Chapter 10: Inference for regression models

Introductory Statistics for Engineering Experimentation

Peter R. Nelson, Marie Coffin and Karen A.F. Copeland

Slides by Douglas Bates

Outline

10.1 Inference for a regression line

10.2 Inference for other regression models

Section 10.1: Inference for a regression line

- Recall that the simple linear regression model is

$$\mathcal{Y}_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, \ldots, n \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2)$$

- The least squares estimates, $\widehat{\beta}_0$ and $\widehat{\beta}_1$, of the coefficients are functions of the data and hence are random variables. We associate *standard errors* with these estimates.
- The text derives formulas for the variance of the estimators. The formulas can be interesting but do not easily extend to more complex models. It is easier to simply read the standard error from the output.
- In the `R` output each coefficient estimate is accompanied by a `Std. Error` (standard error), a `t value` (the ratio of the estimate and its standard error) and a `Pr(>|t|)`, which is the p-value for the two-sided hypothesis test. The `confint` extractor can be used to determine confidence intervals.

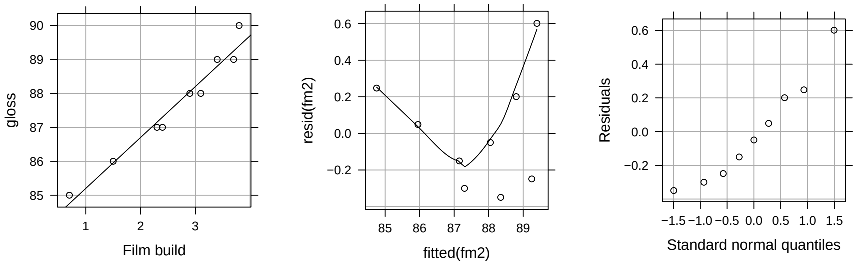
Examples 10.1.1 and 10.1.2

```
> summary(fm1 <- lm(time ~ temp, timetemp,
+                   subset = type == "Repaired"))
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -37.4902     2.7833  -13.47 3.52e-08
temp         -1.8643     0.1069  -17.45 2.30e-09
Residual standard error: 0.7699 on 11 degrees of freedom
Multiple R-squared: 0.9651, Adjusted R-squared: 0.9619
F-statistic: 304.3 on 1 and 11 DF,  p-value: 2.303e-09
> confint(fm1)

                 2.5 %     97.5 %
(Intercept) -43.616328 -31.364111
temp         -2.099460  -1.629053
```

- The confidence interval ($[-2.099, -1.629]$) on β_1, the slope, is of interest. The other confidence interval is not of interest because β_0 is meaningless for these data.

Example 10.1.3



```
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  83.7031     0.3184  262.86 3.04e-15
build         1.4988     0.1130   13.27 3.23e-06
Residual standard error: 0.3306 on 7 degrees of freedom
Multiple R-squared: 0.9618, Adjusted R-squared: 0.9563
F-statistic:   176 on 1 and 7 DF,  p-value: 3.234e-06
> confint(fm2)

                2.5 %    97.5 %
(Intercept) 82.950116 84.456061
build        1.231687  1.765977
```

Inference for coefficients

- As seen in the previous slides, we can evaluate confidence intervals on the coefficients, β_0 and β_1, with the `confint` extractor function.
- The formula for the $(1-\alpha)$ confidence interval on β_1 is

  $$\widehat{\beta}_1 \pm t(\alpha/2, \nu)\, s_{\beta_1}$$

  where ν is the degrees of freedom for residuals ($n-2$ for a simple linear regression) and s_{β_1} is the standard error for the coefficient.
- The observed t statistic, $\widehat{\beta}_1/s_{\beta_1}$, is used to perform tests of the hypothesis $H_0 : \beta_1 = 0$. The p-value for the two-sided alternative is given in the coefficient table. The p-value for the one-sided alternative that is indicated by the data will be half this value. By "indicated by the data" I mean the alternative $H_a : \beta_1 > 0$, if $\widehat{\beta}_1 > 0$ and vice versa.

More on inference for coefficients

- Testing $H_0 : \beta_1 = 0$ versus the appropriate alternative is usually of interest. Tests on β_0 are not always of interest as the intercept may not represent a meaningful response.
- For a simple linear regression the F test reported in the summary compares the model that was fit to a trivial model in which all the fitted values are equal to $\bar{y}$. You can also obtain this test as

```
> anova(fm1)

Analysis of Variance Table
Response: time
          Df  Sum Sq Mean Sq F value    Pr(>F)
temp       1 180.403 180.403  304.34 2.303e-09
Residuals 11   6.520   0.593
```

  This test is equivalent to the t-test of $H_0 : \beta_1 = 0$ *vs.* $H_a : \beta_1 \neq 0$.

