Unit 3: Linear models: Misspecification

Eugene Katsevich

October 4, 2021

In our discussion of linear model inference in Unit 2, we assumed the normal linear model throughout:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \text{where } \boldsymbol{\epsilon} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_n). \tag{1}$$

In this unit, we will discuss what happens when this model is misspecified:

- Non-normality (Section 1): $\boldsymbol{\epsilon} \sim (0, \sigma^2 \boldsymbol{I}_n)$ but not $N(0, \sigma^2 \boldsymbol{I}_n)$.
- Heteroskedastic errors (Section 2): $\epsilon_i \overset{\text{ind}}{\sim} N(0, \sigma_i^2)$, where it is not the case that $\sigma_1^2 = \cdots = \sigma_n^2$.
- Correlated errors (Section 3): It is not the case that $(\epsilon_1, \dots, \epsilon_n)$ are independent.
- Model bias (Section 4): It is not the case that $\mathbb{E}[\boldsymbol{y}] = \boldsymbol{X}\boldsymbol{\beta}$ for some $\boldsymbol{\beta} \in \mathbb{R}^p$.
- Outliers (Section 5): For one or more i, it is not the case that $y_i \sim N(\boldsymbol{x}_{i*}^T \boldsymbol{\beta}, \sigma^2)$.

For each type of misspecification, we will discuss its origins, consequences, detection, and fixes (Sections 1-5). We conclude with an R demo (Section 7).

1 Non-normality

1.1 Origin

Non-normality occurs when the distribution of $y|\boldsymbol{x}$ is either skewed or has heavier tails than the normal distribution. This may happen, for example, if there is some discreteness in y.

1.2 Consequences

Non-normality is the most benign of linear model misspecifications. While we derived linear model inferences under the normality assumption, all the corresponding statements hold asymptotically without this assumption. Recall Homework 2 Question 1, or take for example the simpler problem of estimating the mean μ of a distribution based on n samples from it: We can test $H_0 : \mu = 0$ and build a confidence interval for μ even if the underlying distribution is not normal. So if n is relatively large and p is relatively small, you need not worry too much. If n is small and the errors are highly skewed or heavy-tailed, there might be an issue.

1.3 Detection

Non-normality is a property of the error-terms ϵ_i. We do not observe these directly, but we can approximate these using the residuals

$$\widehat{\epsilon}_i = y_i - \boldsymbol{x}_{i*}^T \widehat{\boldsymbol{\beta}}. \tag{2}$$

Recall from Unit 2 that $\text{Var}[\widehat{\boldsymbol{\epsilon}}] = \sigma^2(\boldsymbol{I} - \boldsymbol{H})$. Letting h_i be the ith diagonal entry of $\boldsymbol{H}$, it follows that $\widehat{\epsilon}_i \sim (0, \sigma^2(1 - h_i))$. The *standardized residuals* are defined as

$$r_i = \frac{\widehat{\epsilon}_i}{\widehat{\sigma}\sqrt{1 - h_i}}. \tag{3}$$

Under normality, we would expect $r_i \dot{\sim} N(0, 1)$. We can therefore assess normality by producing a histogram or normal QQ-plot of these residuals (see Figure 1).

