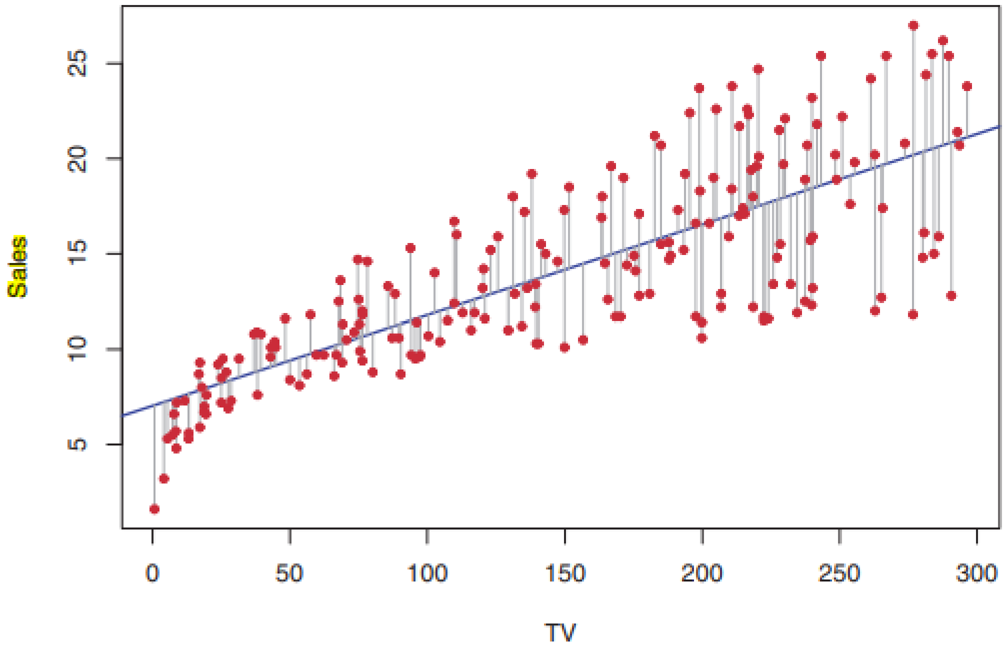
Linear Regression-2

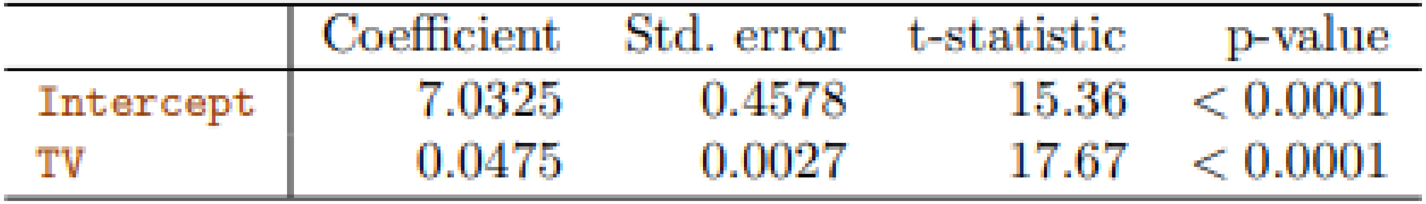
Prof. Asim Tewari
IIT Bombay

Simple Linear regression



For the Advertising data, the least squares fit for the regression of sales onto TV is shown. The fit is found by minimizing the sum of squared errors. Each grey line segment represents an error, and the fit makes a compromise by averaging their squares. In this case a linear fit captures the essence of the relationship, although it is somewhat deficient in the left of the plot.

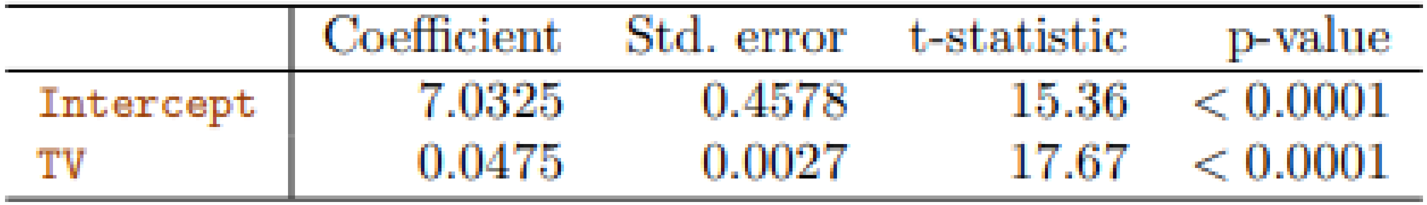
Simple Linear regression



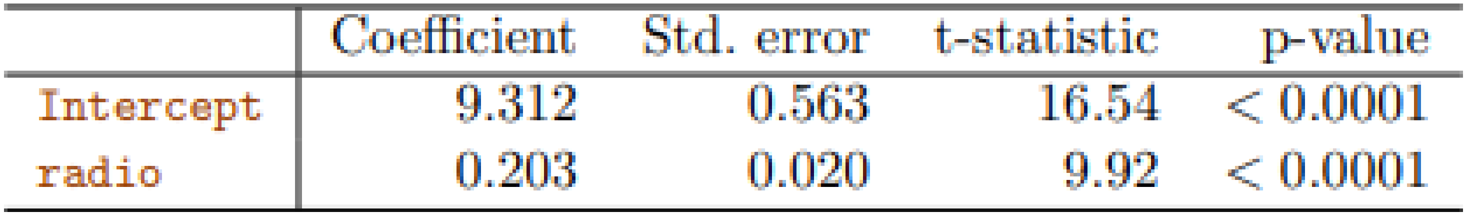
	Coefficient	Std. error	t-statistic	p-value
Intercept	7.0325	0.4578	15.36	< 0.0001
TV	0.0475	0.0027	17.67	< 0.0001

For the Advertising data, coefficients of the least squares model for the regression of number of units sold on TV advertising budget. An increase of $1,000 in the TV advertising budget is associated with an increase in sales by around 50 units (Recall that the sales variable is in thousands of units, and the TV variable is in thousands of dollars)

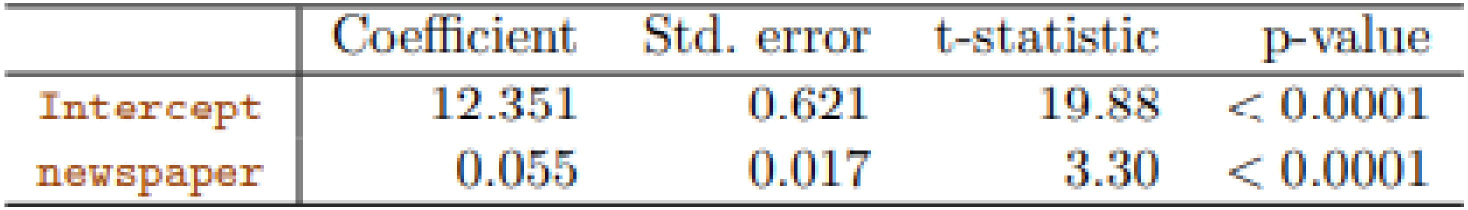
Simple Linear regression



	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	< 0.0001

Multiple Linear Regression

Multiple Linear Regression assumes

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$$

The model can be expressed as

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_p x_p$$

with its coefficients being derived by minimizing RSS

$$\begin{aligned} \text{RSS} &= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \\ &= \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \cdots - \hat{\beta}_p x_{ip})^2 \end{aligned}$$

$$\sum_{i=1}^{n} |y_i - \hat{y}_i|$$

Multiple Linear Regression

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ \vdots \\ y_n \end{bmatrix} = \underset{n \times p+1}{\begin{bmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_n^T \end{bmatrix}} \underset{(p+1)\times 1}{\begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \vdots \\ \beta_p \end{bmatrix}}$$

$$x_1 = \underset{p \times 1}{\begin{bmatrix} 1 \\ x_1^1 \\ x_1^2 \\ \vdots \\ x_1^p \end{bmatrix}}$$

$$\underset{n\times 1}{Y} = \underset{n \times (p+1)}{X}\ \underset{(p+1)\times 1}{\beta}$$

# of Features = p

# of sample points = n

Multiple Linear Regression

$$RSS = \|\underbrace{X\beta}_{\hat{Y}} - Y\|^2$$

$$RSS = f(\beta)$$

Find β^* such that $RSS(\beta^*)$ is minimum.

