21.

22. PCA:reduce your set of variables into the most informative.

23.

1. There should be a linear relationship between dependent (response) variable and independent (predictor) variable(s). A linear relationship suggests that a change in response Y due to one unit change in X^1 is constant, independent of other variables.

2. There should be no correlation between the residual (error) terms. (No auto-correlation, the estimated standard errors tend to underestimate the true standard error.)

3. The independent variables should not be correlated. (not multicollinearity)

4. The error terms must have constant variance. (homoskedasticity)

5. The error terms must be normally distributed.

24.

Here the coefficient $\hat{\beta}_1$ is interpreted as the average increase in y for unit increase in $x^{(1)}$, *provided all other explanatory variables* $x^{(2)}, \ldots, x^{(p)}$ *are kept constant.*

log(x): the average increase in y associated with a unit change in x., kept everything else constant

log(y)

Consider regression equation

$$\log y = \beta_0 + \beta_1 x.$$

Evaluating at $x + 1$, we have

$$\log(y + \Delta y) = \beta_0 + \beta_1(x + 1).$$

Subtracting the first equation from the second gives

$$\log\left(\frac{y + \Delta y}{y}\right) = \beta_1 \Rightarrow \log\left(1 + \frac{\Delta y}{y}\right) = \beta_1 \Rightarrow \frac{\Delta y}{y} \approx \beta_1$$

Since $\Delta y/y$ is the change in y divided by the original value of y, $\beta_1 \times 100\%$ can be interpreted as the percent change in y associated with a unit change in x.

both: the percent change in y associated with a percent change in x.

25.

$$\hat{y}_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \ldots + \beta_p x_{i,p}$$

n Least squares regression, the sum of the squares of the errors is minimized.

$$SSE = \sum_{i=1}^{n} (e_i)^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i,1} - \beta_2 x_{i,2} - \ldots - \beta_p x_{i,p})^2$$

ake the partial derivative of SSE with respect to β_0 and setting it to zero.

$$\frac{\partial SSE}{\partial \beta_0} = \sum_{i=1}^{n} 2(y_i - \beta_0 - \beta_1 x_{i,1} - \beta_2 x_{i,2} - \ldots - \beta_p x_{i,p})^1(-1) = -2\sum_{i=1}^{n} e_i = 0$$

TSS always equals to or larger than RSS.

TSS means using the average response to estimate y, which is the worst case. Add one more variable, RSS always decrease.

$$\begin{aligned}
\text{TSS} &= \sum_i (y_i - \bar{y})^2 \\
&= \sum_i (y_i - \hat{y}_i + \hat{y}_i - \bar{y})^2 \\
&= \sum_i (y_i - \hat{y}_i)^2 + \sum_i (\hat{y}_i - \bar{y})^2 + 2\sum_i (y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) \\
&= \text{ESS} + \text{RSS} + 2\sum_i (y_i - \hat{y}_i)(\hat{y}_i - \bar{y}).
\end{aligned}$$

26.

T lower P-value means the variable is significant

F R^2(how much of the response is explained by variables) always increase, adjusted R^2 not always increase since it has to divide dof

$$R^2_{adj}(M) = 1 - \frac{RSS(M)/(n-p-1)}{TSS/(n-1)}$$

T F-test is only valid for comparing submodels.You can't compare disjoint sets of variables with an F-test.

T happens when multicollinearity(the predictors are highly correlated). Imagine a situation where there are only two predictors with very high correlation. Individually, they both also correlate closely with the response variable. Consequently, the F-test has a low p-value (it is saying that the predictors together are highly significant in explaining the variation in the response variable). But the t-test for each predictor has a high p-value because after allowing for the effect of the other predictor, there is not much left to explain.

27.

T: there's a linear relationship between education and residuals of fitted values

T: the fitted values cannot be fully explained by prestige

29. total = 12

1 intercept for basemodel, 3 for dummy variables; 2 slopes for numeric variables, 6 slope changes for interaction variables

30.

The F-test of overall significance indicates whether your regression model provides a better fit than a model that contains no independent variables.

$$F = \frac{(RSS_0 - RSS_1)/p}{RSS_1/(n-p-1)} \quad RSS_1/(n-p-1) = \hat{\sigma}^2 \quad F = \frac{(TSS - RSS)/p}{\hat{\sigma}^2}$$

```
curve(df(x, df1 = p, df2 = n - p - 1),
```

T

31.

Permutation test:

1. Permute the order of the y_i values, so that the y_i are paired up with different x_i.
2. Fit the regression model on the permuted data
3. Calculate R^2_b
4. Repeat B times to get $R^2_1, \ldots, R^2_B$.
5. Determine the p-value of the *observed* R^2 as compared to the compute null distribution

assumption:the observations are exchangeable. Basically this means that the labels don't matter.