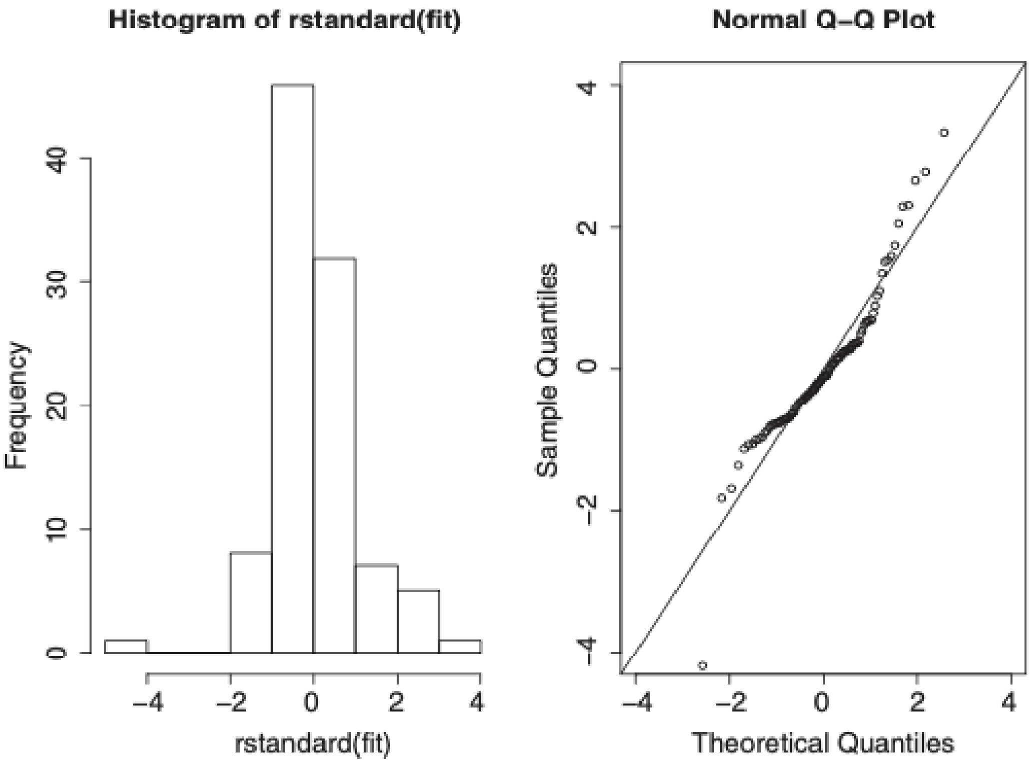

Figure 1: Histogram and normal QQ plot of standardized residuals.

1.4 Fixes

As mentioned in Section 1.2, non-normality is not necessarily a problem that needs to be fixed, except in small samples. In small samples, we can apply the bootstrap (Section 6.2.2) for robust standard error computation and a few different strategies (Section 6.3) for robust hypothesis testing.

2 Heteroskedastic errors

2.1 Origin

Suppose each observation y_i is actually the average of n_i underlying observations, each with variance σ^2. Then, the variance of y_i is σ^2/n_i, which will differ across i if n_i differ. It is also common to see the variance of a distribution increase as the mean increases (as in Figure 2), whereas for a linear model the variance of y stays constant as the mean of y varies.

2.2 Consequences

All normal linear model inference from Unit 2 hinges on the assumption that $\epsilon_i \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2)$. The coverage of confidence intervals and the levels of hypothesis tests may depart from their nominal levels. This is easiest to see if we consider the width of confidence intervals for $\boldsymbol{x}_0^T \boldsymbol{\beta}$; see Figure 2 for intuition.

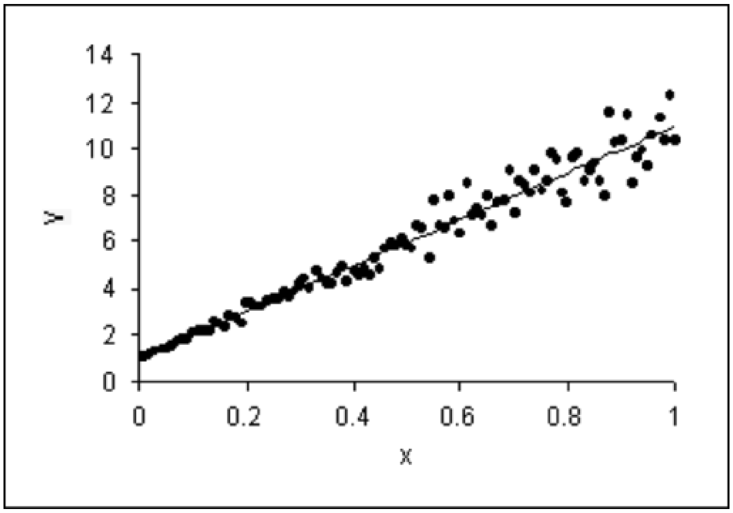

Figure 2: Heteroskedasticity in a simple bivariate linear model (image source).

2.3 Detection

Heteroskedasticity is usually assessed via the *residual plot* (Figure 3). In this plot, the standardized residuals r_i (3) are plotted against the fitted values $\widehat{\mu}_i$. In the absence of heteroskedasticity, the spread of the points around the origin should be roughly constant as a function of $\widehat{\mu}$ (Figure 3(a)). A common sign of heteroskedasticity is the fan shape where variance increases as a function of $\widehat{\mu}$ (Figure 3(c)).

