Evaluating Model Fit

Chapter 7, Lab 2: Solutions

OpenIntro Biostatistics

Topics

- Checking model assumptions
- Using R^2 and adjusted R^2

Methods for evaluating the fit of a multiple regression model are similar to those for a simple regression model. Residual plots become essential tools for checking modeling assumptions since it is not possible to make a two-dimensional plot of a response variable against several predictors simultaneously. This lab discusses the use of residual plots to check assumptions for multiple linear regression and introduces adjusted R^2.

The material in this lab corresponds to Section 7.3 of *OpenIntro Biostatistics*.

Introduction

Assumptions for multiple linear regression

1. *Linearity*: For each predictor variable x_i, change in the predictor is linearly related to change in the response variable when the value of all other predictors is held constant.
2. *Constant variability*: The residuals have approximately constant variance.
3. *Independent observations*: Each set of observations $(y, x_1, x_2, \ldots, x_p)$ is independent.
4. *Approximate normality of residuals*: The residuals are approximately normally distributed.

Plots for checking assumptions

The linearity assumption is assessed with respect to each predictor variable. For each predictor, examine a residual plot in which the values of the predictor variable are on the x-axis and the model residuals are on the y-axis. Any patterns or curvature are indicative of non-linearity, as in the residual plot for simple linear regression.

Since each case in a dataset has one residual value and one predicted (i.e., fitted) value, regardless of the number of predictors in the model, the constant variance assumption can still be assessed with the same method as for simple linear regression: examining a scatterplot of predicted values on the x-axis and residual values on the y-axis.

Normal probability plots can be used to check the normality of the residuals.

R^2 and adjusted R^2

Adding a variable to a regression model always increases the value of R^2. The **adjusted** R^2 imposes a penalty for including predictors that do not contribute much towards explaining observed variation in the response variable.

RFFT, statin use, and age in the prevend data

The questions in this section use data from `prevend.samp`, a random sample of $n = 500$ individuals from the `prevend` dataset. Run the code shown in the template to load `prevend.samp` from the `oibiostat` package and convert the statin use variable into a factor.

```
#load the data
library(oibiostat)
data("prevend.samp")

#convert Statin into a factor
prevend.samp$Statin = factor(prevend.samp$Statin, levels = c(0, 1),
                             labels = c("NonUser", "User"))
```

1. Fit a multiple regression model predicting RFFT score from statin use and age. Check the assumptions for multiple linear regression.

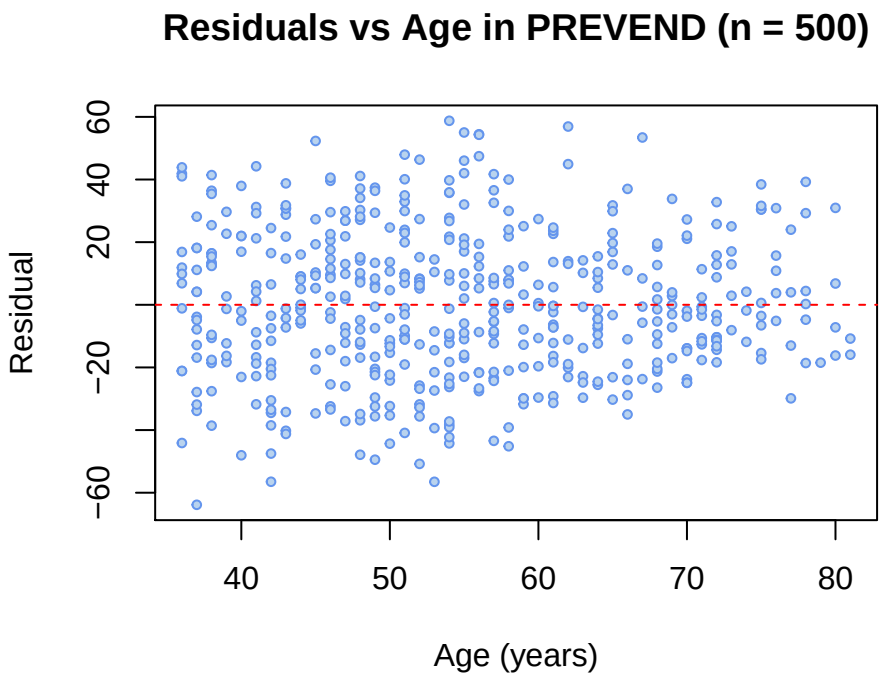
```
#fit a multiple regression model
model1 = lm(RFFT ~ Statin + Age, data = prevend.samp)
```

   a) Assess linearity with respect to age using a scatterplot with residual values on the y-axis and values of age on the x-axis. Is it necessary to assess linearity with respect to statin use?

