Topic 4. Linear Regression

Simple Linear Regression

- Data $(x_1, y_1), \dots, (x_n, y_n)$, where
  - $x_i \in R$ is the predictor (independent variable, input, feature)
  - $y_i \in R$ is the response (dependent variable, output, outcome)

- We denote the regression function as

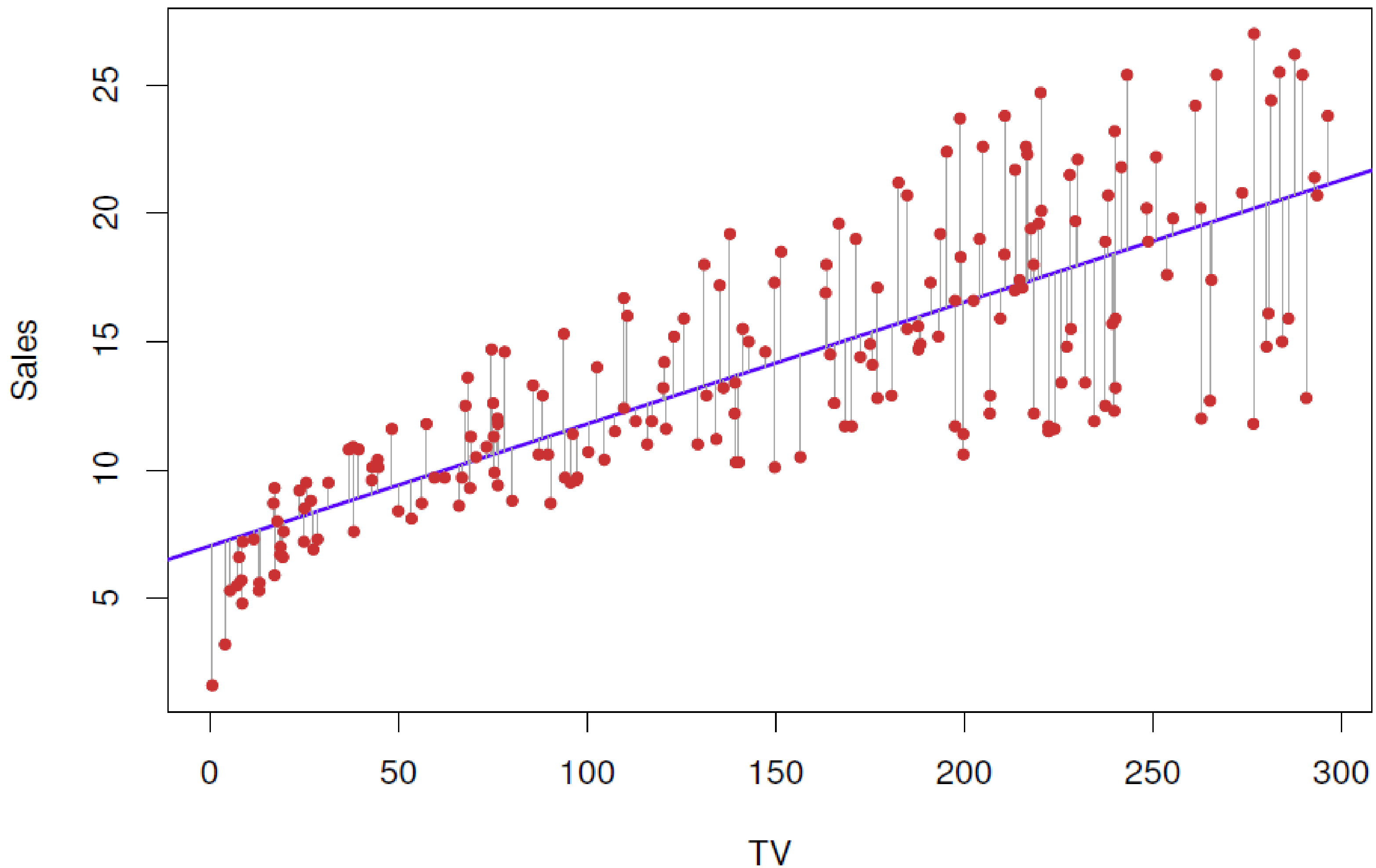
$$f(x) = E(Y|X = x)$$

- Linear regression model assumes that

$$f(x) = \beta_0 + \beta_1 x$$

which is usually viewed as an approximation to the truth.

A Toy Example



Least-squared Fitting

- Minimize the least square error

$$(\hat{\beta}_0, \hat{\beta}_1) = \arg\min_{\beta_0, \beta_1} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

- Solution is

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

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

- $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ are the fitted values.
- $e_i = y_i - \hat{y}_i$ are the residuals.

Point Estimation of β_0 and β_1

- Assume that

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

where ϵ_i's are iid from $N(0, \sigma^2)$.

- It can be shown that

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{\sum_i (x_i - \bar{x})^2}\right)$$

$$\hat{\beta}_0 \sim N\left(\beta_0, \left\{\frac{1}{n} + \frac{\bar{x}^2}{\sum_i (x_i - \bar{x})^2}\right\} \sigma^2\right)$$

Some Remarks

- We can think of $\hat{\beta}_0$ and $\hat{\beta}_1$ as functions of Y_i's, and thus they are also random variables and have distributions.

