Session 7, Part 1 - Robust SEs and Analysis of Categorical Data

Brian Leroux

Wednesday, February 27, 2019

Outline

1. Robust Standard Errors
2. Analysis of Categorical Data
3. Logistic Regression

1. Robust Standard Errors

- The concept of robustness
- Robust (“sandwich”) standard errors for linear regression

The concept of robustness

Robustness can have several different meanings in Statistics. Here we will use the most general meaning: methods are *robust* if they are approximately valid even if certain assumptions do not hold. For example, a test for comparing means is robust against unequal variances if it provides a valid test even if the variances are unequal.

Types of robustness correspond to the 4 different types of assumptions for regression analysis, as detailed below. Some robust methods are applicable to departures from specific assumptions, whereas others are applicable to more than one type of assumption.

Types of Robustness

1. Robustness against misspecified models for the mean response:

If we fit a linear regression model (e.g., $E[Y] = \alpha + \beta X$) using least squares but this model is not true then our estimate of β might still be meaningful but can be misleading, depending on the degree of departure from linearity and the goal of the analysis. Methods that are explicitly robust to misspecifying the form of a regression model include non-parametric regression (“scatterplot smoothing”). A smoother can provide a useful summary of the association between X and Y (but does not provide an easily interpretable summary of the association as a regression coefficient does).

2. Robustness against dependence

All of the methods we have seen assume independence of observations. This is a critical assumption. If observations are not independent, specialized methods that accommodate non-independent observations must be used, e.g., Generalized Estimating Equations and random effects models. The bootstrap and jackknife can also be applied to account for non-independence in certain situations. (We will introduce some of these methods next week.)

Types of robustness (continued)

3. Robustness against misspecification of the constant variance assumption

For linear regression with non-constant variance, we can use “robust” (or “sandwich”) standard errors to get valid inference. (We will see in Part 2 that these also apply to Generalized Linear Models.) The bootstrap and jackknife can also be used in these situations.

4. Robustness against small samples

Most statistical methods require large sample sizes to be valid. One exception is the case of linear regression where we can get away with small sample sizes if we can assume that the errors are normally distributed (and constant variance). Note that t-tests, ANOVA, and least-squares regression are fairly robust against non-normality for moderate sample sizes due to the CLT. For non-normal errors, there are special methods that give valid results for small samples, including exact methods and permutation methods.

Robust standard errors

Robust standard errors are used for linear regression when we cannot assume constant variance (they are also used for GLMs - see Part 2).

The robust SEs are based on a specialized formula that avoids the constant variance assumption. Recall

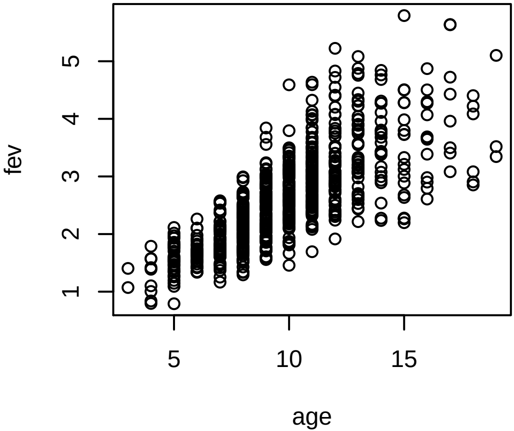
$$\text{cov}(\underline{\hat{\beta}}) = (\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{X}'\text{cov}(\underline{\mathbf{Y}})\mathbf{X}(\mathbf{X}'\mathbf{X})^{-\mathbf{1}}$$

Under the constant variance assumption, this simplies to $\sigma^2(\mathbf{X}'\mathbf{X})^{-\mathbf{1}}$. The robuse SEs are obtained by using the "empirical" covariance for the residuals in place of cov($\underline{Y}$) (this is the "meat" part of the sandwich).

In R robust SEs can be obtained using the function 'vcovHC' in the "sandwich" package. Note that "HC" stands for "heteroscedasticity consistent."

Example: the FEV data

The FEV data exhibits heteroscedasticity.



```
             Estimate  Std. Error   t value      Pr(>|t|)
(Intercept) 0.4316481 0.077895393  5.541382  4.35888e-08
age         0.2220410 0.007518462 29.532766 2.45396e-122
```

The Robust Variance-Covariance Matrix

```
round(vcovHC(model),6)
```

```
            (Intercept)       age
(Intercept)    0.006473 -0.000709
age           -0.000709  0.000082
```

The output from vcovHC is the estimated variance-covariance matrix of variances and covariances of the parameter estimates.

