Chapter 3 Linear Regression

Question we should ask:

1. Is there a relationship between X and Y?
2. How strong is the relationship between ....?
3. Which things are associated with Y ?
   we must find a way to separate out the individual contribution of each medium to Y
4. How large is the association between each medium and Y?
5. How accurately can we predict future sales?
6. Is the relationship linear?
7. Is there synergy among the X?

Sample Linear regression

$Y \approx \beta 0 + \beta 1X$.

You might read “≈” as “is approximately modeled as”.
We will sometimes describe it by saying that we are regressing Y on X (or Y onto X).

- β0 and β1 are two unknown constants that represent
  the intercept and slope terms in the linear model.
- β0 and β1 are intercept known as the model coefficients or parameters.

3.1.1 Estimate the coefficient

In practice, β_0 and β_1 are unknown. So before we can use (3.1) to make predictions, we must use data to estimate the coefficients. Let

$$(x_1, y_1),\ (x_2, y_2), \ldots,\ (x_n, y_n)$$

represent n observation pairs, each of which consists of a measurement of X and a measurement of Y. In the `Advertising` example, this data set consists of the TV advertising budget and product sales in $n = 200$ different markets. (Recall that the data are displayed in Figure 2.1.) Our goal is to obtain coefficient estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ such that the linear model (3.1) fits the available data well—that is, so that $y_i \approx \hat{\beta}_0 + \hat{\beta}_1 x_i$ for $i = 1, \ldots, n$. In other words, we want to find an intercept $\hat{\beta}_0$ and a slope $\hat{\beta}_1$ such that the resulting line is as close as possible to the $n = 200$ data points. There are a number of ways of measuring *closeness*. However, by far the most common approach involves minimizing the *least squares* criterion, and we take that approach in this chapter. Alternative approaches will be considered in Chapter 6.

least squares

Let $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ be the prediction for Y based on the ith value of X. Then $e_i = y_i - \hat{y}_i$ represents the ith *residual*—this is the difference between the ith observed response value and the ith response value that is predicted by our linear model. We define the *residual sum of squares* (RSS) as

residual

residual sum of squares

$$\text{RSS} = e_1^2 + e_2^2 + \cdots + e_n^2,$$

or equivalently as

$$\text{RSS} = (y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1)^2 + (y_2 - \hat{\beta}_0 - \hat{\beta}_1 x_2)^2 + \cdots + (y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n)^2. \quad (3.3)$$

The least squares approach chooses $\hat{\beta}_0$ and $\hat{\beta}_1$ to minimize the RSS. Using some calculus, one can show that the minimizers are

$$\begin{aligned} \hat{\beta}_1 &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \\ \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x}, \end{aligned} \quad (3.4)$$

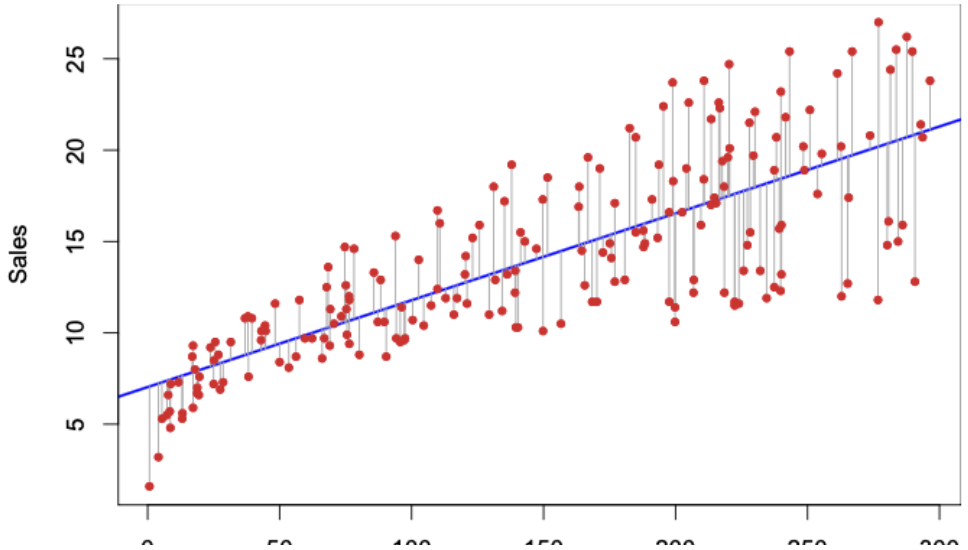

FIGURE 3.1. *For the* Advertising *data, the least squares fit for the regression of* sales *onto* TV *is shown. The fit is found by minimizing the residual sum of squares. Each grey line segment represents a residual. In this case a linear fit captures the essence of the relationship, although it overestimates the trend in the left of the plot.*

3.1.2 Assessing the accuracy of the coefficient estimate

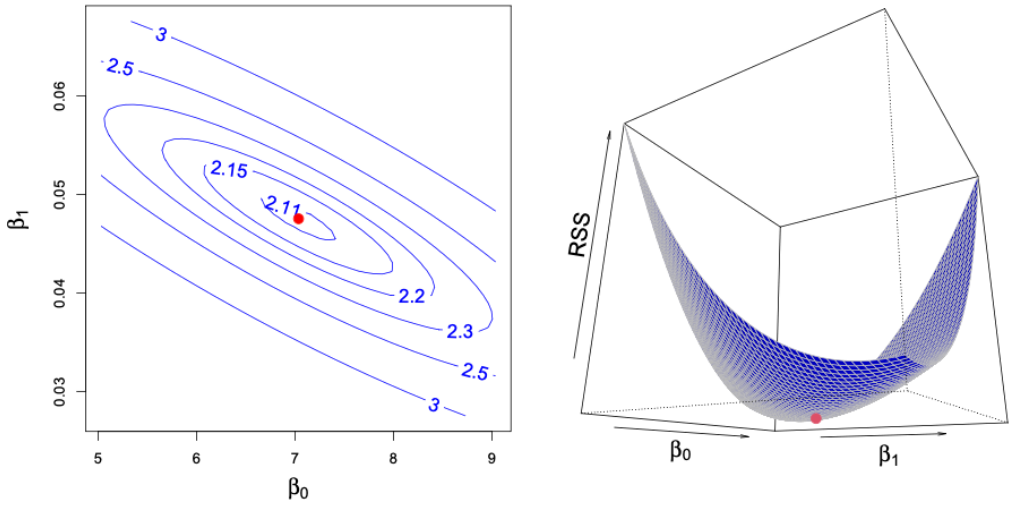

FIGURE 3.2. *Contour and three-dimensional plots of the RSS on the* Advertising *data, using* sales *as the response and* TV *as the predictor. The red dots correspond to the least squares estimates $\hat{\beta}_0$ and $\hat{\beta}_1$, given by (3.4).*

$$Y= \beta0 + \beta1X + \epsilon.$$

- The sample mean and the population means are different but in general the sample mean will provide a good estimate of the population means.
- Also, the unknown coefficient β0 and β1 in linear regression define the population regression line.
- population mean U (mui) = Y^ (simple mean)

@@@ When we fit the model
Y= β0 + β1X + ϵ.

we get estimate of β0 and β1 from our particular sample of data, but if we had collect a different sample, we would get slightly different β0 and β1.

