MASM22/FMSN30: Linear and Logistic Regression, 7.5 hp
FMSN40: ... with Data Gathering, 9 hp

Lecture 8, spring 2023
Likelihood ratio test, influence, residuals, variable selection

Mathematical Statistics / Centre for Mathematical Sciences
Lund University

25/4-23

Oslo: more variables

Model: $Y_i = 1$ if the concentration of PM_{10} particles lies above $50\,\mu g/m^3$, else 0. $Y_i \sim Bin(1, p_i)$ with

$$\ln \frac{p_i}{1-p_i} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \beta_4 x_{\text{neg}} + \beta_5 x_{\text{pos}}$$

where

$x_1 =$	number of cars/1000
$x_2 =$	wind speed (m/s)
$x_{\text{neg}} =$	1 if temperature difference between 25 m and 2 m above the ground is $< -1\,^\circ$C, else 0.
$x_{\text{pos}} =$	1 if ... is $> +1\,^\circ$C, else 0.
$x_{\text{zero}} =$	... is between -1 and $+1\,^\circ$C (reference category)

A large temperature difference can create a "lid" that keeps the particles close to the ground.

Oslo: estimates

	β_j	$\hat{\beta}_j$	s.e.	P-value	95 % C.I.	e^{β_j}	95 % C.I.
Intercept	β_0	-1.3	0.5	0.009	$(-2.3, -0.4)$	0.27	$(0.10, 0.70)$
cars/1000	β_1	0.2	0.2	0.37	$(-0.2, 0.6)$	1.20	$(0.81, 1.80)$
wind	β_2	-0.5	0.2	0.005	$(-0.8, -0.2)$	0.62	$(0.44, 0.85)$
cars:wind	β_3	0.1	0.1	0.03	$(0.01, 0.3)$	1.15	$(1.01, 1.31)$
diff < -1	β_4	1.8	0.4	< 0.001	$(1.04, 2.6)$	5.96	$(2.83, 12.9)$
diff $> +1$	β_5	1.0	0.3	0.001	$(0.4, 1.6)$	2.77	$(1.50, 5.10)$

- The number of cars is not significant(?!) but its interaction with the wind speed is.
- Higher wind speeds give less probability of large PM_{10} concentrations ($\beta_2 < 0 \Leftrightarrow e^{\beta_2} < 1$) ...
- ... but the effect is smaller when the number of cars is large ($\beta_3 > 0 \Leftrightarrow e^{\beta_3} > 1$).
- Non-zero temperature differences, in either direction, increase the probability of high concentration levels ($\beta_4 > 0 \Leftrightarrow e^{\beta_4} > 1$ and $\beta_5 > 0 \Leftrightarrow e^{\beta_5} > 1$)

Model selection

- In logistic regression (and in general for nonlinear models) we do not have sums of squares and cannot do an ANOVA decomposition. We can thus not use either the global or the partial F-test to test our model.
- In the F-test we used $\text{SS(Error)} = Q(\hat{\boldsymbol{\beta}})$ which is the minimized loss function.
- We have now the maximized likelihood function, $L(\hat{\boldsymbol{\beta}})$. Use that instead!

Deviance

The **Deviance**, D, of a fitted model is defined as

$$D = 2\left(\ln L(\hat{\boldsymbol{\mu}}_{\text{sat}}; \boldsymbol{Y}) - \ln L(\hat{\boldsymbol{\beta}}; \boldsymbol{Y})\right)$$

where $\hat{\boldsymbol{\mu}}_{\text{sat}}$ is the estimates in the saturated model with one parameter for each observation.

Saturated model in logistic regression

If we estimate p_i using $\hat{\mu}_i = \hat{p}_i = Y_i$ $(= 0$ or $1)$ we get

$$\begin{aligned}\ln L(\hat{\boldsymbol{\mu}}_{\text{sat}}) &= \ln \prod_{i=1}^{n} \hat{p}_i^{Y_i}(1-\hat{p}_i)^{1-Y_i} \\ &= \sum_{i;Y_i=1} Y_i \ln Y_i + \sum_{i;Y_i=0} (1-Y_i)\ln(1-Y_i) = 0.\end{aligned}$$

Null model, D_0

The deviance for the null model with only the intercept β_0 and $\hat{p}_i = \frac{e^{\hat{\beta}_0}}{1+e^{\hat{\beta}_0}} = \bar{Y}$ then becomes

$$\begin{aligned}D_0 &= 2\Big(0 - \ln L(\hat{\beta}_0)\Big) = -2\sum_{i=1}^{n}\big(Y_i \ln \hat{p}_i + (1-Y_i)\ln(1-\hat{p}_i)\big) \\ &= -2n\big(\bar{Y}\ln\bar{Y} + (1-\bar{Y})\ln(1-\bar{Y})\big).\end{aligned}$$

General model, D

For a general model with $\hat{p}_i = \dfrac{e^{\mathbf{x}_i\hat{\boldsymbol{\beta}}}}{1+e^{\mathbf{x}_i\hat{\boldsymbol{\beta}}}}$ the deviance becomes

$$D = 2\left(0 - \ln L(\hat{\boldsymbol{\beta}})\right) = -2\ln L(\hat{\boldsymbol{\beta}})$$
$$= -2\sum_{i=1}^{n}\left(Y_i \ln \hat{p}_i + (1-Y_i)\ln(1-\hat{p}_i)\right).$$

- The deviance measures the distance (in log-likelihood terms) between our model and the saturated, model.
- A large likelihood means a large probability for the estimated model to produce our data, which is good.
- So a small deviance is good.

