COM 525000 – Statistical Learning

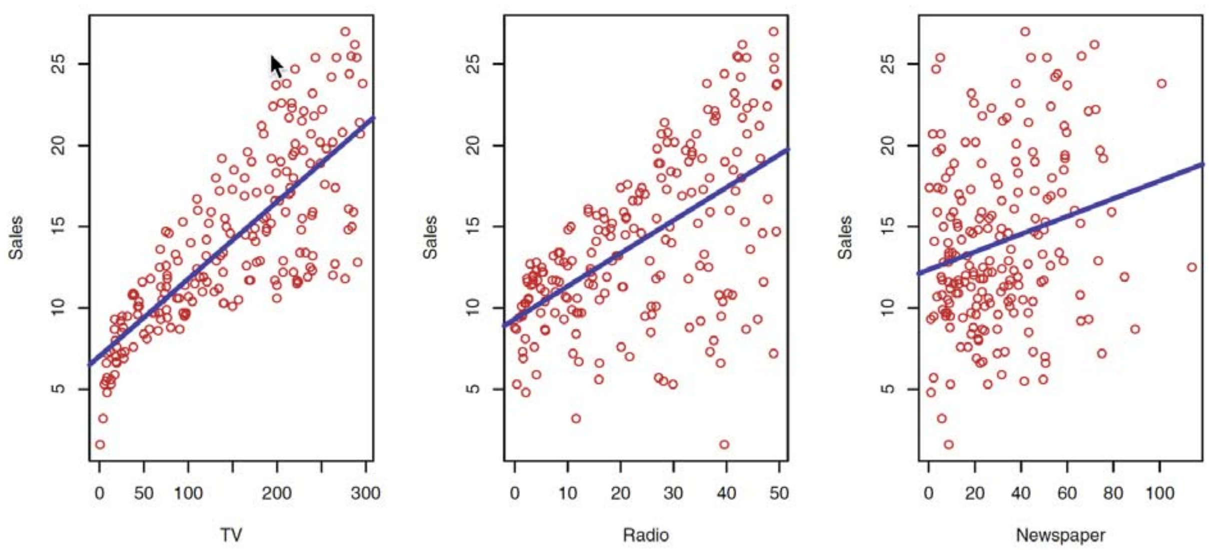
Lecture 3 – Linear Regression

Y.-W. Peter Hong

Recall Example

Example: (Advertising Data Set)

- Sales in 200 different markets versus the advertising budgets for TV, radio, and newspaper.



Simple Linear Regression

- **Goal:** Predict a quantitative response Y on the basis of a single predictor X under a *linear model*.
- Linear Model:

$$Y = \beta_0 + \beta_1 X + \epsilon$$

where β_0, β_1 are the model coefficients and ϵ is the error that captures noise and model mismatches.

E.g., $\texttt{sales} \approx \beta_0 + \beta_1 \times \texttt{TV}$

➔ β_0 is the intercept and β_1 is the slope.

- With estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ obtained from training data, a prediction of Y on the basis of $X = x$ is given by

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x.$$

Estimating Coefficients

- Let $\mathcal{D} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ be a set of n training data points.
- Find coefficient estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ such that the predictions

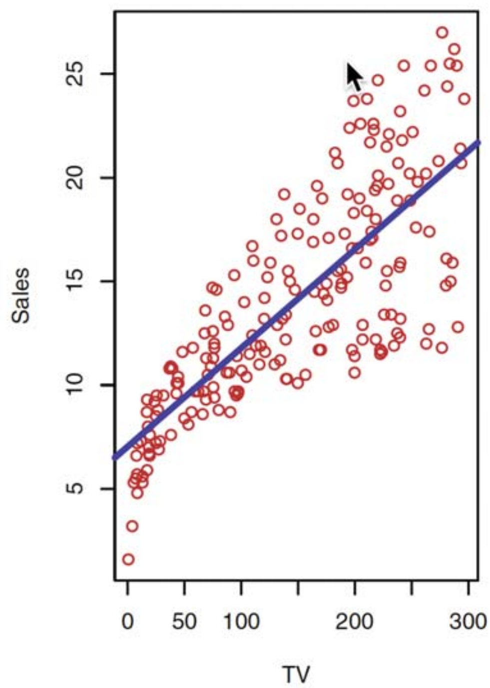
$$\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n$$

where $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ are as *close* as possible to observed responses

$$y_1, y_2, \ldots, y_n.$$

- Closeness can be measured using **residual sum of squares (RSS)**:

$$\mathrm{RSS}(\hat{\beta}_0, \hat{\beta}_1) = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2.$$



Least Squares Approach

- **Least squares (LS) approach:** Choose $\hat{\beta}_0$ and $\hat{\beta}_1$ to minimize the RSS. The solution is given by

$$\hat{\beta}_1^{(\mathcal{D})} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \quad \text{and} \quad \hat{\beta}_0^{(\mathcal{D})} = \bar{y} - \hat{\beta}_1^{(\mathcal{D})} \bar{x}$$

where $\bar{x} \triangleq \frac{1}{n} \sum_{i=1}^n x_i$ and $\bar{y} \triangleq \frac{1}{n} \sum_{i=1}^n y_i$.

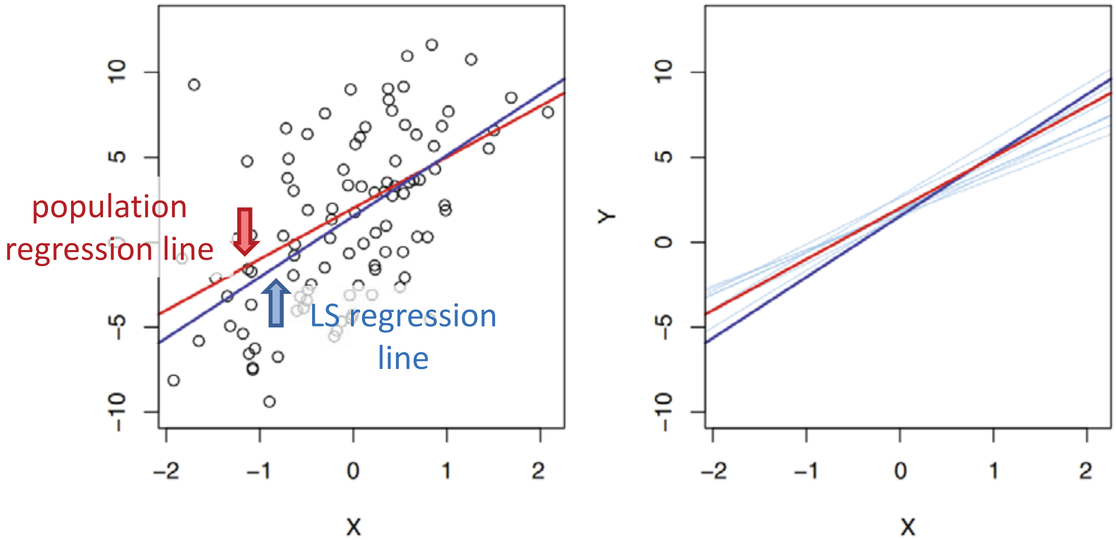
Proof:

Population vs LS Regression Lines

- Recall that, in linear regression, the model is given by

$$Y = \beta_0 + \beta_1 X + \epsilon.$$

- $f(X) = \beta_0 + \beta_1 X$ is the population regression line, and $\hat{f}(X;\mathcal{D}) = \hat{\beta}_0^{(\mathcal{D})} + \hat{\beta}_1^{(\mathcal{D})} X$ is the LS regression line.