- The distributions of $\hat{\beta}_0$ and $\hat{\beta}_1$ refer to their different values that would be obtained with repeated sampling when X_i's are held constant from sample to sample.

- $\hat{\beta}_1$ has minimum variance among all unbiased linear estimators of the form: $b_1 = \sum_i c_i y_i$ (so called the BLUE estimator).

CI's for β_0 and β_1

- Note that σ^2 can be estimated by the unbiased MSE

$$\hat{\sigma}^2 = \frac{\sum_i (y_i - \hat{y}_i)^2}{n-2}, \qquad \frac{(n-2)\hat{\sigma}^2}{\sigma^2} \sim \chi^2(n-2)$$

- By replacing σ^2 with $\hat{\sigma}^2$, we have

$$s(\hat{\beta}_1) = \left(\frac{\hat{\sigma}^2}{\sum_i (x_i - \bar{x})^2}\right)^{1/2}$$

$$s(\hat{\beta}_0) = \left(\frac{\hat{\sigma}^2}{n} + \frac{\hat{\sigma}^2 \bar{x}^2}{\sum_i (x_i - \bar{x})^2}\right)^{1/2}$$

CI's for β_0 and β_1 (Cont.)

- Cochran's Theorem implies that $(\hat{\beta}_0, \hat{\beta}_1)$ and $\hat{\sigma}^2$ are independent, and thus

$$\frac{\hat{\beta}_1 - \beta_1}{s(\hat{\beta}_1)} \sim t_{n-2} \qquad \frac{\hat{\beta}_0 - \beta_0}{s(\hat{\beta}_0)} \sim t_{n-2}$$

- Then the CI's for $\hat{\beta}_0$ and $\hat{\beta}_1$ are

$$\hat{\beta}_1 \pm t\left(\frac{\alpha}{2}, n-2\right) s(\hat{\beta}_1)$$
$$\hat{\beta}_0 \pm t\left(\frac{\alpha}{2}, n-2\right) s(\hat{\beta}_0)$$

Hypothesis Test for β_0 and β_1

- To test $H_0: \beta_0 = 0 \;\; vs. \;\; H_1: \beta_0 \neq 0$, we have

$$t_0^* = \frac{\hat{\beta}_0 - 0}{s(\hat{\beta}_0)} \sim t_{n-2} \qquad \text{under } H_0$$

and we reject H_0 if $|t_0^*| > t\left(\frac{\alpha}{2}, n-2\right)$.

- To test $\beta_1 = 0 \;\; vs. \;\; H_1: \beta_1 \neq 0$, we have

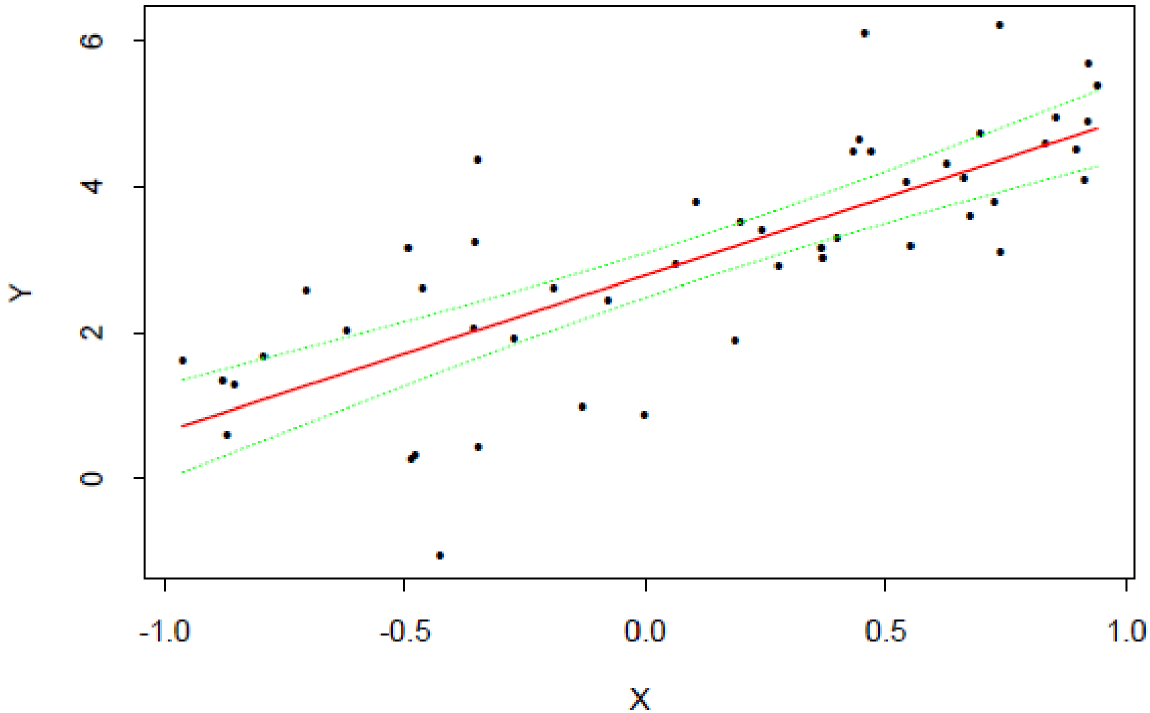
$$t_1^* = \frac{\hat{\beta}_1 - 0}{s(\hat{\beta}_1)} \sim t_{n-2} \qquad \text{under } H_0$$

and we reject H_0 if $|t_1^*| > t\left(\frac{\alpha}{2}, n-2\right)$.