Bootstrap CI:

1. We create a bootstrap sample by sampling *with replacement* N times from our data $(x_1, y_1), \ldots, (x_N, y_N)$
2. This gives us a sample $(x^*_1, y^*_1), \ldots, (x^*_N, y^*_N)$ (where, remember some data points will be there multiple times)
3. Run regression on $(x^*_1, y^*_1), \ldots, (x^*_N, y^*_N)$ to get $\hat{\beta}^*_1$ and $\hat{\beta}^*_0$
4. Repeat this B times, to get

$$(\hat{\beta}^{(1)*}_0, \hat{\beta}^{(1)*}_1), \ldots, (\hat{\beta}^{(B)*}_0, \hat{\beta}^{(B)*}_1)$$

assumption:1.xi are i.i.d and bootstrap samples are S.R.S 2. sample size are large enough so taht

sample dsn~population dsn 3.test statistic is unbiased

Difference: Bootstrap keeps the combination of (xi,yi), Permutation mixed up yi with different xi

32.

prediction interval predict the average value of yi given xi i.e. E(yi|xi)

CI predict a particular observation yi for a given xi

prediction interval is larger than CI because it consider SE(ei)+SE(β j）

33.

A function in R that is useful for variable selection is `regsubsets` in the R package `leaps`. For each value of $k = 1, \ldots, p$, this function gives the best model with k variables according to the residual sum of squares.

cross-validation：

we want to minimize estimated predicted error, that is $minE(y_0 - \hat{y}(x_0))^2.$

The problem is that when you use the same data to estimate both coefficients and the prediction error, the estimate of the prediction error will underestimate the true prediction error. The lager the model is, the more it underestimate. So that we divide the data into 10 parts, and use 9 of the parts to fit/train the model and 1 part to estimate prediction error, and repeat over all 10 partitions.

34.

Why do selection:

1.Removing unnecessary variables results in a simpler model.

2.Unnecessary explanatory variables will add noise to the estimation of quantities that we are interested in/causing over-fitting

3. Collinearity (i.e. strong linear relationships in the variables) is a problem with having too many variables trying to do the same job.

4.We can save time and/or money by not measuring redundant explanatory variables.

step-wise

pros:simpler to compute

cons:one at a time approach might miss the optimal model;

regression subset:provide more information

Stepwise regression does not fit all models but instead assesses the statistical significance of the variables one at a time and arrives at a single model. Best subsets regression fits all possible models and displays some of the best candidates based on adjusted R-squared or Mallows' Cp

$$LOOCV(M) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}^{(-i)})^2$$

In fact, LOOCV can be computed very quickly in linear regression from our residuals of the model without a lot of coding using algebraic facts about regression that we won't get into.[58]

- **Mallows Cp**

$$C_p(M) = RSS(M)/n + \frac{2\hat{\sigma}^2(p+1)}{n}$$

There are other ways of writing C_p as well. $\hat{\sigma}^2$ in this equation is the estimate based on the *full* mode (with all predictors included.)

In fact, $C_p(M)$ becomes equivalent to the LOOCV as n gets large (i.e. asymptotically).

- **Akaike Information Criterion (AIC)**

$$AIC(M) = nlog(RSS(M)/n) + 2(p+1)$$

In regression, AIC is equivalent to using C_p above, only with $\hat{\sigma}^2(M)$, i.e. the estimate of σ based o the model M.

- **Bayes Information Criterion (BIC)**

$$BIC(M) = nlog(RSS(M)/n) + (p+1)log(n)$$

We would note that all of these measures, except for C_p can be used for models that are more complicated than just regression models, though AIC and BIC are calculated differently depending on the prediction model.

35.

remove variables with large p-values(not significant)

p-value is not valid because of multiple testing(while the data is correlated,);don't do data fishing(fishing variables you think important);variables you drop may still correlated with the response but not adding exploratory results

36.

(1)We want to do classification, the response variable y is binary (takes only two values; for our purposes we assume it is coded as 0 and 1) while in regression, the response variable is continuous.

The logistic regression model, for $p = P(y = 1)$, is given as:

$$\log(\frac{p}{1-p}) = \log(odds(y=1)) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p.$$

This means that we are modeling the probabilities as

$$p(x) = \frac{\exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p)}{1 + \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p)}$$

Note that this can be thought of as our model for the random process of how the data were generated: for a given value of x, you calculate the $p(x)$, and then toss a coin that has probability $p(x)$ of heads. If you get a head, $y = 1$, otherwise $y = 0$.[1] The coin toss provides the randomness – $p(x)$ is an unknown but fixed quantity.

37.

```
plot.dat <- data.frame(prob = menarche$Menarche/menarche$Total,
                       age = menarche$Age,
                       fit = predict(m, menarche))
#convert those logit values to probabilities
plot.dat$fit_prob <- exp(plot.dat$fit)/(1+exp(plot.dat$fit))
```

38.