Bias & Standard Error

- In general, the estimate $\hat{\theta}^{(\mathcal{D})}$ of an unknown parameter θ has mean squared error (MSE) $E\left[(\theta - \hat{\theta}^{(\mathcal{D})})^2\right] =$

- $\hat{\theta}^{(\mathcal{D})}$ is an *unbiased estimator* if $E[\hat{\theta}^{(\mathcal{D})}] = \theta$.
  - It does not systematically over- or under-estimate.
  - For independent data sets $\mathcal{D}_1, \mathcal{D}_2, \ldots$, we have
    $$\frac{1}{M}\sum_{m=1}^{M} \hat{\theta}^{(\mathcal{D}_m)} \to \theta \quad \text{as} \quad M \to \infty.$$
- *Standard error* $\mathrm{SE}(\hat{\theta}^{(\mathcal{D})})$ measures how far off a single estimate may be.

Example

Example: Let μ be the (population) mean of random variable Z, and let $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} z_i$ (sample mean) be an estimate μ, where $z_1, \ldots, z_n$ are n i.i.d. observations of Z. Show that $\hat{\mu}$ is unbiased and its standard error is

$$\mathrm{SE}(\hat{\mu}) = \sqrt{\mathrm{Var}(\hat{\mu})} = \sqrt{\frac{\sigma^2}{n}}$$

where $\sigma^2 = \mathrm{Var}(Z)$.

Standard Errors of LS Coefficients

- **Assumptions:** (i) in the model $Y = \beta_0 + \beta_1 X + \epsilon$, ϵ is independent of X; (ii) among observations $y_1, \ldots, y_n$, where $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, $\epsilon_1, \ldots, \epsilon_n$ are i.i.d. $\mathcal{N}(0, \sigma^2)$.
- The LS coefficients are *unbiased*, i.e., $E[\hat{\beta}_i] = \beta_i$, $i = 1, 2$,
- The standard errors (squared) are

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

$$\mathrm{SE}(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

where $\sigma^2 = \mathrm{Var}(\epsilon)$. (➔ shown later)

Confidence Intervals (1/2)

- In practice, σ is unknown, but can be estimated by the *residual standard error (RSE)*

$$\hat{\sigma} \triangleq \sqrt{\frac{\mathrm{RSS}}{n-2}} \left(= \mathrm{RSE} \right).$$

➔ $\hat{\sigma}^2 = \frac{\mathrm{RSS}}{(n-2)}$ is an unbiased estimate of σ^2. $(\Rightarrow \widehat{\mathrm{SE}}(\hat{\beta}_i)^2)$

- A ***h%-confidence interval*** is a range of values that contains the true unknown parameter with *h*% prob.
- For linear regression, the 95%-confidence interval is

$$[\hat{\beta}_i - 2 \cdot \mathrm{SE}(\hat{\beta}_i), \hat{\beta}_i + 2 \cdot \mathrm{SE}(\hat{\beta}_i)]$$

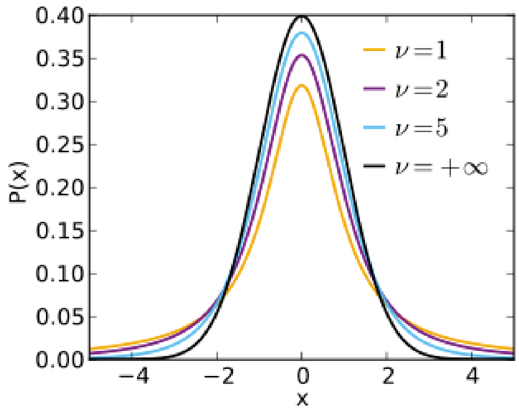
$$\approx \quad [\hat{\beta}_i - 2 \cdot \widehat{\mathrm{SE}}(\hat{\beta}_i), \hat{\beta}_i + 2 \cdot \widehat{\mathrm{SE}}(\hat{\beta}_i)], \text{ for } i = 1, 2.$$

  - E.g., in sales vs TV advertising, they are [6.130, 7.935] and [0.042, 0.053] for $\hat{\beta}_0$ and $\hat{\beta}_1$, respectively.

Confidence Intervals (2/2)

- More precisely, we want to find c such that

$$\Pr\left(\hat{\beta}_i - c \cdot \widehat{\mathrm{SE}}(\hat{\beta}_i) \leq \beta_i \leq \hat{\beta}_i + c \cdot \widehat{\mathrm{SE}}(\hat{\beta}_i)\right) = h\%$$



➔ t statistic: $\dfrac{\hat{\beta}_i - \beta_i}{\widehat{\mathrm{SE}}(\hat{\beta}_i)}$ has student *t* distribution with $n-2$ degrees of freedom.

➔ Approaches $\mathcal{N}(0,1)$ as degrees of freedom increases.

Testing Significance of the Input (1/2)

- To test the significance of the input on the output, we perform the hypothesis test:

$$\begin{aligned} \mathcal{H}_0 &: \text{ No relation between } X \text{ and } Y \text{ (i.e., } \beta_1 = 0) \\ \mathcal{H}_1 &: \text{ Some relation between } X \text{ and } Y \text{ (i.e., } \beta_1 \neq 0) \end{aligned}$$

- That is, each hypothesis assumes different models:

$$\begin{aligned} \mathcal{H}_0 &: \ Y = \beta_0 + \epsilon \\ \mathcal{H}_1 &: \ Y = \beta_0 + \beta_1 X + \epsilon \end{aligned}$$

- <u>Key idea:</u> Observe how far $\hat{\beta}_1$ is to 0 relative to $\widehat{\mathrm{SE}}(\hat{\beta}_1)$.

➔ ***t-statistic:*** $t = \dfrac{\hat{\beta}_1 - 0}{\widehat{\mathrm{SE}}(\hat{\beta}_1)}$ (under $\mathcal{H}_0$).

  – Under $\mathcal{H}_0$, t follows the *student t-distribution* with $n-2$ degrees of freedom.

Testing Significance of the Input (2/2)

- **Question:** How large should t be in order to reject the null hypothesis?
- Answer: Compute the *p-value*

$$p = \Pr(|T| \geq |t|; \mathcal{H}_0).$$

and reject $\mathcal{H}_0$ if *p* is small.

➔ Typical cutoff values: 5% or 1% (i.e., $|t| = 2 \text{ or } 2.75$).

- E.g., sales vs TV advertising example:

$$\texttt{sales} \approx \beta_0 + \beta_1 \times \texttt{TV}$$

	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

Assess Model Accuracy (1/2)

- The above discussion is based on the linear model

$$Y = \beta_0 + \beta_1 X + \epsilon$$

➔ **Question:** Does the linear model fit the data well?