Solve $\nabla RSS(\beta^*) = 0$ to get β^*.

Multiple Linear Regression

$$RSS(\beta) = \| X\beta - Y \|^2 = (X\beta - Y)^T (X\beta - Y)$$

$$= (X\beta)^T X\beta - (X\beta)^T Y - Y^T X\beta + Y^T Y$$

$$= \beta^T X^T X \beta - 2\beta^T X^T Y + Y^T Y$$

$$\ulcorner \; \nabla_x (a^T x) = a \text{ and } \nabla_x (x^T A x) = (A + A^T) x \; \lrcorner$$

Multiple Linear Regression

$$RSS(\beta) = \beta^T X^T X \beta - 2\beta^T X^T y + y^T y$$

$$\nabla_\beta RSS(\beta) = (X^T X + X^T X)\beta - 2X^T y$$

$$\nabla_x (a^T x) = a$$

$$\nabla_x (x^T A x) = (A + A^T)x$$

$$= 2X^T X\beta - 2X^T y$$

$\therefore$ Solving $\nabla_\beta RSS(\beta^*) = 0$ for β^*

$$2X^T X \beta^* - 2X^T y = 0$$

$$\Rightarrow \boxed{\underset{(p+1)\times 1}{\underset{\uparrow}{\beta^*}} = (X^T X)^{-1}(X^T y)}$$

Multiple Linear Regression

$$\beta^* = (X^T X)^{-1} X^T Y$$

↳ Is this min or max?

Is this local or global?

$$\because \quad \underbrace{\nabla_\beta RSS(\beta)}_{} = 2X^TX\beta - 2X^TY$$

$$\therefore \quad \nabla^2 RSS(\beta) = \underbrace{2\,\underset{\uparrow\,(p+1)n}{X^T}\;\underset{\uparrow\,n\times(p+1)}{X}}_{(p+1)\times(p+1)} - 0$$

Multiple Linear Regression

$$\underbrace{\nabla^2 RSS(\beta)}_{\text{Hessian of RSS}} = 2 X^T X$$

$$\forall \beta \quad \beta^T (2 X^T X) \beta = 2 (X\beta)^T X\beta$$

$$= 2 \| X\beta \|^2 \geq 0$$

$\therefore$ β^* is a global min

Multiple Linear Regression

$$\beta^* = \underbrace{(X^T X)}^{-1} \underbrace{X^T Y}$$

$$X = \begin{bmatrix} 1 & x_1^1 & x_1^2 & \cdots & x_1^p \\ \vdots & & & & \\ 1 & x_n^1 & x_n^2 & \cdots & x_n^p \end{bmatrix}$$

$n \times (p+1)$

$$\underbrace{(p+1) \times n \quad n \times (p+1)}$$

$(p+1)\ (p+1)$

$$\underbrace{(p+1) \times n \quad n \times 1}$$

$(p+1)\ 1$

$(p+1) \times 1$

Linear Regression

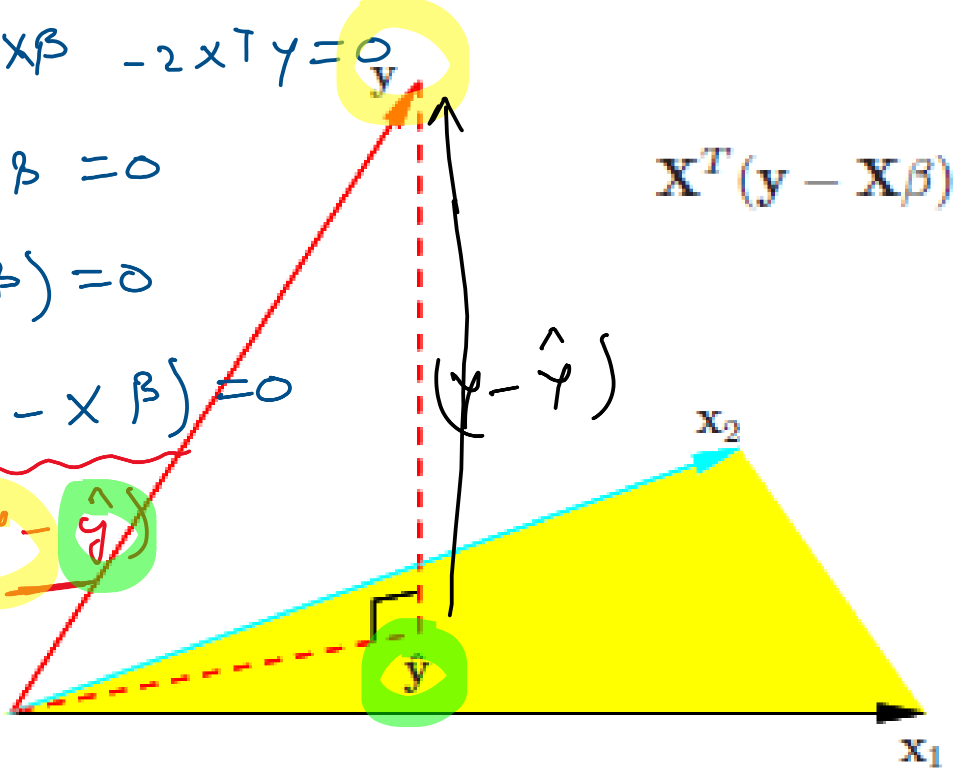
$$\nabla_\beta RSS(\beta) = 2X^T X\beta - 2X^T y = 0$$

