Lecture 30: Review

Pratheepa Jeganathan

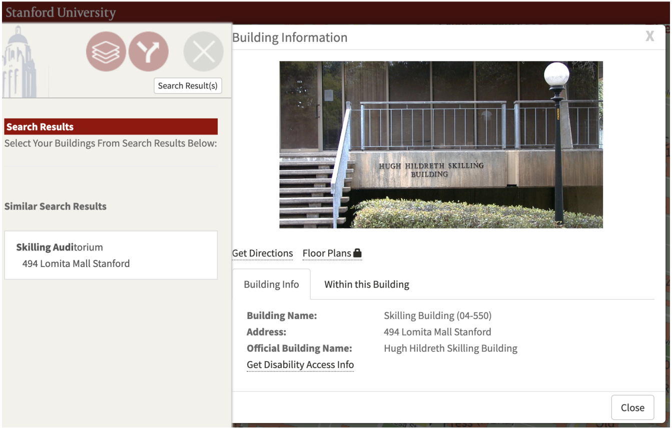
12/06/2019

Course Evaluations Now Open

- Axess is now open to complete end-term course evaluations.
- You can find it on
  - Axess > Student > Course and Section Evaluations
- If you complete all of the feedback you will see your grades by 12/13/2019 otherwise 12/17/2019

Final examination

- Final examination information
  - In-class examination.
  - Time: According the Stanford calendar: Wednesday, December 11, 2019 @ 3:30PM-6:30 PM.
  - Location: Skilling Auditorium.



Final examination

- Students **are not allowed** to take final examinations earlier than the scheduled date and time (except for the event of extraordinary circumstance that is determined solely by me).
- What to bring
  - A CALCULATOR.
  - FOUR SINGLE-SIDED PAGES OF NOTES.

Expected outcomes

By the end of the course, students should be able to:

- Enter tabular data using R.
- Plot data using R, to help in exploratory data analysis.
- Formulate regression models for the data, while understanding some of the limitations and assumptions implicit in using these models.
- Fit models using R and interpret the output.
- Test for associations in a given model.

Expected outcomes (cont.)

- Use diagnostic plots and tests to assess the adequacy of a particular model.
- Find confidence intervals for the effects of different explanatory variables in the model.
- Use some basic model selection procedures, as found in R, to find a *best* model in a class of models.
- Fit simple ANOVA models in R, treating them as special cases of multiple regression models.
- Fit simple logistic and Poisson regression models.

Evaluation

The final letter grade for this course will be determined by each method of assessment weighted as follows:

- 7 weekly homework assignments (55%)
- Midterm examination (15%)
- Final examination (30%)
- Quiz and Bonus points (5%+5.2%)

The final percentage to letter grade conversion:

A+ = 97-110.2	A = 96-94	A- = 90-93
B+ = 87-89	B = 84-86	B- = 80-83
C+ = 77-79	C = 74-76	C- = 70-73
D+ = 67-69	D = 64-66	D- = 60-63

Topics covered

- Simple linear regression.
- Diagnostics for simple linear regression.
- Multiple linear regression.
- Diagnostics.
- Interactions and ANOVA.
- Weighted Least Squares.
- Autocorrelation.
- Bootstrapping `lm`.
- Model selection.
- Multicollinearity.
- Penalized regression.
- Logistic regression.

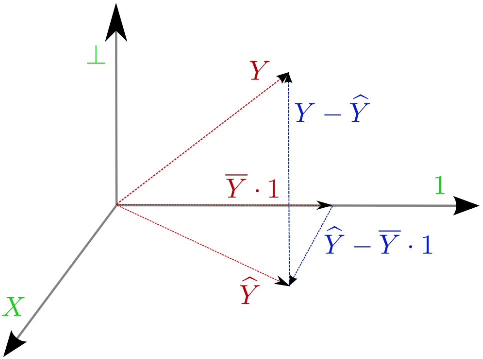
Simple linear regression

Least squares

- We used "least squares" regression. This measures the goodness of fit of a line by the sum of squared errors, SSE.
- Least squares regression chooses the line that minimizes $SSE(\beta_0, \beta_1) = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 \cdot X_i)^2$.

Geometry of least squares

- The following picture depicts the geometry involved in least squares regression.



What is a t-statistic?

- Start with $Z \sim N(0,1)$ is standard normal and $S^2 \sim \chi^2_\nu$, independent of Z.
- Compute $T = \frac{Z}{\sqrt{\frac{S^2}{\nu}}}$.
- Then, $T \sim t_\nu$ has a t-distribution with ν degrees of freedom.
- Generally, a t-statistic has the form

$$T = \frac{\hat{\theta} - \theta}{SE(\hat{\theta})}$$

Interval Estimates

- A $(1-\alpha)\cdot 100\%$ confidence interval:

$$\widehat{\beta}_1 \pm SE(\widehat{\beta}_1)\cdot t_{n-2,1-\alpha/2}.$$

- Interval for regression line $\beta_0 + \beta_1 \cdot X$
  - $(1-\alpha)\cdot 100\%$ confidence interval for $\beta_0 + \beta_1 X$:

$$\widehat{\beta}_0 + \widehat{\beta}_1 X \pm SE(\widehat{\beta}_0 + \widehat{\beta}_1 X)\cdot t_{n-2,1-\alpha/2}$$

where $SE(\widehat{\beta}_0 + \widehat{\beta}_1) = \widehat{\sigma}\sqrt{\frac{1}{n} + \frac{(\overline{X}-X)^2}{\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2}}$

Interval Estimates

- Prediction intervals for $\beta_0 + \beta_1 X_{new} + \epsilon_{new}$
  - $(1-\alpha)\cdot 100\%$ prediction interval for $\beta_0 + \beta_1 X_{new} + \epsilon_{new}$ is

$$\widehat{\beta}_0 + \widehat{\beta}_1 X_{\text{new}} \pm t_{n-2,1-\alpha/2} \cdot SE(\widehat{\beta}_0 + \widehat{\beta}_1 X_{\text{new}} + \varepsilon_{\text{new}}),$$

where $SE(\widehat{\beta}_0 + \widehat{\beta}_1 X_{\text{new}} + \varepsilon_{\text{new}}) = \widehat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(\overline{X} - X_{\text{new}})^2}{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}}.$

