# Regresyonlar

Fuat Can Beylunioğlu

December 21, 2017

# Giriş

- Regresyon bir bağımlı değişkenin $Y_i$ ile bağımsız değişken $X_i$ ya da değişkenlerle $X$ arasındaki ilişkiyi ölçer.
- Bu ilişki tek yönlüdür, yani $Y_i$'nin $X_i$'deki değişikliklerden nasıl etkilendiği üzerine bir modeldir.
- Model lineerdir, yani ilişkiyi hata terimlerini dışarıda bırakarak doğrusal olarak modellemeyi hedefler.
- Temel regresyon denklemi şu şekildedir:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \tag{1}$$

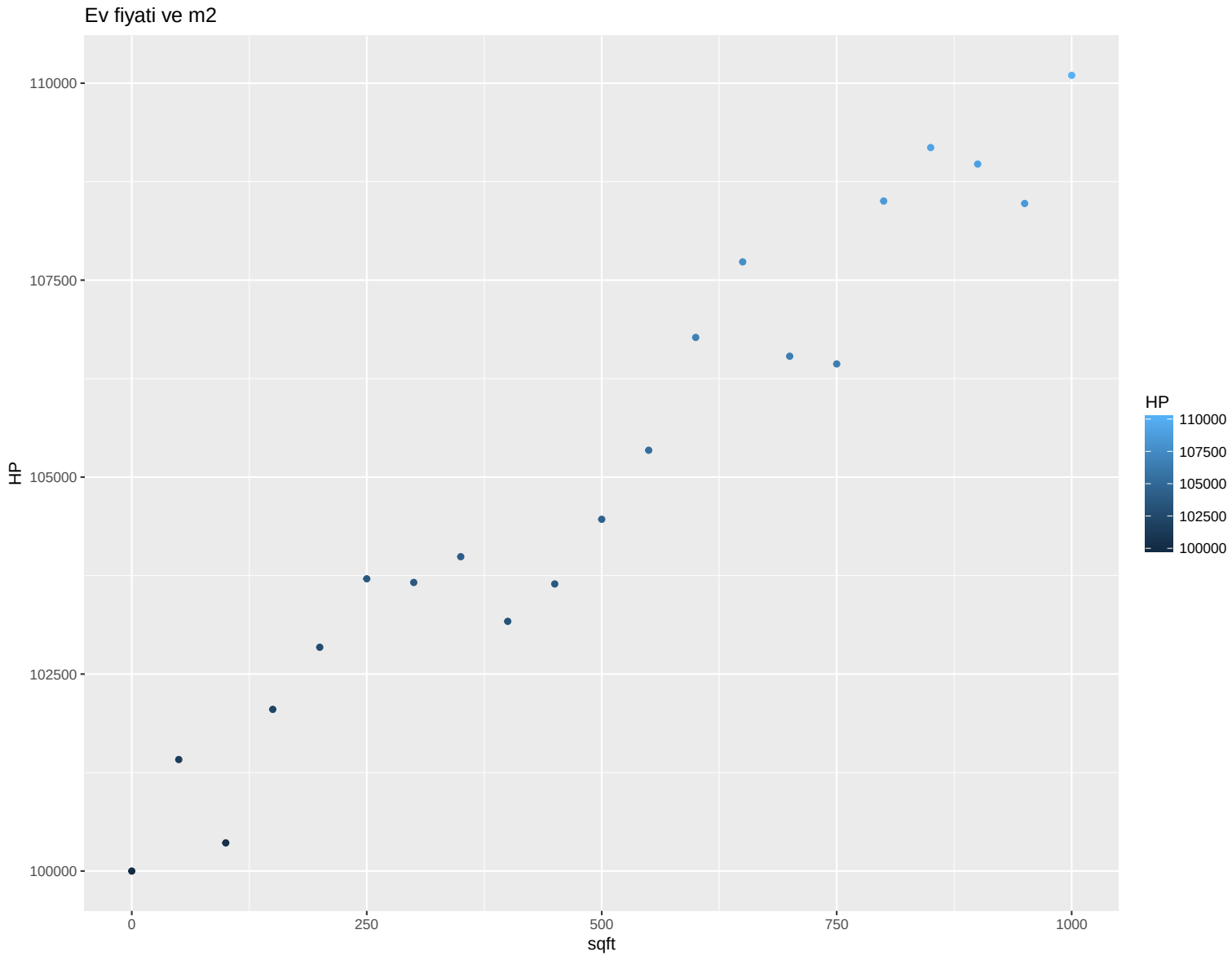
- Burada $\epsilon_i$'nin tamamen rassal olduğu varsayılmaktadır. Matematiksel olarak ifade etmek gerekirse $\epsilon_i \sim N(0, \sigma^2)$ olmalıdır. Aksi taktirde denklem taraflı (biased) olacaktır.

# Regresyon: Ev Fiyatı Örneği

- Örneğin ev fiyatının, evin $m^2$ genişliği ile ilişkili olduğunu varsayalım. Bu durumda evin teorik değeri aşağıdaki gibidir:

