Stats 401 Lab 6

401 GSI team

2/8/2018 and 2/9/2018

Review - Expectation

Recall for stats 250, if X is a discrete random variable, which takes value c_i with probability p_i, we have

$$\mathbb{E}(X) = \sum c_i p_i$$

If X is a continuous random variable with density f, we have

$$\mathbb{E}(X) = \int x f(x) dx$$

Expectation is a linear operator. For example:

$$\mathbb{E}(aX + bY + c) = a\mathbb{E}(X) + b\mathbb{E}(Y) + c$$

Let $\mathbf{X} = (X_1, ..., X_p)^\top$ be a p dimension random vector,

$$\mathbb{E}(\mathbf{X}) = (\mathbb{E}(X_1), ..., \mathbb{E}(X_p))^\top$$

For q by p matrix $\mathbb{A}$ and $\mathbb{C}$, we have

$$\mathbb{E}(\mathbb{A}\mathbf{X} + \mathbb{C}) = \mathbb{A}\mathbb{E}(\mathbf{X}) + \mathbb{C}$$

Review - Variance and covariance

For random variable X and Y, we have,

$$Var(X) = \mathbb{E}[(X - \mathbb{E}(X))^2] = \mathbb{E}(X^2) - [\mathbb{E}(X)]^2$$

$$Cov(X, Y) = \mathbb{E}[(X - \mathbb{E}(X))(Y - \mathbb{E}(Y))] = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)$$

Some useful results:

- $Var(aX + c) = a^2 Var(X)$
- $Var(X + Y) = Var(X) + Var(Y) + 2Cov(X, Y)$
- $Cov(aX + bY, Z) = aCov(X, Z) + bCov(Y, Z)$

You should be able to show the above results using definition of variance/covariance and propeties of expectation

Review - Variance and covariance

For p dimentional random varaible $\mathbf{X} = (X_1, ..., X_p)^\top$

$$Var(\mathbf{X}) = \begin{bmatrix} Var(X_1) & Cov(X_1, X_2) & \dots & Cov(X_1, X_p) \\ Cov(X_2, X_1) & Var(X_2) & \dots & Cov(X_2, X_p) \\ & \dots & \dots & \\ Cov(X_p, X_1) & Cov(X_p, X_2) & \dots & Var(X_p) \end{bmatrix}$$

For q by p matrix $\mathbb{A}$ and $\mathbb{C}$, we have

$$Var(\mathbb{A}\mathbf{X} + \mathbb{C}) = \mathbb{A}Var(\mathbf{X})\mathbb{A}^\top$$

In lab activity 1

- Compute $Cov(aX + bY + c, X - Y)$
- Suppose $\mathbf{X} = (X_1, X_2)^\top$ where X_1 and X_1 independently follows standard normal distribution. Calculate $Var(\mathbb{A}\mathbf{X})$ where

$$\mathbb{A} = \begin{bmatrix} 1 & 2 \\ -1 & 2 \end{bmatrix}$$

Review - linear model

Population model:

$$\mathbf{Y} = \mathbb{X}\beta + \epsilon$$

β can be estimated by $\widehat{\beta} = (\mathbb{X}^\top \mathbb{X})^{-1}\mathbb{X}^\top \mathbf{Y}$
We have seen in class that $\mathbb{E}(\widehat{\beta}) = \beta$ and $Var(\widehat{\beta}) = \sigma^2(\mathbb{X}^\top \mathbb{X})^{-1}$
Sample version:

$$\mathbf{y} = \mathbb{X}\mathbf{b} + \mathbf{e}$$

where $\mathbf{b} = (\mathbb{X}^\top \mathbb{X})^{-1}\mathbb{X}^\top \mathbf{y}$
Let $s^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-p}$ be an estimator of σ^2, then $Var(\widehat{\beta})$ can be estimated by $s^2(\mathbb{X}^\top \mathbb{X})^{-1}$

Review - linear model

```r
# Look at birthwt data
library(MASS)
data(birthwt)
head(birthwt, n=4)
```

```
##     low age lwt race smoke ptl ht ui ftv  bwt
## 85    0  19 182    2     0   0  0  1   0 2523
## 86    0  33 155    3     0   0  0  0   3 2551
## 87    0  20 105    1     1   0  0  0   1 2557
## 88    0  21 108    1     1   0  0  1   2 2594
```

```r
# Transform the data. Want to look at log of birth weight
birthwt$log_bwt <- log(birthwt$bwt)
# Fit the linear regression with log_bwt as response;
# age,lwt,smoke as predictor
fit1 <- lm(log_bwt ~ age + lwt + smoke, data = birthwt)
```

```r
summary(fit1)
```

```
##
## Call:
## lm(formula = log_bwt ~ age + lwt + smoke, data = birthwt)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.31748 -0.14279  0.04364  0.19970  0.53810
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.800e+00  1.175e-01  66.380   <2e-16 ***
## age         -5.657e-05  3.878e-03  -0.015   0.9884
## lwt          1.460e-03  6.720e-04   2.172   0.0311 *
## smoke       -9.233e-02  4.135e-02  -2.233   0.0267 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
##
## Residual standard error: 0.277 on 185 degrees of freedom
## Multiple R-squared:  0.05279,    Adjusted R-squared:  0.03743
## F-statistic: 3.437 on 3 and 185 DF,  p-value: 0.01806
```

```r
# design matrix
X <- model.matrix(fit1)
y <- birthwt$log_bwt

b <- solve(t(X) %*% X) %*% t(X) %*% y;b
```

```
##                      [,1]
## (Intercept)  7.800052e+00
## age         -5.657257e-05
## lwt          1.459999e-03
## smoke       -9.233030e-02
```

```r
# find residual standard error
y_hat <- X %*% b
s <-sqrt(sum((y - y_hat)^2)/(nrow(birthwt)-4));s
```

```
## [1] 0.2769809
```

```r
# find standard error for b
b_se <- s*sqrt(diag(solve(t(X) %*% X)));b_se
```

```
##  (Intercept)          age          lwt        smoke
## 0.1175058373 0.0038784596 0.0006720396 0.0413464378
```

In lab activity 2

Still use the birthwt data.

1. Use subset function to construct a sub-dataset that only contains observations "race==1"
2. Within this sub-dataset, use lm() to fit a linear model using age and log(lwt) as predictors for the response log(bwt)
3. Use the design matrix and the response variable to compute the standard error of b. Compare your result with (2).

Lab ticket

- Calculate $\mathbb{E}(\hat{\mathbf{Y}})$ and $Var(\hat{\mathbf{Y}})$, where $\hat{\mathbf{Y}} = \mathbb{X}\hat{\beta}$.
- What about $\mathbb{E}(\hat{\mathbf{y}})$ and $Var(\hat{\mathbf{y}})$, where $\hat{\mathbf{y}} = \mathbb{X}\mathbf{b}$. (You can actully answer this question without any computation)