Imperial College
London

Problem Sheet 8 Solutions

MATH50011
Statistical Modelling 1

Week 10

Lecture 17: Inference with Normal Theory Assumptions

1. Consider the simple linear model

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, \dots, n.$$

Assume that the full rank and normal theory assumptions hold.

(a) Describe how to test $H_0 : \beta_1 = 0$ against $H_1 : \beta_1 \neq 0$ at level α using (i) a t-test; and (ii) an F-test.

(b) Show that the p-values for the tests in (a) are equal.

(c) Derive a $(1-\alpha) \times 100\%$ confidence interval for $E(Y|x_0)$, where x_0 is a fixed value of the covariate.

Solution.

(a) Part (i) amounts to applying Lemma 22 using results for the simple linear model studied in Example 52. We use the test statistic

$$t = \frac{\hat{\beta}_1}{\sqrt{\hat{\sigma}^2 / \sum_{i=1}^n (x_i - \bar{x})^2}} \sim t_{n-2}$$

where $\hat{\sigma}^2 = RSS/(n-2)$. Let $Z \sim t_{n-2}$ and let $t_{n-2,\alpha/2}$ be such that $P(Z > t_{n-2,\alpha/2}) < \alpha$. If $|t| \geq t_{n-2,\alpha/2}$ then we reject H_0 at level α.
In part (ii), we instead use the statistic

$$F = \frac{RSS_0 - RSS}{RSS/(n-2)} \sim F_{1,n-2}$$

obtained by applying Lemma 23 with $r = 2$ and $s = 1$. Noting that $RSS_0 = \sum_{i=1}^n (Y_i - \bar{Y})^2$ and $RSS = \sum_{i=1}^n (Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$, some algebra yields

$$F = \frac{\hat{\beta}_1^2 \sum_{i=1}^n (x_i - \bar{x})^2}{\hat{\sigma}^2}.$$

Let $W \sim F_{1,n-2}$ and let $F_{1,n-2,\alpha/2}$ be such that $P(W > F_{1,n-2,\alpha/2}) < \alpha$. If $F \geq F_{1,n-2,\alpha/2}$, then we reject H_0 at the α level.

(b) To see that the p-values for the tests in (a) are equal, we note that

$$t^2 = F$$

and hence that the p-values satisfy $P(|t_{n-2}| > |t|) = P(t_{n-2}^2 > t^2) = P(F_{1,n-2} > F)$.

(c) We will apply Lemma 22 with $c = (1, x_0)^T$. Note that, making use of results from Example 52, we have

$$c^T(X^TX)^{-1}c = \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}.$$

From Lemma 22, we have that

$$\frac{\hat{\beta}_0 + \hat{\beta}_1 x_0 - E(Y|x_0)}{\sqrt{\hat{\sigma}^2\left(\frac{1}{n} + \frac{(x_0-\bar{x})^2}{\sum_{i=1}^n (x_i-\bar{x})^2}\right)}} \sim t_{n-2}.$$

Let τ be the value such that $P(-\tau < t_{n-2} < \tau) = 1 - \alpha$. Using standard rearrangements for pivotal statistics, we have that

$$\hat{\beta}_0 + \hat{\beta}_1 x_0 \pm \tau\sqrt{\hat{\sigma}^2\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right)}$$

is a $(1 - \alpha) \times 100\%$ confidence interval for $E(Y|x_0)$.

2. Suppose we believe that the distribution of Y depends on covariates x_1 and x_2, and that the relationship between Y and x_1 depends on the value of x_2. That is, we assume there is an *interaction* between x_1 and x_2.

To allow for the interaction, we use the following linear model:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1} x_{i2} + \epsilon_i.$$

The term $\beta_3 x_1 x_2$ is called an *interaction term*.

Now, assuming that the full rank and normal theory assumptions hold:

(a) Derive expressions for $E(Y_i|x_{i1} = x, x_{i2} = 0)$ and $E(Y_i|x_{i1} = x, x_{i2} = 1)$.

Solution. We have

$$\begin{aligned} E(Y_i|x_{i1} = x, x_{i2} = 0) &= \beta_0 + \beta_1 x_{i1} \\ E(Y_i|x_{i1} = x, x_{i2} = 1) &= \beta_0 + \beta_1 x_{i1} + \beta_2 + \beta_3 x_{i1} = \beta_0 + \beta_2 + (\beta_1 + \beta_3)x_{i1} \end{aligned}$$

which illustrates how, depending on the value of x_{i2}, we obtain different lines of the form $a + bx_{i1}$ to describe the relationship between Y and x_1.

(b) State a hypothesis in terms of the parameter vector that could be used to test for the presence of an interaction between x_1 and x_2. Construct a level α test of your hypothesis, clearly identifying the form of the test statistic and its distribution under the null hypothesis.

Solution. The null hypothesis is $H_0 : \beta_3 = 0$. A t-test based on Lemma 22 can be used with test statistic

$$t = \frac{\hat{\beta}_3}{\sqrt{\hat{Var}(\hat{\beta}_3)}} \sim t_{n-4}$$

where $\hat{Var}(\hat{\beta}_3) = c^T(X^TX)^{-1}cRSS/(n-4)$ for $c = (0,0,0,1)^T$, X the design matrix, and RSS the residual sum of squares.