Sums of squares

$$
\begin{aligned}
SSE &= \sum_{i=1}^{n} (Y_i - \widehat{Y}_i)^2 = \sum_{i=1}^{n} (Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 X_i)^2 \\
SSR &= \sum_{i=1}^{n} (\overline{Y} - \widehat{Y}_i)^2 = \sum_{i=1}^{n} (\overline{Y} - \widehat{\beta}_0 - \widehat{\beta}_1 X_i)^2 \\
SST &= \sum_{i=1}^{n} (Y_i - \overline{Y})^2 = SSE + SSR \\
R^2 &= \frac{SSR}{SST} = 1 - \frac{SSE}{SST} = \widehat{Cor}(\boldsymbol{X}, \boldsymbol{Y})^2.
\end{aligned}
$$

F-test in simple linear regression

- *Full (bigger) model : FM* : $\quad Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$
- *Reduced (smaller) model: RM* : $\quad Y_i = \beta_0 + \varepsilon_i$
- The F-statistic has the form $F = \frac{(SSE(RM) - SSE(FM))/(df_{RM} - df_{FM})}{SSE(FM)/df_{FM}}$.
- Reject H_0 : RM is correct, if

$$F > F_{1-\alpha,1,n-2}$$

or

$$P - value = P\left(F_{1,n-2} > F\right) \leq \alpha.$$

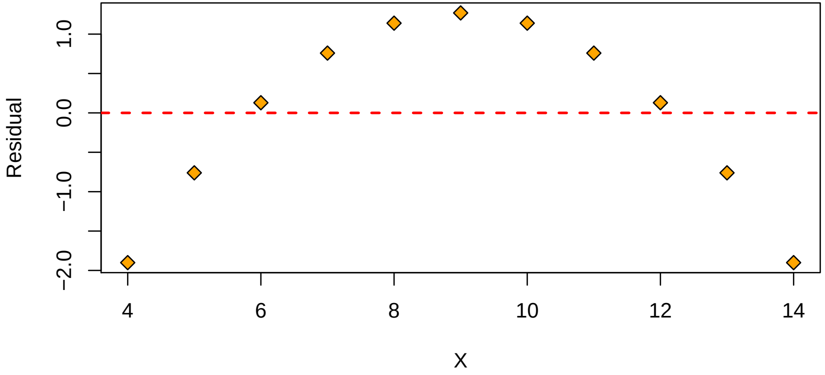
Assumptions in the simple linear regression model

- $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$
  - Errors ε_i are assumed independent $N(0, \sigma^2)$.

Diagnostics for simple linear regression

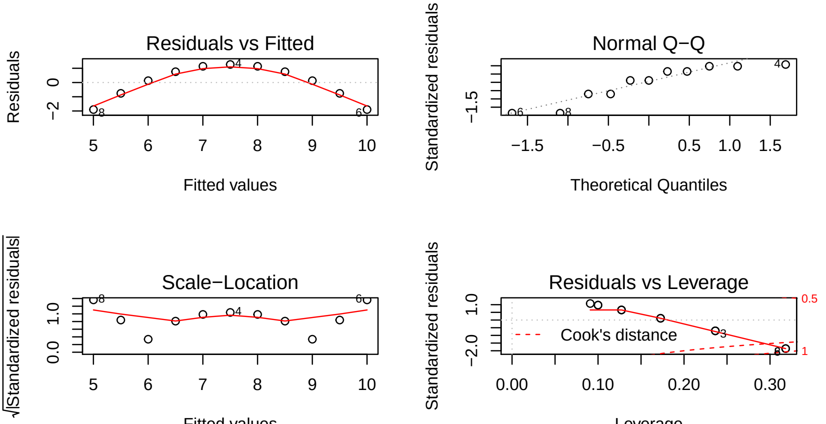
Diagnostic plots for linearity

```
simple.lm = lm(y2 ~ x2, data=anscombe)
plot(anscombe$x2, resid(simple.lm),
     ylab='Residual', xlab='X',
     pch=23, bg='orange', cex=1.2)
abline(h=0, lwd=2, col='red', lty=2)
```



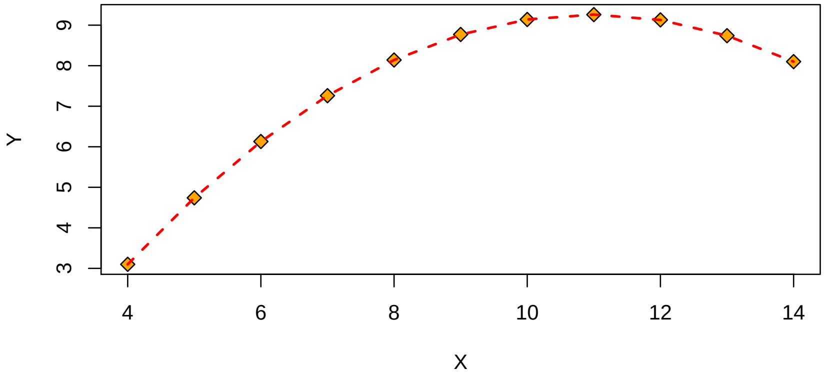
Diagnostic plots for linearity

```
par(mfrow=c(2,2))
plot(simple.lm)
```



