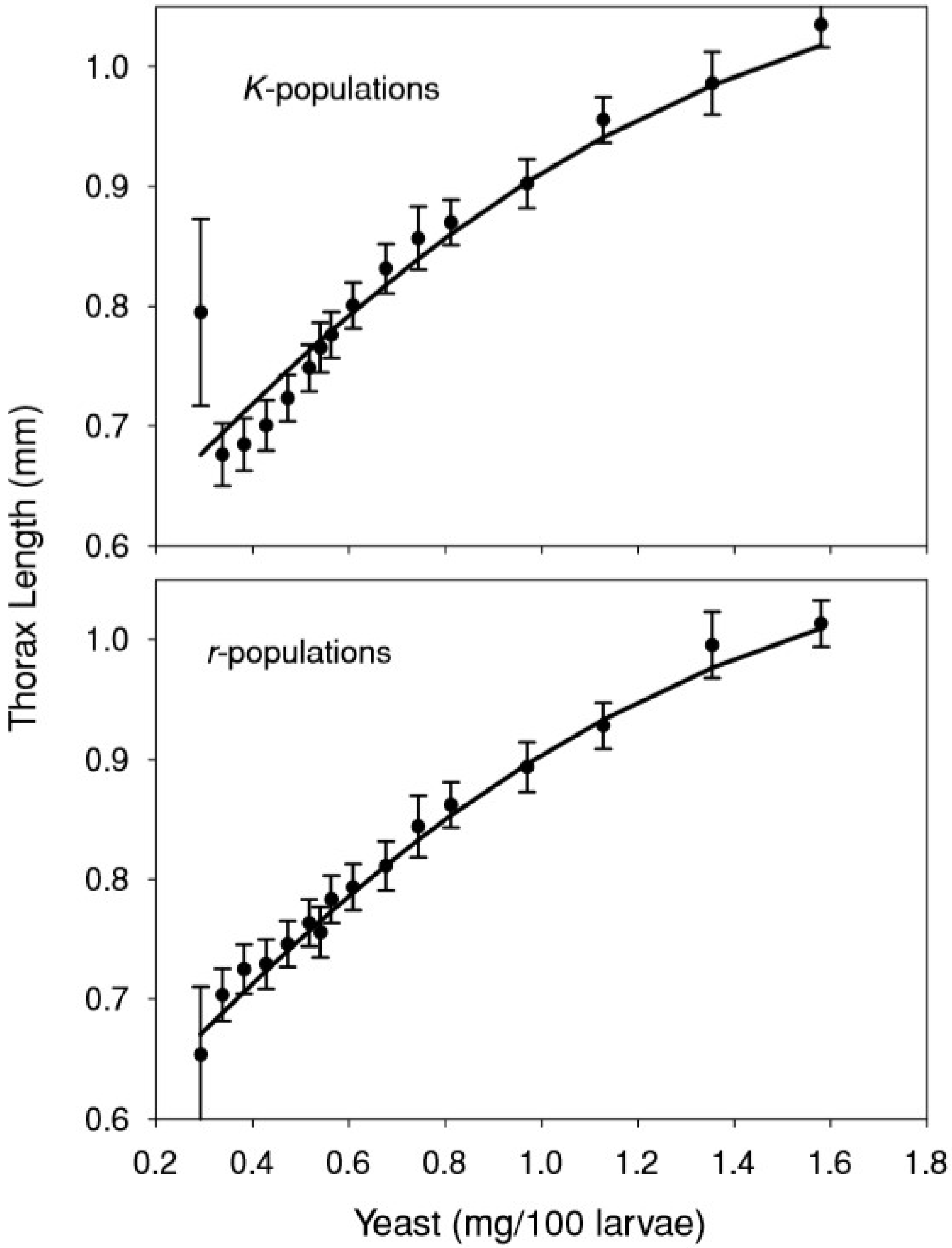
Chapter 3: Linear Regression

- Why do regression?
- Prediction and understanding.
- Linear regression is used as long as the coefficients enter in linearly like, $a_0 + a_1x + a_2x^2$
- But not, $a_0 + e^{a_1x}$



Estimating the coefficients

- Data look like this, (x_1,y_1), (x_2,y_2),…, (x_n,y_n)
- Fit the linear model, $y_i = \beta_0 + \beta_1 x_i$
- Choose estimates, $\hat{\beta}_0\,\hat{\beta}_1$, such that the sum of squared residuals, $r_i = \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)$, are minimized.
- Residual Sum of Squares = RSS $= r_1^2 + \cdots + r_n^2$
- The least squares estimates can be found by applying some standard calculus, which we do for the general case in the next slide.