Calculating the robust SEs

The diagonal elements of the variance-covariance matrix are the variances of the coefficients, so their square-roots are the SEs.

Let's compare them to the standard SEs from the lm function.

```r
v <- vcovHC(model)
robust.se <- sqrt(diag(v))
round(cbind(summary(model)$coef,robust.se),4)
```

```
            Estimate Std. Error t value Pr(>|t|) robust.se
(Intercept)   0.4316     0.0779  5.5414        0    0.0805
age           0.2220     0.0075 29.5328        0    0.0091
```

The robust SEs are larger than the standard SEs, particularly for the age coefficient. In general, robust SEs may be larger or smaller (or about the same), compared to the standard ones.

Comparison for a larger model

```r
model2 <- lm(fev ~ age + ht + male + smoke, data=d)
round(cbind(summary(model2)$coef,robust.se=sqrt(diag(vcovHC(model2)))),4)
```

```
            Estimate Std. Error  t value Pr(>|t|) robust.se
(Intercept)  -4.4570     0.2228 -20.0008   0.0000    0.2412
age           0.0655     0.0095   6.9040   0.0000    0.0105
ht            0.1042     0.0048  21.9011   0.0000    0.0052
male          0.1571     0.0332   4.7310   0.0000    0.0322
smoke        -0.0872     0.0593  -1.4724   0.1414    0.0778
```

The robust SE is smaller than the standard SE for "male" but larger for other terms, especially "smoke".

Robust SEs for the Sales data

Fit a linear model to the Sales data and calculate robust SEs.

```r
s=read.csv("Sales.csv")
fit=lm(LAST_SALE_PRICE ~ SQFT, data=s)
v <- vcovHC(fit)
robust.se <- sqrt(diag(v))
round(cbind(summary(fit)$coef,robust.se),4)
```

```
              Estimate Std. Error t value Pr(>|t|)  robust.se
(Intercept) -13574.815 11452.8601 -1.1853    0.236 22753.1774
SQFT           340.383     4.7936 71.0078    0.000    11.6986
```

Note that the robust SE for the coefficient of SQFT is quite similar to the jackknife and bootstrap estimates of the SEs.

The Robust Wald Test

The t-statistic for a regression coefficient (estimate divided by its SE) can be used evaluate statistical significance of a regression coefficient. It is sometimes expressed as the square of the t-statistic and referred to an F distribution (or chi-squared for large sample sizes. This test is called a **Wald** test. If the robust SEs are used in place of the standard SEs then we call it a **robust Wald test**.

There are also robust Wald Tests for composite hypotheses in linear regression that can be used in place of the F-test, when model assumptions about the variance do not hold.

When to use robust SEs

For large sample sizes (e.g., the FEV data or Sales data) we usually use robust SEs. We should still examine the residuals because we will learn a lot about our data and the fit of our model.

If we are confident about the homoscedasticity (constant variance) assumption, we can use the usual SEs. For small sample sizes, they can be more accurate than the robust SEs.

Generalized Estimating Equations (GEE)

GEE is a very powerful method for fitting linear regression and generalized linear models that can accommodate non-independence.

A more general type of robust (“sandwich”) SEs are available with GEE that yield valid inference for certain types of correlated data structures as long as the sample size is sufficiently large.

2. Introduction to Analysis of Categorical Data

- Z-test and chi-squared test
- Confidence intervals for a difference between proportions
- Continuity correction
- Fisher's exact test

Example: Randomized Clinical Trial

In a clinical trial, we might wish to compare the proportions of patients who survive when given a new treatment compared with an older treatment. We would want to test the null hypothesis that the probability of surviving is the same under the two treatments versus the alternative hypothesis that the survival probabilities are not equal.

Note: even if the new treatment is thought to be better than the old, we typically use a 2-sided alternative hypothesis because we want to be able to detect if the new treatment is in fact inferior (perhaps due to unsuspected side effects).

Randomly assinging patients to one or the other treatment will ensure an unbiased estimate of the difference in survival probabilities and allow us to make a valid test of the hypothesis.