- **Residual Standard Error (RSE):**

$$\mathrm{RSE} \triangleq \sqrt{\frac{\mathrm{RSS}}{n-2}} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2}}.$$

➔ RSE measures the *lack of fit* of the model.

(Note: The true model may be nonlinear, or may depend on other predictors not considered here.)

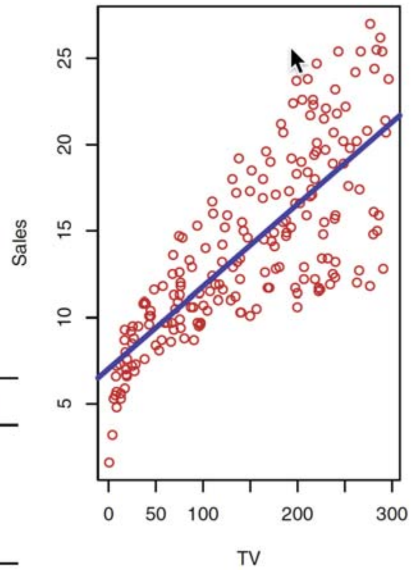
Assess Model Accuracy (2/2)

- R^2 **Statistic:**

$$R^2 \triangleq \frac{\text{TSS} - \text{RSS}}{\text{TSS}} = 1 - \frac{\text{RSS}}{\text{TSS}} \in [0, 1]$$

where $\text{TSS} \triangleq \sum_{i=1}^{n} (y_i - \bar{y})^2$ is the total sum of squares.

➔ R^2 measures *the proportion of variability in* Y that is explained by performing regression.



- F **Statistic:** $F \triangleq \dfrac{\text{TSS} - \text{RSS}}{\text{RSS}/(n-2)}$.

E.g., sales vs TV advertising

Quantity	Value
Residual standard error	3.26
R^2	0.612
F-statistic	312.1

Multiple Linear Regression

- Linear model with p predictors:

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon$$

(e.g., $\texttt{sales} = \beta_0 + \beta_1 \times \texttt{TV} + \beta_2 \times \texttt{radio} + \beta_3 \times \texttt{newspaper} + \epsilon$)

- By letting $X = (1, X_1, \ldots, X_p)^T$ and $\beta = (\beta_0, \beta_1, \ldots, \beta_p)^T$, we can express the model as $Y = X^T \beta + \epsilon$.
- **Goal:** Find $\hat{\beta} = (\hat{\beta}_0, \ldots, \hat{\beta}_p)^T$ that minimizes the RSS

$$\text{RSS}(\hat{\beta}) = \sum_{i=1}^{n} (y_i - x_i^T \hat{\beta})^2 = \|\mathbf{y} - \mathbf{X}\hat{\beta}\|^2$$

where (x_i, y_i)'s are the training data points.

➔ Here, $\mathbf{y} = (y_1, \ldots, y_n)^T$, $\mathbf{X} = (x_1, \ldots, x_n)^T$, and $x_i = (1, x_{i1}, \ldots, x_{ip})^T$.

LS Solution for Multiple Linear Regression

- For $\mathbf{X}$ with full column rank, minimizing RSS yields the least squares (LS) solution

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

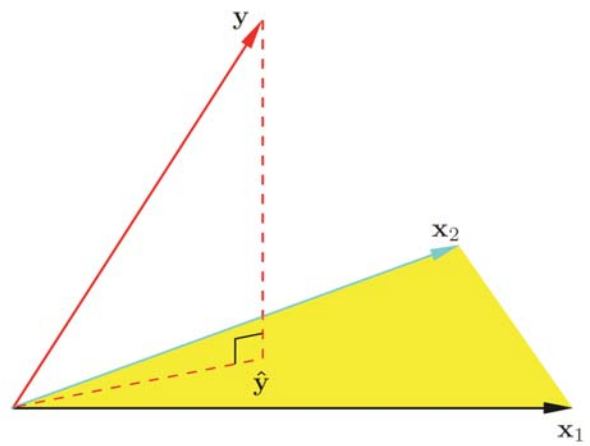
Proof: (Differentiate RSS w.r.t. $\hat{\beta}$ and set it to zeros.)

➔ The fitted values at the n training inputs are

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

Geometric Interpretation

- Let $\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_p$ (with $\mathbf{x}_0 = \mathbf{1}$) be the columns of $\mathbf{X}$.
- Finding $\hat{\mathbf{y}} = \mathbf{X}\hat{\beta}$ to minimize $\mathrm{RSS} = \|\mathbf{y} - \hat{\mathbf{y}}\|^2$ is equivalent to finding the orthogonal projection of $\mathbf{y}$ onto the subspace $\mathrm{span}\{\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_p\}$.



➔ Insight applies even if $\mathbf{X}$ is not full column rank.

Simple vs Multiple Linear Regression

Simple regression of `sales` on `radio`

	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

Simple regression of `sales` on `newspaper`

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

	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

Properties of $\hat{\beta}$

- **Assumption:** (i) ϵ is independent of X; (ii) $\epsilon_1, \dots, \epsilon_n$ are i.i.d. $\mathcal{N}(0, \sigma^2)$ (i.e., $\boldsymbol{\epsilon} = (\epsilon_1, \dots, \epsilon_n)^T \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I})$).
- Given $\mathcal{D}$ (or $\mathbf{X}$), we have $\hat{\beta} \sim \mathcal{N}(\beta, (\mathbf{X}^T\mathbf{X})^{-1}\sigma^2)$.

  Proof:

➔ The standard error of $\hat{\beta}_j$ is

$$\mathrm{SE}(\hat{\beta}_j) = \sqrt{\mathrm{Var}(\hat{\beta}_j)} = \sqrt{\{(\mathbf{X}^T\mathbf{X})^{-1}\sigma^2\}_{jj}}.$$

Noise Variance Estimation

- In practice, σ^2 is unknown, but can be estimated by

$$\hat{\sigma}^2 = \frac{1}{n-p-1}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2.$$

- It can be shown that $(n-p-1)\hat{\sigma}^2 \sim \sigma^2 \chi^2_{n-p-1}$ (and thus $E[\hat{\sigma}^2] = \sigma^2$).

Testing Significance of Inputs (1/2)

- For each coefficient β_j, we can test the hypothesis that $\beta_j = 0$ by computing the *t*-statistic

$$t_j = \frac{\hat{\beta}_j - 0}{\widehat{\mathrm{SE}}(\hat{\beta}_j)} = \frac{\hat{\beta}_j}{\hat{\sigma}\sqrt{\{(\mathbf{X}^T\mathbf{X})^{-1}\}_{jj}}}.$$

t-distribution with $n-p-1$ degrees of freedom under $\mathcal{H}_0$

- **Testing Significance of Groups of Coefficients:**

$$\mathcal{H}_0:\ \beta_j = 0,\ j = p-q+1, \ldots, p$$
$$\mathcal{H}_1:\ \exists j \in \{p-q+1, \ldots, p\},\ \beta_j \neq 0$$

Use **F-statistic:** $F \triangleq \dfrac{[\mathrm{RSS}(\hat{\hat{\beta}}_{1:p-q}) - \mathrm{RSS}(\hat{\beta})]/q}{\mathrm{RSS}(\hat{\beta})/(n-p-1)}.$

➔ The normalized RSS reduction due to the q inputs.