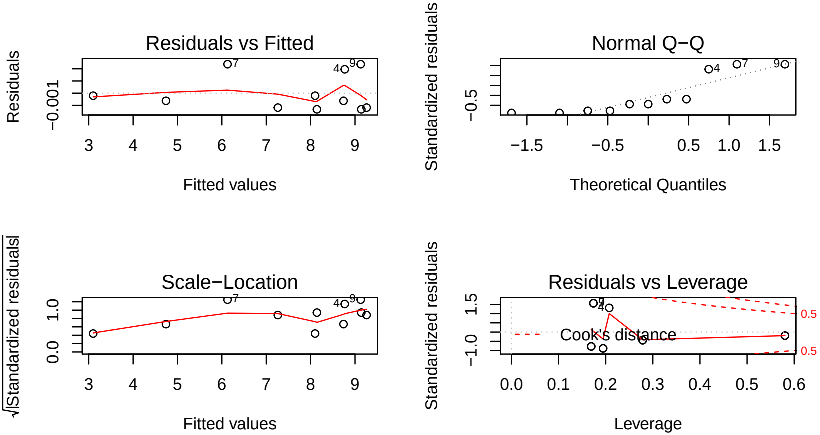
Diagnostic plots for linearity (Quadratic model)

```
quadratic.lm = lm(y2 ~ poly(x2, 2), data=anscombe)
Xsort = sort(anscombe$x2)
plot(anscombe$x2, anscombe$y2, pch=23,
     bg='orange', cex=1.2, ylab='Y', xlab='X')
lines(Xsort, predict(quadratic.lm, list(x2=Xsort)),
      col='red', lty=2, lwd=2)
```



Diagnostic plots for linearity

```
par(mfrow=c(2,2))
plot(quadratic.lm)
```

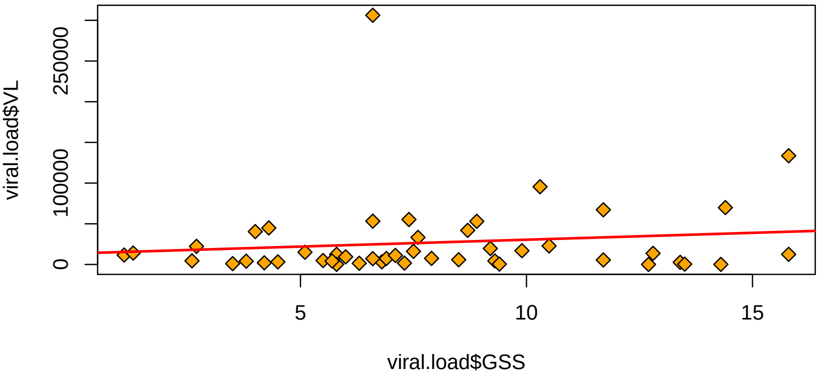


Simple linear diagnostics

- Outliers
- Nonconstant variance

Simple linear diagnostics

```
url = 'http://stats191.stanford.edu/data/HIV.VL.table'
viral.load = read.table(url, header=T)
plot(viral.load$GSS, viral.load$VL, pch=23,
     bg='orange', cex=1.2)
viral.lm = lm(VL ~ GSS, data=viral.load)
abline(viral.lm, col='red', lwd=2)
```



Multiple linear regression

Multiple linear regression model

- Rather than one predictor, we have $p = 6$ predictors.
- $Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip} + \varepsilon_i$
  - Errors ε are assumed independent $N(0, \sigma^2)$, as in simple linear regression.
  - Coefficients are called (partial) regression coefficients because they "allow" for the effect of other variables.

Overall F-test

- *Full (bigger) model :*

$$Y_i = \beta_0 + \beta_1 X_{i1} + \ldots \beta_p X_{ip} + \varepsilon_i$$

- *Reduced (smaller) model:*

$$Y_i = \beta_0 + \varepsilon_i$$

- The F-statistic has the form $F = \frac{(SSE(R) - SSE(F))/(df_R - df_F)}{SSE(F)/df_F}$.

Matrix formulation

- $\boldsymbol{Y}_{n\times 1} = \mathbf{X}_{n\times(p+1)}\boldsymbol{\beta}_{(p+1)\times 1} + \varepsilon_{n\times 1}$
- $\mathbf{X}$ is called the *design matrix* of the model
- $\boldsymbol{\epsilon} \sim N(0, \sigma^2\mathbf{I}_{n\times n})$ is multivariate normal *SSE* in matrix form

$$SSE(\beta) = (\boldsymbol{Y} - \mathbf{X}\boldsymbol{\beta})'(\boldsymbol{Y} - \mathbf{X}\boldsymbol{\beta})$$

OLS estimators

- Normal equations yield

$$\widehat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{Y}$$

- Properties:

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

Confidence interval for $\sum_{j=0}^{p} a_j \beta_j$

- Suppose we want a $(1-\alpha) \cdot 100\%$ CI for $\sum_{j=0}^{p} a_j \beta_j$.
- Just as in simple linear regression:
  - $\sum_{j=0}^{p} a_j \widehat{\beta}_j \pm t_{1-\alpha/2, n-p-1} \cdot SE\left(\sum_{j=0}^{p} a_j \widehat{\beta}_j\right)$.
  - Standard error:

$$SE\left(\sum_{j=0}^{p} a_j \widehat{\beta}_j\right) = \sqrt{\hat{\sigma}^2 \boldsymbol{a}^T (\mathbf{X}^T \mathbf{X})^{-1} \boldsymbol{a}}$$

General F-tests

- Given two models $R \subset F$ (i.e. R is a subspace of F), we can consider testing

$$H_0 : R \text{ is adequate (i.e. } \mathbb{E}(Y) \in R)$$

vs.

$$H_a : F \text{ is adequate (i.e. } \mathbb{E}(Y) \in F)$$

  - The test statistic is

$$F = \frac{(SSE(R) - SSE(F))/(df_R - df_F)}{SSE(F)/df_F}$$

  - If H_0 is true, $F \sim F_{df_R - df_F, df_F}$ so we reject H_0 at level α if $F > F_{df_R - df_F, df_F, 1-\alpha}$.

Diagnostics: What can go wrong?

- Regression function can be wrong: maybe regression function should have some other form (see diagnostics for simple linear regression).
- Model for the errors may be incorrect:
  - may not be normally distributed.
  - may not be independent.
  - may not have the same variance.
- Detecting problems is more *art* then *science*, i.e. we cannot *test* for all possible problems in a regression model.
- Basic idea of diagnostic measures: if model is correct then residuals $e_i = Y_i - \widehat{Y}_i, 1 \leq i \leq n$ should look like a sample of (not quite independent) $N(0, \sigma^2)$ random variables.