→ the question is
How much do these estimate vary from sample to sample?
How precise/reliable are they?
→ This uncertainly is measured by the Standard Error (SE)

=== SE to accuracy of the population mean (u^)

+++Standard Error of the mean (SE)
If you estimate the mean μ of a population with sample mean $\hat{\mu} = (y_1 + y_2 + \ldots + y_n)/n$

Then the standard error of the mean is:
$SE(\hat{\mu}) = \sigma / \sqrt{n}$

where σ = population standard deviation → more data (larger n) → smaller uncertainty → **This tell us the average amount that this estimate μ^ differ from the actual value of μ**

+++Standard Error for regression coefficients $\hat{\beta_0}$ and $\hat{\beta_1}$
In simple linear regression we have **two** similar formulas (equation 3.8):

$$SE(\hat{\beta}_1)^2 = \sigma^2 / \Sigma(x_i - \bar{x})^2$$

$$SE(\hat{\beta}_0)^2 = \sigma^2 \left[1/n + \bar{x}^2 / \Sigma(x_i - \bar{x})^2 \right]$$

Where σ^2 = variance of the errors (how much the points scatter around the true line)

Very important observations:

Coefficient	What mainly affects its SE?	Intuitive meaning
$\hat{\beta_1}$ (slope)	$\Sigma(x_i - \bar{x})^2$ ← **spread of the x-values**	The more spread out your x-values are, the more accurately you can estimate the slope
		→ **leverage** (big range in X helps a lot)
$\hat{\beta_0}$ (intercept)	n + spread of x + how far $\bar{x}$ is from 0	Intercept is harder to estimate when x-values are far from zero
		If $\bar{x} = 0$ → $SE(\hat{\beta_0})$ behaves like SE(mean)

Key intuition (very important):

- If your x-values are all very similar (small spread) → slope is **hard** to estimate reliably
- If x-values are very spread out (large range) → slope is **much easier/more precise** to estimate → That's why experimenters/designers try to make x vary a lot when they can

---

+++We almost never know σ

In real life σ (true error standard deviation) is unknown.

We estimate it using the **Residual Standard Error** (RSE):

RSE = √(RSS / (n − 2))

where RSS = Residual Sum of Squares = $\Sigma(e_i^2)$ = sum of squared errors

→ RSE is basically the "typical size of the residuals" → We use RSE instead of σ in the SE formulas in practice

---

+++Why do we care about Standard Errors?

Main two uses (explained on this page and the next):

1. **Confidence intervals** Roughly: $\hat{\beta_1} \pm 2 \times SE(\hat{\beta_1})$ → 95% confidence interval (the ±2 is an approximation that works well for most sample sizes)
2. **Hypothesis testing** Test whether $\beta_1 = 0$ (i.e. is there really any relationship?) → This is the most common question in regression

```
t-test       t= β1/SE(β1)
```

find the p-value using t-distribution
p = 2P(T> |t|)
Df = n-2

```python
# Python example (using scipy)
from scipy import stats

t_stat = 17.67
df = 198
p_value = 2 * stats.t.cdf(-abs(t_stat), df)   # or 2 * (1 -
stats.t.cdf(abs(t_stat), df))
print(p_value)   # → extremely small number, like 1.4e-42
```

```r
# R example
t_stat ← 17.67
df ← 198
p_value ← 2 * pt(-abs(t_stat), df)
p_value   # → tiny number
```

Large p-value (>0.05) ⟶ cannot reject null hypothesis
Small p-value (<0.05) ⟶ reject the null hypothesis
Quick Summary Table

Concept	Meaning in plain English	Key formula / fact	Big practical message
Standard Error (SE)	How much the estimate would change if we had different data	$SE(\hat{\beta_1}) = \sigma / \sqrt{\Sigma(x_i - \bar{x})^2}$	Measures uncertainty of $\hat{\beta}$
Spread of X $(\Sigma(x_i - \bar{x})^2)$	How much leverage we have to estimate the slope	Bigger spread → much smaller $SE(\hat{\beta_1})$	Spread your predictor values as much as possible!
Residual Standard Error (RSE)	Estimated typical size of error $(\hat{\sigma})$	$\sqrt{(RSS/(n-2))}$	Our best guess of how noisy the data really is
Why $SE(\hat{\beta_1})$ gets smaller	More spread in X or more observations	—	Design studies / collect data to maximize information

---

The quality of a linear regression fit is assessed by:

1. Residual Standard Error
2. R statistic

+++Residual Standard Error (RSE)

What it measures

RSE is basically:

"On average, how far off (in the original units of Y) are our predictions from the actual values?"

Formula:

```
RSE = √( RSS / (n − 2) )
```

- RSS = Residual Sum of Squares = $\Sigma(y_i - \hat{y}_i)^2$ = total squared error
- n - 2 = degrees of freedom (because we estimated 2 parameters: intercept + slope)

Advertising example (TV vs Sales):

From Table 3.2:

- RSE = **3.26**
- Sales is measured in **thousands of units** → 3.26 means ≈ **3,260 units**

Interpretation:

Our model's predictions are wrong by about **3,260 units** on average (in either

direction).
Since the average sales ≈ 14,000 units, this is roughly a **23% error** on average.

When is RSE good or bad?

- Small RSE → model predicts quite accurately
- Large RSE → model has big prediction errors

Problem with RSE:
It's in the **same units as Y**, so it's hard to compare across different problems (sales in thousands vs house prices in millions, etc.).

+++R^2 (R-squared) — the most popular measure

R^2 = proportion of the total variation in Y that is explained by the model (using X).

Formula (two equivalent ways):

```
R² = (TSS − RSS) / TSS     =     1 − (RSS / TSS)
```

Where:

- **TSS** = Total Sum of Squares = $\Sigma(y_i - \bar{y})^2$
  → total variability in Y **before** doing any regression (how spread out Y is around its mean)
- **RSS** = Residual Sum of Squares = $\Sigma(y_i - \hat{y}_i)^2$
  → leftover (left unexplaned) variability **after** fitting the model

So:

- $R^2 \approx 1$ → model explains almost all the variation → excellent fit
- $R^2 \approx 0$ → model explains almost nothing → terrible fit (basically as good as just predicting the mean)

Advertising example (TV vs Sales):
From Table 3.2:

- **$R^2 = 0.612$** (or 61.2%)

Interpretation (Should memorize):
61.2% of the variability in sales can be explained by a linear relationship **with TV advertising budget.** The remaining 38.8% is due to other factors (or just random noise).

The correlation

The R^2 statistic is a measure of the linear relationship between X and Y. Recall that *correlation*, defined as

correlation

$$\text{Cor}(X, Y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}, \qquad (3.18)$$

R^2 = r^2, while r = Cor(X, Y)

+++Comparison Table – RSE vs R^2

Measure	What it tells you	Units	Range	Easy to interpret?	Example from TV model
RSE	Average prediction error	Same as Y	0 to ∞	Depends on context	~3,260 units (~23% error)
R^2	% of variance in Y explained by X	Proportion	0 to 1	Very easy	61.2% of sales variation explained

Additional Tables from the Book (Radio & Newspaper)

Predictor	Slope (β_1)	Interpretation	R^2 (implied)	Strength of relationship
TV	0.0475	+$1k TV → +47.5 units sales	~0.612	Strong
Radio	0.203	+$1k radio → +203 units sales	higher	Very strong
Newspaper	0.055	+$1k newspaper → +55 units sales	much lower	Weak

→ Radio has the **strongest individual relationship** with sales (biggest slope and likely highest R^2).
→ Newspaper has the **weakest** (small slope, small t-statistic 3.30, p=0.00115 still significant but much less impressive).

Multiple Linear Regression (MLR)

- **Model Definition**: MLR predicts a response variable Y using p predictors (X1 to Xp). The population model is:

```
Y = β0 + β1 X1 + β2 X2 + ... + βp Xp + ε
```

- β0 is the intercept (expected Y when all X's are 0).
- βj is the slope for predictor Xj: average change in Y for a 1-unit increase in Xj, holding other predictors constant (key for isolating effects).
- ε is the error term (random variation).
- Example: Sales = β0 + β1 (TV) + β2 (Radio) + β3 (Newspaper) + ε. This shows synergy between media types.

- **Why Use MLR?** → Simple linear regression (one predictor) is limited; MLR handles real-world scenarios with multiple factors. In advertising, running separate simple regressions ignores interactions (e.g., TV and radio might amplify each other).
- **Assumptions**: Errors are normally distributed (for inference); predictors are not perfectly correlated (to avoid multicollinearity issues).

+++**Estimating Coefficients (Least Squares Method)**

- **Goal**: Find estimates b0, b1, ..., bp that minimize the Residual Sum of Squares (RSS):

```
RSS = Σ (yi − ŷi)^2
```

  - yi is observed Y, ŷi = b0 + b1 xi1 + ... + bp xip is predicted Y.
  - This fits a "plane" (in 3D for p=2) or hyperplane (higher dimensions) to minimize vertical distances from data points.
- **Process**: Use matrix algebra or software to solve. No closed-form for p > 1 like in simple regression, but computers handle it.
- **Example from Data**:
  - For advertising: Sales = 2.939 + 0.046 (TV) + 0.189 (Radio) - 0.001 (Newspaper).
  - Interpretation: A $1,000 increase in TV budget boosts sales by ~46 units (holding others fixed). Radio is stronger (189 units), newspaper negligible (-1 unit, possibly negative due to correlation with others).
- **Key Table Insight**: Table 3.4 shows coefficients: TV (0.046, t=17.67, p<0.001), Radio (0.189, t=9.20, p<0.001), Newspaper (-0.001, t=-0.18, p=0.860). Newspaper is insignificant.

