

# Linear Model

## Fitted Model

```
(catheter.df <- read.table("catheter-1.data", header = TRUE))
```

```
##      ht   wt  ca
## 1  42.8 40.0 37
## 2  63.5 93.5 50
## 3  37.5 35.5 34
## 4  39.5 30.0 36
## 5  45.5 52.0 43
## 6  38.5 17.0 28
## 7  43.0 38.5 37
## 8  22.5  8.5 20
## 9  37.0 33.0 34
## 10 23.5  9.5 30
## 11 33.0 21.0 38
## 12 58.0 79.0 47
```

```
catheter.lm <- lm(ca~., data=catheter.df)
summary(catheter.lm)
```

```
##
## Call:
## lm(formula = ca ~ ., data = catheter.df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -6.7419 -1.2034 -0.2595  1.8892  6.6566
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  20.3758      8.3859   2.430   0.038 *
## ht           0.2107      0.3455   0.610   0.557
## wt           0.1911      0.1583   1.207   0.258
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.778 on 9 degrees of freedom
## Multiple R-squared:  0.8254, Adjusted R-squared:  0.7865
## F-statistic: 21.27 on 2 and 9 DF,  p-value: 0.0003888
```

## Estimated Coefficients

For linear model  $Y \sim N(\mu, \sigma^2 I)$ , the link function is  $\mu = X\beta$ , and the estimated coefficients is  $\hat{\beta} = (X^T X)^{-1} X^T Y$ .

Given  $U=MV$ , then  $\mu_U = M\mu_V$ , we can get  $\mu_B = (X^T X)^{-1} X^T (X\beta) = \beta$

```
x1 <- catheter.df[,1]
x2 <- catheter.df[,2]
X <- cbind(1,x1,x2)
y <- matrix(catheter.df[,3],12,1)
(BETAhat <- solve(t(X)%*%X)%*%t(X)%*%y)
```

```
##           [,1]
## 20.3757645
## x1  0.2107473
## x2  0.1910949
```

# Estimated $\sigma^2$ ( $=RSE^2$ )

From theory,  $\frac{RSS}{\sigma^2} \sim \chi_{n-k-1}^2$ , thus  $\frac{E(RSS)}{\sigma^2} = n - k - 1 \rightarrow \sigma^2 = \frac{E(RSS)}{n-k-1}$

```
res = residuals(catheter.lm)
(sig2hat = sum(res^2)/(12-2-1))
```

```
## [1] 14.27543
```

```
(rse = sqrt(sig2hat))
```

```
## [1] 3.778284
```

## Covariance Matrix

Given  $U=MV$ , then  $\Sigma_U = M\Sigma_V M^t$ , we can get  $\Sigma_{\hat{\beta}} = (X^t X)^{-1} X^t \sigma^2 I ((X^t X)^{-1} X^t)^t = (X^t X)^{-1} X^t \sigma^2 I X (X^t X)^{-1} = \sigma^2 (X^t X)^{-1}$

```
(XtXinv <- solve(t(X)%*%X))
```

```
##                x1                x2
##      4.9262358 -0.197172444  0.081695704
## x1 -0.1971724  0.008363701 -0.003681904
## x2  0.0816957 -0.003681904  0.001754749
```

```
summary(catheter.lm)$cov.unscaled
```

```
##      (Intercept)      ht      wt
## (Intercept)  4.9262358 -0.197172444  0.081695704
## ht          -0.1971724  0.008363701 -0.003681904
## wt           0.0816957 -0.003681904  0.001754749
```

```
(catheter.cov <- sig2hat*XtXinv)
```

```
##              x1              x2
##    70.324134 -2.81472140  1.16624129
## x1 -2.814721  0.11939543 -0.05256076
## x2  1.166241 -0.05256076  0.02504980
```

## Hypothesis Test

Say, if we want to test  $H_0 : \hat{\beta}_1 = 0$ , we can calculate  $t - stat = \frac{\hat{\beta}_1 - 0}{se(\hat{\beta}_1)}$ , then  $p - value = 2 * Pr(t_{12-9-1} \geq |t - stat|)$

```
2*(1-pt(BETAhat[2,1]/sqrt(catheter.cov[2,2]), 12-2-1))
```

```
##              x1
## 0.5570028
```

Say, if we want to get a  $(1-\alpha)\%$  confidence interval for  $\hat{\beta}_1$ , then  $\hat{\beta}_1 \pm t_{12-9-1}(1 - \alpha/2) * se(\hat{\beta}_1)$

```
(max = BETAhat[2,1] + sqrt(catheter.cov[2,2]) * qt(1-0.05/2, 12-2-1))
```

```
##              x1
## 0.992405
```

```
(min = BETAhat[2,1] - sqrt(catheter.cov[2,2]) * qt(1-0.05/2, 12-2-1))
```

```
##              x1
## -0.5709104
```

$R^2$

Correlation (otherwise known as “R”) is a number between 1 and -1 where a value of +1 implies that an increase in x results in some increase in y, -1 implies that an increase in x results in a decrease in y, and 0 means that there isn’t any relationship between x and y. Like correlation,  $R^2$  tells you how related two things are. However, we tend to use  $R^2$  because it’s easier to interpret.  $R^2$  measures how much of the total variability is explained by our model.

