MODEL DIAGNOSTICS

Model diagnostics

We'll discuss some tools for checking the adequacy of the single-factor model:

$$y_{ij} = \mu_i + e_{ij}, \quad (i = 1, \ldots, t; j = 1, \ldots, r)$$

where μ_i is the ith treatment mean and $e_{ij} \sim N(0, \sigma^2)$

Question: What are the basic assumptions for a linear regression model?

We'll discuss some tools for checking the adequacy of the single-factor model where each observation is modeled as a group mean plus an error term.

What are the basic assumptions for a linear regression model?

Assumptions for a linear regression model

Recall that there are three assumptions—on the random errors e_{ij}—for the linear regression model: normality, independence, and equal variance.

If these assumptions are violated, our inferences (confidence intervals for the parameters, ANOVA test, tests for group means and so on) may be invalid.

Among the three assumptions, which one is the most crucial?

Recall that there are three assumptions on the random errors for the linear regression model: normality, independence, and equal variance.

If these assumptions are violated, our inferences, such as confidence intervals for the parameters, ANOVA test, tests for group means and so on, may be invalid.

Among the three assumptions, which one is the most crucial?

The independence assumption is the most crucial

The independence assumption is the most crucial: our inferences are reasonably robust regarding to the normality assumption and there are some remedies for non-constant variance.

The independence assumption is the most crucial: our inferences are reasonably robust regarding to the normality assumption and there are some remedies for non-constant variance.

The fixed and random components in a regression model

Note that in a regression model there is a fixed part and random part.

For example, in the one-way ANOVA model,

$$y_{ij} = \mu_i + e_{ij}, \quad (i = 1, \dots, t; j = 1, \dots, r)$$

- The group mean parameter μ_i is the fixed part (also called the systematic part): μ_i is fixed, not random, but its value is unknown to us.
- The error term e_{ij} is the random part. It is often assumed to have a normal i.i.d. distribution.

Note that error term e_{ij} is simply $y_{ij} - \mu_i$. The e_{ij}'s represent the amount of data variation that is not explained the fixed part of the model.

When we say the model is not adequate, it means the fixed part it does not adequately explain the variation in the data. One possible consequence is that there is still systematic pattern in e_{ij}'s—potentially leading to dependence among e_{ij}'s.

Note that in a regression model there is a fixed part and random part.

For example, in the one-way ANOVA model, the group mean parameter (mu i) is the fixed part (also called the systematic part) of the model. (mu i)is fixed, not random, but its value is unknown to us. The error term (e i j) is the random part. It is often assumed to have a normal i.i.d. distribution.

Note that error term (e i j) is simply (y i j - mu i). The (e i j)’s represent the amount of data variation that is not explained the fixed part of the model.

When we say the model is not adequate, it means that the fixed part does not adequately explain the variation in the data. One possible consequence is that there is still systematic pattern in (e i j)’s—potentially leading to potential statistical dependence among (e i j)’s.

The basic of model diagnostics

The basic unit of model diagnostics is the residual:

$$\hat{e}_{ij} = \text{observed} - \text{predicted} = y_{ij} - \hat{y}_{ij}.$$

For the one-way ANOVA model, $\hat{y}_{ij} = \overline{y}_{i\cdot}$.

The first thing to remember is that the basic unit of model diagnostics is the residual: the observed value minus the fitted or predicted value.

For the one-way ANOVA model, the fitted value for each observation is the corresponding group sample mean.

Check the normality assumption with a normal Q-Q plot

The e_{ij}'s are supposed to be normally distributed.

To check this assumption, we can plot a histogram of the $\hat{e}_{ij}$'s or do a normal probability plot or a normal quantile-quantile plot (a normal Q-Q plot)—see Kuehl, pp. 125-127.

Note that e_{ij}'s are not observable, so we use the estimates—the residuals:

$$\hat{e}_{ij} = \text{observed} - \text{predicted} = y_{ij} - \hat{y}_{ij}.$$

If the residuals are normal, a scatterplot of the residuals vs. the corresponding normal quantiles should be reasonably linear.

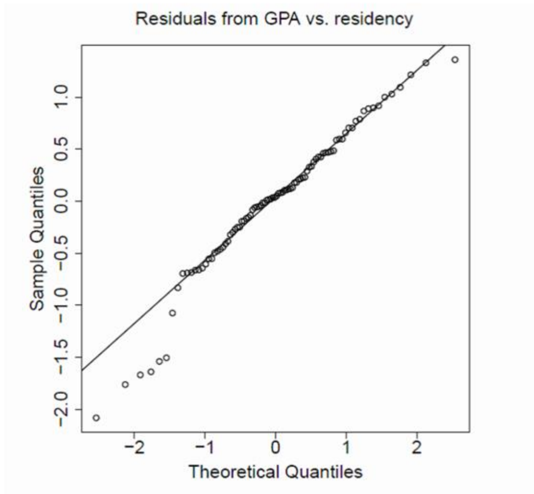
The (e i j)'s are supposed to be normally distributed.

