ADVANCED BAYESIAN MODELING

Bread and Peace Example: Model Checking

Diagnostics

Regression *diagnostics* are used to check the fit.

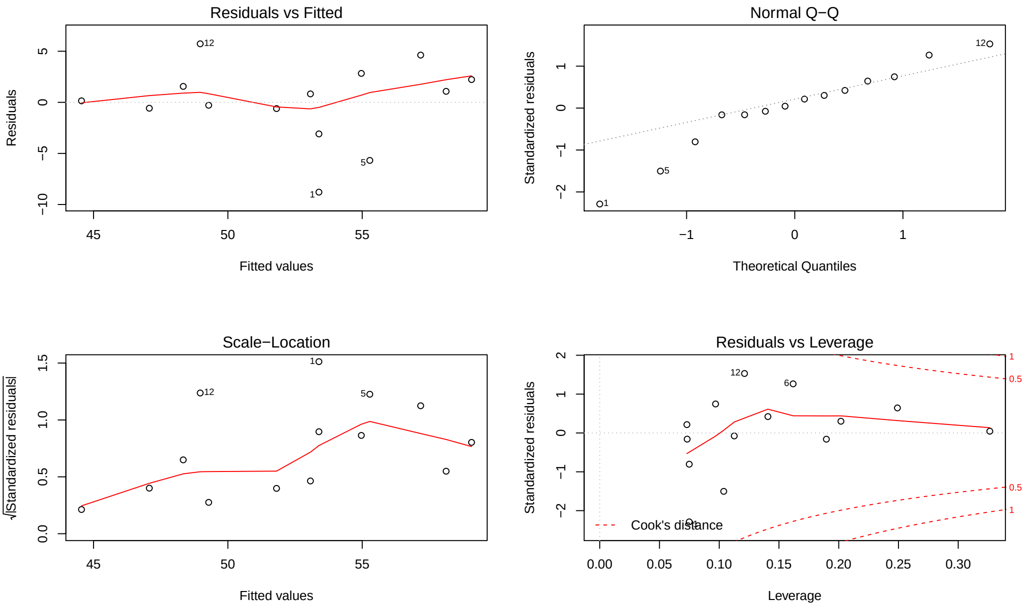
Classical regression diagnostics are based on *residuals*

$$y_i - X_i\hat{\beta} \qquad\qquad i = 1, \ldots, n$$

or their "standardized" versions.

Fit is often checked visually with *residual plots*:

```
> par(mfrow=c(2,2))
> plot(mod)  # four classical diagnostic plots of residuals
```



Residuals vs Fitted
Residuals
Fitted values
12
5
1
Normal Q−Q
Standardized residuals
Theoretical Quantiles
12
5
1
Scale−Location
√|Standardized residuals|
Fitted values
1
12
5
Residuals vs Leverage
Standardized residuals
Leverage
12
6
1
0.5
Cook's distance

Residuals are proxies for the errors

$$\varepsilon_i \;=\; y_i - X_i\beta \qquad\qquad i = 1, \ldots, n$$

and "standardized" residuals for the standardized errors

$$\varepsilon_i/\sigma \;=\; (y_i - X_i\beta)/\sigma \qquad\qquad i = 1, \ldots, n$$

which have a standard normal sampling distribution.

Though they depend on β and σ^2, they can still be used directly in Bayesian test quantities for diagnostic purposes.

For replicated y^{rep}, the replicated standardized errors in vector

$$\varepsilon^{\mathsf{rep}}/\sigma \;=\; (y^{\mathsf{rep}} - X\beta)/\sigma$$

have independent standard normal distributions, conditional on any β and σ^2.

Thus, they are easy to simulate when approximating posterior predictive p-values:

$$\Pr\big(T(y^{\mathsf{rep}}, X, \theta) \;\geq\; T(y, X, \theta) \mid y\big)$$

(Note: The same explanatory variable values are reused for the replicated data.)

First, create a matrix of posterior-simulated standardized errors (rows = simulations, columns = observations):

```
> error.std.sim <- matrix(NA, Nsim, nrow(bp))
> for(s in 1:Nsim)
+   error.std.sim[s,] <- (bp$IncumbentPct - X %*% cbind(post.beta.sim[s,])) /
+                          sqrt(post.sigma.2.sim[s])
```

Then, as a reference, create a matching matrix of independent standard normals:

```
> ref.std.normal <- matrix(rnorm(Nsim*nrow(bp)), Nsim, nrow(bp))
```

To check for outliers, might use test quantity (discrepancy)

$$T(y, X, \theta) \;=\; \max_i |\varepsilon_i/\sigma|$$

```
> mean(apply(abs(ref.std.normal), 1, max) >= apply(abs(error.std.sim), 1, max))
[1] 0.411
```

No evidence of outliers here, but might be more efficient to compare the largest-magnitude residual with the median absolute value ...

Consider test quantity (discrepancy)

$$T(y, X, \theta) \quad = \quad \frac{\max_i |\varepsilon_i / \sigma|}{\text{median}_i |\varepsilon_i / \sigma|}$$

```
> mean(apply(abs(ref.std.normal), 1, max) /
+        apply(abs(ref.std.normal), 1, median) >=
+      apply(abs(error.std.sim), 1, max) /
+        apply(abs(error.std.sim), 1, median))
[1] 0.322
```

Still no evidence of outliers.

Is there any systematic relationship between errors and time t_i (election year)?

Consider test quantity (discrepancy)

$$T(y, X, \theta) \;=\; \left|\widehat{\mathrm{cor}}(\varepsilon, t)\right|$$

where $\widehat{\mathrm{cor}}$ is sample correlation:

```
> mean(abs(cor(t(ref.std.normal), bp$Election)) >=
+       abs(cor(t(error.std.sim), bp$Election)))
[1] 0.509
```

So no evidence of a relationship.

Remark: The simple linear regression also passes tests meant to check for non-constant variance, non-normality, and serial correlation.