+++**Model Fit and Evaluation Measures**

- **Residual Standard Error (RSE)**: Measures average deviation of predictions from actuals.

```
RSE = √[RSS / (n − p − 1)]
```

  - Lower is better. In example, RSE=1.69 (sales in thousands), meaning predictions are off by ~1,690 units on average.

- **R^2 (Coefficient of Determination)**: Proportion of variance in Y explained by the model.

```
R² = 1 - (RSS / TSS), where TSS = Σ (yi - ȳ)^2 (total sum of squares)
```

  - Ranges 0-1; higher means better fit. In example, full model R^2=0.897 (89.7% variance explained), vs. TV-only R^2=0.612.
  - Adjusted R^2: Penalizes adding useless predictors: Adjusted R^2 = 1 - [RSS/(n-p-1)] / [TSS/(n-1)]. Use this for model comparison; it can decrease if predictors add noise.
- **F-Statistic**: Tests overall model significance (H0: all βj=0 except β0).

```
F = [ (TSS - RSS)/p ] / [ RSS/(n-p-1) ]
```

  - High F rejects H0. In example, F=570 (p<0.001), model is significant. Partial F tests subsets (e.g., adding radio to TV model: F=79.4, p<0.001).
- **Correlation Matrix (Table 3.5)**: Shows relationships between variables.
  - TV & Sales: 0.782 (strong positive).
  - Radio & Sales: 0.576.
  - Newspaper & Sales: 0.228 (weak).
  - TV & Radio: 0.054 (low correlation, good—no multicollinearity).
  - High correlations (e.g., >0.8) can inflate variances and make coefficients unstable.

+++**Hypothesis Testing and Inference**

- **t-Test for Individual Coefficients**: Tests H0: βj = 0 (predictor j useless).

```
t = bj / SE(bj), p-value from t-distribution.
```

  - Low p (<0.05) rejects H0. In example, TV and radio p<0.001 (useful), newspaper p=0.86 (not useful—drop it?).
- **F-Test**: For overall or subsets (e.g., test if β2=β3=0).
- **Confidence Intervals (CI)**: For coefficients, e.g., 95% CI for β1: b1 ± t*SE(b1).
  - Narrow CI means precise estimate. In example, for radio: [0.162, 0.216].
- **Prediction Intervals**: Wider than CI; account for both coefficient uncertainty and error ε.
  - Example: For $100k TV + $20k radio, predicted sales=20,000 units; 95% CI [19,985, 20,015], prediction interval [16,735, 23,265] (uncertainty in new observation).
- **Key Question**: Even with high R^2, check if p-values are low—high R^2 alone doesn't mean causation or usefulness.

+++**Variable Selection Methods**

- **Why Select?** Too many predictors overfit (high R^2 but poor predictions); aim for parsimony.
- **Forward Selection**: Start with null model (just intercept), add one variable at a time with lowest p-value or highest F-improvement. Stop when no additions help (e.g., $p>0.05$).
  - Example: Start null, add TV (best), then radio (improves R^2 from 0.612 to 0.897), newspaper doesn't help ($p=0.86$).
- **Backward Selection**: Start with full model, remove least significant (highest p) one by one.
  - Example: Full model, remove newspaper ($p=0.86$), left with TV+radio (R^2 still 0.897, adjusted R^2 increases slightly).
- **Mixed Selection**: Combine forward/backward—add forward, then check/remove if needed.
- **Other Criteria**: Use AIC, BIC, or adjusted R^2 to compare models (lower AIC/BIC better).
- **Insight**: In advertising data, best model is TV + radio ($R^2=0.897$, $F=859$, newspaper adds noise due to correlation with radio: 0.354).

+++**Important Questions to Ask in MLR**

- Is there a relationship between predictors and response? (F-test > critical value, $p<0.05$).
- Do all predictors help explain Y, or only a subset? (t-tests, variable selection).
- How well does the model fit the data? (R^2, adjusted R^2, RSE).
- Given predictors, what response should we predict, and how accurate? (Use $\hat{y}$ with prediction intervals).
- **Non-Linearity Note**: Data may show interactions (e.g., TV*radio term) or non-linear patterns (Figures 3.4-3.5 show slight curvature—consider adding quadratic terms).

+++**Practical Insights and Warnings**

- **Synergy/Interaction**: In example, radio amplifies TV—combined effect > sum of individuals. Full model captures this; separate simple regressions miss it (e.g., radio alone: 203 units/$1k, but with TV: 189).
- **Overfitting**: Adding irrelevant predictors (like newspaper) decreases adjusted R^2, increases RSE.
- **Causation vs. Correlation**: High R^2 doesn't prove cause (e.g., newspaper correlates with sales but via radio).
- **Data Example Summary**: 200 markets, sales avg. ~14k units. Model predicts well but prediction intervals wide due to variability.
- **Limitations**: Assumes linearity; check residuals for normality, constant variance. If violated, transform variables or use other models.

3.3.1 Qualitative Predictors (Categorical / Factor Variables)

Key ideas to remember:

- Quantitative predictors → numbers with meaningful magnitude (age, income, horsepower)
- Qualitative predictors → categories (own house? student? region?)

Two-level qualitative variable → use dummy variable (0/1 coding)

Example: `own` (house ownership)

- x_i = 1 if owns house, 0 otherwise
- Model: balance = $\beta_0 + \beta_1$ own + ε
- Interpretation options (both correct, just different viewpoint):

  - β_0 = average balance for non-owners
  - $\beta_0 + \beta_1$ = average balance for owners
    → β_1 = average difference (owners − non-owners)

  OR

  - β_0 = overall average (if roughly balanced groups)
  - β_1 = how much extra/less owners have compared to non-owners

More than 2 levels → k−1 dummies (k = number of categories)

Example: region (East, West, South)

- Choose baseline (usually arbitrary — often the one with most observations or reference group)
- Example coding (East = baseline):

  x_South = 1 if South, 0 else
  x_West = 1 if West, 0 else

  Model: balance = $\beta_0 + \beta_1$ South + β_2 West + ε

  Interpretation:

  - β_0 = average balance in East
  - $\beta_0 + \beta_1$ = average balance in South
  - $\beta_0 + \beta_2$ = average balance in West
  - β_1 = South vs East difference
  - β_2 = West vs East difference

Important principle: Choice of baseline changes intercept & dummy coefficients, **but does NOT change model fit, predictions, or p-values of overall group effect** (use F-test for that).

3.3.2 Extensions of the Linear Model

A. Interaction / Synergy Effect (most important concept in this section!)

Additive model assumption → effect of X_1 is constant regardless of X_2

Interaction relaxes this → effect of X_1 depends on value of X_2

Mathematical form:

$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 (X_1 \times X_2) + \varepsilon$

Two classic examples from the text:

1. Advertising (TV × Radio)
   - Main effects only: $R^2 \approx 89.7\%$
   - With interaction: $R^2 \approx 96.8\%$
   - β_3 (TV×radio) very significant → strong synergy
     → Spending on both TV and radio together is more effective than the sum of separate effects
2. Credit data — Income × Student

   Without interaction → two **parallel** lines (same slope)

   With interaction → two lines with **different slopes**

   Important observation from Figure 3.7:

   - Students have **lower slope** (income increases balance less strongly for students)
   - Non-students have steeper slope

   → Including interaction is often necessary even if main effects look insignificant

Hierarchical principle / guideline taught in the book:

> If you include an **interaction** X_1:X_2, **you should also include the main effects** X_1 and X_2 — even if their p-values are large / not significant.

Reason: $X_1 \times X_2$ is correlated with X_1 and X_2 → omitting main effects changes the meaning of the interaction.

B. Non-linear relationships → Polynomial regression

Still uses linear regression software!