$$\mathsf{HP}_i = \beta_0 + \beta_1 \mathsf{sqft}_i + \epsilon_i \tag{2}$$

$\beta_0 = 100000, \beta_1 = 1000$



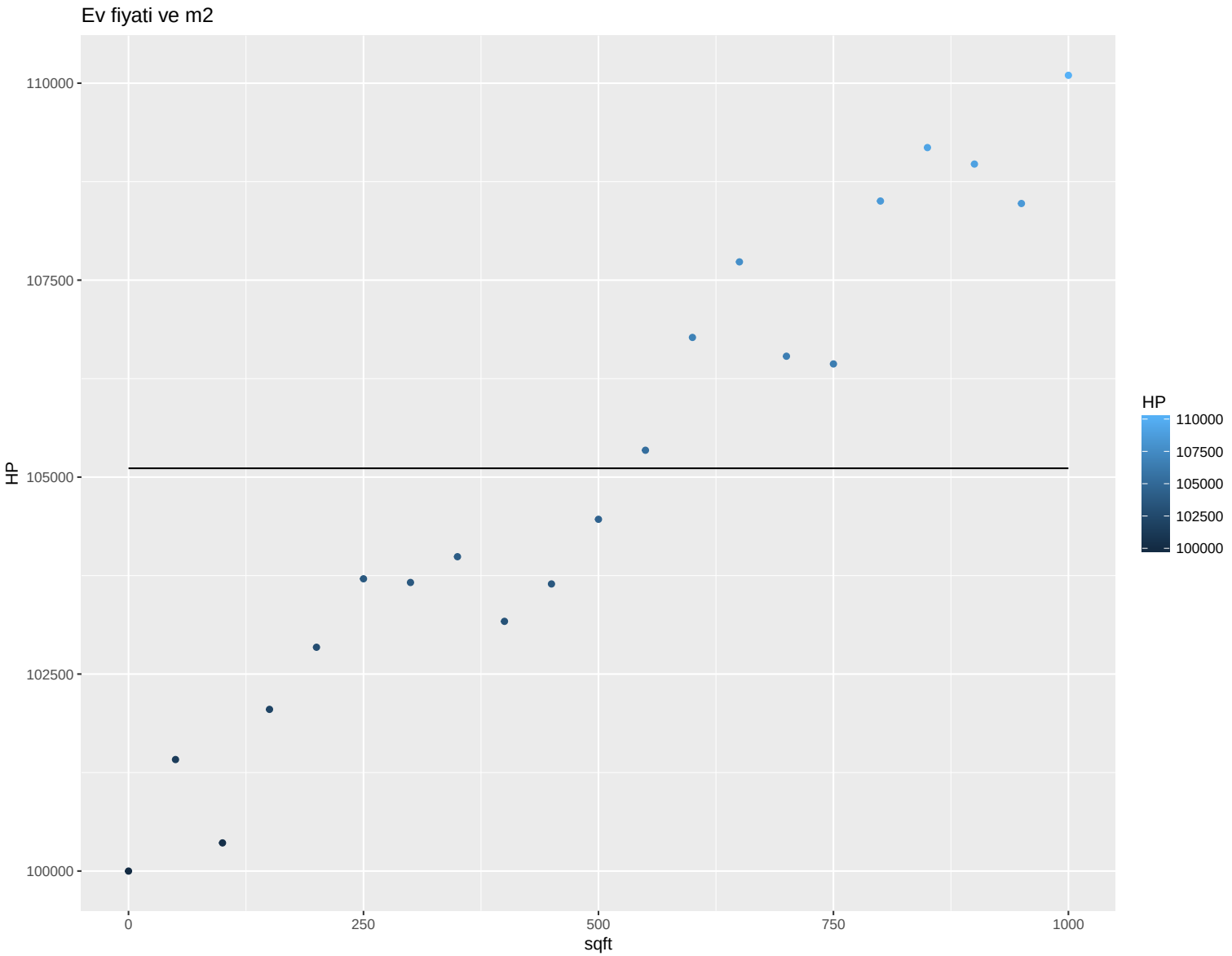
# Regresyon: Ev Fiyatı Örneği

- Örneğin ev fiyatını, evin $m^2$ genişliği ile tahmin eden aşağıdaki modeli ele alalım:

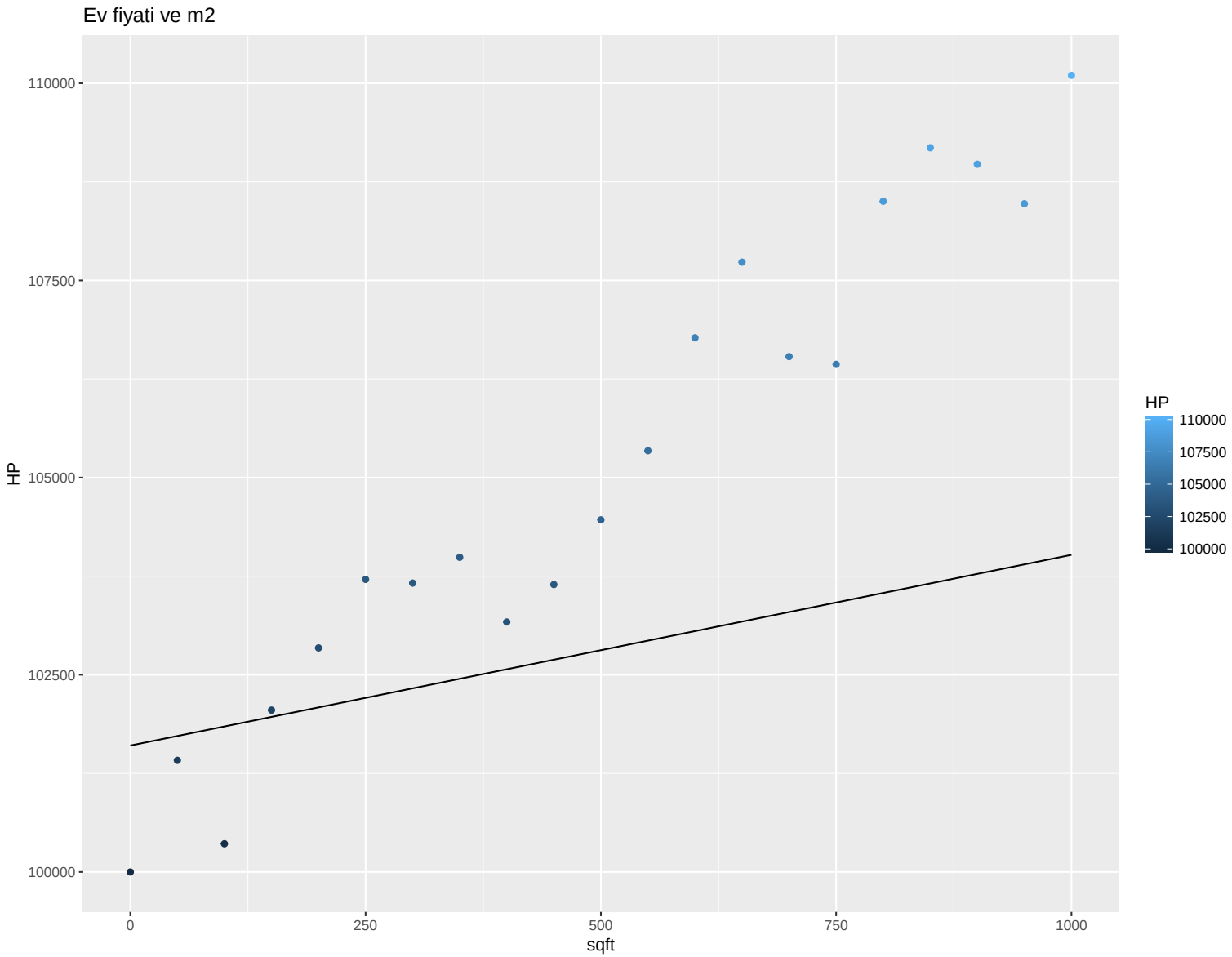
$$\text{HP}_i = \beta_0 + \beta_1 \text{sqft}_i \tag{3}$$

