Recommended Exercise 7 in Statistical Linear Models, Spring 2021

alexaoh

17 mars, 2021

Gå skikkelig gjennom LF senere!

Problem 1 Inference about a new observation in multiple linear regression

a)

Since $\hat{\boldsymbol{\beta}} = (X^TX)^{-1}X^T\boldsymbol{Y}$, we know that $\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \sigma^2(X^TX)^{-1})$. Since the Gaussian distribution is closed under linear transformations, this means that $\boldsymbol{x}_0^T\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{x}_0^T\boldsymbol{\beta}, \sigma^2\boldsymbol{x}_0^T(X^TX)^{-1}\boldsymbol{x}_0)$, which is univariate. Now, since $E\boldsymbol{x}_0^T\hat{\boldsymbol{\beta}} = \boldsymbol{x}_0^T\boldsymbol{\beta}$, $\boldsymbol{x}_0^T\hat{\boldsymbol{\beta}}$ is an unbiased estimator of $EY_0 = \boldsymbol{x}_0^T\boldsymbol{\beta}$, with distribution as given above.

b)

Generally, we know that if $Z \sim N(0,1)$ and $U \sim \chi^2_q$, where Z and U are independent, then $\frac{Z}{\sqrt{U/q}} \sim t_q$. In this case, we can set $Z = \frac{\boldsymbol{x}_0^T\hat{\boldsymbol{\beta}} - \boldsymbol{x}_0^T\boldsymbol{\beta}}{\sqrt{\sigma^2\boldsymbol{x}_0^T(X^TX)^{-1}\boldsymbol{x}_0)}}$ and $U = \frac{1}{\sigma^2}\text{SSE} = \frac{(n-p)\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-p}$. Moreover, Z and U are independent, because $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$ are independent. Hence, the estimator

$$T = \frac{Z}{\sqrt{U/q}} = \frac{\frac{\boldsymbol{x}_0^T\hat{\boldsymbol{\beta}} - \boldsymbol{x}_0^T\boldsymbol{\beta}}{\sqrt{\sigma^2\boldsymbol{x}_0^T(X^TX)^{-1}\boldsymbol{x}_0)}}}{\sqrt{\frac{(n-p)\hat{\sigma}^2}{\sigma^2}/(n-p)}} = \frac{\boldsymbol{x}_0^T\hat{\boldsymbol{\beta}} - \boldsymbol{x}_0^T\boldsymbol{\beta}}{\sqrt{\hat{\sigma}^2\boldsymbol{x}_0^T(X^TX)^{-1}\boldsymbol{x}_0)}} \sim t_{n-p}.$$

This means that

$$\begin{aligned}
&-t \leq \frac{\boldsymbol{x}_0^T\hat{\boldsymbol{\beta}} - \boldsymbol{x}_0^T\boldsymbol{\beta}}{\sqrt{\hat{\sigma}^2\boldsymbol{x}_0^T(X^TX)^{-1}\boldsymbol{x}_0}} \leq t, \\
&\iff \boldsymbol{x}_0^T\hat{\boldsymbol{\beta}} - t\hat{\sigma}\sqrt{\boldsymbol{x}_0^T(X^TX)^{-1}\boldsymbol{x}_0)} \leq \boldsymbol{x}_0^T\boldsymbol{\beta} \leq \boldsymbol{x}_0^T\hat{\boldsymbol{\beta}} + t\hat{\sigma}\sqrt{\boldsymbol{x}_0^T(X^TX)^{-1}\boldsymbol{x}_0)}, \\
&\iff \boldsymbol{x}_0^T\hat{\boldsymbol{\beta}} - t\hat{\sigma}\sqrt{\boldsymbol{x}_0^T(X^TX)^{-1}\boldsymbol{x}_0)} \leq EY_0 \leq \boldsymbol{x}_0^T\hat{\boldsymbol{\beta}} + t\hat{\sigma}\sqrt{\boldsymbol{x}_0^T(X^TX)^{-1}\boldsymbol{x}_0)}.
\end{aligned}$$

Hence,

$$\begin{aligned}
1-\alpha &= P\left(-t_{\alpha/2} \leq \frac{\boldsymbol{x}_0^T\hat{\boldsymbol{\beta}} - \boldsymbol{x}_0^T\boldsymbol{\beta}}{\sqrt{\hat{\sigma}^2\boldsymbol{x}_0^T(X^TX)^{-1}\boldsymbol{x}_0}} \leq t_{\alpha/2}\right) \\
&= P\left(\boldsymbol{x}_0^T\hat{\boldsymbol{\beta}} - t_{\alpha/2}\hat{\sigma}\sqrt{\boldsymbol{x}_0^T(X^TX)^{-1}\boldsymbol{x}_0} \leq EY_0 \leq \boldsymbol{x}_0^T\hat{\boldsymbol{\beta}} + t_{\alpha/2}\hat{\sigma}\sqrt{\boldsymbol{x}_0^T(X^TX)^{-1}\boldsymbol{x}_0}\right),
\end{aligned}$$

which means that $\boldsymbol{x}_0^T\hat{\boldsymbol{\beta}} \pm t_{\alpha/2}\hat{\sigma}\sqrt{\boldsymbol{x}_0^T(X^TX)^{-1}\boldsymbol{x}_0)}$ is a $100(1-\alpha)\%$ confidence interval for EY_0.

