BDA 23: Multiple linear regression

We're going to:

1. Perform a multiple linear fit to some data according to least squares, pointing out the assumptions made in such a fit as we do so.
2. Estimate the common variance,
3. Use the estimated common variance to compute the standard errors for each coefficient, and plot these against the estimated coefficients.
4. Evaluate the F-statistic for all coefficients in the model.

The dataset we'll use to do this is from a study examining the correlation between the level of prostate-specific antigen and several clinical measurements in men who were about to have their prostates removed.

```
data <- read.table(file = 'prostate.tsv')
data <- data[ , c(1, 2, 9)] # Keep lcavol, lweight, lpsa
```

We scale and center all variables, so that it's possible to directly compare effect sizes.

```
unscaled_data <- data
data          <- data.frame(apply(data, 2, FUN = function(col) (col - mean(col))/sqrt(var(col))))
print(apply(data, 2, function(col) var(col))) # Check the variances are all 1
```

```
##  lcavol lweight    lpsa
##       1       1       1
```

Next we fit a linear model. We're going to do this "by hand" to make it really clear what's going on. Of course, R has built-in functions to fit linear models and assess the sampling properties of their coefficients. Here though, we're going to set up the design matrix $\mathbf{X}$ and the response vector $\mathbf{y}$, then estimate the coefficients using the normal equation:

$$\hat{\beta} = (\mathbf{X^T X})^{-1} \mathbf{X^T y} \tag{1}$$

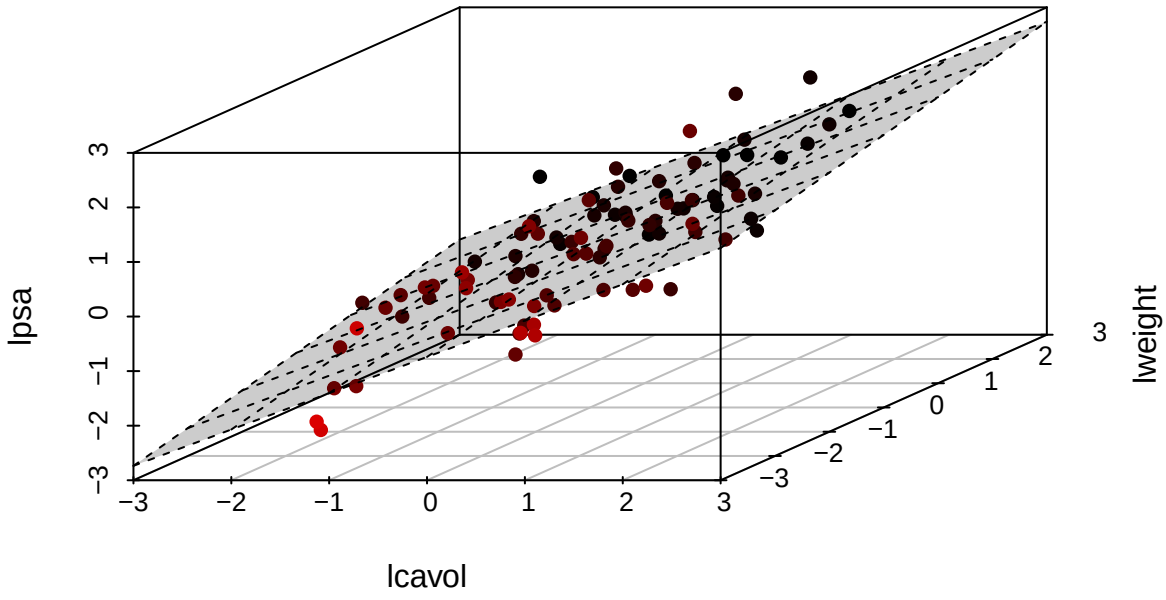
This is the best linear unbiased estimator under the assumptions that:

- the relationship between the response and inputs is indeed linear, $E[y|\mathbf{X}] = \mathbf{X}\beta$,
- the responses are independent conditional on the inputs (i.e. values drawn for the same input are independent - there is no systematic deviation in the response produced),
- and the two predictors aren't collinear (this is necessary to...).