➔ $F = t^2$ when $p = q = 1$.

Testing Significance of Inputs (2/2)

- Under the null hypothesis $\mathcal{H}_0$,

$$\mathrm{RSS}(\hat{\hat{\beta}}_{1:p-q}) - \mathrm{RSS}(\hat{\beta}) \sim \sigma^2\chi^2_q$$
$$\mathrm{RSS}(\hat{\beta}) \sim \sigma^2\chi^2_{n-p-1}$$

and, thus, F follows an F-distribution with $(q, n-p-1)$ degrees of freedom.

➔ This implies that $E\big[[\mathrm{RSS}(\hat{\hat{\beta}}_{1:p-q}) - \mathrm{RSS}(\hat{\beta})]/q\big] = \sigma^2$ and $E\big[\mathrm{RSS}(\hat{\beta})/(n-p-1)\big] = \sigma^2$ under $\mathcal{H}_0$.

- Reject null hypothesis $\mathcal{H}_0 : \beta_j = 0,\ j = p-q+1, \ldots, p$ if

$$p = \Pr(\mathcal{F} \geq F; \mathcal{H}_0) \leq 0.05 \text{ (for example)}.$$

Joint vs Individual Tests of Significance

Question: Why do we need to test multiple inputs together as opposed to separately?

Gauss-Markov Theorem (1/2)

- **Theorem (Gauss-Markov):** Suppose that

$$\mathbf{y} = \mathbf{X}\beta + \boldsymbol{\epsilon}$$

where $E(\boldsymbol{\epsilon}) = \mathbf{0}$ and $\mathrm{Var}(\boldsymbol{\epsilon}) = E[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T] = \sigma^2\mathbf{I}$ (i.e., *uncorrelated* and *equal variance* (homoscedastic)). The LS solution $\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$ is the **best linear unbiased estimator (BLUE)**. That is, for any $\tilde{\beta} = \mathbf{C}\mathbf{y}$ such that $E[\tilde{\beta}] = \beta$, we have $\mathrm{Var}(\tilde{\beta}) - \mathrm{Var}(\hat{\beta}) \succeq \mathbf{0}$.

Proof:

Gauss-Markov Theorem (2/2)

Questions for Linear Regression (1/3)

Question 1: Is at least one of a set/subset of predictors useful in predicting the response?

$$\mathcal{H}_0 : \ \beta_j = 0, \ j \in \mathcal{Q}$$
$$\mathcal{H}_1 : \ \exists j \in \mathcal{Q}, \ \beta_j \neq 0$$

➔ Compute F-statistic and see if p-value is less than 5%.

Question 2: Given that the subset of predictors may be useful, which ones are most important?

➔**Variable selection** (Chapter 6)

➔Mallow's C_p, Akaike information criterion (AIC), Bayesian information criterion (BIC), adjusted R^2 etc.

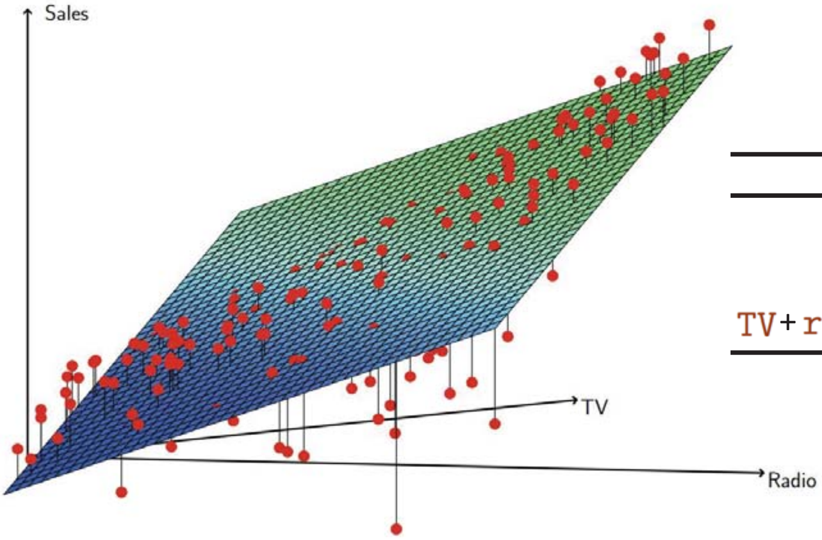
➔Forward, backward, and mixed selection.

Questions for Linear Regression (2/3)

Question 3: How well does the model fit the data?

➔Measured by

$$R^2 \triangleq \frac{\text{TSS} - \text{RSS}}{\text{TSS}} \quad \text{and} \quad \text{RSE} = \sqrt{\frac{1}{n-p-1}\text{RSS}}.$$



	R^2	RSE
TV	0.61	3.26
TV+radio	0.89719	1.681
TV+radio+newspaper	0.8972	1.686

Questions for Linear Regression (3/3)

Question 4: Given new predictor values, what should our prediction be? How accurate is it?

➔Let $Y = f(X) + \epsilon$ be the true model, $X^T\beta$ be the best linear approx. to $f(X)$, and $X^T\hat{\beta}^{(\mathcal{D})}$ be our prediction.

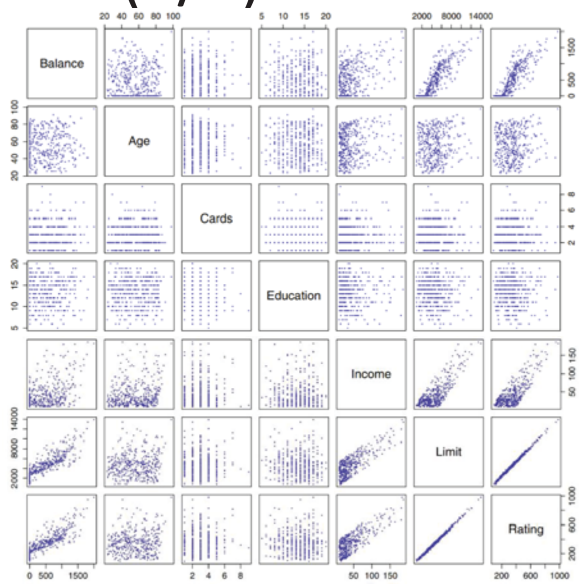
The mean squared error (MSE) is

$$E\left[(Y - X^T\hat{\beta}^{(\mathcal{D})})^2\right] = E\left[(X^T\hat{\beta}^{(\mathcal{D})} - E_{\mathcal{D}}[X^T\hat{\beta}^{(\mathcal{D})}|X])^2\right] + E\left[(f(X) - E_{\mathcal{D}}[X^T\hat{\beta}^{(\mathcal{D})}|X])^2\right] + \text{Var}(\epsilon).$$

➔ $E_{\mathcal{D}}[X^T\hat{\beta}^{(\mathcal{D})}|X] = X^T\beta$ since $\hat{\beta}^{(\mathcal{D})}$ is unbiased.

