Now consider including more than one covariate, so the model is

$$\begin{aligned} Y_i &= \sum_{j=1}^{k} \beta_j X_{ij} + \epsilon_i \\ &= x_i' \beta + \epsilon_i \end{aligned}$$

where $x_i = (X_{i1}\ X_{i2}\ \ldots\ X_{ik})'$ and $\beta = (\beta_1\ \beta_2\ \ldots\ \beta_k)'$.

Then the model for the vector of the observations $Y = (Y_1\ Y_2\ \ldots\ Y_n)'$ is

$$Y = X\beta + \epsilon$$

where X is the $n \times k$ matrix with i^{th} row x_i' and $\epsilon = (\epsilon_1\ \epsilon_2\ \ldots\ \epsilon_n)'$. Usually the model will contain an intercept, with $X_{i1} = 1$ for all i. That is, the first column of X contains all 1's.

Example (Hamilton, 1983)

In the late 1970's, the city of Concord New Hampshire experienced a growing demand for water, despite having roughly stable population. In 1979 and 1980 there was a shortage of water, leading to a media campaign to persuade citizens to use less. Over the next year, water use declined by about 15%. The 1981 Concord Water Study examined what variables were associated with water use.

$Y =$ Summer 1981 water use (cubic feet)
$X_1 =$ Household income (thousands of dollars)
$X_2 =$ Summer 1980 water use (cubic feet)
$X_3 =$ Highest household education (years)
$X_4 =$ Retired: 1 for yes, 0 for no
$X_5 =$ People living in the house in 1981
$X_6 =$ Increase in number of people from 1980

```
summary(lm(log(water81)~income+water80+educat+
                        retire+peop81+cpeop, data = water))
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.629e+00  1.072e-01  61.841  < 2e-16 ***
income       7.972e-03  1.795e-03   4.442 1.10e-05 ***
water80      2.035e-04  1.365e-05  14.908  < 2e-16 ***
educat      -9.795e-03  6.850e-03  -1.430   0.1534
retireyes   -1.028e-01  4.924e-02  -2.088   0.0373 *
peop81       1.083e-01  1.488e-02   7.278 1.36e-12 ***
cpeop        8.113e-02  4.172e-02   1.945   0.0524 .
```

How do we fit the model, interpret it, and decide which variables to keep?

We will consider the case that the ϵ_i's are iid normal. However, it will be convenient to write distributions in terms of the multivariate normal distribution.

Recall that we write $Z \sim N(\mu, \Sigma)$ to denote that Z is multivariate normal with $E[Z_i] = \mu_i$ and $Cov(Z_i, Z_j) = \Sigma_{ij}$.

We will make use of the fact that if $Z \sim N(\mu, \Sigma)$, then

$$AZ \sim N(A\mu, A\Sigma A').$$

The multivariate regression model with normal errors is

$$Y \sim N(X\beta, \sigma^2 I_n)$$

where I_n is the $n \times n$ identity matrix.

The likelihood is

$$f(y|\beta, \sigma^2) = (2\pi\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}(y - X\beta)'(y - X\beta)\right\}$$

Note that the likelihood involves $(Y - X\beta)'(Y - X\beta)$, which is the residual sum of squares (RSS) for this model.

$$RSS = (Y - X\beta)'(Y - X\beta) = ||Y - X\beta||^2 = \sum_{i=1}^{N}(Y_i - x_i'\beta)^2$$

As before, maximizing the likelihood with respect to β is equivalent to minimizing RSS.

Setting $\frac{\partial}{\partial\beta}RSS = -2X'Y + 2X'X\beta \equiv 0_k$ implies

$$X'X\beta = X'Y$$

These are called the "normal equations." They have a unique solution if and only if $X'X$ is nonsingular, which is true if and only if the rank of X is k. That is, the regressors may not be linear combinations of one another.