Olso: Deviance

Running `summary(model.oslo)` gives

```
Null deviance: 536.85  on 499  degrees of freedom
Residual deviance: 463.40  on 494  degrees of freedom
AIC: 475.4
```

- `Null deviance` is D_0 for a model having intercept only:
  $\hat{p}_i = \bar{Y} = 0.228$ gives
  $D_0 = -2 \cdot 500 \cdot (0.228 \ln 0.228 + (1 - 0.228) \ln(1 - 0.228)) = 536.85$
- `Residual deviance` is D for our model.

How can we use this to test if our model is significantly different from (better than) the null model (similar to a global F-test)?

Likelihood ratio test (global)

For nested models we can compare the likelihoods through the deviance. If H_0: $\beta_1 = \cdots = \beta_p = 0$ is false the full model is better than the null model and should give larger probabilities of getting the observations that we actually got, i.e.

$$L(\hat{\boldsymbol{\beta}}) > L(\hat{\beta}_0) \Leftrightarrow \ln L(\hat{\boldsymbol{\beta}}) > \ln L(\hat{\beta}_0) \Leftrightarrow \underbrace{-2 \ln L(\hat{\beta}_0)}_{D_0} > \underbrace{-2 \ln L(\hat{\boldsymbol{\beta}})}_{D}$$

$$\Leftrightarrow D_0 - D = \sum_{i=1}^{n} \left(Y_i \ln(\frac{\hat{p}_i}{\bar{Y}})^2 + (1 - Y_i) \ln(\frac{1 - \hat{p}_i}{1 - \bar{Y}})^2 \right) > 0$$

If H_0 is true it can be proven that, asymptotically as $n \to \infty$, $D_0 - D \sim \chi^2(p)$ and we should reject H_0 at significance level α if

$$D_0 - D > \chi^2_\alpha(p)$$

Likelihood ratio test (partial)

The likelihood ratio test also does the job of a partial F-test:

- Test H_0: k specific β-parameters (e.g. the last k) $= 0$.
- If H_0 is false then

$$D_{\text{red}} - D_{\text{full}} = \sum_{i;Y_i=1} \ln(\frac{\hat{p}_{i,\text{full}}}{\hat{p}_{i,\text{red}}})^2 + \sum_{i;Y_i=0} \ln(\frac{1-\hat{p}_{i,\text{full}}}{1-\hat{p}_{i,\text{red}}})^2 \gg 0.$$

- For the successes, when $Y_i = 1$, the probability p_i should be close to 1 and we want $\hat{p}_{i,\text{full}} > \hat{p}_{i,\text{red}}$.
- For the failures, when $Y_i = 0$, the probability p_i should be close to 0 and we want $1 - \hat{p}_{i,\text{full}} > 1 - \hat{p}_{i,\text{red}}$.
- If H_0 is true then $D_{\text{red}} - D_{\text{full}} \sim \chi^2(k)$, asymptotically, and we should reject H_0 if

$$D_{\text{red}} - D_{\text{full}} > \chi^2_\alpha(k).$$

Example: Oslo

- Is our model better than the null model?
  H_0: $\beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$.
  $D_0 - D = 536.85 - 463.40 = 73.45 > \chi^2_{0.05}(5) = 11.1$. Reject H_0 at $\alpha = 0.05$ significance level. Yes, our model is better than nothing.
- Do we need to take the number of cars and/or the wind speed into account?
  H_0: $\beta_1 = \beta_2 = \beta_3 = 0$.
  Reduced model: $\ln p/(1-p) = \beta_0 + \beta_4 x_{\text{neg}} + \beta_5 x_{\text{pos}}$
  $D_{\text{red}} - D_{\text{full}} = 499.80 - 463.40 = 36.40 > \chi^2_{0.05}(3) = 7.81$.
  Reject H_0 at $\alpha = 0.05$ significance level. Yes, cars and/or wind speed matters.

R is intelligent and we can use `anova(model.red, model.full)`. When `model` is estimated using `glm()`, the `anova()`-function knows to produce a deviance table instead of an ANOVA table.

Likelihood based test and confidence interval for β_j

- When n is small or medium sized the normal approximation of $\hat{\beta}_j$ and $d(\hat{\beta}_j)$ are bad and we should *not* use the Wald test or the Wald test based confidence interval for β_j.
- The rate of convergence for $D_{\text{red}} - D_{\text{full}} \sim \chi^2(k)$ is faster so this approximation is better and should be used instead.

Likelihood Ratio test for β_j

Test H_0: $\beta_j = 0$ against H_1: $\beta_j \neq 0$ using an LR-test with $k = 1$.