Most common: quadratic

$\text{mpg} = \beta_0 + \beta_1 \text{ horsepower} + \beta_2 \text{ horsepower}^2 + \varepsilon$

- This is still a **linear model** in the parameters $\beta_0, \beta_1, \beta_2$
- But now the relationship between mpg and horsepower is curved

From Figure 3.8 (Auto data):

- Linear fit → clearly bad (orange line)
- Quadratic fit → much better (blue curve)
- Higher polynomials (degree 5 green line) → overfitting, wiggly, unnecessary

Rule of thumb from text:

Start with quadratic. If residual plot still shows pattern → consider cubic or other transformations (covered more in Ch. 7).

Quick Visual Summary Table (most exam-tested concepts)

Concept	When to use	Mathematical form	Key Figure (text)	Important note / trap
Dummy (2 levels)	Binary categorical	$Y = \beta_0 + \beta_1 D$	Fig 3.6	Coding 0/1 or 1/-1 → only interpretation changes
Multiple categories	>2 levels	k−1 dummies	Table 3.8	Baseline arbitrary, use F-test for overall effect
Interaction	Effect of one var depends on another	$+ \beta_3 X_1X_2$	Fig 3.7	Include main effects even if p large!
Polynomial (quadratic)	Clear curvature in scatterplot/residuals	$+ \beta_2 X^2$	Fig 3.8	Still linear model — just transformed X

- **Always look at Figure first** — if lines are parallel → no interaction needed
  if slopes clearly different → interaction almost certainly needed
  if scatter looks curved → think polynomial / transformation
- **Residual plots are your best friend** for detecting non-linearity
- Exam favorite questions:
  - Interpret coefficient with interaction present
  - Explain why we include main effects when interaction is in model
  - Compare R^2 with vs without interaction
  - Decide whether to add X^2 term

3.3.3 Potential Problems

Key problems that can go wrong when fitting linear regression:

1. **Non-linearity** of response–predictor relationship
2. **Correlation** of error terms (esp. time series)
3. **Non-constant variance** of errors = heteroscedasticity
4. **Outliers**
5. **High-leverage** points
6. **Collinearity** (multicollinearity when p > 1)

Non-linearity of the Response–Predictor Relationships

Core idea

Linear regression assumes a straight-line (or flat hyperplane in multiple regression) relationship between predictors and the response.
If the true relationship is curved (e.g., quadratic, exponential, logarithmic), the model will systematically miss patterns → residuals will show structure instead of random scatter → unreliable inference & poor predictions.

Diagnostic tool

Plot **residuals $e_i = y_i - \hat{y}_i$** vs fitted values $\hat{y}_i$ (or sometimes vs individual predictors X_j).
Look for a smooth trend line (often loess/red line in the book) — any clear pattern (U-shape, curve, etc.) signals non-linearity.

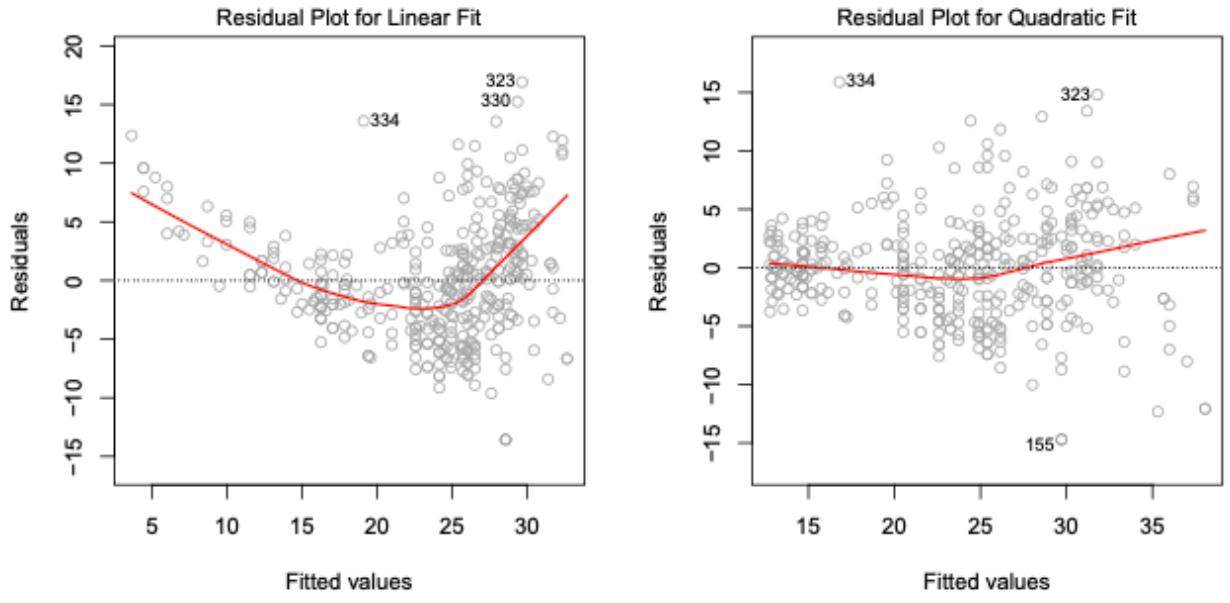

Figure 3.9 (from the book — two side-by-side panels, Auto dataset: mpg ~ horsepower)

- **Left panel** — Residual Plot for Linear Fit
  - X-axis: Fitted values ($\hat{y}$) — range roughly 5 to 30+ mpg
  - Y-axis: Residuals — range about −15 to +15

    - Points: large cloud of gray circles (many observations)
    - Red smooth line: clear **U-shaped** (or inverted-U) pattern
      → residuals start positive at low fitted values → dip negative in the middle (around 20–25 mpg) → rise positive again at high fitted values
    - Numbers like 323, 330, 334 are labeled outliers/points of interest
      → Strong visual evidence that a pure linear model is inadequate — systematic curvature remains in residuals.
- **Right panel** — Residual Plot for Quadratic Fit (mpg ~ horsepower + horsepower²)
    - X-axis: Fitted values (now from the quadratic model) — wider range ~15 to 35 mpg
    - Y-axis: Residuals — much tighter spread, roughly −10 to +10
    - Red smooth line: almost perfectly flat around zero (very slight gentle curve at most)
    - Points scattered randomly with no obvious pattern
      → Adding the quadratic term has largely removed the non-linearity — residuals now look much closer to what we expect under a good linear(ized) model.

Key takeaway from Figure 3.9

- A curved smooth line in residuals vs fitted = classic sign of non-linearity
- Simple fix: include polynomial terms (X^2, X^3), splines, log(X), √X, etc.
- If non-linearity remains even after transformations → may need more flexible methods (trees, GAMs, etc. — later chapters)

Correlation of Error Terms

Core idea

Linear regression assumes errors $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ are **independent** (uncorrelated).
When errors are correlated (most common in **time series** or spatial data), standard errors are underestimated → confidence intervals too narrow, p-values too small → we overstate certainty in the model.

Why it happens

- Time series: adjacent measurements often similar (e.g., stock prices, temperature over days)
- Spatial data, repeated measures on same subject, etc.
- If εt *is positively correlated with* $\varepsilon\{t+1\}$, we see **tracking** or clustering in residuals over time.

Diagnostic tool

Plot residuals vs time/order of observation (especially for time-series or ordered data).

Look for patterns like runs of positive/negative values, oscillation, or obvious dependence between adjacent points.

