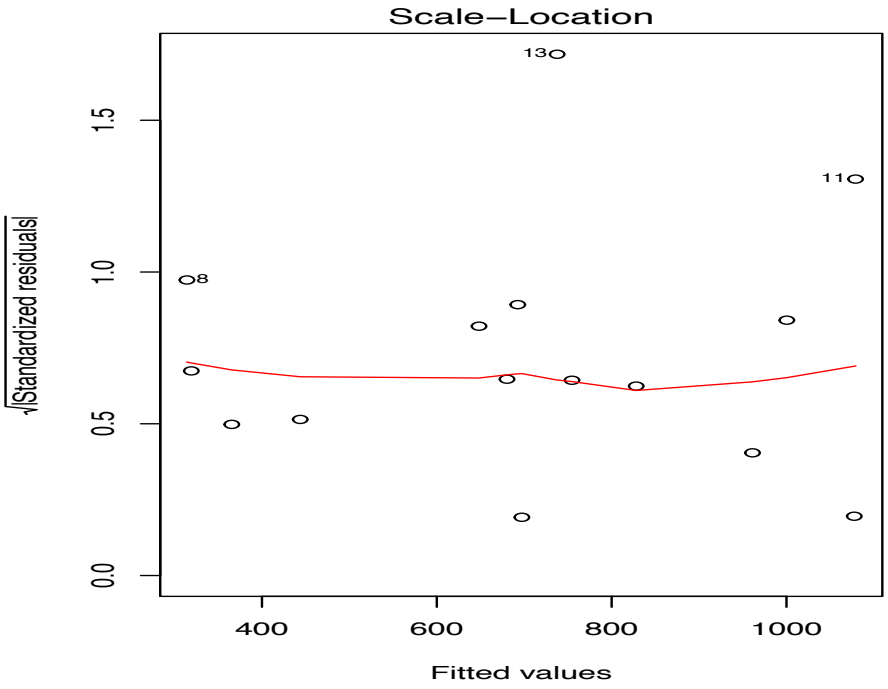
Plot the weight vs. volume for books

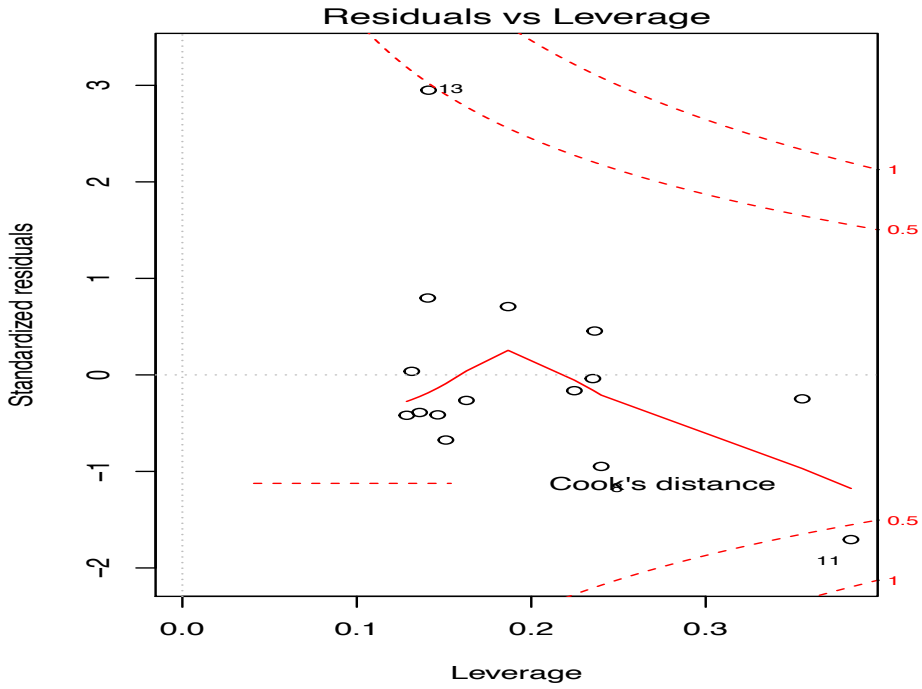
```
library(DAAG)
plot(weight ~ volume,
data=allbacks,pch=c(16,1)[unclass(cover)])
# unclass(cover) gives the integer codes that
identify levels
with(allbacks, text(weight ~ volume,
labels=paste(1:15),pos=c(2,4)[unclass(cover)]))
```

