Zusammenfassung F-Test und Modellselektion

- Beispiel: $y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$

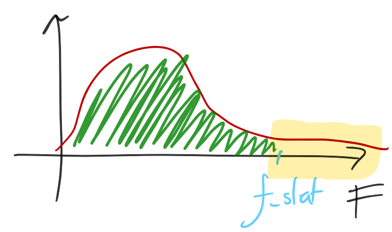
- F-Test: Verfahren, bei dem getestet wird, ob die geprüften Variablen gemeinsam einen signifikanten Erklärungsgehalt für das Modell (y) aufweisen.

↳ Intuition: Wir haben gesehen, dass wir wegen Korrelationen zwischen $(X_i)_{i=1}^{3}$ hohe Standardfehler erhalten können, was die marginalen Tests verzerrt. Zudem berücksichtigen die marginalen Tests die Korrelation zwischen $(X_i)_{i=1}^{3}$ nicht.

↳ "gemeinsamer Erklärungsgehalt"

Marginale Tests:
1. $H_0: \beta_1 = 0$, $H_A: \beta_1 \neq 0$
2. $H_0: \beta_2 = 0$, $H_A: \beta_2 \neq 0$

↳ F-Test:
- $H_0: \beta_1 = \beta_2 = 0$ (1)
  $H_A: \beta_1 \neq 0$ und/oder $\beta_2 \neq 0$



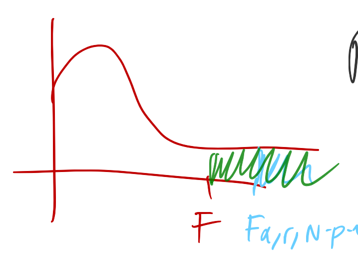
- $$F = \frac{(SQR_0 - SQR_1)/r}{SQR_1/(N-p-1)} \sim F_{\alpha,r,N-p-1} \quad (2)$$

pf(f-stat,..) = W

p-Wert = 1 − pf(f-stat,..

- Verwerfen auf Niveau α, falls

$$F > F_{\alpha,r,N-p-1}$$



oder p-Wert $< \alpha$.

$$R^2_{p+1} \geq R^2_p$$

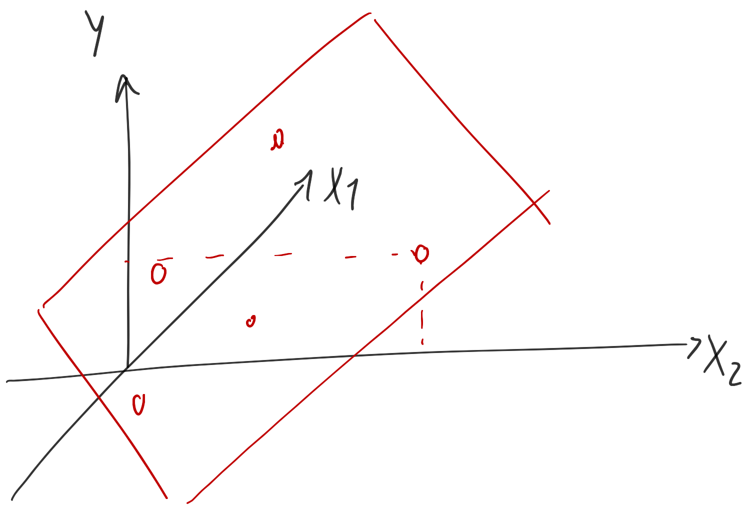
$$R^2_{p+1} = 1 - \frac{SQR_{p+1}}{TQS}$$

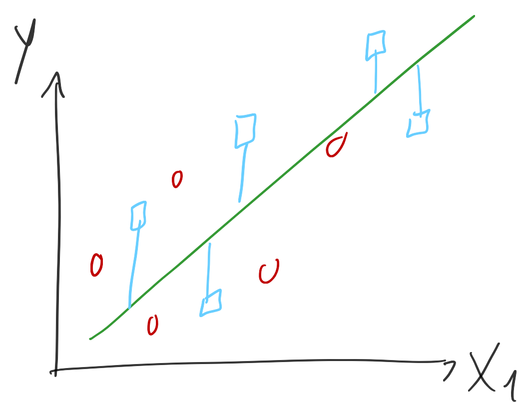
· **Variablenselektion:** Wir haben gesehen, dass $R_p^2 \leq R_{p+1}^2$ (auf Trainingsdaten). Wir wollen aber ein einfaches Modell, dass die tatsächlich wichtigen Faktoren einfängt und eine hohe Anpassungsgüte hat. Wir führen dazu einen Trade-off zw. Anpassungsgüte und Anzahl von Regressoren ein, indem p gewählt wird durch die Minimierung von:

