STA 108 Notes - J. Jiang
Dylan M Ang
March 13, 2022

Contents

1 Simple Linear Regression

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i \quad i = 1 \dots n \tag{1}$$

If assumptions hold true,

- Y_i is normally distributed

$$E(Y_i) = \beta_0 + \beta_1 x_i \tag{2}$$
$$Var(Y_i) = \sigma^2 \tag{3}$$

- Mean: $\beta_0 + \beta_1 x_i$
- Variance: σ^2

Assumptions

- $\epsilon_1 \dots \epsilon_n$
- $E(\epsilon_i) = 0, var(\epsilon_i) = 0$, where σ^2 is an unknown constant.
- ϵ_i is normal. (normality assumption)

1.1 Least Squares Estimate

$$\Sigma_{i=1}^n (Y_i - \beta_0 - \beta_1 x_i)^2 \tag{4}$$

Take the first order derivative with respect to β_0, β_1 to minimize equation (4) to find optimal β_0, β_1.

LS estimators

$$\hat{\beta}_1 = \frac{\Sigma_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y})}{\Sigma_{i=1}^n (x_i - \bar{x})^2} \tag{5}$$
$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x} \tag{6}$$

- $\bar{x} = \frac{1}{n}\Sigma_{i=1}^n x_i$ (sample mean of the x_i's)
- $\bar{Y} = \frac{1}{n}\Sigma_{i=1}^n Y_i$ (sample mean of the Y_i's)
- Regression Line: $y = \hat{\beta}_0 + \hat{\beta}_1 x$

Properties of LS Estimators

- $E(\hat{\beta}_0) = \beta_0$, $E(\hat{\beta}_1) = \beta_1$. The average of many sample beta values will approach the true beta values.

Fitted (or predicted) values are estimates. The fitted value for Y_i is $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$
$\hat{Y}$ is an unbiased estimator of $E(Y) = \beta_0 + \beta_1 x$ so $E(\hat{Y}) = E(Y)$

1.2 Residuals

Residuals: $\hat{\epsilon}_i = Y_i - \hat{Y}_i, i = 1 \dots n$
Properties of Residuals

- $\Sigma_{i=1}^n \hat{\epsilon}_i = 0$
- The residuals are not independent.
- If one residual is positive, another residual has to compensate.

1.3 Variance

Estimation of σ^2, the variance of the errors (which is the same as the variance of Y_i)

$$s^2 = \frac{1}{n-2}\Sigma_{i=1}^n (Y_i - \hat{Y}_i)^2 \tag{7}$$

where $\hat{Y}_i$ is the estimate of $E(Y_i)$.

Notes

- $\hat{Y}_i$ is an estimator of $E(Y_i) = \beta_0 + \beta_1 x_i$ in which two parameters are estimated (β_0 and β_1) $\implies$ 2 degrees of freedoms are subtracted.
- $E(s^2) = \sigma^2$

When errors are normally distributed, the LS estimators of β_0, β_1 is equal to the MLEs (Maximum Likelihood Estimators) of β_0, β_1, but the MLE of $\sigma^2, \hat{\sigma}^2$, is different from s^2
s^2 is just (7)

$$\hat{\sigma}^2 = \frac{1}{n}\Sigma_{i=1}^n \hat{\epsilon}_i^2 \tag{8}$$

2 Inference in regression and correlation analysis

2.1 Inference about β_1

For testing β_1
$H_0 : \beta_1 = \beta_{10}, \beta_{10}$ is a given value such as 0.
$H_a : \beta_1 \neq \beta_{10}, \beta_1 > \beta_{10}$, or $\beta_1 < \beta_{10}$
Test statistic: A statistic whose distribution is known under the null hypothesis.

$$t = \frac{\hat{\beta}_1 - \beta_{10}}{s.e.(\hat{\beta}_1)} \tag{9}$$

where $\hat{\beta}_1$ is the LS estimate of β_1, and

$$s.e.(\hat{\beta}_1) = \sqrt{\frac{MSE}{\Sigma_i (x_i - \bar{x})^2}} \tag{10}$$
$$MSE = s^2 \tag{11}$$