Observational Medical Studies

If it is not feasible to randomly assign treatments to patients, then we could perform an observational study by comparing results for those patients who choose to receive the new treatment versus those who choose the old treatment. This suffers from bias due to the fact that patients who choose one treatment are likely different than patients who choose the other treatment in ways that may affect their outcome (for example, sicker patients might choose the new treatment).

Epidemiological studies of health effects of environmental or dietary exposures are examples of this type of study. For example, we might compare incidence rates of cancer between people with high versus low fat diets. The possibility of bias in such a study is very high. Special regression methods have been developed to try to control such bias, including *logistic regression* and *Generalized Linear Models*.

Example: A/B Testing with Dichotomous Response

In A/B testing we might record whether or not a visitor to a web site follows a certain link.

We would like to test the null hypothesis that the proportion of users who select the link is the same in the two versions of the site.

To design the experiment we need a test of the null hypothesis that has a fixed type I error and with adequate power to detect a meaningful alternative hypothesis.

Comparison of Proportions from Two Independent Samples

We need a test procedure that is analogous to the 2-sample t-test to compare two population proportions from independent samples. The special nature of the data (a binary variable that takes values 0 or 1) means that we have to modify our approach for comparing means.

Suppose that we have two independent random samples of binary data with population proportions p_A and p_B. We want to test $H_0 : p_A = p_B$ against the alternative hypothesis $H_1 : p_A \neq p_B$.

We base our test on the difference between the sample proportions $\hat{p}_A - \hat{p}_B$, where the sampling proportions are the proportions of 1s in each sample. Thus, we need to determine the sampling distribution of this statistic.

The Sampling Distribution of the Difference Between Two Sample Proportions

We know from before that each of the sample proportions has expected value equal to its respective population proportion, i.e.,

$$E(\hat{p}_A) = p_A,$$

and variance equal to

$$\mathrm{var}(\hat{p}_A) = p_A(1 - p_A)/n,$$

and similarly for $\hat{p}_B$.

As before, using the properties of expected values and variances of sums of independent random variables, we get

$$E(\hat{p}_A - \hat{p}_B) = p_A - p_B.$$

and

$$\mathrm{var}(\hat{p}_A - \hat{p}_B) = p_A(1 - p_A)/n_A + p_B(1 - p_B)/n_B.$$

Using the Pooled Sample Proportion

So far, things are following the same pattern as for the t-test for comparison of means. However, the estimation of the variance of the difference between sample proportions has a special form compared to what we had for the difference between sample means. Because of the special $p(1-p)$ form of the variance of a binary variable, we don't have any σ^2 parameters to deal with. This means that *under the null hypothesis*, which says that $p_A = p_B$, the variance formula takes the special form

$$\text{var}(\hat{p}_A - \hat{p}_B) = p(1-p)(1/n_A + 1/n_B),$$

where p is the common (under H_0) value of p_A and p_B.

This means that we must estimate an assumed common value of p. This is done simply by calculating the overall proportion of 1s in the pooled sample, i.e.,

$$\hat{p} = \frac{n_A\hat{p}_A + n_B\hat{p}_B}{n_A + n_B}$$

Thus, we have our estimate of the variance of the difference between the sample proportions equal to $\hat{p}(1-\hat{p})(1/n_A + 1/n_B)$, which implies that the estimated SE of the difference between proportions is

$$SE(\hat{p}_A - \hat{p}_B) = \sqrt{\hat{p}(1-\hat{p})(1/n_A + 1/n_B)}$$

The Test Statistic for Comparison of Proportions

The test statistic follows the general recipe we saw before:

$$\text{Test Statistic} \;=\; \frac{\text{Sample Estimate} - \text{Hypothesized Value}}{\text{SE of the Sample Estimate}}$$

$$Z = \frac{\hat{p}_A - \hat{p}_B}{\sqrt{\hat{p}(1-\hat{p})(1/n_A + 1/n_B)}}.$$

The rules for interpreting Z are similar to those for the comparison of sample means, except that there is no consideration of normality for the population.

Interpretation of Z

Summary of Tests for Comparison of Two Population Means from Independent Samples

We have two *independent* random samples from populations with proportions p_A and p_B. We want to test $H_0 : p_A = p_B$ versus $H_1 : p_A \neq p_B$.

1. Calculate the test statistic:

$$Z = \frac{\hat{p}_A - \hat{p}_B}{\sqrt{\hat{p}(1-\hat{p})(1/n_A + 1/n_B)}}.$$

2. Calculate the critical value for the desired significance level (α) using the standard normal distribution, (e.g., 1.96) for $\alpha = 0.05$.

