STATS 111/202

Lecture 13

Ana Maria Kenney

(Thanks to Koko for slight
support in this lecture)

Statistical Modeling: GLM Diagnostics

- In general, residuals are set to be $e_i = Y_i - \hat{\mu}_i$.
- The issue is that the residuals do not have constant variance in a generalized linear model (even if the model is specified correctly).
- As a result it is better to standardize the residuals.

Statistical Modeling: GLM Diagnostics

- Now set $e_i \equiv \frac{Y_i - \hat{\mu}_i}{\sqrt{V(\hat{\mu}_i)(1-h_{ii})}}$.
- This is called the Pearson residual.
- Where h_{ii} is the i-th diagonal element of the hat matrix.
- $V(\mu_i)$ is the specified variance of Y_i in the model.

Statistical Modeling: GLM Diagnostics

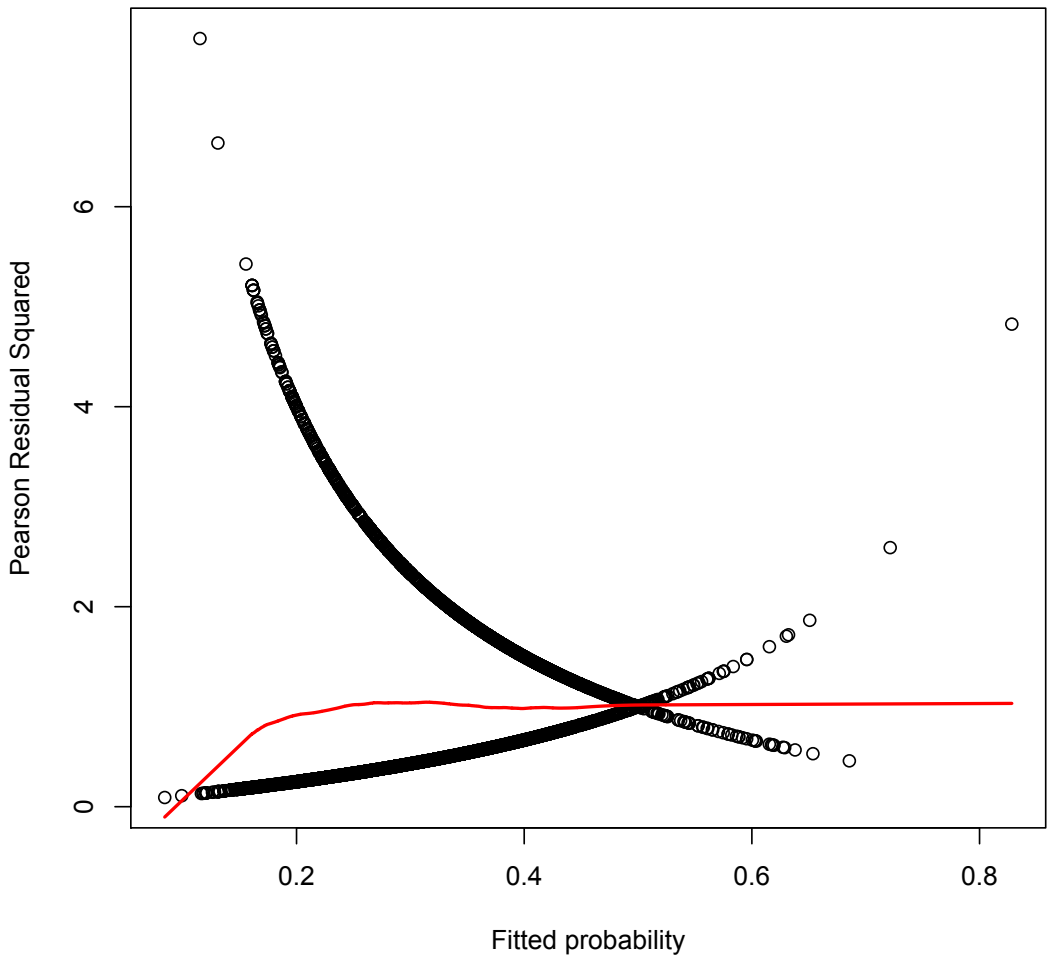
- The hat matrix is: $H = \hat{W}^{1/2} X (X^T \hat{W} X)^{-1} X^T \hat{W}^{1/2}$
- Where the square root of a square matrix is defined to be $W^{1/2} W^{1/2} = W$.
- We saw this matrix W previously (diagonal matrix with elements of the variances on the diagonal).
- For the logistic regression specifically, the matrix $W^{1/2}$ is a diagonal matrix that has elements $\sqrt{\mu_i(1-\mu_i)}$ on the diagonals.
- These residuals are useful for detecting variance misspecification or outlier detection in some cases.