If normal, $T \sim t_{n-1}$

$$T = \frac{\hat{\beta}_1 - \beta_1}{s.e.(\hat{\beta}_1)} \tag{12}$$

Therefore under the $H_0 : \beta_1 = \beta_{10}, t \sim t_{n-2}$
Decision Rules

$$H_1 : \beta_1 \neq \beta_{10}, reject\ H_0\ if\ |t| > t_{n-2,\alpha/2}$$
$$H_1 : \beta_1 > \beta_{10}, reject\ H_0\ if\ |t| > t_{n-2;\alpha}$$
$$H_1 : \beta_1 < \beta_{10}, reject\ H_0\ if\ |t| < -t_{n-2;\alpha}$$

Alternatively, Reject H_0 if the p-value of t is $\leq \alpha$
Error

- Type I: Reject H_0 when it is true.
- Type II: Fail to reject H_0 when it is false.

Level of Significance α
α is the upper bound for the probability of Type I error.
P-value
p-value is the observed level of significance: the actual probability that the test statistic is as extreme as observed given H_0 is true.
Power
Power is the probability of rejecting H_0 when the alternative holds at a given value.
If $\beta_{10} = 0, \beta_1 = 1, s.d.(\hat{\beta}_1) = 0.5$, we have $\delta = \frac{1}{0.5} = 2$ Let $\alpha = 0.05$. From table B.5 we find the power is 0.48.
Confidence interval for β_k
Assuming normality, a $100(1-\alpha)\%$ c.i. for β_k is

$$\hat{\beta}_k \pm t_{n-2}(1 - \frac{alpha}{2} * s.e.(\hat{\beta}_j) \tag{13}$$
$$k = 0, 1 \tag{14}$$

where $s.e.(\hat{\beta}_1)$ can be found with eq(10) and $s.e.(\hat{\beta}_0)$ can be found with eq(15).

2.2 Inference about β_0

$$s.e.(\beta_0) = \sqrt{mse(\frac{1}{n} + \frac{\bar{x}^2}{\Sigma_{i=1}^n (x_i - \bar{x})^2})} \tag{15}$$

Confidence intervals for β_0 can be found with (13)

2.3 Inference about $\hat{Y}$

Confidence Interval for $E(Y) = \beta_0 + \beta_1 x$

$$\hat{Y} \pm t_{n-2}(1 - \frac{alpha}{2}) * s.e.(\hat{Y}) \tag{16}$$
$$s.e.(\hat{Y}) = \sqrt{MSE(\frac{1}{n} + \frac{(x - \bar{x})^2}{\Sigma_{i=1}^n (x_i - \bar{x})^2})} \tag{17}$$

2.4 Prediction interval for $\hat{Y}$

A $100(1-\alpha)\%$ prediction interval for $Y = E(Y) + \epsilon = \beta_0 + \beta_1 x + \epsilon$, where Y is the future observation and ϵ is the new error:

$$\hat{Y} \pm t_{n-2}(1 - \frac{\alpha}{2}) * p.s.e.(\hat{Y}) \tag{18}$$
$$p.s.e.(\hat{Y}) = \sqrt{MSE(1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{\Sigma_{i=1}^n (x_i - \bar{x})^2})} \tag{19}$$

Where p.s.e. is the percent standard error.
The 1 in the p.s.e is because the variance of $\epsilon = \sigma^2$. If $var(\epsilon) = \frac{\sigma^2}{2}$ change the 1 to $\frac{1}{2}$.