$$AIC = N \log(SQR) + 2p$$

oder

$$BIC = N \log(SQR) + p \log(N)$$





Einführung in die Empirische Wirtschaftsforschung

Übungsaufgaben 9

Modellvergleich via F-Test - Modelselektion - Zeitreihen

1. Betrachten Sie erneut in R das Datenset ”miete”. Die Variable “rent” ist die Zielvariable.

   a) Die Variablen “size” und “rooms” kommen als Prediktoren in Frage. Verwenden Sie die AIC-Methode der Variablenselektion, um unter allen möglichen gebildetenen Modellen das beste zu bestimmen.

   b) Erklären Sie den Unterschied zwischen Variablenselektion und einem F-Test.

   c) Lohnt es sich, zusätzlich zu “size” und “rooms” die qualitativen Prädiktoren für die Regionen in das Modell hinein zu nehmen. Beantworten Sie die Frage mit einem geeigneten F-Test.
   `Tipp`: Sie können dazu die R-Funktion `anova(fit1, fit2)` verwenden.

   d) Wiederholen Sie den F-Test in b), aber diesmal, indem Sie direkt die Formel für den F-Test verwenden und zudem den $p-$Wert selber mit der in R eingebauten F-Verteilungsfunktion berechnen.

2. Wir betrachten in R die Zeitreihe der Lufttemperatur im Datensatz ”luft”. In den folgenden Teilaufgaben bezeichnen wir die Lufttemperatur zum Zeitpunkt t als Y_t.

   a) Plotten Sie Y_t über die Zeit. Handelt es sich um einen stationären Prozess?

   b) Plotten Sie Y_t gegen Y_{t-1}. Was fällt Ihnen auf?
   Tipp: Beim kreieren von Y_{t-d} verlieren Sie d Beobachtungen der original Zeitreihe.

   c) Generieren Sie zusätzlich die Variable Y_{t-365} und speichern Sie die drei Variablen Y_t, Y_{t-1} und Y_{t-365} in einem `data.frame`. Regressieren Sie nun Y_t auf die beiden anderen Variablen, aber lassen Sie dazu die letzten 32 Beobachtungen in Ihrem `data.frame` fürs training weg. Sagen Sie nun genau diese letzten 32 Beobachtungen vorher und berechnen Sie den MSE.

   d) Wie ändert sich der MSE wenn Sie nur Y_{t-1} als Prediktor verwendet hätten?

   e) Kommt ein F-Test auf das gleiche Ergebnis?

1. Betrachten Sie erneut in R das Datenset ”miete”. Die Variable “rent” ist die Zielvariable.

   a) Die Variablen “size” und “rooms” kommen als Prediktoren in Frage. Verwenden Sie die AIC-Methode der Variablenselektion, um unter allen möglichen gebildetenen Modellen das beste zu bestimmen.

   b) Erklären Sie den Unterschied zwischen Variablenselektion und einem F-Test.

   c) Lohnt es sich, zusätzlich zu “size” und “rooms” die qualitativen Prädiktoren für die Regionen in das Modell hinein zu nehmen. Beantworten Sie die Frage mit einem geeigneten F-Test.
   `Tipp`: Sie können dazu die R-Funktion `anova(fit1, fit2)` verwenden.

   d) Wiederholen Sie den F-Test in b), aber diesmal, indem Sie direkt die Formel für den F-Test verwenden und zudem den $p-$Wert selber mit der in R eingebauten F-Verteilungsfunktion berechnen.

2. Wir betrachten in R die Zeitreihe der Lufttemperatur im Datensatz "luft". In den folgenden Teilaufgaben bezeichnen wir die Lufttemperatur zum Zeitpunkt t als Y_t.

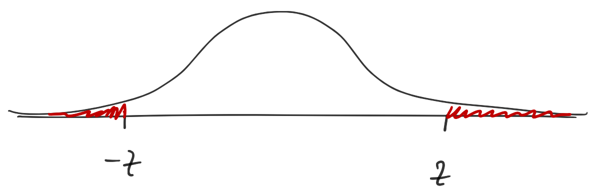
a) Plotten Sie Y_t über die Zeit. Handelt es sich um einen stationären Prozess?

