# Supervised2
# Linear regression

資料科學 Data Science
張家銘 Jia-Ming Chang
政治大學資訊科學系/NCCU.CS

The slide isonly for educational purposes. If any infringement, please contact me, we will correct immediately.

# Copyright declaration 版權說明

- Some of the figures in this presentation are taken from "An Introduction to Statistical Learning, with applications in R" (Springer, 2013) with permission from the authors: G. James, D. Witten, T. Hastie and R. Tibshirani.
- The web site of the book
- The credit of individual is indicated in the bottom part of the slide.
  - ie.,

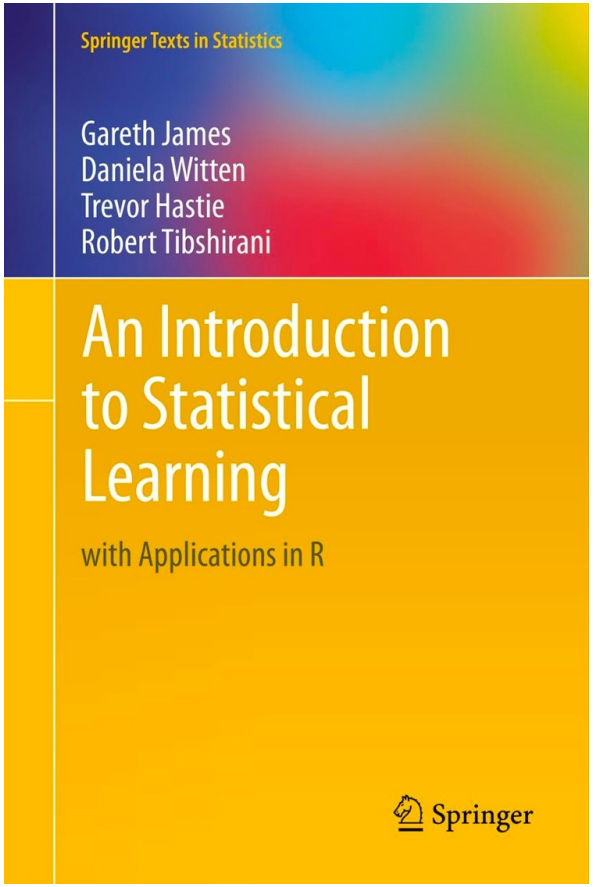



Figure 3.18, *An Introduction to Statistical Learning with Applications in R,* by Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani

# Copyright declaration 版權說明

- Some of the figures in this presentation are taken from "Practical Data Science with R (Manning, 2019)"
- The web site of the book
- The credit of individual is indicated in the bottom part of the slide.
  - ie.,

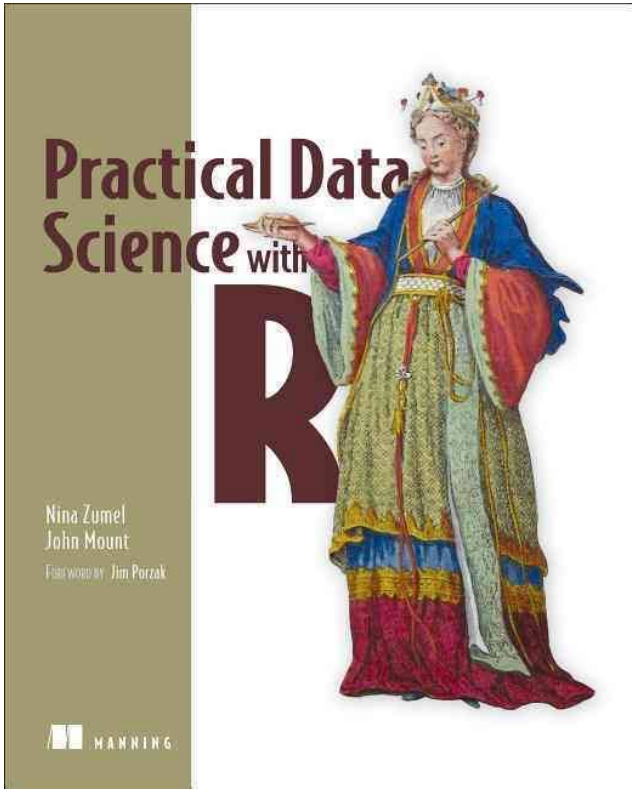
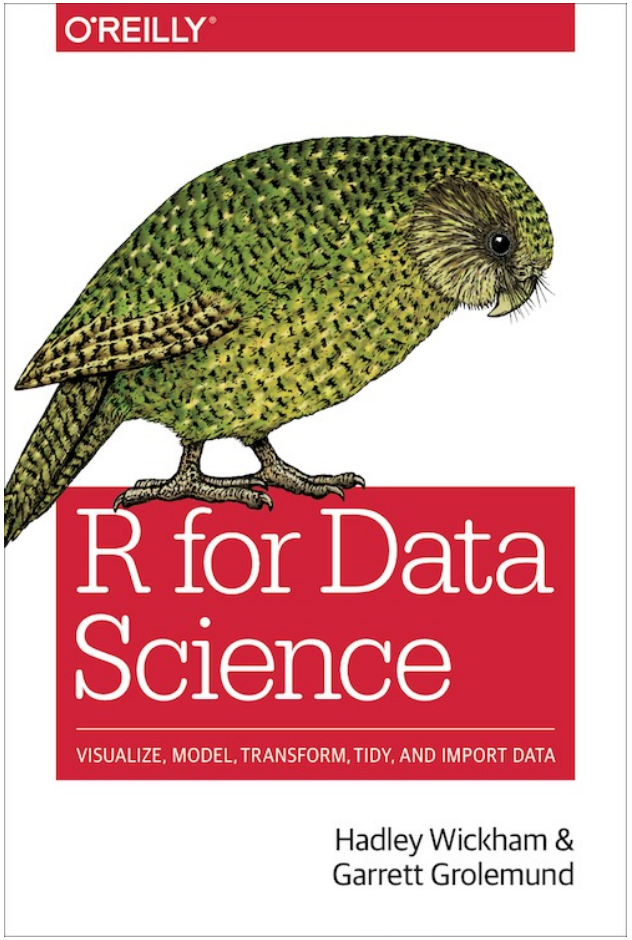



Figure 7.6, *Practical Data Science with R* by Nina Zumel and John Mount

# Copyright declaration 版權說明

- Some of the figures in this presentation are taken from "R for Data Science" under the Creative Commons Attribution-NonCommercial-NoDerivs 3.0 License.
- The web site of the book
- The credit of individual is indicated in the bottom part.
  - ie.,

*R for Data Science* by Garrett Grolemund, Hadley Wickham



# Recap for the last week

# Memorization methods

- Building single-variable models
- Cross-validated variable selection
- Building basic multivariable models
  - nearest neighbor
  - naive Bayes

# Building single-variable models

Using categorical features

Using numeric features

# Building single-variable models - Using categorical features

# Take Var218 as example

```
pPos <- sum(outCol==pos)/length(outCol)

naTab <- table(as.factor(outCol[is.na(varCol)]))

pPosWna <- (naTab/sum(naTab))[pos]

vTab <- table(as.factor(outCol),varCol)

pPosWv <- (vTab[pos,]+1.0e-
3*pPos)/(colSums(vTab)+1.0e-3)

pred <- pPosWv[appCol]

pred[is.na(appCol)] <- pPosWna

pred[is.na(pred)] <- pPos
```

# Function to build single-variable models for categorical variables

```
mkPredC <- function(outCol,varCol,appCol) {
   pPos <- sum(outCol==pos)/length(outCol)
   naTab <- table(as.factor(outCol[is.na(varCol)]))
   pPosWna <- (naTab/sum(naTab))[pos]
   vTab <- table(as.factor(outCol),varCol)
   pPosWv <- (vTab[pos,]+1.0e-
   3*pPos)/(colSums(vTab)+1.0e-3)
   pred <- pPosWv[appCol]
   pred[is.na(appCol)] <- pPosWna
   pred[is.na(pred)] <- pPos
   pred
}
```

# Building single-variable models - Using numeric features

# Take Var7 as example

```
v<-"Var7"
outCol<-dTrain[,outcome]
varCol<-dTrain[,v]
appCol<-dTrain[,v]
cuts <- unique(as.numeric(quantile(varCol,
probs=seq(0, 1, 0.1),na.rm=T)))
varC <- cut(varCol,cuts)
appC <- cut(appCol,cuts)
mkPredC(outCol,varC,appC)
```

# Using cross-validation to estimate effects of overfitting

```
var <- 'Var217'

aucs <- rep(0,100)

for(rep in 1:length(aucs)) {
    useForCalRep <-
    rbinom(n=dim(dTrainAll)[[1]],size=1,prob=0.1)>0
    predRep <- mkPredC(dTrainAll[!useForCalRep,outcome],
    dTrainAll[!useForCalRep,var],
    dTrainAll[useForCalRep,var])
    aucs[rep] <-
    calcAUC(predRep,dTrainAll[useForCalRep,outcome])
}

mean(aucs)

sd(aucs)
```

# Using cross-validation to estimate effects of overfitting

- without `for()` loop

```
fCross <- function() {
    useForCalRep <-
    rbinom(n=dim(dTrainAll)[[1]],size=1,prob=0.1)>0
    predRep <- mkPredC(dTrainAll[!useForCalRep,outcome],
    dTrainAll[!useForCalRep,var],
    dTrainAll[useForCalRep,var])
    calcAUC(predRep,dTrainAll[useForCalRep,outcome])
}
aucs <- replicate(100,fCross())
```

# Building models using many variables

*k*-nearest neighbor

Naive Bayes

decision trees :  Supervised learning  - random forest

# Building models using many variables - Naive Bayes

# Bayes' law

$$\mathrm{P}(y==T \mid ev_1) = \frac{\mathrm{P}(y==T) \times \mathrm{P}(ev_1 \mid y==T)}{\mathrm{P}(ev_1)}$$

$$\mathrm{P}(y==F \mid ev_1) = \frac{\mathrm{P}(y==F) \times (\mathrm{P}(ev_1 \mid y==F)}{\mathrm{P}(ev_1)}$$

# Naive Bayes assumption

- $\mathrm{P}(y==T|\text{evidence}) + \mathrm{P}(y==F|\text{evidence}) = 1$

$$\mathrm{P}(y==T \mid ev_1 \& \cdots ev_N) \approx \frac{\mathrm{P}(y==T) \times (\mathrm{P}(ev_1 \mid y==T) \times \cdots \mathrm{P}(ev_N \mid y==T))}{\mathrm{P}(ev_1 \& \cdots ev_N)}$$

$$\mathrm{P}(y==F \mid ev_1 \& \cdots ev_N) \approx \frac{\mathrm{P}(y==F) \times (\mathrm{P}(ev_1 \mid y==F) \times \cdots \mathrm{P}(ev_N \mid y==F))}{\mathrm{P}(ev_1 \& \cdots ev_N)}$$

# Naive Bayes

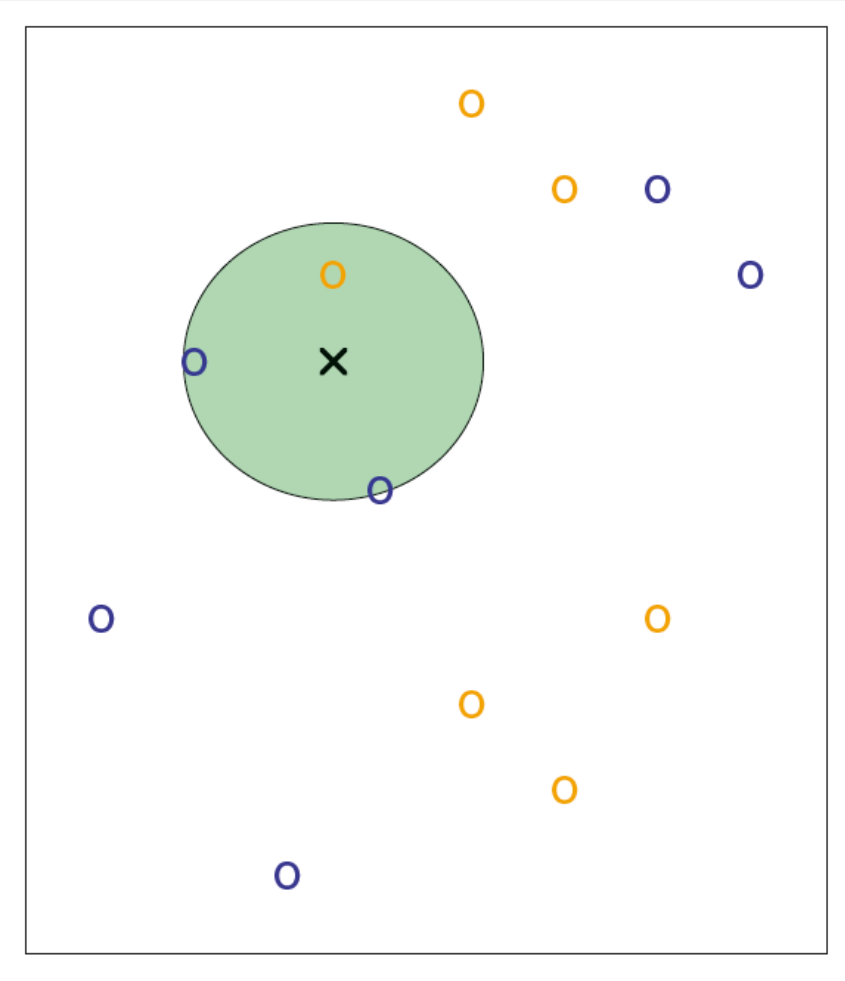
- For numerical reasons, it's better to convert the products into sums,

$$\text{score}(T|\, ev_1 \& \cdots\, ev_N) = log(\text{P}(y==T)) + log(\text{P}(ev_1 \,|\, y==T)) + \cdots\, log(\text{P}(ev_N \,|\, y==T))$$

$$\text{score}(F|\, ev_1 \& \cdots\, ev_N) = log(\text{P}(y==F)) + log(\text{P}(ev_1 \,|\, y==F)) + \cdots\, log(\text{P}(ev_N \,|\, y==F))$$

# Building models using many variables - *k*-nearest neighbor

# K=3



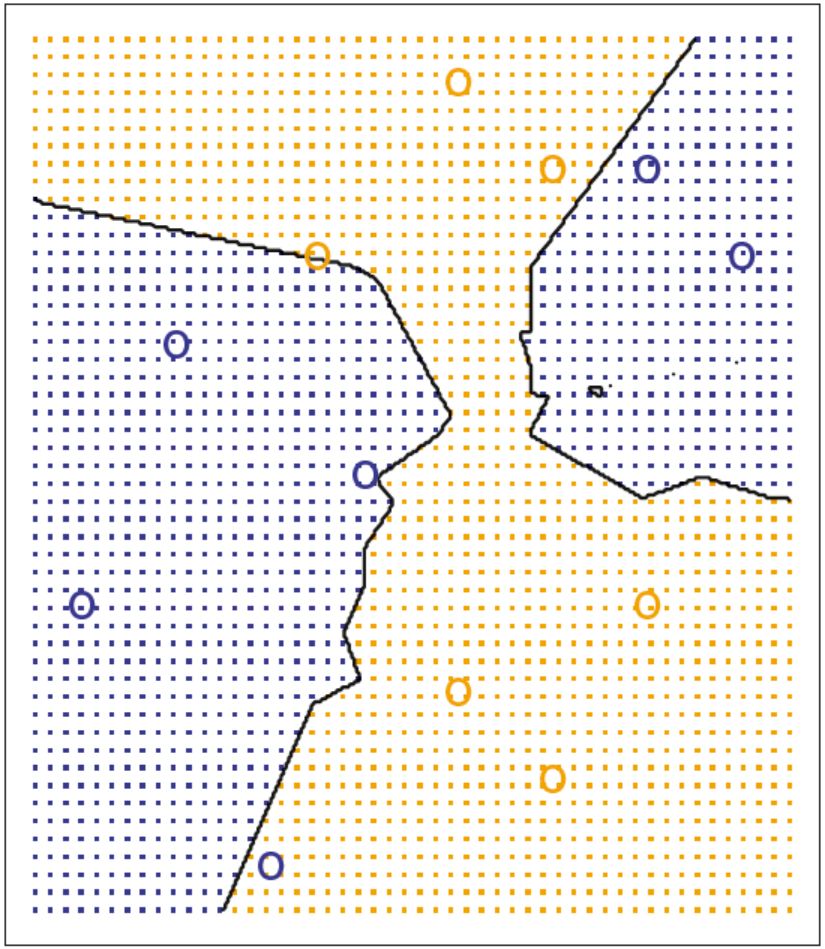


Figure2.14, *An Introduction to Statistical Learning with Applications in R* by Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani

# Simulation data by $K$=10

- error rate
  - KNN:0. 1363
  - Bayes: 0. 1304

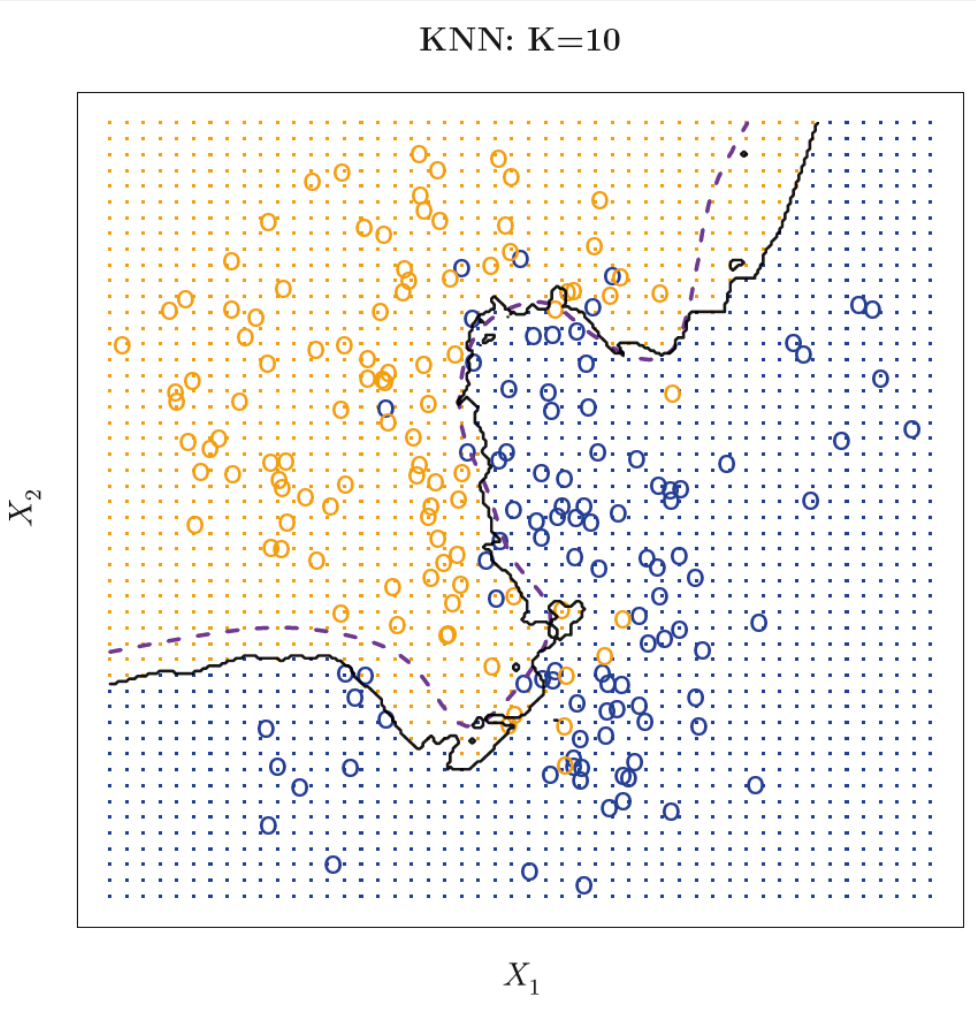
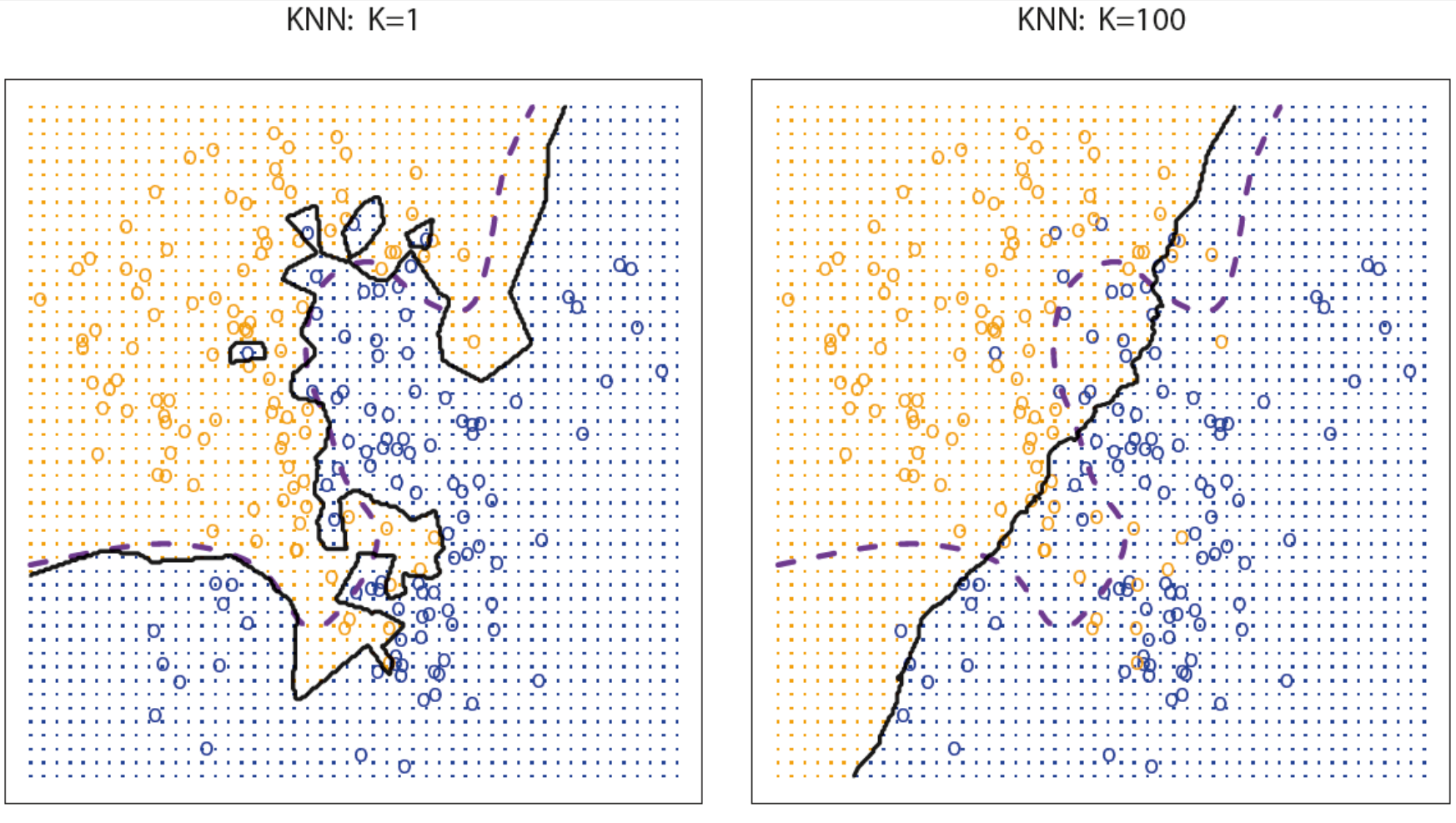



Figure2.15, *An Introduction to Statistical Learning with Applications in R* by Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani

# The bigger *K*, the better?




low-bias but very high-variance

low-variance but high-bias

Figure2.16, *An Introduction to Statistical Learning with Applications in R* by Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani6

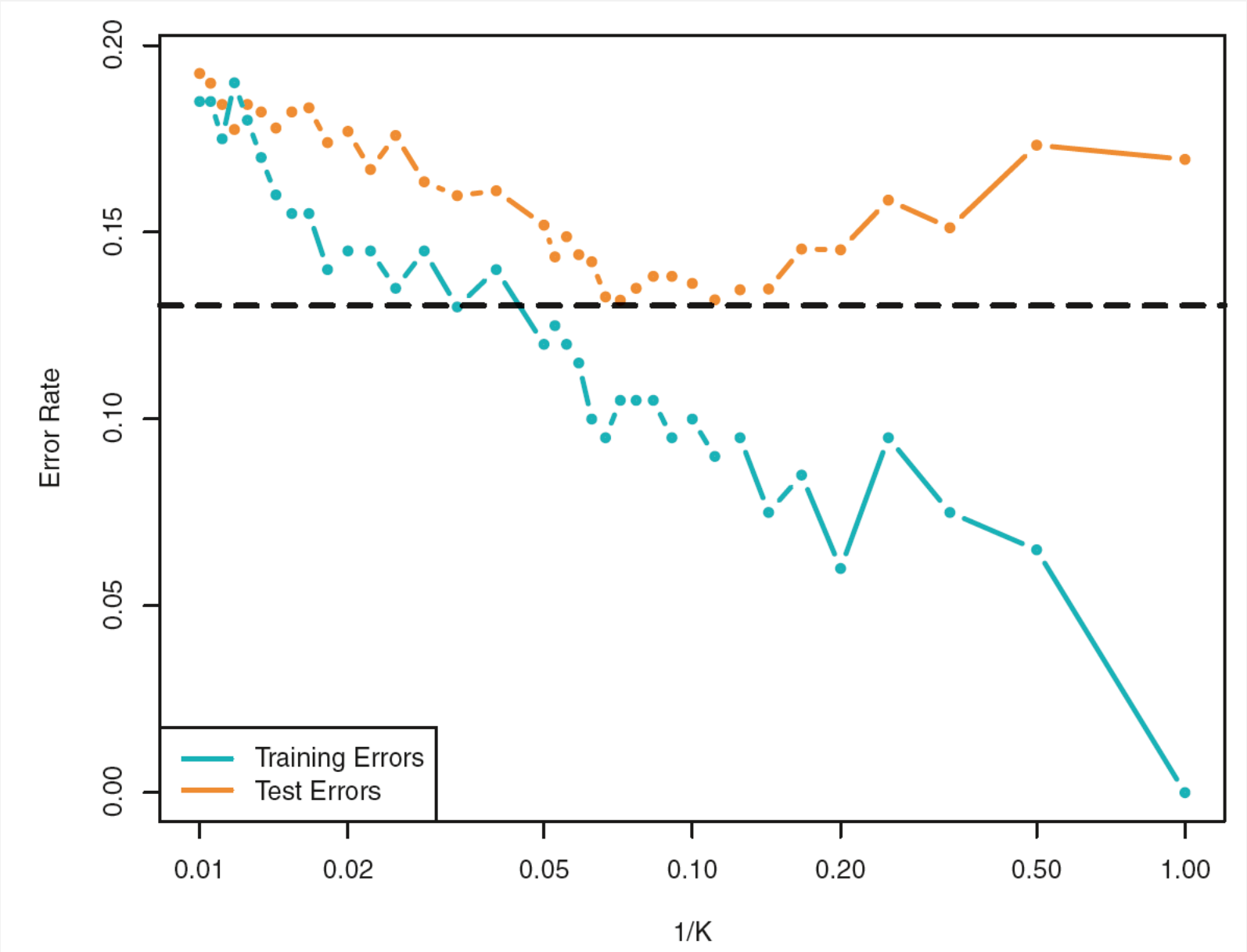



Figure2.17, *An Introduction to Statistical Learning with Applications in R* by Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani

# Using nearest neighbor methods

- events with unbalanced outcomes (that probabilities not near 50%)
  - using a large *k* so KNN can express a useful range of probabilities
  - have a good chance of seeing 10 positive examples in each neighborhood
  - 10/0.07 = 142

# Single-variable model

- Single-variable models can be thought of as being simple summaries of the training data (categorical variables)
- => the model is essentially a contingency table or pivot table

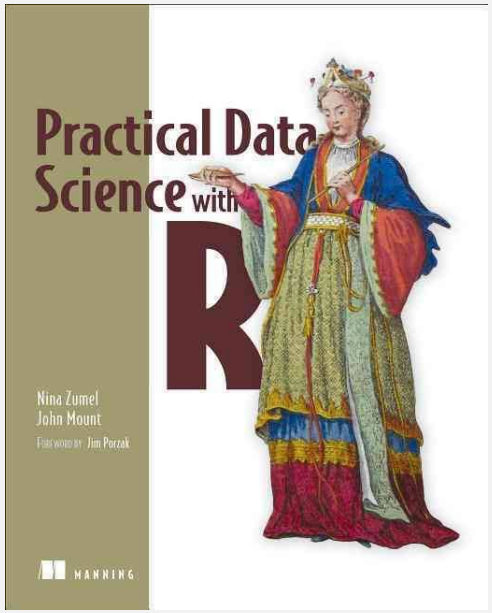
# Naive Bayes

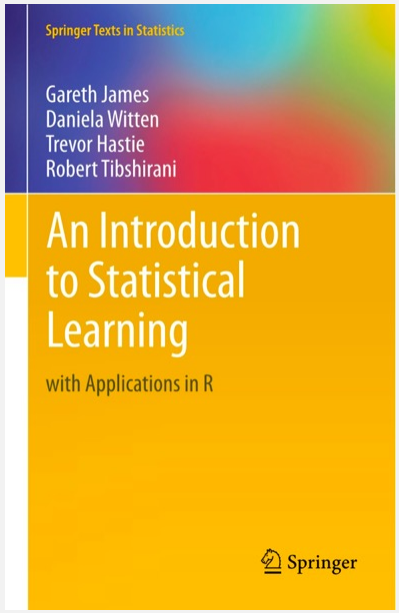
- form their decision by building a large collection of independent single-variable models
- Prediction for a given example is just the product of all the applicable single variable model adjustments => sums of appropriate summaries of the original training data.
- Naive Bayes doesn't perform any clever optimization, so it can be outperformed by methods like logistic regression and support vector machines.

# *K*-nearest neighbor

- summaries of the *k* pieces of training data that are closest to the example to be scored.
- *KNN* models usually store all of their original training data instead of an efficient summary
- => they truly do memorize the training data

# Chp. 7: Linear and logistic regression




3. Linear regression

# Linear and logistic regression

- Extracting relations and advice from functional models
- linear regression
  - predict quantities
  - Interpreting the diagnostics from `lm` call
- logistic regression
  - predict probabilities or categories
  - Interpreting the diagnostics from `glm` call

# Linear and logistic regression

- This class of methods is especially useful when you don't just want to predict an outcome, but you also want to know the relationship between the input variables and the outcome.
- This knowledge can prove useful because this relationship can often be used as advice on how to get the outcome that you want.

# Linear regression

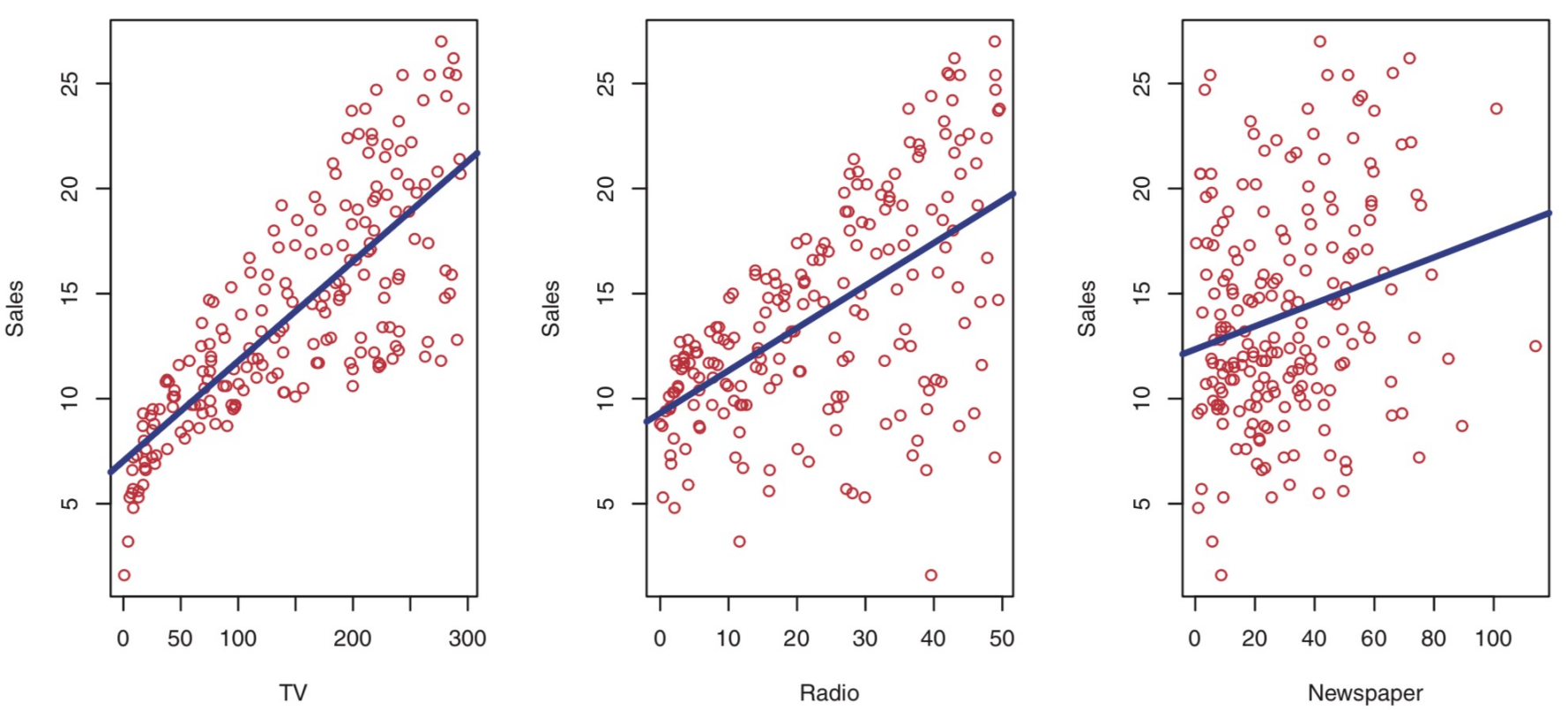



Figure 2.1, p16, An Introduction to Statistical Learning, with applications in R" (Springer, 2013) with permission from the authors: G. James, D. Witten, T. Hastie and R. Tibshirani

- Is there a relationship between advertising budget and sales?
- How strong is the relationship between advertising budget and sales?
- Which media contribute to sales?
- How accurately can we estimate the effect of each medium on sales?
- How accurately can we predict future sales?
- Is the relationship linear?
- Is there synergy (*interaction*) among the advertising media?
  - $50,000 on television + $50,000 on radio >= $100,000 to either television or radio individually?

# How linear regression is *used in the field*?

- we're using a linear model to predict something that is itself not linear.

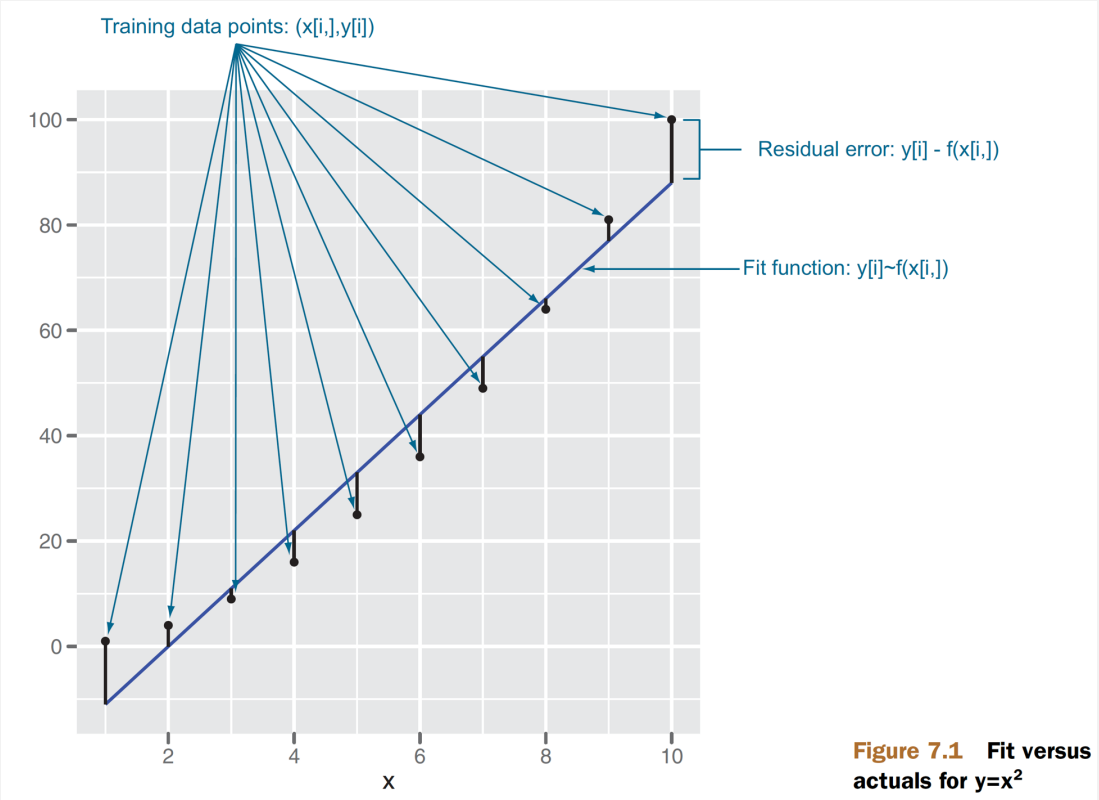



**Figure 7.1** **Fit versus actuals for $y=x^2$**

Figure7.1, p143, Practical Data Science with R by Nina Zumel, John Mount

# Simple Linear Regression

- predicting a quantitative response $Y$ on the basis of a single predictor variable $X$
- $Y \approx \beta_0 + \beta_1 X$
  - regressing *Y* on *X*
  - *Y* onto *X*
  - $\beta_0$: intercept
  - $\beta_1$: slope
  - $\beta_0, \beta_1$: coefficients or parameters
- $\text{sales} \approx \beta_0 + \beta_1 \text{TV}$

# Simple Linear Regression

- $\hat{y} \approx \widehat{\beta_0} + \widehat{\beta_1} x$
  - a prediction of *Y* on the basis of *X* = *x*
  - ˆ denote the estimated value for an unknown parameter or coefficient, or to denote the predicted value of the response

# Linear regression

- for every person *i*, we want to predict `pounds.lost[i]` based on `daily.cals[i]` and `daily.exercise[i]`
- `pounds.lost[i] = b.cals * daily.cals[i] + b.exercise * daily.exercise[i]`
  - *pounds.lost* : dependent, response variable
  - *daily.cals*, *daily.exercise* : independent, explanatory variables
  - *b.cals*, *b.exercise* : coefficients, betas

# Estimating the Coefficients

- `y[i] ~ f(x[i,]) = b[1] x[i,1] + b[2] x[i,2] + ... b[n] x[i,n] + e[i]`
  - We want numbers b[1],...,b[n] (called the coefficients or betas ) such that f(x[i,]) is as near as possible to y[i] for all (x[i,],y[i]) pairs in our training data.
  - e[i] : unsystematic errors
    - average to 0
    - uncorrelated with x[i,] and y[i] .