2.5 ANOVA and F-test

$$SSTO = SSR + SSE \tag{20}$$
$$= \Sigma_{i=1}^n (Y_i - \bar{Y})^2 \tag{21}$$
$$SSR = \Sigma_{i=1}^n (\hat{Y}_i - \bar{Y})^2 \tag{22}$$
$$SSE = \Sigma_{i=1}^n (Y_i - \hat{Y})^2 \tag{23}$$
$$\tag{24}$$

Sum of Squares of Regression (SSR) explains the variability in Y due to the regression model compared to the baseline model. Sum of Squares of Errors (SSE) is the remaining unexplained variability of Y found from SSTO - SSR.
Degrees of Freedom

$$SSRdf = 1$$
$$SSEdf = n - 2$$
$$SSTOdf = n - 2 + 1 = n - 1$$

Mean Squares
Mean squares is SS divided by its degrees of freedom.

$$MSR = \frac{SSR}{1} \tag{25}$$
$$MSE = \frac{SSE}{n-2} \tag{26}$$
$$\tag{27}$$

F-Statistic

$$F = \frac{MSR}{MSE} = \frac{SSR * (n-2)}{SSE} \tag{28}$$

ANOVA table: Analysis of variance.
The distribution of F under the null hypothesis $H_0 : \beta_1 = 0$ is $F_{1,n-2}$.

Source	SS	df	MS	F
Regression	SSR	1	MSR	F
Error	SSE	n-2	MSE	
Total	SSTO	n-1		

2.6 Inference about ρ

R^2: a measure of goodness of fit, which is the proportion of variation in Y explained by the regression (i.e. by x).

$$R^2 = \frac{SSR}{SSTO} = 1 - \frac{SSE}{SSTO} \tag{29}$$

Coefficient of correlation:

$$r = \pm\sqrt{R^2} = \begin{cases} +\sqrt{R^2} & if\ \hat{\beta}_1 > 0 \\ -\sqrt{R^2} & if\ \hat{\beta}_1 < 0 \end{cases} \tag{30}$$
$$r = \frac{\Sigma_i (Y_i - \bar{Y})(x_i - \bar{x})}{\sqrt{\Sigma_i (Y_i - \bar{Y})^2 \Sigma_i (x_i - \bar{x})^2}} \tag{31}$$

Properties of R^2 and r

- $0 \leq R^2 \leq 1 \quad -1 \leq r \leq 1$
- $R^2 \approx 1$ or $r \approx \pm 1$, if there is a strong linear association between x and Y.
- $R^2 \approx 0$, or $r \approx 0$, if there is a weak or no linear association between x and Y.
- Both R^2 and r are measures of linear association only.

Covariance and correlation between two random variables

$$cov(X, Y) = E\{(X - \mu_X)(Y - \mu_Y)\} \tag{32}$$
$$= E(XY) - E(X)E(Y) \tag{33}$$
$$cor(X, Y) = \frac{cov(X, Y)}{sd(X)sd(Y)} \tag{34}$$

where $\mu_X = E(X), sd(X) = \sqrt{var(X)}$,etc.
Special case: (X, Y) has a bivariate normal distribution.
Testing for ρ
Assume that the bivariate normal distribution holds for (X, Y).
$H_0 : \rho = 0$
$H_a : \rho \neq 0($ or $\rho > 0$ or $\rho < 0)$

$$t^* = \frac{r\sqrt{n-2}}{\sqrt{1 - r^2}} \sim t_{n-2} \text{ under } H_0 \tag{35}$$

3 Diagnostics

The goal of diagnostics is to examine the departures from the simple linear regression model with normal errors. Typical departures and corresponding diagnostic plots/tests are:

- The regression is not linear - residual plots(residual against the predictor variable, or against the fitted values), lack of fit test.
- The error terms are not normally distributed - histogram, boxplot/dot plot of residuals, normal probability plot (aka QQ plot), Shapiro-Wilk's test, correlation test for normality.
- The error terms do not have constant variance - residual plots, Brown - Forsythe (BF) test.
- The error terms are not independent - residual against time.
- The model fits all but one or a few outlier observations - (semistudentized) residual plots, box plots, dot plots, stem and leaf plots.
- Some important predictors are missing - residual plots (residual against other possibly important predictors).

3.1 Residual Plots

Residuals can be used to check whether

- The regression function is not linear.
- The variance of the errors is not constant.
- The errors are not independent.
- Outliers
- The errors are not normal.
- Some important predictors are missing.