To check this assumption, we can plot a histogram of the (e i j) hats or do a normal probability plot or a normal quantile-quantile plot.

Note that (e i j)'s are not observable, so we use the estimates—the residuals.

If the residuals are normal, a scatterplot of the residuals vs. the corresponding normal quantiles should be reasonably linear.

Example. GPA/residency data (data set resgpa).



The distribution of residuals from a model of cumulative GPA as a function of residency appears to have a heavy left tail.

In this plot, we see the normal Q-Q plot for the residuals from a one-way ANOVA model fitted to the GPA/residency data.

We see that there are a cluster of points on the lower left corner. This means that a few residuals are more negative than expected from a normal distribution of residuals.

In the context of the problem, it indicates a few students' score are lower than expected from a normal distribution of errors after accounting for the group means.

But in my opinion, in this example, the deviation from normal is not much.

Note on the normal Q-Q plot

Note that there will always be a small departure from the 45 degree line in the normal Q-Q plot near the tails even when the data are perfectly normal.

One can simulate a Q-Q plot of the same sample size to get a feel of what a “normal” normal Q-Q plot looks like (see the plot on next page):

```
set.seed(99);
qqnorm(rnorm(90), main="Normal Q-Q plot for simulated i.i.d N(0,1) data")
abline(0, 1);
```

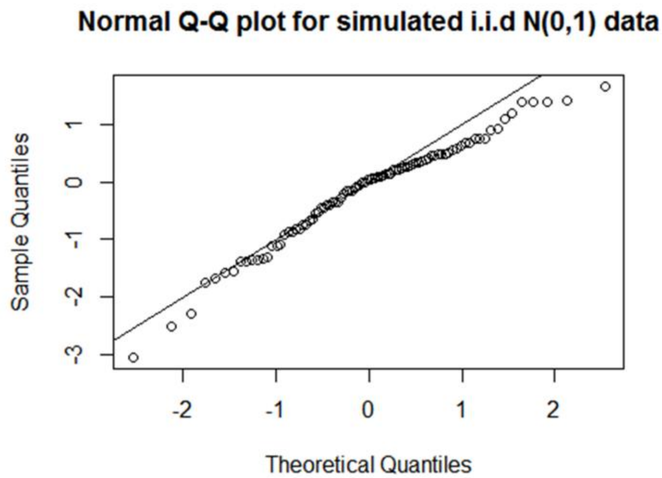
It is actually notoriously difficult to test for non-normality. So use the normal Q-Q plot only as a guideline.

Note that there will always be a small departure from the 45 degree line in the Q-Q plot near the tails even when the data are perfectly normal.

One can simulate a Q-Q plot of the same sample size to get a feel of what a “normal” normal Q-Q plot looks like.

It is actually notoriously difficult to test for non-normality. So use the Q-Q plot only as a guideline.

A normal Q-Q plot for simulated N(0, 1) data (n = 90)



How to fix non-normality?

The ANOVA F-test is quite robust with respect to non-normality, i.e., it performs well even when the normality assumption is not strictly met.

Consider transformation for highly skewed data:

- More correctly speaking: consider transformation for data when the residuals are highly skewed
- I sometimes see people do a histogram of the entire data set—before fitting the one-way ANOVA model—and find the data skewed. But that could be due the differences in group means.

How to fix non-normality?

The ANOVA F-test is quite robust with respect to non-normality, i.e., it performs well even when the normality assumption is not strictly met.

Consider transformation for highly skewed data.

More correctly speaking: consider transformation for data when the residuals are highly skewed

I sometimes see people do a histogram of the entire data set—before fitting the one-way ANOVA model—and find the data skewed. But that could be due the differences in group means.

Homogeneous variance

The linear regression model assumes that $\mathrm{Var}(e_{ij}) = \sigma^2$, for all i (treatments). Tools for checking for non-constant variance:

1. Scatterplot or boxplot of residuals vs. fitted values (group means).
2. Levene's test (Kuehl, p. 128)
   a. Let $\tilde{y}_i$ be the median of the observations in group i. Calculate $z_{ij} = |y_{ij} - \tilde{y}_i|$, for $i = 1, \ldots, t$ and $j = 1, \ldots, r_i$.
   b. Use an ANOVA to test the hypothesis that the mean z is the same for all groups.
   c. If the p-value is small, reject $H_0: \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_t^2$. That is, we have evidence of non-constant variance.