Try to add area to the explanatory variables


Residuals vs Fitted
Residuals
Fitted values
13
8
11
200
100
0
−100
400
600
800
1000


Normal Q−Q
Standardized residuals
Theoretical Quantiles
13
8
11
3
2
1
0
−1
−1
0
1


Scale−Location
√|Standardized residuals|
Fitted values
13
11
8
1.5
1.0
0.5
0.0
400
600
800
1000


Residuals vs Leverage
Standardized residuals
Leverage
Cook's distance
13
11
1
0.5
0.5
1
3
2
1
0
−1
−2
0.0
0.1
0.2
0.3

Cook's distance or Cook's D

- Used estimate of the influence of a data point when performing a least-squares regression analysis

$$D_i = \frac{\sum_{j=1}^{n} \left(\hat{y}_j - \hat{y}_{j(i)} \right)^2}{ps^2}$$

$\hat{y}_{j(i)}$ is the fitted response value obtained when excluding i

p: # of parameters in the regression model

s^2: mean squared error of the regression, i.e., SSE

Cook's D

- Cook's distance and leverage are used to detect highly influential data points.
- D follows F(p, N-p) distribution. We need to calculate the quantile of each observation's D value.
- For large sample sizes, a rough guideline is to consider Cook's distance values above 1 to indicate highly influential points.
- Leverage values greater than 2*p/N indicate high leverage observations.

Leverage

- Leverage is a measure of how far an independent variable deviates from its mean.
- These leverage points can have an effect on the estimate of regression coefficients.
- **槓桿量H hat**
- 超過2P/N代表該樣本點可能為離群值，此迴歸模式參數個數p=6，樣本數=16，所以計算出來的2P/N=0.75，因此H值超過0.75可能為離群值。

Try with mtcars data

32 observations on 11 (numeric) variables.

- Test the relationship between mpg and (wt, disq, hp, drat).

[, 1]	mpg	Miles/(US) gallon
[, 2]	cyl	Number of cylinders
[, 3]	disp	Displacement (cu.in.)
[, 4]	hp	Gross horsepower
[, 5]	drat	Rear axle ratio
[, 6]	wt	Weight (1000 lbs)
[, 7]	qsec	1/4 mile time
[, 8]	vs	Engine (0 = V-shaped, 1 = straight)
[, 9]	am	Transmission (0 = automatic, 1 = manual)
[,10]	gear	Number of forward gears
[,11]	carb	Number of carburetors

US States

- We will build a model to predict the % of the state that voted for Obama (out of the two party vote) in the 2012 US presidential election, using the 50 states as cases.
- This can help us to understand how certain features of a state are associated with political beliefs.

Interpreting R^2

A regression where the cases are states, the response variable is % vote for Obama in 2012 election (ObamaPer), and the explanatory variable is region of the country (Region) gives $R^2 = 0.36$.

Which of the following is true?

(a) The correlation between ObamaPer and Region is 0.36

(b) 36% of the variability in ObamaPer is explained by Region

(c) The correlation between ObamaPer and Region is $\sqrt{0.36}$

(d) $\sqrt{36\%}$ of the variability in ObamaPer is explained by Region

Categorical Variables

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \varepsilon_i$$

- For this to make any sense, each x value has to be a number.
- How do we include categorical variables in a regression setting?

Categorical Variables

- Take one categorical variable, and replace it with several "dummy" variables.
- A ***dummy variable*** is 1 if the case falls into the category represented by the dummy variable, and 0 otherwise.
- Create one dummy variable for each category of the categorical variable.

