Lecture 10.1

Diagnostics

Linear Model Assumptions

Recall that we assumed $Y_i \overset{\text{indep}}{\sim} \mathcal{N}(\beta_0^\star + \beta_1^\star x_i, \sigma^2)$, i.e.,

- **Independence:** Y_i's are independent.
- **Normality:** Y_i's have normal distribution.
- **Linearity:** $\mathbb{E}(Y_i)$ is a linear function of x_i.
- **Common Variance:** $\text{Var}(Y_i)$ does not depend on x or i.

The conclusions we make using the linear regression model are only appropriate if the assumptions are satisfied.

How can we check the validity of our assumptions?

Linear Model Assumptions

These assumptions are equivalent to

$$Y_i - (\beta_0^\star + \beta_1^\star x_i) = \varepsilon_i \overset{\text{indep}}{\sim} \mathcal{N}(0, \sigma^2).$$

- **Independence:** ε_i's are independent.
- **Normality:** ε_i's have normal distribution.
- **Linearity:** $\mathbb{E}(\varepsilon_i) = 0$.
- **Common Variance:** $\text{Var}(\varepsilon_i)$ does not depend on x or i.

We do not know ε_i's, but we can estimate them by the residuals as

$$e_i = y_i - (\underbrace{\hat{\beta}_0 + \hat{\beta}_1 x_i}_{\hat{y}_i}), \quad i = 1, 2, \ldots, n.$$

So, we can analyze the residuals to see whether most (if not all) underlying assumptions of the linear regression model are verified.

Residuals

The residual e_i is an observation of the underlying random variable E_i. Since

$$\mathbb{E}(E_i) = \mathbb{E}(Y_i - (\overbrace{\hat{\beta}_0 + \hat{\beta}_1 x_i}^{\hat{Y}_i})) = \mathbb{E}(Y_i) - \mathbb{E}(\hat{\beta}_0 + \hat{\beta}_1 x_i)$$
$$= (\beta_0^\star + \beta_1^\star x_i) - (\beta_0^\star + \beta_1^\star x_i) = 0,$$

and it can be shown that

$$\text{Var}(Y_i - \hat{Y}_i) = \sigma^2 \left(1 - \frac{1}{n} - \frac{(x_i - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right),$$

then we can also look at the **standardized residuals** as

$$\tilde{e}_i = \frac{y_i - \hat{y}_i}{\hat{\sigma}\sqrt{1 - \frac{1}{n} - \frac{(x_i - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}}.$$

Linearity

We can check the linearity assumption by plotting the residuals e_i (or $\tilde{e}_i$) against the explanatory variable(s), x_i, or against the predicted values $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$. In simple regression with single predictor, both work equivalently since $\hat{y}_i$ is just scaled and shifted version of x_i.

If the linearity assumption is holds, the residuals should evenly spread above and below zero. Indeed, under the linearity assumption, $\mathbb{E}(\varepsilon_i) = 0$, and so we expect that a flat line through $e = 0$ should pass through the middle of the residuals $\{e_i\}_{i=1}^n$ all the way along.

Normality

We can plot the residuals e_i (or $\tilde{e}_i$) against the quantiles of the standard normal distribution to check the normality assumption.

Under the normality assumption the points should lie approximately on the straight line.

One could also look for normality on a histogram or density plot of the residuals.

Constant variance

We can use the plot of the residuals e_i (or $\tilde{e}_i$) against the explanatory variable(s) or against the predicted values $\hat{y}$ to check the constant variance assumption.

If the constant variance assumption holds, since $\text{Var}(\varepsilon_i) = \sigma^2$, then the spread of the residuals $\{e_i\}_{i=1}^n$ should remain constant through the plot.