- $\beta_0 = 100000$ ve $\beta_1 = 1000$ olduğu durumda, modele göre evin baz değerinin 100000 olduğu ve her bir $m^2$'lik artışın evin değerine 1000TL değer eklediği söylenebilir.
- Yukarıdaki model regresyon çizgisinin formülüdür, $\epsilon_i$ terimi yazılmaz.

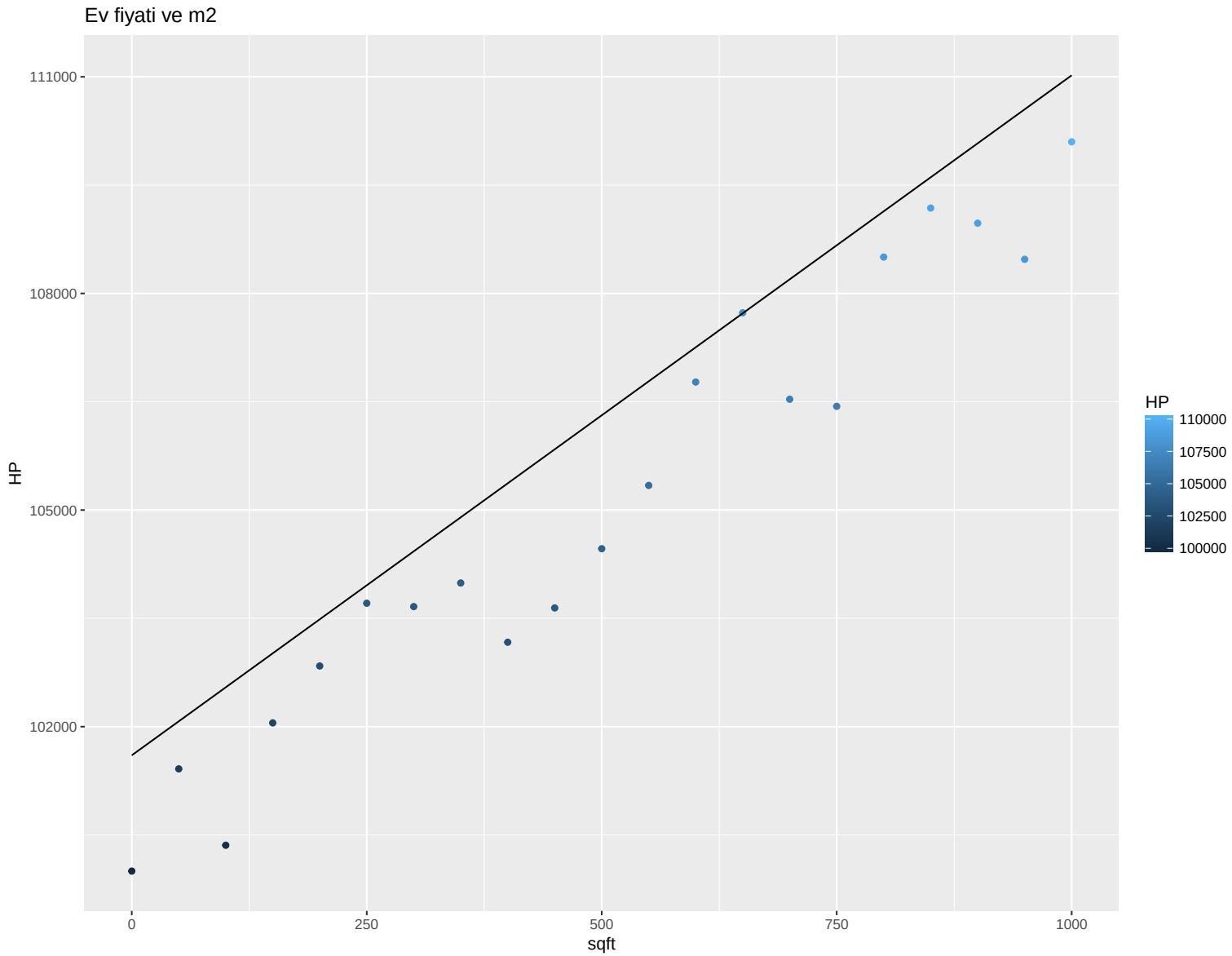
# Hangi Model


Ev fiyati ve m2
110000
107500
105000
102500
100000
HP
0
250
500
750
1000
sqft
HP
110000
107500
105000
102500
100000