Fitted Line

$$\hat{f}(x) = \hat{\beta}_0 + \hat{\beta}_1 x = \bar{y} + \hat{\beta}_1 (x - \bar{x})$$

$$s\left(\hat{f}(x)\right) = \left(\mathrm{var}(\bar{y}) + \mathrm{var}(\hat{\beta}_1)(x - \bar{x})^2\right)^{1/2} = \left(\frac{\hat{\sigma}^2}{n} + \frac{\hat{\sigma}^2 (x - \bar{x})^2}{\sum_i (x_i - \bar{x})^2}\right)^{1/2}$$



Multiple Linear Regression

- The regression model is $y_i = f(\mathbf{x}_i) + \epsilon_i$ with

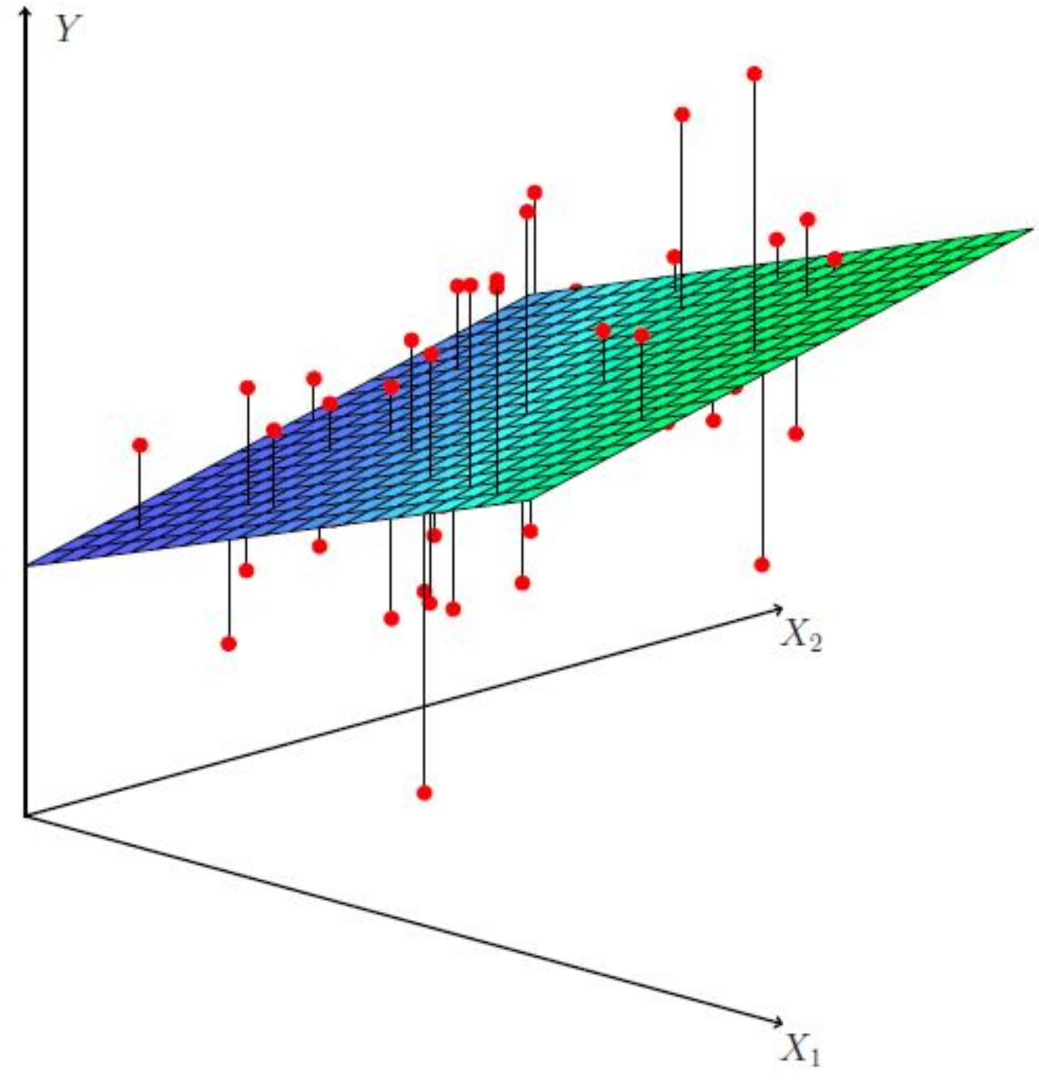
$$f(\mathbf{x}_i) = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}$$

or equivalently,

$$\mathbf{f} = \mathbf{X}\beta$$

- $\mathbf{f}$ is n-vector of fitted values.
- $\mathbf{X}$ is $n \times (p+1)$ matrix, with all ones in the first column.
- $\beta = \left(\beta_0, \dots, \beta_p\right)^T$ is a $(p+1)$-vector of parameters.