Extracting the coefficients table only

- The analysis of variance table can also be obtained by explicitly comparing the model that was fit to the trivial model.

```
> anova(update(fm1, . ~ . - temp), fm1)

Analysis of Variance Table
Model 1: time ~ 1
Model 2: time ~ temp
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1     12 186.92
2     11   6.52  1    180.40 304.34 2.303e-09
```

- Sometimes it is convenient to extract the table of coefficients, standard errors and test statistics. You can do this by

```
> coef(summary(fm1))

              Estimate Std. Error   t value     Pr(>|t|)
(Intercept) -37.490219  2.7833482 -13.46947 3.518561e-08
temp         -1.864256  0.1068628 -17.44532 2.302911e-09
```

Inference for $\mu_{\mathcal{Y}|x}$

- In a regression model we consider the response as having a normal distribution conditional on a particular value of the covariate, $x = x_o$.
- This distribution has an expected value, which we write as $\mu_{\mathcal{Y}|x=x_o}$ or $\mathrm{E}(\mathcal{Y}|x = x_0)$. Our estimate of this conditional mean is $\widehat{\beta}_0 + \widehat{\beta}_1 x_0$.
- Just as $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are random variables with standard errors, our estimate $\widehat{\mu}_{\mathcal{Y}|x=x_o}$ has a standard error.
- The estimate and its standard error can be evaluated with predict and the optional argument se.fit = TRUE.

```
> str(predict(fm2, list(build = 2.6), se.fit = TRUE)) # Ex 10.1.
```

```
List of 4
 $ fit           : Named num 87.6
  ..- attr(*, "names")= chr "1"
 $ se.fit        : num 0.11
 $ df            : int 7
 $ residual.scale: num 0.331
```

Confidence intervals on the mean response

- Typically we use the standard errors to form a confidence interval on $\mu_{\mathcal{Y}|x=x_0}$, which we can create with the optional argument interval = "conf" to predict.
- In example 10.1.7 we wish to form a 90% confidence interval on the mean gloss when the film build is 2.6 mm

```
> predict(fm2, list(build = 2.6), int = "conf", level = 0.90)
```

```
       fit      lwr      upr
1 87.60005 87.39106 87.80904
```

- We can use the estimate and its standard error to conduct hypothesis tests but generally we are more interested in the confidence intervals. Occasionally we want to test $H_0 : \beta_0 = 0$ versus one of the alternatives and this is a test on the mean response when $x = 0$. We can obtain the p-value for this test from the table of coefficients.

Inference for a future value of $\mathcal{Y}$

- Note that the confidence interval on $\mu_{\mathcal{Y}|x=x_0}$ refers to the mean response at $x = x_0$, not the response that we will observe.
- If we want a prediction interval at $x = x_0$ then we must formulate it as $\mathcal{Y}_0 = \mu_{\mathcal{Y}|x=x_0} + \epsilon_0$ which we estimate as

  $$\mathrm{E}[\mathcal{Y}_0] = \widehat{\beta}_0 + \widehat{\beta}_1 x_0$$

  with a standard error of $\sqrt{s^2_{\hat{\mu}} + s^2}$.
- A 90% prediction interval on the gloss at a build of 3 mm. is

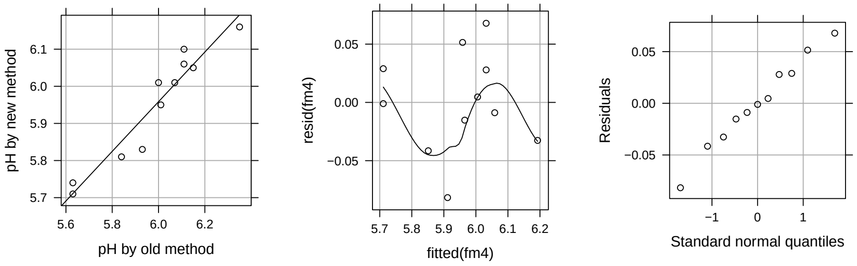
```
> predict(fm2, list(build = 3:4), int = "pred", level = 0.90)
```

```
       fit      lwr      upr
1 88.19958 87.53502 88.86415
2 89.69842 88.97728 90.41955
```

Testing for lack of fit

- One of the assumptions in a simple linear regression is that the relationship between $\mathcal{Y}$ and x is reasonably close to a straight line over the range of interest.
- If we have replicates in the data then we can check this assumption by evaluating the sum of squares due to replication (the pooled sum of squares of the deviations about the average within replicates) and what is called the *mean square for lack of fit*.
- There are various ways of calculating these quantities, some with unsatisfactory numerical properties. A simple way of doing this test is to compare the linear model to a model with the covariate x as a factor.