Profile likelihood based confidence interval for β_j

Define the confidence interval for β_j as all values θ_0 such that H_0: $\beta_j = \theta_0$ cannot be rejected against H_1: $\beta_j \neq \theta_0$ at significance level α when using a Likelihood Ratio test (instead of a Wald test).

Note: This is what `confint(model)` does for a `glm()`-model if the `MASS`-package is installed (it probably is), hence the `Waiting for profiling to be done...`

(*) Profile likelihood confidence interval for β_j

- Notation: $\theta = \beta_j$ while $\boldsymbol{\delta} = \boldsymbol{\beta}_{-j}$ is the vector of the other β-parameters.
- Then $L(\boldsymbol{\beta})$ can be rewritten as $L(\theta, \boldsymbol{\delta})$ and its maximized value $L(\hat{\boldsymbol{\beta}})$ as $L(\hat{\theta}, \hat{\boldsymbol{\delta}})$ with deviance $D_{\text{full}} = -2 \ln L(\hat{\theta}, \hat{\boldsymbol{\delta}})$.
- The profile likelihood function for θ is defined as

$$L_1(\theta) = \max_{\boldsymbol{\delta}} L(\theta, \boldsymbol{\delta})$$

  i.e., the maximum of the likelihood function over the remaining β-parameters, when β_j is set to the value θ.
- The maximized value of the profile likelihood when $\beta_j = \theta_0$ is then $L(\theta_0, \hat{\boldsymbol{\delta}}_0)$ with deviance $D_{\text{red},\theta_0} = -2 \ln L(\theta_0, \hat{\boldsymbol{\delta}}_0)$.
- Using an LR-test, the reduced model with H_0: $\beta_j = \theta_0$ is rejected against the full model if $D_{\text{red},\theta_0} - D_{\text{full}} > \chi^2_\alpha(1)$.
- The confidence interval for β_j consists of the values of θ_0 that cannot be rejected, i.e., where $D_{\text{red},\theta_0} - D_{\text{full}} \leq \chi^2_\alpha(1)$.

Wald vs profile likelihood — when to use what

- A Wald test of H_0: $\beta_j = 0$ based on the approximate $\hat{\beta}_j \sim N(\beta_j, d(\hat{\beta}_j))$ is OK if it is obvious whether H_0 should be rejected or not, i.e.
  - $\hat{\beta}_j$ is close to 0 and the P-value is large or
  - $\hat{\beta}_j$ is far from 0 and the P-value is small.

  If the P-value is close to α we should use the LR-test instead.
- For a confidence interval for β_j we want to know what the value could be, not just whether it is 0 or not. Then we should always use the profile likelihood method.
- For the linear predictor $\mathbf{x}_0\hat{\boldsymbol{\beta}} \sim N(\mathbf{x}_0\boldsymbol{\beta}, d(\mathbf{x}_0\hat{\boldsymbol{\beta}}))$ we are not interested in a test, but only in a confidence interval. However, there is no suitable null hypothesis to base a reasonable test on so we cannot use the profile likelihood, but have to use the Wald-based confidence interval.

Leverage (again)

In logistic regression we have the linear predictors

$$\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\mathbf{Z} = \mathbf{P}\mathbf{Z}$$

where the projection matrix can be re-written as the hat-matrix

$$\mathbf{P} = \mathbf{W}^{1/2}\mathbf{X}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}^{1/2}$$

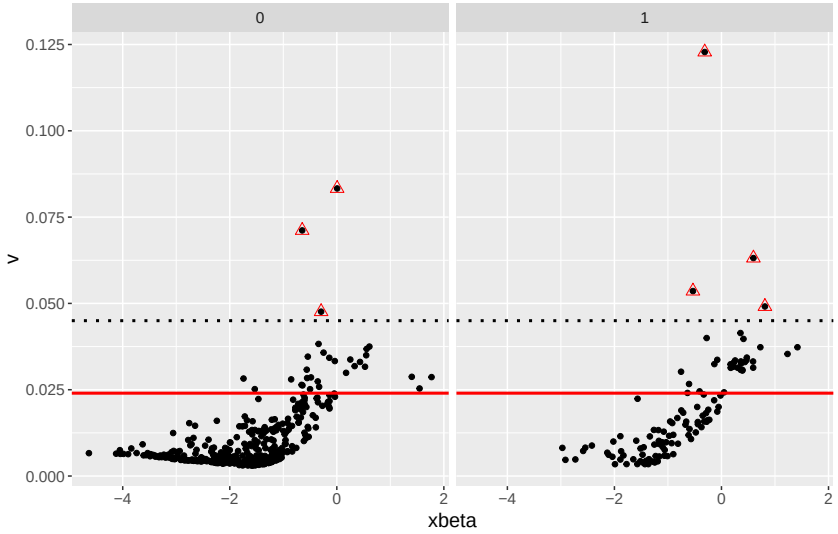
and $\mathbf{W}^{1/2}$ is a diagonal matrix with elements