Dummy Variables

dummy variables

State	Region	South	West	Northeast	Midwest
Alabama	South	1	0	0	0
Alaska	West	0	1	0	0
Arkansas	South	1	0	0	0
California	West	0	1	0	0
Colorado	West	0	1	0	0
Connecticut	Northeast	0	0	1	0
Delaware	Northeast	0	0	1	0
Florida	South	1	0	0	0
Georgia	South	1	0	0	0
Hawaii	West	0	1	0	0
…	…	…	…	…	…

Dummy Variables

- When using dummy variables, one has to be left out of the model
- The dummy variable left out is called the ***reference level***
- When using region of the country (Northeast, South, Midwest, West) to predict % Obama vote, how many dummy variables will be included?
- One b) Two c) Three d) Four

Dummy Variables

- Predicting % vote for Obama with one categorical variable: region of the country
- If "midwest" is the reference level:

$$\%\ \text{Obama vote} = \beta_0 + \beta_1 Northeast + \beta_2 South + \beta_3 West + \varepsilon$$

Predicted percentage vote for midwest state

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.46971    0.02383  19.710  < 2e-16 ***
RegionNE      0.13085    0.03520   3.717 0.000545 ***
RegionS      -0.04283    0.03370  -1.271 0.210139
RegionW       0.01249    0.03370   0.371 0.712699
```

Increase in vote for a West state, compared to a Midwest state

Voting by Region

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.46971    0.02383  19.710  < 2e-16 ***
RegionNE     0.13085    0.03520   3.717 0.000545 ***
RegionS     -0.04283    0.03370  -1.271 0.210139
RegionW      0.01249    0.03370   0.371 0.712699
```

Based on the output above, which region had the highest percent vote for Obama?

a) Midwest
b) Northeast
c) South
d) West

Voting by Region

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.46971    0.02383  19.710  < 2e-16 ***
RegionNE     0.13085    0.03520   3.717 0.000545 ***
RegionS     -0.04283    0.03370  -1.271 0.210139
RegionW      0.01249    0.03370   0.371 0.712699
```

What is the predicted % Obama vote for a state in the northeast?

a) 13%
b) 47%
c) 55%
d) 60%

Voting by Region

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.46971    0.02383  19.710  < 2e-16 ***
RegionNE     0.13085    0.03520   3.717 0.000545 ***
RegionS     -0.04283    0.03370  -1.271 0.210139
RegionW      0.01249    0.03370   0.371 0.712699
```

What is the predicted % Obama vote for a state in the midwest?

a) 50%
b) 47%
c) 0%
d) 45%

Categorical Variables

ANOVA for Regression:

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.46971    0.02383  19.710  < 2e-16 ***
RegionNE     0.13085    0.03520   3.717 0.000545 ***
RegionS     -0.04283    0.03370  -1.271 0.210139
RegionW      0.01249    0.03370   0.371 0.712699
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 '

Residual standard error: 0.08592 on 46 degrees of freedom
Multiple R-squared: 0.3616,     Adjusted R-squared:  0.32
F-statistic: 8.685 on 3 and 46 DF,  p-value: 0.0001126
```

ANOVA for Difference in Means:

```
            Df Sum Sq Mean Sq F value   Pr(>F)
Region       3 0.1924 0.06412   8.685 0.000113 ***
Residuals   46 0.3396 0.00738
```

p-values

Do p-values make sense to use here?

a) Yes
b) No

Categorical Variables in R

- R automatically creates dummy variables for you if you include a categorical explanatory variable.
- The first level alphabetically is usually the reference level.

Categorical Variables

- Either all dummy variables associated with a categorical variable have to be included in the model, or none of them.
- RegionS and RegionW are not significant, but leaving them out would clump the South and the West with the reference level, Midwest, which does not make sense.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.46971    0.02383  19.710  < 2e-16 ***
RegionNE     0.13085    0.03520   3.717 0.000545 ***
RegionS     -0.04283    0.03370  -1.271 0.210139
RegionW      0.01249    0.03370   0.371 0.712699
```