Example 10.1.10



```
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.94538    0.39539    4.92 0.000824
phold        0.66886    0.06603   10.13 3.21e-06
Residual standard error: 0.04555 on 9 degrees of freedom
Multiple R-squared: 0.9194, Adjusted R-squared: 0.9104
F-statistic: 102.6 on 1 and 9 DF,  p-value: 3.214e-06
```

To perform the lack of fit test we compare this model fit to one with `phold` treated as a factor.

Example 10.1.10 (cont'd)

```
> anova(fm4, lm(phnew ~ factor(phold), phmeas))
```

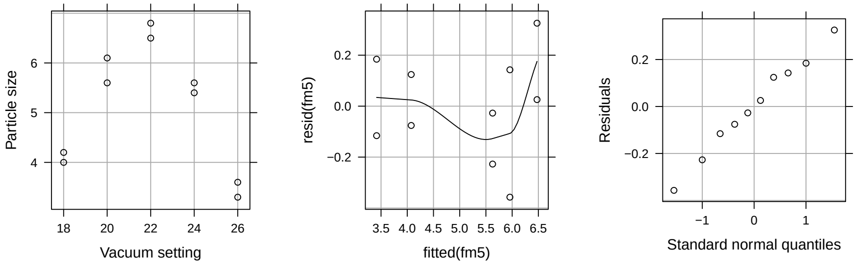
```
Analysis of Variance Table
Model 1: phnew ~ phold
Model 2: phnew ~ factor(phold)
  Res.Df      RSS Df Sum of Sq      F Pr(>F)
1      9 0.018673
2      2 0.001250  7  0.017423 3.9823 0.2153
```

- Note that this result is different from the result shown in the text. In the text they use only one of the sets of replicates. Here we use both.
- The computer is better at picking up repetitions in the covariate than are humans.
- In either calculation there is no significant evidence of lack of fit. We prefer the calculation with more denominator degrees of freedom (the one shown above). More denominator degrees of freedom produces a more powerful test.

Section 10.2: Inference for other regression models

- As seen in chapter 3, regression models can incorporate many different types of terms (see p. 386).
- Inferences on the coefficients in such a model can be performed using the information in the coefficients table.
- We must, however, be careful of the interpretation of the tests. For example, if we fit a quadratic (next slide) then we generally are not interested in testing $H_0 : \beta_1 = 0$ in the presence of the quadratic term.
- The general rule is that the t-test in the coef table is a test of removing only that term and keeping all the other terms in the model. Ask yourself if it would be a sensible model with that term omitted.

Example 10.2.1



```
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -74.25000    5.24071  -14.17 2.07e-06
x             7.42107    0.48221   15.39 1.18e-06
I(x^2)       -0.17054    0.01094  -15.58 1.08e-06
Residual standard error: 0.2316 on 7 degrees of freedom
Multiple R-squared: 0.9731, Adjusted R-squared: 0.9654
F-statistic: 126.5 on 2 and 7 DF,  p-value: 3.203e-06
> confint(fm5)
```

```
                  2.5 %      97.5 %
(Intercept) -86.6423011 -61.8576989
x             6.2808209   8.5613219
I(x^2)       -0.1964131  -0.1446583
```

Prediction intervals and confidence intervals on $\mu_{\mathcal{Y}|x=x_0}$

- Prediction intervals and confidence intervals on $\mu_{\mathcal{Y}|x=x_0}$ are calculated as before. We must specify values for all the covariates in the model but we do not need to specify both x and x^2. Higher-order terms are evaluated from the formula.

```
> predict(fm5, list(x = 21), int = "pred")

       fit      lwr      upr
1 6.38625 5.780808 6.991692

> predict(fm5, list(x = 21), int = "conf")

       fit      lwr      upr
1 6.38625 6.128254 6.644246
```

Testing lack of fit

- We test lack of fit as before. If we have more than one covariate we must use a cell-means model with all of the covariates as factors.

```
> anova(fm5, lm(y ~ factor(x), ex336))

Analysis of Variance Table
Model 1: y ~ x + I(x^2)
Model 2: y ~ factor(x)
  Res.Df     RSS Df Sum of Sq      F Pr(>F)
1      7 0.37557
2      5 0.25500  2   0.12057 1.1821 0.3799
```