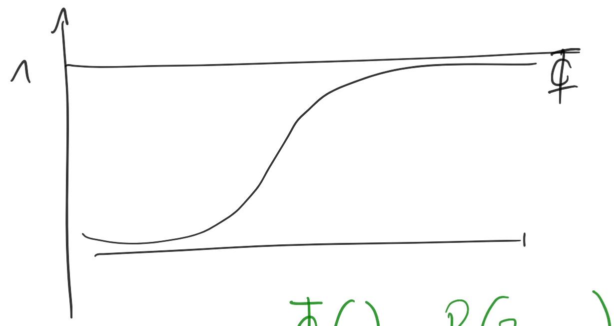
b) Plotten Sie Y_t gegen Y_{t-1}. Was fällt Ihnen auf?
Tipp: Beim kreieren von Y_{t-d} verlieren Sie d Beobachtungen der original Zeitreihe.

c) Generieren Sie zusätzlich die Variable Y_{t-365} und speichern Sie die drei Variablen Y_t, Y_{t-1} und Y_{t-365} in einem `data.frame`. Regressieren Sie nun Y_t auf die beiden anderen Variablen, aber lassen Sie dazu die letzten 32 Beobachtungen in Ihrem `data.frame` fürs training weg. Sagen Sie nun genau diese letzten 32 Beobachtungen vorher und berechnen Sie den MSE.

d) Wie ändert sich der MSE wenn Sie nur Y_{t-1} als Prediktor verwendet hätten?

e) Kommt ein F-Test auf das gleiche Ergebnis?





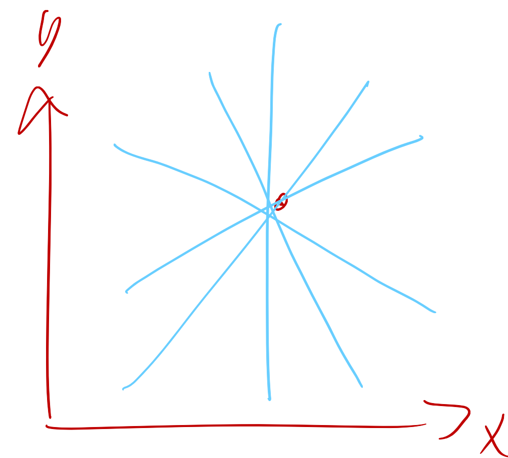
$\Phi(z) := P(Z \leq z)$

$\Phi(-1) = P(Z \leq -1)$

Φ

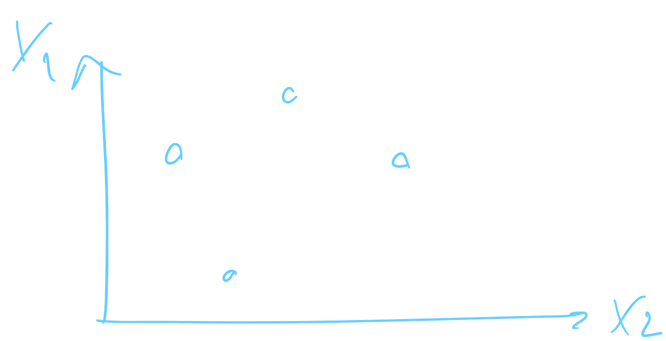
$$p\text{-Wert} = 2P(Z \geq |z|) = P(Z \leq -z) + P(Z \geq z)$$
$$= 2\Phi(-z)$$
$$= 2P(Z \leq -z) = P(Z \leq -z) + P(Z \leq -z)$$

$$t = \frac{b_1 - 0}{SE(b_1)} =$$



fit 2 : $X_1 \sim X_2 \quad (\hat{X}_1 = a_0 + a_1 X_2)$

$\hat{X}_1 - X_1$



$\hat{y} = 0.5 + 0.7X_1 + 0.3X_2$

$\hat{y}_1 = g_0 + g_1X_1$

$$\frac{Cov(X_1, Y)}{Var(X_1)} = Cov(X_1, Y)$$

- Klicker, Kap 4. S. 29 anschauen

$Y = f(X) + \varepsilon$

$$\boxed{Y = \beta_0 + \beta_1 X + \varepsilon}$$

Alter