The linear regression model assumes that the error variance is the same for all observations in all groups.

For checking for non-constant variance, we can use a scatterplot or boxplot of residuals vs. fitted values—the group means.

We can also consider using Levene's test. Here we describe one version of the Levene's test:

Let (y i tilde) be the median of the observations in group i and let (z i j) be the absolution difference between (y i j) and (y i tilde).

Use an ANOVA to test the hypothesis that the mean z is the same for all groups.

If the p-value is small, reject the null hypothesis that all groups have the same variance. That is , we have evidence of non-constant variance.

F-max test

Kuehl (p. 130) also discussed a F-max test. This test is sensitive to non-normality … don't use it!

Kuehl also discussed a F-max test. This test is sensitive to non-normality … don't use it!

Remedies for non-constant variance

The ANOVA model is not very robust with respect to unequal variances, especially when sample sizes are unequal. p-values may be too high or too low.

We will discuss a couple of possible solutions.

The ANOVA model is not very robust with respect to unequal variances, especially when sample sizes are unequal. p-values may be too high or too low. We will discuss a couple of possible solutions.

Transform the response.

Theory lets us identify "variance-stabilizing" transformations for certain kinds of data, i.e., transformations that make the variance less dependent on the mean.

The first approach is to transform the data using "variance-stabilizing" transformations.

Theory lets us identify "variance-stabilizing" transformations for certain kinds of data, i.e., transformations that make the variance less dependent on the mean.

A few commonly used transformations

Type of data	*Transformation*
Skewed to the right; wide range of values	$\log(Y)$
Poisson data: counts (e.g., no. of plants in a quadrat)	$\sqrt{Y}$
Binomial data: no. of 'successes' in N trials	$\log\left(\frac{Y}{1-Y}\right)$, or $\sin^{-1}(\sqrt{Y})$

A few commonly used transformations

In this table, we show a few commonly used variation-stabilizing transformations.

My own experience is that the log transformation and the logit transformation (i.e., log y over 1 - y) are used a lot in practice.

If you want to use a transformation such as the sin inverse square root transformation, you have to think about how to interpret the transformed data.

Welch's one-way ANOVA

Welch (1951) developed an alternative to the usual F-test that works well even when variances differ among groups. Both the F-statistic and the associated degrees of freedom are modified (see Welch 1951.pdf).

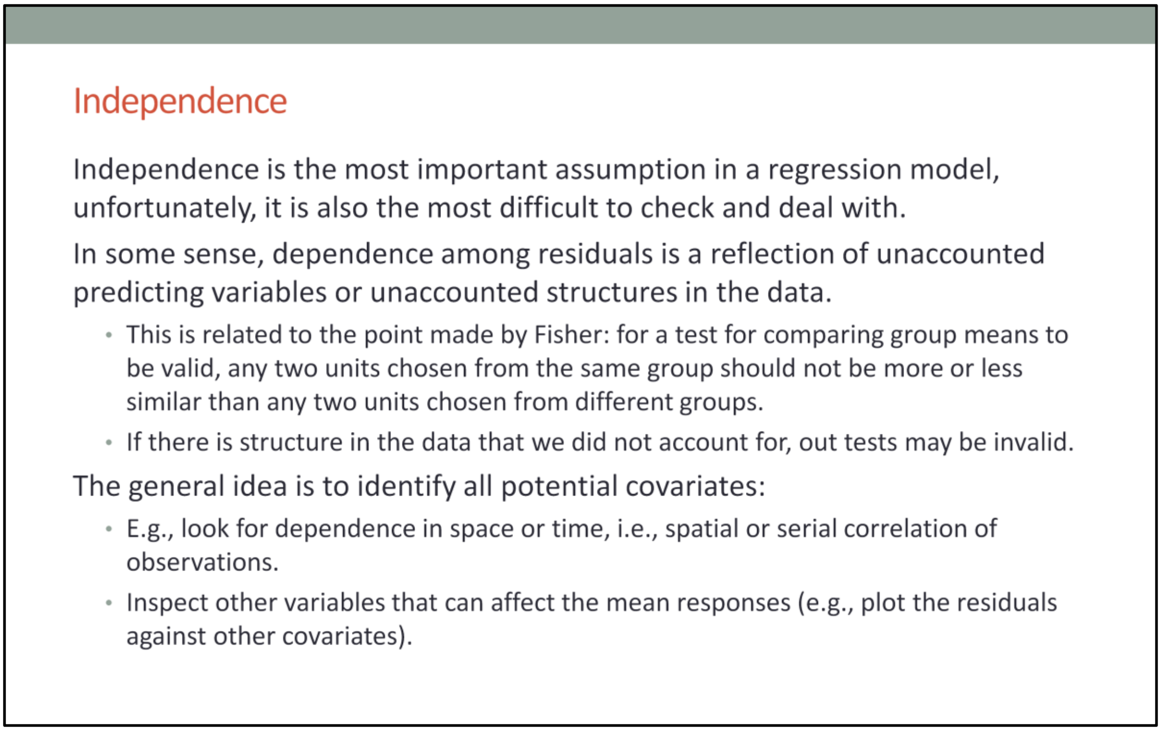
In R:

```
tmp <- oneway.test(y ~ group)
```