   There are no apparent trends; the data scatter evenly above and below the horizontal line. There does not seem to be remaining nonlinearity with respect to age after the model is fit.

   It is not necessary to assess linearity with respect to statin use since statin use is measured as a categorical variable. A line drawn through two points (that is, the mean of the two groups defined by a binary variable) is necessarily linear.

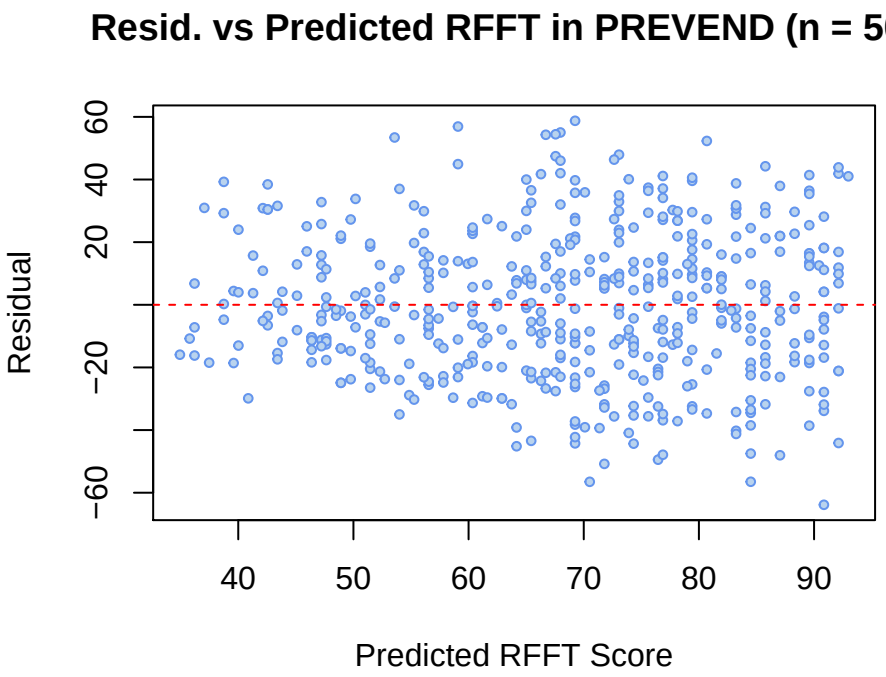
```
#assess linearity
plot(resid(model1) ~ prevend.samp$Age,
 main = "Residuals vs Age in PREVEND (n = 500)",
 xlab = "Age (years)", ylab = "Residual",
 pch = 21, col = "cornflowerblue", bg = "slategray2",
 cex = 0.60)
abline(h = 0, col = "red", lty = 2)
```



b) Assess whether the residuals have approximately constant variance.

The variance of the residuals is somewhat smaller for lower predicted values of RFFT score, but this may simply be an artifact from observing few individuals with relatively low predicted scores. It seems reasonable to assume approximately constant variance.

```
#assess constant variance of residuals
plot(resid(model1) ~ fitted(model1),
 main = "Resid. vs Predicted RFFT in PREVEND (n = 500)",
 xlab = "Predicted RFFT Score", ylab = "Residual",
 pch = 21, col = "cornflowerblue", bg = "slategray2",
 cex = 0.60)
abline(h = 0, col = "red", lty = 2)
```



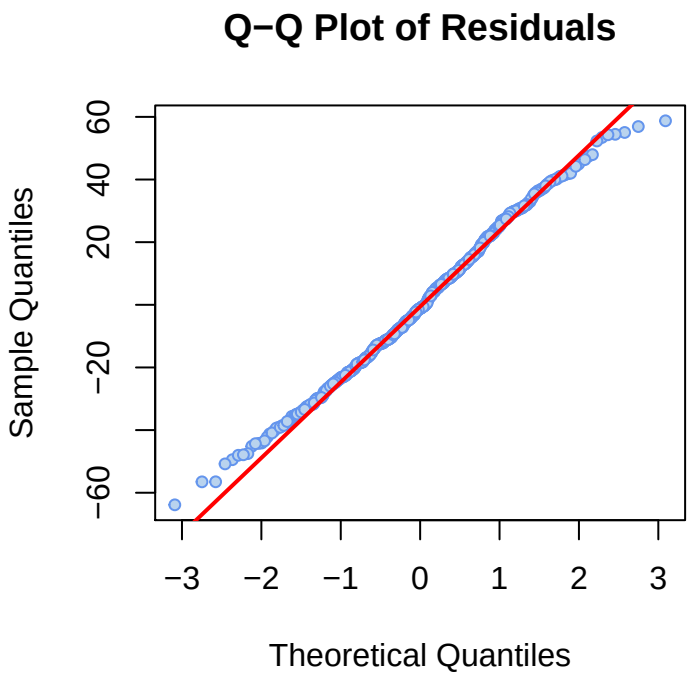
c) Is it reasonable to assume that each set of observations is independent of the others?