Statistical Modeling: GLM Diagnostics

- Can plot residuals against the fitted values, as was done in the linear regression setting.
- Now, the residuals are the standardized Pearson residuals from the previous slides.
- The fitted values in a logistic regression are the estimated probabilities (of Y=1).
- Will additionally fit a smoothed line to check to see if our variance specification, $V(\mu)$, is correct.
- If $V(\mu)$ is correctly specified, then the Pearson residuals should have variance approximately equal to 1.

Statistical Modeling: GLM Diagnostics

Example: Framingham data. Let the response be heart disease status (yes/no) and the explanatory variables are blood pressure (sbp), age (in years) and an interaction between sbp and age.



Statistical Modeling: GLM Diagnostics

- The smoothed (red) line is roughly linear and around 1.
- This implies that our variance specification ,$V(\mu)$, is appropriate.
- If $V(\mu)$ is correctly specified, then the Pearson residuals should have variance approximately equal to 1.
- The smoothed red line is showing an approximation to the variance of the residuals given the fitted probability μ.

- Can use the hat matrix to find observations that can be considered to be outliers.
- Remember that:

$$\hat{Y}_i = \sum_{j=1}^{n} h_{ij} Y_j = h_{ii} Y_i + \sum_{i \neq j} h_{ij} Y_j.$$

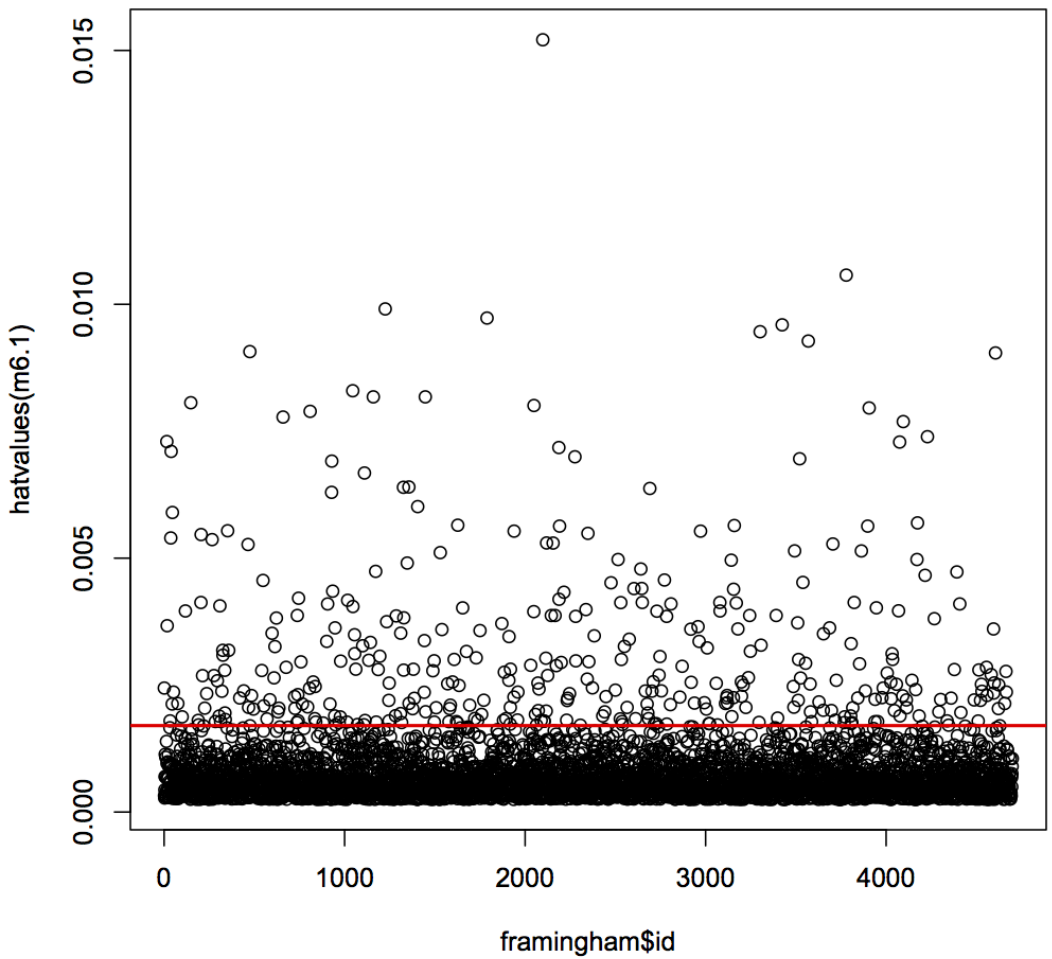
- h_{ii} can viewed as something like a weight that is assigned to the i-th response when calculating the i-th fitted value.
- h_{ii} is called the *leverage* of the i-th case.