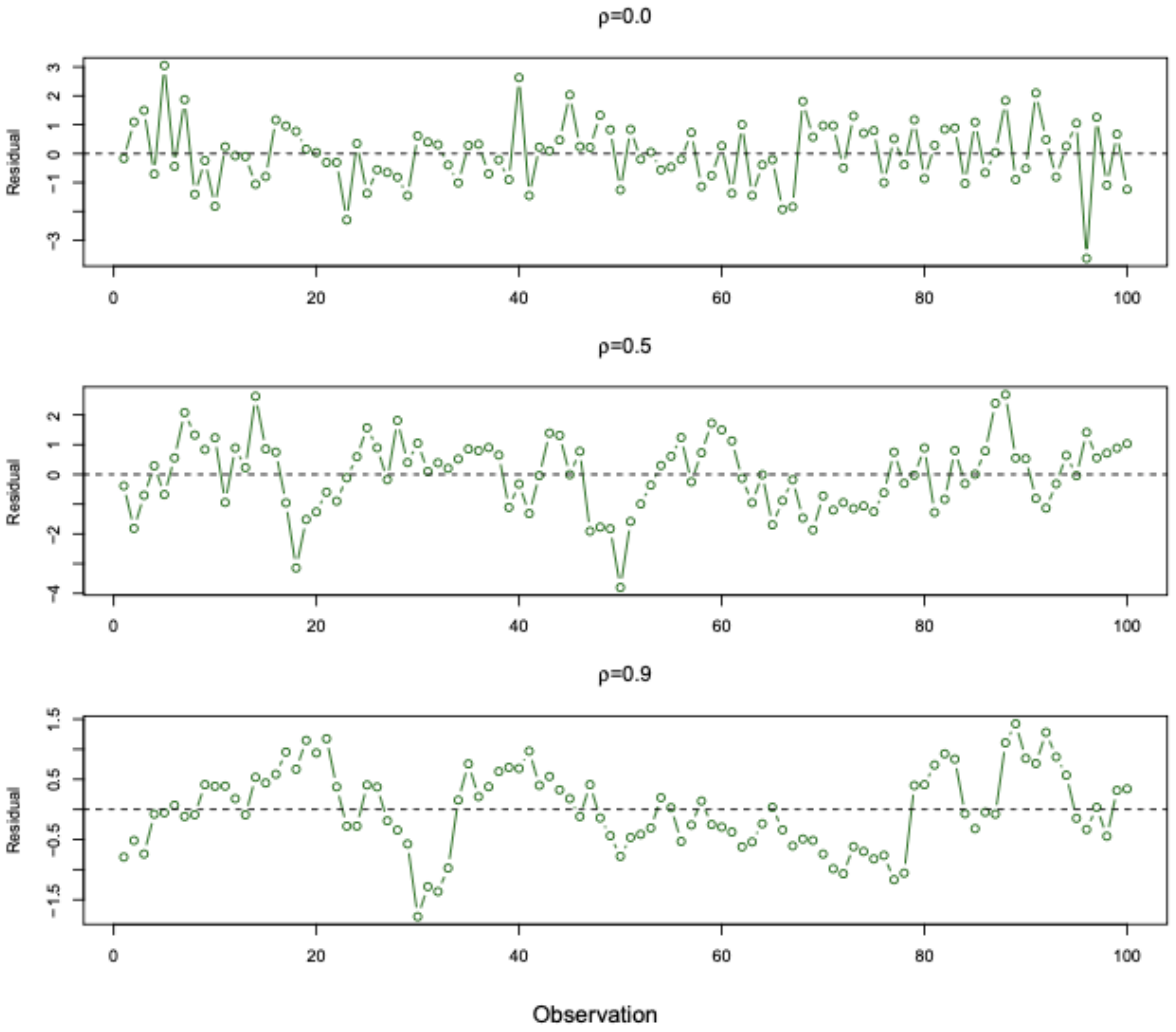

Figure 3.10 (from the book — four panels of simulated time-series residual plots)

- All panels: X-axis = observation index (time/point 0 to 100)
  - Y-axis = Residual value (roughly −3 to +3 range across plots)
  - Black dashed line at residual = 0
- **Top panel** (ρ = 0.0)
  → Residuals look completely random — no pattern, no clustering
  → Adjacent residuals show no relationship (ideal case)
- **Second panel** (ρ = 0.5)
  → Mild positive correlation visible
  → Runs of positive residuals followed by runs of negative ones
  → Some "tracking" — values tend to stay on the same side of zero for a few consecutive points
- **Third panel** (ρ = 0.9)
  → Very strong positive correlation
  → Long stretches where residuals stay positive or negative

→ Clear **tracking** pattern — adjacent residuals are very similar
→ Looks almost like a slowly wandering line rather than independent noise

- **Bottom panel** (another ρ = 0.9 case or variant)
  → Similar strong tracking — emphasizes how high correlation makes residuals behave like a smooth process rather than white noise

Key takeaway from Figure 3.10

- As correlation ρ increases → residuals show longer runs & less randomness
- Consequence: ordinary least squares standard errors are too optimistic
- Fixes: use time-series models (AR, ARIMA), generalized least squares, Newey-West standard errors, etc. (covered more in later time-series chapters)

Non-constant Variance of Error Terms = Heteroscedasticity

Core idea

Linear regression assumes $\text{Var}(\varepsilon_i) = \sigma^2$ (constant for all i).
When variance changes with X or with the level of Y (heteroscedasticity), standard errors are biased → inference unreliable.

Common patterns

- Variance increases with fitted value (funnel shape — most common)
- Variance decreases with fitted value (reverse funnel)
- Variance depends on a predictor (e.g., larger spread for higher income)

Diagnostic tool

Residuals vs fitted values plot — look for changing spread (fan/funnel) rather than constant band.

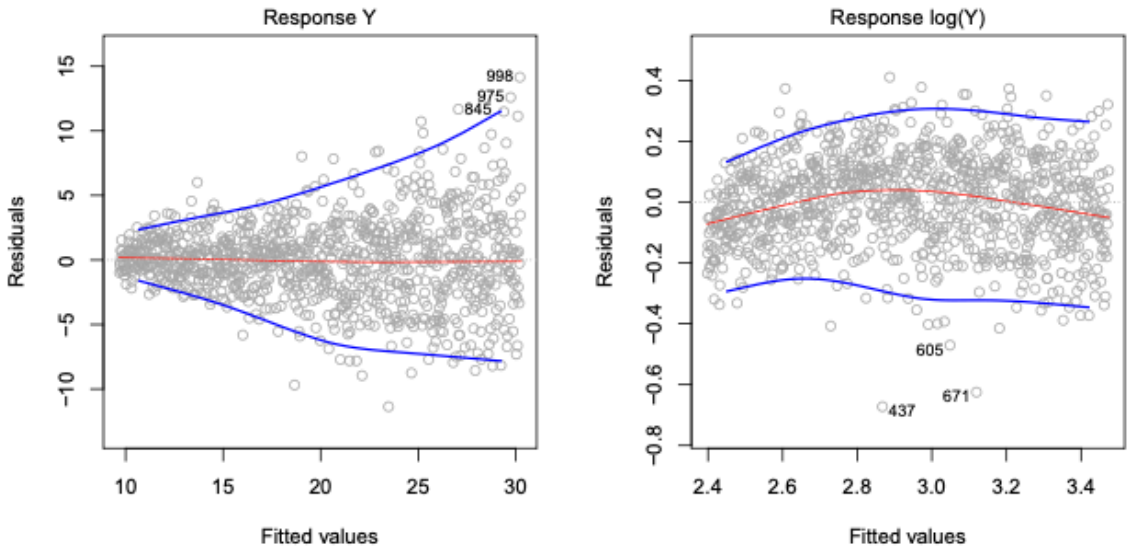

Figure 3.11 (from the book — two main panels + small inset examples)

- **Left panel** — Residuals vs Fitted for original response Y
  - X-axis: Fitted values (roughly 10 to 30)

- Y-axis: Residuals (roughly −10 to +15)
- Clear **funnel shape**:
  → Small spread (tight cluster) at low fitted values
  → Much wider spread (large residuals) at high fitted values
- Red smooth line tracks outer quantiles → emphasizes increasing spread
  → Strong heteroscedasticity — variance grows with mean response

- **Right panel** — Residuals vs Fitted after log(Y) transformation
  - X-axis: Fitted values on log scale (roughly 2.4 to 3.4)
  - Y-axis: Residuals (now roughly −0.8 to +0.4)
  - Spread is much more constant across the range of fitted values
  - Red smooth line almost horizontal → variance stabilized
    → Log transformation is a very effective remedy here
- **Small inset plots** (often shown nearby):
  → Examples of other transformations (e.g., $\sqrt{Y}$, Y^2) applied in different situations
  → Illustrate that the right transformation depends on how variance behaves

Key takeaway from Figure 3.11

- Funnel = heteroscedasticity → violates constant variance assumption
- Common fix: transform the response (log Y most frequent when variance increases with mean)
- Alternative: weighted least squares (if you know the form of variance) or robust standard errors