Least Squares Estimates

- The general linear model is, $y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$, for *i*=1,…,*n*
- In matrix notation this is written, $\boldsymbol{y}=\boldsymbol{X}\beta$, where $\boldsymbol{y}$ is $n\times1$ and β is $(p+1)\times1$ and $\boldsymbol{X}$ is $n\times(p+1)$ ($\boldsymbol{X}$ called the design matrix)
- Fact: $h(\boldsymbol{x}) = (\boldsymbol{a} - \boldsymbol{c}\boldsymbol{x})^T K(\boldsymbol{a} - \boldsymbol{c}\boldsymbol{x})$ then $\frac{dh(\boldsymbol{x})}{d\boldsymbol{x}} = -2\boldsymbol{c}^T K(\boldsymbol{a} - \boldsymbol{c}\boldsymbol{x})$
- RSS=$(\boldsymbol{y}\text{-}\boldsymbol{X}\beta)^T(\boldsymbol{y}\text{-}\boldsymbol{X}\beta)=(\boldsymbol{y}\text{-}\boldsymbol{X}\beta)^T\mathbf{I}(\boldsymbol{y}\text{-}\boldsymbol{X}\beta)$, now use Fact to find derivative
- $\frac{dRSS}{d\boldsymbol{\beta}} = -2\boldsymbol{X}^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$, set this to 0 and solve for β
- $\boldsymbol{X}^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) = 0$
  $\boldsymbol{X}^T\boldsymbol{y}$- $\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta}$=0
  $\boldsymbol{X}^T\boldsymbol{y}$= $\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta}$
- $(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$= $(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta}$
- $(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$=$\widehat{\boldsymbol{\beta}}$

Variance Estimates of the least squares parameters

- Facts:
- 1. $E(a^T y) = a^T E(y) and\ Var(a^T y) = \ a^T Var(y) a$
- 2. $(A^T)^{-1} = (A^{-1})^T$
- 3. $(AB)^T \ = B^T A^T$
- $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{y}$
- $Var\left(\widehat{\boldsymbol{\beta}}\right) = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T Var(\boldsymbol{y}) [(\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T]^T,$ by Fact 1
  $= (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T [(\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T]^T \sigma^2$, since σ^2 is a constant
- $= (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \{\boldsymbol{X}[(\boldsymbol{X}^T \boldsymbol{X})^{-1}]^T\} \sigma^2$, by Fact 3
- $= (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \{\boldsymbol{X}[(\boldsymbol{X}^T \boldsymbol{X})^T]^{-1}\} \sigma^2$, by Fact 2
- $= (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{X}\ (\boldsymbol{X}^T \boldsymbol{X})^{-1} \sigma^2$, by Fact 3
- $= (\boldsymbol{X}^T \boldsymbol{X})^{-1}\ \sigma^2$, since $(\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{X} = \mathbf{I}$

  where σ=$\sqrt{\frac{1}{n-p-1} RSS}$

How to solve for $\hat{\beta}$

- Although the equation above suggests that we would use the inverse of $(X^TX)^{-1}$ to find the estimates of β in practice this would not be done. If the condition number of X^TX is very large then the solution to the Linear equations above will be sensitive to round off errors and these will be magnified if we simply take the inverse of X^TX.
- Rather you should use methods with orthogonal factorization, partial pivoting and compact elimination, which I had to do in the old days (page 463 of Mueller, L.D., F. González-Candelas and V.F. Sweet, 1991. Components of density-dependent population dynamics: models and tests with Drosophila. The American Naturalist 137: 457).
- Nowadays, thanks to R we can use functions like "solve" that will implement these techniques.

Example of solving linear equations the wrong way

```
H10<- NULL #create an ill-conditioned Hilbert matrix. The condition number
of the 10x10 matrix if 35 x 10^6, while the condition number for the
identity matrix is 1. Condition number=||H10||_1 ||H10^-1||_1
n<- 10
for (i in 1:n){
temp<- NULL
for (j in 1:n)
temp<- c(temp,1/(i+(j-1)))
H10<- rbind(H10,temp)}
a<- 1:n #This is going to be the unknown vector
b<- H10%*%a #This is the right side of the equation: H10%*%a=b
sol.1<- solve(H10,b) #Solve for a
#ANSWER:
> sol.1
[,1]
[1,] 1.000000
[2,] 2.000000
[3,] 3.000009
[4,] 3.999920
[5,] 5.000381
[6,] 5.998951
[7,] 7.001725
[8,] 7.998330
[9,] 9.000878
[10,] 9.999807
```

```
Library(MASS)
H10.i<- ginv(H10)# Estimate the inverse of H10
H10.i%*%b #Solve for a
 [,1]
 [1,] 1.000002
 [2,] 1.999910
 [3,] 3.000945
 [4,] 3.996272
 [5,] 5.005708
 [6,] 5.999289
 [7,] 6.994882
 [8,] 8.000046
 [9,] 9.006133
[10,] 9.996813
```