Full Regression Model

```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       2.940e-01  7.708e-01   0.381  0.70546
HouseholdIncome  -2.605e-07  2.310e-06  -0.113  0.91093
IQ               -7.944e-03  1.374e-02  -0.578  0.56712
RegionNE          4.545e-02  3.622e-02   1.255  0.21867
RegionS          -8.204e-02  2.947e-02  -2.784  0.00894 **
RegionW          -7.738e-02  3.840e-02  -2.015  0.05238 .
Population        1.290e-03  2.114e-03   0.610  0.54586
EighthGradeMath   4.862e-04  3.646e-03   0.133  0.89476
HighSchool       -5.364e-03  4.198e-03  -1.278  0.21050
GSP              -2.175e-06  1.705e-06  -1.275  0.21132
FiveVegetables    3.828e-03  5.079e-03   0.754  0.45659
Smokers           1.046e-02  4.419e-03   2.368  0.02409 *
PhysicalActivity  9.882e-03  4.895e-03   2.019  0.05195 .
Obese            -2.056e-03  5.841e-03  -0.352  0.72712
College           8.277e-03  4.637e-03   1.785  0.08374 .
NonWhite          2.415e-03  1.599e-03   1.510  0.14074
HeavyDrinkers     2.270e-02  8.986e-03   2.526  0.01668 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.05165 on 32 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared: 0.822,      Adjusted R-squared: 0.733
F-statistic: 9.237 on 16 and 32 DF,  p-value: 6.909e-08
```

West Region

- With only region as an explanatory variable, interpret the positive coefficient of RegionW.

In this data set, states in the West voted more for Obama than states in the Midwest.

- With all the other explanatory variables included, interpret the negative coefficient of RegionW.

States in the West voted less for Obama than would be expected based on the other variables in the model, as compared to states in the Midwest.

Smoking

Smokers	1.046e-02	4.419e-03	2.368	0.02409 *

Given all the other variables in the model, states with a higher percentage of smokers are more likely to vote

(a) Republican
(b) Democratic
(c) Impossible to tell

Smoking

Smokers	1.046e-02	4.419e-03	2.368	0.02409 *

The correlation between percent of people smoking in a state and the percent of people voting for Obama in 2012 was

(a) Positive
(b) Negative
(c) Impossible to tell

Smokers

- If smoking was banned in a state, the percentage of smokers would most likely decrease.

- In that case, the percentage voting Democratic would…
  - (a) increase
  - (b) decrease
  - (c) impossible to tell

Causation

- A significant explanatory variable in a regression model indicates association, but not necessarily causation

- CAUSALITY CAN ONLY BE INFERRED FROM A RANDOMIZED EXPERIMENT!!!!

Goal of the Model?

- If the goal of the model is to see what and how each variable is associated with a state's voting patterns, given all the other variables in the model, then we are done.
- If the goal is to predict the % of the vote that will be for the democrat, say in the 2016 election, we want to prune out insignificant variables to improve the model.

Over-fitting

- It is possible to ***over-fit*** a model: to include too many explanatory variables.
- The fewer the coefficients being estimated, the better they will be estimated.
- Usually, a good model has pruned out explanatory variables that are not helping.

R^2

- Adding more explanatory variables will only make R^2 increase or stay the same
- Adding another explanatory variable can not make the model explain less, because the other variables are all still in the model
- Is the best model always the one with the highest proportion of variability explained, and so the highest R^2?
    - Yes (b) No

Adjusted R^2

- ***Adjusted R^2*** is like R^2, but takes into account the number of explanatory variables
- As the number of explanatory variables increases, adjusted R^2 gets smaller than R^2
- One way to choose a model is to choose the model with the highest adjusted R^2

Adjusted R^2

The formula for the adjusted R^2 is

$$R^2_{\text{adj}} = 1 - \frac{\left(1 - R^2\right)\left(n - 1\right)}{n - k - 1}$$

$$= 1 - \frac{SSE / (n - k - 1)}{SST / (n - 1)}$$

Adjusted R^2

```
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       5.127e-01  8.152e-01   0.629  0.53383
HouseholdIncome   3.433e-07  2.443e-06   0.141  0.88913
IQ               -9.714e-04  1.453e-02  -0.067  0.94710
RegionNE         -3.623e-02  3.830e-02  -0.946  0.35135
RegionS           7.821e-02  3.116e-02   2.510  0.01733 *
RegionW           7.579e-02  4.061e-02   1.866  0.07119 .
Population       -2.276e-03  2.235e-03  -1.018  0.31619
X8thGradeMath     3.355e-03  3.856e-03   0.870  0.39072
HighSchool        7.909e-03  4.439e-03   1.782  0.08431 .
GSP               1.844e-06  1.803e-06   1.022  0.31434
FiveVegetables   -2.572e-03  5.371e-03  -0.479  0.63530
Smokers          -9.291e-03  4.673e-03  -1.988  0.05541 .
PhysicalActivity -1.461e-02  5.176e-03  -2.822  0.00814 **
Obese             1.542e-03  6.177e-03   0.250  0.80451
College          -6.655e-03  4.904e-03  -1.357  0.18425
NonWhite         -1.098e-03  1.691e-03  -0.649  0.52077
HeavyDrinkers    -1.757e-02  9.502e-03  -1.849  0.07378 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.05462 on 32 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared: 0.7545,    Adjusted R-squared: 0.6317
F-statistic: 6.146 on 16 and 32 DF,  p-value: 6.835e-06
```