# Estimating the Coefficients

- $n$ observation pairs : $(x_1,y_1),(x_2,y_2), \ldots,(x_n,y_n)$
- $y_i \approx \widehat{\beta_0} + \widehat{\beta_1} x_i$ for $i = 1, \ldots, n$
  - obtain coefficient estimates $\widehat{\beta_0}$ and $\widehat{\beta_1}$ such that the linear model fits the available data
  - find an intercept $\widehat{\beta_0}$ and a slope $\widehat{\beta_1}$ such that the resulting line is as close as possible to the n = 200 data points
- measuring closeness : Residual sum of squares

# Residual sum of squares (RSS)

- Residual: $e_i = y_i - \hat{y}_i$
- Residual sum of squares (RSS)
  - $e_1^2 + e_2^2 + \ldots + e_n^2$
  - $\left(y_1 - \left(\widehat{\beta_0} + \widehat{\beta_1} x_1\right)\right)^2 + \left(y_2 - \left(\widehat{\beta_0} + \widehat{\beta_1} x_2\right)\right)^2 + \cdots + \left(y_n - \left(\widehat{\beta_0} + \widehat{\beta_1} x_n\right)\right)^2$

# the least squares fit for the regression of sales onto TV

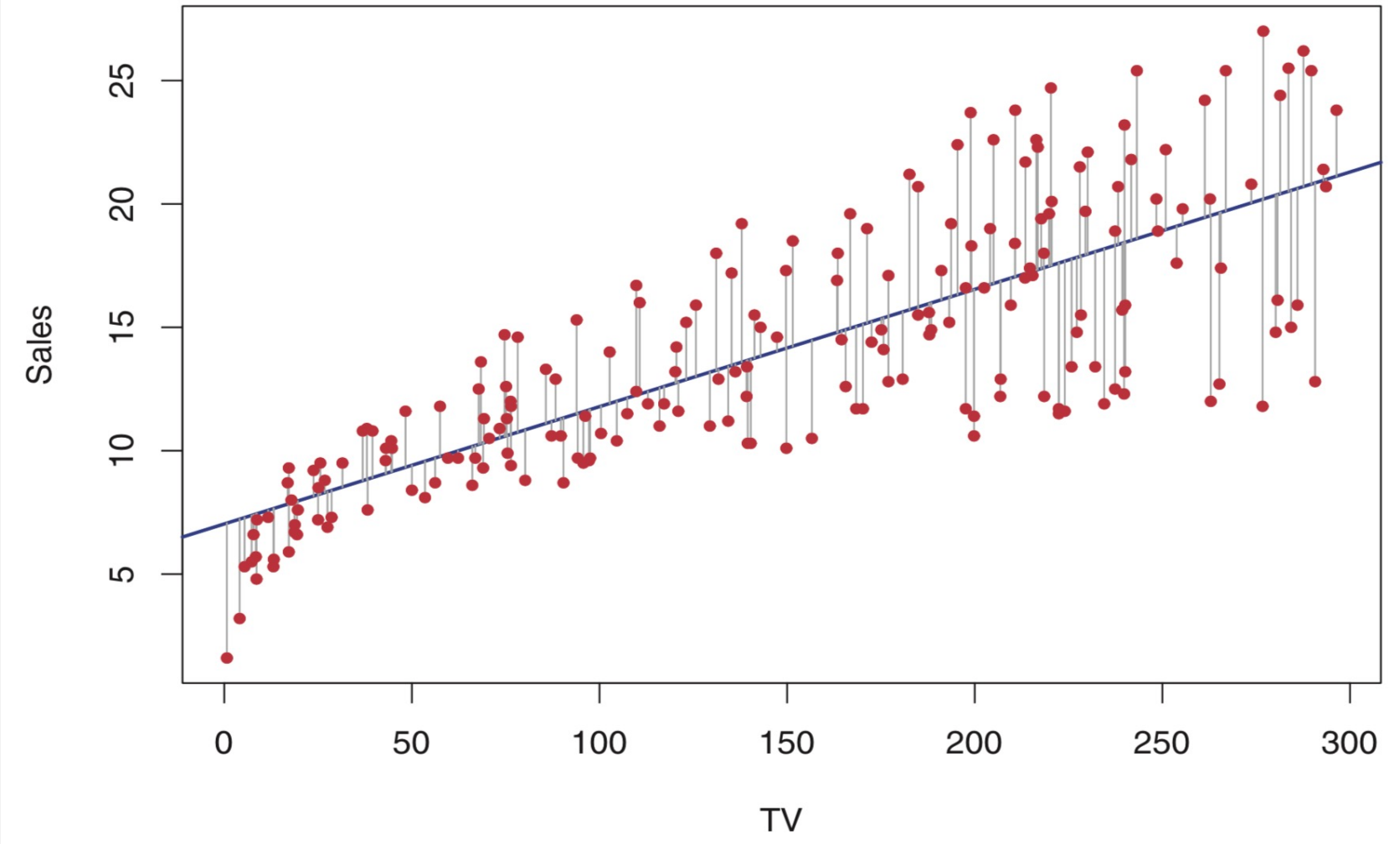



Figure 3.1, p62, An Introduction to Statistical Learning, with applications in R" (Springer, 2013) with permission from the authors: G. James, D. Witten, T. Hastie and R. Tibshirani

# least squares : minimize the RSS

- $\widehat{\beta_1} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$
- $\widehat{\beta_0} = \bar{y} - \widehat{\beta_1}\bar{x}$
  - $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$
  - $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

$\widehat{\beta_0} = 7\ .03\ ,\ \widehat{\beta_1} = 0\ .0475$

$1,000 on TV advertising ~= 47.5 additional selling

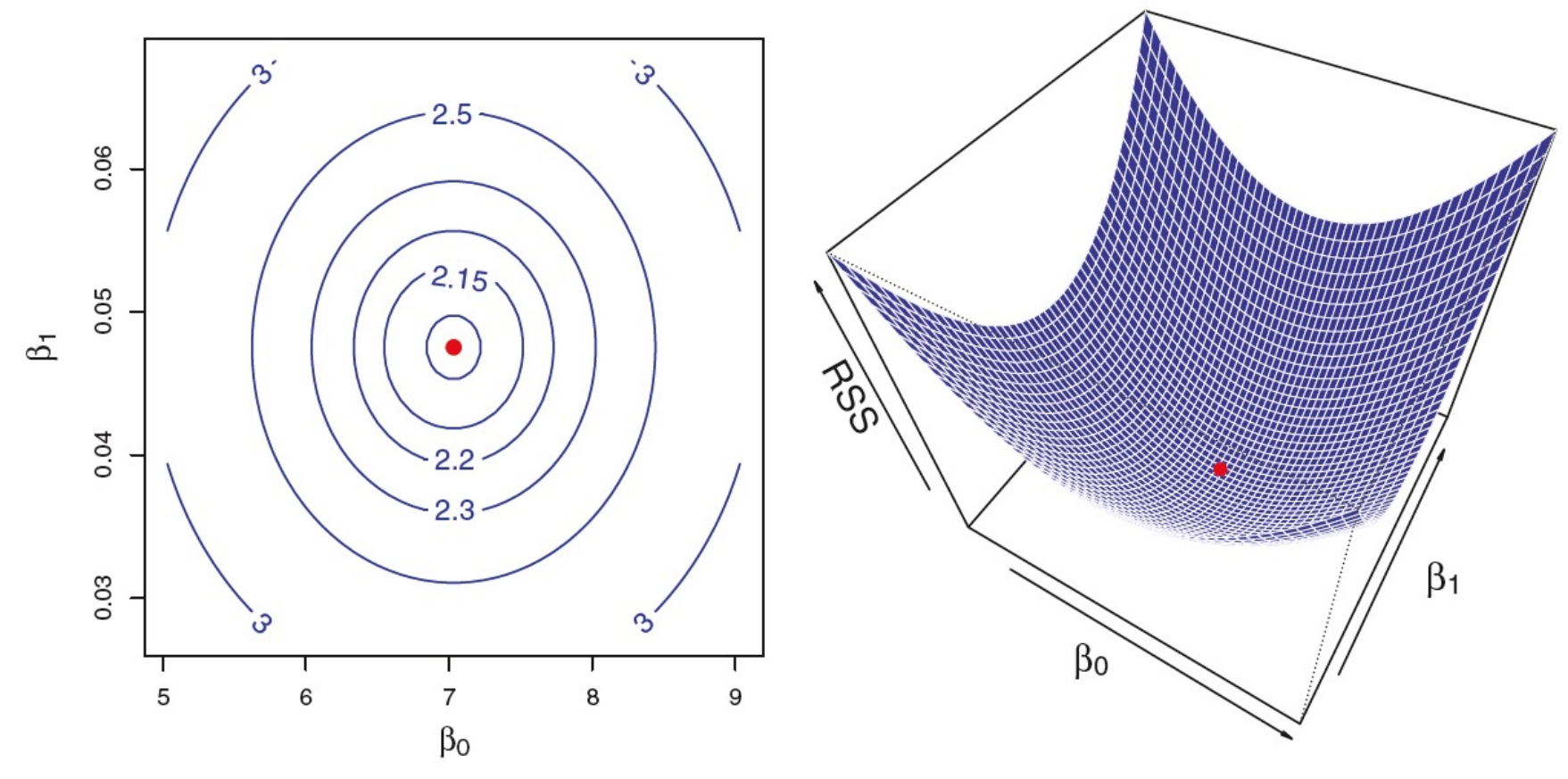



Figure 3.2, p63, An Introduction to Statistical Learning, with applications in R" (Springer, 2013) with permission from the authors: G. James, D. Witten, T. Hastie and R. Tibshirani

# Prepare data

- 2011 US Census PUMS data
  - predict personal income from other demographic variables such as age and education
- pums.R

```
load("psub.RData")
dtrain <- subset(psub,ORIGRANDGROUP >= 500)
dtest <- subset(psub,ORIGRANDGROUP < 500)
```

# Building a linear model using the lm() command

- Predict the log base 10 of income as a function of age, sex, employment class, and education.
  - Output
    - Personal income (PINCP)
  - Input
    - age (AGEP)
    - Sex (SEX) : reference level =M
    - class of worker (COW) : reference level = Employee of a private for-profit
    - level of education (SCHL) : reference level = no high school diploma

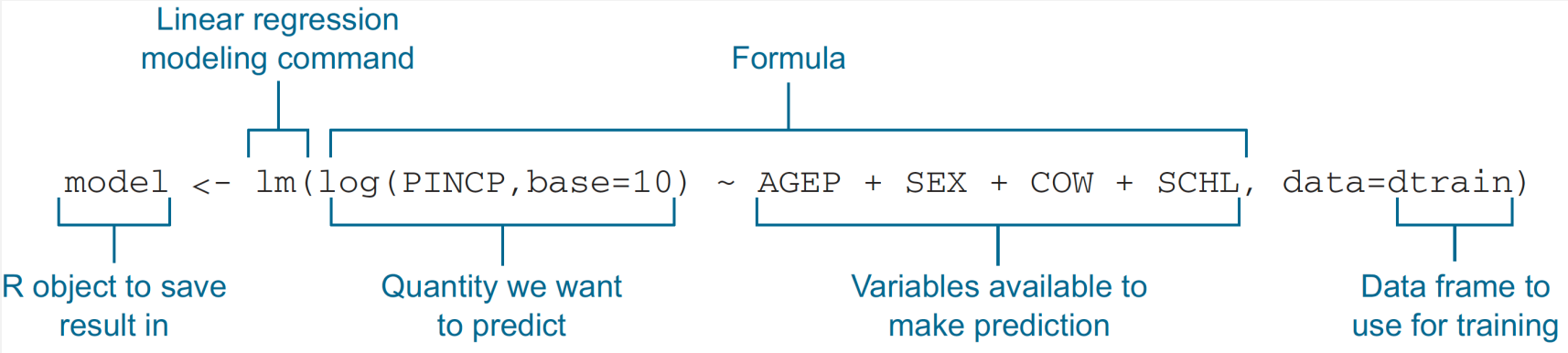
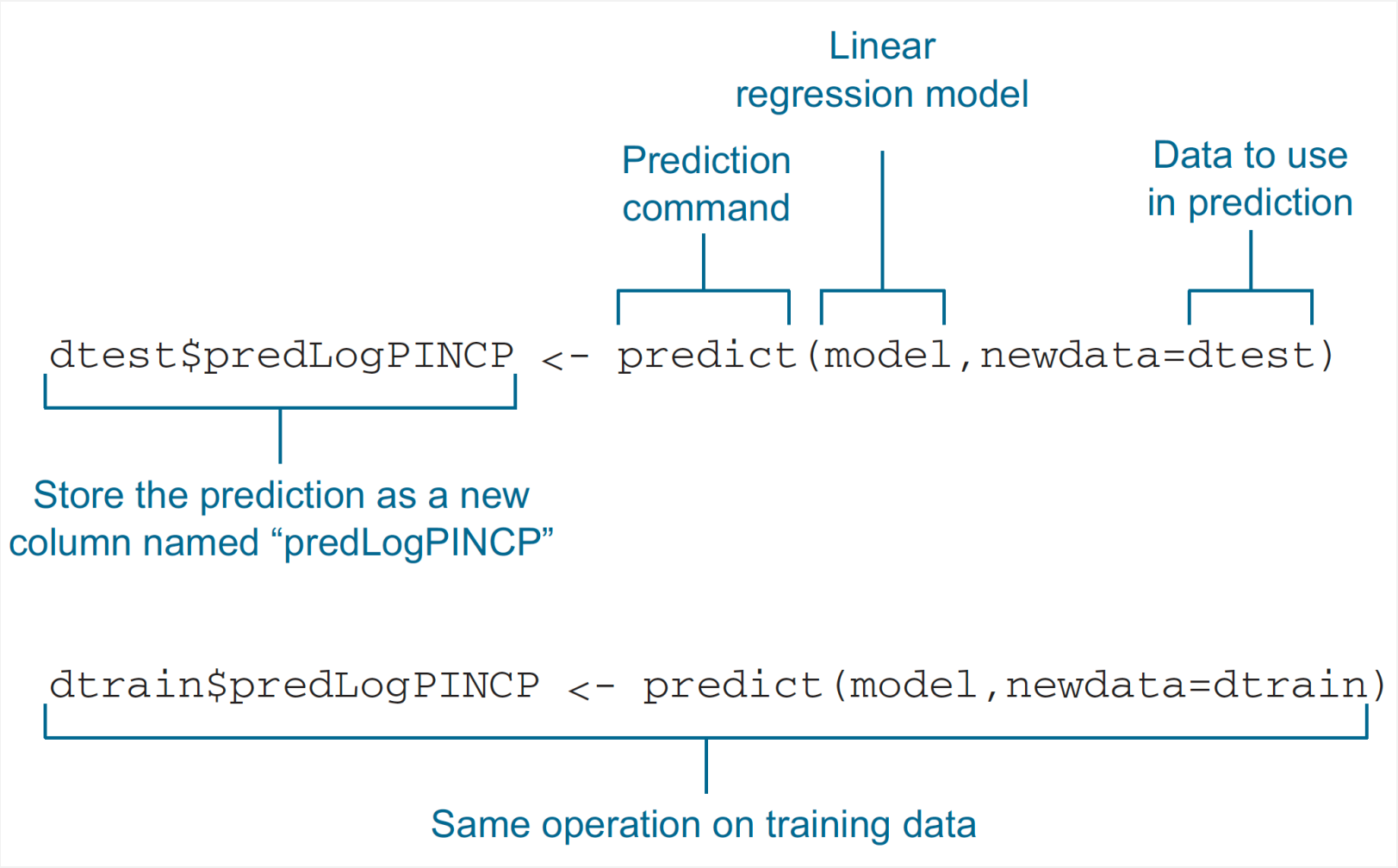



Figure7.2, p145, Practical Data Science with R by Nina Zumel, John Mount

# Making predictions with a linear regression model



# Check coefficients?

## Linear regression

# Finding relations and extracting advice

- coefficients(model)
  - The level that isn't shown is called *the reference level*
  - What is the reference level for SCHL?

```
levels(dtrain$SCHL)
```

```
> coefficients(model)
                        (Intercept)                                AGEP
                         3.97328317                          0.01171695
                               SEXF           COWFederal government employee
                        -0.09313262                          0.05972743
      COWLocal government employee      COWPrivate not-for-profit employee
                        -0.03441690                         -0.13352439
     COWSelf-employed incorporated       COWSelf-employed not incorporated
                        -0.07242727                         -0.06060878
      COWState government employee                  SCHLAssociate's degree
                        -0.08128385                          0.22143196
          SCHLBachelor's degree                   SCHLDoctorate degree
                         0.39383406                          0.30655376
 SCHLGED or alternative credential                    SCHLMaster's degree
                         0.13776759                          0.47728422
        SCHLProfessional degree                 SCHLRegular high school diploma
                         0.67877416                          0.10174598
SCHLsome college credit, no degree
                         0.16315219
```

The modeled values of a bachelor's degree versus a high school diploma

**Figure 7.6** **The model coefficients**

# Finding relations and extracting advice

- The modeled relation between the bachelor's degree holder's expected income and high school graduate's ?

```
> coefficients(model)
                        (Intercept)                                AGEP
                         3.97328317                          0.01171695
                               SEXF  COWFederal government employee
                        -0.09313262                          0.05972743
 COWLocal government employee COWPrivate not-for-profit employee
                        -0.03441690                         -0.13352439
 COWSelf-employed incorporated  COWSelf-employed not incorporated
                        -0.07242727                         -0.06060878
 COWState government employee          SCHLAssociate's degree
                        -0.08128385                          0.22143196
      SCHLBachelor's degree                SCHLDoctorate degree
                         0.39383406                          0.30655376
SCHLGED or alternative credential           SCHLMaster's degree
                         0.13776759                          0.47728422
     SCHLProfessional degree          SCHLRegular high school diploma
                         0.67877416                          0.10174598
SCHLsome college credit, no degree
                         0.16315219
```

The modeled values of a bachelor's degree versus a high school diploma

**Figure 7.6 The model coefficients**

# SCHLBachelor's degree the coefficient = 0.39

- The model gives a *0.39* bonus to log income for having a bachelor's degree, relative to not having a high school degree.

```
> coefficients(model)
                          (Intercept)                                  AGEP
                           3.97328317                            0.01171695
                                 SEXF     COWFederal government employee
                          -0.09313262                            0.05972743
      COWLocal government employee COWPrivate not-for-profit employee
                          -0.03441690                           -0.13352439
     COWSelf-employed incorporated  COWSelf-employed not incorporated
                          -0.07242727                           -0.06060878
      COWState government employee            SCHLAssociate's degree
                          -0.08128385                            0.22143196
          SCHLBachelor's degree                SCHLDoctorate degree
                           0.39383406                            0.30655376
 SCHLGED or alternative credential            SCHLMaster's degree
                           0.13776759                            0.47728422
         SCHLProfessional degree     SCHLRegular high school diploma
                           0.67877416                            0.10174598
SCHLsome college credit, no degree
                           0.16315219
```

The modeled values of a bachelor's degree versus a high school diploma

**Figure 7.6 The model coefficients**

# Finding relations and extracting advice

- AGEP is a continuous variable with coefficient 0.0117?
  - a one-year increase in age, adding a 0.0117 bonus to log income

```
> coefficients(model)
                        (Intercept)                                AGEP
                         3.97328317                          0.01171695
                               SEXF           COWFederal government employee
                        -0.09313262                          0.05972743
      COWLocal government employee      COWPrivate not-for-profit employee
                        -0.03441690                         -0.13352439
     COWSelf-employed incorporated       COWSelf-employed not incorporated
                        -0.07242727                         -0.06060878
      COWState government employee                 SCHLAssociate's degree
                        -0.08128385                          0.22143196
          SCHLBachelor's degree                   SCHLDoctorate degree
                         0.39383406                          0.30655376
 SCHLGED or alternative credential                    SCHLMaster's degree
                         0.13776759                          0.47728422
        SCHLProfessional degree                 SCHLRegular high school diploma
                         0.67877416                          0.10174598
SCHLsome college credit, no degree
                         0.16315219
```