An Illustrative Plot



Least Squares Fitting

- Minimize the least square error,

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ji} \right)^2$$

$$= \underset{\beta}{\operatorname{argmin}} (\mathbf{y} - \mathbf{X}\beta)^T (\mathbf{y} - \mathbf{X}\beta)$$

- Solution is $\hat{\beta} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1} \mathbf{X}^T\mathbf{y}$ with

$$E\left(\hat{\beta}\right) = \beta \qquad \operatorname{cov}\left(\hat{\beta}\right) = \sigma^2\left(\mathbf{X}^T\mathbf{X}\right)^{-1}$$

- Fitted values are $\hat{\mathbf{y}} = \mathbf{X}\hat{\beta} = \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1} \mathbf{X}^T\mathbf{y} = \mathbf{H}\mathbf{y}$ with

$$E(\hat{\mathbf{y}}) = \mathbf{x}\beta \qquad \operatorname{cov}(\hat{\mathbf{y}}) = \sigma^2\mathbf{H}$$

Residual Properties

- Residuals are $\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} = (\mathbf{I} - \mathbf{H})\mathbf{y}$ with

$$E(\mathbf{e}) = 0 \qquad \operatorname{cov}(\mathbf{e}) = \sigma^2(\mathbf{I} - \mathbf{H})$$

- Furthermore,

$$\begin{aligned} E(\mathbf{e}^T\mathbf{e}) &= E\left(tr(\mathbf{e}^T\mathbf{e})\right) = E\left(tr(\mathbf{e}\mathbf{e}^T)\right) = tr\left(E(\mathbf{e}\mathbf{e}^T)\right) \\ &= tr\left(\sigma^2(\mathbf{I} - \mathbf{H})\right) = \sigma^2(n - p - 1) \end{aligned}$$

with $\quad tr(\mathbf{H}) = tr\left(\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\right) = tr(\mathbf{I}_{p+1}) = p + 1$

- σ^2 can be estimated by $\hat{\sigma}^2 = MSE = \frac{\mathbf{e}^T\mathbf{e}}{n-p-1}$

Analysis of Variance (ANOVA)

- The ANOVA decomposition is $SSTO = SSE + SSR$, where

$$SSTO = \mathbf{y}^T\left(\mathbf{I} - \frac{\mathbf{J}}{n}\right)\mathbf{y}$$

$\mathbf{J}$ is the matrix with all ones.

$$SSE = \mathbf{e}^T\mathbf{e} = \mathbf{y}^T(\mathbf{I} - \mathbf{H})\mathbf{y} = \mathbf{y}^T\mathbf{y} - \hat{\beta}^T\mathbf{X}^T\mathbf{y}$$

$$SSR = \mathbf{y}^T\left(\mathbf{H} - \frac{\mathbf{J}}{n}\right)\mathbf{y} = \mathbf{y}^T(\mathbf{I} - \mathbf{H})\mathbf{y} = \mathbf{y}^T\mathbf{y} - \hat{\beta}^T\mathbf{X}^T\mathbf{y}$$

- Note that $E(SSE) = \sigma^2(n - p - 1)$, and one can also show that

$$E(SSTO) = (n-1)\sigma^2 + \beta^T\mathbf{X}^T\left(\mathbf{I} - \frac{\mathbf{J}}{n}\right)\mathbf{X}\beta$$

$$E(SSR) = p\sigma^2 + \beta^T\mathbf{X}^T\left(\mathbf{I} - \frac{\mathbf{J}}{n}\right)\mathbf{X}\beta$$

R^2 for Regression

- The **coefficient of multiple determination** is

$$R^2 = \frac{SSR}{SSTO} = 1 - \frac{SSE}{SSTO}$$

which measures the proportionate reduction of total variation in Y associated with the use of X.

- R^2 is not a good measure for comparing different models, as it always increases with more variables in the model.

- The **adjusted** R^2 accounts for the effects of multiple predictors

$$R_a^2 = 1 - \frac{\frac{SSE}{n-p-1}}{\frac{SSTO}{n-1}} = 1 - \frac{n-1}{n-p-1}\frac{SSE}{SSTO}$$

Test for Linear Model

- To test

$$H_0\colon\ \beta_1 = \beta_2 = \cdots = \beta_p = 0$$

$$H_a\colon \text{not all } \beta_k = 0 \ (k \geq 1)$$

- We often use the F-test,

$$F^* = \frac{MSR}{MSE} = \frac{SSR/p}{SSE/(n-p-1)}$$

- Decision: reject H_0 if $F^* > F(\alpha; p, n-p-1)$.

Test for Coefficients

- The covariance matrix for $\hat{\beta}$ is

$$\text{cov}(\hat{\beta}) = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$$

- The estimated covariance matrix of $\hat{\beta}$ is

$$s^2(\hat{\beta}) = MSE(\mathbf{X}^T\mathbf{X})^{-1}$$

- Under normal error assumption, we have

$$\frac{\hat{\beta}_k - \beta_k}{s(\hat{\beta}_k)} \sim t(n-p-1); k = 0, \ldots, p$$

where $s(\hat{\beta}_k)$ is the corresponding diagonal element of $s(\hat{\beta})$.

Test for Coefficients (Cont.)

- So the $100(1-\alpha)\%$ CI for β_k is

$$\hat{\beta}_k \pm t\left(\frac{\alpha}{2}, n-p-1\right) s(\hat{\beta}_k)$$

- For hypothesis test

$$H_0\colon\ \beta_k = 0, H_a\colon\ \beta_k \neq 0$$

we can use the test statistic $t^* = \hat{\beta}_k / s(\hat{\beta}_k)$.

