

Biostat II: Lab 10, Checking model assumptions for linear regression

Date: 8 May 2008

1. Part II of our last lab consisted of performing linear regression to study the relationship between blood plasma levels of retinol as a function of dietary intake. Today, we'll do some diagnostic plots to assess the validity of our fitted model.

- (a) Reload the same data set as in part II of lab 9:

```
nutrition <- read.table("http://lib.stat.cmu.edu/datasets/Plasma_Retinol",
  skip=30, nrow=315)
```

If you didn't have a chance to read the data description last time, go to the website:

http://lib.stat.cmu.edu/datasets/Plasma_Retinol

- (b) Again, assign names to the `nutrition` data set, so we may refer to the existing variables more conveniently:

```
names(nutrition) <- c("age", "sex", "smokstat", "quetelet",
  "vituse", "calories", "fat", "fiber", "alcohol", "cholesterol",
  "betadiet", "retdiet", "betaplasma", "retplasma")
```

- (c) Make a scatter plot of `nutrition$retplasma` versus `nutrition$retdiet`. Put `nutrition$retplasma` on the y-axis of your plot and `nutrition$retdiet` on the x-axis. Comment on whether or not you these variables appear to be linearly related.
- (d) State the independence assumption for linear regression as it applies to our simple linear regression of plasma retinol on dietary retinol. Although we cannot assess the independence assumption graphically, use your judgement to evaluate whether or not the assumption is valid.
- (e) Perform regression of `retplasma` on `retdiet`:

```
out.ret <- lm(retplasma ~ retdiet, data=nutrition)
summary(out.ret)
```

- (f) Check to see that the residuals are normally distributed around zero. To do so, look at a histogram of residuals from our regression: `hist(out.ret$resid)`. Is it reasonable to assume the errors are normally distributed here?
- (g) Plot the residuals versus predicted values and put a horizontal line at $y = 0$:

- `yhat <- nutrition$retplasma - out.ret$resid`
- `plot(yhat, out.ret$resid)`
- `abline(h=0, lty=3)`

Add a lowess line over our plot to get a better look at the overall trend.

```
lines(lowess(out.ret$resid ~ yhat))
```

Lowess gives us a smoothed look at `out.ret$resid` as a function of `yhat`, by taking averages across a moving window of the predictor. We can think of lowess as a form of non-parametric regression. Do the residuals appear to have the same distribution across the range of predicted values, `yhat`.

- (h) Do you see any outliers in the set of predicted values, `yhat`? Look at the distribution of the predictor itself to see why: `boxplot(nutrition$retdiet)`.
- (i) Summarize your overall judgement of how well the regression assumptions are met here. Remember the LINE acronym, which tells us the regression assumptions: linear relationship, independent, normally distributed errors and equal variance.

2. We also want to study the relationship between blood plasma levels of beta carotene compared to dietary intake of beta carotene. Our first inclination is to perform linear regression of `nutrition$betaplasma` on `nutrition$betadiet`.

Repeat the steps in question 1 for the beta carotene variables (follow the instructions below). Is it reasonable to perform linear regression of `nutrition$betaplasma` on `nutrition$betadiet`?

- (a) Make a scatter plot to assess visually whether it appears that our variables are linearly related:

```
plot(nutrition$betadiet, nutrition$betaplasma)
```

- (b) Perform regression of `betaplasma` on `betadiet`:

```
out.beta <- lm(betaplasma ~ betadiet, data=nutrition)
summary(out.beta)
```

- (c) Check to see that the residuals are normally distributed around zero. To do so, look at a histogram of residuals from our regression: `hist(out.beta$resid)`. Is it reasonable to assume the errors are normally distributed here?

- (d) Plot the residuals versus predicted values and put a horizontal line at $y = 0$:

- `yhat <- nutrition$betaplasma - out.beta$resid`
- `plot(yhat, out.beta$resid)`
- `abline(h=0, lty=3)`

Add a lowess line over our plot to get a better look at the overall trend.

```
lines(lowess(out.beta$resid ~ yhat))
```

Do the residuals appear to have the same distribution across the range of predicted values, `yhat`.

- (e) Do you see any outliers in the set of predicted values, `yhat`?
- (f) Summarize your overall judgement of how well the regression assumptions are met here. Remember the LINE acronym, which tells us the regression assumptions: linear relationship, independent, normally distributed errors and equal variance.

3. In exercise 2, we saw worrisome patterns in the diagnostic plots for regression of `nutrition$betaplasma` on `nutrition$betadiet`. The source of this problem is apparent by looking at the overall distribution of the outcome `nutrition$betaplasma`.

- (a) Make a histogram of `nutrition$betaplasma`. Is this variable normally distributed?

- (b) Now, make a histogram of the log-transformed variable `log(nutrition$betaplasma + 10)`. Is this variable normally distributed?

- (c) Repeat steps 1(c-i) for regression of `log(nutrition$betaplasma + 10)` on `nutrition$betadiet`:

```
out.beta <- lm(log(nutrition$betaplasma + 10) ~ betadiet, data=nutrition)
summary(out.beta)
```

- i. Check to see that the residuals are normally distributed around zero. To do so, look at a histogram of residuals from our regression: `hist(out.beta$resid)`. Is it reasonable to assume the errors are normally distributed here?

- ii. Plot the residuals versus predicted values and put a horizontal line at $y = 0$:

- `yhat <- log(nutrition$betaplasma + 10) - out.beta$resid`
- `plot(yhat, out.beta$resid)`
- `abline(h=0, lty=3)`

Add a lowess line over our plot to get a better look at the overall trend.

```
lines(lowess(out.beta$resid ~ yhat))
```

Do the residuals appear to have the same distribution across the range of predicted values, `yhat`.

- iii. Do you see any outliers in the set of predicted values, `yhat`?
 - iv. Summarize your overall judgement of how well the regression assumptions are met here. Remember the LINE acronym, which tells us the regression assumptions: linear relationship, independent, normally distributed errors and equal variance.
- (d) Are the results of our regression for the log-transformed variable valid?
 - (e) Interpret the coefficient for the `betadiet` term in regression. Is it statistically significant? What is the 95% confidence interval?
 - (f) What is the R^2 value for this regression? Interpret this value in a sentence.
 - (g) Explain why our regression coefficient can be statistically significant even when the R^2 value is relatively small. Discuss the difference between statistical significance and clinical significance.