Diagnostics

- Assumption: $\mathbf{y} \sim \mathsf{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n)$.

- Error: assumed to be iid $\sim \mathsf{N}(0, \sigma^2)$.

- Model: assumed to be linear, i.e., $\mathbb{E}\mathbf{y} = \mathbf{X}\boldsymbol{\beta}$.

- Unusual observations

- We will use both graphical and numerical tools for diagnosis.

Diagnostics: Finding Unusual Observations

Why we discuss unusual observations first?

Least squares regression is very sensitive to individual data points. (Yes, this is why we need to discuss robust regression procedures later.)

It is possible the inference, p-values, parameter estimation, CI's are all driven by a single data point.

Types of Unusual Observations

- High leverage points: We'll define some measure called "leverage" which quantifies how far a data point is from the center of the whole sample. Points with a large value of leverage are flagged as the high leverage points. High leverage points could be "good" or "bad".
- Outliers: data point that does not fit the model as the other data points. We will introduce a formal testing procedure to identify outliers.

- High influential points: How does each individual observation affect the estimation of the model?

  Sometimes, the estimated parameters and other related statistics (such as R^2) depend heavily on one observation, in the sense that if that observation were removed, the result of the analysis would change.

  We will define some measure, “Cook’s distance”, to quantify the aforementioned change for each data point and data points with large value of Cook’s distance are called high influential points.

Leverages

- The diagonal elements of $\mathbf{H}$,

  $$h_i = H_{ii},$$

  are called leverages and are very useful diagnostics.

- h_i gives a measure (invariant under any affine transformation of X) of how far the i-th observation is from the center of the data (in the X-space). This measure also arises in our discussion on the width of CI and standard error of prediction/estimation at $\mathbf{x}_i$.

- For simple regression,

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_i (x_i - \bar{x})^2}.$$

- In general

$$\begin{aligned} h_i &= \mathbf{x}_i^t (\mathbf{X}^t \mathbf{X})^{-1} \mathbf{x}_i \\ &= \frac{1}{n} + \frac{1}{n-1} (\mathbf{z}_i - \bar{\mathbf{z}})^t \hat{\Sigma}^{-1} (\mathbf{z}_i - \bar{\mathbf{z}}) \qquad (1) \end{aligned}$$

where $\hat{\Sigma}_{(p-1)\times(p-1)} = \frac{1}{n-1} \sum_{i=1}^n (\mathbf{z}_i - \bar{z})(\mathbf{z}_i - \bar{z})^t$ is the sample covariance of the $(p-1)$ predictor variables. The 2nd term in the right hand side of (1) is the so-called *Mahalanobis distance* from $\mathbf{z}_i$ to the data center $\bar{\mathbf{z}}$.

The following two properties of $\mathbf{H}$

$$\text{tr}(\mathbf{H}) = p, \quad \mathbf{H} = \mathbf{H}\mathbf{H}^T$$

imply that $\sum_i h_i = p$ and $\sum_j H_{ij}^2 = h_i$.

$$\sum_j H_{ij}^2 = H_{ii}^2 + \sum_{j \neq i} H_{ij}^2 = h_i^2 + \sum_{j \neq i} H_{ij}^2 = h_i$$

$$\Longrightarrow \sum_{j \neq i} H_{ij}^2 = h_i(1 - h_i) \Longrightarrow h_i(1 - h_i) > 0$$

Properties of h_i:

$$0 < h_i < 1, \quad \sum_i h_i = p.$$

Recall $\hat{\mathbf{y}} = \mathbf{H}\mathbf{y}$, that is,

$$\begin{pmatrix} \hat{y}_1 \\ \ldots \\ \hat{y}_i \\ \ldots \\ \hat{y}_n \end{pmatrix} = \begin{pmatrix} H_{i1} & \cdots & H_{n1} \\ \ldots & \ldots & \ldots \\ H_{i1} & \ldots & H_{in} \\ \ldots & \ldots & \ldots \\ H_{n1} & \cdots & H_{nn} \end{pmatrix} \begin{pmatrix} y_1 \\ \ldots \\ y_i \\ \ldots \\ y_n \end{pmatrix}$$

$$\begin{aligned} \hat{y}_i &= H_{11}y_1 + \cdots + H_{ii}y_i + \cdots + H_{in}y_n \\ &= H_{11}y_1 + \cdots + h_i y_i + \cdots + H_{in}y_n \end{aligned}$$