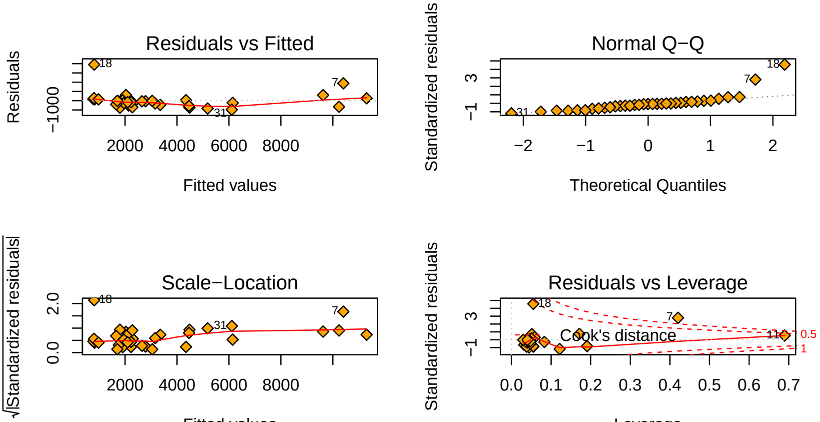
Diagnostics

Diagnostics

```
url = 'http://stats191.stanford.edu/data/scottish_races.tab
races.table = read.table(url, header=T)
attach(races.table)
races.lm = lm(Time ~ Distance + Climb)
```

Diagnostics

```r
par(mfrow=c(2,2))
plot(races.lm, pch=23 ,bg='orange',cex=1.2)
```



Diagnostics measures

- DFFITS:

$$DFFITS_i = \frac{\widehat{Y}_i - \widehat{Y}_{i(i)}}{\widehat{\sigma}_{(i)}\sqrt{H_{ii}}}$$

- Cook's Distance:

$$D_i = \frac{\sum_{j=1}^{n}(\widehat{Y}_j - \widehat{Y}_{j(i)})^2}{(p+1)\,\widehat{\sigma}^2}$$

- DFBETAS:

$$DFBETAS_{j(i)} = \frac{\widehat{\beta}_j - \widehat{\beta}_{j(i)}}{\sqrt{\widehat{\sigma}^2_{(i)}(X^TX)^{-1}_{jj}}}.$$

Diagnostics measures

```r
influence.measures(races.lm)
```

```
## Influence measures of
##   lm(formula = Time ~ Distance + Climb) :
##
##      dfb.1_  dfb.Dstn  dfb.Clmb    dffit  cov.r   cook.c
## 1   0.03781 -0.016613 -0.004743  0.03861 1.1595 5.13e-04
## 2  -0.05959  0.067223 -0.073404 -0.11957 1.1269 4.88e-03
## 3  -0.04858 -0.006707  0.028036 -0.06310 1.1329 1.37e-03
## 4  -0.00767 -0.005677  0.008766 -0.01368 1.1556 6.44e-05
## 5  -0.05047  0.084718 -0.145019 -0.20949 1.0837 1.47e-02
## 6   0.00348 -0.004311  0.007567  0.01219 1.1536 5.12e-05
## 7  -0.89062 -0.712743  2.364517  2.69897 0.8179 1.89e+00
## 8  -0.00845 -0.001650  0.005567 -0.01116 1.1467 4.29e-05
## 9  -0.01437  0.000913  0.006163 -0.01664 1.1453 9.52e-05
## 10  0.04703  0.013056 -0.036517  0.06399 1.1431 1.41e-03
## 11 -0.30124  0.768854 -0.479935  0.78583 3.4524 2.11e-01
## 12 -0.01150  0.009662 -0.007493 -0.01673 1.1492 9.62e-05
```

Outliers

- Observations $(Y, X_1, \dots, X_p)$ that do not follow the model, while most other observations seem to follow the model.
- One solution: Bonferroni correction, threshold at $t_{1-\alpha/(2*n), n-p-2}$.
- Bonferroni: if we are doing many t (or other) tests, say $m >> 1$ we can control overall false positive rate at α by testing each one at level α/m.

Outliers

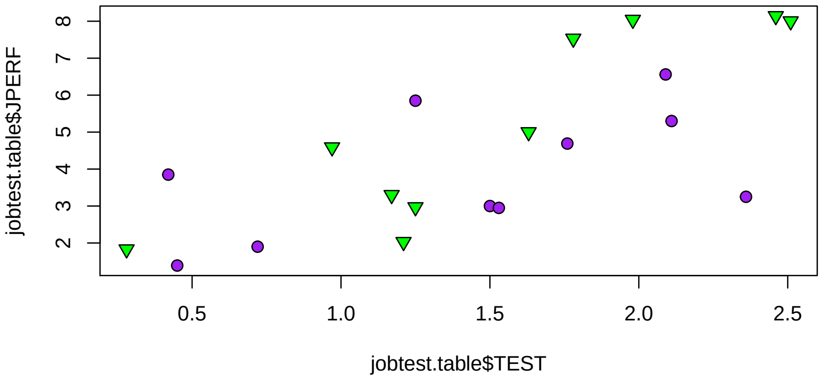
```
library(car)
outlierTest(races.lm)
```

```
##    rstudent unadjusted p-value Bonferroni p
## 18 7.610958         1.3968e-08    4.889e-07
```