This test requires large sample sizes. We call it the large-sample Z-test for comparing proportions.

Example: A Randomized Clinical Trial

Suppose that an experiment was done to compare two different types of CPR for victims of cardiac arrest and had the following results: 1000 patients were randomized to treatment A, of whom 100 survived, and 1050 were randomized to treatment B, of whom 90 survived. Is there evidence that the probability of survival is different for the two treatments?

```r
n = c(1000,1050)
y = c(100,90)
phat=y/n
diff=diff(phat)
pooled.p=sum(n*phat)/sum(n)
se=sqrt(pooled.p*(1-pooled.p)*sum(1/n))
z=diff/se
data.frame(phat[1],phat[2],diff,se,z,p=2*(1-pnorm(abs(z))))
```

```
  phat.1.    phat.2.        diff         se         z         p
1     0.1 0.08571429 -0.01428571 0.01281332 -1.114911 0.2648885
```

There is not evidence that the two survival probabilities are different.

95% confidence interval for the difference between survival probabilities

```
lower = diff-1.96*se
upper = diff+1.96*se
lower
```

```
[1] -0.03939982
```

```
upper
```

```
[1] 0.01082839
```

Comparing proportions with the R function 'prop.test'

This function is based on the cross-tab of treatment group and survival.

```
cross.tab=data.frame(rbind(y,n-y))
names(cross.tab)=c("A","B")
row.names(cross.tab)=c("survived","died")
cross.tab
```

```
           A   B
survived 100  90
died     900 960
```

```
prop.test(cbind(y,n-y),correct=F)
```

```
    2-sample test for equality of proportions without continuity
    correction

data:  cbind(y, n - y)
X-squared = 1.243, df = 1, p-value = 0.2649
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.01086266  0.03943409
sample estimates:
    prop 1     prop 2 
0.10000000 0.08571429 
```

The Chi-Squared Test

The chi-squared test is a general test procedure for testing hypotheses with categorical data. In the simple case of comparing proportions the chi-squared test is equivalent to the Z test for comparing proportions.

```
chisq.test(cbind(y,n-y),correct=F)
```

```

	Pearson's Chi-squared test

data:  cbind(y, n - y)
X-squared = 1.243, df = 1, p-value = 0.2649
```

Note that the p-values for the two tests are the same. The test statistic “X-squared” is equal to the square of the Z statistic. The chi-squared test is useful for more general types of hypothesis testing, such as comparison of 3 or more population proportions.

Yates's continuity correction

This is a correction that can improve the normal approximation to the distribution and result in a more accurate p-value with small sample sizes.

```r
prop.test(cbind(y,n-y),correct=T)
```

```
    2-sample test for equality of proportions with continuity
    correction

data:  cbind(y, n - y)
X-squared = 1.079, df = 1, p-value = 0.2989
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.01183885  0.04041028
sample estimates:
    prop 1     prop 2 
0.10000000 0.08571429 
```

The correction can be applied to the chi-squared test as well.

An alternative to this correction is to apply Fisher's exact test which gives an exact p-value.

Fisher's Exact Test

Fisher's exact test is based on the *hypergeometric* distribution for contingency tables. It provides an **exact test** in the sense that the p-value is the (exact) probability of getting a result as or more extreme than the one observed under the null hypothesis (not an approximate probability as with most test procedures).[1]

Sometimes simulation is used to get a "close-enough" approximation to the exact p-value.

```r
chisq.test(cbind(y,n-y),simulate.p.value=TRUE)
```

```
    Pearson's Chi-squared test with simulated p-value (based on 2000
    replicates)

data:  cbind(y, n - y)
X-squared = 1.243, df = NA, p-value = 0.2849
```

In many practical situations (i.e., with moderately large sample sizes) all of the different tests described above give very similar results. By convention, the chi-squared test is the most commonly reported test.

[1] Just like the p-value for the coin tossing experiment.