As discussed in previous labs using the PREVEND data, it is reasonable to assume that the observations in this dataset are independent. The participants were recruited from a large city in the Netherlands for a study focusing on factors associated with renal and cardiovascular disease.

d) Assess whether the residuals are approximately normally distributed.

The residuals are reasonably normally distributed, with only slight departures from normality in the tails.

```
#assess normality of residuals
qqnorm(resid(model1),
   pch = 21, col = "cornflowerblue", bg = "slategray2", cex = 0.75,
   main = "Q-Q Plot of Residuals")
qqline(resid(model1), col = "red", lwd = 2)
```



2. How well does the model explain the variability in observed RFFT score?

   The R^2 is 0.285; the model explains 28.5% of the observed variation in RFFT score. The moderately low R^2 suggests that the model is missing other predictors of RFFT score.

```
summary(model1)$r.squared
```

```
## [1] 0.2851629
```

The **adjusted** R^2 is computed as

$$R^2_{adj} = 1 - \left(\frac{\text{Var}(e_i)}{\text{Var}(y_i)} \times \frac{n-1}{n-p-1} \right),$$

where n is the number of cases and p is the number of predictor variables.

3. Using the formula, calculate the adjusted R^2 for the multiple regression model predicting RFFT score from statin use and age.

   The adjusted R^2 for the model is 0.282.

```
#use r as a calculator
var.resid = var(resid(model1))
var.y = var(prevend.samp$RFFT)
n = nrow(prevend.samp)
p = 2

1 - ( (var.resid/var.y) * ((n - 1)/(n - p - 1)))
```

```
## [1] 0.2822863
```

```
#check answer with summary(lm( ))$adj.r.squared
summary(lm(model1))$adj.r.squared
```

```
## [1] 0.2822863
```

4. In the previous lab, a multiple regression model was used to estimate the association between statin use and RFFT score while adjusting for age as a potential confounder. Suppose that instead, the goal was to build a model that effectively explains the observed variation in RFFT score; in other words, to build a predictive model for RFFT score. In such a setting, there is no primary predictor of interest.

   The adjusted R^2 is useful as a statistic for comparing models to select a best predictive model. Model selection will be discussed in Chapter 7, Lab 5.

   While R^2 increases with the addition of any predictor to a model, adjusted R^2 only increases substantially when a 'useful' predictor is added; that is, a predictor that contributes to explaining observed variation in the response.

   a) Would you expect the adjusted R^2 for the multiple regression model to be very different from the adjusted R^2 for the simple regression model predicting RFFT score from age? Explain your answer.

      The multiple regression model shows that while age is strongly associated with RFFT score ($p < 0.0001$), statin use is not ($p = 0.74$). Thus, relative to a model that contains age, the multiple regression model that includes statin use will not explain much more variation in RFFT score. Thus, it is reasonable to expect that the adjusted R^2 values will not be very different.

      The adjusted R^2 of a model containing only age is actually slightly higher than the adjusted R^2 of the multiple regression model, at 0.284. If the goal is to build as simple a model as possible that predicts RFFT score well, the model with only age is preferable to the model with age and statin use.

```
#look at coefficients in the multiple regression model
summary(model1)$coef
```

```
##                Estimate Std. Error    t value     Pr(>|t|)
## (Intercept) 137.8822193 5.12208509  26.919158 4.031155e-99
## StatinUser    0.8508699 2.59570660   0.327799 7.432017e-01
## Age          -1.2709945 0.09430356 -13.477693 1.652922e-35
```

```
summary(lm(RFFT ~ Age, data = prevend.samp))$adj.r.squared
```

```
## [1] 0.2835726
```

   b) Would you expect the adjusted R^2 for the multiple regression model to be very different from the adjusted R^2 for the simple regression model predicting RFFT score from statin use? Explain your answer.