What this means is that least-squares will, on average, provide the most accurate estimate of the coefficients β such that $E[y|\mathbf{X}] = \mathbf{X}\beta$, given that $E[\hat{\beta}] = \beta$.

All right, enough waffle: let's calculate the coefficients and plot the resulting plane:

```
library(MASS)
library(scatterplot3d)
X <- as.matrix(cbind(matrix(1, nrow = dim(data)[1], ncol = 1), data[ , -3]))
y <- as.matrix(data[ , 3])
N <- dim(X)[1]
p <- dim(X)[2] - 1
beta_hat <- ginv(t(X) %*% X) %*% t(X) %*% y
```



This plot suggests that the log of prostate-specific antigen increases with the log of prostate volume and the log of prostate density. The increase with density is minor however, so we'll need to look at the amount of uncertainty in the coefficient estimates.

If we assume the variance of the errors, σ^2, is constant across all responses, then the variance of the coefficient estimates will be:

$$\mathrm{Var}(\hat{\beta}) = (\mathbf{X^T X})^{-1}\sigma^2 \tag{2}$$

In reality the common variance won't be known, so it needs to be estimated. Since we're estimating the expected response in $p+1$ dimensions, the unbiased estimate of the response's variance is:

$$\hat{\sigma}^2 = \frac{(\mathbf{y} - E\mathbf{y}|\mathbf{X})^T(\mathbf{y} - E\mathbf{y}|\mathbf{X})}{N - p - 1} \tag{3}$$

where there are N observations and p non-constant inputs.

If we now assume that the errors in the response are Gaussian, then it follows that the distribution of the coefficient estimates will also be Gaussian:

$$\hat{\beta} \sim \mathrm{N}(\beta, (\mathbf{X^T X})^{-1}\sigma^2) \tag{4}$$

The standard error on these estimates is estimated by taking the square root of the coefficient's sample variance:

$$\mathrm{SE}(\hat{\beta}) = (\mathrm{diag}((\mathbf{X^T X})^{-1})\hat{\sigma}^2)^{\circ\frac{1}{2}} \tag{5}$$

The normality assumption means that these standard errors can be used to run t-tests. We have about 100 samples in this dataset and only three features, which means that the normal approximation to to the t-distribution is pretty good.

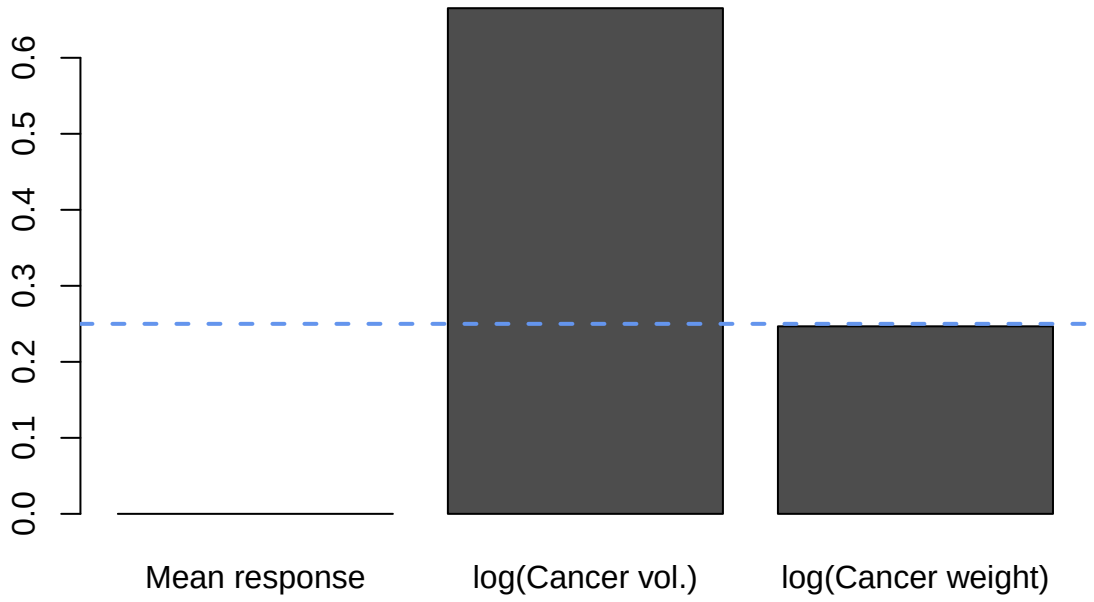
The one-sided 99% confidence limit is

```
n_se <- qnorm(0.99)
print(n_se)
```