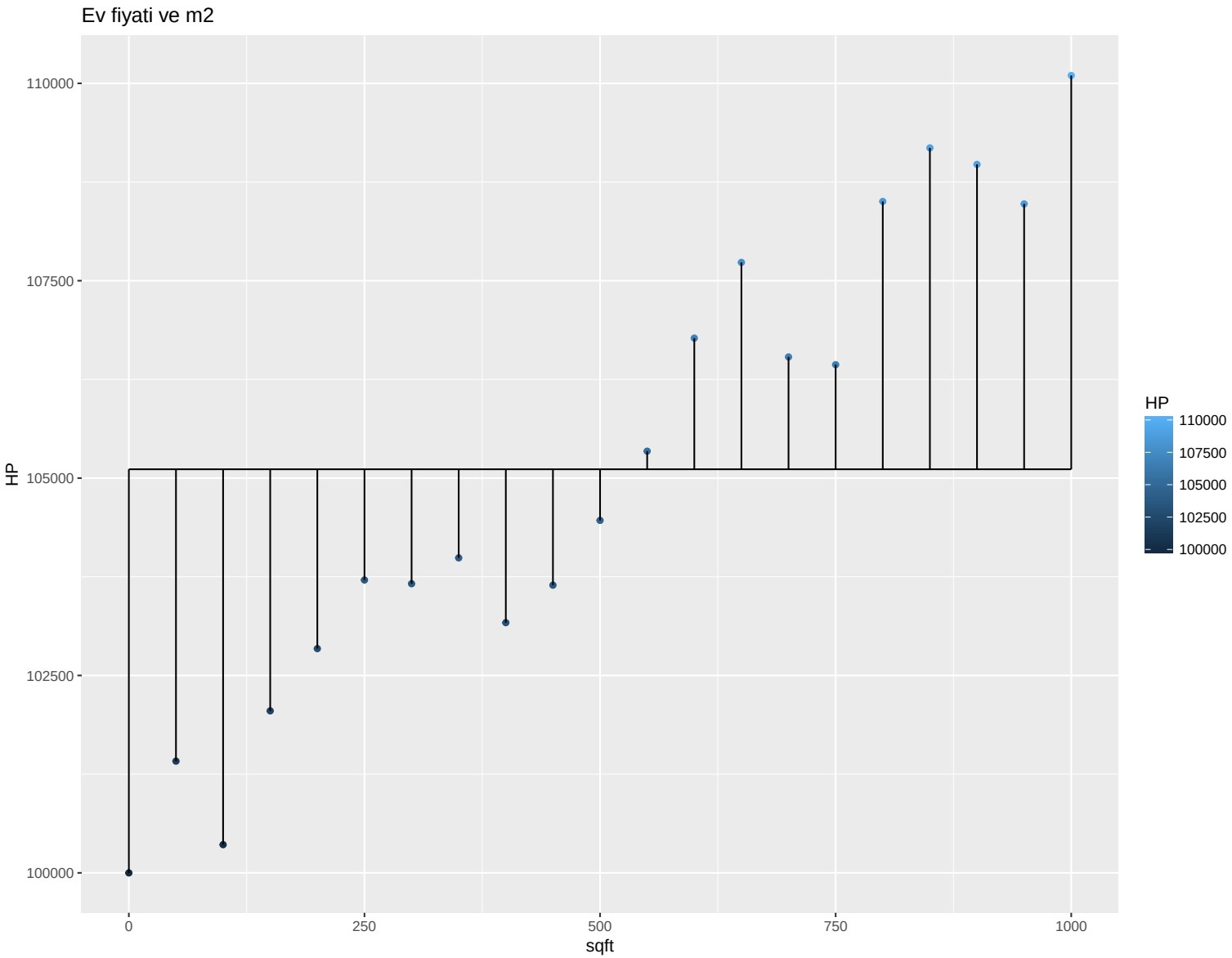
# Hangi Model


Ev fiyati ve m2
HP
110000
107500
105000
102500
100000
0
250
500
750
1000
sqft
HP
110000
107500
105000
102500
100000