      Relative to a model that contains statin use, the multiple regression model that includes age will do much better at explaining variation in RFFT score. Thus, it is reasonable to expect that the adjusted R^2 value for the multiple regression model to be much higher.

This is confirmed below; the adjusted R^2 of the model with only statin use as a predictor is 0.022, which is much lower than 0.282.

Of the three models (statin use only, age only, statin use and age), a model with age as the single predictor of RFFT score most 'efficiently' predicts RFFT score.

```
summary(lm(RFFT ~ Statin, data = prevend.samp))$adj.r.squared
```

```
## [1] 0.02193744
```

Expenditures, race, and age in the dds.discr data

Recall that Chapter 1 introduced a case study examining the evidence for ethnic discrimination in the amount of financial support offered by the State of California to individuals with developmental disabilities. Although an initial look at the data suggested an association between expenditures and ethnicity (specifically between Hispanics and White non-Hispanics), further analysis suggested that age is a confounding variable for the relationship.

The amount of financial support provided to individuals is stored as the `expenditures` variable. The age variable records age in years and the `ethnicity` variable records ethnicity. The data in `dds.discr` represent a random sample of 1,000 individuals who receive financial support from the California Department of Developmental Services (out of a total population of 250,000).

5. Run the following code to load the `dds.discr` data from the `oibiostat` package and subset the data to include only observations from Hispanic and White non-Hispanic consumers.

```
#load the data
data("dds.discr")

#subset the data
dds.subset = dds.discr[dds.discr$ethnicity == "Hispanic" |
                       dds.discr$ethnicity == "White not Hispanic", ]

#drop unused factor levels
dds.subset$ethnicity = droplevels(dds.subset$ethnicity)
```

6. Fit a multiple regression model predicting expenditures from ethnicity and age using the data in `dds.subset` and write the equation of the linear model.

$$\widehat{expenditures} = -3920.07 + 4489.61(ethnicity_{WhitenotHisp}) + 862.48(age)$$

```
#fit the model
model2 <- lm(expenditures ~ ethnicity + age, data = dds.subset)
coef(model2)
```

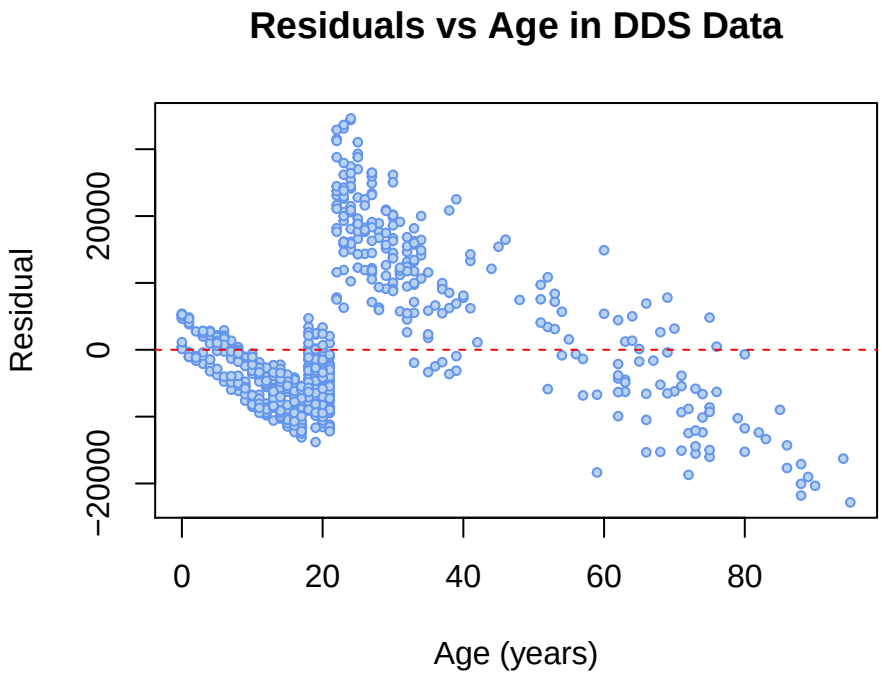
```
##                 (Intercept) ethnicityWhite not Hispanic
##                   -3920.0617                   4489.6128
##                         age
##                    862.4824
```

7. Evaluate the model fit.

   a) Assess linearity with respect to age.