- Decision: reject H_0 if $|t^*| > t(\alpha/2, n-p-1)$.

Model Diagnostics

- Recall the normal error assumption model,

$$y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \epsilon_i \qquad\qquad i = 1, \dots, n$$

  where

  - $\beta_0, \dots, \beta_p$ are parameters.
  - x_i's are treated as fixed constants.
  - ϵ_i's are independent from $N(0, \sigma^2)$.

- Can the model be inadequate? If yes, in what aspect? And how to remedy the model?

Potential Issues

- The regression function is not linear.
- Other important predictors are missed from the model.
- ϵ's have non-constant variance.
- ϵ's are not independent.
- ϵ's are not normally distributed.
- The model fits all but few outlier observations.
- The predictors are correlated.

Residual Properties

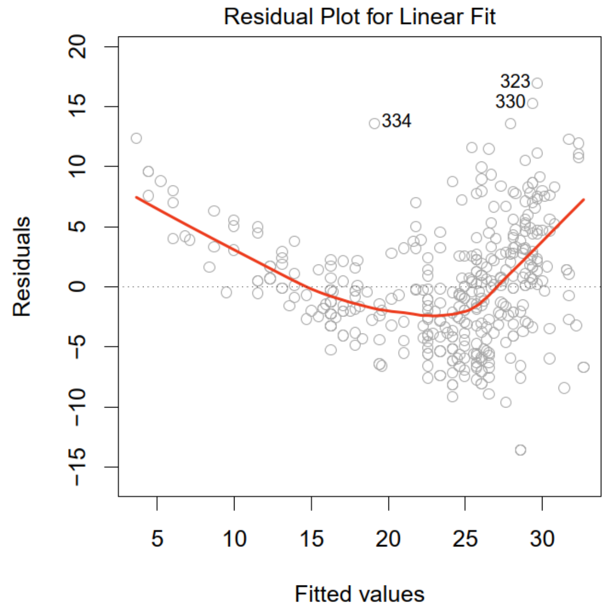
- Most of our diagnostics concern the distribution of ϵ_i's, which are estimated by the residuals $e_i = y_i - \hat{y}_i$.
- Note that e_i's are linear functions of y_i's, and it can be shown that $\mathbf{e} \sim N\left(0, \sigma^2(\mathbf{I} - \mathbf{H})\right)$.
- Even though ϵ_i's are independent, e_i's are not. If n is big, $\operatorname{cov}(e_i, e_j)$ is approximately zero, and thus e_i's can be treated as approximately independent.
- Sometimes, it's useful to standardize the residuals, such as

$$e_i^* = \frac{e_i - \bar{e}}{\sqrt{MSE}} = \frac{e_i}{\sqrt{MSE}}$$

the **semi-studentized** residuals.

Nonlinearity of Regression Function

- We want to detect some nonlinear patterns in the data. We can visually inspect the scatter plot of $(y_i, \hat{y}_i)$ or $(e_i, \hat{y}_i)$. For detecting nonlinear patterns, these two scatter plots are equivalent to each other since e_i's are linear functions of y_i's.



- Ideally, e_i will remove the linear tendency in the original data, and make the detection of nonlinear pattern in the scatter plot of $(e_i, \hat{y}_i)$ much easier.

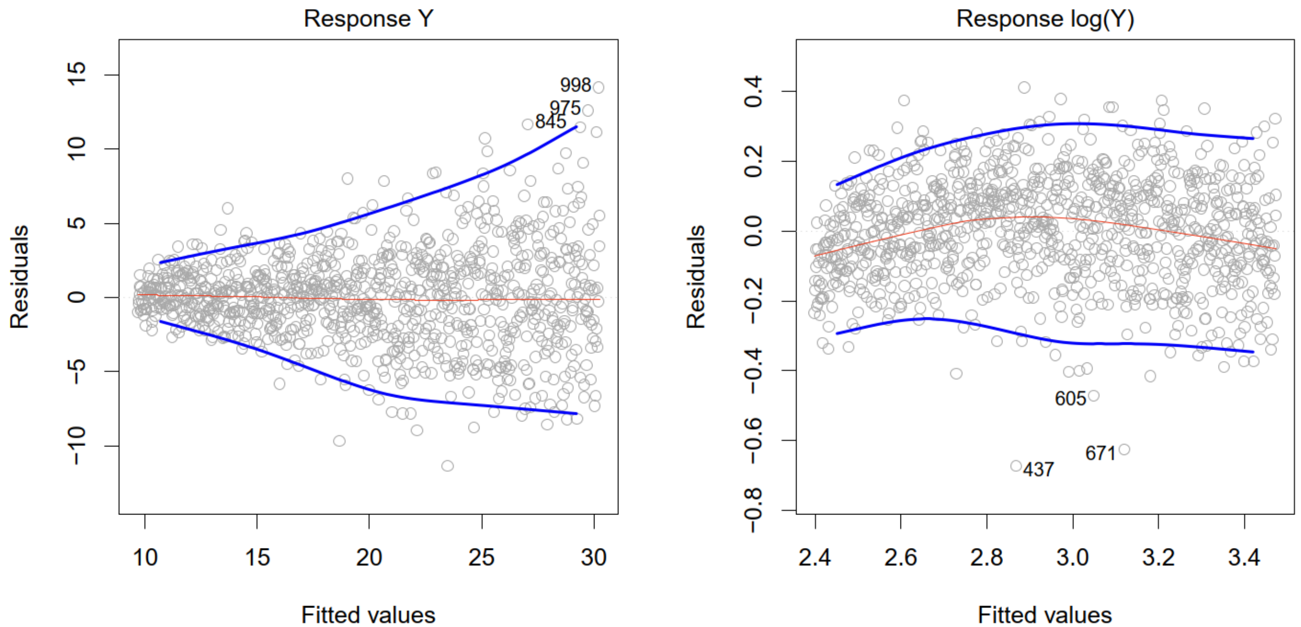
Omission of Important Predictor Variables