(c) State a hypothesis in terms of the parameter vector that could be used to test for the presence of *any* effect of x_1 on the distribution of Y. Construct a level α test of your hypothesis, clearly identifying the form of the test statistic and its distribution under the null hypothesis.

Solution. Now, the null hypothesis is $H_0 : \beta_1 = \beta_3 = 0$. An F-test based on Lemma 23 can be used with statistic

$$F = \frac{RSS_0 - RSS}{RSS}\frac{n-4}{2} \sim F_{2,n-4}$$

where RSS_0 is the residual sum of squares based on the design matrix $X_0 = \begin{pmatrix} 1 & x_2 \end{pmatrix}$ and RSS is the residual sum of squares based on the full model.

Lecture 18: Outliers, Under- and Over-fitting, WLS

3. Let $Y = X\beta + \epsilon$ and assume that $E(\epsilon) = 0$ and $Cov(\epsilon) = \sigma^2 I$. Moreover, assume that X is an $n \times p$ matrix with full column rank. Let $\hat{\beta} = (X^TX)^{-1}X^TY$ denote the least squares estimator of β under $E(Y) = X\beta$.

(a) Suppose that you fit the model in which $E(Y) = X\beta$ when the true model is such that $E(Y) = X\beta + Z\gamma$. That is, the model is under fitted.

i. Show that $\hat{\beta}$ is typically a biased estimator of β.

ii. Under which conditions on Z we have that $\hat{\beta}$ is an unbiased estimator of β,

iii. Compute $Cov(\hat{\beta})$.

Solution. Using properties of expectations, we find that

$$\begin{aligned} E(\hat{\beta}) &= E[(X^TX)^{-1}X^TY] = (X^TX)^{-1}X^TE(Y) \\ &= (X^TX)^{-1}X^T(X\beta + Z\gamma) \\ &= \beta + (X^TX)^{-1}X^TZ\gamma. \end{aligned}$$

Because $E(\hat{\beta}) \neq \beta$, we conclude that $\hat{\beta}$ is typically a biased estimator of β. However, if the columns of X are orthogonal to the columns of Z we have $X^TZ = 0$. Thus, in this case, the estimator would be unbiased. (Also, if $\gamma = 0$, the larger model reduces to $E(Y) = X\beta$.)

Moreover, using properties of covariance,

$$Cov(\hat{\beta}) = (X^TX)^{-1}X^TCov(Y)X(X^TX)^{-1} = \sigma^2(X^TX)^{-1}.$$

(b) Let $X = (X_1, X_2)$, where X_1 denotes the matrix with the first k columns of X. Suppose that you fit the model $E(Y) = X\beta$ when the true model is $E(Y) = X_1\beta$. That is, the model is over fitted.

i. Show that the fitted model provides an unbiased estimator of the true model.

ii. Show that the elements of $\hat{\beta}$ have in general higher variance than would result from fitting the true (reduced) model.

iii. Under which conditions on X the elements of $\hat{\beta}$ do have higher variance than would result from fitting the true (reduced) model.

Solution. Using properties of expectations

$$\begin{aligned} E(\hat{\beta}) &= (X^TX)^{-1}X^TE(Y) \\ &= (X^TX)^{-1}X^TX_1\beta_1 \\ &= (X^TX)^{-1}X^T(X_1, X_2)\begin{pmatrix}\beta_1 \\ 0\end{pmatrix} \\ &= \begin{pmatrix}\beta_1 \\ 0\end{pmatrix}. \end{aligned}$$

This implies that $E(X\hat{\beta}) = X_1\beta_1$ as claimed.
Furthermore, we have

$$\begin{aligned} Cov(\hat{\beta}) &= \sigma^2(X^TX)^{-1} \\ &= \sigma^2\begin{pmatrix} X_1^TX_1 & X_1^TX_2 \\ X_2^TX_1 & X_2^TX_2 \end{pmatrix}^{-1} \\ &= \sigma^2\begin{pmatrix} (X_1^TX_1)^{-1} + FE^{-1}F^T & -FE^{-1} \\ -E^{-1}F^T & E^{-1} \end{pmatrix} \end{aligned}$$

where $F = (X_1^TX_1)^{-1}X_1^TX_2$, and

$$E = X_2^TX_2 - X_2^TX_1(X_1^TX_1)^{-1}X_1^TX_2 = X_2^T(I - P_{span(X_1)})X_2.$$

Therefore, $Cov(\hat{\beta}) = \sigma^2[(X_1^TX_1)^{-1} + FE^{-1}F^T]$, compared with $\sigma^2(X_1^TX_1)^{-1}$ which would result from fitting the true model $E(Y) = X_1\beta_1$. using the fact that $FE^{-1}F^T$ is positive definite unless $X_1^TX_2 = 0$, this implies that the variance of individual components of $\hat{\beta}_1$ will be inflated by overfitting unless the unnecessary fitted terms are orthogonal to the other terms in the model. The lesson is that over fitting does not introduce bias into regression coefficient estimates, but it does inflate their variances.