The unique solution to the normal equations in this case (and correspondingly, the MLE for β when the data are normal) is

$$\hat{\beta} = (X'X)^{-1}X'Y$$

Note that $\hat{\beta} = AY$, where $A = (X'X)^{-1}X'$. Therefore,

$$\hat{\beta} \sim N(AX\beta, A[\sigma^2 I_n]A') = N(\beta, \sigma^2(X'X)^{-1})$$

The interpretation for $\hat{\beta}_j$ is a change in the mean of Y, holding all the other covariates constant. This is important, because it means that the interpretation of $\hat{\beta}_j$ will change, depending on which variables are in the model.

To form confidence intervals for the elements of β, we can use the result on the previous slide, but replacing σ^2 by an estimate, which we will derive next. Take

$$\hat{\beta}_j \pm z_{\alpha/2}\widehat{se}(\hat{\beta}_j),$$

where

$$\widehat{se}(\hat{\beta}_j) = \hat{\sigma}^2[(X'X)^{-1}]_{jj}.$$

Slutzky's Theorem gives us $\frac{\hat{\beta}_j - \beta_j}{\widehat{se}(\hat{\beta}_j)} \xrightarrow{D} N(0,1)$, which we can use to construct a Wald test.

The MLE for σ^2 is $\hat{\sigma}^2 = \frac{1}{n}(Y - X\hat{\beta})'(Y - X\hat{\beta})$.

Note that

$$\begin{aligned} Y - X\hat{\beta} &= Y - X(X'X)^{-1}X'Y \\ &= (I_n - X(X'X)^{-1}X')Y \\ &\equiv PY \end{aligned}$$

P is a symmetric and idempotent matrix, meaning $P' = P$ and $PP = P$. Therefore,

$$\hat{\sigma}^2 = \frac{1}{n}(PY)'(PY) = \frac{1}{n}Y'P'PY = \frac{1}{n}Y'PY$$

The term $Y'PY$ is known as a quadratic form in Y. A general result for quadratic forms gives us that