- This can be visually inspected by plotting e_i's versus other predictor variables. If we detect any pattern between them, the variables shall then be included in the model to improve the estimation.

- When multiple predictor variables are available, variable selection, deciding which variables to be included in the model, is an old and still very active research field.

Non-constancy of Error Variance (Heteroscedasticity)

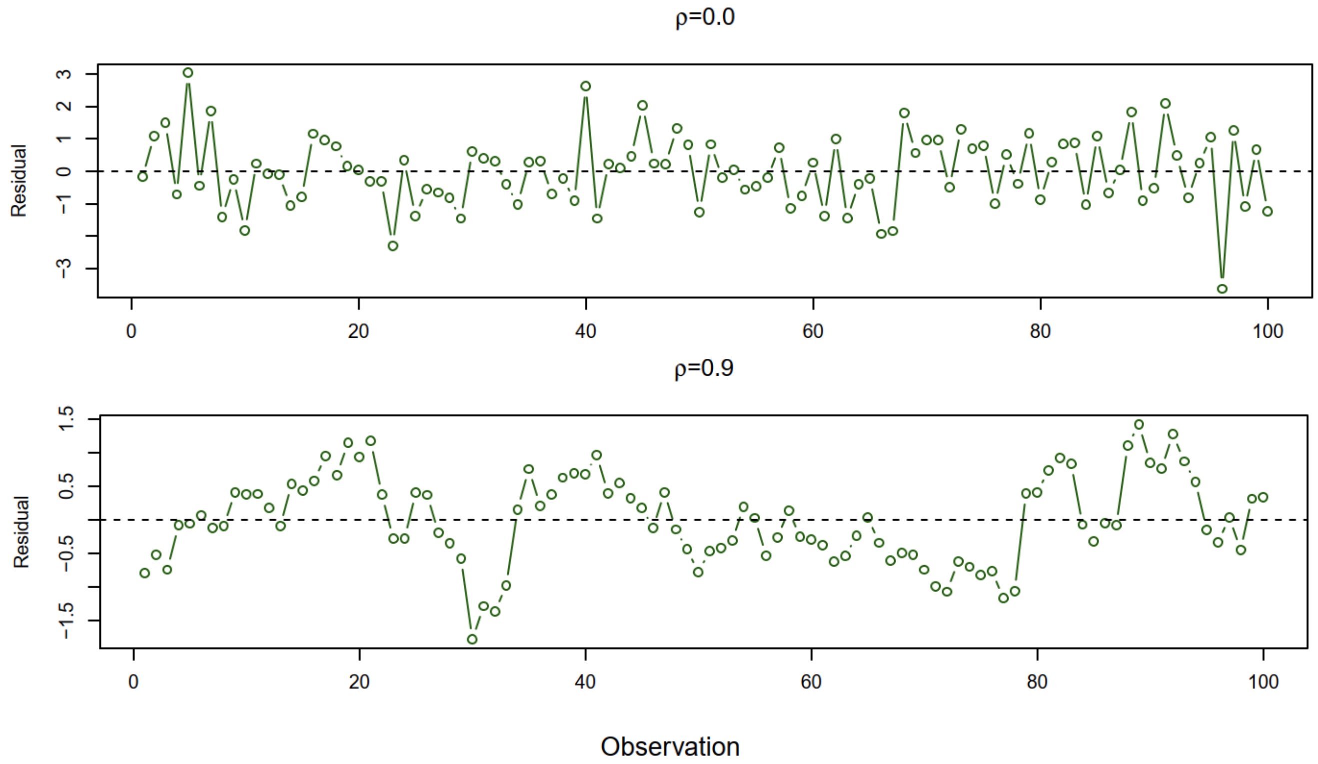
- We can inspect the scatter plot of $(e_i, \hat{y}_i)$. If the error variance is indeed constant, all residuals e_i's should have roughly the same magnitude.



- Since the signs of e_i's are not much meaningful for examining the constant error variance, it is often useful to plot $|e_i|$ against $\hat{y}_i$.

Dependence of Error Terms

- In time series or spatial data, it is always useful to inspect the scatter plot of e_i's versus time or geographical locations.
- The goal is to see if there is any correlation between e_i's that are near each other in the sequence.