3. Logistic Regression

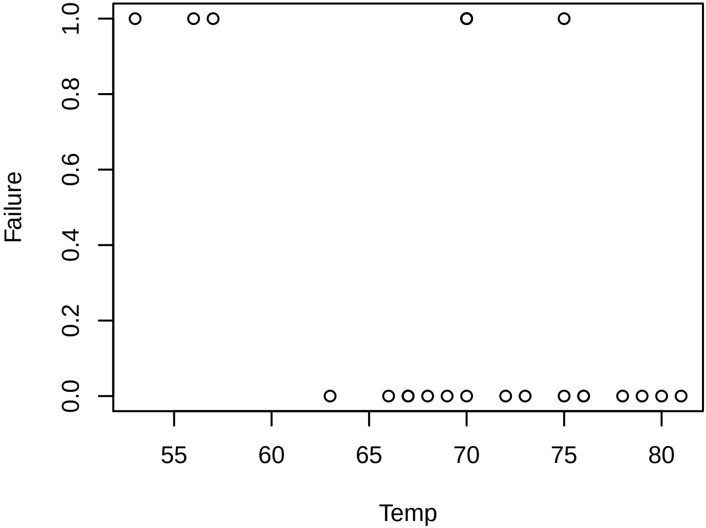
- Logistic regression model
- Odds-ratios
- Confidence interval and hypothesis test for the odds-ratio

The Logistic Regression Model

Logistic regression is a useful method for analyzing binary responses. Other methods, including linear regression and *Poisson regression* (see part 2) might be reasonable to apply to binary responses (remember the CLT!). However, logistic regression has an important advantage in the fact that it takes into account the constraint on a probability to be between 0 and 1.

The O-ring data

For example, the "O-ring" data lists the indicator of O-ring failure on 24 launches or tests prior to the 1986 Challenger Space Shuttle disaster.



A logistic regression model for the O-ring data

We would like to have a model that describes how the probability of O-ring failure depends on temperature. Ideally, this model would give a predicted probability of failure that was between 0 and 1 for any input value of temperature. The logistic regression model does that by using the "logit" transform:

$$\log \frac{p}{1-p} = \beta_0 + \beta_1 X,$$

where X is the temperature and $\log p/(1-p)$ is called the logit (log-odds) of p. To see that p must be in the interval (0,1), we derive the inverse function of the logit function:

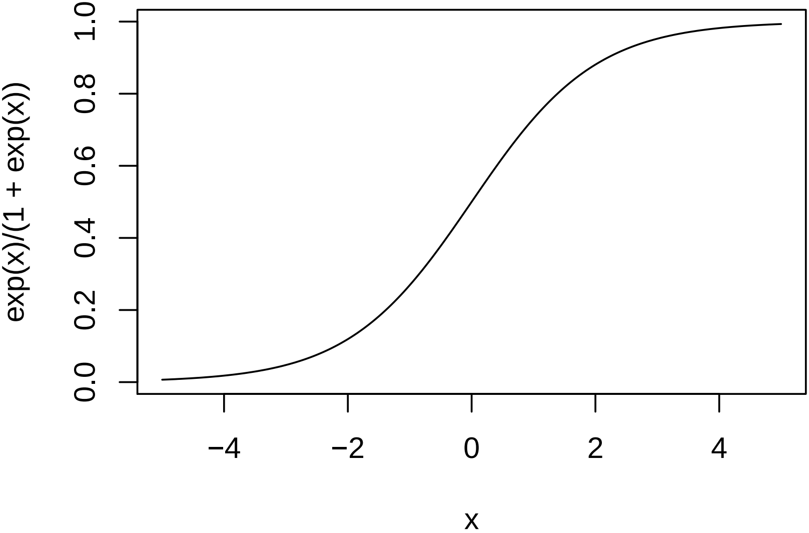
$$\log \frac{p}{1-p} = x \ \rightarrow \ \frac{p}{1-p} = e^x \ \rightarrow p = \frac{e^x}{1+e^x}.$$

If g is the logit function then its inverse is g^{-1} is defined as $g^{-1}(x) = e^x/(1+e^x)$. This function is called the "expit" function.