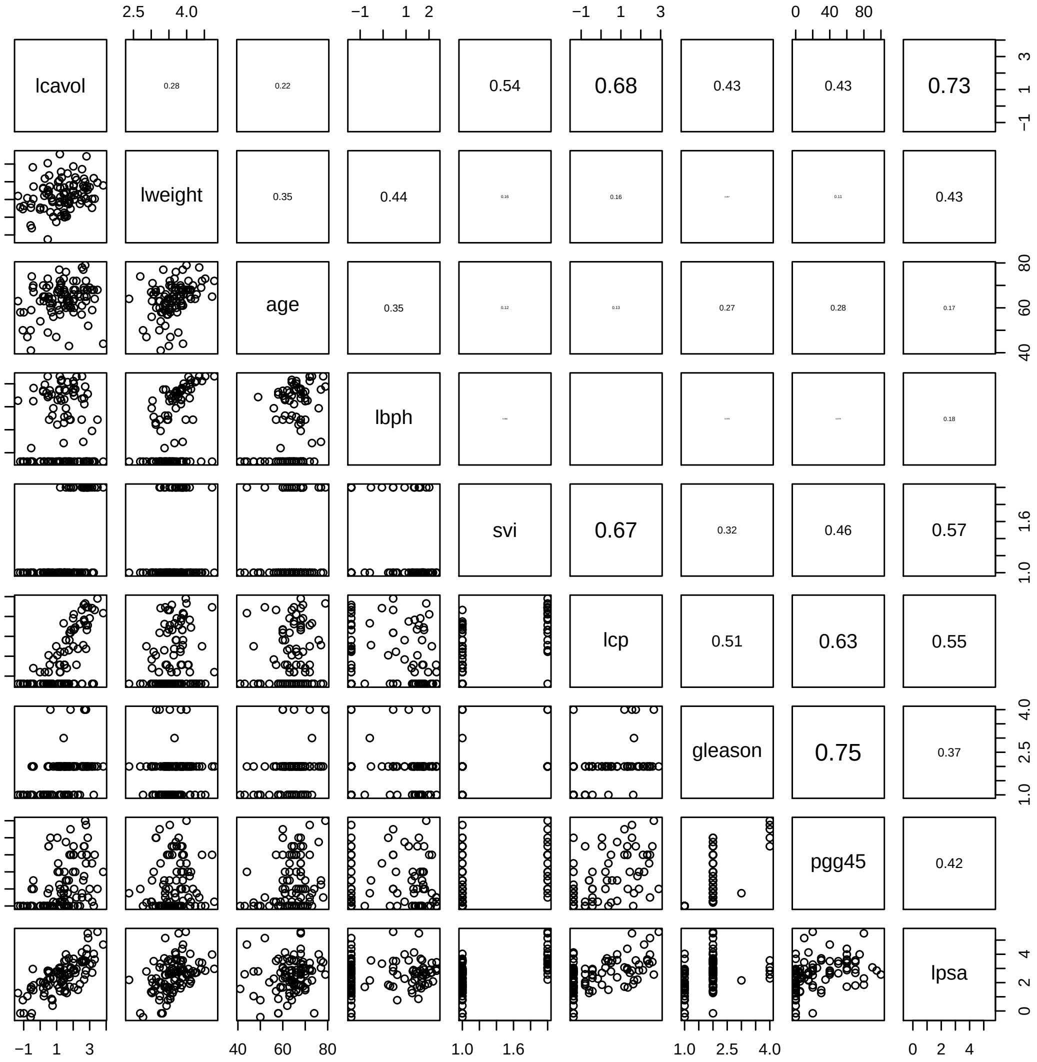
$$\begin{aligned}
E[Y'PY] &= tr\{PCov(Y)\} + E[Y]'PE[Y] = \sigma^2(n-k) \\
&= tr\{P\sigma^2 I_n\} + (X\beta)'P(X\beta) \\
&= \sigma^2 tr\{I_n - X(X'X)^{-1}X'\} + (X\beta)'[I - X(X'X)^{-1}X'](X\beta) \\
&= \sigma^2[tr\{I_n\} - tr\{X(X'X)^{-1}X'\}] \\
&= \sigma^2[n - tr\{X'X(X'X)^{-1}\}] \\
&= \sigma^2(n-k)
\end{aligned}$$

Therefore, $E[\hat{\sigma}^2] = \frac{n-k}{n}\sigma^2$, and we can also use this to form an unbiased estimator.

The term “model selection” refers to choosing a single model from within a class of models under consideration, in this case, linear regression models.

Here is an example, taken from “The Elements of Statistical Learning” by Hastie, Tibshirani, and Friedman. Stamey et al. (1989) measured the level of prostate-specific antigen and other clinical measures in men who were about to receive a radical prostatectomy. The variables are

`lpsa` = log prostate-specific antigen
`locavol` = log cancer volume
`lweight` = log prostate weight
`age` = patient’s age
`lbph` = log benign prostatic hyperplasia
`svi` = seminal vesicle invasion (categories 0 or 1)
`lcp` = log capsular penetration
`gleason` = Gleason score (categories 6, 7, 8, 9)
`ppg45` = percent of Gleason scores 4 or 5


lcavol
lweight
age
lbph
svi
lcp
gleason
pgg45
lpsa
0.28
0.22
0.54
0.68
0.43
0.43
0.73
0.35
0.44
0.16
0.16
0.11
0.43
0.35
0.12
0.13
0.27
0.28
0.17
0.18
0.67
0.32
0.46
0.57
0.51
0.63
0.55
0.75
0.37
0.42
2.5
4.0
−1
1
2
−1
1
3
0
40
80
3
1
−1
4.0
2.5
80
60
40
2
1
−1
1.6
1.0
3
1
−1
4.0
2.5
1.0
80
40
0
4
2
0
−1
1
3
40
60
80
1.0
1.6
1.0
2.5
4.0
0
2
4