Statistical Modeling: GLM Diagnostics

- In general,
  $\sum_{i=1}^{n} h_{ii} = p + 1$ and $0 \leq h_{ii} \leq 1$.
- h_{ii} also measures the distance between the X values for the i-th sampling unit and the means of the X values for all n sampling units.
- If the leverage of the i-th sampling unit is much greater than twice the mean leverage for all cases, 2(p + 1)/n, then might want to look closer at that case.

Statistical Modeling: GLM Diagnostics

Example: Framingham data. Let the response be heart disease status (yes/no) and the explanatory variables are blood pressure (sbp), age (in years) and an interaction between sbp and age.



Statistical Modeling: GLM Diagnostics

- Note that this observation point has high leverage because its explanatory variable levels are far from the mean of those explanatory variable.
- Let's just look at the most extreme hat value.
- This subject has sbp=260 and age=59.
- sbp has mean 132 and inter-quartile range from 116 to 144. Age has mean 46 with inter-quartile range from 39 to 53.

Statistical Modeling: GLM Diagnostics

- With a dataset that is not small, will have several points that can be considered a potential outlier, and therefore a potential influential point.
- Can consider the values with the largest hat values as potential outliers.
- Can remove these outliers and refit the model, to see how much influence the outliers in question had on the original model.
- Would only remove outliers if they are data input errors, not to get a better fitting/looking model.

Statistical Modeling: GLM Diagnostics

Overdispersion.

- Overdispersion is the situation when the actual variance of the response, var(Y_i), exceeds the variance specified by the GLM, $V(\mu_i)$.
- For example, in Poisson data, can have var(Y_i)>μ_i.
- The variance specified by the GLM (for a Poisson regression) is $V(\mu_i) = \mu_i$.
- We are commonly interested when we have underestimated the variance, which is to say $var(Y_i) > V(\mu_i)$.
- When there is overdispersion present (and we do not account for it), then we will underestimate our variances, which will then lead to inflating (increasing magnitude) of our test statistics, which then will finally lead to us to computing p-values that are lower than what they should be.

Overdispersion.