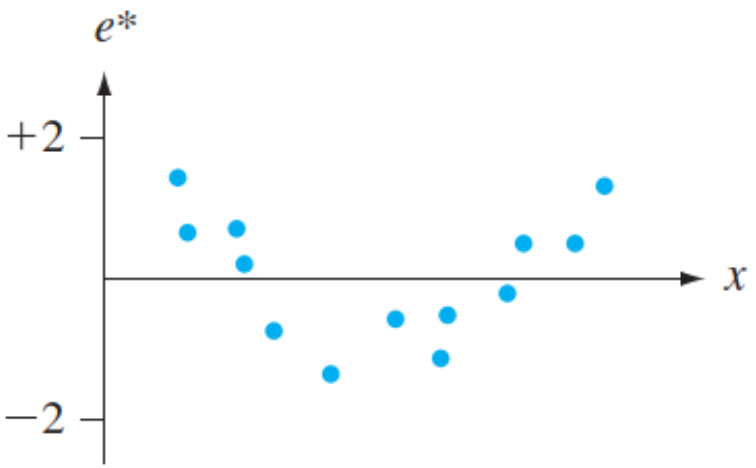
Independence

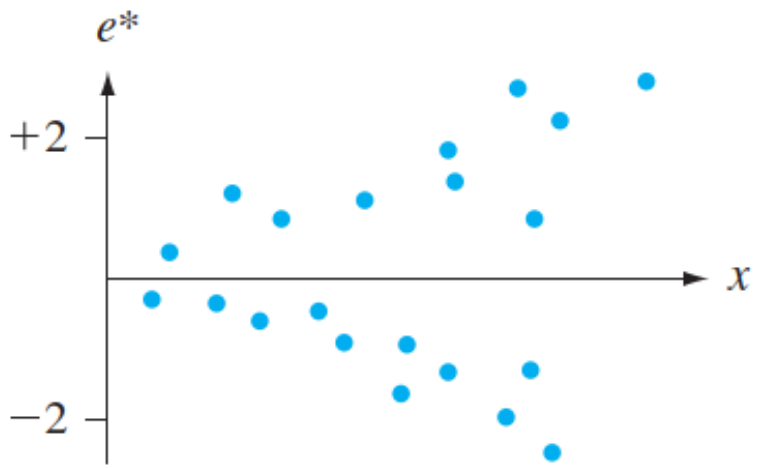
The assumption of independence is difficult to check from data if you don't have any suspicions of how this assumption might be violated.

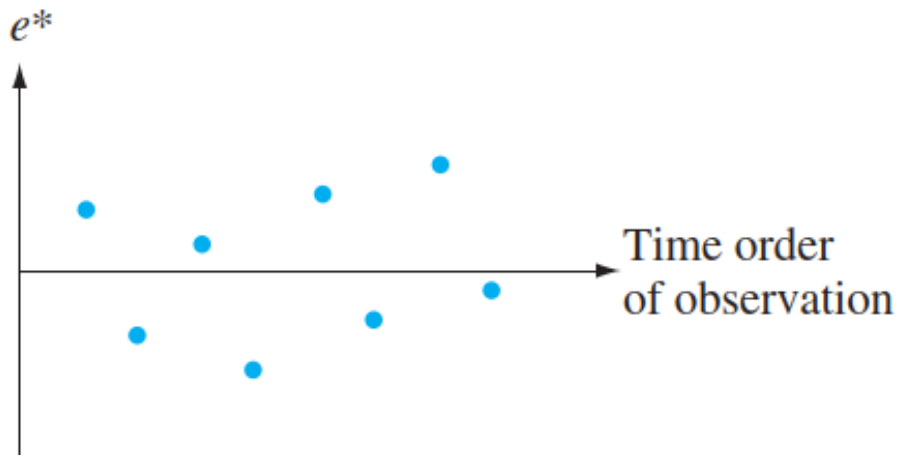
For example, if the data is collected over time, it can display temporal dependence and we can look for that in the plots. This can be checked by plotting the residuals against time. When the residuals are independent, they should be uncorrelated with each other. The residuals should be found above and below the horizontal axis randomly, without any indication that nearby residuals have similar values.

