Regression Analysis

Logistic Regression

Nicoleta Serban, Ph.D.

Professor

Stewart School of Industrial and Systems Engineering

Goodness of Fit Assessment:
Data Examples

Georgia Tech

About This Lesson

Data Example: Smoking

- Between 1972 and 1974, a survey was taken in Whickham, a mixed urban and rural district near Newcastle upon Tyne, United Kingdom.

  - Among the information obtained originally was whether a person was a smoker or not.

- Twenty years later a follow-up study was conducted.

  - 76.12% of the 582 smokers were still alive, while only 68.58% of 732 nonsmokers were still alive.

Smokers had a higher survival rate than nonsmokers!
Call Philip Morris, smoking leads to a longer life span!

Georgia
Tech

GOF Hypothesis Test

Deviance Test for GOF using deviance residuals
c(deviance(smoke2), 1-pchisq(deviance(smoke2),11))
[1] 4.345918e+01 9.033325e-06

Test for goodness-of-fit:
- Using deviance residuals: P-value $\approx$ 0
- Reject the null hypothesis of good fit (thus NOT a good fit)

GOF test using Pearson residuals
pearres2 = residuals(smoke2,type="pearson")
pearson.tvalue = sum(pearres2^2)
c(pearson.tvalue, 1-pchisq(pearson.tvalue,11))
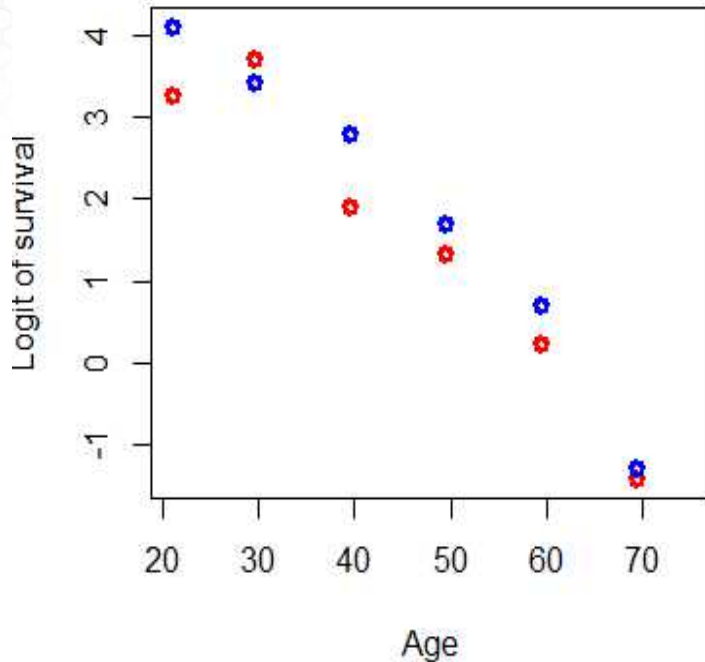[1] 36.751889370 0.000126796

Test for goodness-of-fit:
- Using Pearson residual: P-value $\approx$ 0.0001
- Reject the null hypothesis of good fit (thus NOT a good fit)

Georgia
Tech

Linearity Assumption

plot(Age,log((Survived/At.risk)/(1-Survived/At.risk)), ylab="Logit of survival", main="Scatterplot of logit survival rate vs age", col=c("red","blue"), lwd=3)

The relationship between the logit of survival and age is more quadratic than linear.

Improve the Fit

Age.squared = Age*Age
smoke3 = glm(Survived/At.risk ~ Smoker + Age + Age.squared, weights=At.risk, family=binomial)
summary(smoke3)
Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	2.5190783	1.0248206	2.458	0.0140 *
Smoker	-0.4284561	0.1770581	-2.420	0.0155 *
Age	0.0951102	0.0430095	2.211	0.0270 *
Age.squared	-0.0021673	0.0004309	-5.030	4.91e-07 ***

Null deviance: 641.496 on 13 degrees of freedom
Residual deviance: 19.808 on 10 degrees of freedom

Test for significance: β_{smoker} P-value $\approx$ 0.015, statistically significant at 0.05
Test for significance: $\beta_{\text{Age.squared}}$ P-value $\approx$ 0, statistically significant

Georgia
Tech

GOF Test for Improved Model

round(c(deviance(smoke3), 1-pchisq(deviance(smoke3),10)),2)
[1] 19.81 0.03

pearres3 = residuals(smoke3,type="pearson")
pearson = sum(pearres3^2)
round(c(pearson, 1-pchisq(pearson,10)),2)
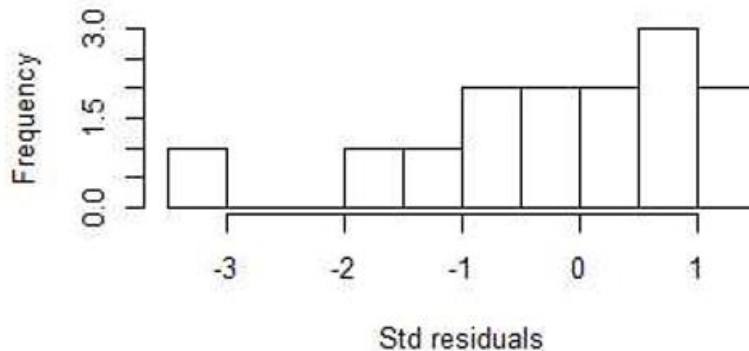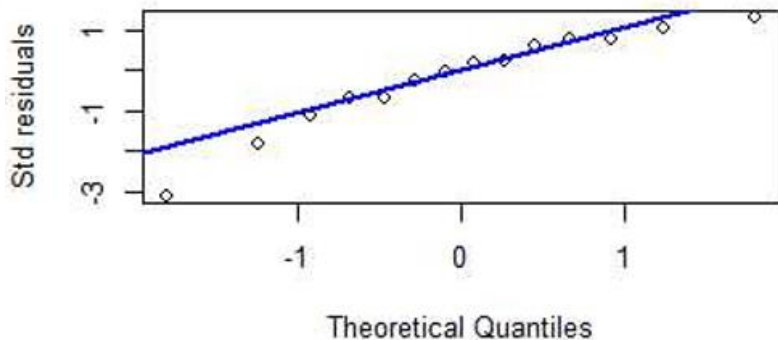[1] 14.79 0.14

Does the goodness of fit improve?
- Using deviance residuals: P-value = 0.03
- Using Pearson residual: P-value = 0.14
- Do not reject the null hypothesis of good fit using Pearson residuals, but do reject using Deviance residuals at the significance level 0.03 or higher.