- The LS fit $\hat{y}_i$ is a linear combination of the n data points,

$$\hat{y}_i = h_i y_i + \sum_{j \neq i} H_{ij} y_j, \qquad \text{i.e. } h_i = \frac{d\hat{y}_i}{dy_i}$$

- When h_i is large (close to 1), $\hat{y}_i$ relies heavily on y_i (instead of using the information from other data points), therefore $\hat{y}_i$ will be "forced" to be close to the observed y_i. Consequently, the variance for the residual r_i will be small, and the variance for the fit $\hat{y}_i$ will be large (since the fit from another data set would be quite different),

$$\text{var}[\hat{y}_i] = \sigma^2 h_i, \quad \text{var}[r_i] = \sigma^2(1 - h_i).$$

High-Leverage Points

high-leverage points: since $\sum_i h_i = p$, a rule-of-thumb is that observations with leverages more than $2p/n$ should be flagged as high-leverage points and should be examined closely.

Good high-leverage points: points are from the model as the rest sample, but with an x_i value that is far away from the sample mean. What're the advantages of including good high-leverage points?

Bad high-leverage points: do not follow the pattern suggested by the rest of the data; the LS fitting would change a lot if we remove this point.

Standardize Residuals r_i's

Residuals $r_i = y_i - \hat{y}_i$ does NOT have a constant variance. So they need to be standardized. There are two versions:

- Standardized residuals r_i^*: internally standardized; does not follow t nor normal distribution.
- Studentized residuals t_i: externally standardized; follows t distribution; will be used in our outlier test.

Residuals are very useful diagnostics. Some recommend using some standardized version of the residual instead of the raw residual r_i in all diagnostic plots.

Difference between $\mathbf{e}$ and $\mathbf{r}$

- Both are normally distributed, but

$$\mathbf{e} \sim \mathsf{N}_n\Big(\mathbf{0}, \sigma^2 \mathbf{I}_n\Big), \quad \mathbf{r} \sim \mathsf{N}_n\Big(\mathbf{0}, \sigma^2 (\mathbf{I}_n - \mathbf{H})\Big),$$

  where $\mathbf{H}$ is the projection/hat matrix.

- The errors e_i's have equal variance and are independent, while the residuals, r_i's have unequal variance and are correlated.

- $\mathbb{E}\mathbf{e} = \mathbb{E}\mathbf{r} = \mathbf{0}$. But

$$\sum e_i \neq 0, \quad \sum_i r_i = 0$$

  (by default, we assume an intercept is included in the model).

Standerdized Residuals

Since $r_i \sim \mathsf{N}(0, (1 - h_i)\sigma^2)$, consider a standardization of the residual

$$r_i^* = \frac{r_i}{\hat{\sigma}\sqrt{1 - h_i}}, \quad i = 1, \ldots, n.$$

- $\sum_i r_i^*$ is no longer zero.
- Each r_i isn't distributed as student's t distribution, since r_i isn't independent of $\hat{\sigma}$.
- As an approximation, we can view r_i^*'s iid $\mathsf{N}(0, 1)$, although they are not normally distributed and they are slightly correlated.

Studentized Residuals

- The studentized residuals are based on the idea of leave one out (also known as jackknife).

- Here is the idea: run a regression model on the $(n-1)$ samples with the i-th sample (x_i, y_i) removed. Denote the leave-one-out estimates of the regression coefficient and error variance by $\hat{\boldsymbol{\beta}}_{(i)}$ and $\hat{\sigma}_{(i)}$, where the notation (i) means "excluding the i-th observation."

- Then, check the discrepancy between y_i and $\hat{y}_{(i)} = \mathbf{x}_i^t \hat{\boldsymbol{\beta}}_{(i)}$.

- Define the studentized residuals as

$$t_i = \frac{y_i - \hat{y}_{(i)}}{\hat{\sigma}_{(i)}\left[1 + x_i^t(\mathbf{X}_{(i)}^t\mathbf{X}_{(i)})^{-1}x_i\right]^{1/2}} = \frac{r_i}{\hat{\sigma}_{(i)}\sqrt{1-h_i}}$$

which follows t_{n-p-1} if $y_i \sim \mathsf{N}(\mathbf{x}_i^T\boldsymbol{\beta}, \sigma^2)$.