4. The file psa.csv is a comma-separated file containing data on 28 men having hormonally treated prostate cancer. The first line of the file contains the following variable names. Each successive line contains data pertinent to one of the 28 patients.

nadirpsa	= lowest PSA value attained post therapy (ng/ml)
grade	= tumor grade (1= least aggressive, 3= most)
age	= patient's age (years)
obstime	= time in remission (months)

(a) Define a linear model for $E(\log Y_i)$ where Y_i is the time spent in remission by the ith patient. In your model, include *nadirpsa*, *age*, as well as *grade* (treated as a 3-level categorical variable). How many parameters are in your model?

Solution. There are several correct solutions with different parametrizations (especially for *grade*). One option is to define

$$E(\log Y_i) = \beta_0 + \beta_1 \mathit{nadirpsa}_i + \beta_2 \mathit{age}_i + \beta_3 \mathit{grade2}_i + \beta_4 \mathit{grade3}_i$$

where $\mathit{grade2}_i = 1$ if the ith individual has a grade 2 tumor and is 0 otherwise and $\mathit{grade3}_i = 1$ if the ith individual has a grade 3 tumor and is 0 otherwise.

There are $p = 5$ parameters in the linear model.

(b) Using the software of your choice, obtain the least squares estimates corresponding to the linear model you specified in part (a).

Solution. We can use R to fit the above model by running the commands

```
psa <- read.csv("psa.csv")
fit <- lm(log(obstime)~nadirpsa+age+factor(grade), data=psa)
summary(fit)
```

We find the following estimates

$\hat\beta_0$	$\hat\beta_1$	$\hat\beta_2$	$\hat\beta_3$	$\hat\beta_4$
3.149	-0.020	0.005	-0.752	-0.424

(c) Is there evidence that *nadirpsa* is associated with time in remission? Implement a hypothesis test to justify your conclusion.

Solution. We will test the hypothesis that $H_0 : \beta_1 = 0$ against $H_1 : \beta \neq 0$ at level $\alpha = 0.05$. This can be read directly from the summary(fit) output, where we see

```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     3.148707   2.116496   1.488   0.1504
nadirpsa       -0.020094   0.003673  -5.470 1.46e-05 ***
age             0.004545   0.033651   0.135   0.8937
factor(grade)2 -0.751607   0.386583  -1.944   0.0642 .
factor(grade)3 -0.423992   0.397401  -1.067   0.2971
```

The last column of the row for `nadirpsa` in the output provides the p-value from the test we are interested in. Since 1.46×10^{-5} is less than 0.05, we reject the null hypothesis.

(d) Modify the linear model in (a) to allow for an interaction between tumor grade and age. How many parameters are in your model?

Solution. There are several correct solutions with different parametrizations (especially for *grade*). One option is to define

$$E(\log Y_i) = \beta_0 + \beta_1 \mathit{nadirpsa}_i + \beta_2 \mathit{age}_i + \beta_3 \mathit{grade2}_i + \beta_4 \mathit{grade3}_i + \beta_5 \mathit{age}_i \mathit{grade2}_i + \beta_6 \mathit{age}_i \mathit{grade3}_i$$

where $\mathit{grade2}_i = 1$ if the ith individual has a grade 2 tumor and is 0 otherwise and $\mathit{grade3}_i = 1$ if the ith individual has a grade 3 tumor and is 0 otherwise.

There are $p = 7$ parameters in the linear model.

(e) Using the software of your choice, obtain the least squares estimates corresponding to the linear model you specified in part (d).

Solution. We can use R to fit the above model by running the commands

```
psa <- read.csv("psa.csv")
fit <- lm(log(obstime)~nadirpsa+age*factor(grade), data=psa)
summary(fit)
```

We find the following estimates

$\hat\beta_0$	$\hat\beta_1$	$\hat\beta_2$	$\hat\beta_3$	$\hat\beta_4$	$\hat\beta_5$	$\hat\beta_6$
6.913	-0.019	-0.056	-4.584	-6.570	0.061	0.095

(f) Is there evidence of an interaction between tumor grade and age? Implement a hypothesis test to justify your conclusion.

Solution. We will test the hypothesis that $H_0 : \beta_5 = \beta_6$ at level $\alpha = 0.05$. We can do this in R using the commands

```
psa <- read.csv("psa.csv")
fit <- lm(log(obstime)~nadirpsa+age*factor(grade), data=psa)
fit0 <- lm(log(obstime)~nadirpsa+age+factor(grade), data=psa)

# "by hand" based on Lemma 23 with n=28, r=7, s=5
RSS <- sum(fit$residuals^2)
RSS0 <- sum(fit0$residuals^2)
F.stat <- ((RSS0-RSS)/2) / (RSS/21)
p.value <- pf(F.stat, df1 = 2, df2 = 21, lower=FALSE)

# with built-in functions only
anova(fit0,fit)
```

Whether we use the "by hand" or built-in implementation, the resulting p-value is 0.64 means that the estimates we obtained (or more extreme estimates) are fairly common under H_0. We do not reject the null hypothesis.