Another way to deal with non-constant variance is to use Welch's one-way ANOVA test, which does not assume equal variance among groups.

Welch developed an alternative to the usual F-test that works well even when variances differ among groups. Both the F-statistic and the associated degrees of freedom are modified.

In R, we can use the function oneway.test to perform Welch's one-way ANOVA test.



Independence is the most important assumption in a regression model, unfortunately, it is also the most difficult to check and deal with.

In some sense, dependence among residuals is reflection of unaccounted predicting variables or unaccounted structures in the data.

> This is related to the point made by Fisher: for a test for comparing group means to be valid, any two units chosen from the same group should not be more or less similar than any two units chosen from different groups.
>
> If there is structure in the data that we did not account for, out tests may be invalid.

The general idea is to identify all potential covariates:

> E.g., Look for dependence in space or time, i.e., spatial or serial correlation of observations.
>
> Inspect other variables that can affect the mean responses (e.g., plot the residuals against other covariates).

Outliers

The use of ‘studentized’ residuals simplifies outlier detection. In general,

$$\text{studentized residual} = \frac{\text{ordinary residual}}{\sqrt{MSE(1-\text{leverage})}}$$

In single-factor ANOVA, this becomes

$$\text{studentized residual} = \frac{\text{ordinary residual}}{\sqrt{MSE(1-1/r_i)}}$$

where r_i is the number of observations in group i. The studentized residuals should be approximately standard normal. Examine observations with values exceeding 3 or 4.

In R, if lm.out is the output from a call to lm, the studentized residuals can be obtained using rstandard(lm.out).

We can used “studentized’ residuals” to detect potential outliers.

The studentized residuals should be approximately standard normal. Examine observations with values exceeding 3 or 4.

In R, if lm.out is the output from a call to lm, the studentized residuals can be obtained using rstandard(lm.out).

What to do with detected outliers?

An outlier simply means a data point that is not well explained by the current model.

The practical implication is that **it requires further investigation**: we want to find out why it is outlying, what is special about this data point. That data point can be person, an individual animal, a city … that **requires further attention**.

Simply deleting it from the fitted model is not recommended.

An outlier simply means the data point that is not well explained by the current model.

The practical implication is that it requires further investigation: we want to find out why it is outlying, what is special about this data point.

That data point can be person, an individual animal, a city … that requires further attention.

Simply deleting it from the fitted model is not recommenced.

Summary on model diagnostics:

1. The regression model has a fixed part (the model for the mean) and a random part (the error term).

2. The basic model assumptions for a linear regression model is that the error terms are i.i.d. normal. This corresponds to three model assumptions on the error terms:
   - Normality: Q-Q plot
   - Constant variance: data transformation, Welch's one-way ANOVA test
   - Independence: check spatial, temporal trends, identify potential covariates

3. Outliers,
   - We can detect outliers using studendized residuals.
   - Outliers are experimental units that need further investigation.

In this lecture, we discussed model diagnostics.

The regression model has a fixed part (the model for the mean) and a random part (the error term).

The basic model assumptions for a linear regression model is that the error terms are i.i.d. normal. This corresponds to three model assumptions on the error terms: To check these assumptions:

> We can use a normal Q-Q plot as a guideline for checking normality. The F-test is relatively robust to non-normality.
>
> For non-constant variance, we either transform the data or use a test that does not assume equal variance (such as Welch's one-way ANOVA test).
>
> For independence, we can check spatial, temporal trends in the residuals, and try to identify unaccounted for covariates.

We can detect outliers using studendized residuals. Outliers are experimental units that need further investigation. * Normality * Constant variance * Independence