- This can occur with Poisson data when the rates among certain groups are not equal.
- Say there is a covariate Z that can be 0 or 1, where $P(Z_i = 1) = \pi$ where π is between 0 and 1.
- Let $Y_i|Z_i = 0$ (Y given Z=0) follow a Poisson distribution with $\lambda = \lambda_1$.
- Let $Y_i|Z_i = 1$ (Y given Z=1) follow a Poisson distribution with $\lambda = \lambda_2$.
- Then $E(Y_i) = (1-\pi)\lambda_1 + \pi\lambda_2 = \mu$ and $\text{var}(Y_i) = \mu + (\lambda_1 - \lambda_2)^2\pi(1-\pi)$.

Overdispersion.

- If Z is not observed, or not accounted for, then we would be incorrectly specifying the variance.
- When $\lambda_2 \neq \lambda_1$, then $var(Y_i) > \mu_i$.
- The simplest way to account for this overdispersion is to assume there is scalar overdispersion.
- That is to say that $var(Y_i) = \phi V(\mu_i)$, where $\phi \geq 1$.

Side note.

- Remember the Pearson residuals: $e_i \equiv \frac{Y_i - \hat{\mu}_i}{\sqrt{V(\hat{\mu}_i)(1-h_{ii})}}$.
- Note that $Y - \hat{\mu} = Y - HY$ where H is the hat matrix.
- $var(Y - \hat{\mu}) = var((I - H)Y) = (I - H)var(Y)(I - H)^T$.
- Using properties of the matrix H and (I-H), we have: $var(Y - \hat{\mu}) = var(Y)(I - H)$.
- If there is no overdispersion, then the Pearson residuals will have variance approximately equal to 1.

Statistical Modeling: GLM Diagnostics

- If there is overdispersion , which is to say $var(Y_i) = \phi V(\mu_i)$, then the residuals will have variance approximately equal to ϕ.
- As a result, we plot the Pearson residual against the fitted means from the model to check for overdispersion.
- Note that the residuals have mean/expectation 0, thus to plot variance, we take the squared residuals.
- Remember the variance formula, $var(e) = E(e^2) - [E(e)]^2$, where in our case $E(e) = 0$.

Statistical Modeling: GLM Diagnostics

- That is to say $E(e_i^2) = var(e_i) \approx \phi$.
- The Pearson statistic to test if $\phi = 1$ is: $\chi_p^2 = \sum_{i=1}^{n} e_i^2$.
- Under the null ($\phi = 1$), χ_p^2 is approximately chi-squared with n-p degrees of freedom, where n is our sample size and p is the number of coefficients in the model (slopes plus intercept).
- Formally, the null is $H_0 : \phi = 1$ and alternative hypothesis $H_a : \phi \neq 1$.
- The p-value for this test is the area above χ_p^2 using a chi-squared distribution with n-p degrees of freedom.