The modeled values of a bachelor's degree versus a high school diploma

**Figure 7.6 The model coefficients**

# Reading the model summary and characterizing coefficient quality

`summary(model)`

Model call summary

```
Call:
lm(formula = log(PINCP, base = 10) ~ AGEP + SEX + COW + SCHL,
    data = dtrain)
```

Residuals summary

```
Residuals:
     Min       1Q   Median       3Q      Max
-1.29220 -0.14153  0.02458  0.17632  0.62532
```

Coefficients

```
Coefficients:
                                      Estimate Std. Error t value Pr(>|t|)
(Intercept)                           3.973283   0.059343  66.954  < 2e-16 ***
AGEP                                  0.011717   0.001352   8.666  < 2e-16 ***
SEXF                                 -0.093133   0.023405  -3.979 7.80e-05 ***
COWFederal government employee        0.059727   0.060927   0.980 0.327343
COWLocal government employee         -0.034417   0.048030  -0.717 0.473928
COWPrivate not-for-profit employee   -0.133524   0.039223  -3.404 0.000709 ***
COWSelf-employed incorporated        -0.072427   0.068093  -1.064 0.287928
COWSelf-employed not incorporated    -0.060609   0.069244  -0.875 0.381779
COWState government employee         -0.081284   0.057796  -1.406 0.160146
SCHLAssociate's degree                0.221432   0.052094   4.251 2.49e-05 ***
SCHLBachelor's degree                 0.393834   0.043249   9.106  < 2e-16 ***
SCHLDoctorate degree                  0.306554   0.160127   1.914 0.056058 .
SCHLGED or alternative credential     0.137768   0.078192   1.762 0.078612 .
SCHLMaster's degree                   0.477284   0.050895   9.378  < 2e-16 ***
SCHLProfessional degree               0.678774   0.087321   7.773 3.52e-14 ***
SCHLRegular high school diploma       0.101746   0.042628   2.387 0.017316 *
SCHLsome college credit, no degree    0.163152   0.042729   3.818 0.000149 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Model quality summary

```
Residual standard error: 0.2691 on 578 degrees of freedom
Multiple R-squared: 0.3383, Adjusted R-squared: 0.3199
F-statistic: 18.47 on 16 and 578 DF,  p-value: < 2.2e-16
```

Figure 7.7, p152, Practical Data Science with R by Nina Zumel, John Mount

# The Coefficients Table

```
summary(model)$coefficients
```

```
Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                       3.973283   0.059343  66.954  < 2e-16 ***
AGEP                              0.011717   0.001352   8.666  < 2e-16 ***
SEXF                             -0.093133   0.023405  -3.979 7.80e-05 ***
COWFederal government employee    0.059727   0.060927   0.980 0.327343
COWLocal government employee     -0.034417   0.048030  -0.717 0.473928
COWPrivate not-for-profit employee -0.133524   0.039223  -3.404 0.000709 ***
COWSelf-employed incorporated    -0.072427   0.068093  -1.064 0.287928
COWSelf-employed not incorporated -0.060609   0.069244  -0.875 0.381779
COWState government employee     -0.081284   0.057796  -1.406 0.160146
SCHLAssociate's degree            0.221432   0.052094   4.251 2.49e-05 ***
SCHLBachelor's degree             0.393834   0.043249   9.106  < 2e-16 ***
SCHLDoctorate degree              0.306554   0.160127   1.914 0.056058 .
SCHLGED or alternative credential 0.137768   0.078192   1.762 0.078612 .
SCHLMaster's degree               0.477284   0.050895   9.378  < 2e-16 ***
SCHLProfessional degree           0.678774   0.087321   7.773 3.52e-14 ***
SCHLRegular high school diploma   0.101746   0.042628   2.387 0.017316 *
SCHLsome college credit, no degree 0.163152   0.042729   3.818 0.000149 ***
```

Name of coefficient

Coefficient estimate

Standard error in estimate

t-value: Number of standard errors estimate is away from zero

p-value: Probability of such a large t-value forming by mere chance

# Population regression line

- $Y = \beta_0 + \beta_1 X + \epsilon$
  - best linear approximation to the true relationship between *X* and *Y*
  - $\beta_0$: the intercept term, the expected value of *Y* when *X* = 0
  - $\beta_1$: the slope, the average increase in *Y* associated with a one-unit increase in *X*.
  - $\epsilon$: the error term, a catch-all for what we miss with this simple model
    - the true relationship is probably not linear, there may be other variables that cause variation in *Y*, and there may be measurement error.
    - typically assume that the error term is independent of *X*.

# The population regression line vs least squares estimate

- $Y = \beta_0 + \beta_1 X + \epsilon$ (red)
  - $Y = 2 + 3X + \epsilon$, where $\epsilon$ generated from a normal distribution with mean zero
- $y_i \approx \widehat{\beta_0} + \widehat{\beta_1} x_i$ for $i = 1, \cdots, n$ (blue)

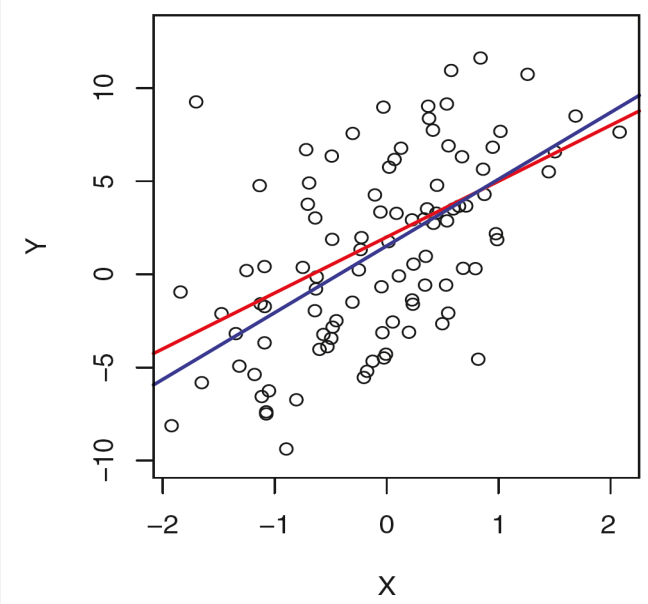



Figure3.3, p64, An Introduction to Statistical Learning, with applications in R" (Springer, 2013) with permission from the authors: G. James, D. Witten, T. Hastie and R. Tibshirani

# How close $\widehat{\beta_0}$ & $\widehat{\beta_1}$ are to the true values $\beta_0$ & $\beta_1$?

- we are interested in knowing the population mean $\mu$ of some random variable $Y$
- Unfortunately, $\mu$ is unknown, but we do have access to $n$ observations from $Y$, $y_1,\cdots,y_n$

# How close $\widehat{\beta_0}$ & $\widehat{\beta_1}$ are to the true values $\beta_0$ & $\beta_1$?

- A reasonable estimate
  - $\hat{\mu} = \bar{y}, \text{where } \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$
- How far off will that single estimate of $\hat{\mu}$ be?
- standard error of $\hat{\mu}$:
  - $\text{Var}(\hat{\mu}) = \text{SE}(\hat{\mu})^2 = \frac{\sigma^2}{n}$
  - $\sigma$: the standard deviation of each of the realizations $y_i$ of $Y$

# The standard errors of $\widehat{\beta_0}$ & $\widehat{\beta_1}$

- $\mathrm{SE}\left(\widehat{\beta_0}\right)^2 = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right]$
- $\mathrm{SE}\left(\widehat{\beta_1}\right)^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$
  - $\sigma^2 = \mathrm{Var}(\epsilon)$, in general unknown
    - estimated by residual standard error, $RSE = \sqrt{RSS/(n-2)}$

# Confidence intervals

- 95% confidence confidence interval: a range of values such that with 95% probability, the range will contain the true unknown value of the parameter.
- 95% confidence interval for $\beta_0$
  - $\widehat{\beta_0} \pm 2 \cdot \mathrm{SE}(\widehat{\beta_0})$
- 95% confidence interval for $\beta_1$
  - $\widehat{\beta_1} \pm 2 \cdot \mathrm{SE}(\widehat{\beta_1})$

# TV @ Sales example

- The 95% confidence interval $\beta_0$: [6130 , 7935]
  - in the absence of any advertising, sales will, on average, fall somewhere between 6130 and 7940 units.
- The 95% confidence interval $\beta_1$: [0 .042 , 0 .053]
  - for each $1,000 increase in television advertising, there will be an average increase in sales of between 42 and 53 units

# *t*-statistic

- measures the number of standard deviations that $\widehat{\beta_1}$ is away from 0
  - $t = \frac{\widehat{\beta_1} - 0}{\mathrm{SE}(\widehat{\beta_1})}$
  - *t*-distribution with *n*−2 degrees of freedom
  - *t*-distribution has a bell shape
  - *n* >= 30: quite similar to the normal distribution

# Null hypothesis

- $H_0$: There is no relationship between $X$ and $Y$
- $H_0 : \beta_1 = 0$
  - depends on $\mathrm{SE}(\widehat{\beta_1})$

# *p*-value

- the probability of observing any number equal to $|t|$ or larger in absolute value, assuming $\beta_1 = 0$
- *p*-value $\leq$ 5% or 1%: reject the null hypothesis, a relationship to exist between $X$ and $Y$
  - *n* = 30, *t*-statistics of around 2 and 2.75, respectively.

# TV @ Sales example

| | Coefficient | Std. error | t-statistic | p-value |
|---|---|---|---|---|
| `Intercept` | 7.0325 | 0.4578 | 15.36 | $< 0.0001$ |
| `TV` | 0.0475 | 0.0027 | 17.67 | $< 0.0001$ |

# *p*-value

- the probability of seeing a coefficient with a magnitude as large as we observe if the true coefficient is really 0 (the variable has no effect on the outcome)

≥ 0.05, is not to be trusted

1e-23 vs 1e-08?

```
Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                       3.973283   0.059343  66.954  < 2e-16 ***
AGEP                              0.011717   0.001352   8.666  < 2e-16 ***
SEXF                             -0.093133   0.023405  -3.979 7.80e-05 ***
COWFederal government employee    0.059727   0.060927   0.980 0.327343
COWLocal government employee     -0.034417   0.048030  -0.717 0.473928
COWPrivate not-for-profit employee -0.133524   0.039223  -3.404 0.000709 ***
COWSelf-employed incorporated    -0.072427   0.068093  -1.064 0.287928
COWSelf-employed not incorporated -0.060609   0.069244  -0.875 0.381779
COWState government employee     -0.081284   0.057796  -1.406 0.160146
SCHLAssociate's degree            0.221432   0.052094   4.251 2.49e-05 ***
SCHLBachelor's degree             0.393834   0.043249   9.106  < 2e-16 ***
SCHLDoctorate degree              0.306554   0.160127   1.914 0.056058 .
SCHLGED or alternative credential 0.137768   0.078192   1.762 0.078612 .
SCHLMaster's degree               0.477284   0.050895   9.378  < 2e-16 ***
SCHLProfessional degree           0.678774   0.087321   7.773 3.52e-14 ***
SCHLRegular high school diploma   0.101746   0.042628   2.387 0.017316 *
SCHLsome college credit, no degree 0.163152   0.042729   3.818 0.000149 ***
```

Name of coefficient

Coefficient estimate

Standard error in estimate

t-value: Number of standard errors estimate is away from zero

p-value: Probability of such a large t-value forming by mere chance

# Multiple Linear Regression
# vs.
# Separate simple linear regression

# Separate simple linear regression

| | Coefficient | Std. error | t-statistic | p-value |
|---|---|---|---|---|
| Intercept | 7.0325 | 0.4578 | 15.36 | $< 0.0001$ |
| TV | 0.0475 | 0.0027 | 17.67 | $< 0.0001$ |

| | Coefficient | Std. error | t-statistic | p-value |
|---|---|---|---|---|
| Intercept | 9.312 | 0.563 | 16.54 | $< 0.0001$ |
| radio | 0.203 | 0.020 | 9.92 | $< 0.0001$ |

| | Coefficient | Std. error | t-statistic | p-value |
|---|---|---|---|---|
| Intercept | 12.351 | 0.621 | 19.88 | $< 0.0001$ |
| newspaper | 0.055 | 0.017 | 3.30 | 0.00115 |

# Multiple Linear Regression

- sales = $\beta_0 + \beta_1$x TV + $\beta_2$ x radio + $\beta_3$ x newspaper + $\epsilon$

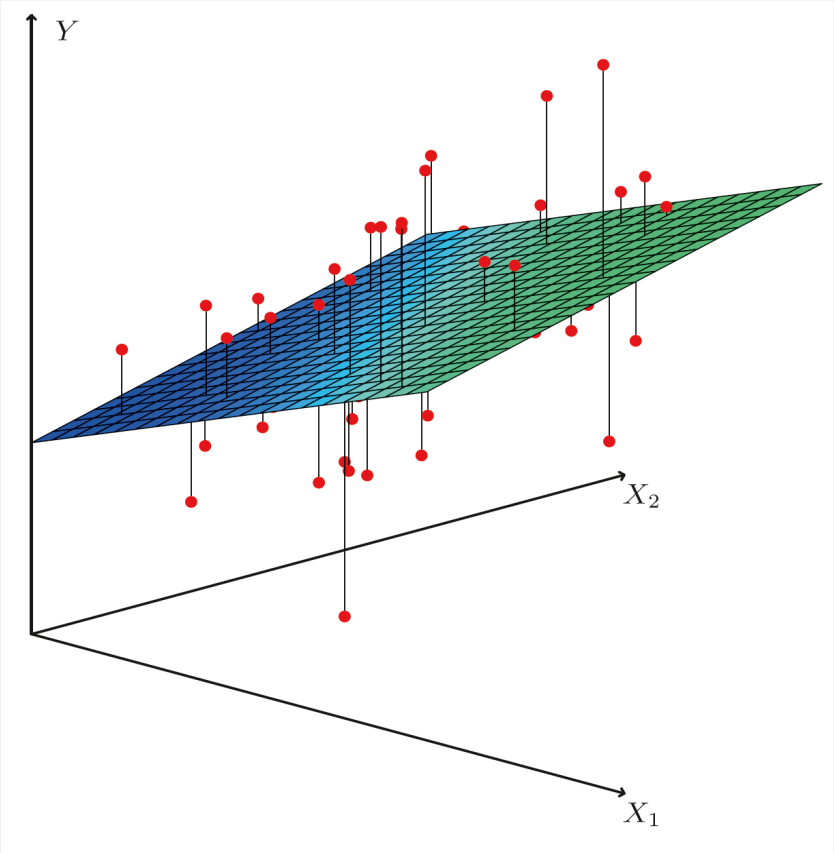



Figure3.4, p64, An Introduction to Statistical Learning, with applications in R" (Springer, 2013) with permission from the authors: G. James, D. Witten, T. Hastie and R. Tibshirani

# Multiple Linear Regression

| | Coefficient | Std. error | t-statistic | p-value |
|---|---|---|---|---|
| Intercept | 2.939 | 0.3119 | 9.42 | $< 0.0001$ |
| TV | 0.046 | 0.0014 | 32.81 | $< 0.0001$ |
| radio | 0.189 | 0.0086 | 21.89 | $< 0.0001$ |
| newspaper | $-0.001$ | 0.0059 | $-0.18$ | 0.8599 |

# Multiple Linear Regression vs. Separate simple linear regression

| | Coefficient | Std. error | t-statistic | p-value |
|---|---|---|---|---|
| Intercept | 2.939 | 0.3119 | 9.42 | $< 0.0001$ |
| TV | 0.046 | 0.0014 | 32.81 | $< 0.0001$ |
| radio | 0.189 | 0.0086 | 21.89 | $< 0.0001$ |
| newspaper | −0.001 | 0.0059 | −0.18 | 0.8599 |

| | Coefficient | Std. error | t-statistic | p-value |
|---|---|---|---|---|
| Intercept | 7.0325 | 0.4578 | 15.36 | $< 0.0001$ |
| TV | 0.0475 | 0.0027 | 17.67 | $< 0.0001$ |

| | Coefficient | Std. error | t-statistic | p-value |
|---|---|---|---|---|
| Intercept | 9.312 | 0.563 | 16.54 | $< 0.0001$ |
| radio | 0.203 | 0.020 | 9.92 | $< 0.0001$ |

| | Coefficient | Std. error | t-statistic | p-value |
|---|---|---|---|---|
| Intercept | 12.351 | 0.621 | 19.88 | $< 0.0001$ |
| newspaper | 0.055 | 0.017 | 3.30 | 0.00115 |

# Multiple Linear Regression vs. Separate simple linear regression

| | Coefficient | Std. error | t-statistic | p-value |
|---|---|---|---|---|
| Intercept | 2.939 | 0.3119 | 9.42 | $< 0.0001$ |
| TV | 0.046 | 0.0014 | 32.81 | $< 0.0001$ |
| radio | 0.189 | 0.0086 | 21.89 | $< 0.0001$ |
| newspaper | −0.001 | 0.0059 | −0.18 | 0.8599 |

the coefficient for newspaper represents the average effect of increasing newspaper spending by $1000 while holding TV and radio fixed.

| | Coefficient | Std. error | t-statistic | p-value |
|---|---|---|---|---|
| Intercept | 12.351 | 0.621 | 19.88 | $< 0.0001$ |
| newspaper | 0.055 | 0.017 | 3.30 | 0.00115 |

The slope term represents the average effect of a $1000 increase in newspaper advertising, ignoring other predictors such as TV and radio

# Correlation matrix

- higher values of newspaper tend to be associated with higher values of sales , even though newspaper advertising does not actually affect sales
- newspaper gets “credit” for the effect of radio on sales.

| | TV | radio | newspaper | sales |
|---|---|---|---|---|
| TV | 1.0000 | 0.0548 | 0.0567 | 0.7822 |
| radio | | 1.0000 | 0.3541 | 0.5762 |
| newspaper | | | 1.0000 | 0.2283 |
| sales | | | | 1.0000 |

# shark attacks vs. ice cream sales

- a regression of shark attacks vs. ice cream sales for data collected at a given beach community over a period of time => a positive relationship
- Higher temperatures cause more people to visit the beach, which in turn results in more ice cream sales and more shark attacks.
- multiple regression
  - attacks vs. ice cream sales + temperature
  - ice cream sales is no longer significant after adjusting for temperature.

# How to evaluate scoring models?

## Linear regression

# Predict based on trained model

```
dtest$predLogPINCP <- predict(model,newdata=dtest)
dtrain$predLogPINCP <- predict(model,newdata=dtrain)
```

# Characterizing prediction quality

```
ggplot(data=dtest,aes(x=predLogPINCP,y=log(PINCP,base=10))) +
geom_point(alpha=0.2,color="black") +
geom_smooth(aes(x=predLogPINCP, y=log(PINCP,base=10)),color="black") +
geom_line(aes(x=log(PINCP,base=10), y=log(PINCP,base=10)),color="blue",linetype=2) +
scale_x_continuous(limits=c(4,5)) + scale_y_continuous(limits=c(3.5,5.5))
```

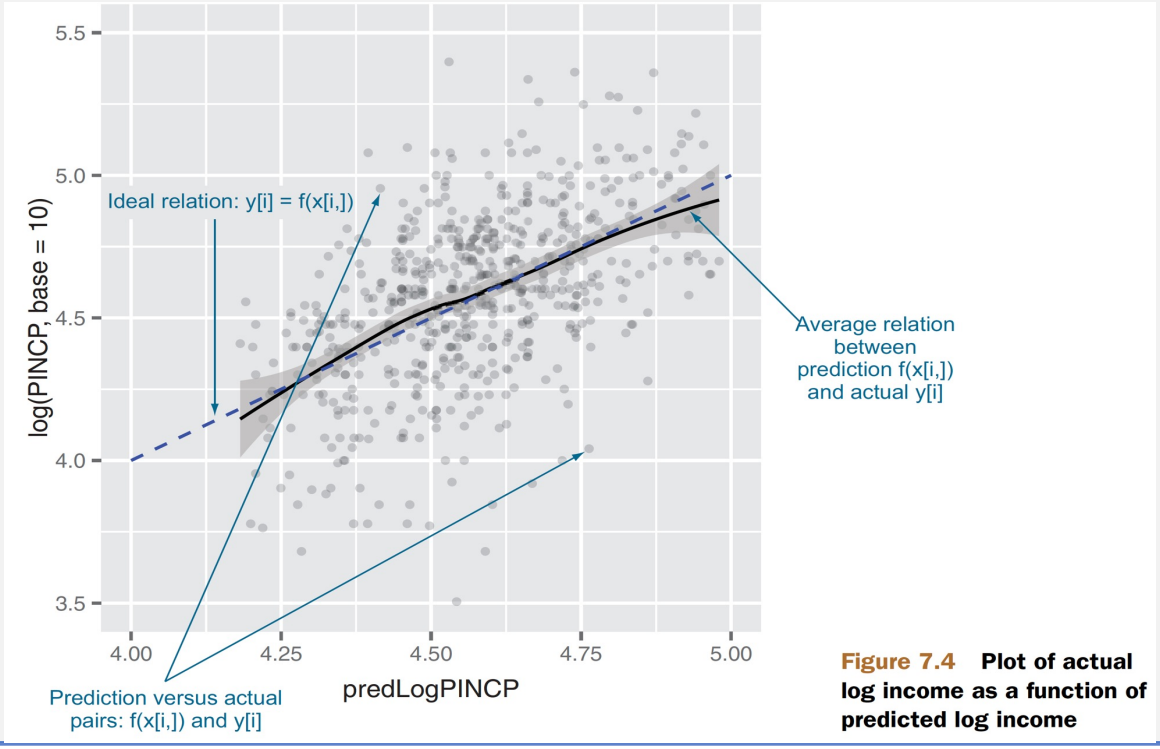



**Figure 7.4** **Plot of actual log income as a function of predicted log income**

# On average, are the predictions correct?