      The residuals clearly show a pattern, rather than random scatter about the $y = 0$ line. There is remaining nonlinearity with respect to age after the model is fit.

```
#evaluate linearity
plot(resid(model2) ~ dds.subset$age,
 main = "Residuals vs Age in DDS Data",
 xlab = "Age (years)", ylab = "Residual",
 pch = 21, col = "cornflowerblue", bg = "slategray2",
 cex = 0.60)
abline(h = 0, col = "red", lty = 2)
```



   b) Assess whether the residuals have approximately constant variance.

      The variance of the residuals is clearly not constant. It is also possible to see in this plot that in the middle range of predicted expenditures, the model consistently under-predicts, while in the upper and lower ranges, the model consistently over-predicts. This is a particularly serious issue with the model fit.

      Refer to the end of these solutions for a further look into why a linear model is not appropriate for modeling the relationship between expenditures, age, and ethnicity.

   c) Is it reasonable to assume that each set of observations is independent of the others?

      Yes, it is reasonable to assume that the values in one set of observation do not influence the values in another set.

d) Assess whether the residuals are approximately normally distributed.

The residuals show marked departures from normality, particularly in the upper tail. There are many more large residuals than expected if the residuals were normally distributed.

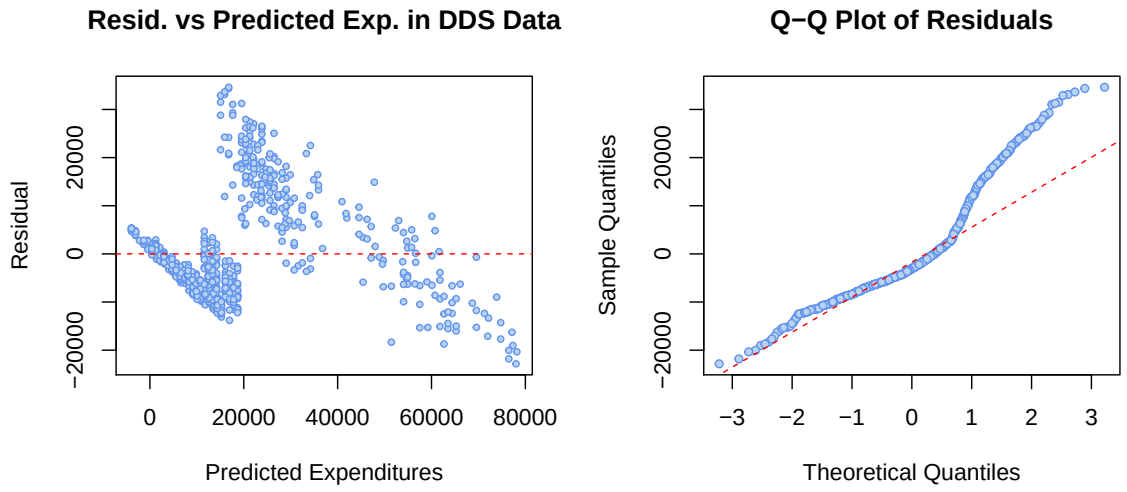
e) Overall, does a linear model seem appropriate for modeling the relationship between expenditures, ethnicity, and age?

A linear regression model is not appropriate for these data.

```
#evaluate constant variance and normality of residuals
par(mfrow = c(1, 2))

plot(resid(model2) ~ fitted(model2),
main = "Resid. vs Predicted Exp. in DDS Data",
 xlab = "Predicted Expenditures", ylab = "Residual",
 pch = 21, col = "cornflowerblue", bg = "slategray2",
 cex = 0.60)
abline(h = 0, col = "red", lty = 2)

qqnorm(resid(model2),
   pch = 21, col = "cornflowerblue", bg = "slategray2", cex = 0.75,
   main = "Q-Q Plot of Residuals")
qqline(resid(model2),
   col = "red", lty = 2)
```