- Variance: $E\left[(X^T\hat{\beta}^{(\mathcal{D})} - X^T\beta)^2\right]$
- Model Bias: $E\left[(f(X) - X^T\beta)^2\right]$

Qualitative Predictors (1/2)



Credit card data set:

age, cards, income ➔quantitative

gender, student, ethnicity ➔qualitative

E.g., predict balance by gender

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

The model becomes

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i, & \text{if } i\text{th person is female,} \\ \beta_0 + \epsilon_i, & \text{if } i\text{th person is male.} \end{cases}$$

	Coefficient	Std. error	t-statistic	p-value
Intercept	509.80	33.13	15.389	< 0.0001
gender[Female]	19.73	46.05	0.429	0.6690

Qualitative Predictors (2/2)

For qualitative predictors with more than two levels, e.g., ethnicity (Asian, Caucasian, African American)

$$x_{i1} = \begin{cases} 1, & \text{if } i\text{th person is Asian,} \\ 0, & \text{if } i\text{th person is not Asian.} \end{cases}$$

$$x_{i2} = \begin{cases} 1, & \text{if } i\text{th person is Caucasian,} \\ 0, & \text{if } i\text{th person is not Caucasian.} \end{cases}$$

The model becomes

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i, & \text{if Asian,} \\ \beta_0 + \beta_2 + \epsilon_i, & \text{if Caucasian,} \\ \beta_0 + \epsilon_i, & \text{if African American.} \end{cases}$$

	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 (1/3)

- Two main assumptions: (i) additivity and (ii) linearity.
- **Removing the additive model:**
  - Additivity implies that the changes in response Y due to changes in some X_j is independent of other predictors.
  - *Interaction* among variables may exist. (E.g., spending money on TV ads may increase effectiveness of radio ads)

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon \qquad (R^2 = 89.7\%)$$
$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \epsilon \qquad (R^2 = 96.8\%)$$

	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

Extensions to the Linear Model (2/3)

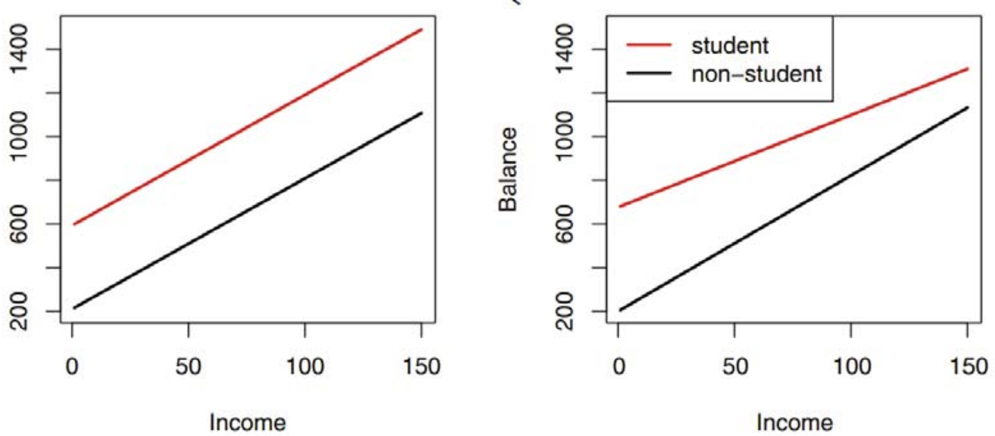
- E.g., predicting balance using income (quantitative) and student status (qualitative)

No interaction:

$$\texttt{balance}_i \approx \beta_0 + \beta_1 \times \texttt{income}_i + \begin{cases} \beta_2 & \text{if } i\text{th person is a student} \\ 0 & \text{if } i\text{th person is not a student} \end{cases}$$

Interaction:

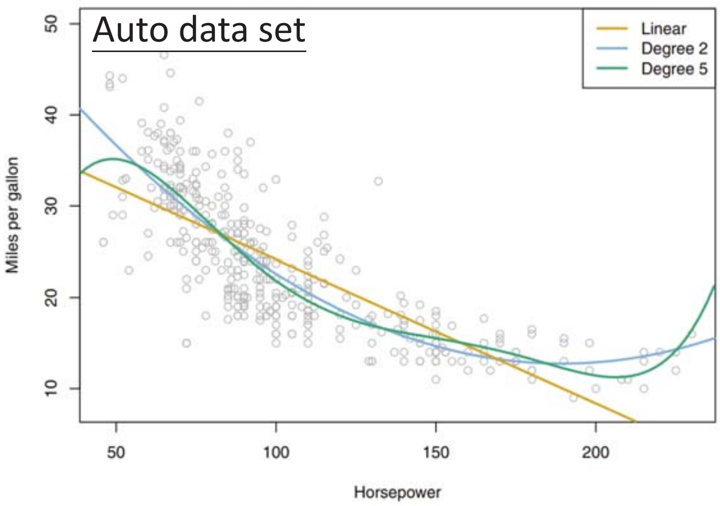
$$\texttt{balance}_i \approx \beta_0 + \beta_1 \times \texttt{income}_i + \begin{cases} \beta_2 + \beta_3 \times \texttt{income}_i & \text{if student} \\ 0 & \text{if not student} \end{cases}$$



Extensions to the Linear Model (3/3)

- **Incorporating nonlinear relationships:**
  - Linearity implies that changes in response Y due to changes in X_j does not depend on its current value.
  - E.g., polynomial regression

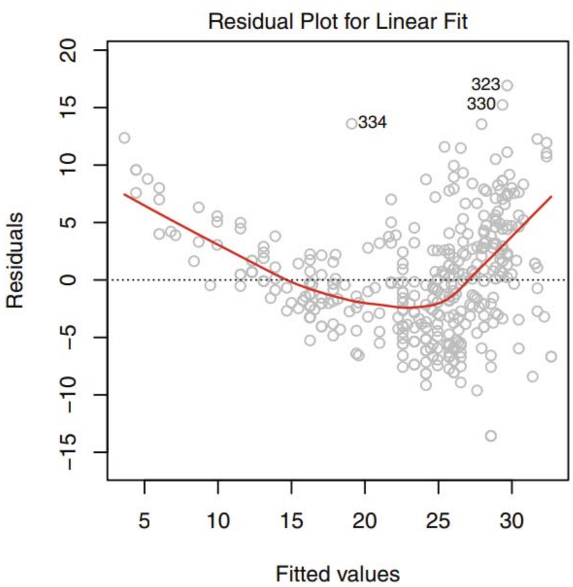
$$\texttt{mpg} = \beta_0 + \beta_1 \times \texttt{horsepower} + \beta_2 \times \texttt{horsepower}^2 + \epsilon$$