Solve SSE = 8 e-06
Inverse SSE = 122 e-06

H10

1.00	0.50	0.33	0.25	0.20	0.17	0.14	0.13	0.11	0.10
0.50	0.33	0.25	0.20	0.17	0.14	0.13	0.11	0.10	0.09
0.33	0.25	0.20	0.17	0.14	0.13	0.11	0.10	0.09	0.08
0.25	0.20	0.17	0.14	0.13	0.11	0.10	0.09	0.08	0.08
0.20	0.17	0.14	0.13	0.11	0.10	0.09	0.08	0.08	0.07
0.17	0.14	0.13	0.11	0.10	0.09	0.08	0.08	0.07	0.07
0.14	0.13	0.11	0.10	0.09	0.08	0.08	0.07	0.07	0.06
0.13	0.11	0.10	0.09	0.08	0.08	0.07	0.07	0.06	0.06
0.11	0.10	0.09	0.08	0.08	0.07	0.07	0.06	0.06	0.06
0.10	0.09	0.08	0.08	0.07	0.07	0.06	0.06	0.06	0.05

Assessing Accuracy

- The residual standard error or RSE=σ, measure of the lack of fit.
- R^2 is the proportion of explained variance.
- Let the total sum of squares (TSS) be $\sum(y_i - \bar{y})^2$ (if the feature variables provided no information to predict *y*, we would just use the average as our best guess.
- Then, $R^2 = \frac{TSS-RSS}{TSS}$
- For a single independent variable R^2=ρ^2, where ρ is the correlations coefficient between *X* and *Y*.

Is there a relationship between the response and predictors?

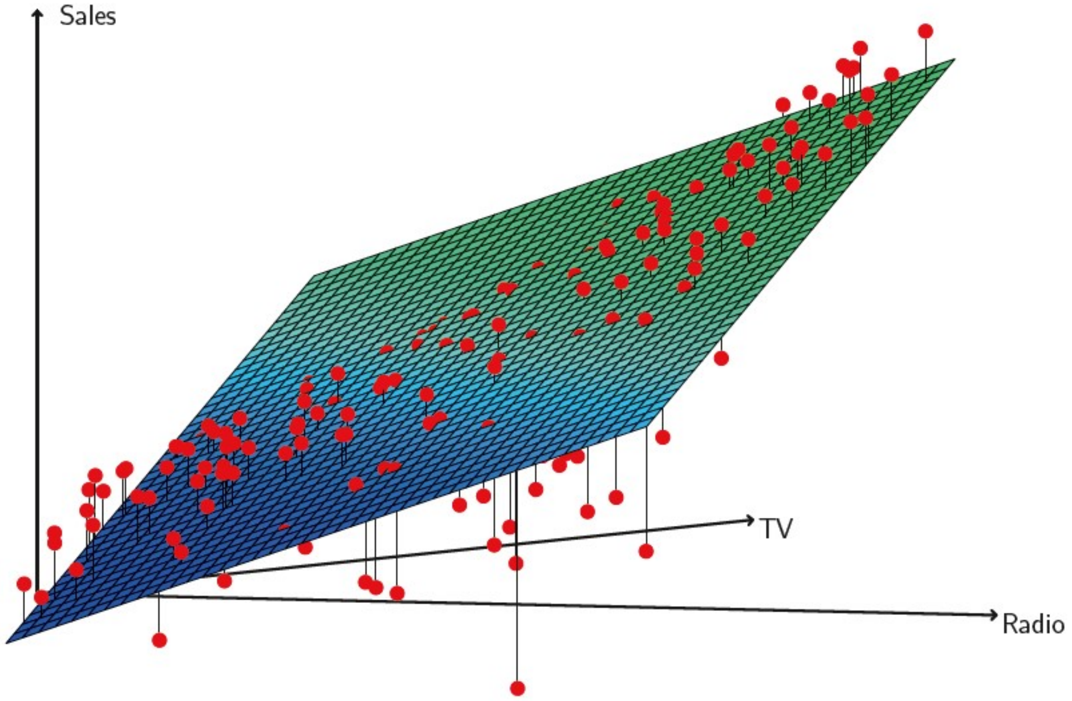
- We can test all parameters; $H_0\colon \beta_1 = \cdots = \beta_p = 0$ with an *F*-test, $F = \frac{(TSS-RSS)/p}{RSS/(n-p-1)}$, which has $F_{p,n\text{-}p\text{-}1}$ distribution. Under the null hypothesis F~1. If any |β|>0, F>1.
- We can get these in R as: if *F*=2, *n*=50, and *p*=10, then pf(2,10,39) = 0.94
- Alternatively we could test if the last q predictors are 0 with, $F = \frac{(RSS_0-RSS)/q}{RSS/(n-p-1)}$ where RSS_0 is fit to the *p-q* predictors.
- Each *F* test will have a type-I error of 5%. However, examining the confidence interval on each parameter involves multiple testing
- Of course, if *p*>*n* then we can't do these tests either.