```
ggplot(data=dtest,aes(x=predLogPINCP,
y=predLogPINCP-log(PINCP,base=10))) +
geom_point(alpha=0.2,color="black") +
geom_smooth(aes(x=predLogPINCP, y=predLogPINCP-log(PINCP,base=10)),
color="black")
```

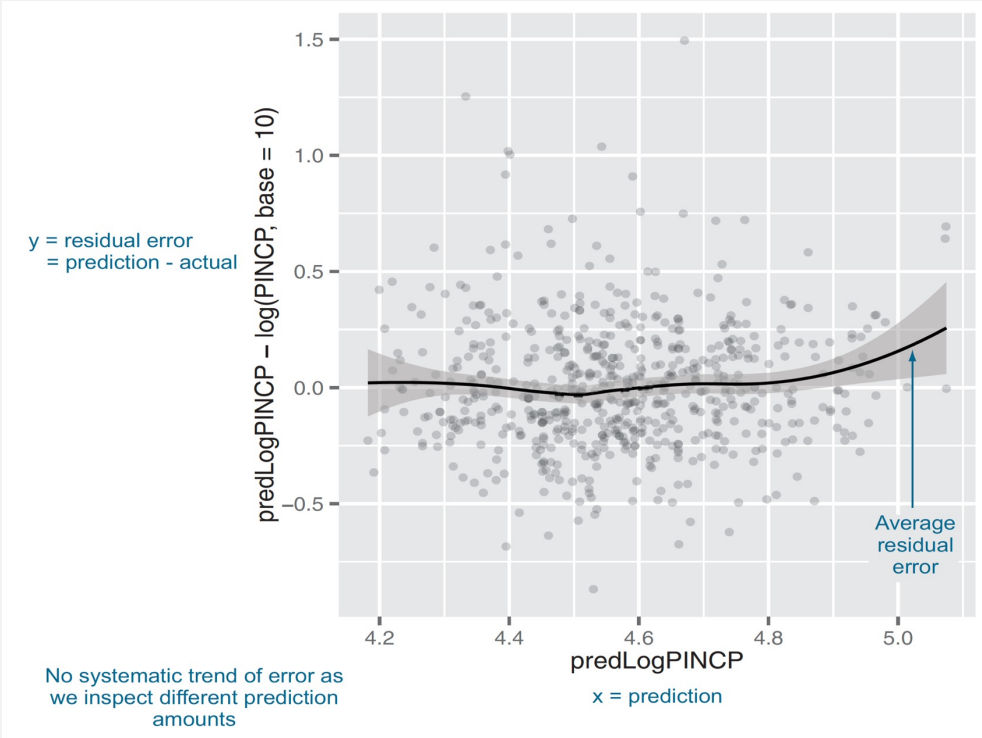



Figure7.5, *Practical Data Science with R* by Nina Zumel, John Mount

# On average, are the predictions correct?

```
summary(log(dtrain$PINCP,base=10) -
predict(model,newdata=dtrain))
```

```
summary(log(dtest$PINCP,base=10) - predict(model,newdata=dtest))
```

- the median near 0
- exactly half of the training data has a residual in 1st. Qu. and 3rd Qu. quantiles

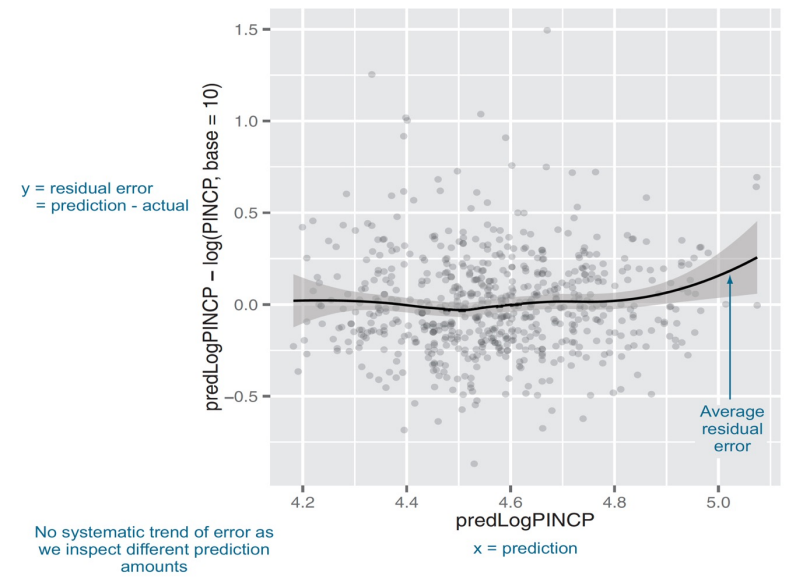



Figure7.5, *Practical Data Science with R* by Nina Zumel, John Mount

# Why are the predictions, not the true values, on the x-axis?

- A residual graph with
  - predictions on the x-axis
    - gives you a sense of when the model may be under- or overpredicting, based on the model's output.
  - the true outcome on the x-axis
    - gives you a sense of where the model under or overpredicts based on the actual outcome.

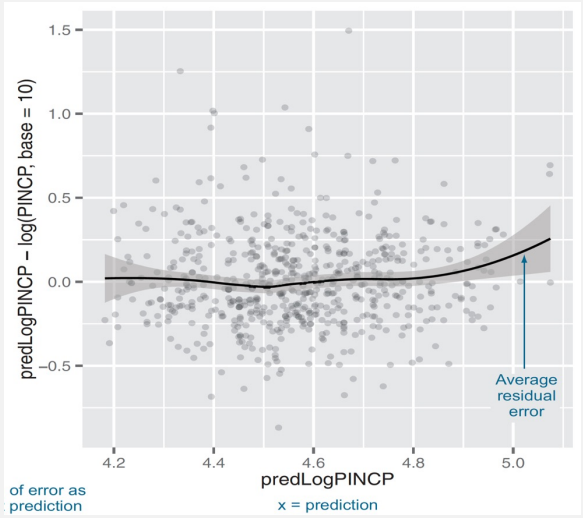



Figure7.5, *Practical Data Science with R* by Nina Zumel, John Mount

# Scoring residuals

- *residuals* or *difference* between our predictions and actual outcomes

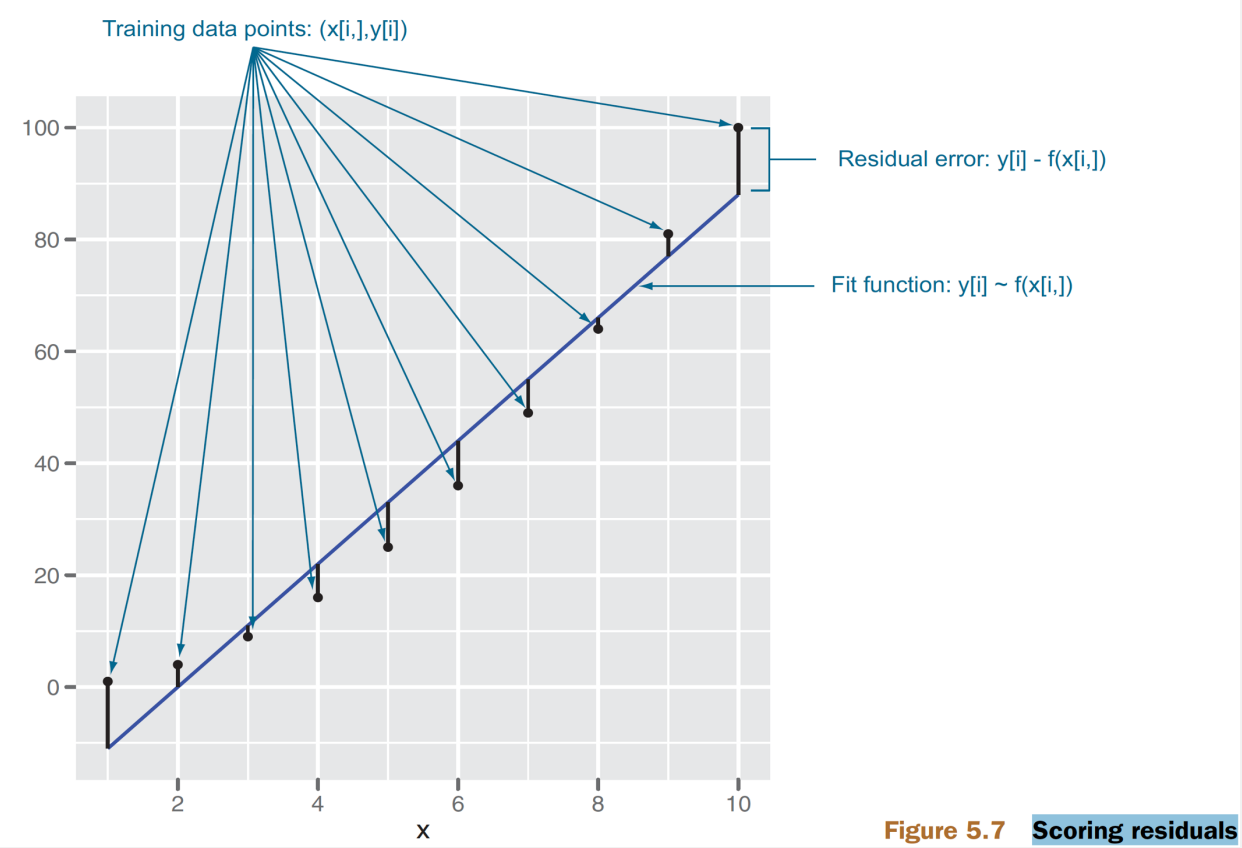



**Figure 5.7** **Scoring residuals**

# Residual Standard Error (RSE)

- Residual sum of squares (RSS)
  - $\mathrm{RSS} = \sum_{i=1}^{n}(y_i - \bar{y})^2$
- Residual Standard Error
  - $\mathrm{RSE} = \sqrt{\frac{1}{n-2}\mathrm{RSS}}$

# Residual Standard Error (RSE)

- TV @ Sales example
  - RSE = 3. 26
  - actual sales in each market deviate from the true regression line by approximately 3260 units, on average
  - if the model were correct and the true values of the unknown coefficients $\beta_0$ and $\beta_1$ were known exactly, any prediction of sales on the basis of TV advertising would still be off by about 3260 units on average.

# R-SQUARED ($R^2$) statistic

- The RSE provides an absolute measure of lack of fit of the model to the data.
  - measured in the units of $Y$ => not always clear what constitutes a good RSE
- $R^2$ statistic: the form of a proportion, the proportion of variance explained, 0~1
  - independent of the scale of $Y$

# $R^2$ statistic

- TSS (Total sum of squares) $= \sum(y_i - \bar{y})^2$
  - variability inherent in the response before the regression is performed
- RSS: variability left unexplained after performing the regression
- TSS-RSS: variability in the response that is explained by performing the regression
- $R^2 = \frac{\text{TSS}-\text{RSS}}{\text{TSS}} = 1 - \frac{\text{RSS}}{\text{TSS}}$
  - the proportion of variability in *Y* that can be explained using *X*

# $R^2$ statistic

- 1 - sum((*pred*-*actVal*)^2)/sum((mean(*actVal*)-*actVal*)^2)
- *R-squared = correlation*$^2$

# Are there systematic errors?

- what fraction of the *y* variation is explained by the model
- well-fit models, *R*-squared is also equal to the square of the correlation between the predicted values and actual training values.
- be fairly large (1.0 is the largest you can achieve)
  - *R*-squareds that are similar on test and training.
- Overfit
  - *R*-squared on test data : lower

# $R^2$ discussion on Sales example

- $R^2 = 0.61$
  - only TV as a predictor
- $R^2 = 0.89719$
  - the model that uses only TV and radio to predict sales
- $R^2 = 0.8972$
  - uses all three advertising media to predict sales
  - tiny increase in $R^2$ => newspaper can be dropped from the model

# Are there systematic errors?

- Code

```
rsq <- function(y,f) { 1 - sum((y-f)^2)/sum((y-mean(y))^2) }
rsq(log(dtrain$PINCP,base=10),predict(model,newdata=dtrain))
rsq(log(dtest$PINCP,base=10),predict(model,newdata=dtest))
```

- How about the model?
  - Quality : low
  - Overfit : not substantially

# *R*-squared can be overoptimistic

- a toy example

```
y <- c(1,2,3,4,5,9,10)
pred <- c(0.5,0.5,0.5, 0.5,0.5,9,10)
```

- What is its *R*-square?

```
rsq(y,pred)
plot(y,pred)
```

# *R*-squared can be overoptimistic

- R-squared is related to correlation, and the correlation can be artificially inflated if the model correctly predicts a few outliers.
- true-versus-fitted graph in addition to checking *R*-squared

# Root Mean Square Error (RMSE )

- sqrt(mean((*prediction*-*actualValues*)^2))

```
rmse <- function(y, f) { sqrt(mean( (y-f)^2 )) }

rmse(log(dtrain$PINCP,base=10),predict(model,newdata=dtrain))

rmse(log(dtest$PINCP,base=10),predict(model,newdata=dtest))
```

# Overall model quality summary

- *degrees of freedom* : the number of training data rows you have after correcting for the number of coefficients you tried to solve
  - # of data rows - # of coefficients fit

```
df <- dim(dtrain)[1] -
dim(summary(model)$coefficients)[1]
```

- *residual standard error*
  - sum of the square of the residuals divided by the degrees of freedom

```
modelResidualError <-
sqrt(sum(residuals(model)^2)/df)
```

```
Residual standard error: 0.2691 on 578 degrees of freedom
Multiple R-squared: 0.3383,     Adjusted R-squared: 0.3199
F-statistic: 18.47 on 16 and 578 DF,  p-value: < 2.2e-16
```

# Overall Model Quality Summary

- *Multiple R-squared* : *R*-squared
- *Adjusted R-squared* : correct that more complex models tend to look better on training data due to overfitting
  - *R*-squared penalized by the ratio of the degrees of freedom to the number of training examples.