Statistical Modeling: GLM Diagnostics

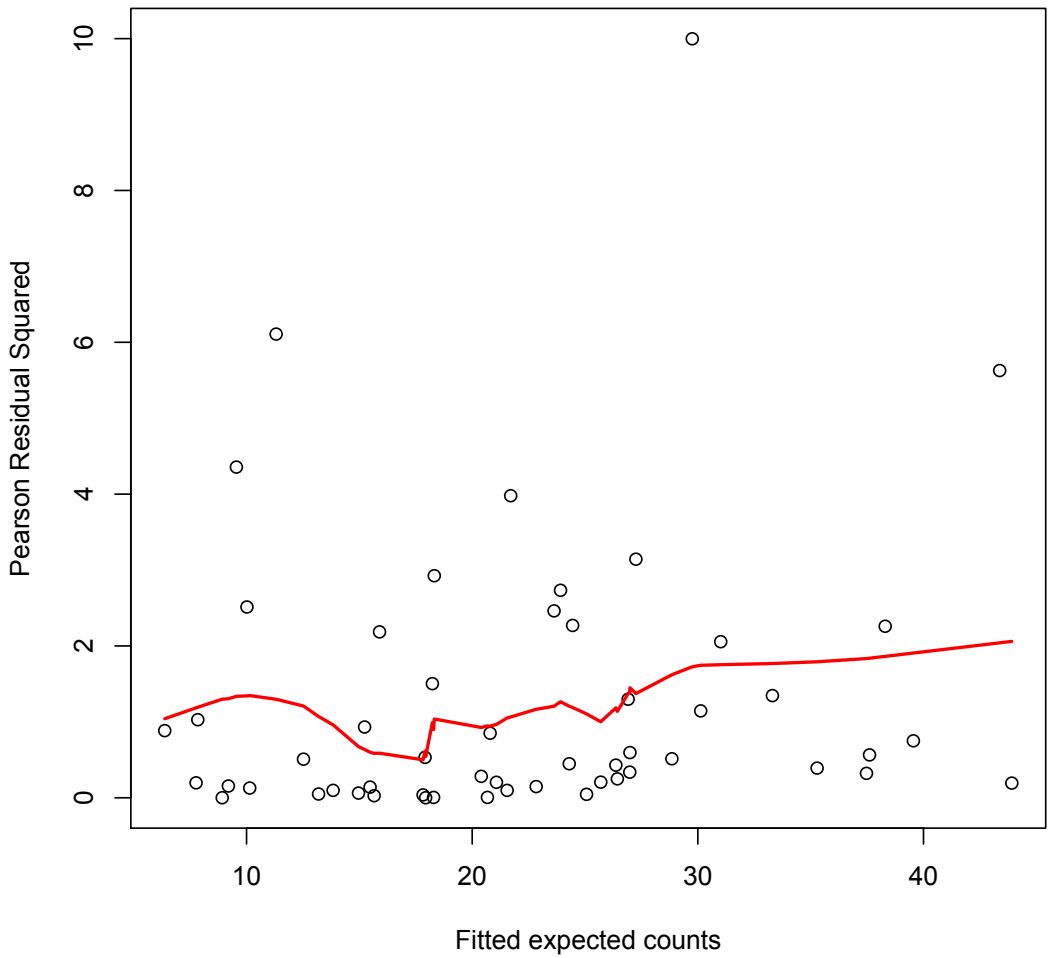
Example: Using the Gail data on the class site.

Fit a model with response being the case counts (inccases) of melanoma, and explanatory variables age group and latitude.

Will fit the model and then plot the Pearson residuals against the fitted expected values of the counts ($\hat{\mu}_i$'s). Using a fitted smooth line will determine if the variance is approximately equal to 1 given the fitted expected counts.

Statistical Modeling: GLM Diagnostics

Example: Gail data. Plot of Pearson residuals squared against fitted expected counts.



Can be argued that the fitted red line is roughly around 1 or that it is not around 1.

If not around 1, this gives evidence of over dispersion. If it is around 1, it implies no over dispersion (model is correctly specifying the variance of the response).

We can refit the model with $V(\mu) \equiv \phi V(\mu)$. The model will now estimate ϕ. Can also conduct formal test of $\phi = 1$.

The function call to R is the same as glm but now with quasipoisson as the family instead of poisson.

The idea is that we will use a quasi likelihood instead of the usual likelihood.

The quasi likelihood is constructed using the exponential family form but with a dispersion parameter $\phi \neq 1$.

With Poisson, we have $E(Y_i) = \mu_i$ and $var(Y_i) = \phi\mu_i$.