Which variables matter?

- We could look at every possible model and assess their goodness of fit with Mallow's C_p, Akaike information criteria, etc. These methods provide a penalty for adding superfluous variables.
- There are 2^p different models to test. When $p=20$, there are over one million models. So, this is not practical.
- Systematic methods like forward and backward selection can be used, although backward selection can't be used when $p>n$.

Model Fit

- In general R^2 will not be a good indicator since it always increases as we add parameters to the model.
- In some cases plotting the data and model fit can reveal problems.
- In this figure sales using mostly TV or radio seem to be overestimated and sales using both underestimated. This suggests a non-linearity the linear model can't pick up.



Predictions

- Two ways to express uncertainty in predictions.
- I. Confidence intervals , (c_1,c_2), to address how close $\hat{Y}$ is to $f(X)$. Thus, upon repeated collections of data from this population we expect 95% of the predictions to include the true f(X).
- II. Prediction intervals, (p_1,p_2), which also include the irreducible error. Thus, upon repeated collections of data from this population we expect 95% of the predictions, $\hat{Y}$, to include the true Y.

Qualitative Predictors

- Two levels, x_i = sex of *ith* person where males (i=0) or females (i=1).
- Or there could be multiple levels: food level, low (i=0), medium (i=1) or high (i=2). Thus, y_i, could be the fecundity of a female receiving the *ith* level of food.
- For this last example the actual model would be for female j at food level i:

$$y_{ij} = \beta_0 + \delta_i \beta_i + \epsilon_j$$

- Where δ_i =0 if i=0 and 1 otherwise. Thus, the model takes on 3 forms: β_0 when i=0, $\beta_0 + \beta_1$ when i=1, and $\beta_0 + \beta_2$ when i=2.

Extending the linear model

- Removing completely additive effects

$$\begin{aligned} Y &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 \\ &= \beta_0 + (\beta_1 + \beta_3 X_2) X_1 + \beta_2 X_2 \\ &= \beta_0 + \tilde{\beta}_1 X_1 + \beta_2 X_2 \end{aligned}$$

- The effect of X_1 on Y now depends on the value of X_2 since $\tilde{\beta}_1$ depends on the value of X_2.
- We can also make Y nonlinear by using polynomials, $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2$
- However, X_1 and X_1^2 are not independent.

Problem: non-linearity

- We may detect non-linearities and non-constant variances by looking at the residuals, e.g. $y_i - \widehat{y}_i$ as a function of $\widehat{y}_i$.