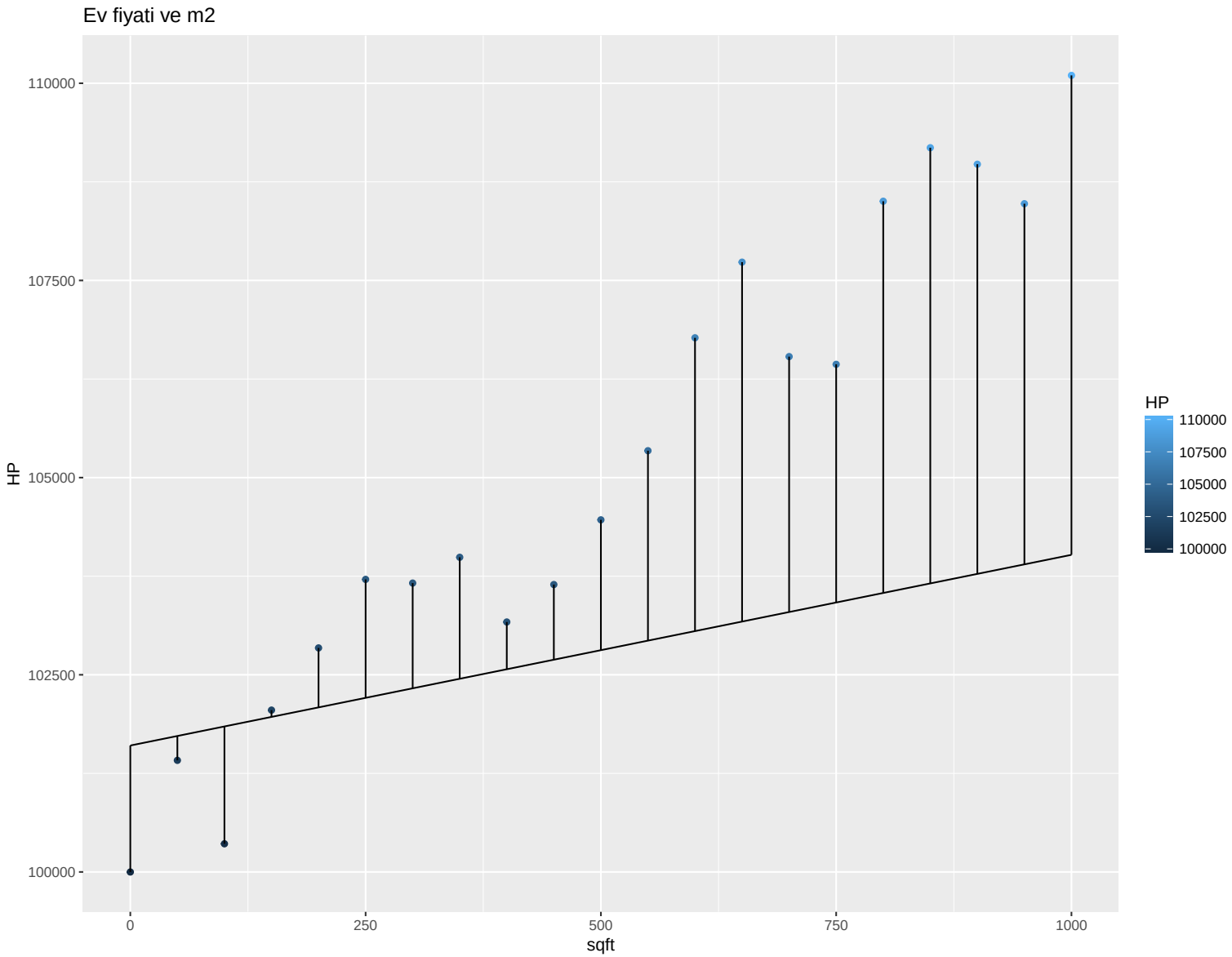
# Hangi Model


Ev fiyati ve m2
HP
111000
108000
105000
102000
0
250
500
750
1000
sqft
HP
110000
107500
105000
102500
100000

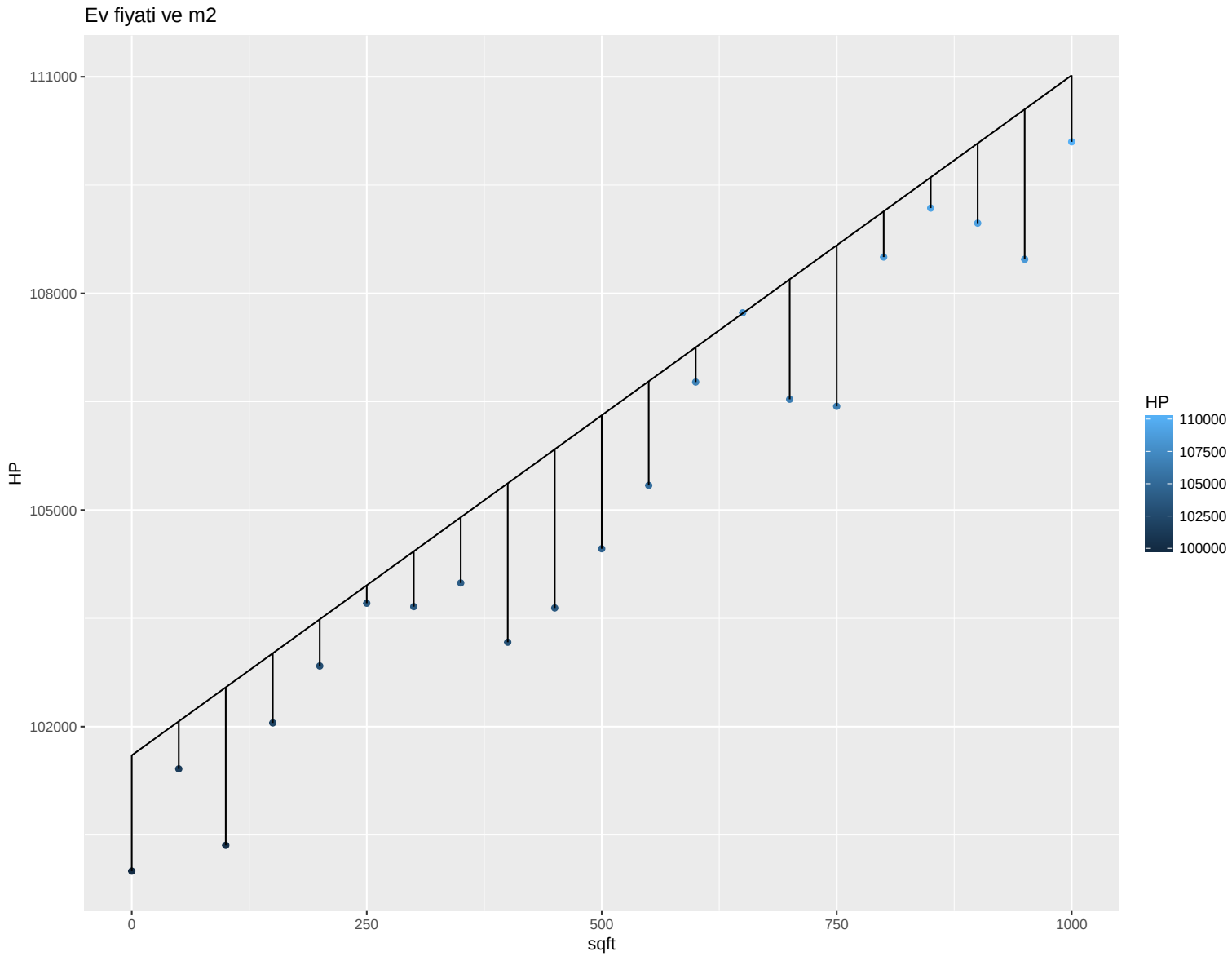
# Hangi Model


Ev fiyati ve m2
110000
107500
105000
102500
100000
HP
0
250
500
750
1000
sqft
HP
110000
107500
105000
102500
100000