```
Residual standard error: 0.2691 on 578 degrees of freedom
Multiple R-squared: 0.3383,     Adjusted R-squared: 0.3199
F-statistic: 18.47 on 16 and 578 DF,  p-value: < 2.2e-16
```

# Is There a Relationship Between the Response and Predictors?

- whether the linear regression model predicts outcome better than the constant mode (the mean value of y)
- Null hypothesis
  - $H_0$: $\beta_1 = \beta_2 = \ldots = \beta_p = 0$
- Alternative hypothesis
  - $H_a$: at least one $\beta_j$ is non-zero.

# *F*-statistic

- $F = \frac{(\text{TSS}-\text{RSS})/p}{\text{RSS}/(n-p-1)}$
  - If the linear model assumptions are correct, $E\{RSS/(n-p-1)\} = \sigma^2$
  - if $H_0$ is true, $E\{(TSS-RSS)/p\} = \sigma^2$
    - *F*-statistic close to 1
  - if $H_a$ is true, $E\{(TSS-RSS)/p\} > \sigma^2$
    - *F*-statistic greater than 1

# *F*-statistic @ Sales example

- *F*-statistic is 570.
- Since this is far larger than 1

=> against the null hypothesis $H_0$

=> at least one of the advertising media must be related to sales

| Quantity | Value |
|---|---|
| Residual standard error | 1.69 |
| $R^2$ | 0.897 |
| F-statistic | 570 |

# Given individual *p*-values for each variable, why need the overall F-statistic?

- Given
  - $p = 100$ (the number of predictors)
  - $H_0: \beta_1 = \beta_2 = \ldots = \beta_p = 0$ is true (no variable is truly associated with the response)
- 5%: the *p*-values associated with each variable will be below 0.05 by chance
  - expect to see ~5 small *p*-values even no true association between the predictors and the response
- if $H_0$ is true, there is only a 5% chance that the F-statistic will result in a *p*-value below 0.05, regardless of the number of predictors or the number of observations
  - F-statistic adjusts for *p*

# F-statistic limit

- $p$ certainly small compared to $n$
- If $p > n$, then there are more coefficients $\beta_j$ to estimate than observations from which to estimate them
  - Cannot even fit the multiple linear regression model using least squares
  - F-statistic cannot be used
  - high-dimensional setting use forward selection

# Other Considerations in the Regression Model

# Qualitative Predictors

- balance vs. Ethnicity
  - Asian
  - Caucasian
  - African American
- create two dummy variables, $x_{i1}$& $x_{i2}$
- $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i =$

$$\begin{cases} \beta_0 + \beta_1 + \epsilon_i, \text{if } i\text{th person is Asian} \\ \beta_0 + \beta_2 + \epsilon_i, \text{if } i\text{th person is Caucasian} \\ \beta_0 + \epsilon_i, \text{if } i\text{th person isAfrican American} \end{cases}$$

# Least squares coefficient estimates associated with the regression of balance onto ethnicity in the Credit data set

- Baseline: African American
  - Balance = $531.00
- Asian: have $18.69 less debt than the African American category, no significant
- Caucasian: have $12.50 less debt than the African American, no significant
- *F*-test to $H_0$: $\beta_1 = \beta_2 = 0$ => *p*-value = 0.96

| | Coefficient | Std. error | t-statistic | p-value |
|---|---|---|---|---|
| Intercept | 531.00 | 46.32 | 11.464 | $< 0.0001$ |
| ethnicity[Asian] | −18.69 | 65.02 | −0.287 | 0.7740 |
| ethnicity[Caucasian] | −12.50 | 56.68 | −0.221 | 0.8260 |

# Extensions of the Linear Model

- 2 most important assumption of linear model
  - Additive
    - the effect of changes in a predictor $X_j$ on the response $Y$ is independent of the values of the other predictors
  - Linear:
    - the change in the response $Y$ due to a one-unit change in $X_j$ is constant, regardless of the value of $X_j$

# Removing the Additive Assumption

- Only *main effects*
  - $Y = \beta_0 + \beta_1 X_1 + \beta_1 X_2 + \epsilon$
    - $\text{Sales} = \beta_0 + \beta_1 \text{TV} + \beta_1 \text{radio} + \epsilon$
- synergy effect, interaction
  - $Y = \beta_0 + \beta_1 X_1 + \beta_1 X_2 + \beta_3 X_1 X_2 + \epsilon$
    - $\text{Sales} = \beta_0 + \beta_1 \text{TV} + \beta_1 \text{radio} + \beta_3 \text{TV} \times \text{radio} + \epsilon$

# Removing the Additive Assumption

- $\text{Sales} = 6.7502 + 0.0191\text{TV} + 0.0289\text{radio} + 0.0011\text{TV}\times\text{radio} + \epsilon$
- increase in TV advertising of \$1, 000
  - $\text{Sales} = 6.7502 + (0.0191 + 0.0011\text{radio})\text{TV} + 0.0289\text{radio} + \epsilon$
  - 19+1.1radio units in sales

| | Coefficient | Std. error | t-statistic | p-value |
|---|---|---|---|---|
| Intercept | 6.7502 | 0.248 | 27.23 | $< 0.0001$ |
| TV | 0.0191 | 0.002 | 12.70 | $< 0.0001$ |
| radio | 0.0289 | 0.009 | 3.24 | 0.0014 |
| TV×radio | 0.0011 | 0.000 | 20.73 | $< 0.0001$ |

# Removing the Additive Assumption

- $R^2$
  - the interaction model = 96.8%,
  - only main effects = 89.7%
- $69\% = (96.8 - 89.7)/(100 - 89.7)$
  - the variability in sales that remains after fitting the additive model has been explained by the interaction term

# The hierarchical principle

- If we include an interaction in a model, we should also include the main effects, even if the *p*-values associated with their coefficients are not significant

# Non-linear Relationships

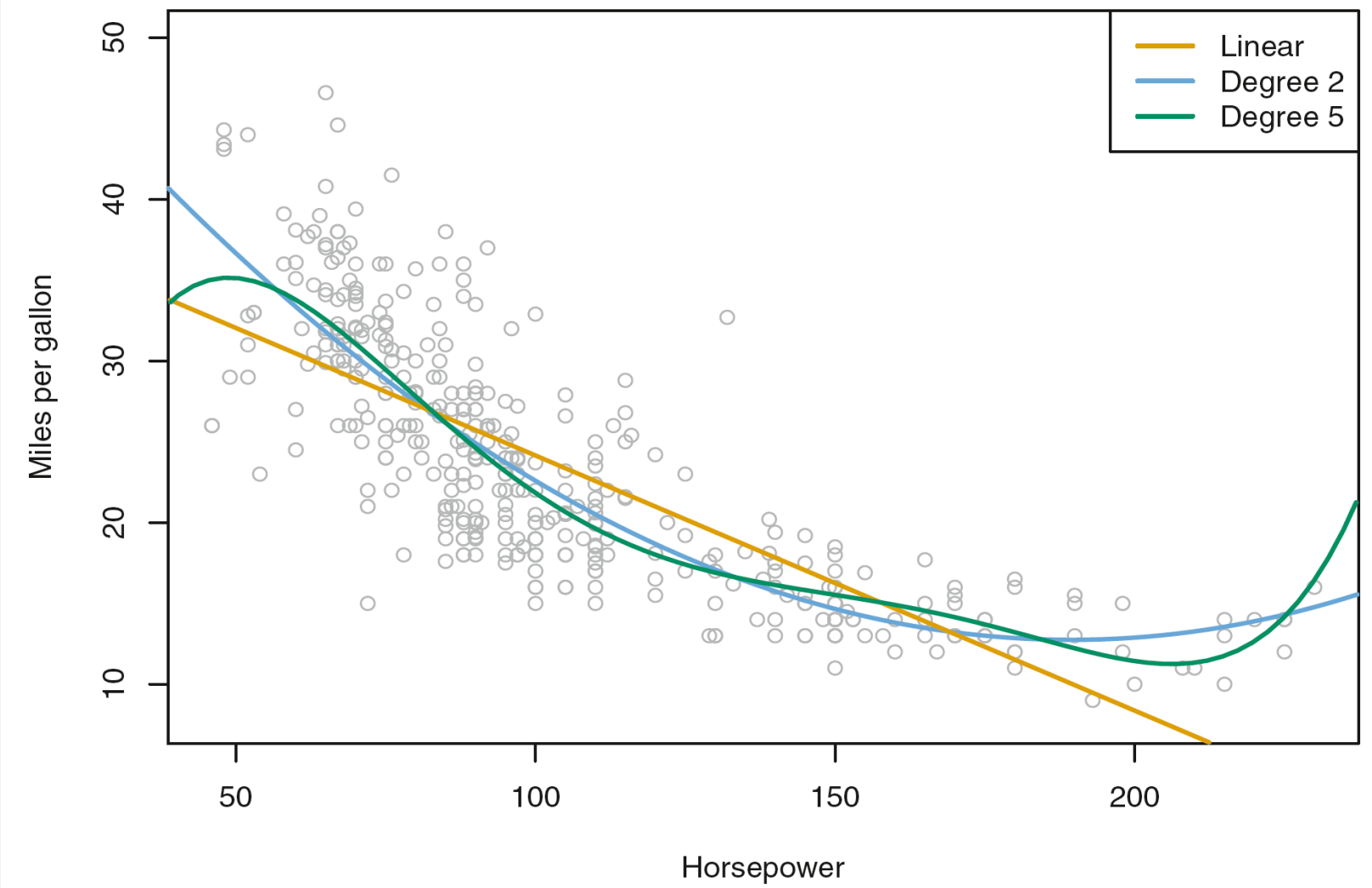



Figure 3.8, *An Introduction to Statistical Learning with Applications in R* by Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani

# Polynomial regression

- Add quadratic
  - $\text{mpg} = \beta_0 + \beta_1 \text{horsepower} + \beta_2 \text{horsepower}^2 + \epsilon$
- $R^2$
  - the quadratic fit = 0 .688
  - the linear fit = 0 .606

| | Coefficient | Std. error | t-statistic | p-value |
|---|---|---|---|---|
| Intercept | 56.9001 | 1.8004 | 31.6 | $< 0.0001$ |