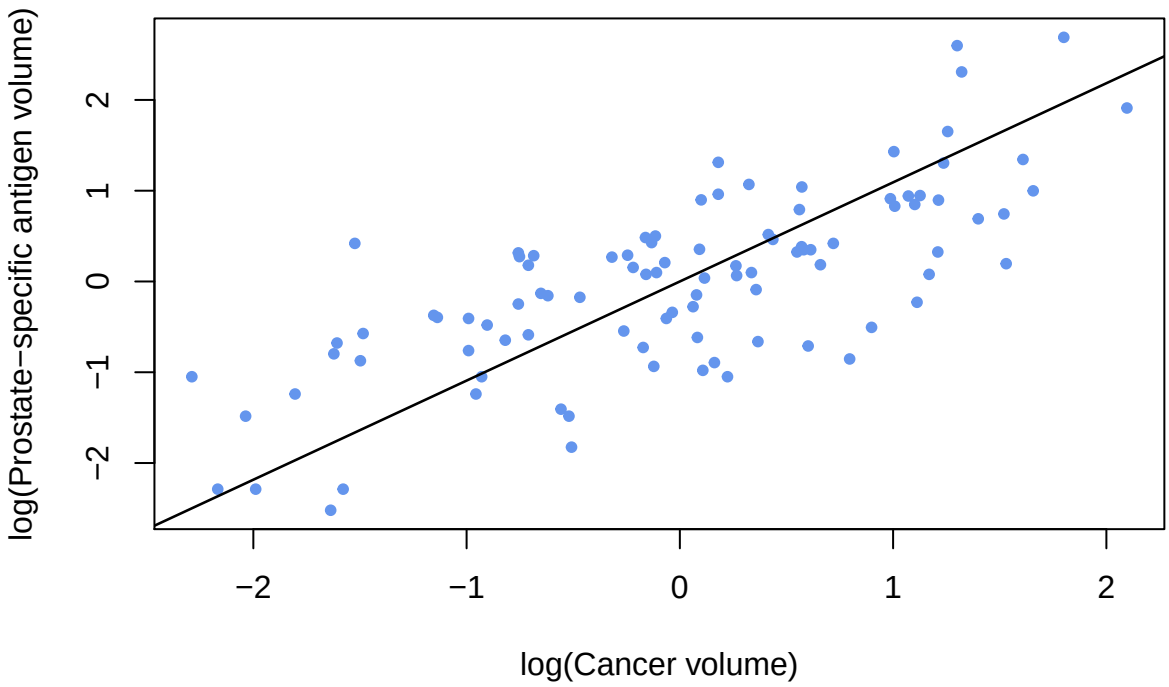
```
## [1] 2.326348
```

standard errors from zero. Our acceptance criteria for the effects to be non-zero then, is that the coefficients are at least 2.33 standard errors from zero.

```
var_hat <- sum(sapply(y, function(t) (t - mean(y))^2))/(N - p - 1)
se <- sqrt(diag(ginv(t(X) %*% X))*var_hat)
z <- beta_hat/(n_se*se)
```



We can see that the size of the cancer influences the amount of antigen it produces. Its density is also quite likely to have an effect, however more data would be needed to verify this. By rescaling the coefficients we can see the rate of increase of log(`psa`) with log(`cavol`).


log(Prostate−specific antigen volume)
−2
−1
0
1
2
log(Cancer volume)