Interactions and ANOVA

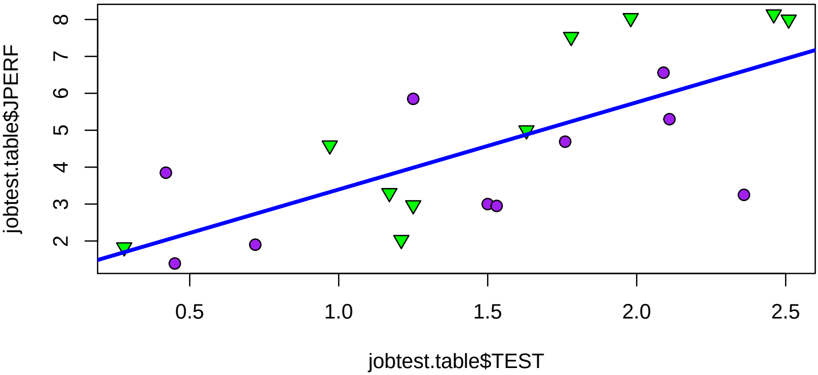
Qualitative variables and interactions

```
url = 'http://stats191.stanford.edu/data/jobtest.table'
jobtest.table = read.table(url, header=T)
jobtest.table$MINORITY = factor(jobtest.table$MINORITY)
plot(jobtest.table$TEST, jobtest.table$JPERF, type='n')
points(jobtest.table$TEST[(jobtest.table$MINORITY == 0)], j
points(jobtest.table$TEST[(jobtest.table$MINORITY == 1)], j
```



Qualitative variables and interactions

```
jobtest.lm1 = lm(JPERF ~ TEST, jobtest.table)
plot(jobtest.table$TEST, jobtest.table$JPERF, type='n')
points(jobtest.table$TEST[(jobtest.table$MINORITY == 0)], j
points(jobtest.table$TEST[(jobtest.table$MINORITY == 1)], j
abline(jobtest.lm1$coef, lwd=3, col='blue')
```



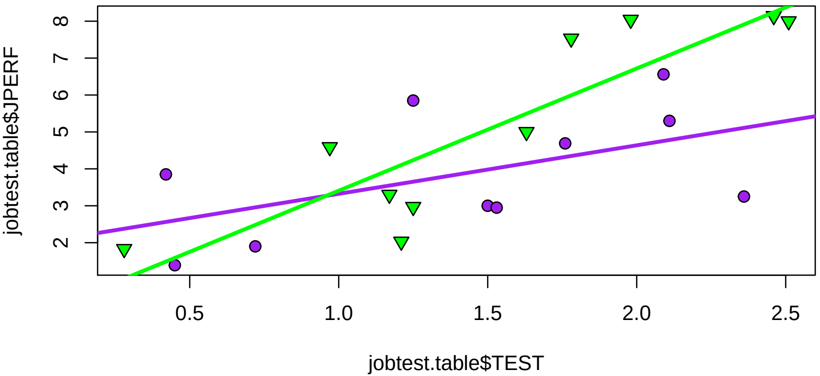
Qualitative variables and interactions

```
jobtest.lm4 = lm(JPERF ~ TEST * MINORITY, data = jobtest.ta
print(summary(jobtest.lm4))
```

```
##
## Call:
## lm(formula = JPERF ~ TEST * MINORITY, data = jobtest.tab
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.0734 -1.0594 -0.2548  1.2830  2.1980
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)       2.0103     1.0501   1.914   0.0736 .
## TEST              1.3134     0.6704   1.959   0.0677 .
## MINORITY1        -1.9132     1.5403  -1.242   0.2321
## TEST:MINORITY1    1.9975     0.9544   2.093   0.0527 .
## ---
```

Qualitative variables and interactions

```
plot(jobtest.table$TEST, jobtest.table$JPERF, type='n')
points(jobtest.table$TEST[(jobtest.table$MINORITY == 0)], j
points(jobtest.table$TEST[(jobtest.table$MINORITY == 1)], j
abline(jobtest.lm4$coef['(Intercept)'], jobtest.lm4$coef['T
abline(jobtest.lm4$coef['(Intercept)'] + jobtest.lm4$coef[
      jobtest.lm4$coef['TEST'] + jobtest.lm4$coef['TEST:MIN
```



ANOVA models: one-way

Source	SS	df	MS	$\mathbb{E}$ (MS)
Treatment	$\text{SSTR} = \sum_{i=1}^{r} n_i \left(\overline{Y}_{i\cdot} - \overline{Y}_{\cdot\cdot}\right)^2$	$r-1$	$\text{MSTR} = \frac{\text{SSTR}}{r-1}$	$\sigma^2 + \frac{\sum_{i=1}^{r} n_i \alpha_i^2}{r-1}$
Error	$\text{SSE} = \sum_{i=1}^{r} \sum_{j=1}^{n_i} (Y_{ij} - \overline{Y}_{i\cdot})^2$	$\sum_{i=1}^{r}(n_i - 1)$	$\text{MSE} = \frac{\text{SSE}}{\sum_{i=1}^{r}(n_i - 1)}$	σ^2

ANOVA models: two-way

- In the balanced case, everything can again be summarized from the ANOVA table

Source	SS	DF	MS
A	$SSA = nm\sum_{i=1}^{r}\left(\overline{Y}_{i\cdot\cdot} - \overline{Y}_{\cdot\cdot\cdot}\right)^2$	r-1	SSA/(r-1)
B	$SSB = nr\sum_{j=1}^{m}\left(\overline{Y}_{\cdot j\cdot} - \overline{Y}_{\cdot\cdot\cdot}\right)^2$	m-1	SSB/(m-1)
A:B	$SSAB = n\sum_{i=1}^{r}\sum_{j=1}^{m}\left(\overline{Y}_{ij\cdot} - \overline{Y}_{i\cdot\cdot} - \overline{Y}_{\cdot j\cdot} + \overline{Y}_{\cdot\cdot\cdot}\right)^2$	(m-1)(r-1)	SSAB/(m-1)(r-1)
ERROR	$SSE = \sum_{i=1}^{r}\sum_{j=1}^{m}\sum_{k=1}^{n}(Y_{ijk} - \overline{Y}_{ij\cdot})^2$	(n-1)mr	SSE/(n-1)mr