The last equality above is not trivial (you can find the proof in the Appendix). One can also show that r_i^* and t_i are monotone transformation of each other.

- Nice part of leave-one-out: to calculate these leave-one-out estimates, such as $\hat{\boldsymbol{\beta}}_{(i)}$ and $\hat{\sigma}_{(i)}$, we don't need to run the model n times. Check some equalities in the Appendix.

- Masking: what if we have two observations which are close, then removing one does not give us an honest estimate of its prediction error since it has a twin brother still in the data? We can leave-two-out, or in general leave-k-out.

An Outlier Test

- Outliers are observations that do not fit the model, but Ourliers $\neq$ Observations with large residuals.

- To check outliers, we should not look at the residuals, but the leave-one-out prediction error, i.e., the studentized residuals.

- Under H_0, $t_i \sim t_{n-p-1}$. So we can use t-test to test whether the i-th observation is an outlier or not.

- Generally, we would want to perform this outlier test for all n observations, doing the tests one at a time. Simply performing the test on the largest observed residuals would be an example of data snooping, unless somehow these cases were identified before data collection.

- In order to be certain that the overall type I error rate is no greater than α, Bonferroni correction may be used. When doing so, each case would be tested at level α/n.

What to do with outliers?

- Delete them.
- But points should not be routinely deleted simply because they do not fit the model. No data snooping.
- Outliers, as well as other unusual observations discussed here, often flag potential problems of the current model. Instead of dropping them, maybe, try a new alternative model. (Outliers are normal points that haven't found their distribution yet.)

Bonferroni Correction

Suppose we are testing m hypotheses simultaneously. For each test, we use significant level α. That is, the chance of making type I error is α. Suppose we want to control the overall type I error rate (for all m tests) to be 95%, then we should set the individual significant level to be $\alpha = 5\%/m$.

$$\begin{aligned} & \mathbb{P}(\text{No type I error for all } m \text{ tests}) \\ = \; & 1 - \mathbb{P}(\text{make a type I error for test 1 OR for test 2 ... OR for test } m) \\ \geq \; & 1 - m\alpha. \end{aligned}$$

Influential Observations

- Observations whose removal greatly affects the analysis are called influential observations

- Define the Cook's distance, as an influence measure of the i-th sample

$$
\begin{aligned}
D_i &= \frac{\|\mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{X}\hat{\boldsymbol{\beta}}_{(i)}\|^2}{p\hat{\sigma}^2} = \frac{\|\hat{\mathbf{y}} - \hat{\mathbf{y}}_{(i)}\|^2}{p\hat{\sigma}^2} \\
&= \frac{(r_i^*)^2}{p}\Big(\frac{h_i}{1-h_i}\Big),
\end{aligned}
$$

  which indicates that high influential points are either outliers (large $|r_i^*|$) or high-leverage points (large h_i) or both.

- A rule-of-thumb: observations with $D_i \geq 1$ are highly influential.

Summary

- High-leverage points: $h_i = H_{ii} > 2p/n$. High-leverage points are far away from the center of the data (in terms of the Mahalanobis distance).

$$\text{var}[\hat{y}_i] = \sigma^2 h_i, \quad \text{var}[r_i] = \sigma^2(1 - h_i).$$

- Outliers: we remove the i-th point, run LS on the remaining $(n-1)$ data points, and then form a PI at $\mathbf{x}_i$; if PI covers y_i, then the i-th point is NOT an outlier.

- High influential points: Cook's distance $D_i > 1$.

$$D_i = \frac{\|\mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{X}\hat{\boldsymbol{\beta}}_{(i)}\|^2}{p\hat{\sigma}^2} = \frac{(r_i^*)^2}{p}\Big(\frac{h_i}{1-h_i}\Big),$$

which indicates that high influential points are either outliers or high-leverage points or both.

Diagnostics: Checking Error Assumptions

- Non-constant variance
- Assessing normality
- Correlated errors
- Graphical tools: residual plots, QQ-plots

  "The importance of producing and analyzing plots as a standard part of statistical analysis cannot be overemphasized." from Weisberg (1980).
- Remedies: transformation, GLS, nonlinear regression.

