```
Residual standard error: 0.193 on 118 degrees of freedom
Multiple R-Squared: 0.702,      Adjusted R-squared: 0.69
F-statistic: 55.7 on 5 and 118 DF,  p-value: <2e-16
```

Notice that the R^2 is higher for this model, but the p-values are similar. Because of the log transformation, we can interpret the coefficients as having a multiplicative effect:

```
> exp(coef(gt)[3:6])
activityisolated    activityone    activitylow    activityhigh
         1.05311        0.88350        1.09189         0.65754
```

Compared to the reference level, we see that the high sexual activity group has 0.66 times the life span (i.e, 34% less).

Why did we include `thorax` in the model? Its effect on longevity was known, but because of the random assignment of the flies to the groups, this variable will not bias the estimates of the effects of the activities. We can verify that `thorax` is unrelated to the activities:

```
> gh <- lm(thorax ~ activity, fruitfly)
> anova(gh)
Analysis of Variance Table

Response: thorax
           Df Sum Sq Mean Sq F value Pr(>F)
activity    4  0.026   0.006    1.11   0.36
Residuals 119  0.685   0.006
```

However, look what happens if we omit `thorax` from the model for `longevity`:

```
> gu <- lm(log(longevity) ~ activity, fruitfly)
> summary(gu)
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)       4.1193     0.0564   72.99  < 2e-16
activityisolated  0.0234     0.0798    0.29     0.77
activityone      -0.1195     0.0798   -1.50     0.14
activitylow       0.0240     0.0806    0.30     0.77
activityhigh     -0.5172     0.0798   -6.48  2.2e-09

Residual standard error: 0.282 on 119 degrees of freedom
Multiple R-Squared: 0.359,      Adjusted R-squared: 0.338
F-statistic: 16.7 on 4 and 119 DF,  p-value: 6.96e-11
```

The magnitude of the effects do not change that much but the standard errors are substantially larger. The value of including `thorax` in this model is to increase the precision of the estimates.

Exercises

1. Using the `teengamb` data, model `gamble` as the response and the other variables as predictors. Take care to investigate the possibility of interactions between `sex` and the other predictors. Interpret your final model.

(e) Suppose we change the cutoff to 0.9 so that $p < 0.9$ is classified as malignant and $p > 0.9$ as benign. Compute the number of errors in this case. Discuss the issues in determining the cutoff.

(f) It is usually misleading to use the same data to fit a model and test its predictive ability. To investigate this, split the data into two parts — assign every third observation to a test set and the remaining two thirds of the data to a training set. Use the training set to determine the model and the test set to assess its predictive performance. Compare the outcome to the previously obtained results.

3. The National Institute of Diabetes and Digestive and Kidney Diseases conducted a study on 768 adult female Pima Indians living near Phoenix. The purpose of the study was to investigate factors related to diabetes. The data may be found in the the dataset `pima`.

(a) Perform simple graphical and numerical summaries of the data. Can you find any obvious irregularities in the data? If you do, take appropriate steps to correct the problems.

(b) Fit a model with the result of the diabetes test as the response and all the other variables as predictors. Can you tell whether this model fits the data?

(c) What is the difference in the odds of testing positive for diabetes for a woman with a BMI at the first quartile compared with a woman at the third quartile, assuming that all other factors are held constant? Give a confidence interval for this difference.

(d) Do women who test positive have higher diastolic blood pressures? Is the diastolic blood pressure significant in the regression model? Explain the distinction between the two questions and discuss why the answers are only apparently contradictory.

(e) Perform diagnostics on the regression model, reporting any potential violations and any suggested improvements to the model.

(f) Predict the outcome for a woman with predictor values 1, 99, 64, 22, 76, 27, 0.25, 25 (same order as in the dataset). Give a confidence interval for your prediction.

4. Aflatoxin B1 was fed to lab animals at various doses and the number responding with liver cancer recorded. The data may be found in the dataset `aflatoxin`.

(a) Build a model to predict the occurrence of liver cancer. Compute the ED50 level.

(b) Discuss the extrapolation properties of your chosen model for low doses.

5. A study was conducted to determine the effectiveness of a new teaching method in economics. The data may be found in the dataset `spector`. Write a report on how well the new method works.

6. Incubation temperature can affect the sex of turtles. An experiment was conducted with three independent replicates for each temperature and the number of male

```
          Theta:  4.397
      Std. Err.:  0.495

 2 x log-likelihood:  -3645.309
```

We see that $\hat{k} = 4.397$ with a standard error of 0.495. We can compare negative binomial models using the usual inferential techniques.

Further Reading: See books by Cameron and Trivedi (1998) and Agresti (2002).

Exercises

1. The dataset `discoveries` lists the numbers of "great" inventions and scientific discoveries in each year from 1860 to 1959. Has the discovery rate remained constant over time?
2. The `salmonella` data was collected in a salmonella reverse mutagenicity assay. The predictor is the dose level of quinoline and the response is the numbers of revertant colonies of TA98 salmonella observed on each of three replicate plates. Show that a Poisson GLM is inadequate and that some overdispersion must be allowed for. Do not forget to check out other reasons for a high deviance.
3. The `ships` dataset found in the `MASS` package gives the number of damage incidents and aggregate months of service for different types of ships broken down by year of construction and period of operation. Develop a model for the rate of incidents, describing the effect of the important predictors.
4. The dataset `africa` gives information about the number of military coups in sub-Saharan Africa and various political and geographical information. Develop a simple but well-fitting model for the number of coups. Give an interpretation of the effect of the variables you include in your model on the response.
5. The `dvisits` data comes from the Australian Health Survey of 1977–78 and consist of 5190 single adults where young and old have been oversampled.
   (a) Build a Poisson regression model with `doctorco` as the response and `sex`, `age`, `agesq`, `income`, `levyplus`, `freepoor`, `freerepa`, `illness`, `actdays`, `hscore`, `chcond1` and `chcond2` as possible predictor variables. Considering the deviance of this model, does this model fit the data?
   (b) Plot the residuals and the fitted values — why are there lines of observations on the plot?
   (c) Use backward elimination with a critical p-value of 5% to reduce the model as much as possible. Report your model.
   (d) What sort of person would be predicted to visit the doctor the most under your selected model?
   (e) For the last person in the dataset, compute the predicted probability distribution for their visits to the doctor, i.e., give the probability they visit 0, 1, 2, etc. times.
   (f) Fit a comparable (Gaussian) linear model and graphically compare the fits. Describe how they differ.
6. Components are attached to an electronic circuit card assembly by a wave-soldering process. The soldering process involves baking and preheating the circuit card and then passing it through a solder wave by conveyor. Defects arise during the process. The design is 2^{7-3} with three replicates. The data is presented in the dataset `wavesolder`.

   Assuming that the replicates are independent, analyze the data. Write a report on the analysis that summarizes the substantive conclusions and includes the highlights of your analysis.
7. The dataset `esdcomp` was recorded on 44 doctors working in an emergency service at a hospital to study the factors affecting the number of complaints received. Build a model for the number of complaints received and write a report on your conclusions.