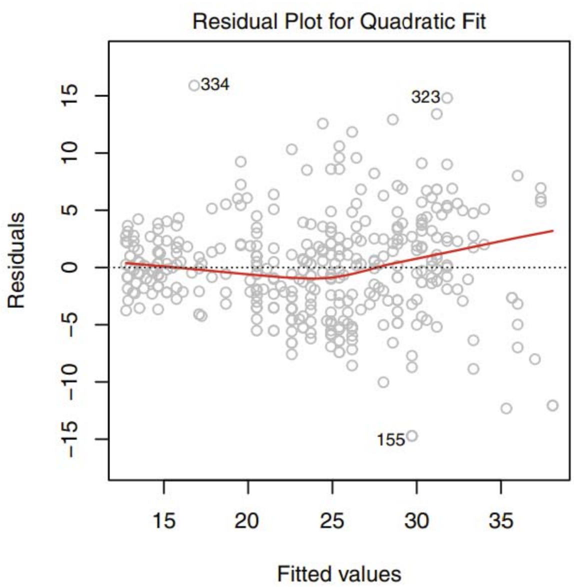


	Coefficient	Std. error	p-value
Intercept	56.9001	1.8004	< 0.0001
horsepower	−0.4662	0.0311	< 0.0001
$\texttt{horsepower}^2$	0.0012	0.0001	< 0.0001

Potential Problems - Nonlinearity

1. Nonlinearity of Data:
  - The true model is far from linear.
  - Residual plot $e_i = y_i - \hat{y}_i$ versus x_i (or fitted value $\hat{y}_i$)

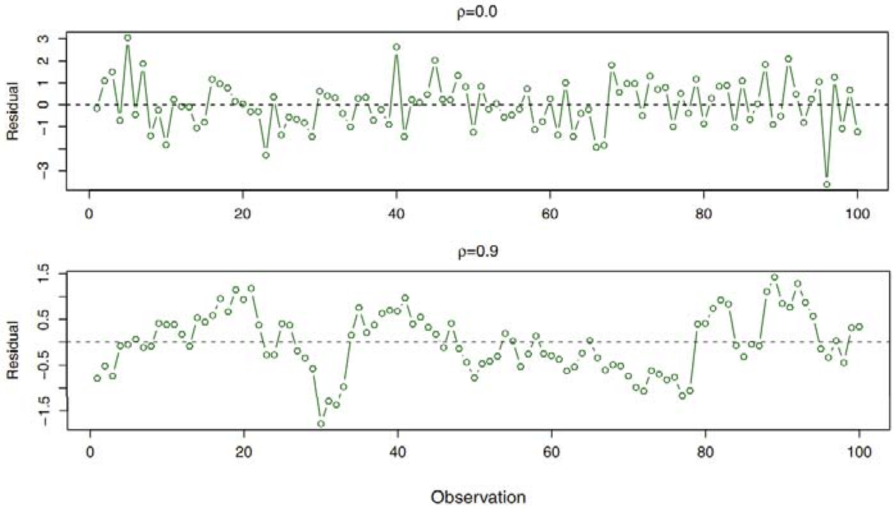




Potential Problems – Error Correlation

2. Correlation of the Error Terms

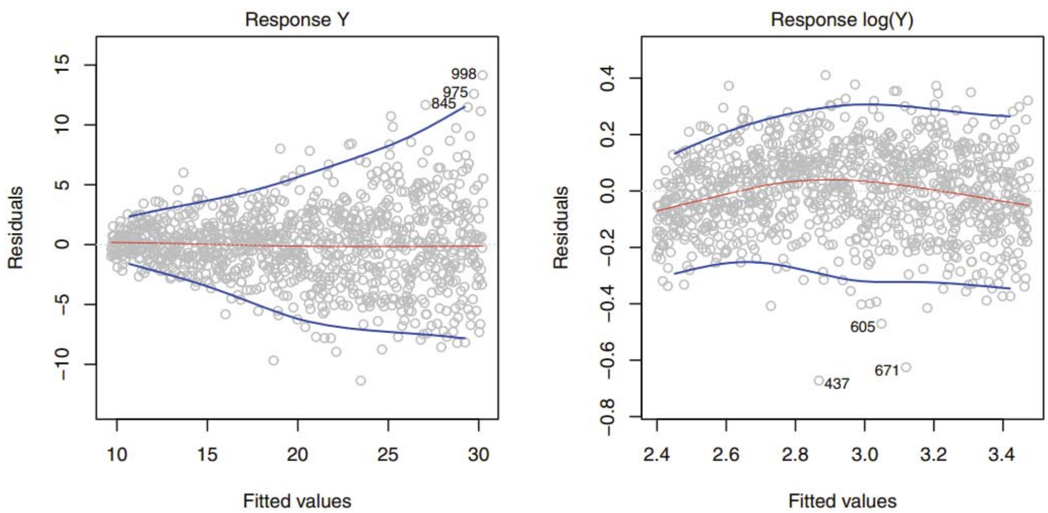
- $\epsilon_1, \ldots, \epsilon_n$ are correlated ➔ under-estimated standard error ➔ narrower confidence interval, and lower p-value.
- E.g., adjacent samples in a time series, samples from the same group (e.g., family), or spatial location (environment)



Potential Problems - Heteroscedasticity

3. Non-Constant Variance of Error Terms:

- $\epsilon_1, \ldots, \epsilon_n$ have different variance (heteroscedasticity)
- Use nonlinear transformations, such as $\log Y$ or $\sqrt{Y}$.



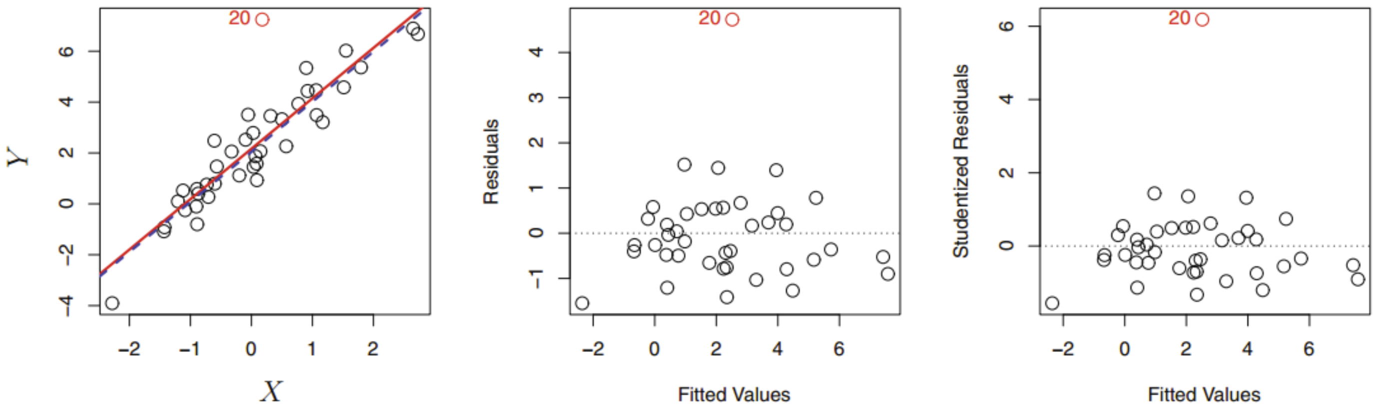
➔weighted least squares, i.e., minimize

$$\mathrm{RSS}(\hat{\beta}) = (\mathbf{y} - \mathbf{X}\hat{\beta})^T \mathbf{W} (\mathbf{y} - \mathbf{X}\hat{\beta})$$