Residual Plots

- Plot the (studentized) residuals r_i (or r_i) against the fitted value $\hat{y}_i$.
- Plot the (studentized) residuals r_i (or r_i) against each predictor x_i.
- Plot the (studentized) residuals r_i (or r_i) against some index variable such as time or case number.
- Look for systemic patterns (non-constant variance, nonlinearity) and large absolute values of residuals.

Non-constant Variance

- Check residual plots.
- A less rigorous but quick way: `lm(abs(res)` $\sim$ `fitted-value)`
- A formal test: Breusch-Pagan Test (`bptest` in package `lmtest`)
- Remedy: transformation.

Variance Stabilizing Transformation

The goal is to find a transformation $h(Y)$ to achieve constant variance. The method for finding these transformations is based on the following. Suppose h is some smooth function. Then by Taylor's Theorem,

$$h(Y) = h(\mathbb{E}[Y]) + h'(\mathbb{E}[Y])(Y - \mathbb{E}[Y]) + \cdots,$$

where $\cdots$ denotes the remainder of this approximation, which is assumed to be reasonably small with high probability (i.e, we'll ignore it). Then

$$\text{var}[h(Y)] \approx (h'(\mathbb{E}[Y]))^2 \text{var}(Y).$$

We want to choose the transformation h such that

$$\mathsf{var}[h(Y)] \approx (h'(\mathbb{E}[Y]))^2 \mathsf{var}(Y)$$

is approximately a constant.

For example, suppose var$(Y) \propto \mathbb{E}[Y]$, then

$$h'(z) = \frac{1}{\sqrt{z}}, \longrightarrow h(z) \propto \sqrt{z}.$$

As another example, suppose var$(Y) \propto \mathbb{E}[Y]^2$, then

$$h'(z) = \frac{1}{z}, , \longrightarrow h(z) = \log z.$$

- $\sqrt{Y}$, var$(e) \propto \mathbb{E}[Y]$

  Suitable for counts from the Poisson distribution.

- $\log Y$ or $\log(Y+1)$, var$(e) \propto \mathbb{E}[Y]^2$

  Suitable for data whose range of Y is very broad, e.g., from 1 to several thousand; suitable for estimating percentage effect $(Y \propto X^{\alpha} C.)$

- $1/Y$ or $1/(Y+1)$, var$(e) \propto \mathbb{E}[Y]^4$

  Suitable for data where Y measures the waiting time or survival time. Taking reciprocals changes the scale from time (time per response) to rate (response per unit time).

Assessing Normality

Suppose we have a sample z_1, z_2, . . . , z_n, and we wish to examine the hypothesis that the z's are a sample from a normal with mean μ and variance σ^2.

A standard graphical method for inspecting the normal assumption is QQ-plot.

1. Order the z's: $z_{(1)} \leq z_{(2)} \leq \cdots \leq z_{(n)}$
2. Compute $u_i = \Phi^{-1}(\frac{i}{n+1})$, where Φ is the cdf of N$(0, 1)$
3. Plot $z_{(i)}$ against u_i. If the z's are normal, the plot should approximately result in a line.

A more formal way: Shapiro-Wilk test.

What to do if the normality assumption doesn't hold?

- For short-tailed distributions, the consequences of non-normality are not very serious.
- A transformation of Y may solve the problem.
- Use other regression methods, such as robust regression (will be discussed later).
- Still use LS, but be careful with the inference such as t-test and CI which are based on the normality assumption. Instead make the inference based on distribution-free methods such as bootstrap or permutation (will be discussed later).

Correlated Errors

- Plot residuals against time or other index such as case number.
- Use formal tests like the Durbin-Waston test ((`dwtest` in package `lmtest`)
- Remedies: use GLS (will be discussed later).

Checking Structure Assumptions (Nonlinearity)

- How do we check whether the assumption $\mathbb{E}\mathbf{y} = \mathbf{X}\boldsymbol{\beta}$ is correct?
- Lack-of-fit Test: when we have replicates. (will be discussed later)
- Partial regression plot
- Partial residual plot
- Remedies: transformation, nonlinear regression (will be discussed later)

Partial Regression Plot (Added Variable Plot)

- We want to know the relationship between response Y and a predictor X_k after the effect of the other predictors has been removed.