Outliers

- **Definition**: point where y_i is far from what the model predicts (large residual)
- Very large residual → strong influence on least-squares fit (especially intercept & slope when n is small)

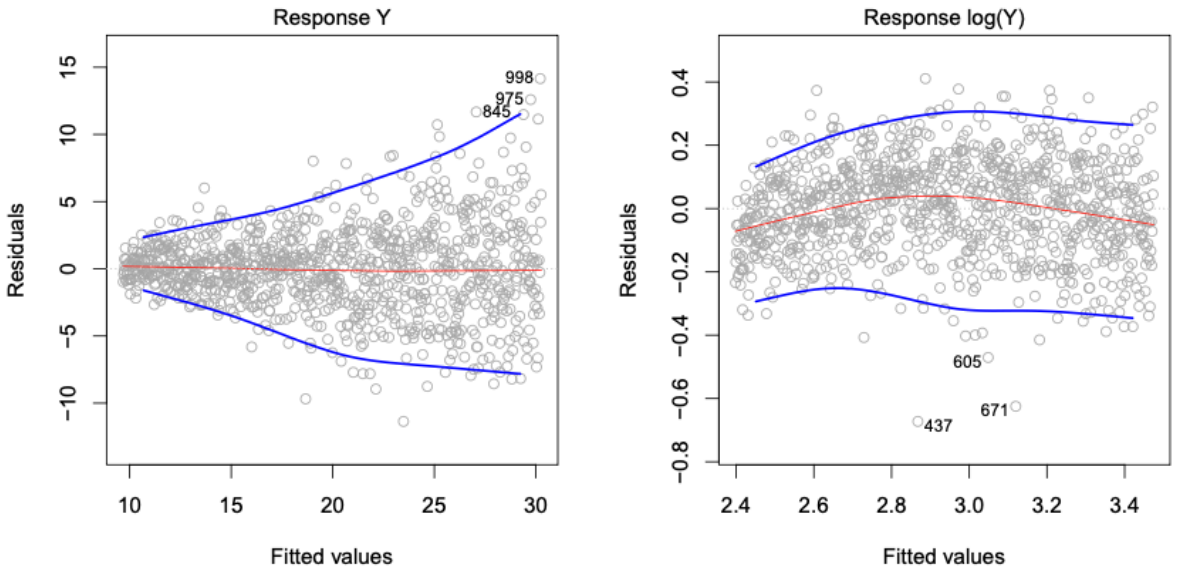

Figure 3.11 (residual plots – funnel shape example)

- Left: raw Y → clear heteroscedasticity + possible mild non-linearity
- Right: log(Y) → residuals much more stable (constant variance), still some curvature

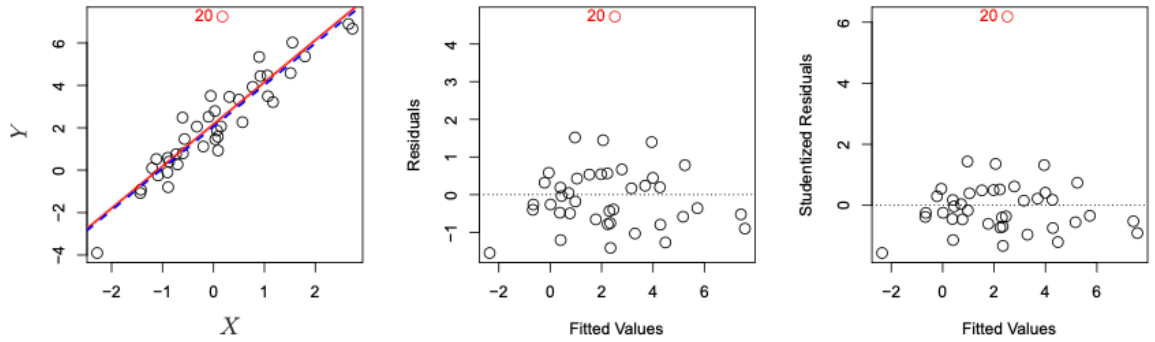

Figure 3.12 (outlier example – observation 20)

- Red line = fit with outlier included
- Blue dashed = fit after removing point 20
- Center panel: residual plot clearly flags the point (large positive residual)
- Right panel: studentized residual ≈ 6 (rule of thumb: |studentized residual| > 3 is suspicious)

Takeaway on outliers

- Do **not** automatically delete — first investigate why it is unusual (data error? interesting subgroup?)
- Studentized residuals are better for detection than raw residuals
- If outlier is legitimate → consider robust regression or weighted least squares later

High Leverage Points

- **Definition**: point with unusual x_i value(s) — far from the center of the predictor cloud
- High **leverage** = potentially large influence on the fitted coefficients (even if residual is small)

Leverage statistic (simple linear regression)
$h_i = 1/n + (x_i - \bar{x})^2 / \Sigma(x_{iv} - \bar{x})^2$ (3.37)

Average leverage = (p+1)/n
Anything much larger than 2× or 3× average is high leverage.

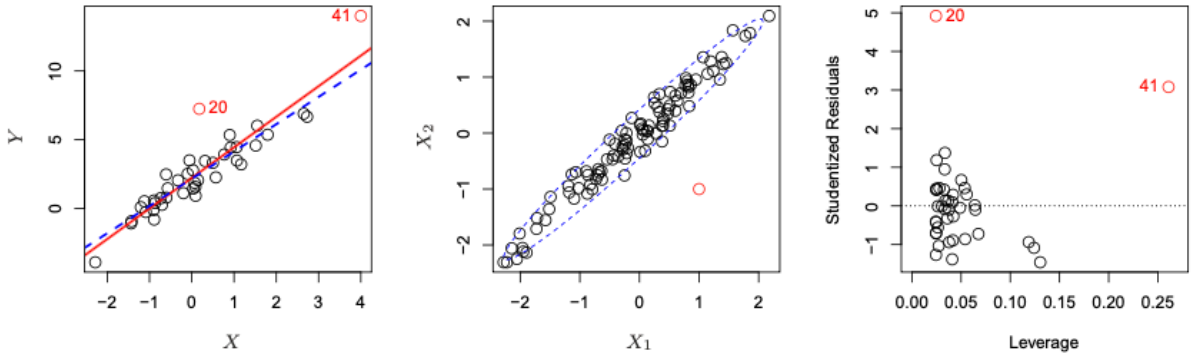

Figure 3.13 (high leverage example – obs 41)

- Left: red line = full fit, blue = fit without obs 41 → huge change in slope!
- Center: obs 41 not extreme in x_1 or x_2 individually, but extreme in the joint (X_1,X_2) space
- Right: same point has very high leverage + large studentized residual → **dangerous combination** (high leverage + outlier)

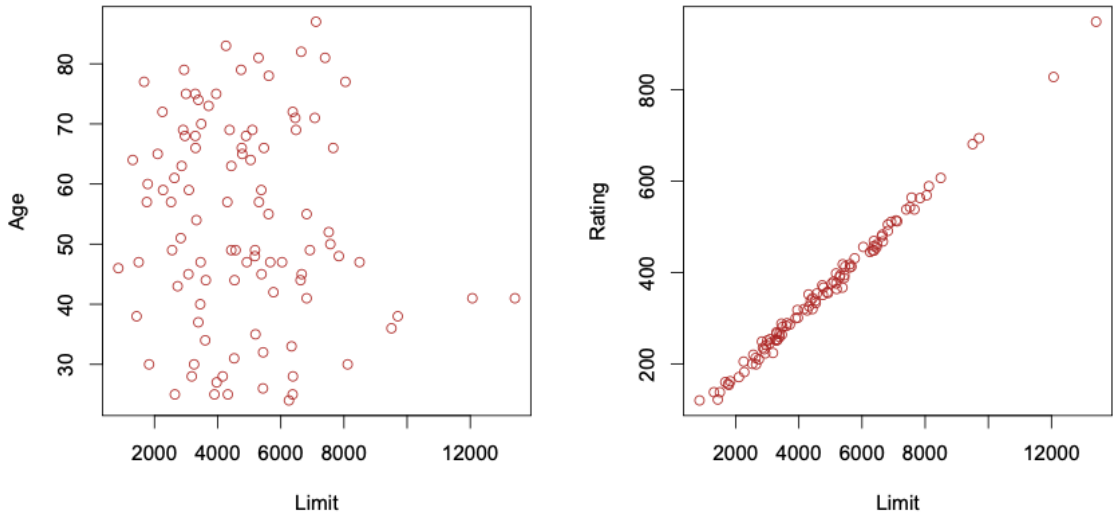

Figure 3.14 (Credit data – collinearity illustration)

- Left: age vs limit → almost no relationship
- Right: rating vs limit → very strong linear relationship → collinearity!