$$\sqrt{w_{ii}} = \sqrt{\hat{p}_i(1-\hat{p}_i)}.$$

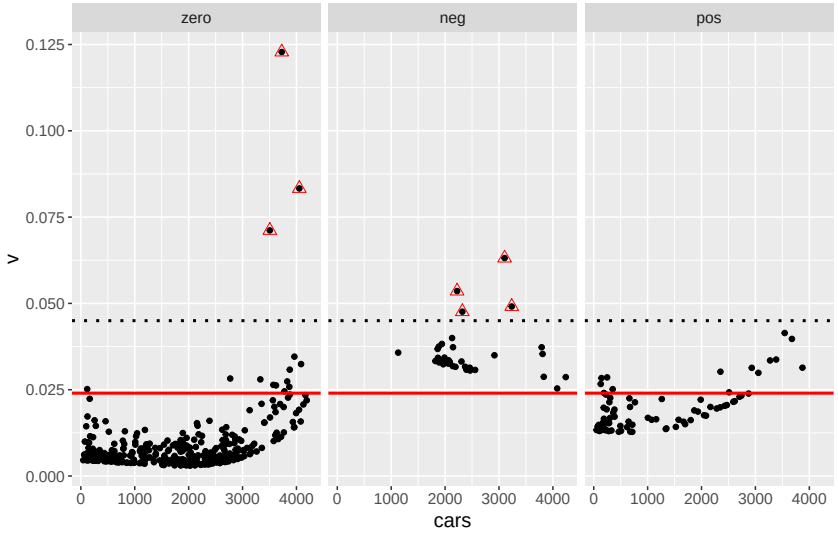
The leverage values in logistic regression are the diagonal values v_{ii} of the hat-matrix They now depend both on $\mathbf{X}$ and $\mathbf{Y}$ and, as such, are no longer indicators of outliers w.r.t. $\mathbf{X}$.

However, they can still be used to standardize residuals.

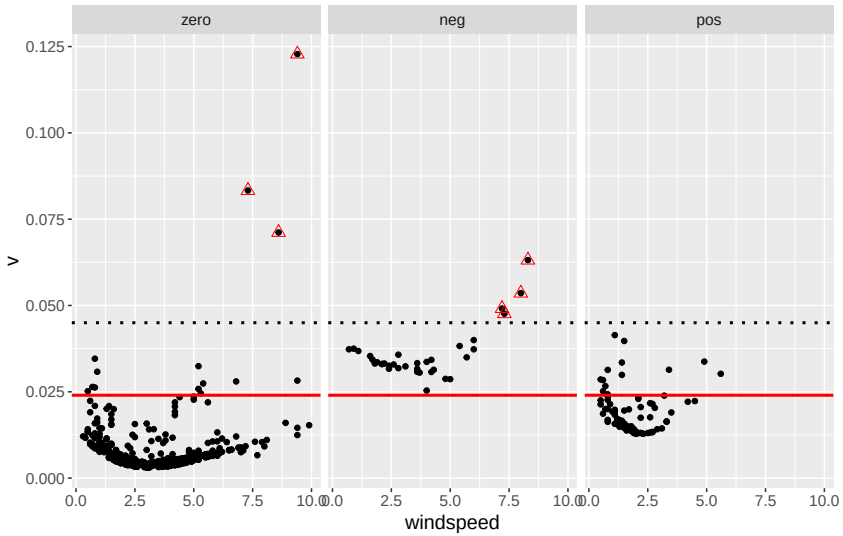
Leverage vs linear predictor, by Y=0 or Y=1



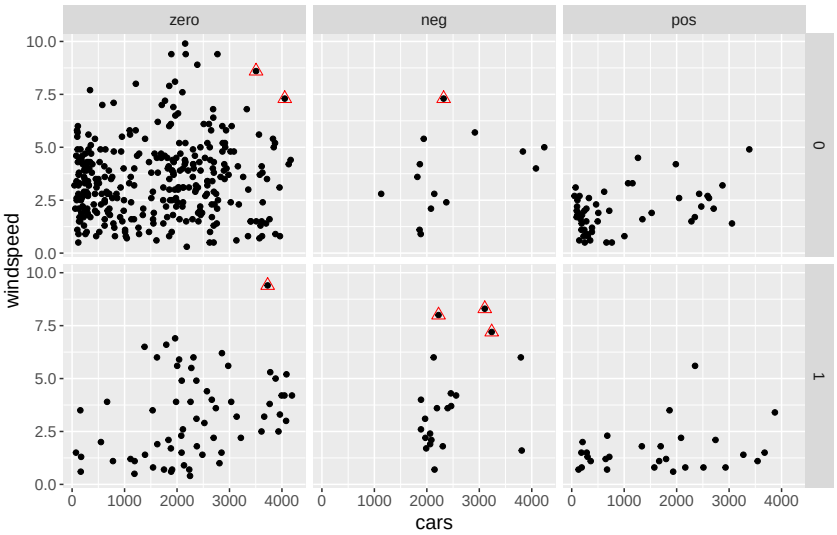
Leverage vs cars, by temp diff



Leverage vs wind speed, by temp diff



wind speed vs cars, by Y=0 or Y=1 and by temp diff



Pearson residuals

Simple standardization, since $Y_i \sim Bin(1, p_i)$ with $E(Y_i) = p_i$ and $V(Y_i) = p_i(1 - p_i)$:

$$\tilde{r}_i = \frac{Y_i - \hat{p}_i}{\sqrt{\hat{p}_i(1 - \hat{p}_i)}} \qquad \text{(not } N(\cdot, \cdot) \quad \text{not even asymptotically!)}$$