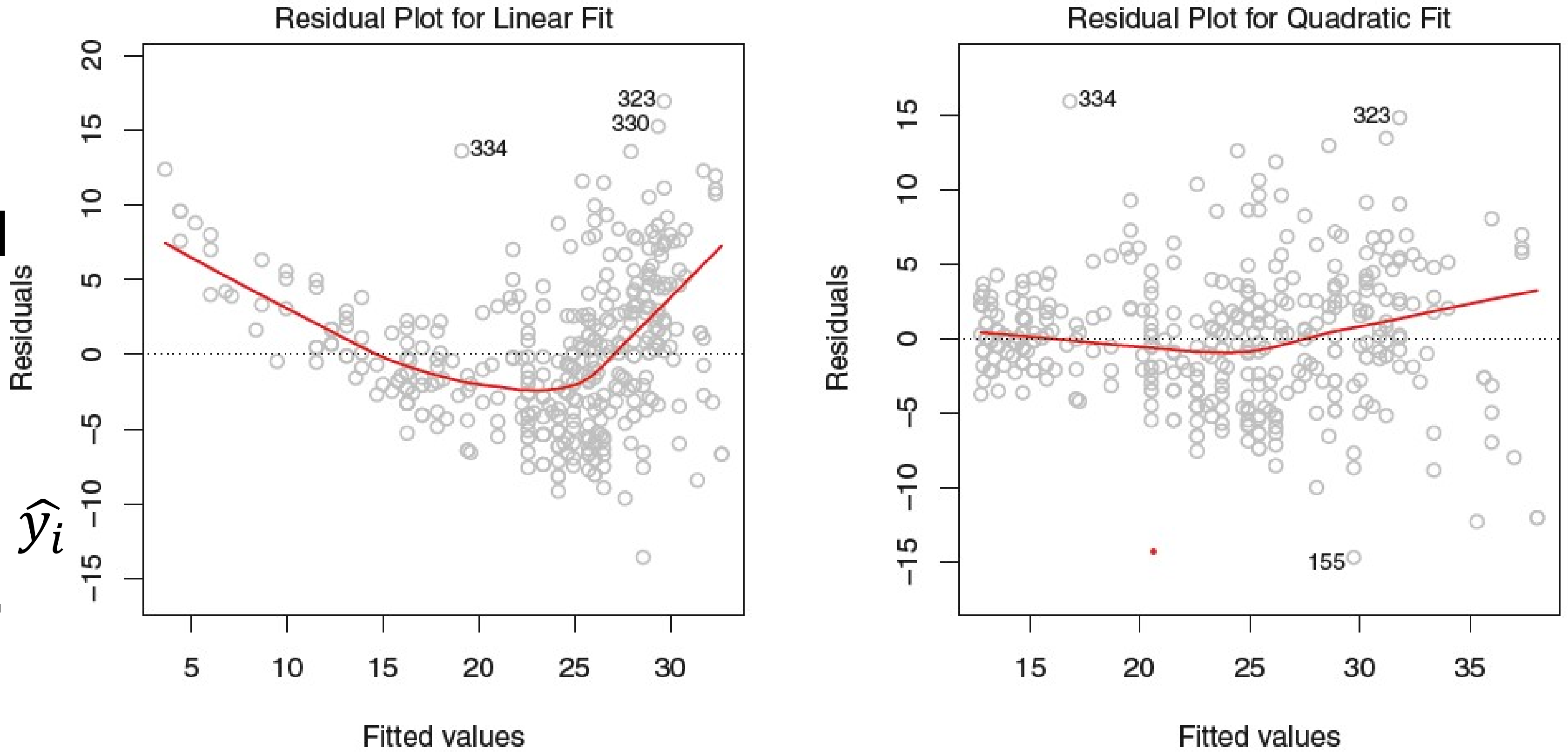


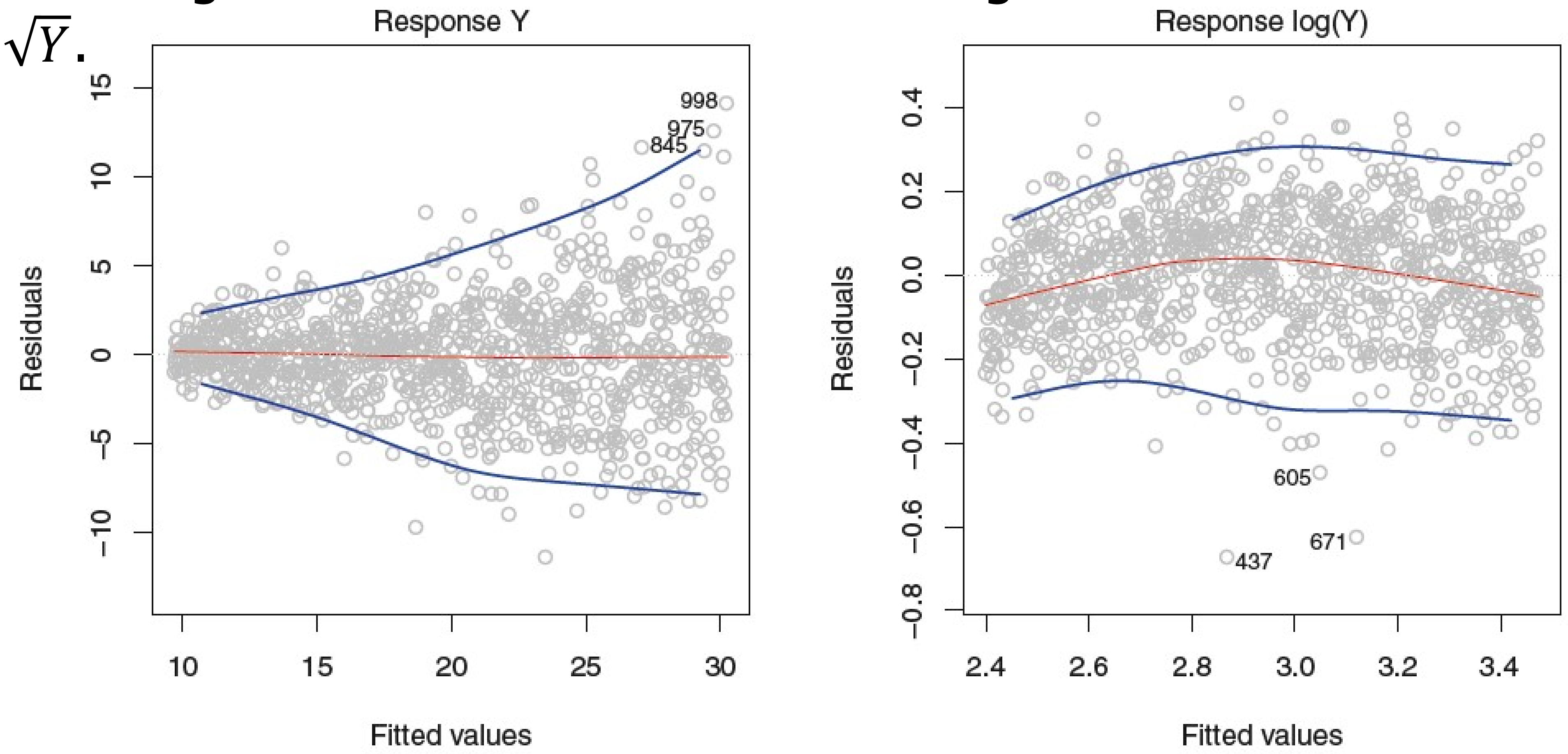
FIGURE 3.9. *Plots of residuals versus predicted (or fitted) values for the* Auto *data set. In each plot, the red line is a smooth fit to the residuals, intended to make it easier to identify a trend.* Left: *A linear regression of* mpg *on* horsepower. *A strong pattern in the residuals indicates non-linearity in the data.* Right: *A linear regression of* mpg *on* horsepower *and* horsepower^2. *There is little pattern in the residuals.*

