Notice that now all parameters are estimated, but the standard errors are very large because we cannot estimate them in a stable way. We deliberately caused this problem so we know the cause but in general we need to be able to identify such situations. We do this in Section 5.3.

Exercises

1. The dataset `teengamb` concerns a study of teenage gambling in Britain. Fit a regression model with the expenditure on gambling as the response and the sex, status, income and verbal score as predictors. Present the output.
   (a) What percentage of variation in the response is explained by these predictors?
   (b) Which observation has the largest (positive) residual? Give the case number.
   (c) Compute the mean and median of the residuals.
   (d) Compute the correlation of the residuals with the fitted values.
   (e) Compute the correlation of the residuals with the income.
   (f) For all other predictors held constant, what would be the difference in predicted expenditure on gambling for a male compared to a female?
2. The dataset `uswages` is drawn as a sample from the Current Population Survey in 1988. Fit a model with weekly wages as the response and years of education and experience as predictors. Report and give a simple interpretation to the regression coefficient for years of education. Now fit the same model but with logged weekly wages. Give an interpretation to the regression coefficient for years of education. Which interpretation is more natural?
3. In this question, we investigate the relative merits of methods for computing the coefficients. Generate some artificial data by:

```
> x <- 1:20
> y <- x+rnorm(20)
```

   Fit a polynomial in `x` for predicting `y`. Compute $\hat{\beta}$ in two ways — by `lm()` and by using the direct calculation described in the chapter. At what degree of polynomial does the direct calculation method fail? (Note the need for the `I()` function in fitting the polynomial, that is, `lm(y ~ x + I(x^2))`.
4. The dataset `prostate` comes from a study on 97 men with prostate cancer who were due to receive a radical prostatectomy. Fit a model with `lpsa` as the response and `lcavol` as the predictor. Record the residual standard error and the R^2. Now add `lweight`, `svi`, `lpph`, `age`, `lcp`, `pgg45` and `gleason` to the model one at a time. For each model record the residual standard error and the R^2. Plot the trends in these two statistics.
5. Using the `prostate` data, plot `lpsa` against `lcavol`. Fit the regressions of `lpsa` on `lcavol` and `lcavol` on `lpsa`. Display both regression lines on the plot. At what point do the two lines intersect?