$$\Rightarrow \quad X^T y - X^T X \beta = 0$$

$$\Rightarrow \quad X^T(y - X\beta) = 0$$

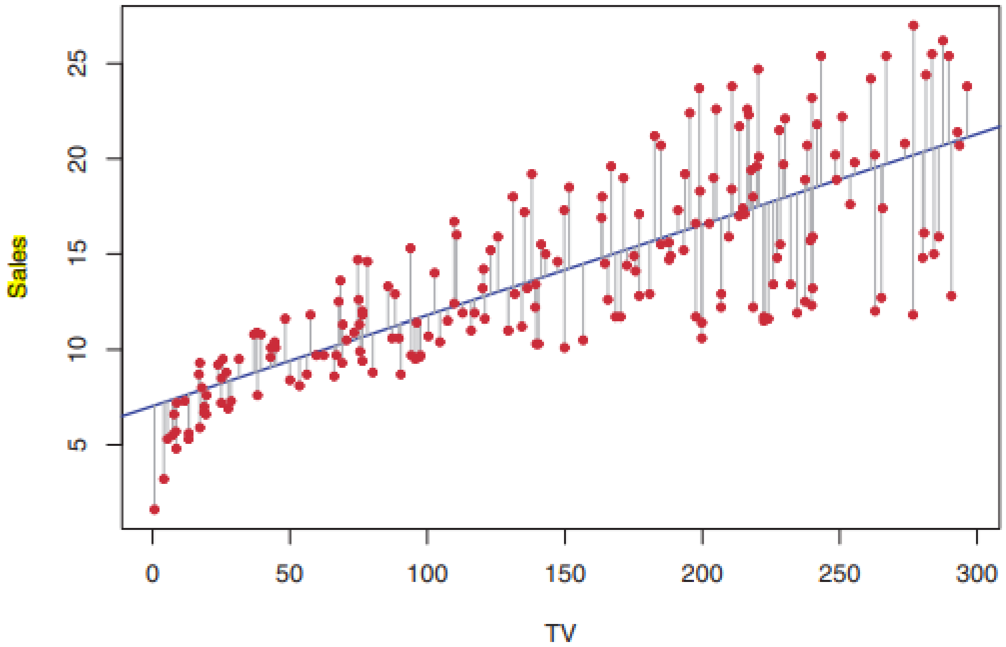
for β^* $$\underset{\uparrow \atop X^T}{X^T} \underbrace{(y - X\beta)}_{(y - \hat{y})} = 0$$

$$\mathbf{X}^T(\mathbf{y} - \mathbf{X}\beta) = 0$$



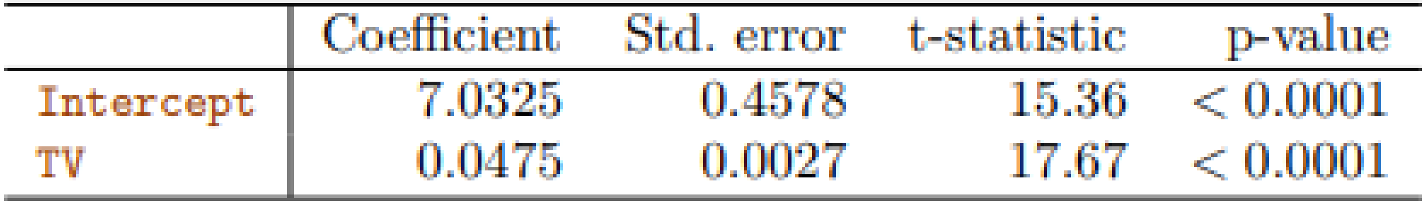
The N-dimensional geometry of least squares regression with two predictors. The outcome vector y is orthogonally projected onto the hyperplane spanned by the input vectors **x₁** and **x₂**. The projection ˆ**y** represents the vector of the least squares predictions

Simple Linear regression



For the Advertising data, the least squares fit for the regression of sales onto TV is shown. The fit is found by minimizing the sum of squared errors. Each grey line segment represents an error, and the fit makes a compromise by averaging their squares. In this case a linear fit captures the essence of the relationship, although it is somewhat deficient in the left of the plot.

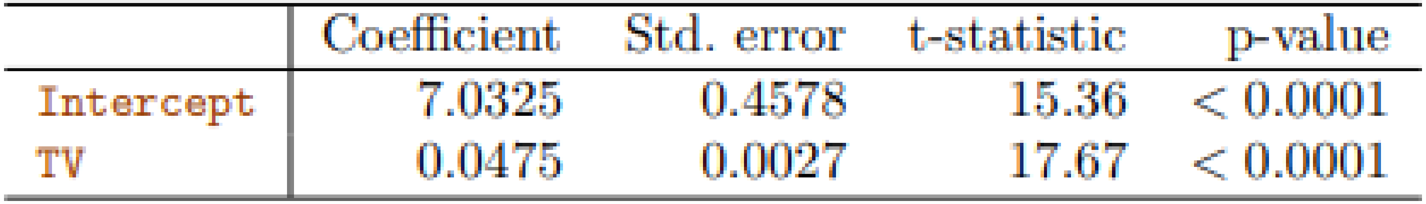
Simple Linear regression



	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

For the Advertising data, coefficients of the least squares model for the regression of number of units sold on TV advertising budget. An increase of $1,000 in the TV advertising budget is associated with an increase in sales by around 50 units (Recall that the sales variable is in thousands of units, and the TV variable is in thousands of dollars)

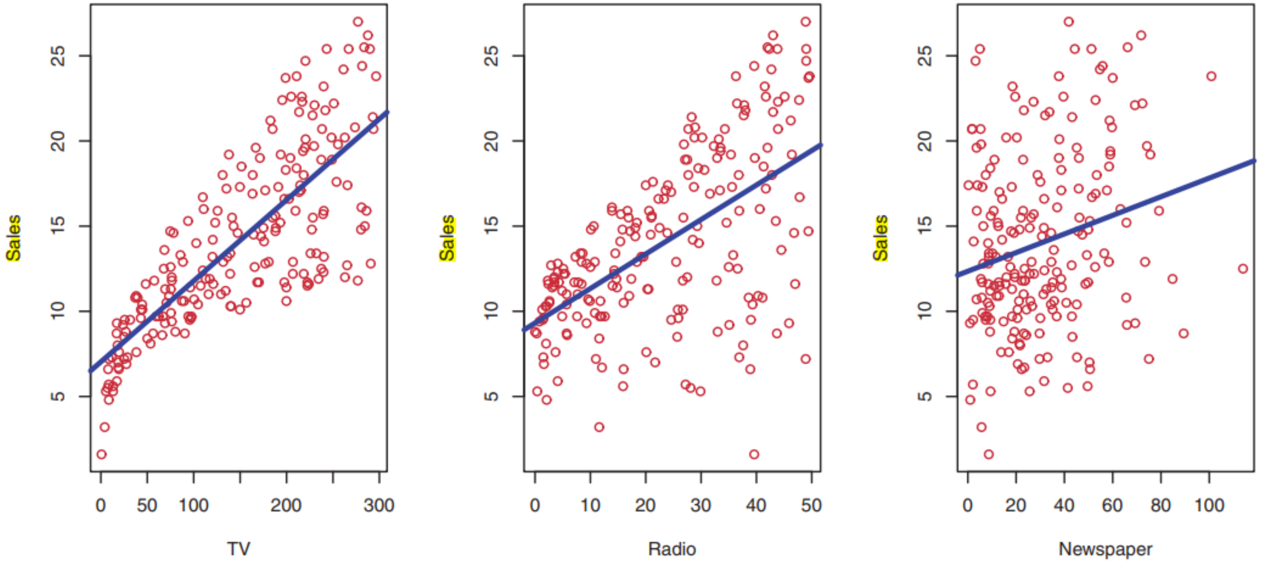
Simple Linear regression



	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

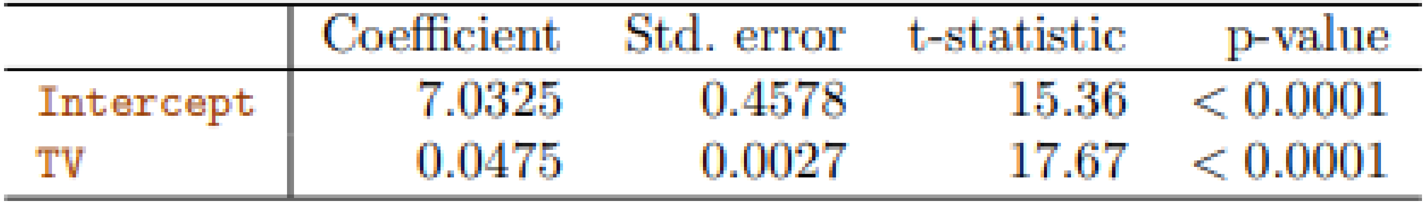
A small p-value for the intercept indicates that we can reject the null hypothesis that $\beta_0 = 0$, and a small p-value for TV indicates that we can reject the null hypothesis that $\beta_1 = 0$. Rejecting the latter null hypothesis allows us to conclude that there is a relationship between TV and sales. Rejecting the former allows us to conclude that in the absence of TV expenditure, sales are non-zero