Problem: correlated errors

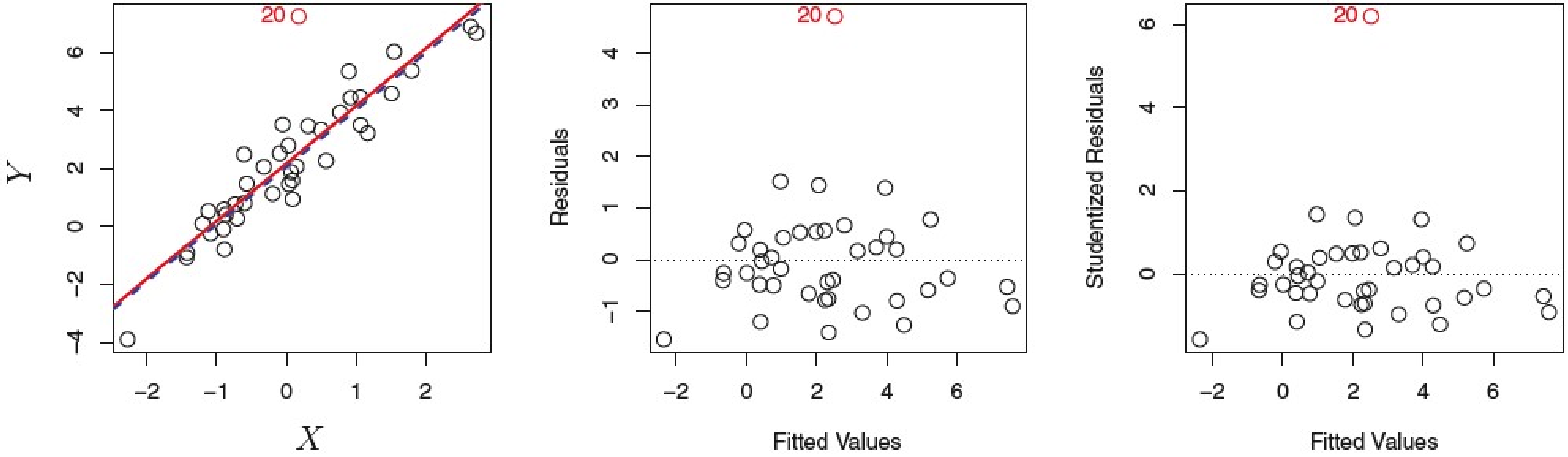
- Error terms are assumed to be independent.
- This assumption can be violated if,
  (1) Some samples are duplicated
  (2) Samples come from an autocorrelated time series.
- Irreducible variance, Var(ε) will be under or over estimated.

Problem: non-constant variance of error terms

- Variances change with Y. The variance may often be proportional to the magnitude of Y. To shrink the larger values of Y use a logY or $\sqrt{Y}$.