| horsepower | $-0.4662$ | 0.0311 | $-15.0$ | $< 0.0001$ |
| horsepower$^2$ | 0.0012 | 0.0001 | 10.1 | $< 0.0001$ |

# Non-linear Relationships

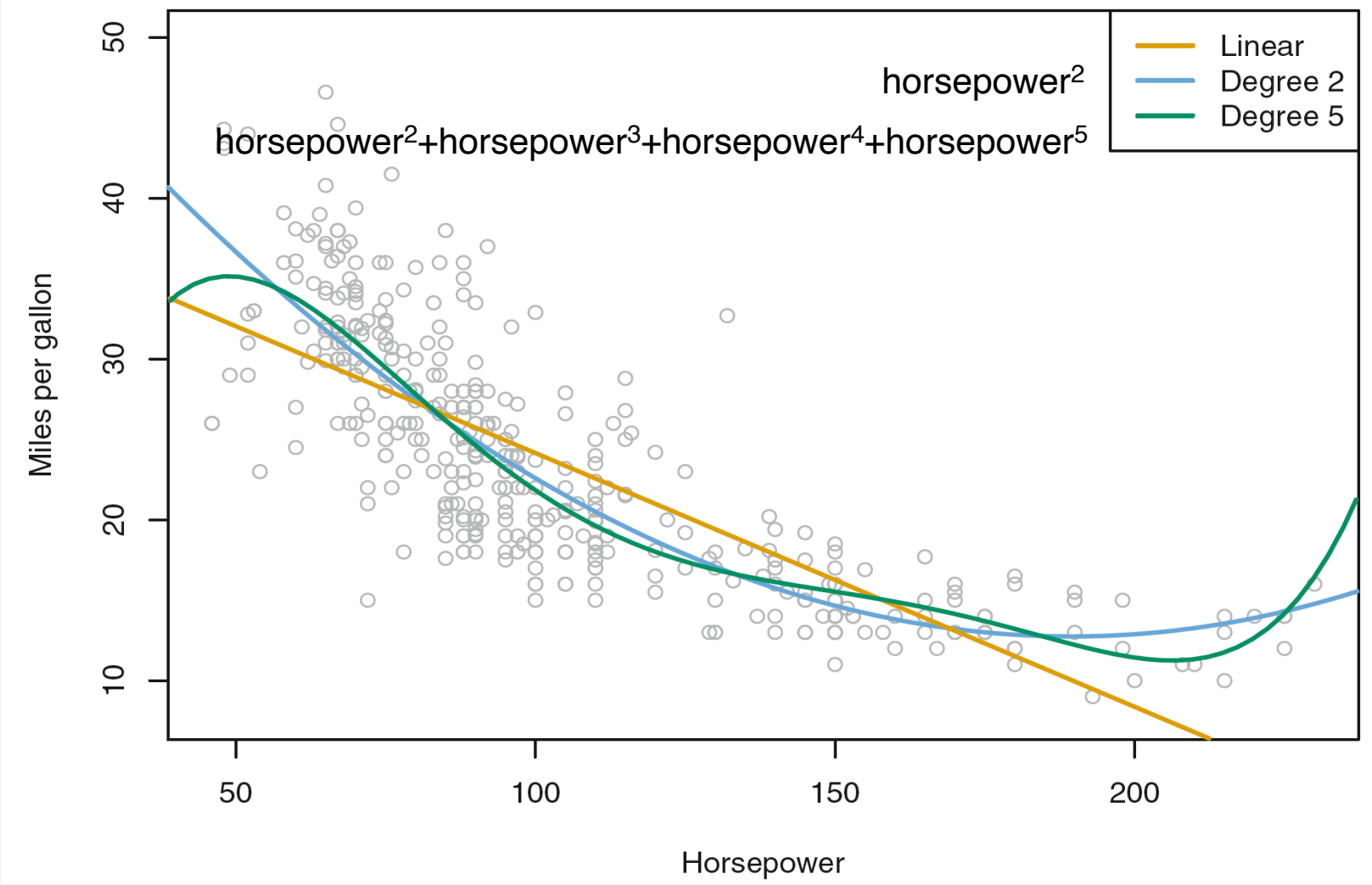
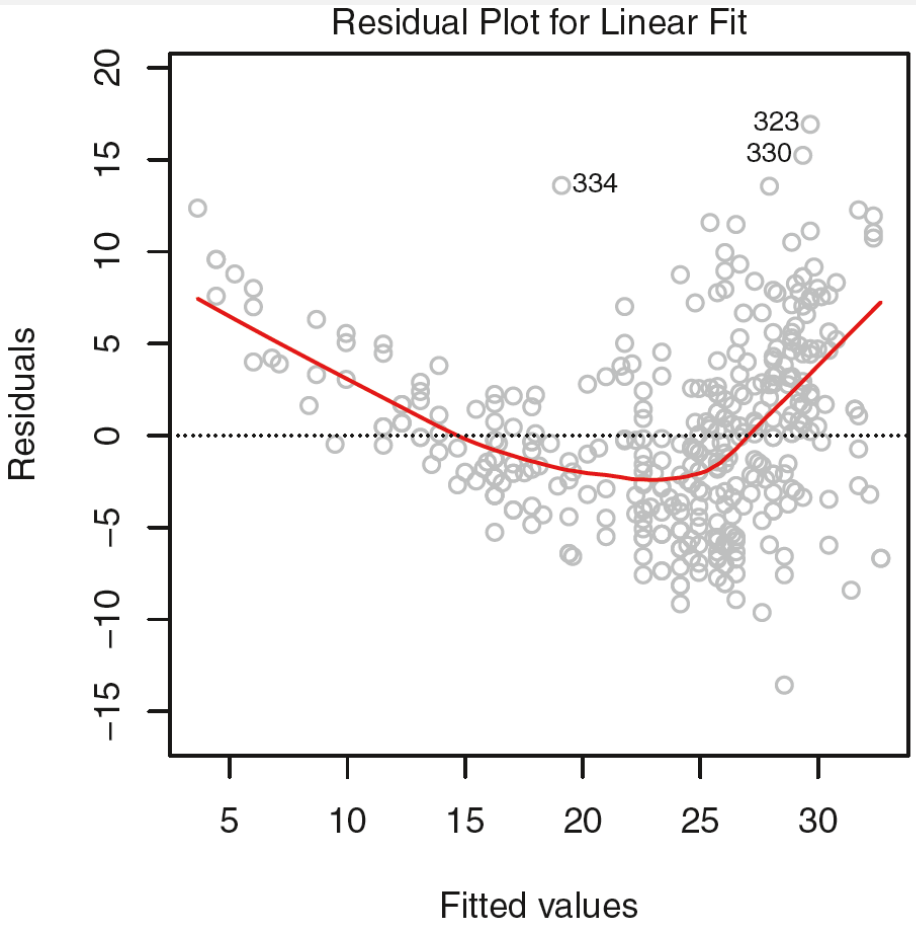



Figure 3.8, *An Introduction to Statistical Learning with Applications in R* by Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani

# Non-linearity of the Data

U-shape

little pattern



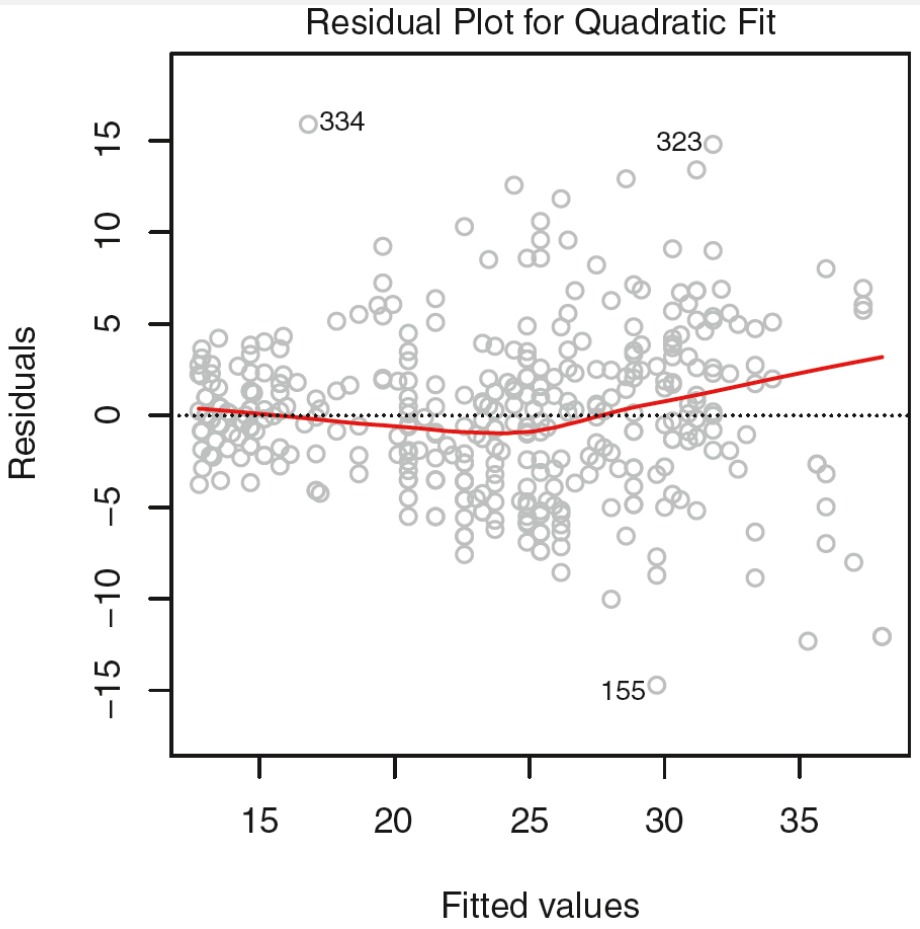




Figure 3.9, *An Introduction to Statistical Learning with Applications in R* by Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani

# Correlation of Error Terms

- $\epsilon_1, \epsilon_2, \cdots, \epsilon_n$ are uncorrelated.
- If they are correlated?
  - we accidentally doubled our data, leading to observations and error terms identical in pairs
  - Our standard error calculations would be as if we had a sample of size 2*n*, when in fact we have only *n* samples
  - confidence intervals will be bias narrower by $\sqrt{2}$

# Non-constant Variance of Error Terms

- the error terms have a constant variance
  - $Var(\epsilon_i) = \sigma^2$
- non-constant variances in the errors
  - a funnel shape in the residual plot

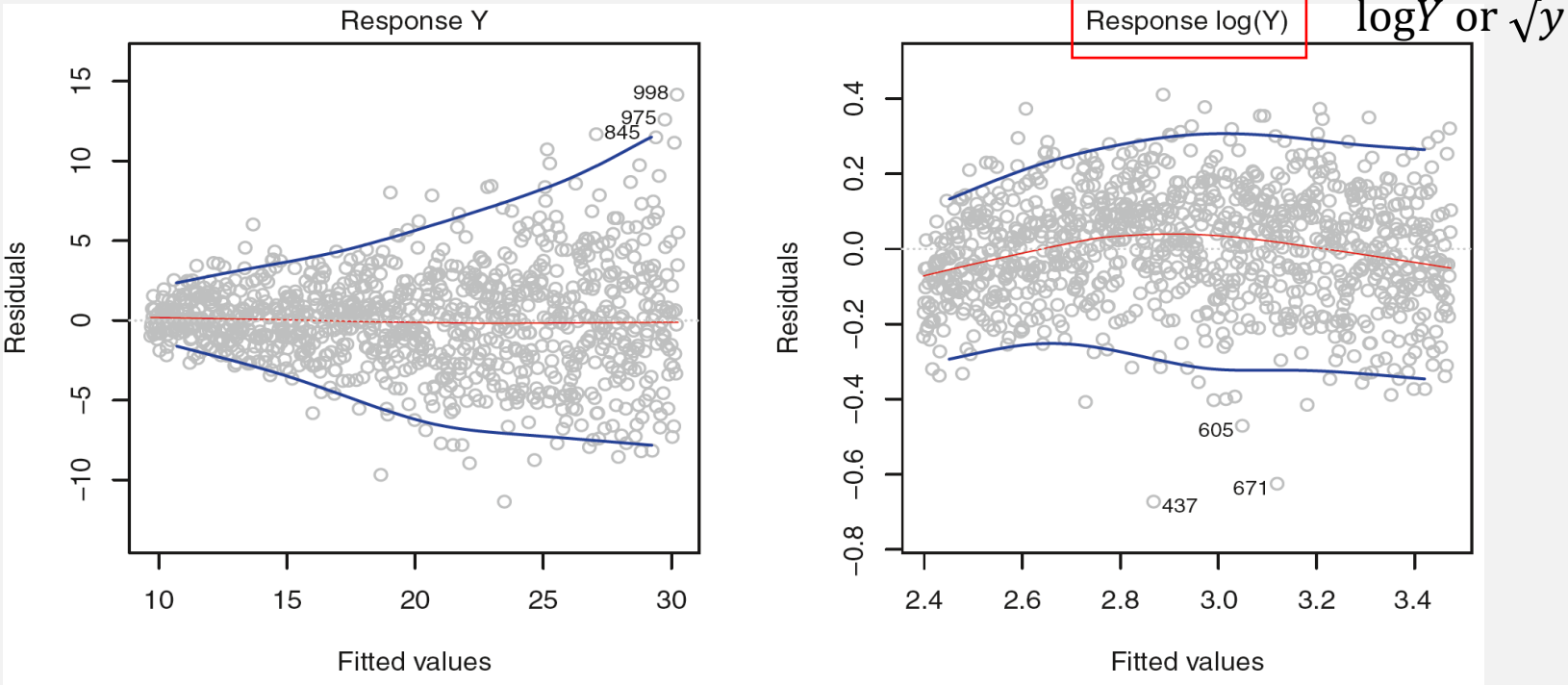
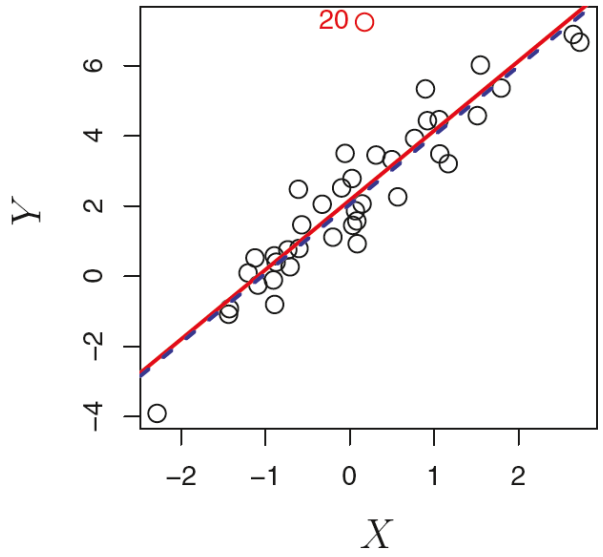
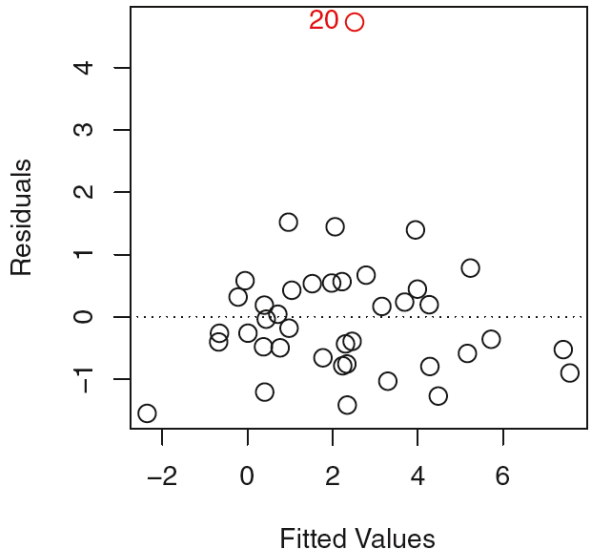



Figure 3.11, *An Introduction to Statistical Learning with Applications in R* by Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani

# Outliers

- Without outlier 20 vs with outlier 20
  - $R^2$: 0.892 vs 0 .805
  - RSE: 1.09 vs 0.77
- Studentized residuals
  - Observations whose absolute studentized residuals >= 3 : possible outliers.



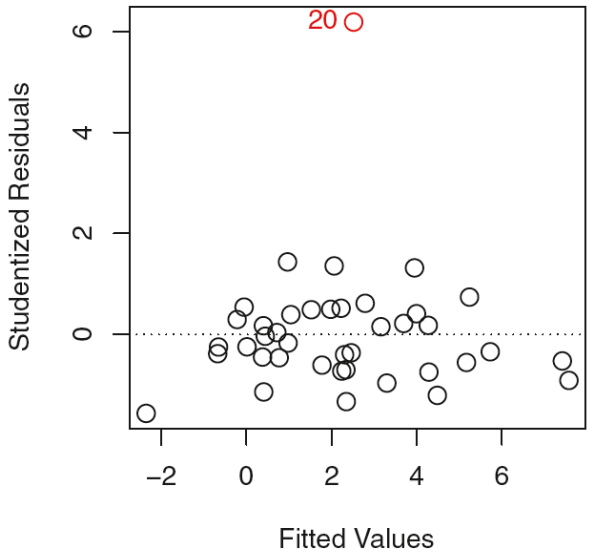




Figure 3.12, *An Introduction to Statistical Learning with Applications in R* by Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani

# High Leverage Points

- Observations with unusual value for $x_i$

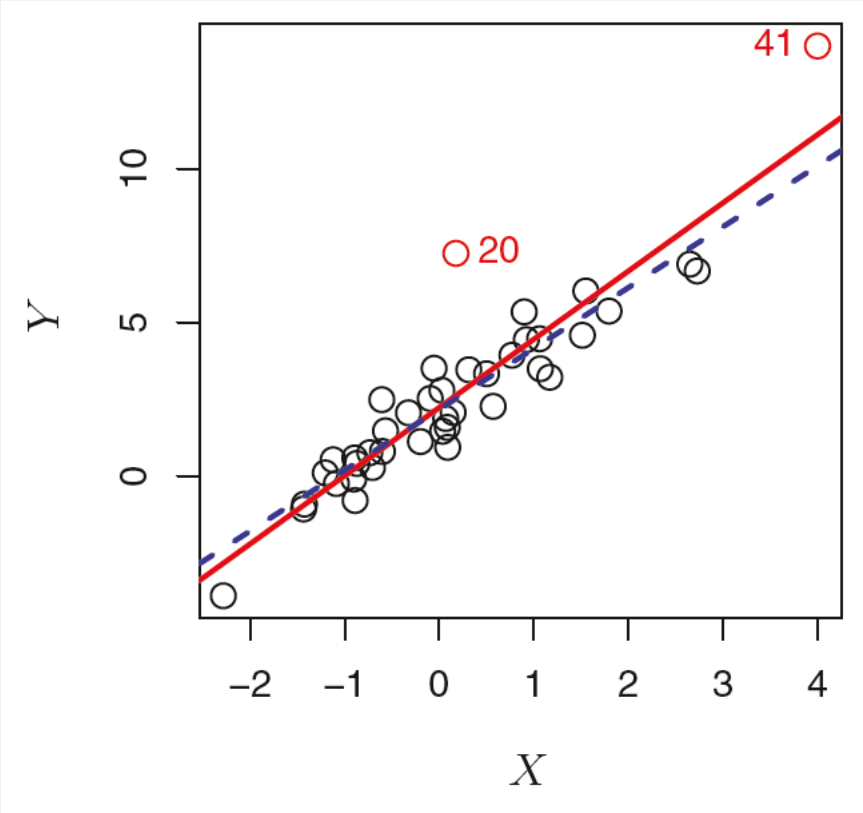



Figure 3.13, *An Introduction to Statistical Learning with Applications in R* by Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani

# High Leverage Points in multiple regression setting

- The red observation is not unusual in terms of its $X_1$ value or its $X_2$ value, but still falls outside the bulk of the data, and hence has high leverage.

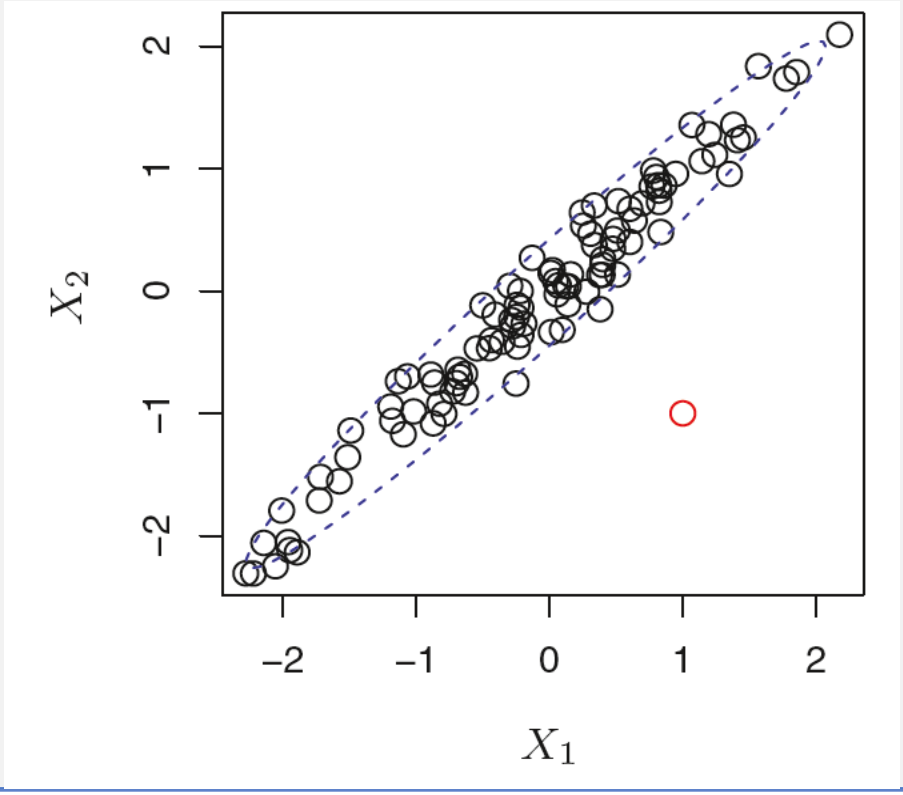



Figure 3.13, *An Introduction to Statistical Learning with Applications in R* by Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani

# Leverage statistic

- $h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i'=1}^{n} (x_{i'} - \bar{x})^2}$
  - range: 1/n ~ 1
  - average = (p+1)/n
  - high leverage : greatly exceeds (p+1)/n

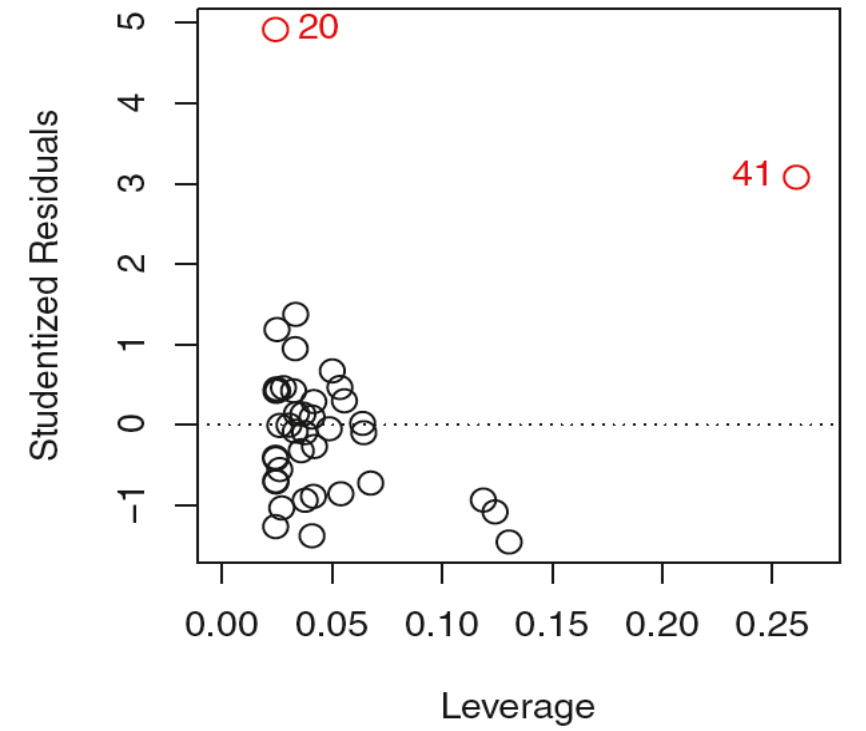
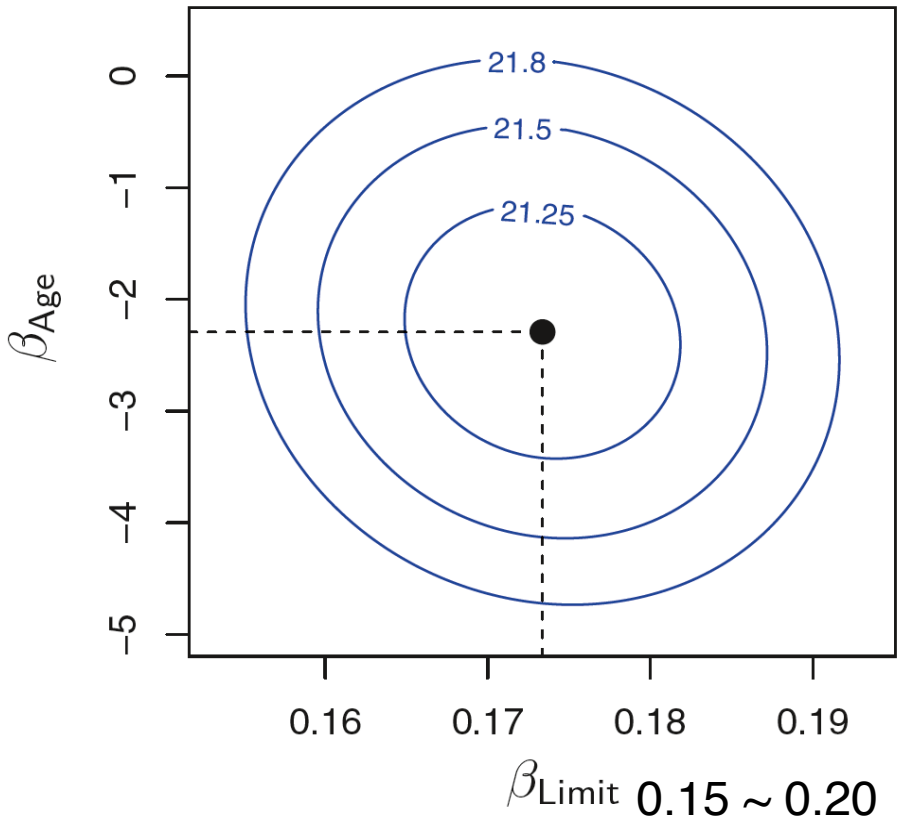



Figure 3.13, *An Introduction to Statistical Learning with Applications in R* by Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani

# Collinearity

- Contour plots for the RSS



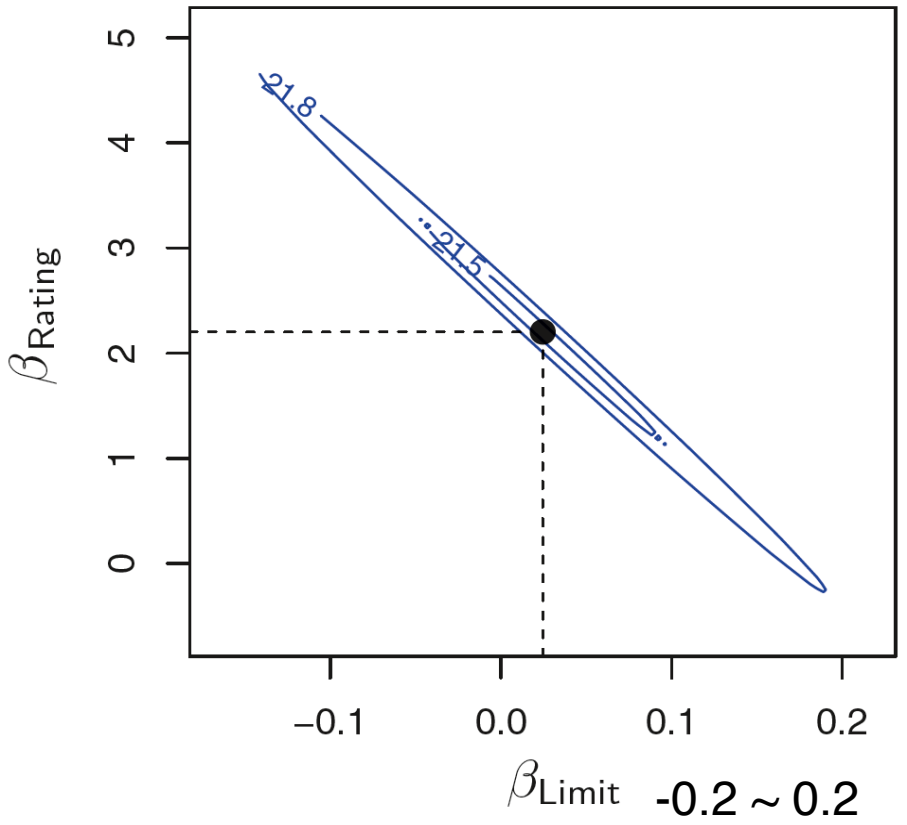




Figure 3.15, *An Introduction to Statistical Learning with Applications in R* by Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani

# Collinearity reduce detecting power

- Since collinearity reduces the accuracy of the estimates of the regression coefficients
- the standard error for $\hat{\beta}_j$ grow
- $t$-statistic for each predictor: dividing $\hat{\beta}_j$ by its standard error => a decline in the $t$-statistic.
- we may fail to reject $H_0$: $\beta_j = 0$
- the probability of correctly power detecting a non-zero coefficient is reduced

# Collinearity reduce detecting power

| | | Coefficient | Std. error | t-statistic | p-value |
|---|---|---|---|---|---|
| Model 1 | Intercept | −173.411 | 43.828 | −3.957 | $< 0.0001$ |
| | age | −2.292 | 0.672 | −3.407 | 0.0007 |
| | limit | 0.173 | 0.005 | 34.496 | $< 0.0001$ |
| Model 2 | Intercept | −377.537 | 45.254 | −8.343 | $< 0.0001$ |
| | rating | 2.202 | 0.952 | 2.312 | 0.0213 |
| | limit | 0.025 | 0.064 | 0.384 | 0.7012 |

# Detect collinearity

- correlation matrix
- multicollinearity
  - variance inflation factor
  - $\mathrm{VIF}(\hat{\beta}_i) = \frac{1}{1-R^2_{x_i|x_{-i}}}$, where $R^2_{x_i|x_{-i}}$ is $R^2$ from a regression of $X_i$ onto all of the other predictors
  - age, rating, and limit with VIF values = 1.01, 160.67, and 160.59

# Solution for collinearity

- drop one of the problematic variables from the regression
  - balance onto age and limit, without rating
- combine the collinear variables together into a single predictor
  - a new variable, credit worthiness: the average of of limit and rating

# Summary with the previous marketing plan

# Is there a relationship between advertising sales and budget?

- Sales $= \beta_0 + \beta_1 \text{TV} + \beta_2 \text{radio} + \beta_3 \text{newspaper} + \varepsilon$
- F-statistic test the hypothesis
  - $H_0: \beta_1 = \beta_2 = \beta_3 = 0$
- the *p*-value corresponding to the F-statistic, 570, is very low
- Reject null hypothesis
- Clear evidence of a relationship between advertising and sales.

# How strong is the relationship?

- RSE: the standard deviation of the response from the population regression line
  - 1681 units, the mean value for the response=14022 => percentage error of roughly 12%
- $R^2$: the percentage of variability in the response that is explained by the predictors
  - 0.897: almost 90% of the variance in sales

# Which media contribute to sales?

- the *p*-values associated with each predictor's *t*-statistic

| | Coefficient | Std. error | t-statistic | p-value |
|---|---|---|---|---|
| Intercept | 2.939 | 0.3119 | 9.42 | $< 0.0001$ |