The Advertising data set

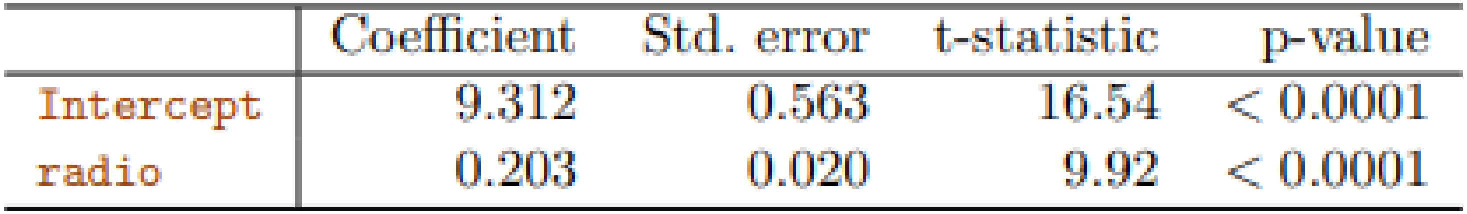


The Advertising data set. The plot displays sales, in thousands of units, as a function of TV, radio, and newspaper budgets, in thousands of dollars, for 200 different markets. In each plot we show the simple least squares fit of sales to that variable. In other words, each blue line represents a simple model that can be used to predict sales using TV, radio, and newspaper, respectively.

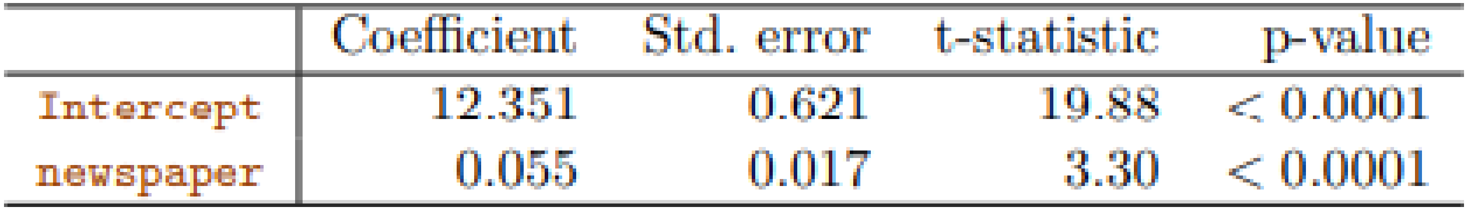
Simple Linear regression



	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

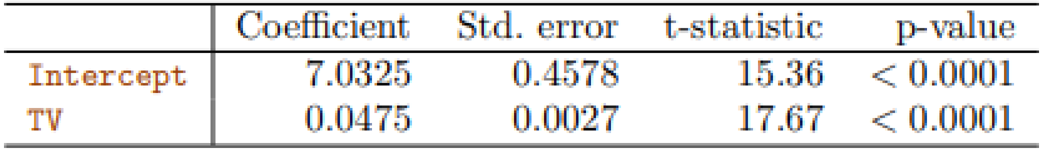


	Coefficient	Std. error	t-statistic	p-value
Intercept	9.312	0.563	16.54	< 0.0001
radio	0.203	0.020	9.92	< 0.0001

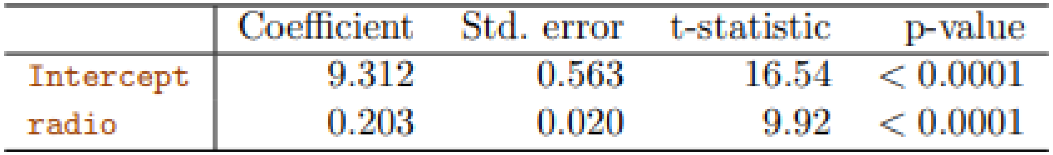


	Coefficient	Std. error	t-statistic	p-value
Intercept	12.351	0.621	19.88	< 0.0001
newspaper	0.055	0.017	3.30	< 0.0001

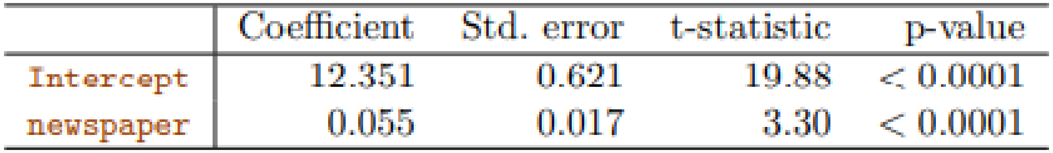
Simple Linear regression



	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

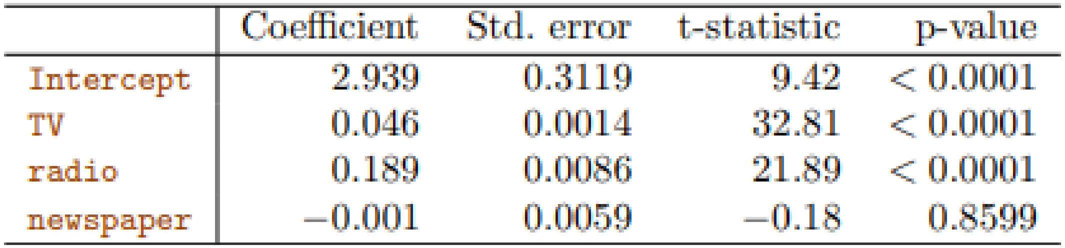


	Coefficient	Std. error	t-statistic	p-value
Intercept	9.312	0.563	16.54	< 0.0001
radio	0.203	0.020	9.92	< 0.0001



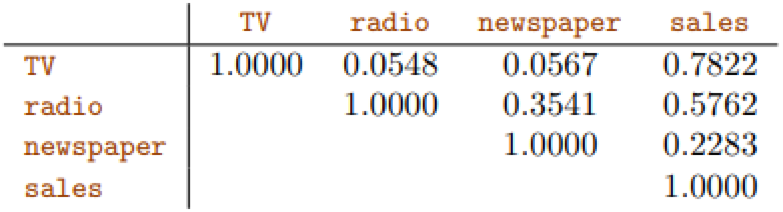
	Coefficient	Std. error	t-statistic	p-value
Intercept	12.351	0.621	19.88	< 0.0001
newspaper	0.055	0.017	3.30	< 0.0001

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



	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

Correlation matrix for TV, radio, newspaper, and sales for the Advertising data.

This reveals a tendency to spend more on newspaper advertising in markets where more is spent on radio advertising.