Here is some output from fitting a multiple regression model in R, including all the covariates.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.486274   0.885929   0.549  0.58451
lcavol       0.549532   0.090087   6.100 2.94e-08 ***
lweight      0.623816   0.200463   3.112  0.00252 **
age         -0.023103   0.011282  -2.048  0.04364 *
lbph         0.091523   0.058454   1.566  0.12108
svi1         0.744537   0.244602   3.044  0.00310 **
lcp         -0.124612   0.094565  -1.318  0.19109
gleason7     0.253874   0.216141   1.175  0.24341
gleason8     0.480520   0.760895   0.632  0.52938
gleason9    -0.036003   0.494866  -0.073  0.94217
pgg45        0.004902   0.004622   1.061  0.29184
```

We already have one possible tool for model selection, which is hypothesis testing. Particularly if the number of possible regressors is small, we can consider testing $H_0 : \beta_j = 0 \; \forall j \in J_0$ for a set of terms J_0.

Here is an `R` function for the likelihood ratio test. Can you see what it is doing?

```
lrt <- function(mod, mod0){
  ll <- sum(dnorm(mod$residuals,
            sd = summary(mod)$sigma, log = TRUE))
  ll.0 <- sum(dnorm(mod0$residuals,
              sd = summary(mod0)$sigma, log = TRUE))
  lambda <- 2*(ll-ll.0)
  return(1 - pchisq(lambda, df = mod$rank - mod0$rank))
}
```

```
> full <- lm(lpsa˜., data = prostate)
> reduced <- lm(lpsa˜lcavol+lweight+svi, data = prostate)
> lrt(full, reduced)
[1] 0.1964859
```

Actually, we do not need the limiting χ^2 distribution in this case; we can modify the test statistic slightly to one that has an exact F distribution. (Similar reasoning applies to using the t distribution for the test statistic for individual regressors.) When the sample size is large, both tests will give similar answers.