Scatter Plot

- Check linearity - residuals normally disributed.
- Check constant variance - residuals are random and dont follow a cone pattern.

Box Plot and Dot Plot

- Normality - residuals should be centered and symmetric about 0.

Normality probability plot - QQ Plot

- QQ plot is linear $\implies$ normal residuals.
- QQ plot is nonlinear $\implies$ non normal residuals.

3.2 Diagnostic Tests

Shapiro Wilk's test
$H_0\ data\ \sim N()$
H_a : data not normal.
$p - val \leq \alpha$ reject normality assumption.
Correlation test for normality
Step 1. Compute the coefficient of correlation between the ordered residuals and their expected values. The latter are given by

$$\sqrt{MSE}z(\frac{k - 0.375}{n + 0.25}), \quad k = 1, \dots, n \tag{36}$$

where $z(p)$ is the pth quantile of the standard normal distribution, that is, $P[Z \leq z(p)] = p$, where Z has the standard normal distribution.
Step 2. Compare the coefficient of correlation on I with the critical value from Table B.6, if the coefficient of correlation exceeds the critical value, accept the normality assumption.
BF test for constant variance
1. Divide the residuals into two parts according to residual pattern (or no pattern)
Let $\hat{\epsilon}_{i1} = 1, \dots, n_1$ be the residuals for the first part, and $\hat{\epsilon}_{i2}, i = 1, \dots, n_2$ be the residuals for the second part, where $n_1 + n_2 = n$.
Compute $m(\hat{\epsilon}_1) =$ *median of* $\hat{\epsilon}_{i1}, i = 1, \dots, n_1$ and $m(\hat{\epsilon}_2)$.
2. Compute $d_{i1} = |\hat{\epsilon}_{i1} - m(\hat{\epsilon}_1)|, i = 1 \dots n_1$ and $d_{i2} = |\hat{\epsilon}_{i2} - m(\hat{\epsilon}_2)|, i = 1 \dots n_2$
3. Compute t score.

$$t_{BF} = \frac{\bar{d}_1 - \bar{d}_2}{s\sqrt{n_1^{-1} + n_2^{-1}}} \tag{37}$$
$$s^2 = \frac{\Sigma_{i=1}^{n_1} (d_{i1} - \bar{d}_1)^2 + \Sigma_{i=1}^{n_2} (d_{i2} - \bar{d}_2)^2}{n - 2} \tag{38}$$

4. Test $H_0 : \sigma_1^2 = \sigma_2^2$ vs $H_a : \sigma_1^2 \neq \sigma_2^2$
$t_{BF} \sim t_{n-2}$ under H_0. Given α, use the critical value (or p-value) to test H_0.
F-test for lack of fit
Regression model: $Y_{ij} = \beta_0 + \beta_1 x_j + \epsilon_{ij}, j = 1 \dots c, i = 1 \dots n_j$ where x_j is the jth value of x, c is the number of different x values, and $Y_{ij}, i = 1 \dots n_j$ are the Y values corresponding to the same x_j.
Full model: $Y_{ij} = \mu_j + \epsilon_{ij}, j = 1 \dots c, i = 1 \dots n_j$
F-statistic:

$$F = \frac{SSE(R) - SSE(F)}{df_R - df_F}\{\frac{SSE(F)}{df_F}\}^{-1} \tag{39}$$

where

$$SSE(R) = \Sigma_j\Sigma_i(Y_{ij} - \hat{Y}_{ij})^2 \tag{40}$$
$$SSE(F) = \Sigma_j\Sigma_i(Y_{ij} - \hat{\mu}_j)^2 \tag{41}$$

with $\hat{Y}_{ij} = \hat{\beta}_0 + \hat{\beta}_1 x_j$ and $\hat{\mu}_j = \bar{Y}_j - n_j^{-1}\Sigma_{i=1}^{n_j} Y_{ij}, df_R = n-2$ with $n = \Sigma_{j=1}^{c} n_j$ and $df_F = n - c$.
Under H_0 : The assumed model is correct, $F \sim F_{c-2,n-c}$.