Potential Problems - Outliers

4. Outliers

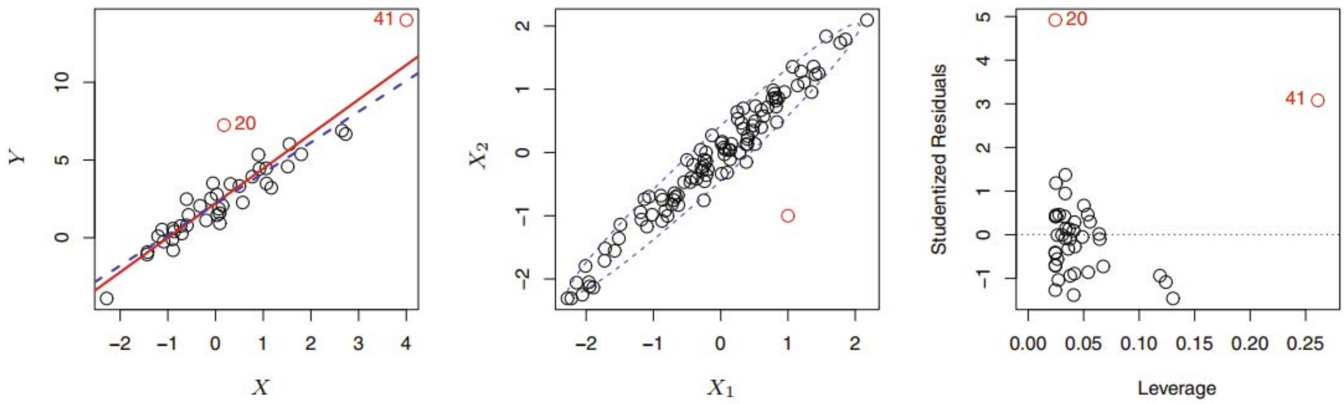
- May cause dramatic increase in RSE and R^2.
  (e.g., RSE is 0.77 (without) or 1.09 (with))
  ➔ Affects confidence interval and p-value computation.



Potential Problems – High Leverage Points

5. High Leverage Points

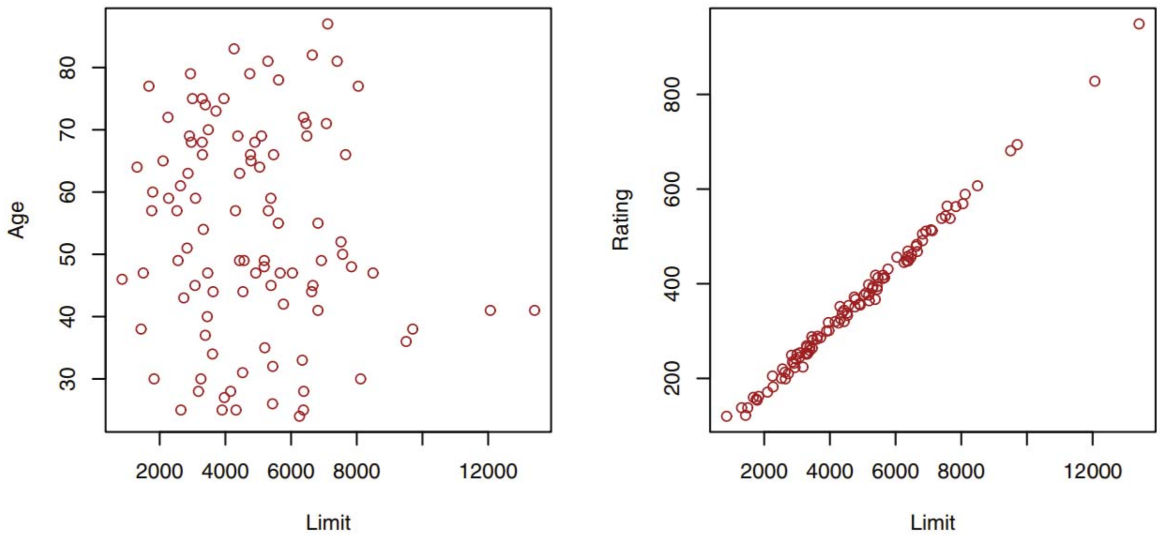
- Leverage statistic for ith data point is $h_i = \{\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\}_{i,i}$.
  For $p = 1$, $h_i = \dfrac{1}{n} + \dfrac{(x_i - \bar{x})^2}{\sum_{i'=1}^{n}(x_i - \bar{x})^2}$.
- Notice that $\mathrm{Var}(e_i) = \mathrm{Var}(y_i - \hat{y}_i) = \sigma^2(1 - h_i)$.
- The studentized residual is $t_i = \dfrac{e_i}{\hat{\sigma}\sqrt{1 - h_i}}$.



Potential Problems – Collinearity (1/2)

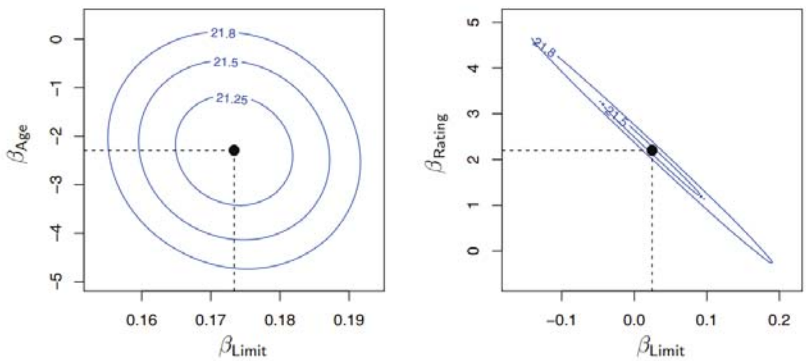
6. Collinearity

- The situation in which two or more predictor variables are closely related to each other.
- Difficult to see how each separately affects the response.



Potential Problems – Collinearity (1/2)

- Increases the uncertainty of the coefficient estimates (i.e., standard error).
  - ➔ smaller t-statistic (null hypothesis more likely)



- Collinearity (or multicollinearity) can be assessed using the variance inflation factor (VIF) defined as

$$\mathrm{VIF}(\hat{\beta}_j) = \frac{1}{1 - R^2_{X_j|X_{-j}}}$$

where $R^2_{X_j|X_{-j}}$ is the R^2 from a regression of X_j onto all of the other predictors

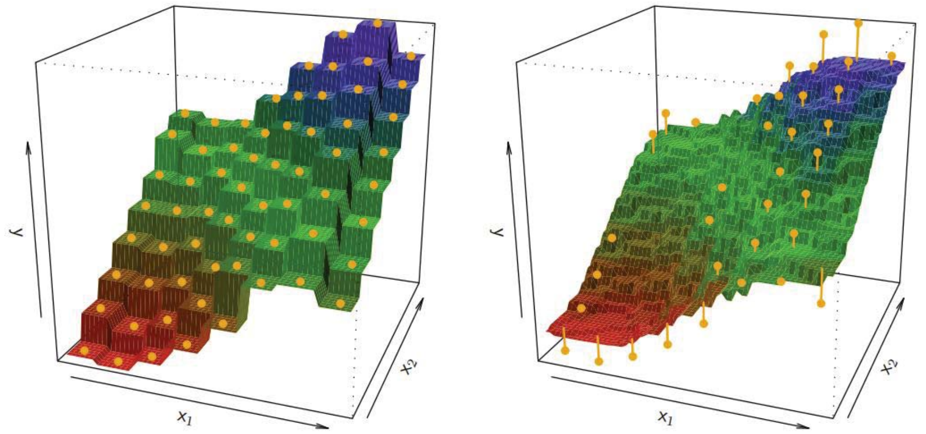
Comparison with kNN (1/4)

- The **k-nearest neighbor (kNN)** prediction for new input x_0 is defined as

$$\hat{y}_0 = \hat{f}(x_0) = \frac{1}{k} \sum_{i:x_i \in \mathcal{N}_k(x_0)} y_i$$

where $\mathcal{N}_k(x_0)$ represents the set of k nearest $x_i's$ in in the vicinity of x_0.