You now know how to interpret all of these numbers!

Model Output

```
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      5.127e-01  8.152e-01   0.629  0.53383
HouseholdIncome  3.433e-07  2.443e-06   0.141  0.88913
IQ              -9.714e-04  1.453e-02  -0.067  0.94710
RegionNE        -3.623e-02  3.830e-02  -0.946  0.35135
RegionS          7.821e-02  3.116e-02   2.510  0.01733 *
RegionW          7.579e-02  4.061e-02   1.866  0.07119 .
Population      -2.276e-03  2.235e-03  -1.018  0.31619
X8thGradeMath    3.355e-03  3.856e-03   0.870  0.39072
HighSchool       7.909e-03  4.439e-03   1.782  0.08431 .
GSP              1.844e-06  1.803e-06   1.022  0.31434
FiveVegetables  -2.572e-03  5.371e-03  -0.479  0.63530
Smokers         -9.291e-03  4.673e-03  -1.988  0.05541 .
PhysicalActivity -1.461e-02  5.176e-03  -2.822  0.00814 **
Obese            1.542e-03  6.177e-03   0.250  0.80451
College         -6.655e-03  4.904e-03  -1.357  0.18425
NonWhite        -1.098e-03  1.691e-03  -0.649  0.52077
HeavyDrinkers   -1.757e-02  9.502e-03  -1.849  0.07378 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.05462 on 32 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared: 0.7545,     Adjusted R-squared: 0.6317
F-statistic: 6.146 on 16 and 32 DF,  p-value: 6.835e-06
```

Variable selection

- (Some) ways of deciding whether a variable should be included in the model or not:

1. Does it improve adjusted R^2?
2. Does it have a low p-value?
3. Is it associated with the response by itself?
4. Is it strongly associated with another explanatory variables? (If yes, then including both may be redundant)
5. Does common sense say it should contribute to the model?

Variable selection

- The p-value for an explanatory variable can be taken as a rough measure for how helpful that explanatory variable is to the model
- Insignificant variables may be pruned from the model, as long as adjusted R2 doesn’t go down too much
- You can also look at relationships between explanatory variables; if two are strongly associated, perhaps both are not necessary.

Stepwise Regression

- We could go through and think hard about which variables to include, or we could automate the process.
- ***Stepwise regression*** drops insignificant variables one by one.
- This is particularly useful if you have many potential explanatory variables/

Full Model

```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       2.940e-01  7.708e-01   0.381  0.70546
HouseholdIncome  -2.605e-07  2.310e-06  -0.113  0.91093
IQ               -7.944e-03  1.374e-02  -0.578  0.56712
RegionNE          4.545e-02  3.622e-02   1.255  0.21867
RegionS          -8.204e-02  2.947e-02  -2.784  0.00894 **
RegionW          -7.738e-02  3.840e-02  -2.015  0.05238 .
Population        1.290e-03  2.114e-03   0.610  0.54586
EighthGradeMath   4.862e-04  3.646e-03   0.133  0.89476
HighSchool       -5.364e-03  4.198e-03  -1.278  0.21050
GSP              -2.175e-06  1.705e-06  -1.275  0.21132
FiveVegetables    3.828e-03  5.079e-03   0.754  0.45659
Smokers           1.046e-02  4.419e-03   2.368  0.02409 *
PhysicalActivity  9.882e-03  4.895e-03   2.019  0.05195 .
Obese            -2.056e-03  5.841e-03  -0.352  0.72712
College           8.277e-03  4.637e-03   1.785  0.08374 .
NonWhite          2.415e-03  1.599e-03   1.510  0.14074
HeavyDrinkers     2.270e-02  8.986e-03   2.526  0.01668 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.05165 on 32 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared: 0.822,      Adjusted R-squared:  0.733
F-statistic: 9.237 on 16 and 32 DF,  p-value: 6.909e-08
```