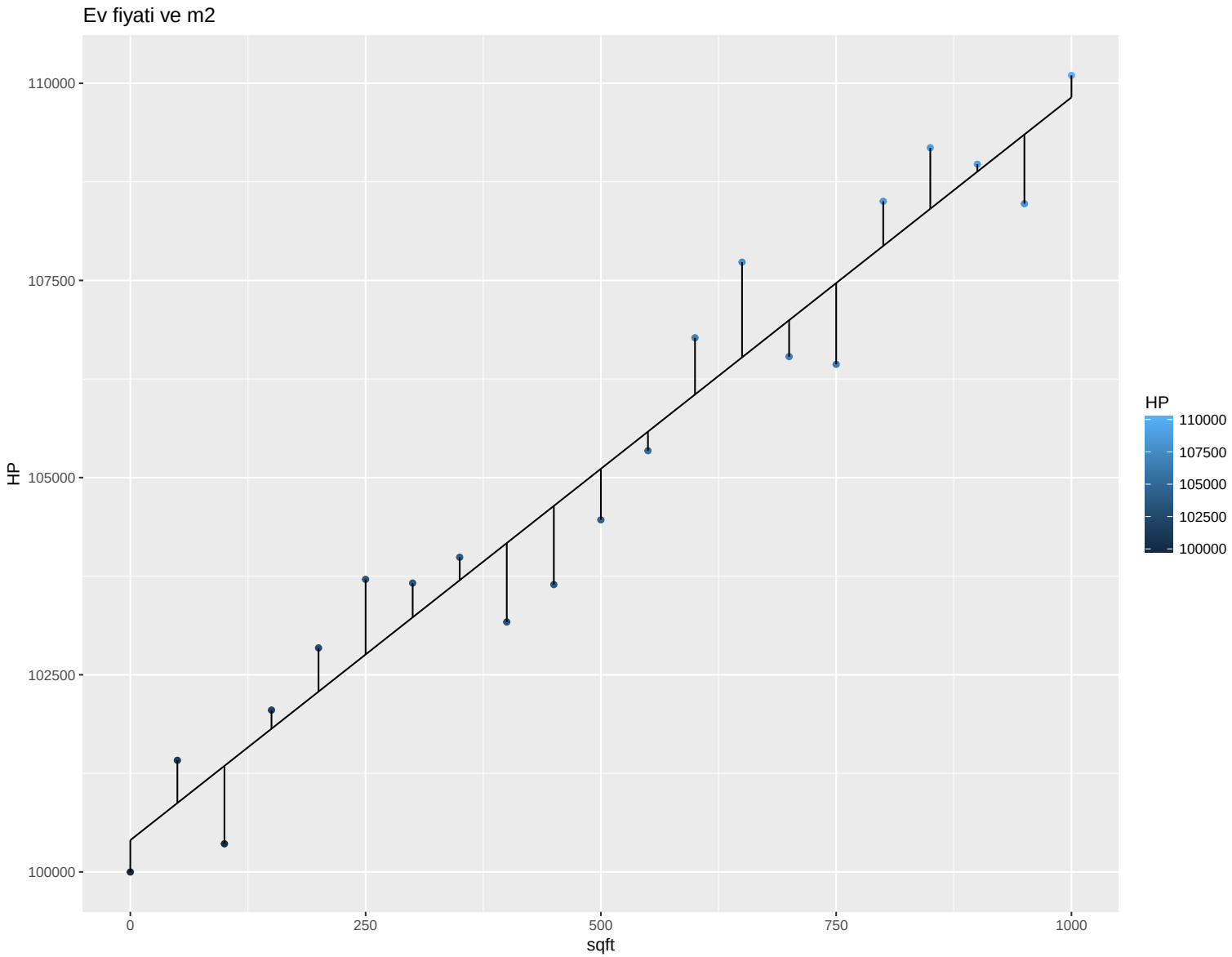
# Hangi Model


Ev fiyati ve m2
HP
110000
107500
105000
102500
100000
0
250
500
750
1000
sqft
HP
110000
107500
105000
102500
100000

# Hangi Model


Ev fiyati ve m2
111000
108000
105000
102000
HP
0
250
500
750
1000
sqft
HP
110000
107500
105000
102500
100000

# Hangi Model


Ev fiyati ve m2
HP
110000
107500
105000
102500
100000
0
250
500
750
1000
sqft
HP
110000
107500
105000
102500
100000

# En İyi Modeli Oluşturmak

- En iyi model hata terimlerinin karelerin toplamının (RSS) hesaplanması ile oluşturulur:

$$\text{RSS} = \sum_{i=1}^{N} \left( Y_i - \hat{Y}_i \right)^2 \tag{4}$$

- RSS'i minimize eden parametre kümesi ($\beta_i$) matematiksel olarak en iyi model(best fit)dir.
- Ancak örneklem kümesi ile elde edilen modelin popülasyon için de geçerli olabilmesi için belli varsayımları sağlaması gerekir.

# Çoklu Regresyon (Multiple Regression)

- Çoklu regresyon aşağıdaki formulle gösterilir:

$$Y_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + \cdots + \beta_n X_{nt} + \epsilon_t \tag{5}$$

- Ayrıca:

$$Y_t = \beta_0 + \sum_{i=1}^{N} \beta_t X_t + \epsilon_t \tag{6}$$

- ya da matrix formunda:

$$Y = X\beta + \epsilon \tag{7}$$

olarak gösterilir.

# Regresyonun Varsayımları

- Normality: $\epsilon_i$'lerin dağılımı $N(0, \sigma^2)$ olmalı,
- Homoskedasticity: Her bir $x_i$ için $\sigma^2$ sabit olmalı,
- No serial autocorrelation: Her bir $\epsilon_i$'nin bir diğeri ile arasında korelasyon olmamalı.
- No multicolinearity: Çoklu regresyon için $X_{ij}$ ve $X_{ik}$ arası korelasyon olmamalı.