3.3 Remedial Measures

Transformation of x: for nonlinear association.
Transformation of Y: for nonnormality/unequal variance.
Box Cox transformation
This is a collection of transformations depending on a "tuning parameter", λ.

$$Y_i' = \begin{cases} K_1(Y_i^\lambda - 1), & \lambda \neq 0 \\ K_2 log(Y_i), & \lambda = 0 \end{cases} \tag{42}$$

where K_1, K_2 are two numbers computed from the data.

$$K_2 = (Y_1 Y_2 \dots Y_n)^{\frac{1}{n}} = e^{\overline{\log Y}} \tag{43}$$
$$K_1 = \frac{1}{\lambda K_2^{\lambda-1}} \tag{44}$$

4 Simultaneous Inference

4.1 Simultaneous Confidence Intervals

An SCI represents the percentage likelihood that a group of confidence intervals will all include the true population parameters or true differences between factor levels if the study were repeated multiple times.
SCI's for $E(Y_h) = \beta_0 + \beta_1 x_h, h \in G, g = |G|$
A $100(1-\alpha)\%$ s.c.i has the following form,

$$\text{Working-Hotelling's} \quad \hat{Y}_h \pm W * se(\hat{Y}_h) \tag{45}$$
$$\text{Bonferroni's} \quad \hat{Y}_h \pm B * se(\hat{Y}_h) \tag{46}$$

Where,

$$se(\hat{Y}_h) = \sqrt{MSE\left(\frac{1}{n} + \frac{(x_h - \bar{x})^2}{\Sigma_i(x_i - \bar{x})^2}\right)} \tag{47}$$
$$W = \sqrt{2 * F_{2,n-2}(1-\alpha)} \tag{48}$$
$$B = t_{n-2}\left(1 - \frac{\alpha}{2g}\right) \tag{49}$$

4.2 Simultaneous Prediction Intervals

The goal of a prediction band is to cover with a prescribed probability the values of one or more future observations from the same population from which a given data set was sampled. Just as prediction intervals are wider than confidence intervals, prediction bands will be wider than confidence bands.

$$\text{Bonferroni's} \quad \hat{Y}_h \pm B * pse(\hat{Y}_h) \tag{50}$$
$$\text{Scheffe's} \quad \hat{Y}_h \pm S * pse(\hat{Y}_h) \tag{51}$$

where $B = (49)$

$$pse(\hat{Y}_h) = \sqrt{MSE\left(1 + \frac{1}{n} + \frac{(x_h - \bar{x})^2}{\Sigma_i(x_i - \bar{x})^2}\right)} \tag{52}$$
$$S = \sqrt{gF_{g,n-2}(1-\alpha)} \tag{53}$$
$$\tag{54}$$

5 Multiple Linear Regression

Matrix expression for multiple linear regression,

$$Y_i = \beta_0 + \beta_1 x_{i,1} + \dots + \beta_p x_{i,p-1} + \epsilon_i, i = 1 \dots n \tag{55}$$

ϵ_i has the same assumptions as simple linear regression.
Multiple linear regression can be expressed as

$$Y = X\beta + \epsilon \tag{56}$$

Given

$$X = \begin{bmatrix} 1 & x_{1,1} & \dots & x_{1,p-1} \\ 1 & x_{2,1} & \dots & x_{2,p-1} \\ \dots & \dots & \dots & \dots \\ 1 & x_{n,1} & \dots & x_{n,p-1} \end{bmatrix} \tag{57}$$

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \dots \\ \beta_{p-1} \end{bmatrix}, Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \dots \\ Y_n \end{bmatrix}, \epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \dots \\ \epsilon_n \end{bmatrix} \tag{58}$$

LS Estimate
Find β that minimizes $|Y - X\beta|^2$, where for a vector $v = (v_1 \dots v_n)', |v|^2 = \Sigma_i^n v_i^2$, the solution is given by

$$\hat{\beta} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \dots \\ \hat{\beta}_{p-1} \end{bmatrix} = (X'X)^{-1}X'Y \tag{59}$$

This can be computed in R. Given an $n * p$ matrix, X.