So newspaper sales are a surrogate for radio advertising; newspaper gets “credit” for the effect of radio on sales.

Key questions in a regression analysis

1. Is at least one of the predictors X1, X2,...,Xp useful in predicting the response?

2. Do all the predictors help to explain Y , or is only a subset of the predictors useful?

3. How well does the model fit the data?

4. Given a set of predictor values, what response value should we predict, and how accurate is our prediction?

Is There a Relationship Between the Response and Predictors?

We test the null hypothesis,

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0 \Big\} \rightarrow$$

If this is true

$(TSS - RSS)/p \rightarrow \sigma^2$

versus the alternative

$$H_a : \text{ at least one } \beta_j \text{ is non-zero.}$$

This hypothesis test is performed by computing the F-statistic,

If H_0 is true

$F \approx 1$

If H_0 is not true

$F >> 1$

$$F = \frac{(\text{TSS} - \text{RSS})/p}{\underbrace{\text{RSS}/(n-p-1)}_{\rightarrow \sigma^2}},$$

Linear model is corret

Value of F-statistic close to 1 when null hypothesis true

Value of F-statistic greater than 1 when alternative hypothesis true

Hypothesis testing in multi linear regression

- F is very close to **one** we cannot reject the null hypothesis (thus, in a sense we accepted the **null hypothesis**)

- If F is **very large** we reject the null hypothesis (thus, in a sense we accepted the **alternate hypothesis**)

How large is large enough?

- This depends upon the values of n and p.
- If n is very large a small value above 1 is also a compelling evidence against the null hypothesis; however if n is a small then F has to be very large for us to reject the null hypothesis.

- When the null hypothesis is true and the error follows a Gaussian distribution, then it can be shown that F-statistic follows F-distribution

Hypothesis testing in multi linear regression

Why do we need F-statistic when t-statistic already exists?

(Given these individual p-values for each variable, why do we need to look at the overall F-statistic? After all, it seems likely that if any one of the p-values for the individual variables is very small, then at least one of the predictors is related to the response.)

- However, the above logic is flawed, especially when the number of predictors p is large.

- For instance, consider an example in which p = 100 and H0 : β1 = β2 = . . . = βp = 0 is true, so no variable is truly associated with the response. In this situation, about 5% of the p-values associated with each variable will be below 0.05 by chance. In other words, we expect to see approximately five small p-values even in the absence of any true association between the predictors and the response.

So which variables are important?

- If alternate hypothesis is true then **at least** one β_j is non-zero.
- But which are the important variables?
- For this we can test another null hypothesis that a particular subset of q of the coefficients are zero.

$$H_0: \quad \beta_{p-q+1} = \beta_{p-q+2} = \ldots = \beta_p = 0,$$

- This is done by putting the q variables (chosen for omission) at the end of the list and fitting a second model that uses all the variables except those last q.
- Suppose that the residual sum of squares for that model is RSS_0. Then the appropriate F-statistic is

$$F = \frac{(\mathrm{RSS}_0 - \mathrm{RSS})/q}{\mathrm{RSS}/(n-p-1)}$$

Extensions of Linear Models

- Removing the Additive Assumption

Introduce the interactive term

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \epsilon.$$

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

- Where $\tilde{\beta}_1 = \beta_1 + \beta_3 X_2.$
- Non-linear Relationships

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

Nonlinear regression

$$\left.\begin{array}{l} y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \ldots \\[2ex] \boxed{y = \underset{\uparrow}{\beta_0} + \underset{\uparrow}{\beta_1} x + \underset{\uparrow}{\beta_2} x^2} \end{array}\right\} \text{Find } \beta_0, \beta_1, \beta_2 \ldots$$

$$\equiv y = \beta_0 + \beta_1 \underset{\underset{x}{\uparrow}}{x_1} + \beta_2 \underset{\underset{x^2}{\uparrow}}{x_2}$$

Basis function regression

$$y = \beta_0 + \beta_1 f_1(\bar{X}) + \beta_2 f_2(\bar{X}) + \beta_3 f_3(\bar{X}) + \cdots + \epsilon$$

$$y = \underline{\beta_0} + \underline{\beta_1}\, \underset{\underset{x}{\uparrow}}{f_1(x)} + \underline{\beta_2}\, \underset{\underset{x^2}{\uparrow}}{f_2(X)} + \cdots + \epsilon$$

e.g. ①

② $f_1(x) = x^{1.5}$, $f_2(x) = \left(1 + \frac{x}{3+x^2}\right)$

Multiple Linear Regression

$y = \beta_0 + \beta_1 x + \epsilon$ → not continuous.

- Predictors with Only Two Levels
- Define new variables

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

The model then takes the form

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

Multiple Linear Regression

- Predictors with more than Two Levels
- Define new variables

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

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

The model then takes the form

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

Potential problems of Linear Regression

1. *Non-linearity of the response-predictor relationships.*
2. *Correlation of error terms.*
3. *Non-constant variance of error terms.*
4. *Outliers.*
5. *High-leverage points.*
6. *Collinearity.*

Potential problems of Linear Regression

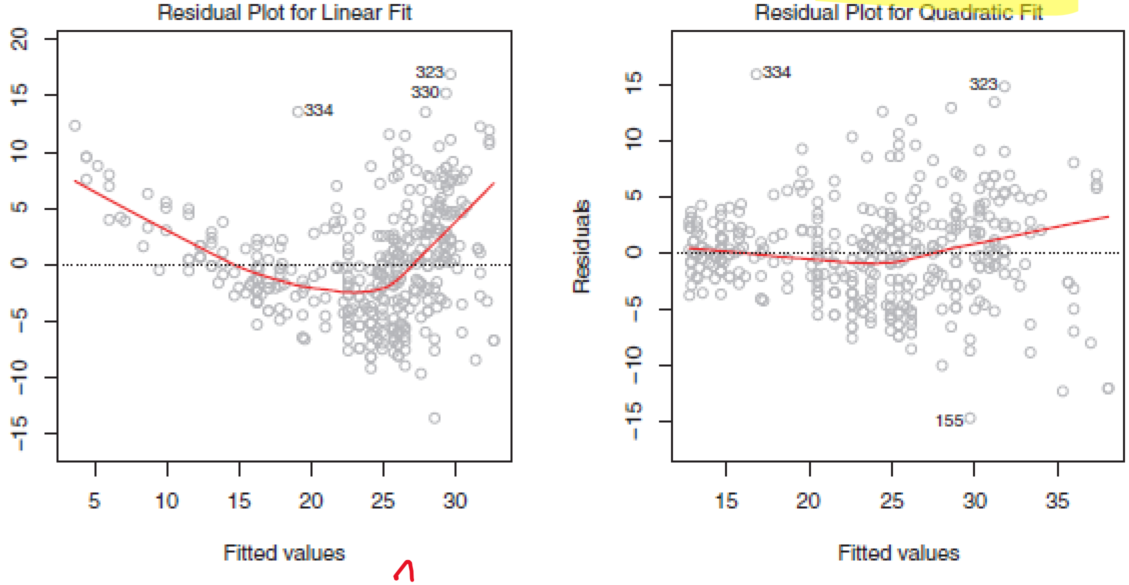
$y = \sum_{i=1}^{k} \beta_k x^k$?