We aim to ensure independence through appropriate sampling.

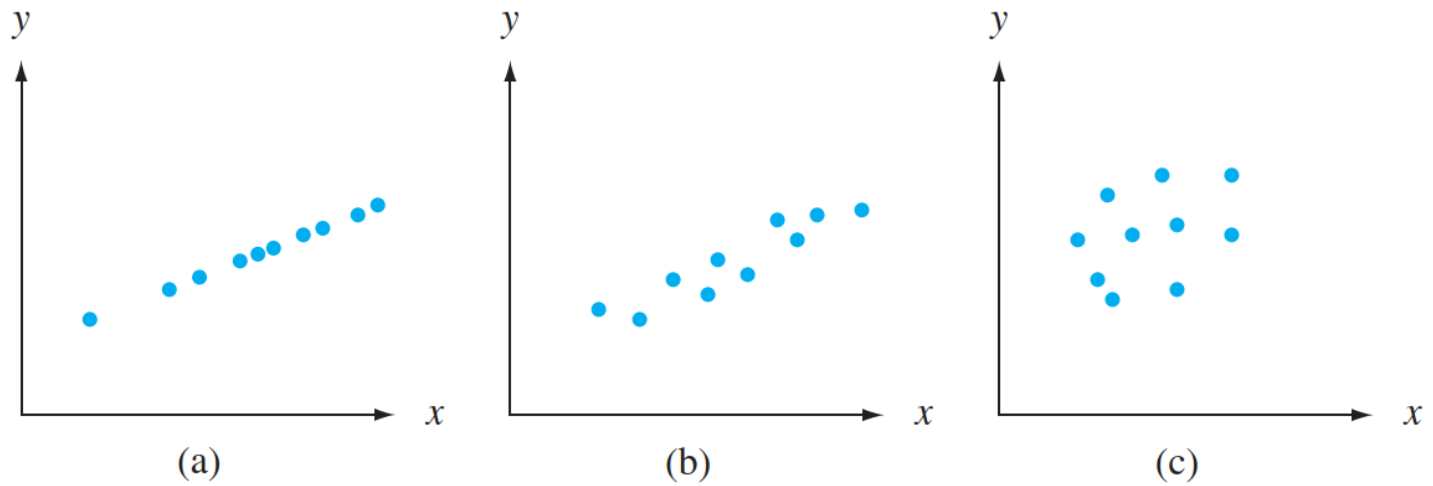
Residuals







Coefficient of Determination



Without considering x, there is similar variability in observed y values in all plots. However, the variability of y with respect to x is not similar. The points in plot (a) all fall exactly on a straight line. Although the points on plot (b) do not fall exactly on a line, but compared to overall y variability, the deviations from the least squares line are small. In plot (c), the simple linear regression model fails to explain variation in y by relating y to x.

Coefficient of Determination

The **sum of squared errors (SSE)** is a measure of how much variation in y is left unexplained by the linear model, i.e., how much cannot be attributed to a linear relationship. But the actual value of SSE is not unit-less, which can make it hard to interpret. **Coefficient of determination** is a unit-less measure to evaluate how well the simple linear regression model can explain the observations.

The coefficient of determination is defined as R^2 =1-SSE/SST where

$$\text{SST} = \sum_{i=1}^{n} (y_i - \bar{y})^2,$$

is the total sum of squares (SST), i.e., SST is the sum of squared deviations about the horizontal line $y = \bar{y}$.

Coefficient of Determination

Clearly, SSE $\leq$ SST (why?).

The ratio SSE/SST is always between 0 and 1 and is the proportion of total variation that cannot be explained by the simple linear regression model.

The coefficient of determination, R^2 is thus the proportion of the variation in the observed y that is explained by the model.

For simple linear regression, R^2 is equal to the square of the sample correlation between X and Y.