1. Manual
   - $\hat{\beta}$ = `solve(t(X) %*% X) %*% (t(X) %*% Y)`
   - `%*%` denotes the matrix product.
2. Using built in functions
   - `result = lsfit(X, Y, intercept = F)`
   - `bhat = result$coef`

5.1 ANOVA Table

$$SSTO = \Sigma_{i=1}^n (Y_i - \bar{Y})^2 \tag{60}$$
$$SSR = \Sigma_{i=1}^n (\hat{Y}_i - \bar{Y})^2 \tag{61}$$
$$SSE = \Sigma_{i=1}^n (Y_i - \hat{Y}_i)^2 \tag{62}$$
$$MSR = \frac{SSR}{(p-1)} \tag{63}$$
$$MSE = \frac{SSE}{(n-p)} \tag{64}$$
$$F = \frac{MSR}{MSE} \tag{65}$$

Where $\hat{Y}_i = (56)$

Source	SS	df	MS	F
Regression	SSR	$p-1$	MSR	F
Error	SSE	$n-p$	MSE	
Total	SSTO	$n-1$		

Under $H_0 : \beta_1 = \dots = \beta_{p-1} = 0$, $F \sim F_{p-1,n-p}$
R^2 has the same interpretation as in SLR.

5.2 Inference about Regression Parameters

Step 1
$H_0 : \beta_k = \beta_{k0}$
$H_1 : \beta_k \neq \beta_{k0} (> \beta_{k0}, < \beta_{k0})$ where β_{k0} is a specified value (e.g. 0).

$$t = \frac{\hat{\beta}_k - \beta_{k0}}{se(\hat{\beta}_k)} \tag{66}$$
$$se(\hat{\beta}_k) = \sqrt{MSE * (X'X)^{-1}_{k,k}} \tag{67}$$

Where $(X'X)^{-1}_{k,k}$ is the kth diagonal element of $(X'X)^{-1}$. $(0 \leq k \leq p-1)$
Under $H_0, t \sim t_{n-p}$
Step 2
A $100(1-\alpha)\%$ sci for $\beta_h, h \in G$ with $g = |G|$

$$\hat{\beta}_h = B * se(\hat{\beta}_h) \tag{68}$$
$$B = t_{n-p}\left(1 - \frac{\alpha}{2g}\right) \tag{69}$$

5.3 Estimation of Mean Responses

$$E(Y_h) = x_h'\beta = \beta_0 + \beta_1 x_{h,1} + \dots + \beta_{p-1} x_{h,p-1} \tag{70}$$

First compute $\hat{Y}_h = x_h'\hat{\beta}$ and

$$se(\hat{Y}_h) = \sqrt{MSE(x_h'(X'X)^{-1}x_h)} \tag{71}$$

A $100(1-\alpha)\%$ sci for $E(Y_h), h \in G, g = |G|$ is

$$W-H: \quad \hat{Y}_h \pm W * se(\hat{Y}_h), W = \sqrt{p * F_{p,n-p}(1-\alpha)} \tag{72}$$
$$Bonf.: \quad \hat{Y}_h \pm B * se(\hat{Y}_h), B = t_{n-p}\left(1 - \frac{\alpha}{2g}\right) \tag{73}$$

5.4 Prediction Interval

$$pse(\hat{Y}_h) = \sqrt{MSE(1 + x_h'(X'X)^{-1}x_h)} \tag{74}$$

A $100(1-\alpha)\%$ spi for Y_h

$$Scheffe: \quad \hat{Y}_h \pm S * pse(\hat{Y}_h), S = \sqrt{g * F_{g,n-p}(1-\alpha)} \tag{75}$$
$$Bonfer: \quad \hat{Y}_h \pm B * pse(\hat{Y}_h), B = t_{n-p}\left(1 - \frac{\alpha}{2g}\right) \tag{76}$$

5.5 Multiple Predictor SS

$$\begin{aligned} SSR(x_2|x_1) &= SSR(x_1, x_2) - SSR(x_1) \\ &= SSE(x_1) - SSE(x_1, x_2) \\ SSR(x_3|x_1, x_2) &= SSR(x_1, x_2, x_3) - SSR(x_1, x_2) \\ &= SSE(x_1, x_2) - SSE(x_1, x_2, x_3) \end{aligned}$$

SSR has (number of predictors on the left of the bars) degrees of freedom.