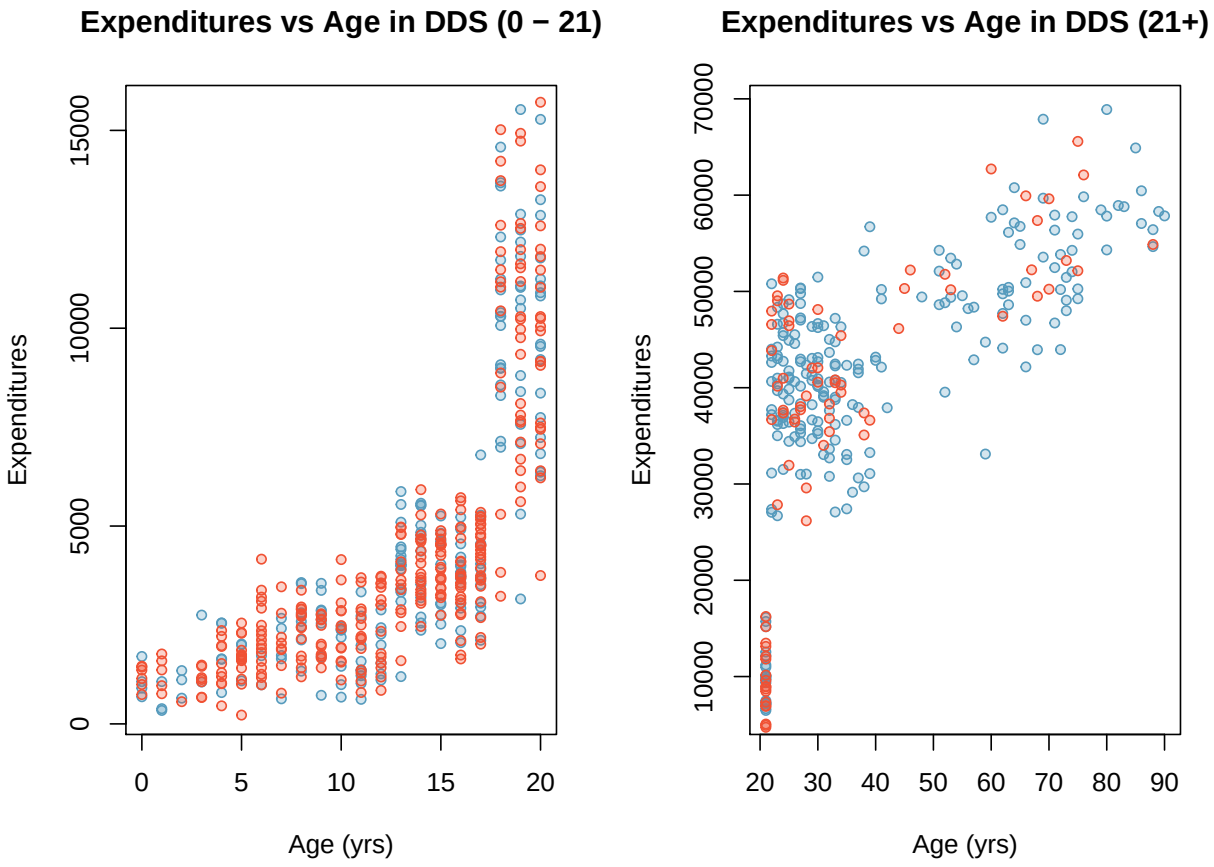


Supplement to Question 7

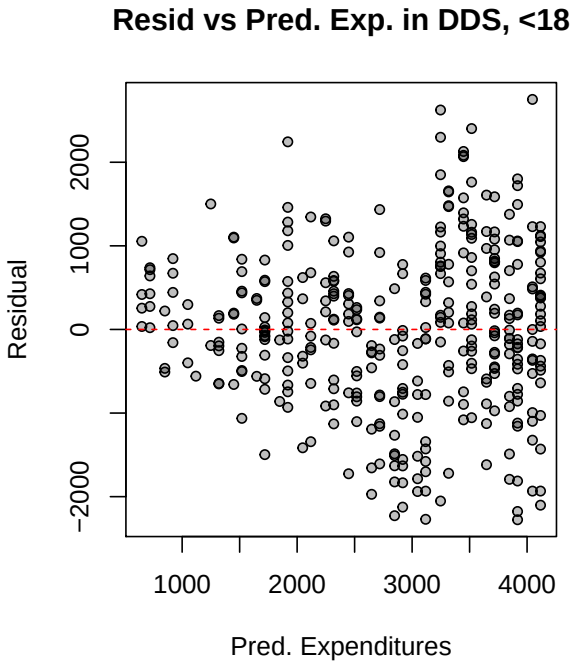
The patterns in the plot of residuals versus age and plot of residual values are very similar; this is explained by the fact that the ordering of the age variable and the predicted expenditures is almost identical. It is not exactly identical because if there are several cases with the same age, the predicted values will vary a bit; also, at a given age, the predicted value will also depend on ethnicity (according to the linear model).