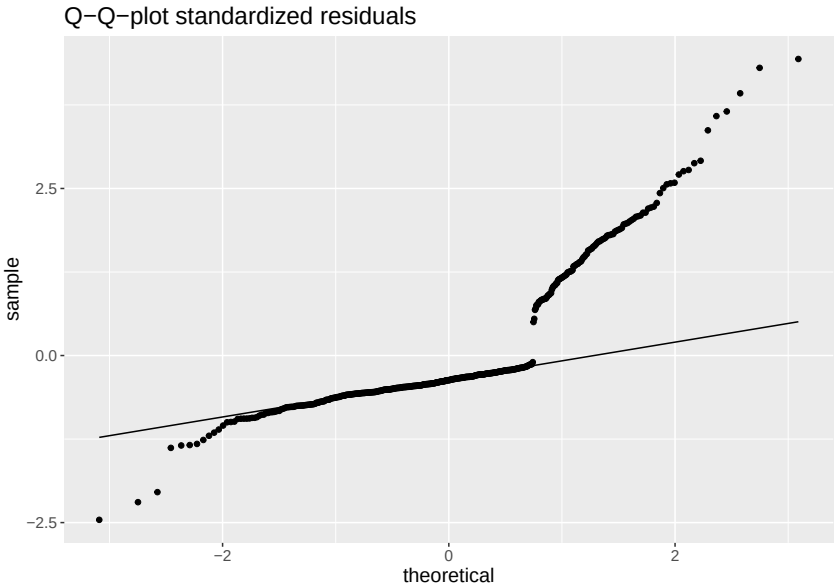
Standardized residuals

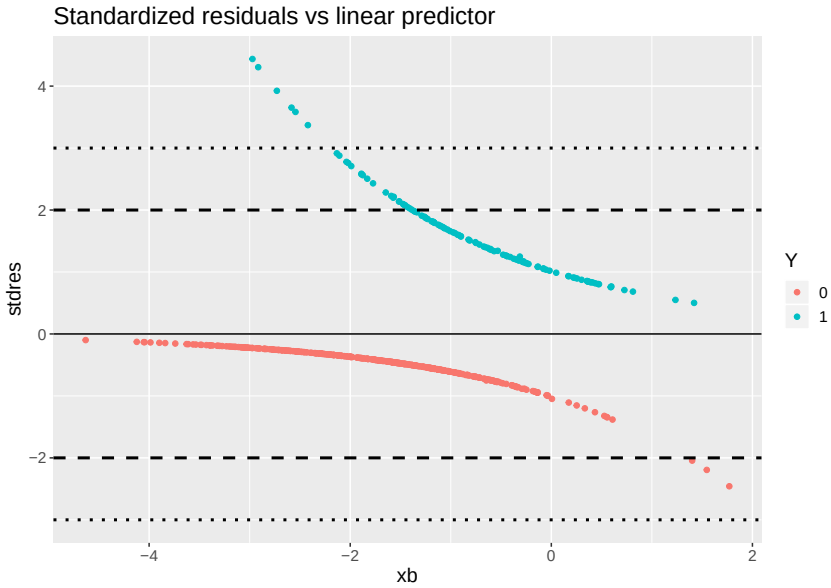
As in linear regression, we can standardize the residuals using the leverage:

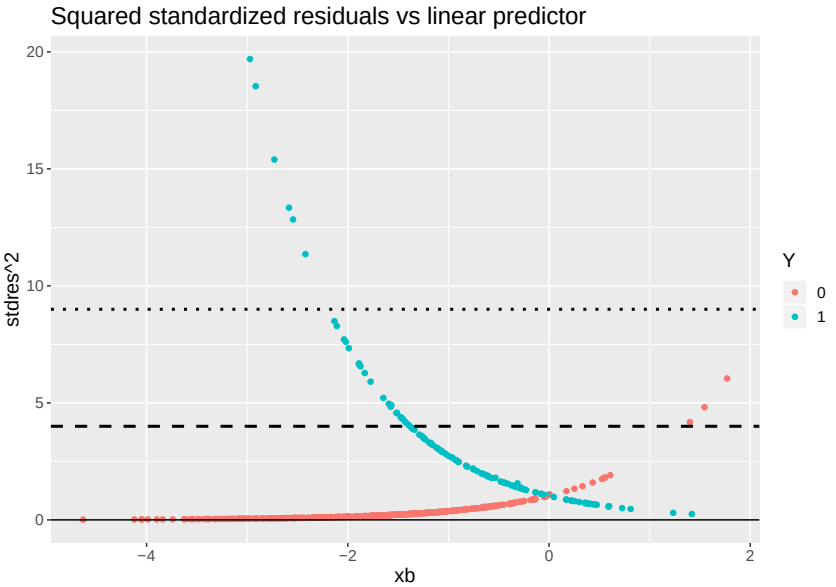
$$r_i = \frac{Y_i - \hat{p}_i}{\sqrt{\hat{p}_i(1 - \hat{p}_i)(1 - v_{ii})}} \approx N(0, 1) \qquad \text{(if 1 had been large)}$$

If $|r_i| > |\lambda_{\alpha/2}| \approx 2$ or > 3 it might be considered suspiciously large.