Problem: outliers



The outlier (unusually high response value) in the left most figure has little effect on the fitted regression. But the RSE is inflated to 1.09 from 0.77 (without the outlier). The studentized residuals on the right could be used to remove any that are greater than |3| (should only be 0.3% of the sample). Studentized = Residual divided by the standard deviation.

Removing points must be done with care and well documented.

Problem: leverage points

- High leverage applies to unusual values of the predictor variable.
- High leverage values will often affect the regression fit.
- Models with a single predictor, leverage for observation-*i* = $\frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{j=1}^{n} (x_j - \bar{x})^2}$, the individual leverage ranges from 1/n to 1, averaged over all observations equals (*p*+1)/*n*

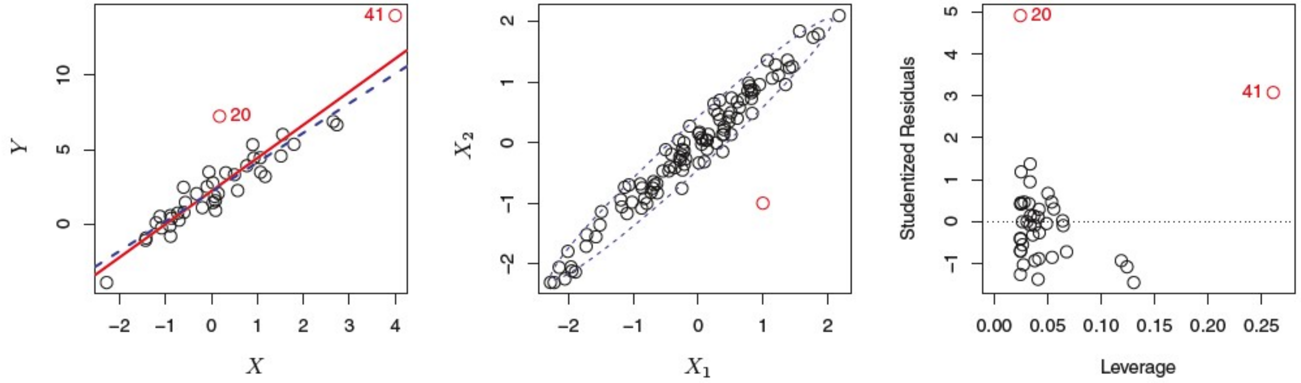
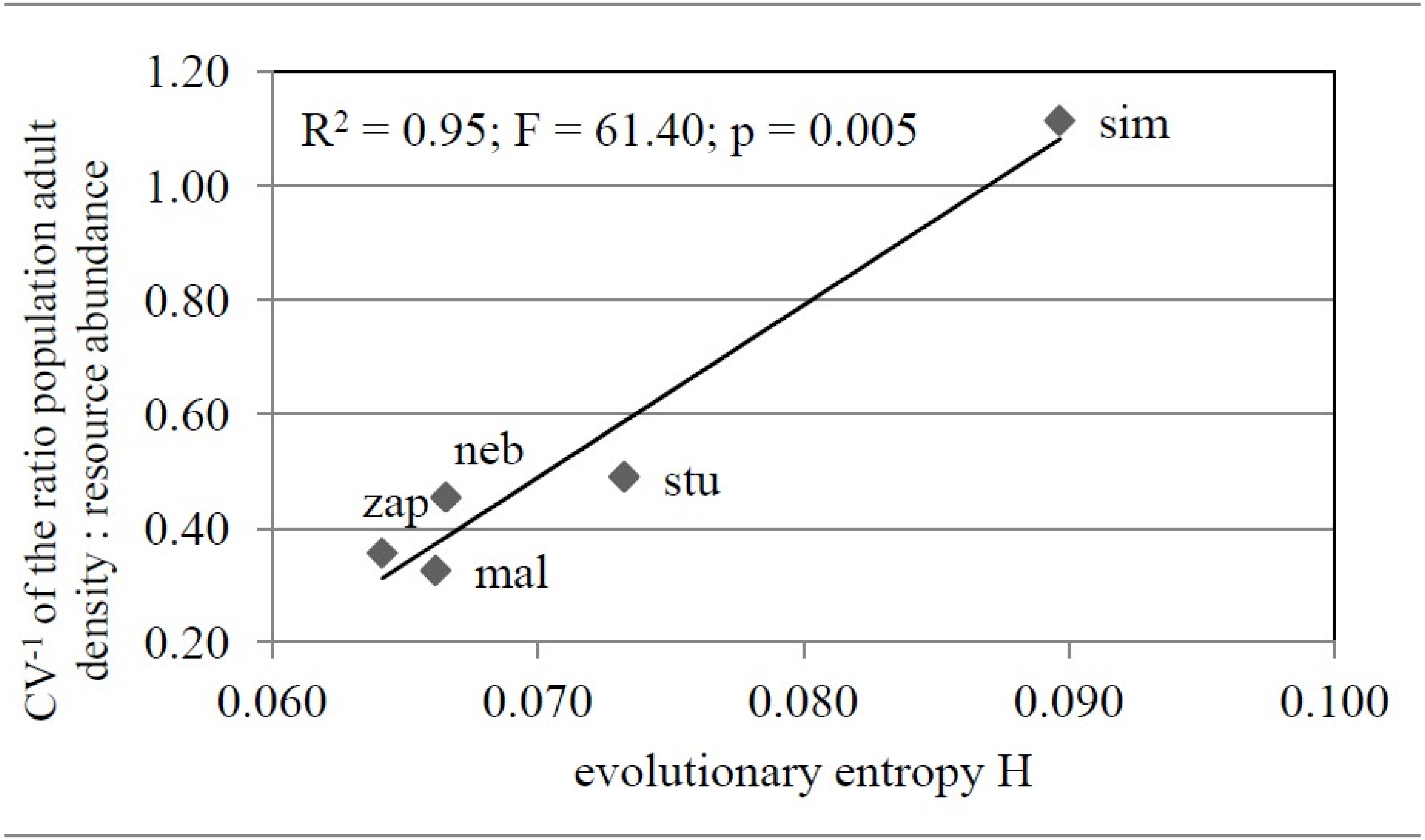