Graph of the expit function

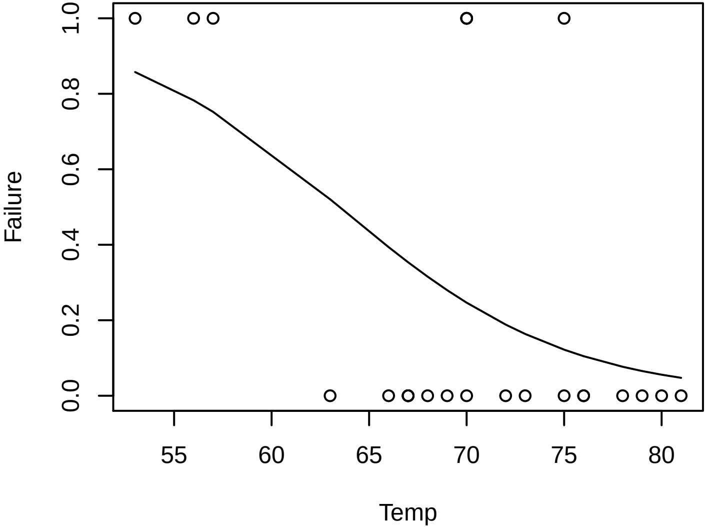
The value of the expit function always lies in (0,1).

```r
x=((-100):100)/20
plot(x,exp(x)/(1+exp(x)),type="l")
```



The fitted logistic regression model for the O-ring data

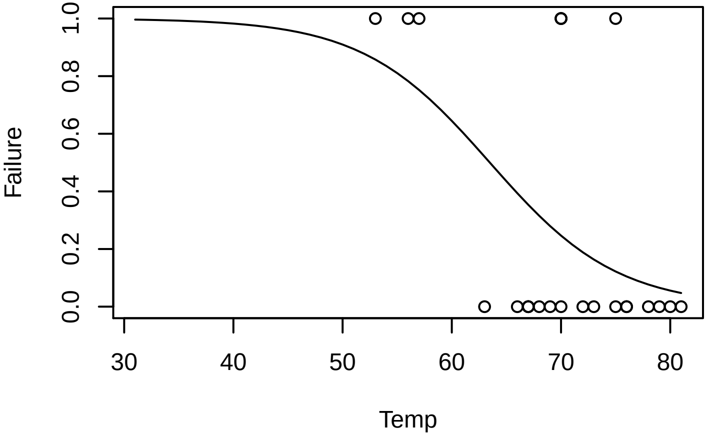
```
plot(Failure ~ Temp,data=o)
fit=glm(Failure ~ Temp, data=o, family=binomial)
p=fit$fitted.values
lines(o$Temp,fit$fitted.values)
```



Extrapolating the logistic regression model to lower temperatures

Recognizing that extrapolation beyond the range of the data is dangerous (never mind the small sample size), it is helpful in this case as a "what-if" pre-cautionary exercise. The temperature on the data of the Challenger explosion was 31 degrees.

```
plot(Failure ~ Temp,data=o,xlim=c(31,max(o$Temp)))
p=predict(fit,newdata=data.frame(Temp=31:81),type="response")
lines(31:81,p)
```



Interpretations of logistic regression coefficients

By analogy with the interpretation of parameters in linear regression,

β_0 is logit(μ) where μ is the mean response (probability) when all predictors are set to 0, i.e., $X_1 = X_2 = \cdots = X_p = 0$.

Translating to the "raw" scale of the responses (i.e., to probabilities),

$$\beta_0 = \text{logit}(\mu) \ \rightarrow \mu = \text{expit}(\beta_0) = \frac{e^{\beta_0}}{1 + e^{\beta_0}}.$$

Therefore, $e^{\beta_0}/(1 + e^{\beta_0})$ is the mean response (probability) when all X's are equal to 0.

The fitted model for the O-ring data

```r
summary(glm(Failure ~ Temp, data=o, family=binomial))$coef
```

```
              Estimate Std. Error   z value   Pr(>|z|)
(Intercept) 10.8753491 5.70291251  1.906982 0.05652297
Temp        -0.1713205 0.08343865 -2.053251 0.04004824
```

For the O-ring data the estimated intercept is 10.88. The expit value of this is essentially equal to 1.

```r
exp(10.88)/(1+exp(10.88))
```

```
[1] 0.9999812
```

Note: the default SEs from the *glm* function are valid for logistic regression with a binary response as long as the sample size is sufficiently large (not true for this example!).

The Odds-Ratio

In the logistic regression model for the O-ring failure, β_1 is the difference in values of logit(μ) per unit difference in temperature:

$$\beta_1 = \log \frac{\mu(X+1)}{1-\mu(X+1)} - \log \frac{\mu(X)}{1-\mu(X)}$$

To put this on the response scale, we exponentiate both sides:

$$e^{\beta_1} = \frac{\mu(X+1)}{1-\mu(X+1)} \Big/ \frac{\mu(X)}{1-\mu(X)}.$$

This is the ratio of the odds of failure at temperature $X+1$ to the odds of failure at temperature X.