Plots of r_i vs $\mathbf{x}_i\hat{\boldsymbol{\beta}}$ can be useful, although it's sometimes more revealing to plot their squares, e.g. r_i^2 vs $\mathbf{x}_i\hat{\boldsymbol{\beta}}$.







- Not at all normally distributed.
- Larger residuals when $Y_i = 1$.
- Weird trends (a feature, not a bug!)

Deviance residuals, d_i

Since $L(\hat{\boldsymbol{\beta}})$ is a product of probabilities the deviance, $D = -2 \ln L(\hat{\boldsymbol{\beta}})$, can be seen as a sum of non-negative values and we can rewrite it as a sum of squares:

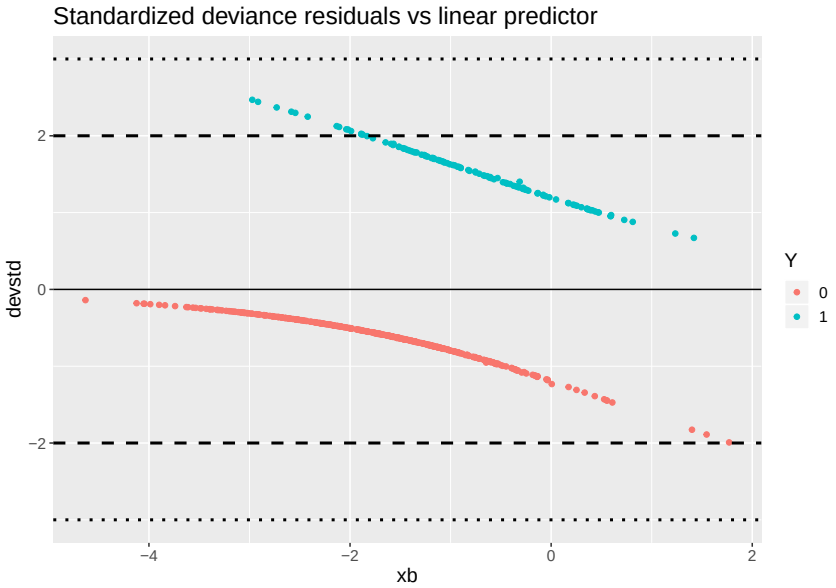
$$D = -2 \sum_{i=1}^{n} \left(Y_i \ln \hat{p}_i + (1 - Y_i) \ln(1 - \hat{p}_i) \right)$$

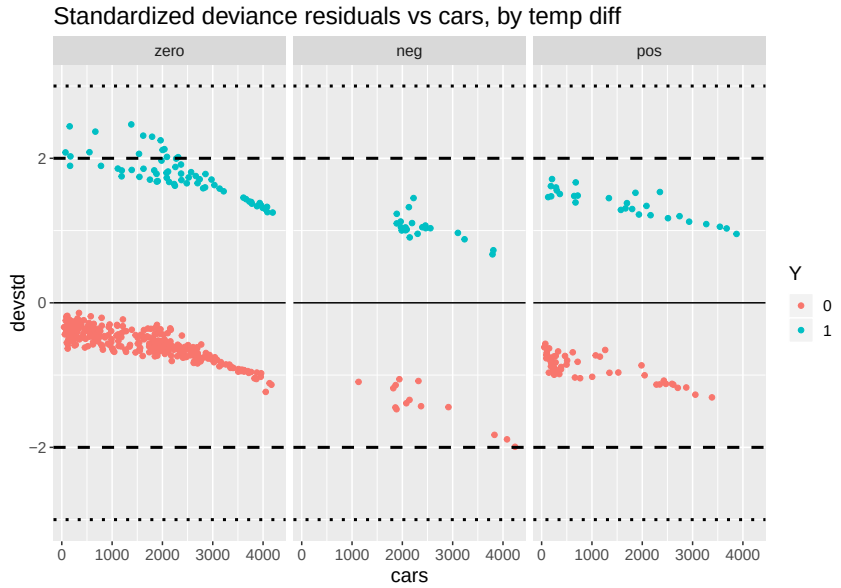
$$= \sum_{i;Y_i=1} 2 \ln \frac{1}{\hat{p}_i} + \sum_{i;Y_i=0} 2 \ln \frac{1}{1 - \hat{p}_i} = \sum_{i=1}^{n} d_i^2 \qquad \text{where}$$