Key lesson

High leverage without large residual → can still strongly pull the line
High leverage + large residual → can completely dominate / destroy the fit

Collinearity

Definition: two or more predictors are highly correlated → redundant information

Consequences

- Coefficients become very unstable (large standard errors)
- Individual t-tests lose power (can't tell which variable is driving the effect)

- Overall model can still fit well (high R^2), but individual $\hat{\beta}_j$ are unreliable

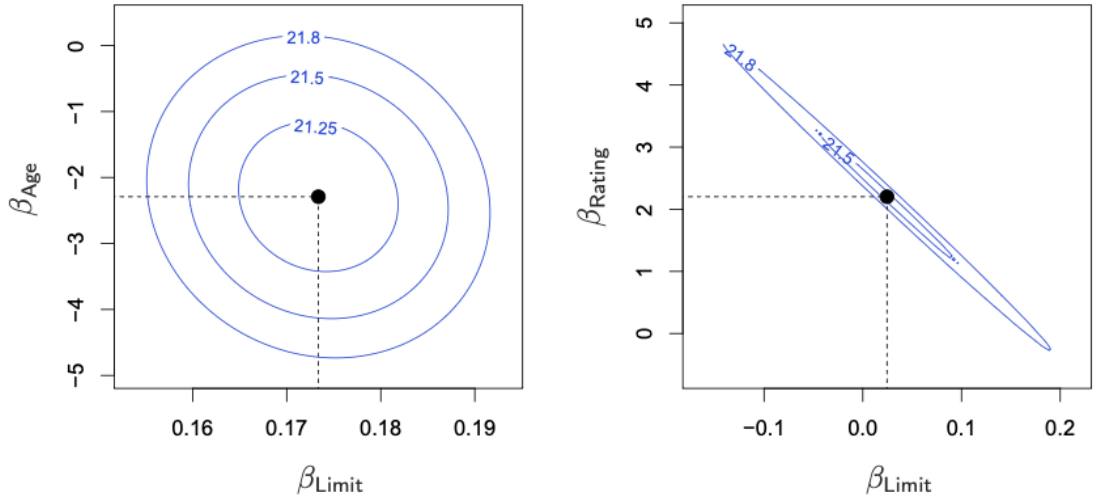

Figure 3.15 (contour plots of RSS – Credit data)

- Left: age & limit → almost circular contours → little collinearity
- Right: rating & limit → very elongated ellipse → strong collinearity
  → many ($\hat{\beta}$_limit, $\hat{\beta}$_rating) pairs give almost the same RSS

Detection

- Correlation matrix (simple, but misses multicollinearity among >2 variables)
- **Variance Inflation Factor** (VIF) — best practical tool

$VIF(\hat{\beta}_j) = 1 / (1 - R^2_{X_j \mid X_{-j}})$
Rule of thumb: VIF > 5–10 signals problematic collinearity

		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

Table 3.11 example (Credit – balance ~ age + limit vs rating + limit)

- Model 1: limit very significant, small std error
- Model 2: limit becomes non-significant, std error ×12 larger → classic collinearity symptom

Remedies

1. Drop one of the collinear predictors (if they carry similar information)
2. Combine them (e.g. average standardized limit + rating → “credit worthiness” index)
3. Ridge regression / lasso (later chapters)

3.3 Summary – Diagnostic Plot Priority

1. Residuals vs fitted values → non-linearity, heteroscedasticity, outliers
2. Studentized residuals → formal outlier detection
3. Leverage (h_i) or (studentized residual × $\sqrt{h_i}$) → high leverage + outlier combo
4. VIF → collinearity

3.5 Comparison: Linear Regression vs K-Nearest Neighbors

Parametric (linear regression) vs **non-parametric** (KNN regression)

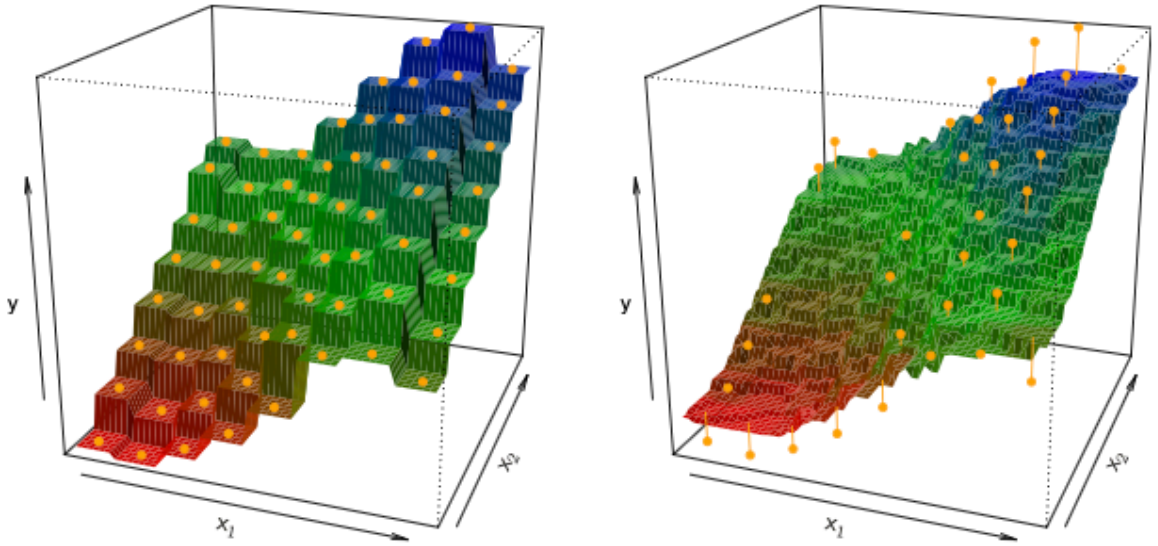

Figure 3.16 (2D toy example)

- K=1 → very wiggly / step-like fit
- K=9 → much smoother surface

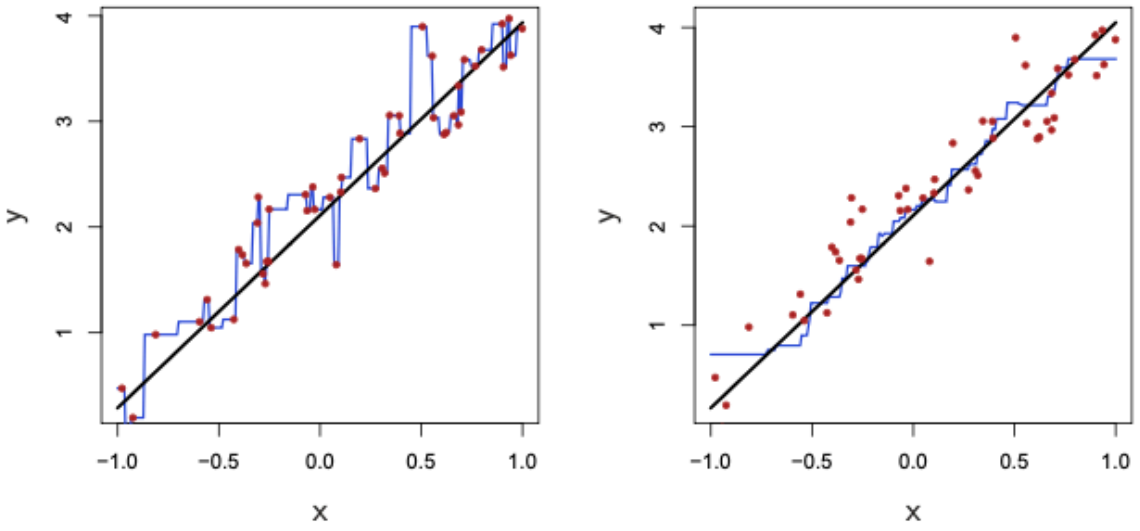

Figure 3.17 (1D example – true f(x) quite smooth & non-linear)