ANOVA models: two-way

Source	$\mathbb{E}(MS)$
A	$\sigma^2 + nm\frac{\sum_{i=1}^{r}\alpha_i^2}{r-1}$
B	$\sigma^2 + nr\frac{\sum_{j=1}^{m}\beta_j^2}{m-1}$
A:B	$\sigma^2 + n\frac{\sum_{i=1}^{r}\sum_{j=1}^{m}(\alpha\beta)_{ij}^2}{(r-1)(m-1)}$
ERROR	σ^2

Weighted Least Squares

Weighted Least Squares

- A way to correct for errors with unequal variance (**but we need a model of the variance**).
- Weighted Least Squares

$$SSE(\beta, w) = \sum_{i=1}^{n} w_i \left(Y_i - \beta_0 - \beta_1 X_i\right)^2 .$$

- In general, weights should be like:

$$w_i = \frac{1}{\mathrm{Var}(\varepsilon_i)}.$$

- WLS estimator:

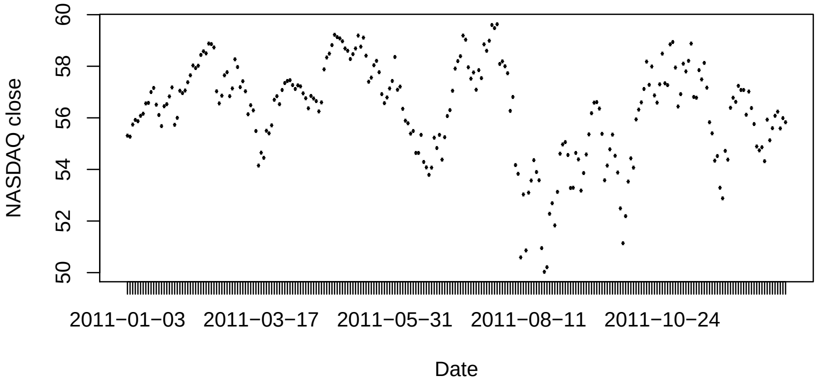
$$\hat{\beta}_W = (X^T W X)^{-1}(X^T W Y).$$

- If weights are ignored standard errors are wrong!
- Briefly talked about efficiency of estimators.

Autocorrelation.

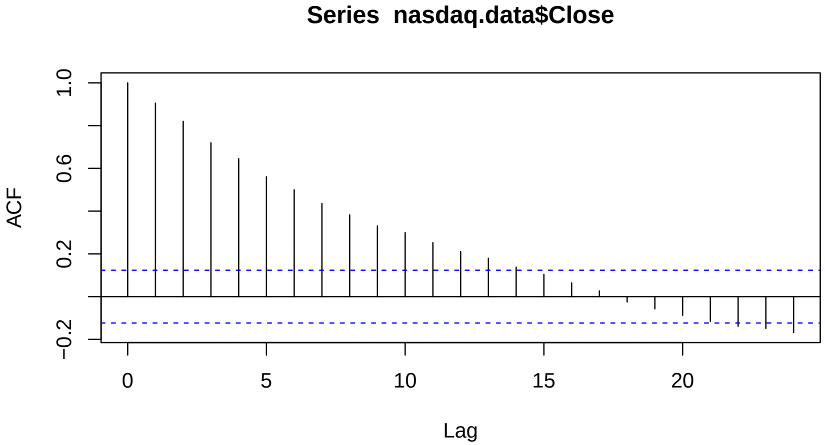
Correlated errors: NASDAQ daily close 2011

```
url = 'http://stats191.stanford.edu/data/nasdaq_2011.csv'
nasdaq.data = read.table(url, header=TRUE, sep=',')

plot(nasdaq.data$Date, nasdaq.data$Close, xlab='Date', ylab
     pch=23, bg='red', cex=1.2)
```



ACF

```r
acf(nasdaq.data$Close)
```



AR(1) noise

- Suppose that, instead of being independent, the errors in our model were $\varepsilon_t = \rho \cdot \varepsilon_{t-1} + \omega_t, \qquad -1 < \rho < 1$ with $\omega_t \sim N(0, \sigma^2)$ independent.
- If ρ is close to 1, then errors are very correlated, $\rho = 0$ is independence.
- This is "Auto-Regressive Order (1)" noise (AR(1)). Many other models of correlation exist: ARMA, ARIMA, ARCH, GARCH, etc.

Correcting for AR(1)

- Suppose we know ρ, if we "whiten" the data and regressors

$$\tilde{Y}_{t+1} = Y_{t+1} - \rho Y_t, t > 1$$
$$\tilde{X}_{(t+1)j} = X_{(t+1)j} - \rho X_{tj}, i > 1$$

  for $1 \leq t \leq n-1$. This model satisfies "usual" assumptions, i.e. the errors $\tilde{\varepsilon}_t = \omega_{t+1} = \varepsilon_{t+1} - \rho \cdot \varepsilon_t$ are independent $N(0, \sigma^2)$.
- For coefficients in new model $\tilde{\beta}$, $\beta_0 = \tilde{\beta}_0/(1-\rho)$, $\beta_j = \tilde{\beta}_j$.
- Problem: in general, we don't know ρ, but estimated it.
- If correlation structure is ignored standard errors are wrong!
- Another example of **whitening when we can model the variance.**

Bootstrap

Bootstrapping `lm`

- Using WLS (weighted least squares) requires a model for the variance of ϵ given X.
- Ignoring this changing variance (heteroskedasticity) and using OLS leads to bad intervals, p-values, etc. **because standard errors are incorrect.**
- The (pairs) bootstrap uses the OLS estimator but is able to get a **correct estimator of standard error**.