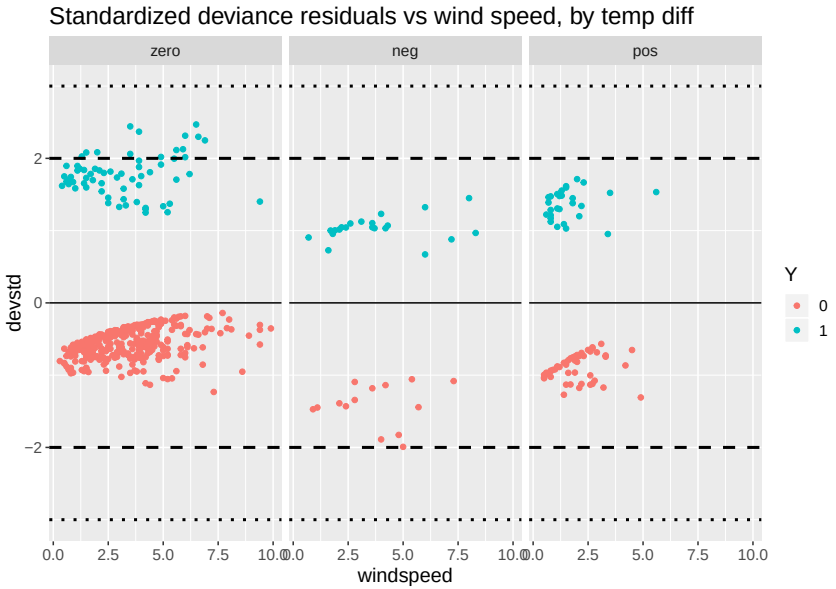
$$d_i = \begin{cases} -\sqrt{2 \ln \dfrac{1}{1 - \hat{p}_i}} & \text{if } Y_i = 0 \\ +\sqrt{2 \ln \dfrac{1}{\hat{p}_i}} & \text{if } Y_i = 1 \end{cases}$$

The deviance residual will be small if $Y_i = 0$ and $\hat{p}_i$ is close to zero, or if $Y_i = 1$ and $\hat{p}_i$ is close to one. Otherwise it will be large.

The deviance residuals can be standardized as $d_i / \sqrt{1 - v_{ii}}$.
If the absolute value is > 2 or > 3 it can be considered to be too large.







- Much more well behaved.
- Some slight problems predicting high concentrations at zero temperature difference, but not alarming.

Influential observations

We can measure the influence of individual observations on the β-estimates in a similar way as in linear regression.

Cook's distance

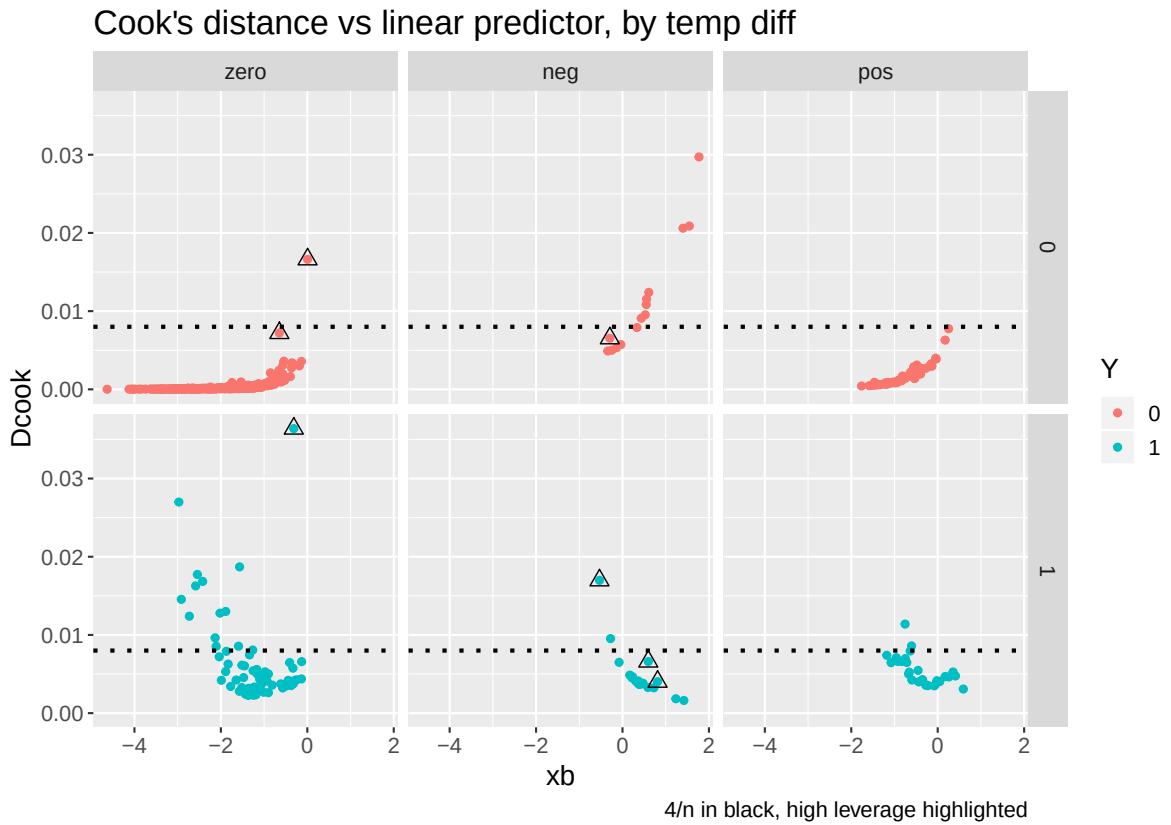
There is a version of Cook's distance for logistic regression:

$$D_i^{\text{Cook}} = \frac{r_i^2}{p+1} \cdot \frac{v_{ii}}{1-v_{ii}}$$

where r_i are the standardized residuals. We might consider influential cases those with $D_i^{\text{Cook}} > 1$ or $> 4/n$.

dfbetas

We also have similar versions of DFBETAS.