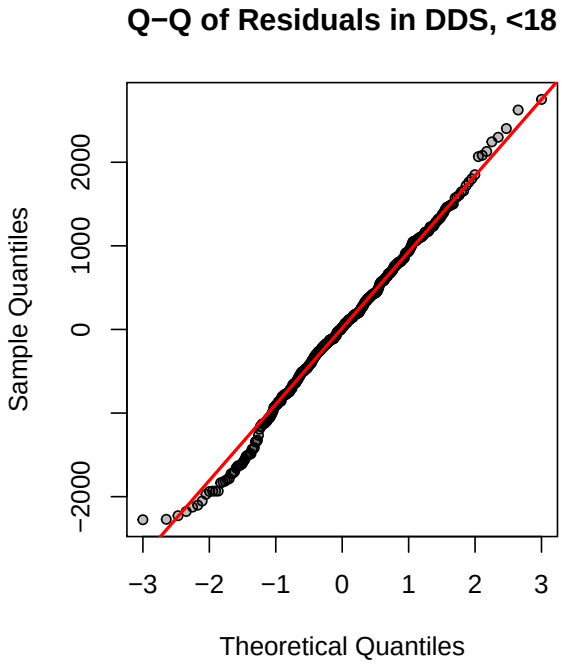
The strange pattern in the first residual plot can be examined further by taking a closer look at the data, and making two plots of expenditures against age: one for ages under 21 and one for ages 21 and older. The scatterplots suggest that the coverage policy determining the amount of financial support differs drastically as individuals transition from adolescence to legal majority age. The plots highlight how the relationship between expenditures and age is certainly not linear, so any conclusions based on the linear model output for either the coefficient of age or ethnicity are unreliable.

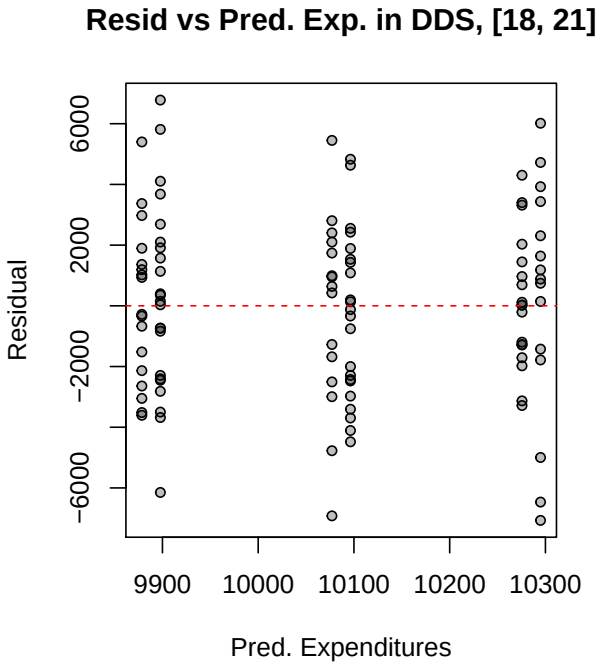
The plots also further support the conclusion from the initial exploratory approach used in Chapter 1—there does not seem to be a systematic difference in the amount of financial support granted to Hispanics versus White non-Hispanics, when adjusting for age. At any given age in the plots, there seems to an even spread of blue points (white non-Hispanics) and red points (Hispanics) across the range of expenditures at that age.