1. *Non-linearity of the response-predictor relationships*

$y = \beta_0 + \beta_1 x + \boxed{\beta_2 x^2}$

$y = \beta_0 + \beta_1 x$

$e = (y - \hat{y})$

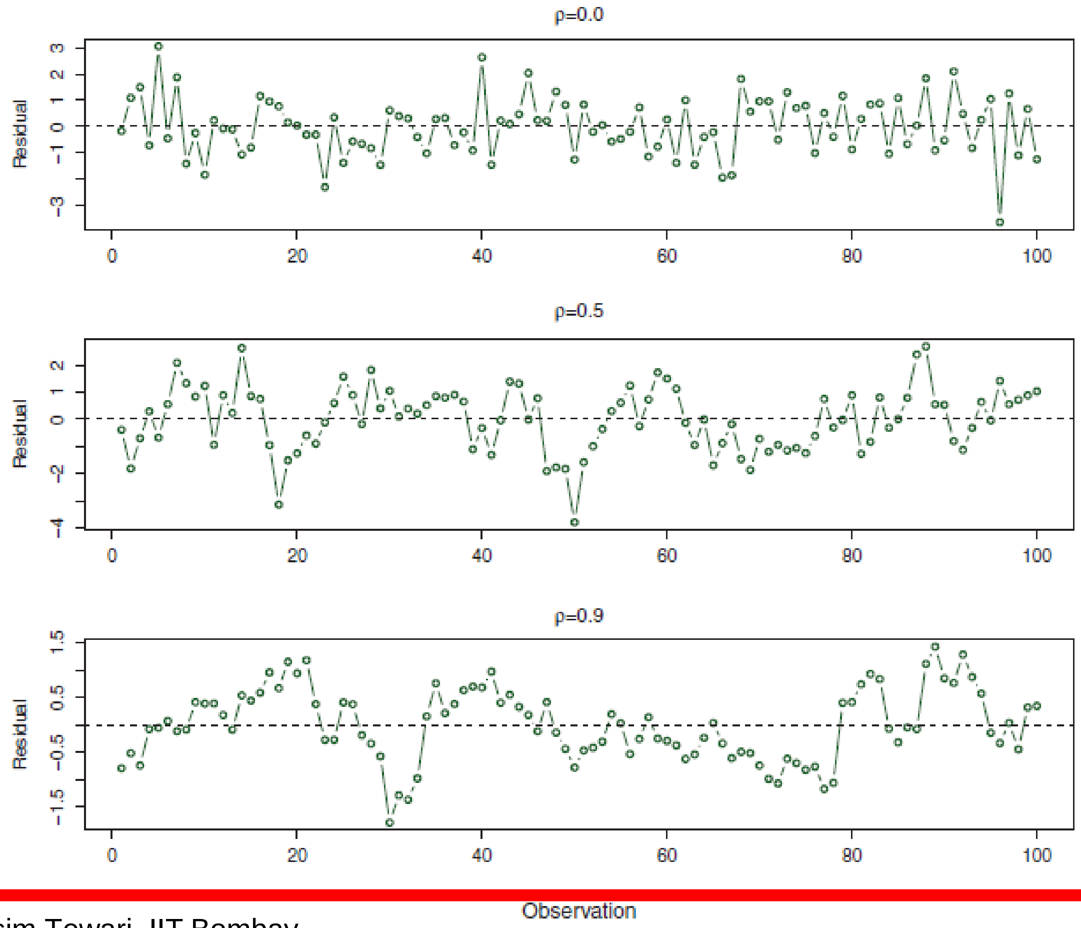


$\longrightarrow \hat{y}$

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.*

Potential problems of Linear Regression

2. *Correlation of error terms*



Plots of residuals from simulated time series data sets generated with differing levels of correlation ρ between error terms for adjacent time points.

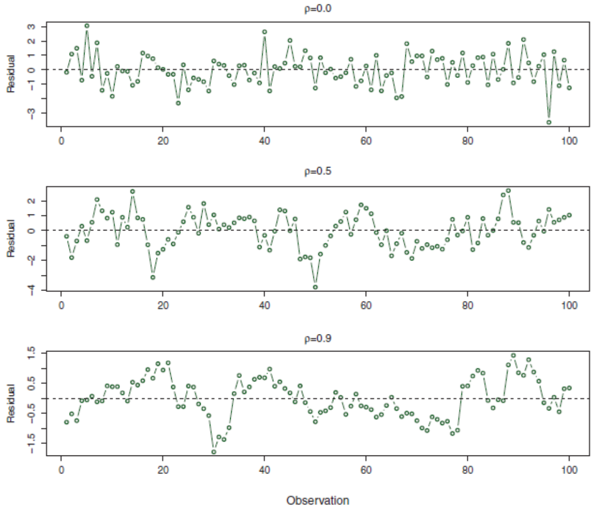
Potential problems of Linear Regression

2. *Correlation of error terms*

*For example, if we a*ccidentally doubled our data, leading to observations and error terms identical in pairs.

If we ignored this, our standard error calculations would be as if we had a sample of size 2n, when in fact we have only n samples.

Our estimated parameters would be the same for the 2n samples as for the n samples, but the confidence intervals would be narrower by a factor of √2



Plots of residuals from simulated time series data sets generated with differing levels of correlation ρ between error terms for adjacent time points.

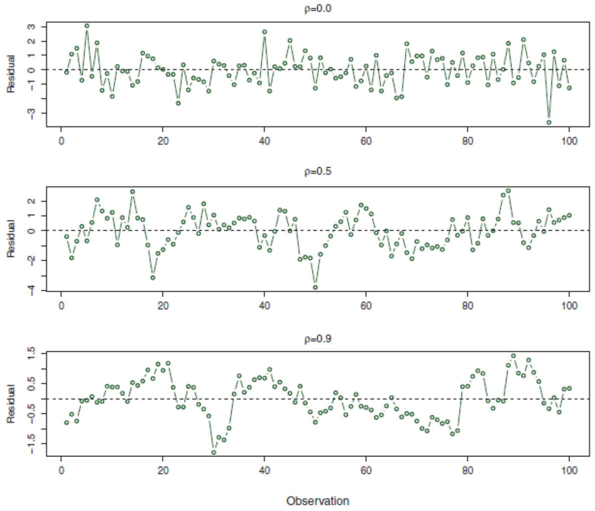
Potential problems of Linear Regression

2. *Correlation of error terms*

If the error is correlated, then the calculated standard errors will tend to underestimate the true standard errors.

As a result, confidence and prediction intervals will be narrower than they should be.
In addition, p-values associated with the model will be lower than they should be; this could cause us to erroneously conclude that a parameter is statistically significant.