Bootstrapping lm

```r
library(car)
n = 50
X = rexp(n)
Y = 3 + 2.5 * X + X * (rexp(n) - 1) # our usual model is fa
Y.lm = lm(Y ~ X)
pairs.Y.lm = Boot(Y.lm, coef, method='case', R=1000)
confint(pairs.Y.lm, type='norm') # using bootstrap SE
```

```
## Bootstrap normal confidence intervals
##
##                 2.5 %   97.5 %
## (Intercept) 2.366167 3.005671
## X           2.550590 3.393426
```

Model selection

Model selection criteria

- Mallow's C_p:

$$C_p(\mathcal{M}) = \frac{SSE(\mathcal{M})}{\widehat{\sigma}^2} + 2 \cdot p(\mathcal{M}) - n.$$

- Akaike (AIC) defined as

$$AIC(\mathcal{M}) = -2 \log L(\mathcal{M}) + 2p(\mathcal{M})$$

  where $L(\mathcal{M})$ is the maximized likelihood of the model.
- Bayes (BIC) defined as

$$BIC(\mathcal{M}) = -2 \log L(\mathcal{M}) + \log n \cdot p(\mathcal{M})$$

- Adjusted R^2
- Stepwise (`step`) vs. best subsets (`leaps`).

K-fold cross-validation

- Fix a model $\mathcal{M}$.
- Break data set into K approximately equal sized groups $(G_1, \dots, G_K)$.
- for (i in 1:K)
  - Use all groups except G_i to fit model, predict outcome in group G_i based on this model $\widehat{Y}_{j,\mathcal{M},G_i}, j \in G_i$.
- Estimate

$$CV(\mathcal{M}) = \frac{1}{n} \sum_{i=1}^{K} \sum_{j \in G_i} (Y_j - \widehat{Y}_{j,\mathcal{M},-G_i})^2.$$

Multicollinearity

Multicollinearity

- Detecting collinearity
  - Large values of pairwise correlation coefficient, the regression results do not conform to prior expectations
  - Variance inflation factors, condition indices
- Working with Collinear data
  - Standardization
  - Principal components regression
  - Penalization

Penalized regression

Shrinkage estimator

- In one sample problem, when trying to estimate μ from $Y_i \sim N(\mu, \sigma^2)$ we looked at the estimator

$$\hat{Y}_\alpha = \alpha \cdot \bar{Y}.$$

- The "quality" of the estimator decomposed as

$$E((\hat{Y}_\alpha - \mu)^2) = \mathrm{Bias}(\hat{Y}_\alpha)^2 + \mathrm{Var}(\hat{Y}_\alpha)$$

Shrinkage estimator

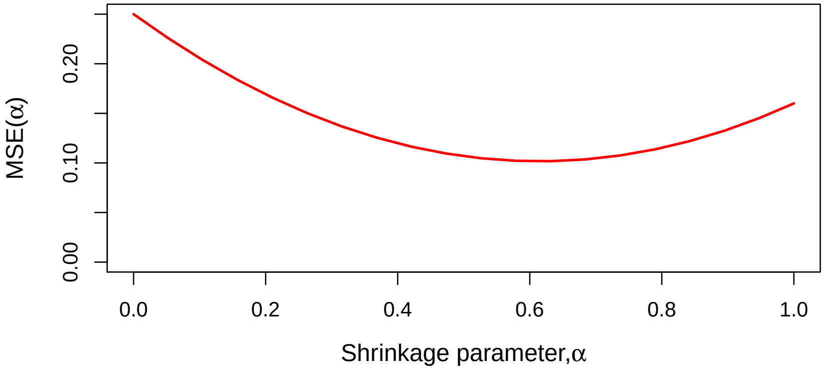
```
nsample = 40
ntrial = 500
mu = 0.5
sigma = 2.5
MSE = function(mu.hat, mu) {
  return(sum((mu.hat - mu)^2) / length(mu))
}

alpha = seq(0.0,1,length=20)

mse = numeric(length(alpha))
```

Shrinkage estimator

```
for (i in 1:ntrial) {
  Z = rnorm(nsample) * sigma + mu
  for (j in 1:length(alpha)) {
    mse[j] = mse[j] + MSE(alpha[j] * mean(Z) * rep(1, nsamp
                                   mu * rep(1, nsample)) / ntrial
  }
}
```

Shrinkage estimator

```
plot(alpha, mse, type='l', lwd=2, col='red',
     ylim=c(0, max(mse)),
     xlab=expression(paste('Shrinkage parameter,', alpha)),
     ylab=expression(paste('MSE(', alpha, ')')),
     cex.lab=1.2)
```

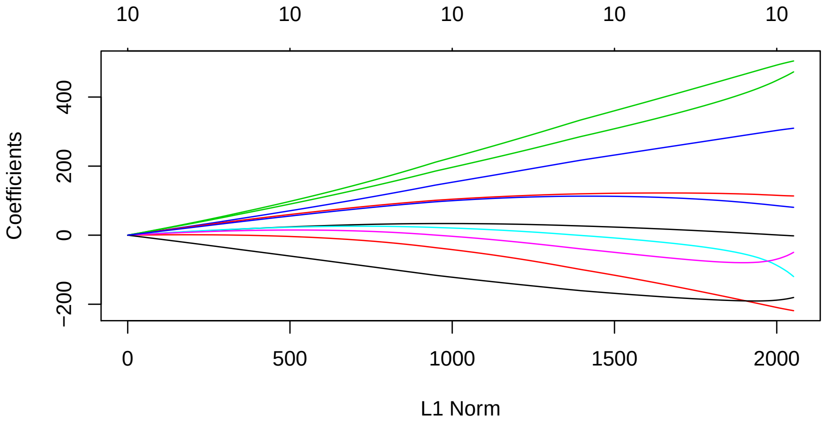


Ridge regression

$$\hat{\beta}_\lambda = \text{argmin}_\beta \frac{1}{2n}\|Y - X\beta\|_2^2 + \lambda\|\beta\|_2^2$$