| TV | 0.046 | 0.0014 | 32.81 | $< 0.0001$ |
| radio | 0.189 | 0.0086 | 21.89 | $< 0.0001$ |
| newspaper | −0.001 | 0.0059 | −0.18 | 0.8599 |

# How large is the effect of each medium on sales?

- the standard error of $\widehat{\beta_i}$ can be used to construct confidence intervals for $\beta_i$
- the 95% confidence intervals
  - TV: (0.043, 0.049)
  - Radio: (0.172, 0.206)
    - TV & radio narrow and far from zero => related to sales
  - Newspaper: (−0.013, 0.011)
    - includes zero, not statistically significant given the values of TV & radio

# How large is the effect of each medium on sales?

- Could collinearity be the reason that the confidence interval associated with newspaper is so wide?
- VIF scores:
  - TV: 1. 005
  - Radio: 1. 145
  - Newspaper: 1. 145
  - no evidence of collinearity

# How large is the effect of each medium on sales?

- three separate simple linear regressions

| | Coefficient | Std. error | t-statistic | p-value |
|---|---|---|---|---|
| Intercept | 7.0325 | 0.4578 | 15.36 | $< 0.0001$ |
| TV | 0.0475 | 0.0027 | 17.67 | $< 0.0001$ |

| | Coefficient | Std. error | t-statistic | p-value |
|---|---|---|---|---|
| Intercept | 9.312 | 0.563 | 16.54 | $< 0.0001$ |
| radio | 0.203 | 0.020 | 9.92 | $< 0.0001$ |

| | Coefficient | Std. error | t-statistic | p-value |
|---|---|---|---|---|
| Intercept | 12.351 | 0.621 | 19.88 | $< 0.0001$ |
| newspaper | 0.055 | 0.017 | 3.30 | 0.00115 |

# How accurately can we predict future sales?

- predict an individual response, $Y = f(X) + \epsilon$
  - prediction interval
- the average response, $f(X)$
  - confidence interval
- Prediction intervals will always be wider than confidence intervals because they account for the uncertainty associated with $\epsilon$, the irreducible error.

# Is the relationship linear?

- the residual plots should display no pattern => linear relationship

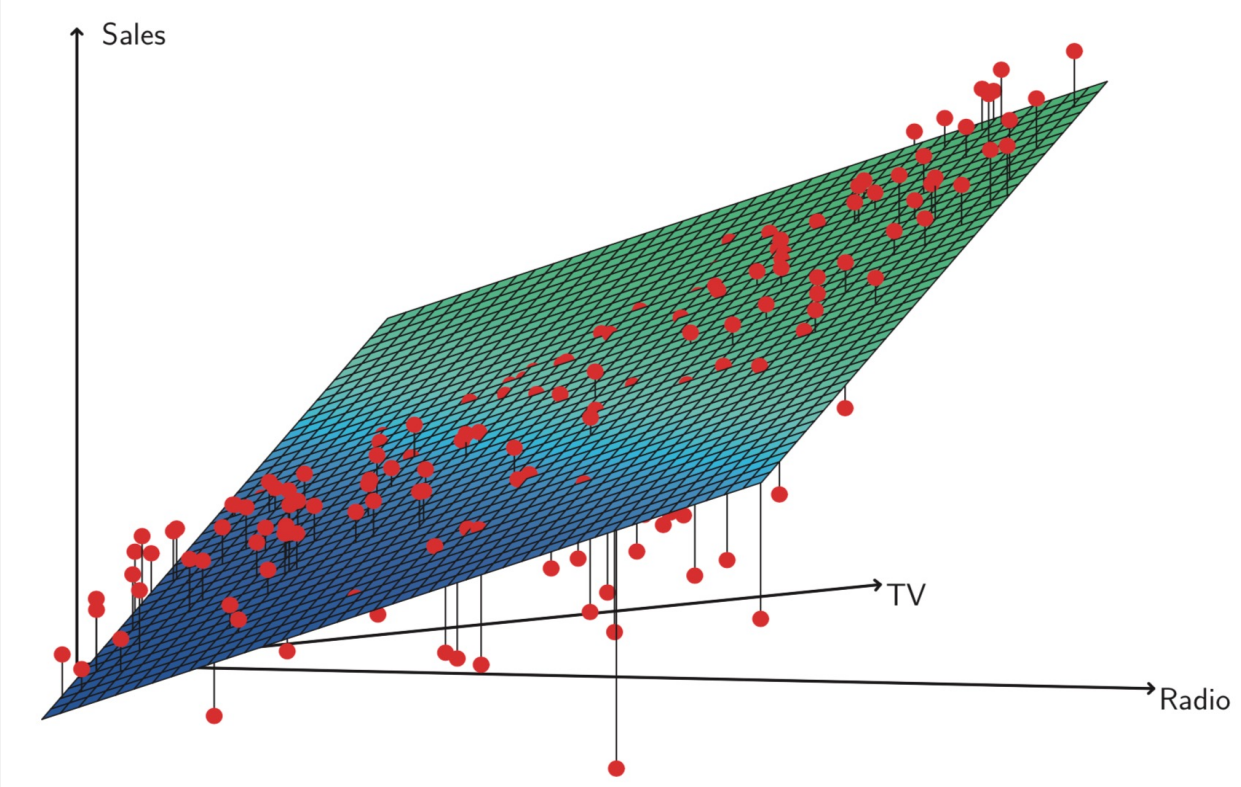



Figure 3.5, *An Introduction to Statistical Learning with Applications in R* by Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani

# Is there synergy among the advertising media?

- *p*-value associated with the interaction term
- $R^2$: ~90% to almost 97%, including an interaction term in the model

# Comparison of Linear Regression with *K*-Nearest Neighbors

- A parametric approach: a linear functional form for $f(X)$
- non-parametric: do not explicitly assume a parametric form for $f(X)$
  - *K*-nearest neighbors regression (KNN-regression)

# The KNN regression

- Given a value for $K$ and a prediction point $x_0$
  - identifies the $K$ training observations that are closest to $x_0$, represented by $N_0$
  - estimates $f(x_0)$ using the average of all the training responses in $N_0$
  - $\hat{f}(x_0) = \frac{1}{K}\sum_{x_{i\in N_0}} y_i$

# Plots of $\hat{f}(x)$using KNN regression on a two-dimensional data

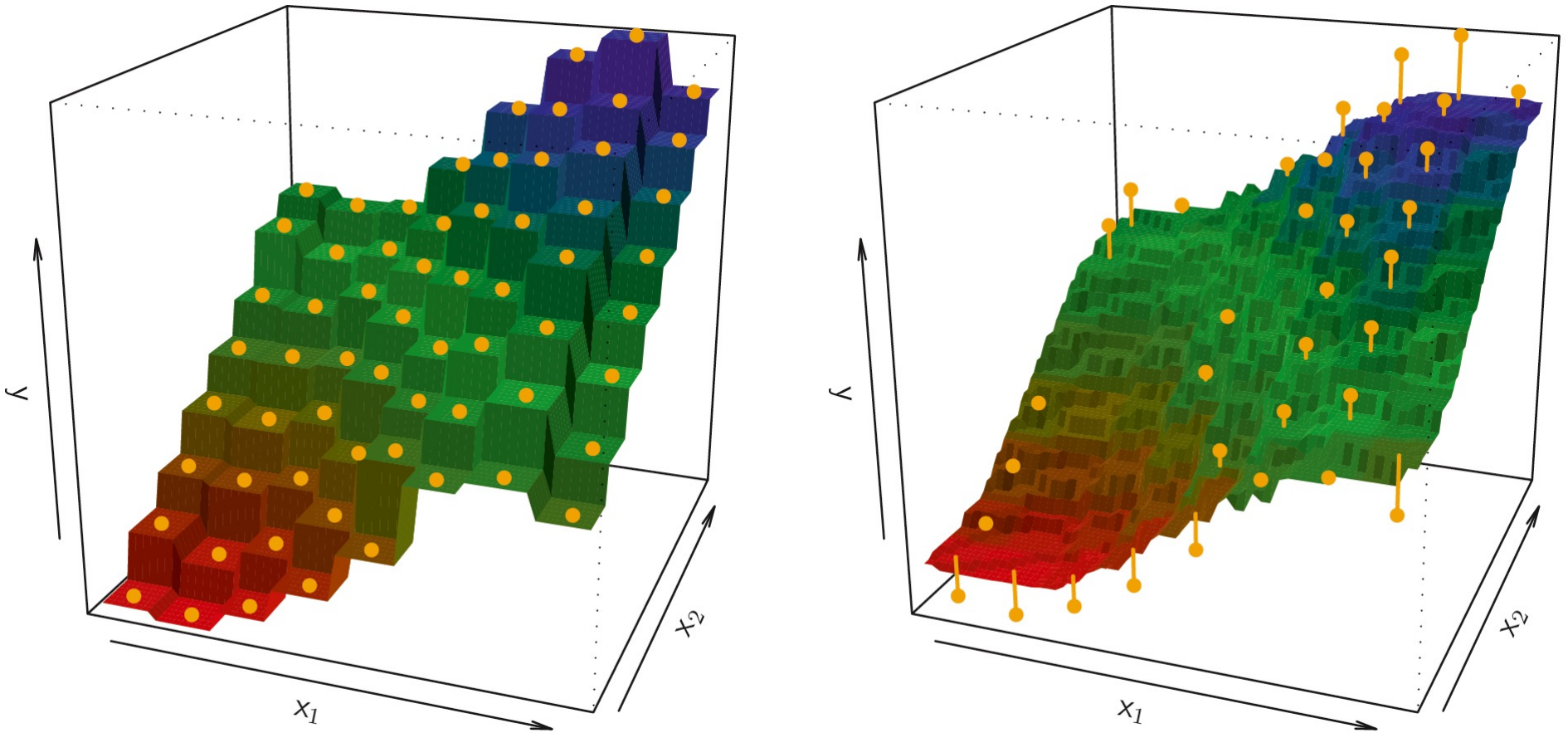



Figure 3.16, *An Introduction to Statistical Learning with Applications in R* by Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani

# When least squares linear regression outperform KNN regression?

- the parametric approach will outperform the nonparametric approach if the parametric form that has been selected is close to the true form of *f*.

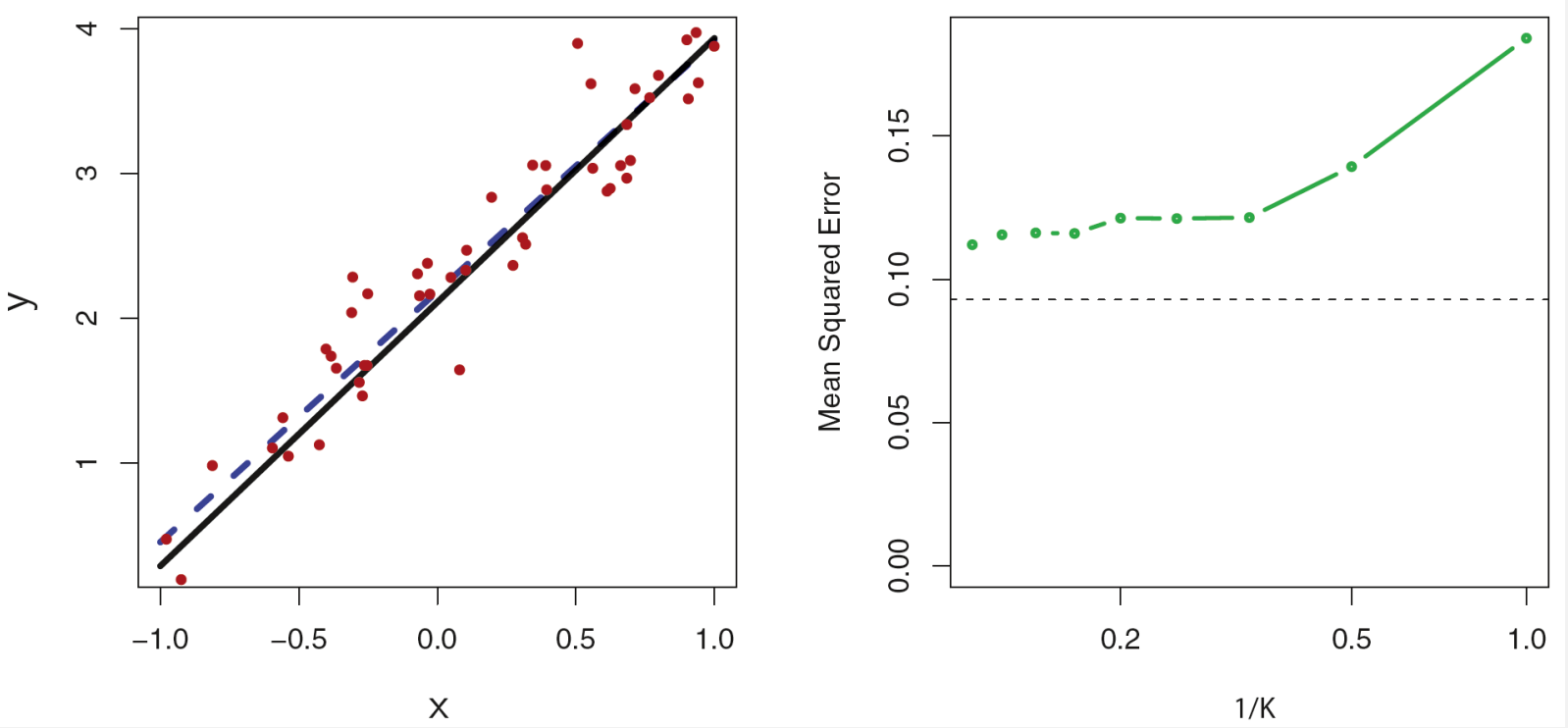



Figure 3.18, *An Introduction to Statistical Learning with Applications in R* by Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani

# KNN performs better than linear regression

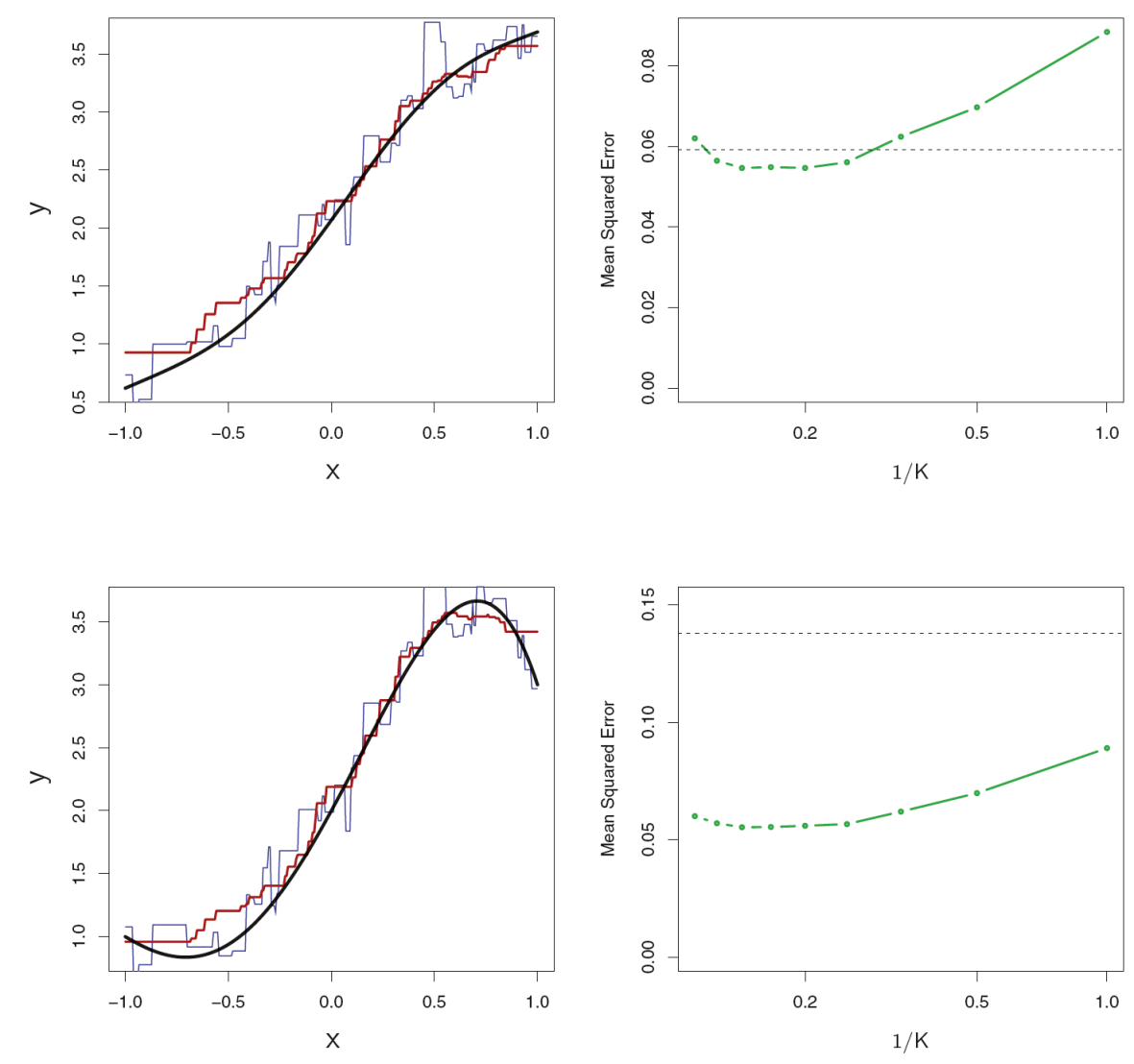



Figure 3.19, *An Introduction to Statistical Learning with Applications in R* by Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani

# Curse of dimensionality

- spreading 100 observations over 20 dimensions
- additional noise *p* predictors

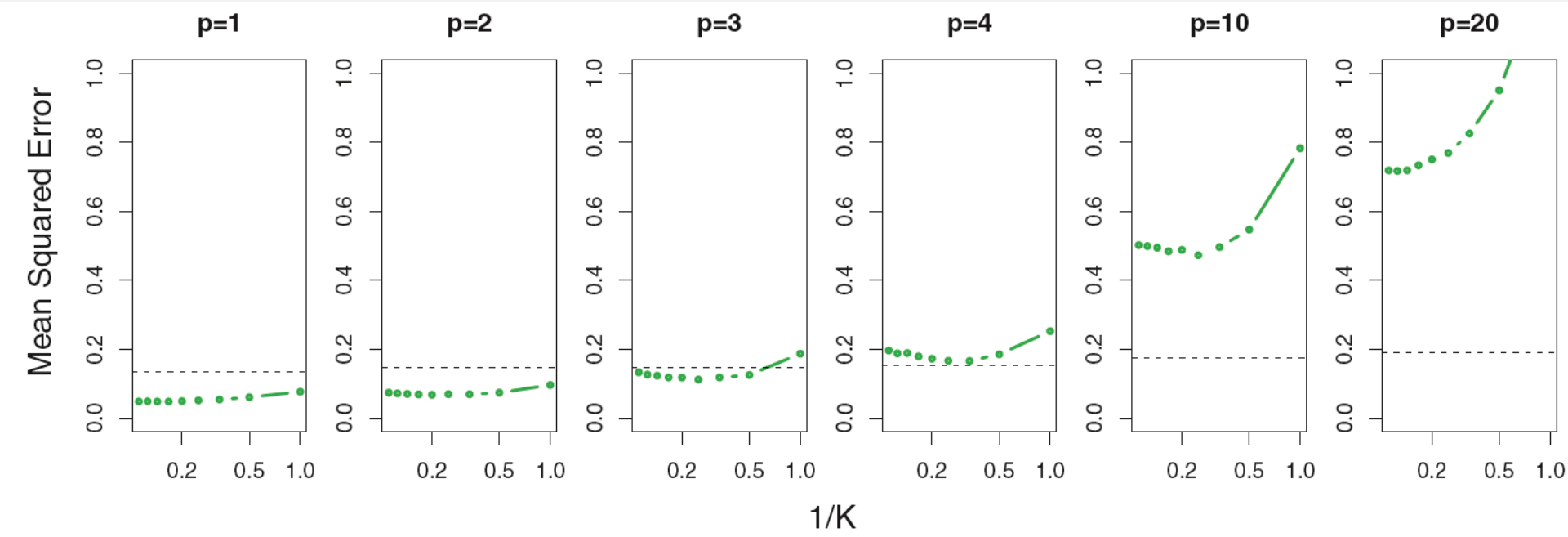



Figure 3.20, *An Introduction to Statistical Learning with Applications in R* by Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani

# Observation ~ predictor

- This decrease in performance as the dimension increases is a common problem for KNN
- Parametric methods will tend to outperform non-parametric approaches when there is a small number of observations per predictor
- The dimension is small, we might prefer linear regression to KNN from an interpretability standpoint.

# Summary

- Linear regression is the go-to statistical modeling method for quantities.
- You should always try linear regression first, and only use more complicated methods if they actually outperform a linear regression model.
- Linear regression will have trouble with problems that have a very large number of variables, or categorical variables with a very large number of levels.
- You can enhance linear regression by adding new variables or transforming variables
  - like we did with the log() transform of y , but always be wary when transforming y as it changes the error model.

# Summary

- With linear regression, you think in terms of residuals. You look for variables that correlate with your errors and add them to try and eliminate systematic modeling errors.
- Linear regression can predict well even in the presence of correlated variables, but correlated variables lower the quality of the advice.
- Overly large coefficient magnitudes, overly large standard errors on the coefficient estimates, and the wrong sign on a coefficient could be indications of correlated inputs.
- Linear regression packages have some of the best built-in diagnostics available, but rechecking your model on test data is still your most effective safety check.


AITC
教育部人工智慧技術及應用人才培育計畫
Artificial Intelligence Talent Cultivation Program