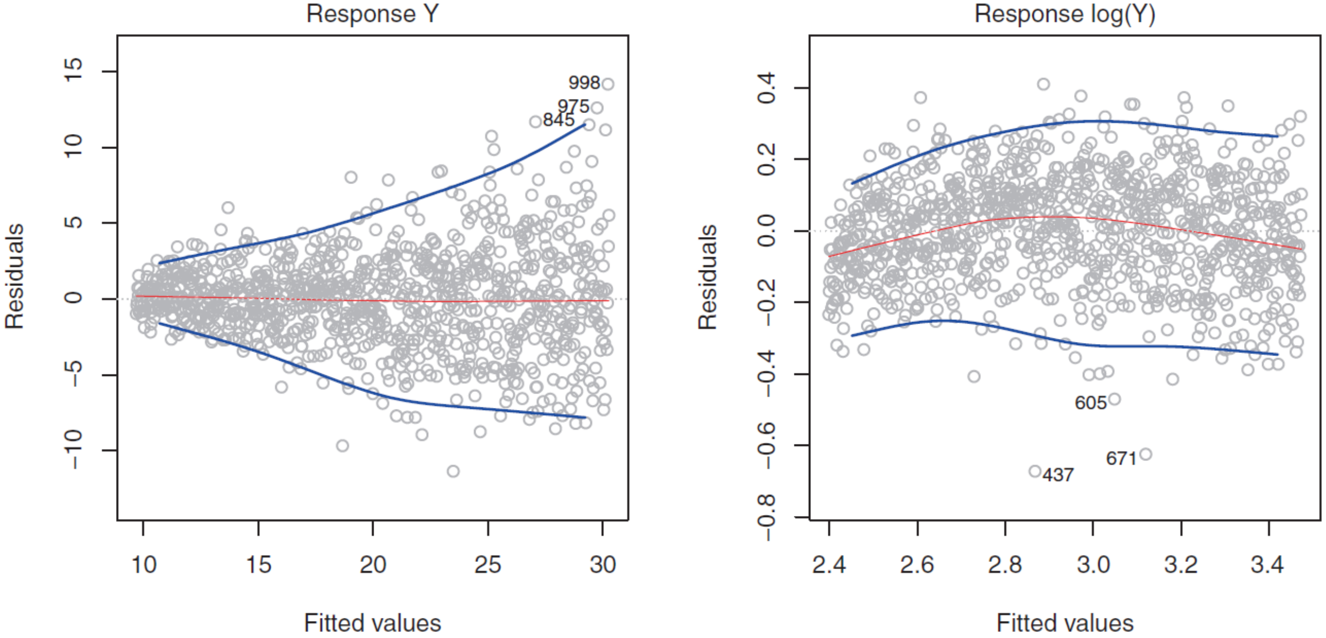
In short, if the error terms are correlated, we may have an unwarranted sense of confidence in our model.



Plots of residuals from simulated time series data sets generated with differing levels of correlation ρ between error terms for adjacent time points.

Potential problems of Linear Regression

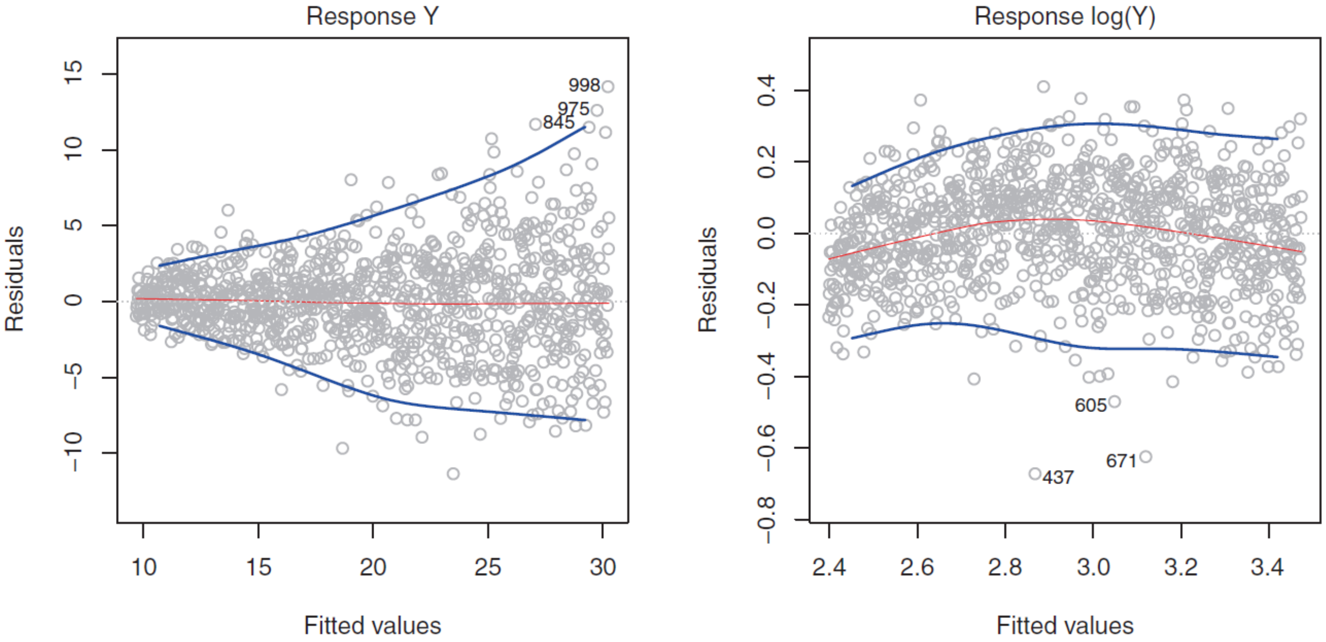
3. *Non-constant variance of error terms.*



Residual plots. In each plot, the red line is a smooth fit to the residuals, intended to make it easier to identify a trend. The blue lines track the outer quantiles of the residuals, and emphasize patterns. Left: *The funnel shape indicates heteroscedasticity.* Right: *The response has been log transformed, and there is now no evidence of heteroscedasticity*

Potential problems of Linear Regression

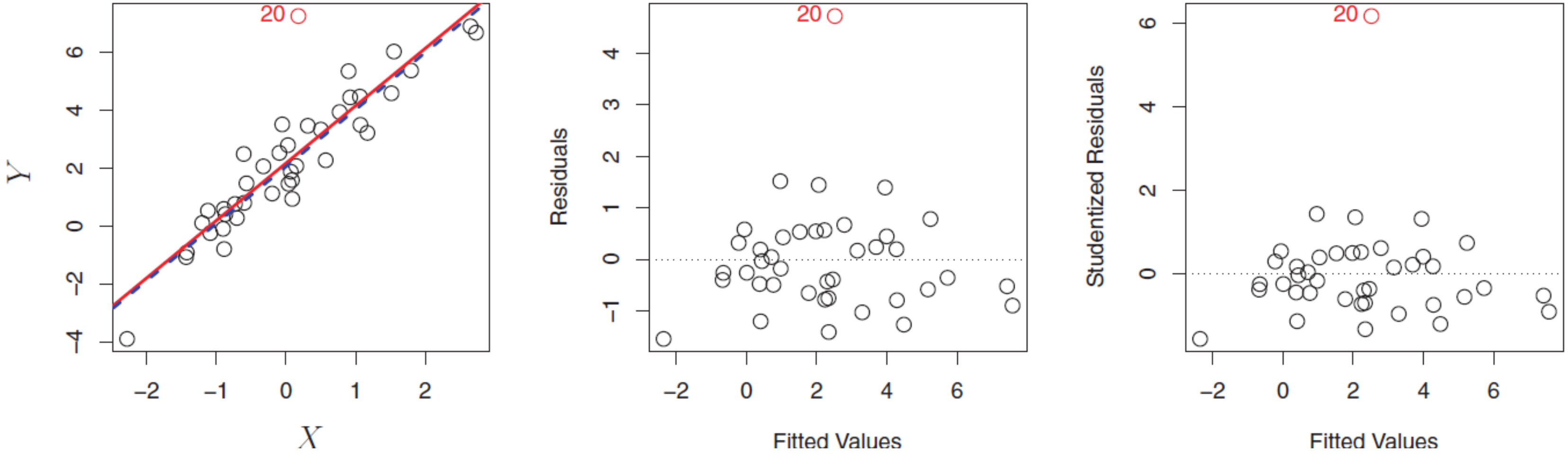
3. *Non-constant variance of error terms.*



- One possible solution is to transform the response Y using a concave function such as log Y or √Y
- Another remedy is to fit our model by weighted least squares, with weights proportional to the inverse weighted least squares variances

Potential problems of Linear Regression

4. *Outliers.*



Left: The least squares regression line is shown in red, and the regression line after removing the outlier is shown in blue. Center: *The residual plot clearly identifies the outlier.* Right: *The outlier has a studentized residual of* 6*; typically we expect values between −*3 *and* 3*.*

Robust Regression

- The average quadratic error functional (RSS) is very sensitive to outliers
- Robust error functionals aim to reduce the influence of outliers.
- Linear regression with robust error functionals is called robust linear regression.