Highest p-value

Pruned Model 1

```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       2.697e-01  7.378e-01   0.366  0.71705
HouseholdIncome  -2.883e-07  2.266e-06  -0.127  0.89953
IQ               -6.459e-03  7.916e-03  -0.816  0.42043
RegionNE          4.440e-02  3.483e-02   1.275  0.21129
RegionS          -8.205e-02  2.903e-02  -2.827  0.00793 **
RegionW          -7.682e-02  3.760e-02  -2.043  0.04910 *
Population        1.282e-03  2.081e-03   0.616  0.54219
HighSchool       -5.458e-03  4.077e-03  -1.339  0.18981
GSP              -2.138e-06  1.657e-06  -1.290  0.20597
FiveVegetables    3.864e-03  4.996e-03   0.773  0.44481
Smokers           1.043e-02  4.347e-03   2.400  0.02218 *
PhysicalActivity  9.967e-03  4.781e-03   2.085  0.04492 *
Obese            -1.778e-03  5.374e-03  -0.331  0.74284
College           8.365e-03  4.522e-03   1.850  0.07329 .
NonWhite          2.471e-03  1.521e-03   1.624  0.11379
HeavyDrinkers     2.293e-02  8.686e-03   2.639  0.01258 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.05088 on 33 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared: 0.8219,     Adjusted R-squared: 0.741
F-statistic: 10.15 on 15 and 33 DF,  p-value: 2.114e-08
```

Highest p-value

Pruned Model 2

```
Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)        2.700e-01  7.270e-01   0.371  0.71270
IQ                -6.312e-03  7.718e-03  -0.818  0.41914
RegionNE           4.397e-02  3.416e-02   1.287  0.20672
RegionS           -8.129e-02  2.799e-02  -2.904  0.00643 **
RegionW           -7.629e-02  3.682e-02  -2.072  0.04594 *
Population         1.267e-03  2.047e-03   0.619  0.54020
HighSchool        -5.621e-03  3.813e-03  -1.474  0.14968
GSP               -2.218e-06  1.509e-06  -1.470  0.15080
FiveVegetables     3.688e-03  4.732e-03   0.779  0.44111
Smokers            1.051e-02  4.242e-03   2.478  0.01835 *
PhysicalActivity   9.893e-03  4.676e-03   2.115  0.04180 *
Obese             -1.830e-03  5.280e-03  -0.347  0.73107
College            8.269e-03  4.394e-03   1.882  0.06841 .
NonWhite           2.457e-03  1.495e-03   1.643  0.10950
HeavyDrinkers      2.318e-02  8.335e-03   2.781  0.00878 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.05014 on 34 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared: 0.8218,     Adjusted R-squared:  0.7485
F-statistic:  11.2 on 14 and 34 DF,  p-value: 6.151e-09
```

Highest p-value

Pruned Model 3

```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       1.873e-01  6.781e-01   0.276  0.78396
IQ               -5.856e-03  7.509e-03  -0.780  0.44072
RegionNE          4.629e-02  3.308e-02   1.399  0.17054
RegionS          -8.262e-02  2.737e-02  -3.019  0.00471 **
RegionW          -6.972e-02  3.117e-02  -2.237  0.03178 *
Population        1.328e-03  2.014e-03   0.659  0.51399
HighSchool       -5.750e-03  3.747e-03  -1.534  0.13390
GSP              -2.366e-06  1.430e-06  -1.654  0.10701
FiveVegetables    3.507e-03  4.643e-03   0.755  0.45512
Smokers           1.043e-02  4.183e-03   2.495  0.01748 *
PhysicalActivity  9.733e-03  4.595e-03   2.118  0.04132 *
College           8.936e-03  3.899e-03   2.292  0.02804 *
NonWhite          2.442e-03  1.476e-03   1.655  0.10683
HeavyDrinkers     2.408e-02  7.815e-03   3.082  0.00400 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.0495 on 35 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared: 0.8212,     Adjusted R-squared: 0.7548
F-statistic: 12.37 on 13 and 35 DF,  p-value: 1.786e-09
```