- Some observations with large influence on the estimates.
- Some of them where observations with high leverage.
- But the negative temperature differences are a problem, especially when the PM_{10} concentration is not high.

AIC and BIC again

Information for a model with $p+1$ parameters:

$$\mathrm{AIC}(p+1) = 2(p+1) - 2\ln L(\hat{\boldsymbol{\beta}}) = 2(p+1) + D$$
$$\mathrm{BIC}(p+1) = \ln n \cdot (p+1) - 2\ln L(\hat{\boldsymbol{\beta}}) = \ln n \cdot (p+1) + D$$

Tradeoff between small deviance and large number of parameters: D decreases and $p+1$ increases with p.

The "best" model is the one with the smallest AIC/BIC.

Model	df	AIC	BIC
0:null	1	535.8	543.1
1:cars	2	515.2	523.7
2:cars+wind	3	507.6	520.3
red:tempdiff	3	505.8	518.4
3:cars+zerodiff	3	480.3	**492.9**
4:cars*wind	4	500.6	517.5
5:cars+tempdiff	4	481.4	498.3
6:cars*wind + zerodiff	5	476.2	497.3
oslo:cars*wind+tempdiff	6	**475.4**	500.7

R^2 for linear regression (again)

For linear regression we could calculate the fraction of the variability in Y that was explained by our model by

$$R^2 = 1 - \frac{\text{SS(Error)}}{\text{SS(Total}_{\text{corr}})}, \qquad R^2_{\text{adj}} = 1 - \frac{\text{MS(Error)}}{\text{MS(Total}_{\text{corr}})}$$

McFadden's pseudo R^2

Use the log-likelihood (or deviance) instead of the sum of squares:

$$R^2_{\text{McF}} = 1 - \frac{\ln L(\hat{\boldsymbol{\beta}})}{\ln L(\hat{\beta}_0)} = 1 - \frac{D}{D_0},$$

$$R^2_{\text{McF,adj}} = 1 - \frac{\ln L(\hat{\boldsymbol{\beta}}) - p/2}{\ln L(\hat{\beta}_0)} = 1 - \frac{D + p}{D_0}$$

with $0 \leq R^2_{\text{McF}} \leq 1$ while $R^2_{\text{McF,adj}} \leq 1$ and negative if $p > D_0 - D$.

Cox-Snell and Nagelkerke pseudo R^2 for logistic regression

$$R^2_{\text{Cox-Snell}} = 1 - \left(\frac{L(\hat{\beta}_0)}{L(\hat{\boldsymbol{\beta}})} \right)^{2/n}$$

$$R^2_{\text{Nagelkerke}} = \frac{R^2_{\text{Cox-Snell}}}{1 - (L(\hat{\beta}_0))^{2/n}} = \frac{1 - \left(\frac{L(\hat{\beta}_0)}{L(\hat{\boldsymbol{\beta}})} \right)^{2/n}}{1 - (L(\hat{\beta}_0))^{2/n}}$$

with $0 \leq R^2_{\text{Cox-Snell}} \leq 1 - (L(\hat{\beta}_0))^{2/n}$ and $0 \leq R^2_{\text{Nagelkerke}} \leq 1$.
These cannot be adjusted to compensate for added variables.

In `R` you can get the likelihood values using `logLik(model)` and calculate the pseudo R^2 yourself.

Example: Olso

Pseudo R^2 as percentages. Maximum value for $R^2_{\text{Cox-Snell}} = 65.8\,\%$.

Model	$p+1$	R^2_{McF}	$R^2_{\text{McF,adj}}$	$R^2_{\text{Cox-Snell}}$	$R^2_{\text{Nagelkerke}}$
0:null	1	0	0	0	0
1:cars	2	4.8 %	4.6 %	5.0 %	7.6 %
2:cars+wind	3	6.6 %	6.2 %	6.8 %	10.3 %
red:tempdiff	3	6.9 %	6.5 %	7.1 %	10.8 %
3:cars+zerodiff	3	11.7 %	11.3 %	11.6 %	17.9 %
4:cars*wind	4	8.2 %	7.1 %	7.7 %	12.9 %
5:cars+tempdiff	4	11.8 %	11.3 %	11.9 %	18.1 %
6:cars*wind+zerodiff	5	13.2 %	12.4 %	13.2 %	20.0 %
oslo:cars*wind+tempdiff	6	13.7 %	**12.7 %**	13.7 %	20.8 %

The value of a pseudo R^2 is not really interpretable but it can be used to compare models.

The adjusted McFadden agrees with AIC that the full model is best.

Better model comparison tools next week!