5.6 ANOVA with extra SS

Source	SS	df
Regression	$SSR(x_1, x_2, x_3)$	3
x_1	$SSR(x_1)$	1
$x_2\|x_1$	$SSR(x_2\|x_1)$	1
$x_3\|x_1, x_2$	$SSR(x_3\|x_1, x_2)$	1
Error	$SSE(x_1, x_2, x_3)$	$n-4$
Total	SSTO	$n-1$

Source	MS
Regression	MSR
x_1	$MSR(x_1) = SSR(x_1)/1$
$x_2\|x_1$	$MSR(x_2\|x_1) = SSR(x_2\|x_1)/1$
$x_3\|x_1, x_2$	$MSR(x_3\|x_1, x_2) = SSR(x_3\|x_1, x_2)/1$
Error	$MSE(x_1, x_2, x_3) = \frac{SSE(x_1, x_2, x_3)}{(n-4)}$
Total	

5.7 F Test for predictors

$H_0 : \beta_3 = 0$
$H_1 : \beta_3 \neq 0$
$SSE(Full) = SSE(x_1, x_2, x_3)$
$SSE(Reduced) = SSE(x_2, x_2)$

$$F = \frac{MSR(x_3|x_1, x_2)}{MSE(x_1, x_2, x_3)} \tag{77}$$

$df_R = n-3, df_F = n-4$
Under $H_0, F \sim F_{1,n-4}$

5.8 Coefficient of Partial Determination

$$R^2_{Y,x_2|x_1} = R^2_{Y,2|1} = \frac{SSR(x_2|x_1)}{SSE(x_1)} = 1 - \frac{SSE(x_1, x_2)}{SSE(x_1} \tag{78}$$

It measures the proportionate reduction in the variation of Y due to adding x_2, given that x_1 is already in the model.
More generally

$$R^2_{Y,x_p,\dots,x_{p+q-1}|x_1,\dots,x_{p-1}} = 1 - \frac{SSE(x_1, \dots, x_{p+1-1})}{SSE(x_1, \dots, x_{p-1})} \tag{79}$$

5.9 Adjusted R

$$R^2 = 1 - \frac{SSE}{SSTO} \tag{80}$$
$$R^2_a = 1 - \frac{MSE}{MSTO} = \frac{SSE/(n-p)}{SSTO/(n-1)} \tag{81}$$

Models with more predictors will always have higher R^2, but R^2_a takes into account the number of predictors. Select the model that maximizes R^2_a.

5.10 Mallow's C

$$C_p = C_p(x_{i_1}, \dots, x_{i_{p-1}}) = \frac{SSE(x_{i_1}, \dots, x_{i_{p-1}})}{MSE(x_1, \dots, x_{K-1})} - (n - 2p) \tag{82}$$

where $SSE(x_{i_1}, \dots, x_{i_{p-1}}) = SSE$ of fitting the regression with $x_{i_1}, \dots, x_{i_{p-1}}$ and $MSE(x_1, \dots, x_{K-1}) = MSE$ of fitting the regression with all candidate predictors.
The best subset of predictors corresponds to the one such that C_p is small and close to p. Note: $C_p = K$.

5.11 AIC and BIC(SBC) criteria

$$AIC_p = n\log(SSE_p\ n) + 2p \tag{83}$$
$$SBC_p = n\log(SSE_p\ n) + (\log n)p \tag{84}$$

Choose a subset of predictors (model) that minimizes $AIC_p(SBC_p)$.

5.12 Forward Stepwise Selection

1. Choose the first predictor (x_1) that has the largest $|t|$ for the slope under a simple linear regression with the predictor.
2. Choose the second predictor $x(2)$ that has the largest $|t|$ for the coefficient under a linear regression with (x_1) and a new predictor.
3. Continue until the p-value of the new predictor is greater than 0.10.
4. After adding new predictors, check existing predictor p-values. If any are greater than 0.15, remove them from the model.