Highest p-value

Pruned Model 4

```
Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      9.266e-02  6.575e-01   0.141  0.88872
IQ              -3.818e-03  6.789e-03  -0.562  0.57736
RegionNE         3.708e-02  2.975e-02   1.246  0.22070
RegionS         -8.831e-02  2.578e-02  -3.426  0.00155 **
RegionW         -7.717e-02  2.882e-02  -2.677  0.01110 *
HighSchool      -6.422e-03  3.577e-03  -1.795  0.08102 .
GSP             -2.115e-06  1.368e-06  -1.546  0.13077
FiveVegetables   5.265e-03  3.772e-03   1.396  0.17126
Smokers          9.381e-03  3.835e-03   2.446  0.01947 *
PhysicalActivity 9.143e-03  4.471e-03   2.045  0.04823 *
College          7.789e-03  3.462e-03   2.250  0.03066 *
NonWhite         3.012e-03  1.186e-03   2.539  0.01559 *
HeavyDrinkers    2.428e-02  7.748e-03   3.133  0.00343 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.04911 on 36 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared: 0.819,      Adjusted R-squared: 0.7586
F-statistic: 13.57 on 12 and 36 DF,  p-value: 5.727e-10
```

Highest p-value

Pruned Model 5

```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)     -2.025e-01  3.925e-01  -0.516  0.60901
RegionNE         3.830e-02  2.940e-02   1.303  0.20063
RegionS         -8.584e-02  2.516e-02  -3.411  0.00158 **
RegionW         -7.020e-02  2.578e-02  -2.723  0.00981 **
HighSchool      -7.027e-03  3.381e-03  -2.078  0.04466 *
GSP             -2.210e-06  1.344e-06  -1.644  0.10864
FiveVegetables   5.068e-03  3.721e-03   1.362  0.18134
Smokers          9.704e-03  3.757e-03   2.583  0.01389 *
PhysicalActivity 8.650e-03  4.344e-03   1.991  0.05387 .
College          7.494e-03  3.390e-03   2.211  0.03333 *
NonWhite         3.375e-03  9.865e-04   3.421  0.00154 **
HeavyDrinkers    2.474e-02  7.633e-03   3.241  0.00252 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.04866 on 37 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared: 0.8174,     Adjusted R-squared: 0.7631
F-statistic: 15.06 on 11 and 37 DF,  p-value: 1.65e-10
```

Highest p-value

Pruned Model 6

```
Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)      -1.864e-01  3.967e-01  -0.470 0.641094
RegionNE          5.326e-02  2.758e-02   1.931 0.060930 .
RegionS          -7.881e-02  2.490e-02  -3.165 0.003054 **
RegionW          -5.980e-02  2.490e-02  -2.402 0.021322 *
HighSchool       -7.492e-03  3.401e-03  -2.203 0.033739 *
GSP              -2.666e-06  1.317e-06  -2.025 0.049906 *
Smokers           1.044e-02  3.759e-03   2.778 0.008445 **
PhysicalActivity  9.382e-03  4.359e-03   2.153 0.037771 *
College           9.499e-03  3.089e-03   3.075 0.003883 **
NonWhite          3.357e-03  9.975e-04   3.366 0.001757 **
HeavyDrinkers     2.879e-02  7.109e-03   4.050 0.000244 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.0492 on 38 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared: 0.8082,     Adjusted R-squared: 0.7578
F-statistic: 16.02 on 10 and 38 DF,  p-value: 9.351e-11
```

Pruned Model 5

```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)     -2.025e-01  3.925e-01  -0.516  0.60901
RegionNE         3.830e-02  2.940e-02   1.303  0.20063
RegionS         -8.584e-02  2.516e-02  -3.411  0.00158 **
RegionW         -7.020e-02  2.578e-02  -2.723  0.00981 **
HighSchool      -7.027e-03  3.381e-03  -2.078  0.04466 *
GSP             -2.210e-06  1.344e-06  -1.644  0.10864
FiveVegetables   5.068e-03  3.721e-03   1.362  0.18134
Smokers          9.704e-03  3.757e-03   2.583  0.01389 *
PhysicalActivity 8.650e-03  4.344e-03   1.991  0.05387 .
College          7.494e-03  3.390e-03   2.211  0.03333 *
NonWhite         3.375e-03  9.865e-04   3.421  0.00154 **
HeavyDrinkers    2.474e-02  7.633e-03   3.241  0.00252 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.04866 on 37 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared: 0.8174,     Adjusted R-squared: 0.7631
F-statistic: 15.06 on 11 and 37 DF,  p-value: 1.65e-10
```