c)

A similar process to the one used in *b)* is used. For a future observation Y_0 we can derive that $Y_0 - \boldsymbol{x}_0^T\hat{\boldsymbol{\beta}} \sim N(0, \sigma^2(1 + \boldsymbol{x}_0^T(X^TX)^{-1}\boldsymbol{x}_0))$, since $Y_0 = \boldsymbol{x}_0^T\boldsymbol{\beta} + \varepsilon_0$ and $\boldsymbol{x}_0^T\hat{\boldsymbol{\beta}}$ is an unbiased estimator of EY_0. Setting $Z = \frac{Y_0 - \boldsymbol{x}_0^T\hat{\boldsymbol{\beta}}}{\sigma\sqrt{(1+\boldsymbol{x}_0^T(X^TX)^{-1}\boldsymbol{x}_0)}}$ and set U as in *b)*. Solving the same inequalities as in *b)* then gives

$$1 - \alpha = P\left(\boldsymbol{x}_0^T\hat{\boldsymbol{\beta}} - t_{\alpha/2}\hat{\sigma}\sqrt{1 + \boldsymbol{x}_0^T(X^TX)^{-1}\boldsymbol{x}_0} \leq Y_0 \leq \boldsymbol{x}_0^T\hat{\boldsymbol{\beta}} + t_{\alpha/2}\hat{\sigma}\sqrt{1 + \boldsymbol{x}_0^T(X^TX)^{-1}\boldsymbol{x}_0}\right),$$

which means that $\boldsymbol{x}_0^T\hat{\boldsymbol{\beta}} \pm t_{\alpha/2}\hat{\sigma}\sqrt{1 + \boldsymbol{x}_0^T(X^TX)^{-1}\boldsymbol{x}_0}$ is a $100(1-\alpha)\%$ prediction interval for Y_0. Why does Y_0 independent of $\boldsymbol{\varepsilon}$ give independence from $\boldsymbol{x}_0^T\hat{\boldsymbol{\beta}}$?

d)

```r
acidrain <-read.table("https://www.math.ntnu.no/emner/TMA4267/2018v/acidrain.txt",header=TRUE)
fit <- lm(y~., data = acidrain)

# Confidence interval: Done manually via the equations found in previous tasks.
X <- model.matrix(fit)
x0 <- c(1, 3, 50, 1, 50, 2, 1, 0)
n <- dim(X)[1]
p <- dim(X)[2]
quantile <- qt(0.025, n-p,lower.tail = FALSE)
prediction <- t(x0)%*%fit$coefficients
upper.conf <- prediction + quantile*summary(fit)$sigma*sqrt(t(x0)%*%solve(t(X)%*%X)%*%x0)
lower.conf <- prediction - quantile*summary(fit)$sigma*sqrt(t(x0)%*%solve(t(X)%*%X)%*%x0)
prediction
```

```
#>          [,1]
#> [1,] 5.531684
```

```r
lower.conf
```

```
#>          [,1]
#> [1,] 5.446329
```

```r
upper.conf
```

```
#>          [,1]
#> [1,] 5.617039
```

```r
# Confidence interval: Done in R.
newdata <- data.frame(x1=3,x2=50,x3=1,x4=50,x5=2,x6=1,x7=0)
predict(fit, newdata, level=.95, interval="confidence")
```

```
#>        fit      lwr      upr
#> 1 5.531684 5.446329 5.617039
```

```r
# Prediction interval: Done manually via the equations found in previous tasks.
upper.pred <- prediction + quantile*summary(fit)$sigma*sqrt(1+t(x0)%*%solve(t(X)%*%X)%*%x0)
lower.pred <- prediction - quantile*summary(fit)$sigma*sqrt(1+t(x0)%*%solve(t(X)%*%X)%*%x0)
prediction
```

```
#>          [,1]
#> [1,] 5.531684
```

```r
lower.pred
```

```
#>          [,1]
#> [1,] 5.272555
```

```r
upper.pred
```

```
#>          [,1]
#> [1,] 5.790813
```

```r
# Prediction interval: Done in R.
predict(fit, newdata, level=.95, interval="prediction")
```

```
#>        fit      lwr      upr
#> 1 5.531684 5.272555 5.790813
```

e)

Calculate $\boldsymbol{x}_0^T(X^TX)^{-1}\boldsymbol{x}_0$ with $\boldsymbol{x}_0^T = (1 \quad x_0)$ and

$$X^T = \begin{pmatrix} 1 & 1 & 1 & \dots & \dots & 1 \\ x_1 & x_2 & x_3 & \dots & \dots & x_n \end{pmatrix}.$$

This, combined with the rest of the expression for the confidence interval from *b)*, gives the desired expression for simple linear regression.