```
> anova(full, reduced) # do the F test
...
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     86 41.811
2     93 46.568 -7   -4.7575 1.3979 0.2166
```

Hypothesis testing is not necessarily a good model selection technique, however. Some possible issues are

- Multiple testing: There are many possible combinations of regressors. How should we adjust for exploring different possibilities? Usually we choose which models to test based on looking at preliminary results, which changes the type I error rate.

- Failing to reject the null hypothesis is not the same as finding evidence for the null hypothesis. It may be that the test has low power, e.g. due to small sample size.

- There is no reason to think that decisions based on testing parameters will lead to good predictions, if that is our goal. On the other hand, if model interpretation is our goal, we often have substantive reasons for keeping non-significant regressors in the model.

Consider predicting a new observation Y^*, for covariates X^*, and let S denote a subset of the covariates in the model. If our loss function is squared error, we could choose S to minimize the frequentist risk, which is $MSPE = E[(\hat{Y}^*(S) - Y^*)^2]$, where the expectation is over both the observed Y and the new Y^*. $MSPE$ stands for "mean squared prediction error."

However, $MSPE$ may vary depending on X_*. What Wasserman calls the "prediction risk",

$$R(S) = \sum_{i=1}^{n} E[(\hat{Y}_i(S) - Y_i^*)^2]$$

sums the MSPE at all of observed values of X. This is somewhat similar to integrating MSPE over the distribution of X^*.

A natural estimate for $R(S)$ is the training error

$$\hat{R}_{tr}(S) = \sum_{i=1}^{n} (\hat{Y}_i(S) - Y_i)^2,$$

which is related to R^2 (for a particular S) by

$$R^2(S) = 1 - \frac{\sum_{i=1}^{n} (\hat{Y}_i(S) - Y_i)^2}{\sum_{i=1}^{n} (Y_i - \bar{Y})} = 1 - \frac{\hat{R}_{tr}(S)}{\sum_{i=1}^{n} (Y_i - \bar{Y})}$$

However $\hat{R}_{tr}(S)$ (or, equivalently, $R^2(S)$) is bad criterion for model selection. In fact, by construction, $\hat{R}_{tr}(S)$ can only decrease (R^2 can only increase) as we add terms to S, whereas the true $R(S)$ tends to eventually increase. More complicated selection criteria penalize extra model complexity and thereby decrease this bias.

“Adjusted R^2” is reported by many statistical packages. This is just

$$1 - \frac{n-1}{n-k} \frac{RSS}{TSS}$$

where n is the number of observations and k is the number of regressors in the model (the dimension of S). However, there is no particular theory guiding this choice.

Mallow’s C_p statistic is

$$\hat{R}(S) = \hat{R}_{tr}(S) + 2k\hat{\sigma}^2$$

where k is the dimension of S and $\hat{\sigma}^2$ is the unbiased estimate of σ^2 *under the full model*. The second term is a bias correction.

The Akaike Information Criterion (AIC) is

$$AIC(S) = \ell_n(\hat{\beta}_S, \hat{\sigma}^2_S) - k$$

where $\hat{\beta}_S$ and $\hat{\sigma}^2_S$ are the MLEs under model S, ℓ_n is the log-likelihood, and k is the number of regressors in S. It is an estimate of the risk function under a different loss function, which is the Kullback Leibler distance between the estimated and true probability distributions.

The Bayesian Information Criterion (BIC) imposes a stronger penalty, with

$$BIC(S) = \ell_n(\hat{\beta}_S, \hat{\sigma}^2_S) - \frac{k}{2}\log n$$

Maximizing BIC is approximately equivalent to choosing the model with highest posterior probability, if the prior distribution assigns equal probability to all models.

Another way of estimating the risk is to set aside a (randomly selected) subset of our data and see how well we can predict it using a model constructed from the rest. In the machine learning literature, these are called the “validation” and “training” data sets.

The risk estimator is

$$\hat{R}_V(S) = \sum_{i=1}^{m} (\hat{Y}_i^*(S) - Y_i^*)^2$$

where m is the size of the validation data set. How big m is will depend on how much data you have “to spare” from your total data set. Common practice is to take m to be $1/4$ to $1/2$ the total data size.

This approach is really only feasible if you have a lot of data to begin with, otherwise the parameters in each model may not be well estimated.

An adaptation of this idea uses almost all the data for fitting, but it approximates the risk by repeatedly fitting the model to all but one data point. This is called leave-one-out cross-validation. The risk estimator is

$$\hat{R}_{CV}(S) = \sum_{i=1}^{n} (Y_i - \hat{Y}_{(i)})^2$$

where $\hat{Y}_{(i)}$ is the prediction for Y_i obtained by fitting the model using all the data *except for* the i^{th} observation.

Luckily, we don't actually have to recalculate the model n times, since

$$\hat{R}_{CV}(S) = \sum_{i=1}^{n} \left(\frac{Y_i - \hat{Y}_i(S)}{1 - U_{ii}(S)} \right)^2$$

where $U(S) = X_S(X_S' X_S)^{-1} X_S$.

When the number of possible regressors is large, it can be computationally infeasible to calculate a criterion for all the possible models. Two popular methods for exploring the space of models are the forward and backward stepwise selection procedures.

Forward selection starts with the intercept-only model. The criterion is then computed for each model with a single covariate, and the one with the best possible criterion is chosen. (The way we've defined each criterion, smaller is better.) Then the criterion is computed for each model with that covariate plus another possible covariate, and so on. Backward selection starts with the full model, with all possible covariates, and at each step deletes the one that leads to the most improvement.

They will not necessarily end up in the same place, and neither is guaranteed to find the best out of all possible models.