Georgia Tech

Residual Analysis

res = resid(smoke3,type="deviance")
qqnorm(res, ylab="Std residuals")
qqline(res,col="blue",lwd=2)
hist(res,10,xlab="Std residuals", main="")

Normal Q-Q Plot

Higher Order Nonlinearity

Fit a logistic regression model with Age as a factor

smoke4 = glm(Survived/At.risk ~ Smoker + factor(Age), weights=At.risk, family=binomial)
summary(smoke4)
Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	3.8601	0.5939	6.500	8.05e-11 ***
Smoker	-0.4274	0.1770	-2.414	0.015762 *
factor(Age)29.5	-0.1201	0.6865	-0.175	0.861178
factor(Age)39.5	-1.3411	0.6286	-2.134	0.032874 *
factor(Age)49.5	-2.1134	0.6121	-3.453	0.000555 ***
factor(Age)59.5	-3.1808	0.6006	-5.296	1.18e-07 ***
factor(Age)69.5	-5.0880	0.6195	-8.213	< 2e-16 ***
factor(Age)75	-27.8073	11293.1437	-0.002	0.998035

   Null deviance: 641.4963 on 13 degrees of freedom
Residual deviance: 2.3809 on 6 degrees of freedom

Higher Order Nonlinearity

Fit a logistic regression model with Age as a factor

smoke4 = glm(Survived/At.risk ~ Smoker + factor(Age), weights=At.risk, family=binomial)
summary(smoke4)
Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	3.8601	0.5939	6.500	8.05e-11 ***
Smoker	-0.4274	0.1770	-2.414	0.015762 *
factor(Age)29.5	-0.1201	0.6865	-0.175	0.861178
factor(Age)39.5	-1.3411	0.6286	-2.134	0.032874 *
factor(Age)49.5	-2.1134	0.6121	-3.453	0.000555 ***
factor(Age)59.5	-3.1808	0.6006	-5.296	1.18e-07 ***
factor(Age)69.5	-5.0880	0.6195	-8.213	< 2e-16 ***
factor(Age)75	-27.8073	11293.1437	-0.002	0.998035

Null deviance: 641.4963 on 13 degrees of freedom
Residual deviance: 2.3809 on 6 degrees of freedom

Test for significance: β_{smoker} P-value ≈ 0.015, statistically significant at 0.05
Test for significance: Not all regression coefficients for the dummy variables for age are statistically significant.

Georgia Tech

Higher Order Nonlinearity: GOF

round(c(deviance(smoke4), 1-pchisq(deviance(smoke4),6)),2)
[1] 2.38 0.88

pearres4 = residuals(smoke4,type="pearson")
pearson = sum(pearres4^2)
round(c(pearson, 1-pchisq(pearson,6)),2)
[1] 2.37 0.88

Does the goodness of fit improve?
- Using deviance residuals: P-value = 0.88
- Using Pearson residual: P-value = 0.88
- Do not reject the null hypothesis of good fit using either Pearson residuals or Deviance residuals.

Georgia Tech

Different Link Function

smoke5 = glm(Survived/At.risk ~ Smoker + Age + Age.squared, weights=At.risk, family=binomial(link = probit))
summary(smoke5)
Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.1033963	0.4904877	2.250	0.02447 *
Smoker	-0.2277451	0.0970191	-2.347	0.01890 *
Age	0.0681279	0.0213095	3.197	0.00139 **
Age.squared	-0.0013767	0.0002173	-6.335	2.37e-10 ***

Null deviance: 641.496 on 13 degrees of freedom
Residual deviance: 18.233 on 10 degrees of freedom

Test for significance: β_{smoker} P-value ≈ 0.018, statistically significant at 0.05
Test for significance: $\beta_{\mathrm{Age.squared}}$ P-value ≈ 0, statistically significant

Georgia Tech

Different Link Function: GOF

round(c(deviance(smoke5), 1-pchisq(deviance(smoke5),10)),2)
[1] 18.23 0.05

pearres5 = residuals(smoke5,type="pearson")
pearson = sum(pearres5^2)
round(c(pearson, 1-pchisq(pearson,10)),2)
[1] 14.00 0.17

Does the goodness of fit improve?
- Using deviance residuals: P-value = 0.05
- Using Pearson residual: P-value = 0.17
- Do not reject the null hypothesis of good fit using Pearson residuals or using deviance residuals at the significance level 0.01.

Georgia Tech

Simpson's Paradox

Simpson's paradox: Reversal of an association when looking at a marginal relationship versus a conditional relationship.

- Smoking is statistically significant with a positive estimated coefficient under the marginal model.
- Smoking has a negative estimated coefficient under the conditional model.

Marginal versus Conditional Relationship

- *Marginal*: Capturing the association of a predicting variable to the response variable without consideration of other factors
- *Conditional*: Capturing the association of a predicting variable to the response variable conditional on other predicting variables in the model

Georgia Tech

Summary