Robust Regression

- One example of a robust error functional is the Huber function where the errors are only squared if they are smaller than a threshold ε > 0, otherwise they have only a linear impact

$$E_H = \sum_{k=1}^{n} \begin{cases} e_k^2 & \text{if } |e_k| < \epsilon \\ 2\epsilon \cdot |e_k| - \epsilon^2 & \text{otherwise} \end{cases}$$

Robust Regression

- Another example of a robust error functional is least trimmed squares which sorts the errors so that

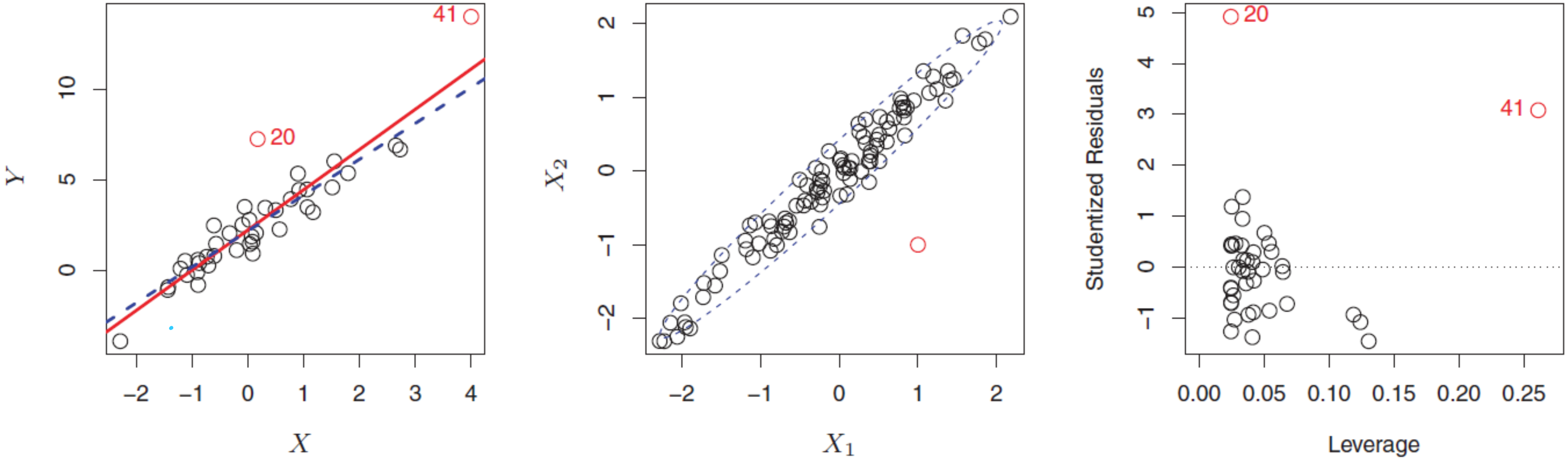
$$e'_1 \leq e'_2 \leq \ldots \leq e'_n$$

and only considers the m smallest errors, $1 \leq m \leq n.$

$$E_{LTS} = \sum_{k=1}^{m} e'^2_k$$

Potential problems of Linear Regression

5. *High-leverage points*



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 X1 value or its X2 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*

High-leverage points

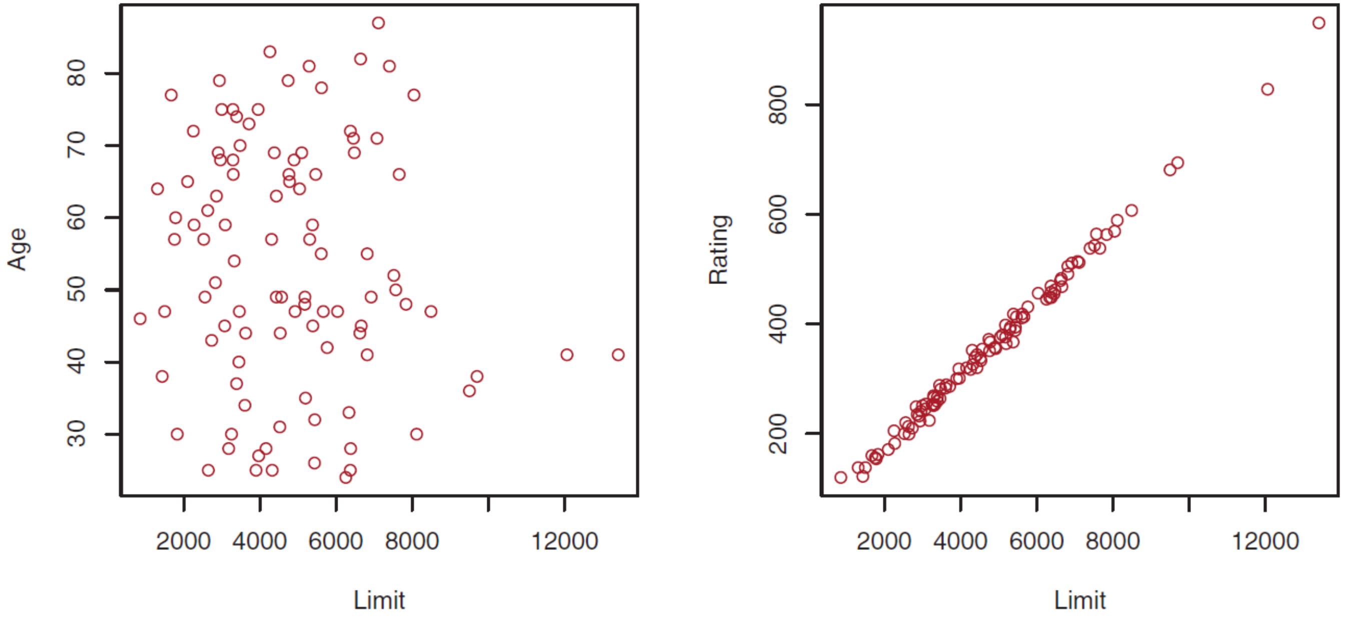
- In order to quantify an observation’s leverage, we compute the *leverage statistic*.
- A large value of this statistic indicates an observation with high leverage.
- For a simple linear regression,

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i'=1}^{n}(x_{i'} - \bar{x})^2}.$$

- The leverage statistic h_i is always between 1/*n* and 1
- The average leverage for all the observations is always equal to (*p*+1)/*n*.

Potential problems of Linear Regression

6. *Collinearity*



Scatterplots of the observations from the Credit *data set.* Left: *A plot of* age *versus* limit. *These two variables are not collinear.* Right: *A plot of* rating *versus* limit. *There is high collinearity.*