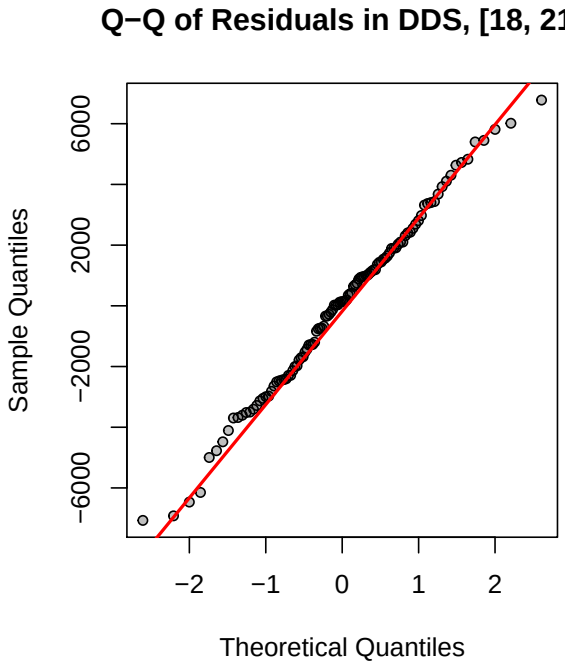


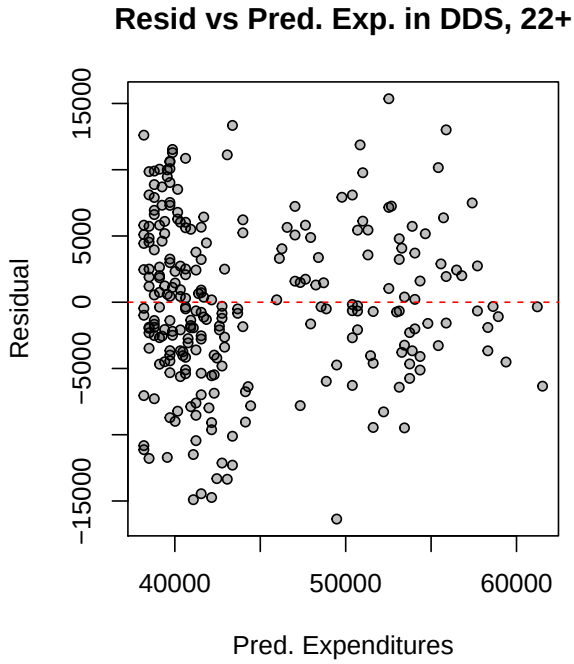
Consider, though, that the relationship between expenditures and age does appear approximately linear for individuals under 18 and for individuals above 21; it may also be approximately linear in the 18-21 age group. The following residual plots are for models predicting expenditures from age and ethnicity, considering only the individuals in a particular age group at a time: under 18 years, between 18 and 21 years (inclusive), and above 21 years.

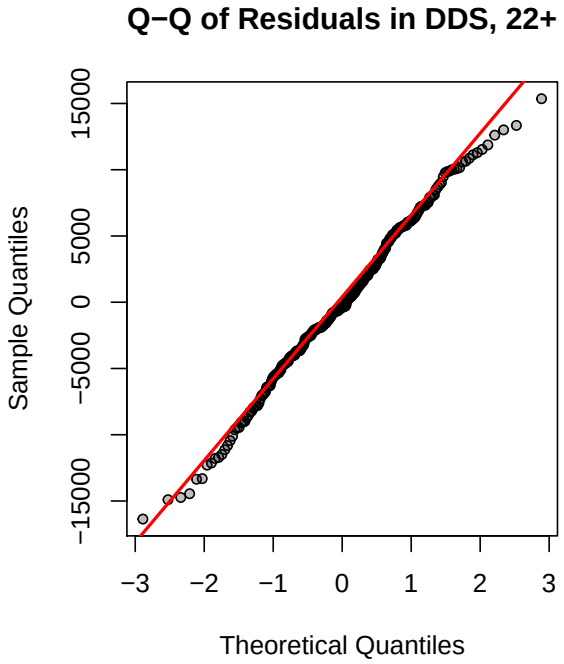












The residual plots indicate that linear regression is appropriate when applied to specific age cohorts at a time. The following tables show the results of t-tests for the model coefficients, showing the regression summary output for the model fit to the youngest participants, the participants in the 18-21 range, and the older participants.[1]

Inference drawn from these models supports the idea that after adjusting for age, there is not evidence of an association between expenditures and ethnicity. In other words, these data are not indicative of ethnic discrimination against Hispanics by the California DDS.

term	estimate	std_error	statistic	p_value	lower_ci	upper_ci
intercept	719.990	126.290	5.701	0.000	471.650	968.330
age[younger]	199.822	10.320	19.363	0.000	179.529	220.114
ethnicity[younger]White not Hispanic	-70.371	103.221	-0.682	0.496	-273.348	132.605

term	estimate	std_error	statistic	p_value	lower_ci	upper_ci
intercept	13868.454	6661.362	2.082	0.040	663.081	27073.827
age[middle]	-198.532	347.586	-0.571	0.569	-887.581	490.518
ethnicity[middle]White not Hispanic	-19.325	566.932	-0.034	0.973	-1143.201	1104.551

term	estimate	std_error	statistic	p_value	lower_ci	upper_ci
intercept	32538.513	1105.150	29.443	0.000	30362.169	34714.857
age[older]	305.098	18.859	16.178	0.000	267.960	342.236
ethnicity[older]White not Hispanic	-1064.835	898.935	-1.185	0.237	-2835.085	705.415

[1] An additional package has been used to format the output from `summary(lm( ))` as a compact table.