- K=1 → interpolates training points → high variance

- K=9 → much closer to truth

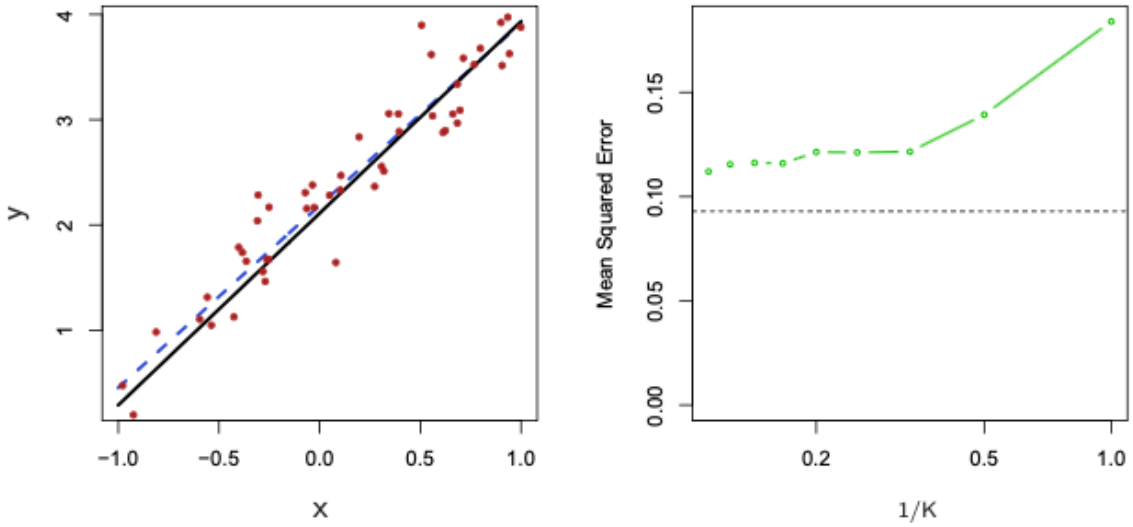

Figure 3.18 (linear case)

- Linear regression almost perfect
- KNN (even K=9) worse — pays unnecessary variance cost

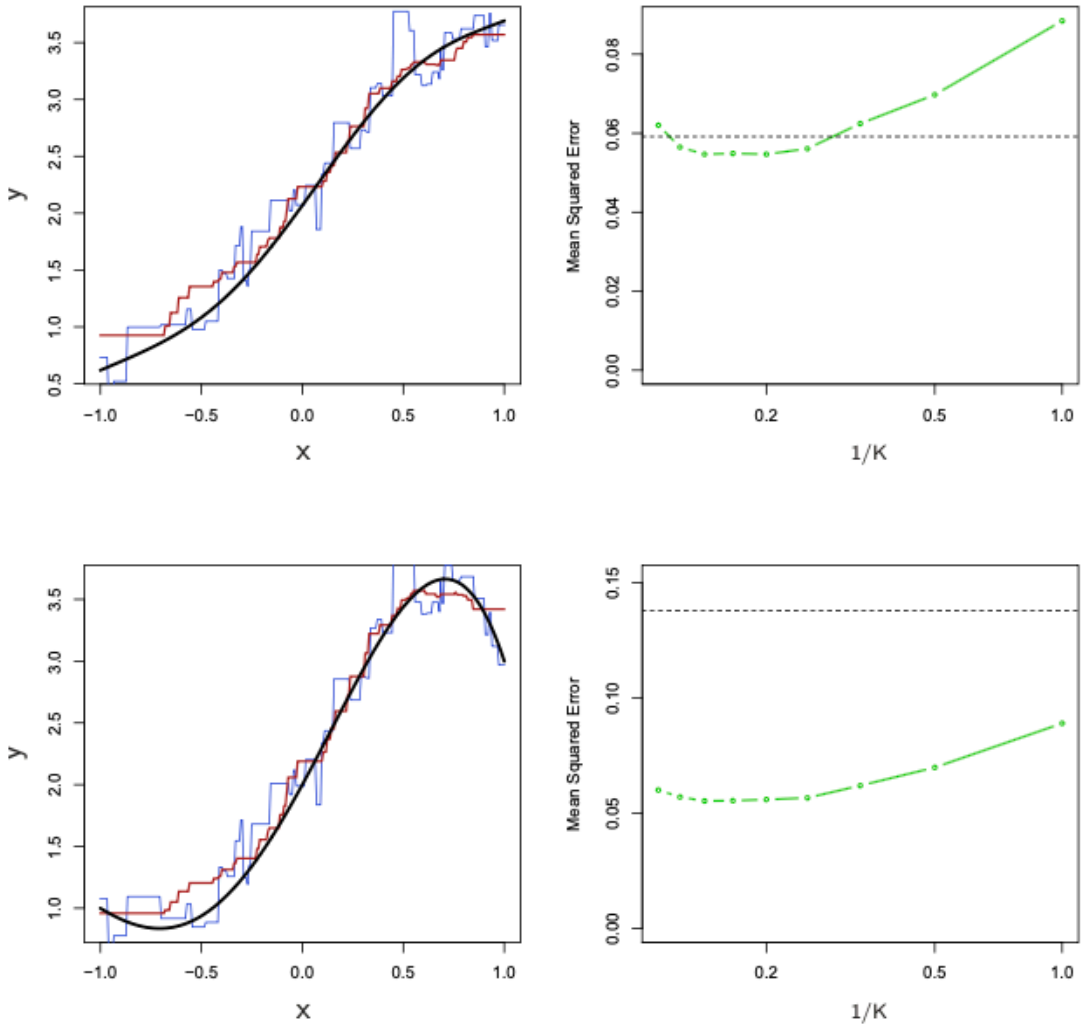

Figure 3.19 (slightly non-linear)

- Low 1/K (large K) → KNN bias ↑ but variance ↓

- High 1/K (small K) → high variance
- Linear regression wins when signal is almost linear

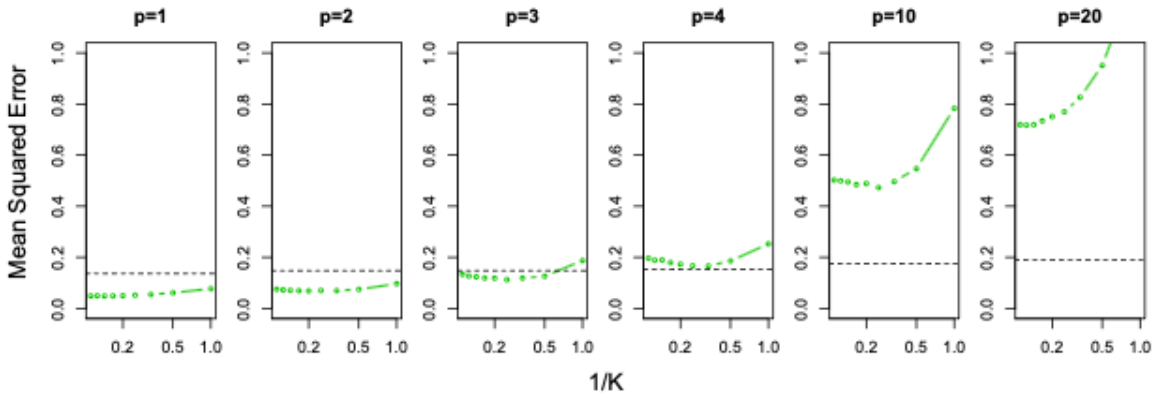

Figure 3.20 (curse of dimensionality demonstration)

- p=1,2 → KNN often beats linear (if true f non-linear)
- p≥4–10 → linear regression much more robust
- KNN MSE explodes as p increases (nearest neighbors become very far away)

Bias-variance takeaway

- Linear regression → high bias, low variance (especially when n small or p large)
- KNN → low bias (flexible), but very high variance unless K is large (and large K → bias ↑)
- **Parametric usually wins** when n is not huge and/or p is moderate-to-large

Interpretability bonus

Linear model → easy coefficients, p-values, confidence intervals
KNN → black box (no simple interpretation)