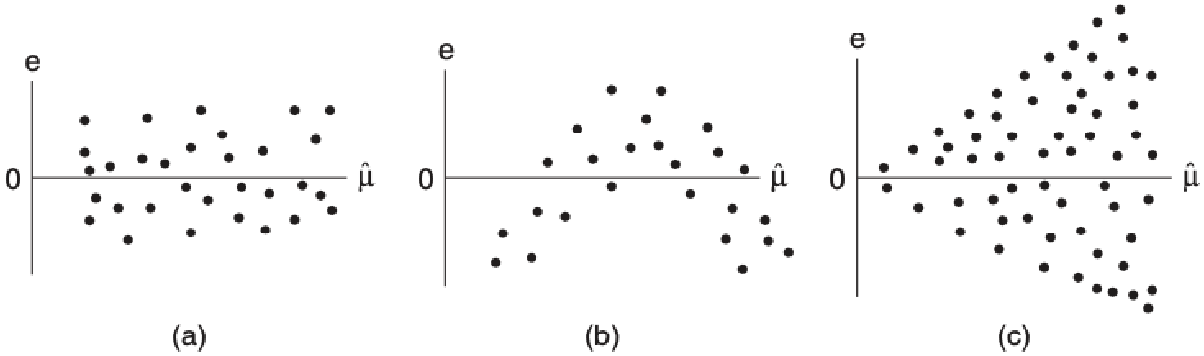

Figure 3: Residuals plotted against linear-model fitted values that reflect (a) model adequacy, (b) quadratic rather than linear relationship, and (c) nonconstant variance(image source: Agresti Figure 2.8).

2.4 Fixes

Heteroskedasticity-robust standard errors for hypothesis testing and confidence intervals can be obtained using a number of strategies, including the Huber-White sandwich estimator 6.2.1, the bootstrap 6.2.2, and permutation tests 6.3.1.

3 Correlated errors

3.1 Origin

Correlated errors can arise when observations have group, spatial, or temporal structure. Below are examples:

- Group structure: We have 10 samples $(\boldsymbol{x}_{i*}, y_i)$ each from 100 schools.
- Spatial structure: We have 100 soil samples from a 10×10 grid on a 1km×1km field.
- Temporal structure: We have 366 COVID positivity rate measurements, one from each day of the year 2020.

The issue arises because there are common sources of variation among sample that are in the same group or spatially/temporally close to one another.

3.2 Consequences

Like with heteroskedastic errors, correlated errors can cause invalid standard errors. In particular, positively correlated errors typically cause standard errors to be smaller than they should be, leading to inflated Type-I error rates. For intuition, consider estimating the mean of a distribution based on n samples. Consider the cases when these samples are independent, compared to when they are perfectly correlated. The effective sample size in the former case is n and in the latter case is 1.

3.3 Detection

Residual plots once again come in handy to detect correlated errors. Instead of plotting the standardized residuals against the fitted values, we should plot the residuals against whatever variables we think might explain variation in the response that the regression does not account for. In the presence of group structures, we can plot residuals versus group (via a boxplot); in the presence of spatial or temporal structure, we can plot residuals as a function of space or time. If the residuals show a dependency on these variables, this suggests they are correlated.

3.4 Fixes

There are a few approaches to addressing correlated errors:

1. Augment the regression model with extra terms. For example, we may want to include an indicator for school in the first example from Section 3.1. Using a fixed effect for each level of the grouping variable may use up too many degrees of freedom, so we can use *random effects* instead. Hopefully we'll have time to cover random/mixed effects models in Unit 6.
2. Estimate the covariance matrix $\boldsymbol{\Sigma}$ of the observations, so that $\boldsymbol{y} \sim N(\boldsymbol{X}\boldsymbol{\beta}, \Sigma)$. This is a *generalized least squares* problem for which inference can be carried out. The generalized least squares estimate is $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{y}$, which is distributed as $\widehat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, (\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1})$.

We can carry out inference based on the latter distributional result analogously to how we did so in Unit 2.