5.13 Conditional residual plots

$e(Y|x_2)$ = residual of fitting Y against x_2.
$e(x_1|x_2)$ = residual of fitting x_1 against x_2.
A linear pattern in the plot of $e(Y|x_1)$ against $e(x_2|x_1)$ suggest that an important predictor, x_2, is missing in the model.

5.14 Identifying outlying Y observations

Internally Studentized (Standardized) residual: Let $\hat{\epsilon}_i$ denote the residual, the studentized residual is,

$$r_i = \frac{\hat{\epsilon}_i}{se(\hat{\epsilon}_i)} = \frac{\hat{\epsilon}_i}{\sqrt{(MSE(1 - h_{ii}))}} \tag{85}$$

The motivation for studentizing is that the variance of residuals for different inputs may differ, even if the variances of the errors are equal.
where h_{ii} is the ith diagonal element of the hat matrix $H = X(X'X)^{-1}X'$, also called the leverage for the ith case.

deleted (jackknife) residual: Fit the regression with the ith case deleted; let $\hat{Y}_{i(-i)}$ denote the predicted value for Y_i, under this regression. The idea behind the deleted residual is that an influential data point i, pulls the regression line towards itself. By removing that data point, the line should bounce back away from the original response, resulting in a large deleted residual.
The deleted residual is,

$$d_i = Y_i - \hat{Y}_{-(-i)} \tag{86}$$

studentized deleted (externally studentized) residual.

$$t_i = \frac{d_i}{se(d_i)} = r_i\left(\frac{n-k-2}{n-k-1-r_i^2}\right)^{1/2} \tag{87}$$
$$se(d_i) = \sqrt{\frac{MSE_i}{1 - h_{ii}}} \tag{88}$$
$$MSE_i = \frac{(1 - h_{ii}SSE - \hat{\epsilon}_i^2)}{(n-p-1)(1-h_{ii})} \tag{89}$$
$$= \frac{n-p}{n-p-1}MSE - \frac{\hat{\epsilon}^2}{(n-p-1)(1-h_{ii})} \tag{90}$$

Under the null hypothesis H_0 : no outliers, $t_i \sim t_{n-p-1}$.

5.15 Bonferonni's method for obtaining critical value for studentized deleted residuals

Decision Rule: Reject H_0 : no outliers, if

$$\max_{1 \leq i \leq n} |t_i| > t_{n-p-1}(1 - \frac{\alpha}{2n}) \tag{91}$$

where p is the number of β's

1. Calculate critical value $t_{n-p-1}(1 - \frac{\alpha}{2n})$.
2. Calculate all studentized residuals t_i.
3. Get max of absolute value $\max_{1 \leq i \leq n} |t_i|$ if residuals.
4. If $\max |t_i| < t^*$, fail to reject H_0 and conclude no outliers.

5.16 Identifying outlier x observations

Recall h_{ii} is the ith diagonal element of the hat matrix $H = Px$, which is called the leverage for the ith case.
A property:

$$\Sigma_{i=1}^n h_{ii} = p \tag{92}$$

If $h_{ii} > 2h = \frac{2p}{n}$, case i is considered outlying in x.

5.17 Identifying influential cases

An outlying case isn't necessarily influential, to identify influential cases, consider Cook's Distance.

$$D = \frac{\Sigma_{j=1}^n (\hat{Y}_j - \hat{Y}_{j(-i)})^2}{p * MSE} \tag{93}$$

where $\hat{Y}_j$ is the predicted value of Y_j via regression with the full data, and $\hat{Y}_{j(-i)}$ is the predicted value of Y_j via regression with the data without the ith case.
Large values of D_i indicate a potentially influential case.
Another more computationally convenient expression is,

$$D_i = \frac{h_{ii}\hat{\epsilon}_i^2}{p(1 - h_{ii})^2 MSE} \tag{94}$$