For the O-ring data, the estimate of β_1 is -0.1713 so $\exp(\hat{\beta}_1)$ is 0.84. This means that the odds of failure are lower by 16% for each additional one degree in temperature (or speaking causally, the odds of failure decrease by 16% per 1 degree increase in temperature).

Confidence Interval for the Odds-Ratio

If the sample size is large, confidence intervals can be formed for logistic regression coefficients just as for linear models. The O-ring data is much too small to support this, but for illustration, a 95% confidence interval for the coefficient of temperature in the O-ring model would be calculated as follows:[2]

$$-0.1713 \pm 1.96 \times 0.0834 = (-0.335, -0.008).$$

Because the OR is obtained by exponentiating the coefficient, we get the confidence interval for the OR by exponentiating the upper and lower limits of the confidence interval for the coefficient.

Confidence interval for OR for O-ring data:

$$exp(-0.1713 \pm 1.96 \times 0.0834) = (e^{-0.335}, e^{-0.008}) = (0.82, 0.99).$$

[2] There are "exact" methods that can be applied in this situation.

Hypothesis Testing

Hypothesis testing for coefficients in logistic regression is almost the same as for linear regression. The test statistic is the coefficient estimate divided by its standard error, which is referred to a N(0,1) distribution to calculate the p-value.

The differences between hypothesis testing in logistic regression and linear models is that for logistic regression:

- we always need a large sample size
- we never assume normality of errors
- we never use a *t* distribution to calculate the *p*-value

Recall that for linear regression, the justification for using the *t* distribution was that the errors are normally distributed, which cannot be true for a binary response variables.

Note that in the output of the 'glm' function the header for the test statistic column is 'z value' rather than 't value' as it is when using the function 'lm'.

Hypothesis tests for the O-ring data

The p-value for the coefficient of temperature is 0.04. This reflects the fact that the confidence interval for the coefficient just barely excluded 0.

What does this mean in terms of the OR? For an OR the null hypothesis value of most interest is 1, i.e., we are interested in testing $H_0 : OR = 1$ vs $H_0 : OR \neq 1$. Fortunately this null hypothesis is equivalent to the null hypothesis that the coefficient is equal to 0.

Based on the O-ring data we would (if the sample size was much larger) say that there is evidence against the null hyopthesis that the OR for the association between temperature and failure is equal to 1.

As with linear regression, tests concerning the intercept are rarely of interest.

Example: ICU Data

The dataset “icu.csv” contains outcomes and patient characteristics on a large sample of patients admitted to a hospital intensive care unit. The response variable is a binary indicator of death within 30 days of hospital admission.

```
i=read.csv("icu.csv")
table(i$income)
```

```
    > $50k    $11-$25k    $25-$50k Under $11k
       368         977         756       2634
```

```
round(summary(glm(dth30 ~ sex+age+edu+factor(income), data=i, family=binomial))$
```

```
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)                -1.7707     0.2474 -7.1583   0.0000
sexMale                     0.0436     0.0625  0.6972   0.4857
age                         0.0118     0.0019  6.1071   0.0000
edu                         0.0166     0.0112  1.4845   0.1377
factor(income)$11-$25k      0.0000     0.1370 -0.0003   0.9998
factor(income)$25-$50k     -0.0393     0.1408 -0.2789   0.7803
factor(income)Under $11k    0.2694     0.1320  2.0410   0.0413
```

An interaction model

Now add the interaction between sex and age to the model.

```r
round(summary(glm(dth30 ~ sex*age+edu+factor(income), data=i, family=binomial))
```

```
                         Estimate Std. Error z value Pr(>|z|)
(Intercept)               -1.7744     0.2753 -6.4461   0.0000
sexMale                    0.0509     0.2473  0.2059   0.8368
age                        0.0118     0.0028  4.2131   0.0000
edu                        0.0166     0.0112  1.4835   0.1379
factor(income)$11-$25k    -0.0002     0.1371 -0.0013   0.9989
factor(income)$25-$50k    -0.0394     0.1409 -0.2798   0.7797
factor(income)Under $11k   0.2691     0.1322  2.0356   0.0418
sexMale:age               -0.0001     0.0038 -0.0307   0.9755
```

The estimated coefficient for the interaction term describes the difference between the *log odds-ratio* for age for men versus women. The estimate of the interaction term is very close to 0 in this case, which says that the OR for age is about the same for men as for women.