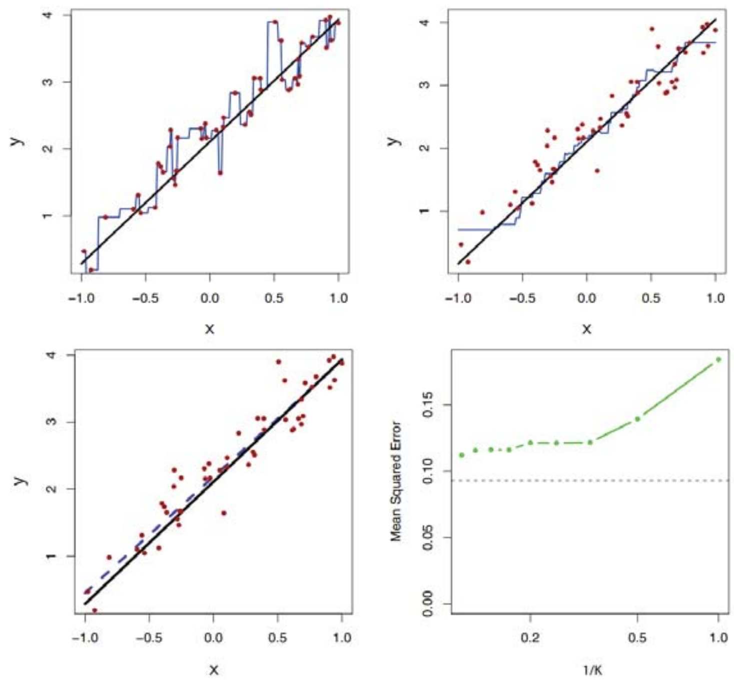
$\underline{k=1}$



$\underline{k=9}$

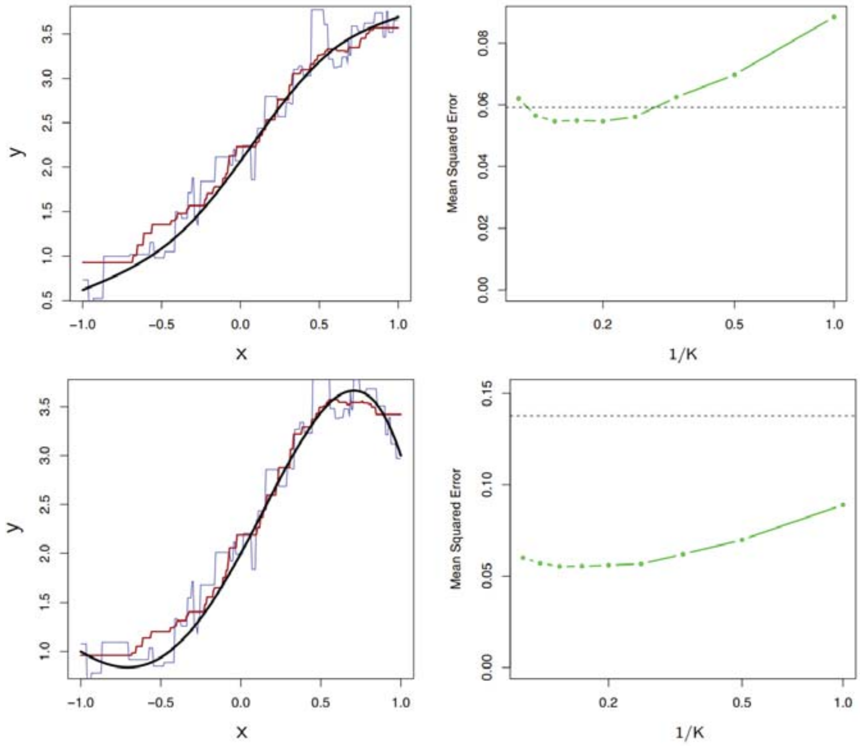
Comparison with kNN (2/4)

- Linear regression (parameter) outperforms kNN (non-parametric) when the true model is linear.



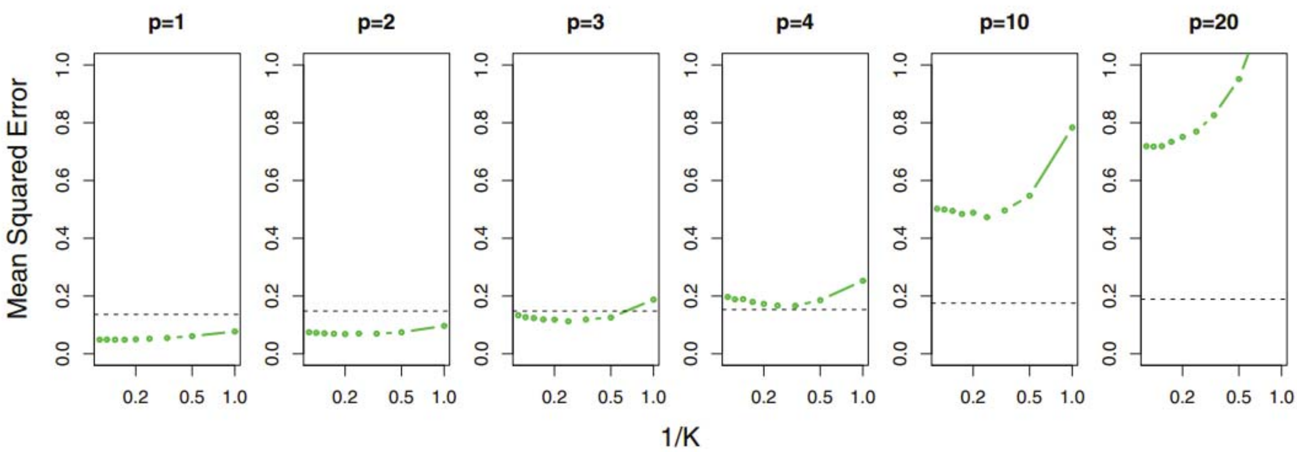
Comparison with kNN (3/4)

- kNN performs well under high nonlinearity.



Comparison with kNN (4/4)

- Linear regression also may perform better in higher dimensions.



- Linear regression also be preferable in terms of interpretability.