Problem 2 Plant stress

a)

- t-value for intercept is missing: This is calculated by the test statistic $\hat{\beta}_{\text{intercept}}/\hat{SE}^2 \approx 16.15942/0.04140 \approx 390.3242$. This is the test statistic for testing $H_0 : \hat{\beta}_{\text{intercept}} = 0$ vs. $H_1 : \hat{\beta}_{\text{intercept}} \neq 0$.
- Std. Error for D:T is missing: This can be calculated from the same test statistic as in the last bullet point, since we know the t-value for D:T. Hence, $\hat{\beta}_{\text{D:T}}/\hat{SE} = t$ gives $\hat{SE} = \hat{\beta}_{\text{D:T}}/t \approx -0.00242/-0.058 \approx 0.0417$.
- p-value for D:F:T is missing: This can be calculated as $2P(T \geq t) =$ `2*pt(2.198, 24, lower.tail = F)` ≈ 0.037836.
- Multiple R-squared is missing: This can be calculated as $1 - \frac{\text{SSE}}{\text{SST}}$. SSE can be calculated from the formula $(n-p)\hat{\sigma}^2 = (32-8)\hat{\sigma}^2 = 24\hat{\sigma}^2 \approx 1.3163914$, where $\hat{\sigma}$ is found as `Residual Standard Error` in the summary-table. We also know the F-statistic, which is calculated as $\frac{\text{SSR}/(p-1)}{\text{SSE}/(n-p)} = \frac{(\text{SST}-\text{SSE})/(p-1)}{\text{SSE}/(n-p)} = \frac{(\text{SST}-\text{SSE})/(8-1)}{\text{SSE}/(32-8)} = \frac{(\text{SST}-\text{SSE})/7}{\text{SSE}/24} \approx 105.6$, which gives that $\text{SST} \approx 41.8612452$. Finally, this gives $R^2 \approx 0.9685535$.

b)

An estimator could be $\hat{\gamma} = 2^{\hat{\beta}_F - \hat{\beta}_D}$. The expected value and variance of this estimator can be calculated using a first order Taylor expansion. Setting $h(x, y) = 2^{x-y}$ and using a Taylor expansion in x and y of first order gives

$$\begin{aligned} h(\hat{\beta}_F, \hat{\beta}_D) &\approx h(\beta_F, \beta_D) + h_{\hat{\beta}_F}(\beta_F, \beta_D)(\beta_F - \hat{\beta}_F) + h_{\hat{\beta}_D}(\beta_F, \beta_D)(\beta_D - \hat{\beta}_D) \\ &= 2^{\beta_F - \beta_D} + 2^{\beta_F - \beta_D}(\beta_F - \hat{\beta}_F)\ln(2) + 2^{\beta_F - \beta_D}(\beta_D - \hat{\beta}_D)\ln(2), \end{aligned}$$

since $h_x(x,y) = \frac{\partial}{\partial x}h(x,y) = \frac{\partial}{\partial x}2^{x-y} = \frac{\partial}{\partial x}\exp\left((x-y)\ln(2)\right) = 2^{x-y}\ln(2)$. This means that $\mathrm{E}(\hat{\gamma}) = \mathrm{E}(2^{\hat{\beta}_F - \hat{\beta}_D}) = \mathrm{E}(h(\hat{\beta}_F, \hat{\beta}_D)) \approx h(\beta_F, \beta_D) = 2^{\beta_F - \beta_D}$, since $\hat{\beta}_F$ and $\hat{\beta}_D$ are unbiased estimators of β_F and β_D. Moreover, $\mathrm{Var}(\hat{\gamma}) = \mathrm{Var}(2^{\hat{\beta}_F - \hat{\beta}_D}) = \mathrm{Var}(h(\hat{\beta}_F, \hat{\beta}_D)) \approx h_{\hat{\beta}_F}(\beta_F, \beta_D)^2\mathrm{Var}(\hat{\beta}_F) + h_{\hat{\beta}_D}(\beta_F, \beta_D)^2\mathrm{Var}(\hat{\beta}_D) = 2^{2(\beta_F - \beta_D)}(\ln(2))^2(\mathrm{Var}(\hat{\beta}_F) - \mathrm{Var}(\hat{\beta}_D))$.

From Figure 1, we can get the numerical estimates of the moments, given as

Problem 3 Multiple testing with plant stress

```r
pvalues <- scan("https://www.math.ntnu.no/emner/TMA4267/2018v/damagePvalues.txt")
m <- length(pvalues)
```

a)

The family-wise error rate (FWER) is the probability of one or more false positive findings $P(V > 0)$, where V is the number of false positive findings among the m tests. The false discovery rate (FDR) gives the expected proportion of false positive results among the m hypotheses tested. A false positive is the same as a type I error, a case where the null hypothesis is rejected despite it being true. In the case of multiple linear regression this means that it is concluded that there is a linear relationship between the predictors and the response, despite the relationship not existing.

b)

Control FWER at level $\alpha = 0.05$ with the Bonferroni method. This gives $\alpha_{\mathrm{loc}} = \frac{0.05}{m} = 5 \times 10^{-6}$. In this case one would reject `length(pvalues[pvalues<5e-6])` $= 19$ null-hypotheses in the data. Requirements when using Bonferronis method are