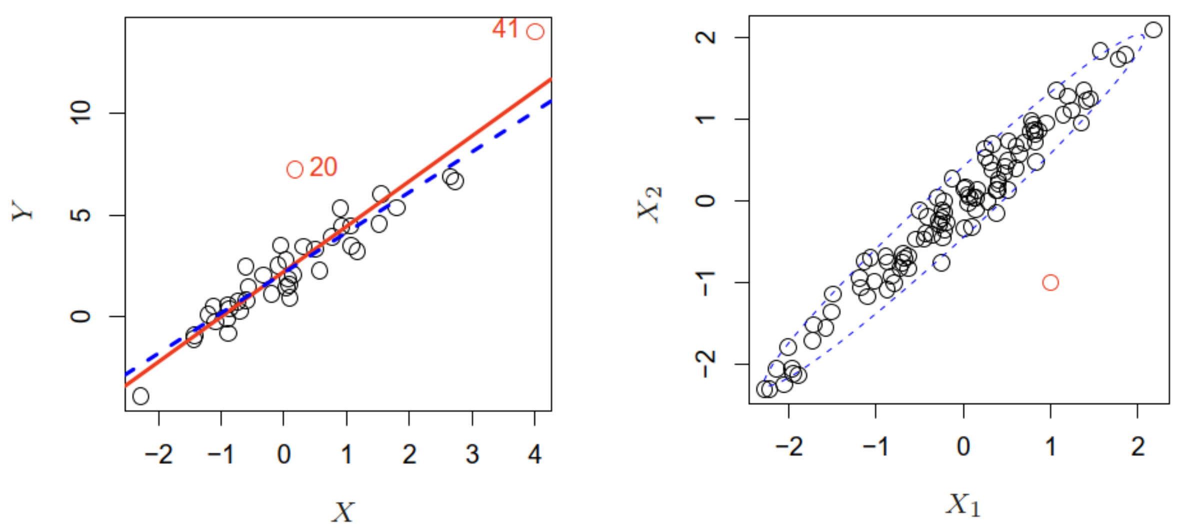


Non-normality of Error Term

- Several ways to inspect the normality of e_i's:
  - Distribution plot: histogram, boxplot, stem-and-leaf for inspecting distribution
  - Cumulative distribution function estimation comparison: sample frequency estimates the cumulative distribution function, which can be compared to the theoretical values.
  - QQ plot: a plot that is nearly linear suggests agreement with normality, whereas a plot that departs substantially from linearity suggests that the errors are not normally distributed.

Outlying Observations

- When some observations are well separated from the majority of the data, these cases are called **outlying**. A case may be outlying with respect to its *Y* value, its *X* value(s), or both.
- **Outlying *Y* observations** (outliers): data point for which y_i is far from the value predicted by the model
- **Outlying *X* observations** (high leverage points): data points that have an unusual *X* value



Outlying *Y* Observations

- Outliers can arise for a variety of reasons such as measurement error and recording error.

- Residual plots can be used to identify outliers. There are different definitions of residuals used to detect outliers.

- Residuals and semi-studentized residuals

$$e_i = y_i - \hat{y}_i \qquad e_i^* = \frac{e_i}{\sqrt{MSE}}$$

- Studentized residuals

$$r_i = \frac{e_i}{s(e_i)} = \frac{e_i}{\sqrt{MSE(1 - h_{ii})}}$$

Outlying Y Observations (Cont.)

- **Deleted residual** is the difference between observations and its fitted value from a regression model without using the observation

$$d_i = y_i - \hat{y}_{i(-i)}$$

where $\hat{y}_{i(-i)}$ is the estimate of $E(Y|X = x_i)$ from a model fitting without using (x_i, y_i).

- We can show that

$$d_i = \frac{e_i}{1 - h_{ii}}$$

Outlying Y Observations (Cont.)

- The deleted residual mimics the prediction error for a new observation.

- Its estimated variance can be obtained as

$$s^2(d_i) = MSE_{(-i)}\left(1 + x_i'\left(\mathbf{X}_{(-i)}'\mathbf{X}_{(-i)}\right)^{-1} x_i\right)$$

where $MSE_{(-i)}$ is computed when the (x_i, y_i) is omitted in the fitting, and $\mathbf{X}_{(-i)}$ is the $\mathbf{X}$ matrix without the x_i row.

- It can be shown that

$$s^2(d_i) = \frac{MSE_{(-i)}}{1 - h_{ii}}$$

Outlying Y Observations (Cont.)

- The **studentized deleted residual** is defined as

$$t_i = \frac{d_i}{s(d_i)} = \frac{e_i}{\sqrt{MSE_{(-i)}(1-h_{ii})}} \sim t_{n-p-2}$$

- We can show that

$$(n-p-1)MSE = (n-p-2)MSE_{(-i)} + \frac{e_i^2}{1-h_{ii}}$$

So

$$t_i = e_i \left\{\frac{n-p-2}{SSE(1-h_{ii})-e_i^2}\right\}^{1/2}$$

- We can formally test for outlying Y observations by comparing $|t_i|$ with $t(1-\frac{\alpha}{2n}, n-p-2)$, which adjusts for the n observations following Bonferroni procedure.

Outlying *X* Observations

- The diagonal elements of **H** are useful indicators of whether an observation is outlying with respect to its *X* values.
- We have $0 \leq h_{ii} \leq 1$ and $\sum_{i=1}^{n} h_{ii} = tr(\mathbf{H}) = p + 1$.
- h_{ii} is called the **leverage** of the ith case, and measures the distance between x_i and the center of all the *X* values.
- Thus a large h_{ii} indicates that x_i is distant from the center of all the *X* values.
- In general, a leverage value greater than $2(p+1)/n$ is considered to indicating outlying observation with regard to its *X* values.