Sum of Squares Total/SST:  $\sum_{i=1}^n (y_i - \bar{y})^2$  = Sum of Squares Regression/SSR:  $\sum_{i=1}^n (\hat{y}_i - \bar{y})^2$  + Sum of Squares Error/SSE/RSS:  
 $\sum_{i=1}^n (y_i - \hat{y}_i)^2$   
 $R^2 = \frac{SSR}{SST}$

```
yhat = X%%BETAhat
sst = sum((y-mean(y))^2)
ssr = sum((yhat-mean(y))^2)
(rs = ssr/sst)
```

```
## [1] 0.8253572
```

Adjusted  $R^2 = 1 - \frac{(1-R^2)*(n-1)}{(n-k-1)}$ . The adjusted  $R^2$  is always smaller than the  $R^2$ , as it penalises excessive use of variables.

```
(1- (1-rs)*(12-1)/(12-2-1))
```

```
## [1] 0.7865477
```

## F-Test

The F-statistic for the added variable test is defined in terms of the residual sums of squares (RSS) for the two models and the number of explanatory variables k in each model:

$$f_0 = \frac{(RSS_S - RSS_F)/(k_F - k_S)}{RSS_F/(n - k_F - 1)}$$

p-value= $\Pr(F \geq f_0)$  where  $F \sim F_{k_F - k_S, n - k_F - 1}$  e.g. the bottom line of the output from summary is an added F-test where the submodel is the null model (just contains the intercept) and the full model contains all of the regressors.

Setup

```
null.lm<-lm(ca~1,data=catheter.df)
ht.lm<-lm(ca~ht,data=catheter.df)
wt.lm<-lm(ca~wt,data=catheter.df)
full.lm<-lm(ca~ht+wt,data=catheter.df)
```

null vs. height

```
anova(null.lm, ht.lm)
```

```
## Analysis of Variance Table
##
## Model 1: ca ~ 1
## Model 2: ca ~ ht
##   Res.Df    RSS Df Sum of Sq    F    Pr(>F)
## 1      11 735.67
## 2      10 149.29  1    586.38 39.278 9.295e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

null vs. weight

```
anova(null.lm, wt.lm)
```

```
## Analysis of Variance Table
##
## Model 1: ca ~ 1
## Model 2: ca ~ wt
##   Res.Df    RSS Df Sum of Sq    F    Pr(>F)
## 1      11 735.67
## 2      10 133.79  1    601.88 44.987 5.317e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

height vs. full

```
anova(ht.lm, full.lm)
```

```
## Analysis of Variance Table
##
## Model 1: ca ~ ht
## Model 2: ca ~ ht + wt
##   Res.Df    RSS Df Sum of Sq    F Pr(>F)
## 1      10 149.29
## 2       9 128.48  1    20.81 1.4578 0.2581
```

weight vs. full

```
anova(wt.lm, full.lm)
```

```
## Analysis of Variance Table
##
## Model 1: ca ~ wt
## Model 2: ca ~ ht + wt
##   Res.Df    RSS Df Sum of Sq    F Pr(>F)
## 1      10 133.79
## 2       9 128.48  1    5.3104 0.372 0.557
```

full vs. null

```
anova(null.lm, full.lm)
```

```
## Analysis of Variance Table
##
## Model 1: ca ~ 1
## Model 2: ca ~ ht + wt
##   Res.Df    RSS Df Sum of Sq    F    Pr(>F)
## 1      11 735.67
## 2       9 128.48  2    607.19 21.267 0.0003888 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We can conclude that we need height or weight, but when one is present, the other provides little additional information (there's probably some correlation between two variables)

## Predicted Values

Applying the result to ht=44 and wt=35, the expected value is

```
new = cbind(1, 44, 35)
(estmate = new %*% BETAhats)
```

```
##           [,1]
## [1,] 36.33697
```

Estimated Variance

```
(variance = new %*% catheter.cov %*% t(new))
```

```
##           [,1]
## [1,] 4.213942
```

95% confidence interval for the expected value

```
(min = estimate - sqrt(variance) * qt(1-0.05/2, 12-2-1))
```

```
##           [,1]
## [1,] 31.69324
```

```
(max = estimate + sqrt(variance) * qt(1-0.05/2, 12-2-1))
```

```
##           [,1]
## [1,] 40.9807
```

95% prediction interval (an interval for a single observation) for the expected value



```
(min = estimate - sqrt(sig2hat+variance) * qt(1-0.05/2, 12-2-1))
```

```
##           [,1]  
## [1,] 26.60986
```

```
(max = estimate + sqrt(sig2hat+variance) * qt(1-0.05/2, 12-2-1))
```

```
##           [,1]  
## [1,] 46.06408
```

Using 'predict' function

```
new.df <- data.frame(ht=44, wt=35)  
predict(catheter.lm, new.df, interval="confidence")
```

```
##           fit      lwr      upr  
## 1 36.33697 31.69324 40.9807
```

```
predict(catheter.lm, new.df, interval="prediction")
```

```
##           fit      lwr      upr  
## 1 36.33697 26.60986 46.06408
```