Pruned Model 7

```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)     -0.1173701  0.3976552  -0.295  0.76948
RegionNE         0.0320078  0.0297918   1.074  0.28943
RegionS         -0.0832146  0.0256689  -3.242  0.00247 **
RegionW         -0.0775162  0.0259578  -2.986  0.00492 **
HighSchool      -0.0074225  0.0034468  -2.153  0.03769 *
FiveVegetables   0.0065921  0.0036831   1.790  0.08146 .
Smokers          0.0076819  0.0036288   2.117  0.04087 *
PhysicalActivity 0.0081657  0.0044297   1.843  0.07308 .
College          0.0048813  0.0030609   1.595  0.11906
NonWhite         0.0030327  0.0009857   3.077  0.00387 **
HeavyDrinkers    0.0236143  0.0077707   3.039  0.00428 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.04973 on 38 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared: 0.8041,     Adjusted R-squared: 0.7525
F-statistic: 15.59 on 10 and 38 DF,  p-value: 1.382e-10
```

FINAL STEPWISE MODEL

```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)     -2.025e-01  3.925e-01  -0.516  0.60901
RegionNE         3.830e-02  2.940e-02   1.303  0.20063
RegionS         -8.584e-02  2.516e-02  -3.411  0.00158 **
RegionW         -7.020e-02  2.578e-02  -2.723  0.00981 **
HighSchool      -7.027e-03  3.381e-03  -2.078  0.04466 *
GSP             -2.210e-06  1.344e-06  -1.644  0.10864
FiveVegetables   5.068e-03  3.721e-03   1.362  0.18134
Smokers          9.704e-03  3.757e-03   2.583  0.01389 *
PhysicalActivity 8.650e-03  4.344e-03   1.991  0.05387 .
College          7.494e-03  3.390e-03   2.211  0.03333 *
NonWhite         3.375e-03  9.865e-04   3.421  0.00154 **
HeavyDrinkers    2.474e-02  7.633e-03   3.241  0.00252 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.04866 on 37 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared: 0.8174,     Adjusted R-squared: 0.7631
F-statistic: 15.06 on 11 and 37 DF,  p-value: 1.65e-10
```

Full Model

```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       2.940e-01  7.708e-01   0.381  0.70546
HouseholdIncome  -2.605e-07  2.310e-06  -0.113  0.91093
IQ               -7.944e-03  1.374e-02  -0.578  0.56712
RegionNE          4.545e-02  3.622e-02   1.255  0.21867
RegionS          -8.204e-02  2.947e-02  -2.784  0.00894 **
RegionW          -7.738e-02  3.840e-02  -2.015  0.05238 .
Population        1.290e-03  2.114e-03   0.610  0.54586
EighthGradeMath   4.862e-04  3.646e-03   0.133  0.89476
HighSchool       -5.364e-03  4.198e-03  -1.278  0.21050
GSP              -2.175e-06  1.705e-06  -1.275  0.21132
FiveVegetables    3.828e-03  5.079e-03   0.754  0.45659
Smokers           1.046e-02  4.419e-03   2.368  0.02409 *
PhysicalActivity  9.882e-03  4.895e-03   2.019  0.05195 .
Obese            -2.056e-03  5.841e-03  -0.352  0.72712
College           8.277e-03  4.637e-03   1.785  0.08374 .
NonWhite          2.415e-03  1.599e-03   1.510  0.14074
HeavyDrinkers     2.270e-02  8.986e-03   2.526  0.01668 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.05165 on 32 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared: 0.822,      Adjusted R-squared: 0.733
F-statistic: 9.237 on 16 and 32 DF,  p-value: 6.909e-08
```

Variation Selection

- There is no one “best” model.
- Choosing a model is just as much an art as a science.
- Adjusted R2 is just one possible criteria.