- To remove the effect of the other predictors, run the following two regression models

$$
\begin{aligned}
Y &\sim X_1 + \cdots + X_{i-1} + X_{i+1} + \ldots, \qquad (1)\\
X_i &\sim X_1 + \cdots + X_{i-1} + X_{i+1} + \ldots, \qquad (2)\\
\mathbf{r}_y &= \text{residuals from (1)}\\
\mathbf{r}_k^X &= \text{residuals from (2)}
\end{aligned}
$$

- Plot $\mathbf{r}_y$ vs. $\mathbf{r}_k^K$: For a valid model, then the added-variable plot should produce points randomly scattered around a line through the origin with slope $\hat{\beta}_k$. Also useful for high influential data points.

Using Transformations to Overcome Nonlinearity

Examples of linearizing transformations:

- $\log Y$ vs $\log X$

  Suitable for $\mathbb{E}(Y) = \alpha X_1^{\beta_1} \cdots X_p^{\beta_p}$.

- $\log Y$ vs X

  Suitable for $\mathbb{E}(Y) = \alpha \exp\{\sum_j X_j \beta_j\}$.

- $1/Y$ vs X

  Suitable for $\mathbb{E}(Y) = \frac{1}{\alpha + \sum_j X_j \beta_j}$.

Box-Cox Transformation of the Y's

- Box and Cox (1964) suggested a family of transformations (for positive response) designed to reduce nonnormality of the errors. In turns out that in doing this, it often reduces nonlinearity as well.

- Suppose each $y_i > 0$, and consider the following transformation[a]

$$g_\lambda(y) = \begin{cases} \frac{y^\lambda - 1}{\lambda}, & \lambda \neq 0 \\ \log y, & \lambda = 0. \end{cases}$$

  Choose λ that maximizes the likelihood of the data, under the normal assumption

$$g_\lambda(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}, \quad \mathbf{e} \sim \mathsf{N}_n(\mathbf{0}, \sigma^2\mathbf{I}).$$

[a] The transformation for $\lambda = 0$ is justified because $\lim_{\lambda \to 0} \frac{y^\lambda - 1}{\lambda} = \log y$.

- The maximum log-likelihood function for $\lambda \neq 0$ is

$$L(\lambda) = -\frac{n}{2}\log(\mathsf{RSS}_\lambda/n) + (\lambda - 1)\sum_{i=1}^{n}\log(y_i),$$

  where RSS_λ is the residual sum of squares when $g_\lambda(y)$ is the response, and for $\lambda = 0$ is

$$L(0) = -\frac{n}{2}\log(\mathsf{RSS}_0/n) - \sum_{i=1}^{n}\log(y_i).$$

  The 2nd term in these log-likelihood function corresponds to the Jacobian of the transformation.

- Note that it doesn't make sense to simply pick λ that maximizes RSS_λ since for each λ, the residual sum of squares are measured in a different scale.

- In `R`, we can graph $L(\lambda)$ versus $\lambda \in (-2, 2)$[a] and then pick the maximizer $\hat{\lambda}$.

- It's common to round $\hat{\lambda}$ to a nearby value like

  $$-1, \ -0.5, \ 0, \ 0.5, \ \text{or } 1,$$

  then the transformation (defined by $\hat{\lambda}$) is easier to interpret.

[a] The method tends to work well for λ in this range.

- To answer the question whether we really need the transformation g_λ, we can do a hypothesis testing ($H_0 : \lambda = 1$) or equivalently construct a CI for λ as follows[a]:

  $$\{\lambda : L(\lambda) > L(\hat{\lambda}) - \frac{1}{2}\chi_1^2(1-\alpha)\}.$$

  If this interval contains 1, then there is no strong evidence supporting the transformation.

[a] This is based on the result that $2(L(\hat{\lambda}) - L(\lambda_0)) \sim \chi_1^2$ under H_0.

Regression Diagnostics for MLR (Sheather chap 6)

- Any unusual patterns of the residuals? Plot standardized residuals vs fitted values.
- Any unusual data points, such as high leverage points, high influential points or outliers?
- Assess the effect of each predictor X_j on Y. Use added variable plot.
- Constant error variance (i.e., heteroscedasticity)?
- Collinearity of X's, correlated errors (will be discussed later)