Multicollinearity (Collinearity)

- Ideally, predictor variables in multiple regression are independent of each other (called “independent variables” in statistics).
- If high correlations among predictor variables are present, this is called multicollinearity.
- Examples:
  - $Y \sim X_1(weight) + X_2(BMI)$ + others
  - $Y \sim X_1(credit\ rating) + X_2(credit\ limit)$ + others
- Serious problems may occur when multicollinearity exists.

Effects of Multicollinearity

- Variance of regression coefficient estimation may become very large.
- Regression coefficients may change signs after deleting one variable.
- The marginal significance of a predictor in a multiple regression highly depends on which other predictors are included in the model.
- The significance of a predictor may be masked by correlated variables in the model.

Variance Inflation Factor (VIF)

- We define the variance inflation factor (VIF) as

$$(VIF)_j = \left(1 - R_j^2\right)^{-1}$$

where R_j^2 is the coefficient of multiple determination when the jth variable is regressed against the other $p-1$ variables in the model.

- If the largest VIF value exceeds 10, then we considered that multicollinearity unduly influences the least squares estimates.

- If the average of all $(VIF)_j$'s is considerably larger than 1, it is also an indication of serious multicollinearity.

Variable Transformation

- Variable transformations can be used to linearize the nonlinear regression function, stabilize error variances, and even normalize the error terms.

- Box-Cox transformation uses y^{λ} with $\lambda \geq 0$ as the response where y^{0} is defined as $\ln(Y)$.

- The selection of optimal λ is based on maximizing the likelihood function

$$L(\lambda; \beta_0, \beta_1, \sigma^2) = \frac{1}{(2\pi\sigma)^{n/2}} exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i^{\lambda} - \beta_0 - \beta_1^T \mathbf{x}_i\right)^2\right)$$

Bias-Variance Tradeoff

- Let $f_0(\mathbf{x})$ be the true regression function at **x**, then a good measure of the quality of $\hat{f}(\mathbf{x})$ is the mean squared error

$$MSE\left(\hat{f}(\mathbf{x})\right) = E\left(\hat{f}(\mathbf{x}) - f_0(\mathbf{x})\right)^2$$

- This can be written as

$$MSE\left(\hat{f}(\mathbf{x})\right) = \mathrm{var}\left(\hat{f}(\mathbf{x})\right) + \left(E(\hat{f}(\mathbf{x})) - f_0(\mathbf{x})\right)^2$$

- Typically, when bias is low, variance will be high and vice-versa; and thus choosing estimators often involves a tradeoff between bias and variance.

Bias-variance Tradeoff (Cont.)

- If the linear model is correct for a given problem, then the least square estimate $\hat{f}$ is unbiased, and has the lowest variance among all unbiased estimators that are linear functions of $\mathbf{y}$ (Gauss-Markov Theorem).
- But there can be biased estimators with smaller MSE
  - Generally, by regularizing the estimator in some way, its variance will be reduced. If the corresponding increase in bias is small, this will be worthwhile.
  - Examples of regularization: subset selection (forward, backward, all subsets), ridge regression, lasso
- In reality, models are almost never correct, so there will be an additional model bias between the "best" linear model and the true regression function.

Qualitative Predictors

- Consider a regression model with one quantitative predictor X_1 and a qualitative predictor with two levels M_1 and M_2.
- We can define a dummy variable

$$X_2 = \begin{cases} 1 & \text{if level } M_1 \\ 0 & \text{if level } M_2 \end{cases}$$

- Then we have the following regression model

$$E(Y|X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

Qualitative Predictors (Cont.)

- The model implies that

$$E(Y|X) = \begin{cases} \beta_0 + \beta_1 X_1 + \beta_2 & \text{if level } M_1 \\ \beta_0 + \beta_1 X_1 & \text{if level } M_2 \end{cases}$$

- So it basically assumes different intercepts but the same slope for two levels (parallel lines), with

$$\beta_2 = E(Y|X_2 = 1) - E(Y|X_2 = 0) = E(Y|M_1) - E(Y|M_2)$$

- Hence β_2 indicates the average difference in the mean response between the two levels.

Interactions Effects

- We can also consider the interaction effect between X_1 and X_2 in the model

$$E(Y|X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2$$

- For this model we have

$$E(Y|X) = \begin{cases} \beta_0 + \beta_2 + (\beta_1 + \beta_3) X_1 & \text{if level } M_1 \\ \beta_0 + \beta_1 X_1 & \text{if level } M_2 \end{cases}$$

- So it assumes different intercepts and slopes for these two levels (nonparallel lines), and β_2 and β_3 are the intercept and slope differences between the two levels.

Further Remarks

- Can have more than one qualitative predictors
- Can have more than two levels
- Imagine a qualitative variable with 5 levels, how to code this variable?
  - Code it as 1, 2, 3, 4, and 5
  - Define X_1, X_2, X_3, and X_4, with $X_j = 1$ if level j and 0 otherwise for $j = 1, 2, 3, 4$
  - Define X_1, X_2, X_3, and X_4, with $X_j = 1$ if level j and -1 otherwise for $j = 1, 2, 3, 4$