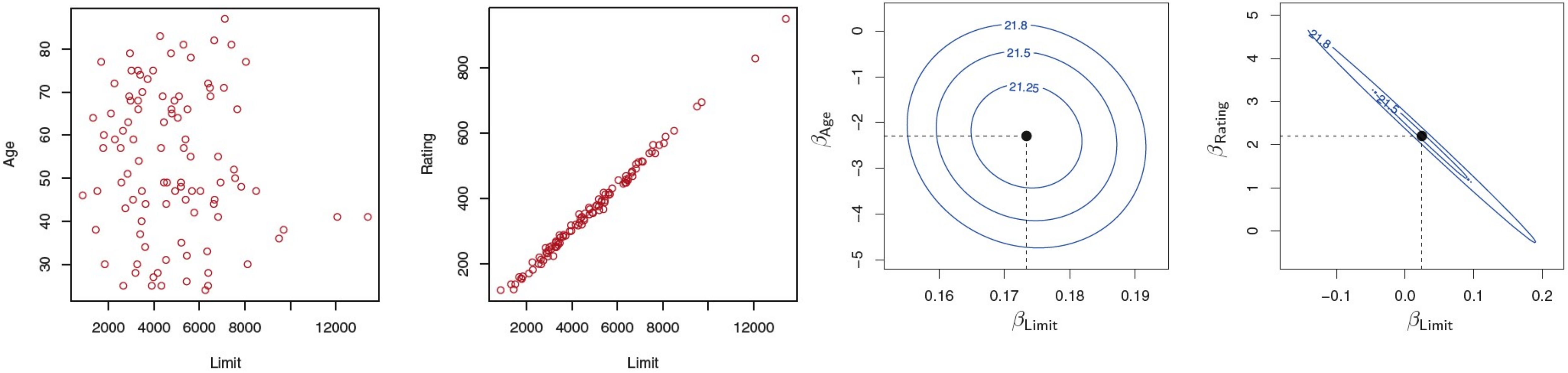
FIGURE 3.13. Left: *Observation 41 is a high leverage point, while 20 is not. The red line is the fit to all the data, and the blue line is the fit with observation 41 removed.* Center: *The red observation is not unusual in terms of its X_1 value or its X_2 value, but still falls outside the bulk of the data, and hence has high leverage.* Right: *Observation 41 has a high leverage and a high residual.*

Example: leverage points



Problem: collinearity

- When predictor variables are highly correlated.
- Credit card balance may depend on age, credit card limit and credit rating. But rating and credit card limit are tightly correlated.
- Thus, small changes in the data can dramatically change the least squares estimates, standard errors of the $\hat{\boldsymbol{\beta}}$ are larger.



Detecting and Fixing Multicollinearity

- Look at the correlation matrix. But there can be multicollinearity between 3 or more variables that won't show up in the correlation matrix.
- Compute the Variance Inflation Factor (VIF) for each predictor-*j*,

$$\frac{1}{1 - R^2_{X_j|X_{-j}}}$$

- Where $R^2_{X_j|X_{-j}}$is R^2 obtained from regressing X_j on all the remaining predictors.
- Eliminate essentially redundant variables or combine several into a single variable.