3. Apply a clustered or block bootstrap; see Section 6.2.2.

4 Model bias

4.1 Origin

Model bias arises when predictors are left out of the regression model:

$$\text{assumed model: } \boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}; \quad \text{actual model: } \boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{\gamma} + \boldsymbol{\epsilon}. \tag{4}$$

We may not always know about or measure all the variables that impact a response $\boldsymbol{y}$.

Model bias can also arise when the predictors do not impact the response on the linear scale. For example:

$$\text{assumed model: } \mathbb{E}[\boldsymbol{y}] = \boldsymbol{X}\boldsymbol{\beta}; \quad \text{actual model: } g(\mathbb{E}[\boldsymbol{y}]) = \boldsymbol{X}\boldsymbol{\beta}. \tag{5}$$

4.2 Consequences

In cases of model bias, the parameters $\boldsymbol{\beta}$ in the assumed linear model lose their meanings. The least squares estimate $\widehat{\boldsymbol{\beta}}$ will be a biased estimate for the parameter we probably actually want to estimate. In the case (4) when predictors are left out of the regression model, these additional predictors $\boldsymbol{Z}$ will act as confounders and create bias in $\widehat{\boldsymbol{\beta}}$ as an estimate of the $\boldsymbol{\beta}$ parameters in the true model, unless $\boldsymbol{X}^T\boldsymbol{Z} = 0$. As discussed in Unit 2, this can lead to misleading conclusions.

4.3 Detection

Similarly to the detection of correlated errors, we can try to identify model bias by plotting the standardized residuals against predictors that may have been left out of the model. A good place to start is to plot standardized residuals against the predictors $\boldsymbol{X}$ (one at a time) that are in the model, since nonlinear transformations of these might have been left out. In this case, you would see something like Figure 3(b).

It is possible to formally test for model bias in cases when we have repeated observations of the response for each value of the predictor vector. In particular, suppose that $\boldsymbol{x}_{i*} = \boldsymbol{x}_c$ for $c = c(i)$ and predictor vectors $\boldsymbol{x}_1, \dots, \boldsymbol{x}_C \in \mathbb{R}^p$. Then, consider testing the following hypothesis:

$$H_0 : y_i = \boldsymbol{x}_{i*}^T\boldsymbol{\beta} + \epsilon_i \quad \text{versus} \quad H_1 : y_i = \beta_{c(i)} + \epsilon_i. \tag{6}$$

The model under H_0 (the linear model) is nested in the model for H_1 (the saturated model), and we can test this hypothesis using an F-test called the *lack of fit F-test.*

4.4 Fixes

To fix model bias in the case (4), ideally we would identify the missing predictors $\boldsymbol{Z}$ and add them to the regression model. This may not always be feasible or possible. To fix model bias in the case (5), it is sometimes advocated to find a transformation g (e.g. a square root or a logarithm) of $\boldsymbol{y}$ such that $\mathbb{E}[g(\boldsymbol{y})] = \boldsymbol{X}\boldsymbol{\beta}$. However, a better solution is to use a *generalized linear model*, which we will discuss starting in Unit 4.

5 Outliers

5.1 Origin

Outliers often arise due to measurement or data entry errors. An observation can be an outlier in $\boldsymbol{x}$, in y, or both.

5.2 Consequences

An outlier can have the effect of biasing the estimate $\widehat{\boldsymbol{\beta}}$. This occurs when an observation has outlying $\boldsymbol{x}$ as well as outlying y.

5.3 Detection

TBD

5.4 Fixes

If outliers can be detected, then the fix is to remove them from the regression. But, we need to be careful. Definitively determining whether observations are outliers can be tricky. Outlier detection can even be used as a way to commit fraud with data, as Theranos is alleged to have done.

As an alternative to removing outliers, we can fit estimators $\widehat{\boldsymbol{\beta}}$ that are less sensitive to outliers; see Section 6.1.

6 Robust inference

TBD

6.1 Robust estimation

6.2 Robust standard error computation

6.2.1 Huber-White sandwich estimator

6.2.2 Bootstrap

Residual bootstrap

Pairs bootstrap

- Vanilla
- Cluster bootstrap
- Block bootstrap

6.3 Robust hypothesis testing

6.3.1 Permutation tests

6.3.2 Rank-based tests

6.3.3 Bootstrap-based tests

7 R demo