Statistical Modeling: GLM Diagnostics

Quasi-Poisson

```
glm(formula = inccases ~ factor(ageg) + latitude.ord,
family = quasipoisson(link = log), data = gail, offset = log(persyrs))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.9430  -0.7380  -0.2058   0.6037   2.9128

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)              -10.91171    0.11808 -92.408  < 2e-16 ***
factor(ageg)>=75  years    2.71086    0.13338  20.324  < 2e-16 ***
factor(ageg)35-44 years    1.66212    0.11900  13.968  < 2e-16 ***
factor(ageg)45-54 years    1.87783    0.11367  16.520  < 2e-16 ***
factor(ageg)55-64 years    2.04197    0.11774  17.343  < 2e-16 ***
factor(ageg)65-74 years    2.19106    0.13112  16.710  < 2e-16 ***
latitude.ord               0.41011    0.04211   9.739 7.48e-13 ***
---
(Dispersion parameter for quasipoisson family taken to be 1.475136)

    Null deviance: 1196.489  on 53  degrees of freedom
Residual deviance:   71.135  on 47  degrees of freedom
AIC: NA
Number of Fisher Scoring iterations: 4
```

Statistical Modeling: GLM Diagnostics

Regular Poisson

```
glm(formula = inccases ~ factor(ageg) + latitude.ord, family = poisson(link = l
    data = gail, offset = log(persyrs))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.9430  -0.7380  -0.2058   0.6037   2.9128

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)              -10.91171    0.09722 -112.23   <2e-16 ***
factor(ageg)>=75  years    2.71086    0.10982   24.68   <2e-16 ***
factor(ageg)35-44 years    1.66212    0.09798   16.96   <2e-16 ***
factor(ageg)45-54 years    1.87783    0.09359   20.06   <2e-16 ***
factor(ageg)55-64 years    2.04197    0.09694   21.06   <2e-16 ***
factor(ageg)65-74 years    2.19106    0.10796   20.30   <2e-16 ***
latitude.ord               0.41011    0.03467   11.83   <2e-16 ***
---

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 1196.489  on 53  degrees of freedom
Residual deviance:   71.135  on 47  degrees of freedom
AIC: 342.91
```

Statistical Modeling: GLM Diagnostics

- The model estimates ϕ to be 1.475 ($\hat{\phi} = 1.475$).
- The Pearson statistic to test $\phi = 1$ is 69.33 with 47 degrees of freedom (we need to compute this ourselves using the formula on slide 18).
- In R, the code to get the test statistics is:

```
sum(residuals(model, "pearson")^2)
```

- P-value is 0.0187 (in R this is 1-pchisq(69.33 , 47)) . At a 0.05 significance level, would reject the null and conclude evidence for the alternative.
- Would conclude that we do have evidence that $\phi \neq 1$ (and thus there is over dispersion).
- Note that the quasi approach has coefficient parameter estimates remain the same but the standard error estimates are different.

Statistical Modeling: GLM Diagnostics

Side notes on a second approach to deal with variance misspecification.

- Another way to account for overdispersion is by using the robust standard errors instead of the usual ones reported by R.
- The robust standard error estimates will be based on the following:

$$(X^T \hat{W} X)^{-1} X^T \hat{W} \hat{\Sigma} \hat{W} X (X^T \hat{W} X)^{-1}.$$

- Where W is the matrix of the weights and Σ is the variance-covariance matrix of the observations (and the hats signify the estimates).
- Similar to the linear regression case, this method assumes the variance structure that is specified may not be the true correct one.
- This method will only update the standard errors of the model, not the estimated coefficients.

Statistical Modeling: GLM Diagnostics

Robust standard errors using Poisson regression.

	Estimate	Robust SE	ci95.lo	ci95.hi	z value	Pr(>|z|)
(Intercept)	-10.912	0.132	-11.171	-10.652	-82.439	0
factor(ageg)>=75 years	2.711	0.148	2.420	3.002	18.267	0
factor(ageg)35-44 years	1.662	0.122	1.423	1.901	13.648	0
factor(ageg)45-54 years	1.878	0.131	1.622	2.134	14.376	0
factor(ageg)55-64 years	2.042	0.128	1.791	2.293	15.929	0
factor(ageg)65-74 years	2.191	0.131	1.935	2.447	16.788	0
latitude.ord	0.410	0.035	0.342	0.478	11.760	0