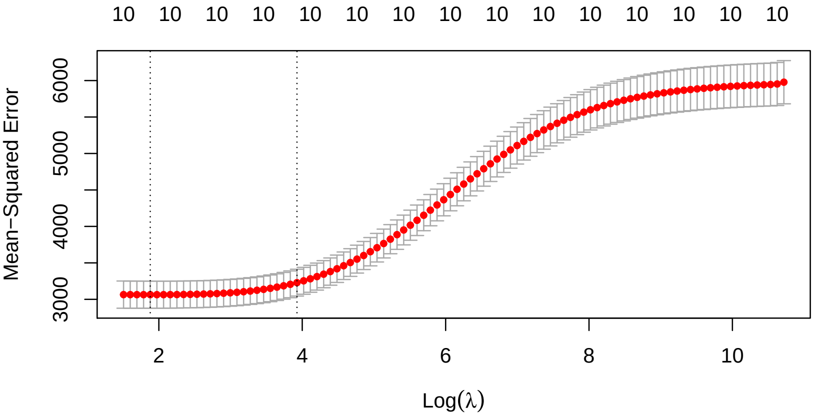
Ridge with glmnet

```
library(lars)
data(diabetes)
plot(glmnet(diabetes$x, diabetes$y, alpha=0))
```



Ridge with glmnet

```
plot(cv.glmnet(diabetes$x, diabetes$y, alpha=0))
```

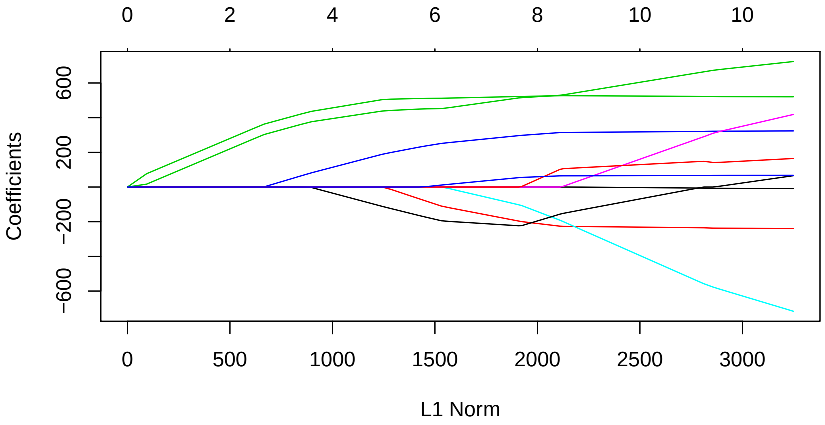


LASSO

$$\hat{\beta}_\lambda = \text{argmin}_\beta \frac{1}{2n} \| Y - X\beta \|_2^2 + \lambda \|\beta\|_1$$

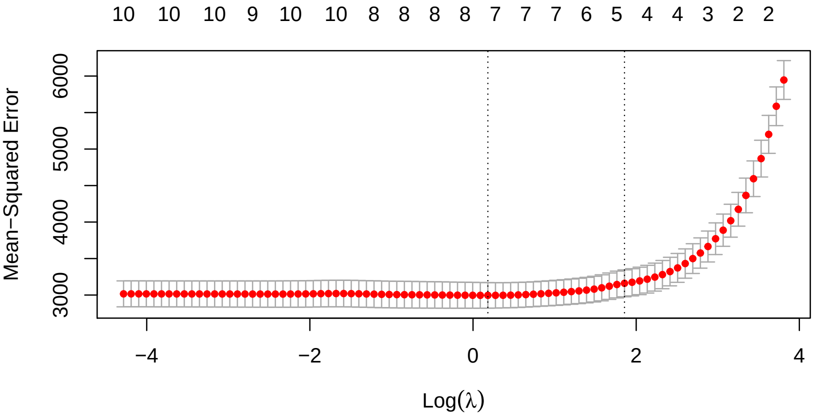
LASSO with glmnet

```r
library(glmnet)
plot(glmnet(diabetes$x, diabetes$y))
```



LASSO with glmnet

```
plot(cv.glmnet(diabetes$x, diabetes$y))
```



Logistic regression

Logistic regression model

- Logistic model

$$E(Y|X) = \pi(X) = \frac{\exp(X^T\beta)}{1+\exp(X^T\beta)}$$

- This automatically fixes $0 \leq E(Y) = \pi(X) \leq 1$.
- The logistic transform: $\text{logit}(\pi(X)) = \log\left(\frac{\pi(X)}{1-\pi(X)}\right) = X^T\beta$
- An example of a *generalized linear model*
  - link function $\text{logit}(\pi(X)) = X^T\beta$
  - Variance function: $\text{Var}(Y|X) = \pi(X)(1-\pi(X))$

Odds Ratios

- One reason logistic models are popular is that the parameters have simple interpretations in terms of odds.
- Logistic model:

$$OR_{X_j} = \frac{ODDS(Y = 1|\ldots, X_j = x_j + h, \ldots)}{ODDS(Y = 1|\ldots, X_j = x_j, \ldots)} = e^{h\beta_j}$$

- If $X_j \in 0, 1$ is dichotomous, then odds for group with $X_j = 1$ are e^{β_j} higher, other parameters being equal.

Deviance

- For logistic regression model $\mathcal{M}$, $DEV(\mathcal{M})$ replaces $SSE(\mathcal{M})/\sigma^2$.
- In least squares regression, we use

$$\frac{SSE(\mathcal{M}_R) - SSE(\mathcal{M}_F)}{\sigma^2} \sim \chi^2_{df_R - df_F}$$

- This is replaced with $DEV(\mathcal{M}_R) - DEV(\mathcal{M}_F) \overset{n \to \infty}{\sim} \chi^2_{df_R - df_F}$
- For Poisson and binary regression, $\sigma^2 = 1$ (dispersion parameter of `glm`).

Poisson regression

Poisson log-linear regression model

- Log-linear model

$$E(Y|X) = \exp(X^T\beta)$$

- This automatically fixes $E(Y|X) \geq 0$.
- An example of a *generalized linear model*
  - link function $\log(E(Y|X)) = X^T\beta$
  - Variance function: $\mathrm{Var}(Y|X) = E(Y|X)$
- Interpretation:

$$\frac{E(Y|\ldots, X_j = x_j + h, \ldots)}{E(Y|\ldots, X_j = x_j, \ldots)} = e^{h\beta_j}$$

Reference

- Lecture notes of Jonathan Taylor .