# Modelin Açıklayıcılığı

- Bir modelin kuvvetini ölçen başlıca istatistikler:
  - $R^2$ ve Adjusted $R^2$
  - t-test, F-test
  - AIC, BIC

# Modelin Açıklayıcılığı, $R^2$

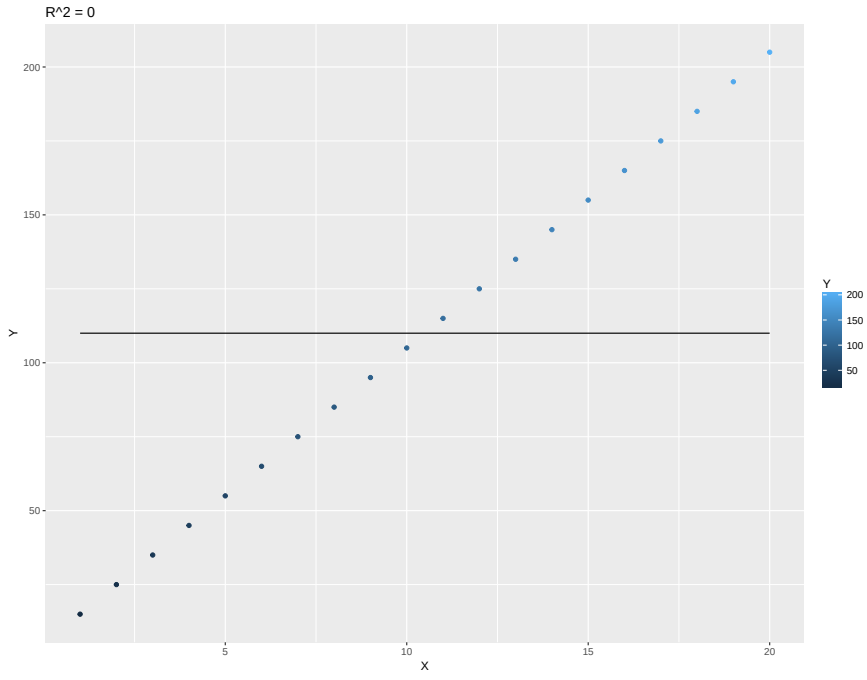



Figure: $R^2 = 0$

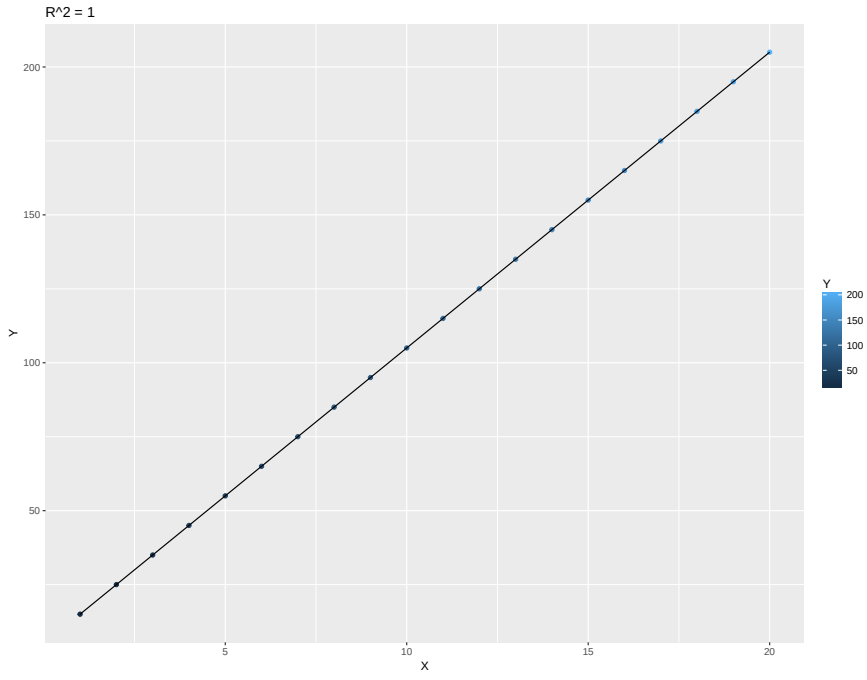
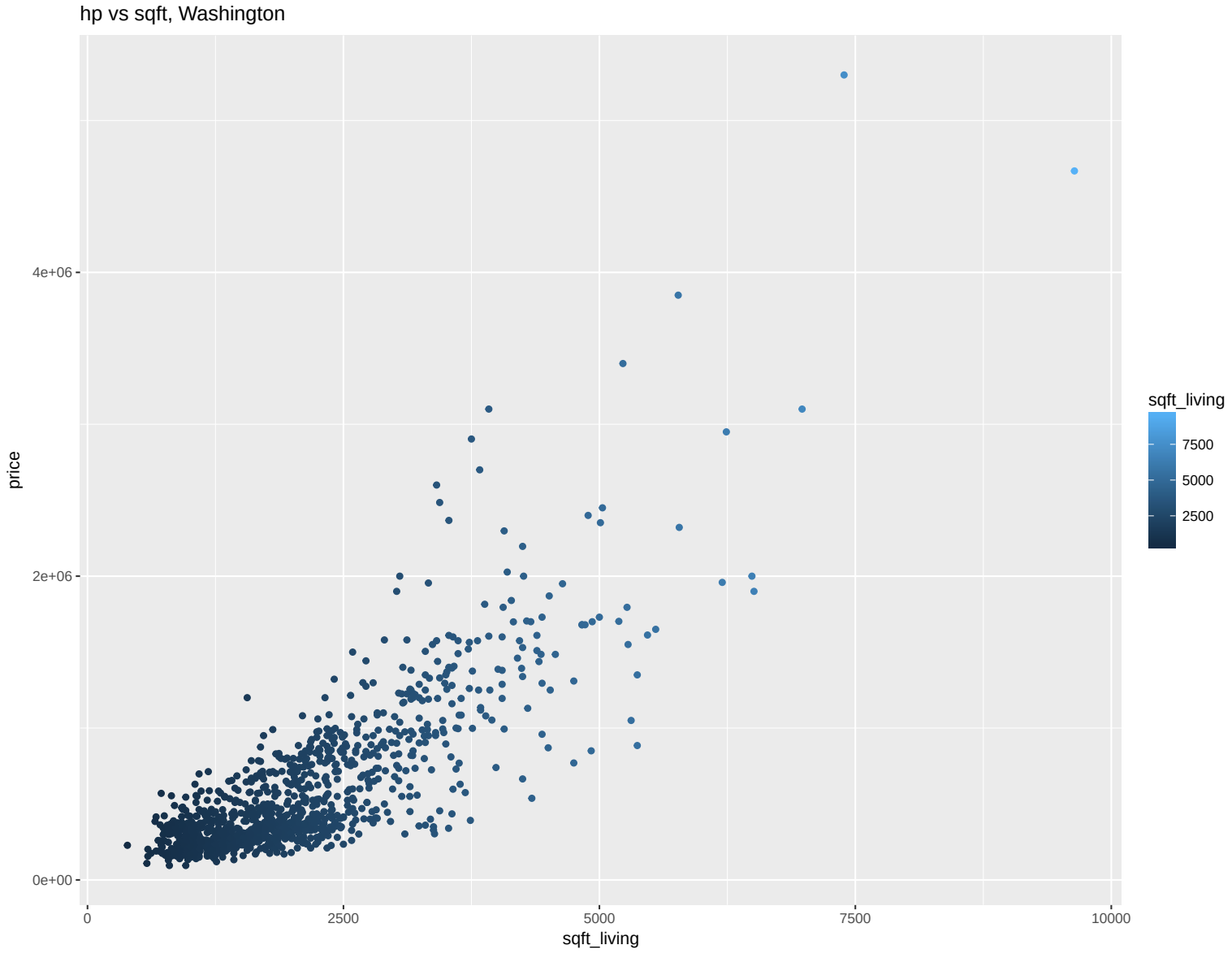
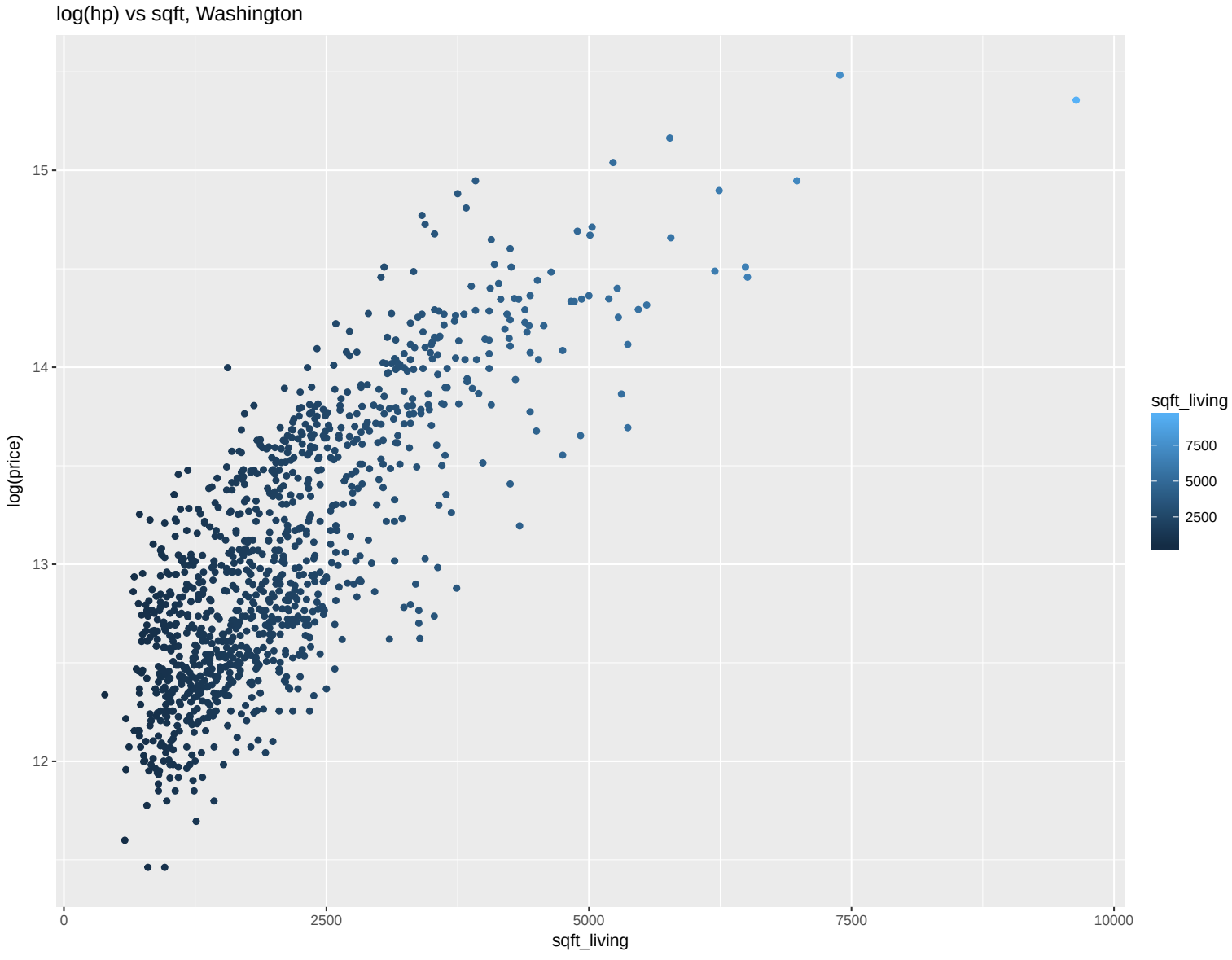


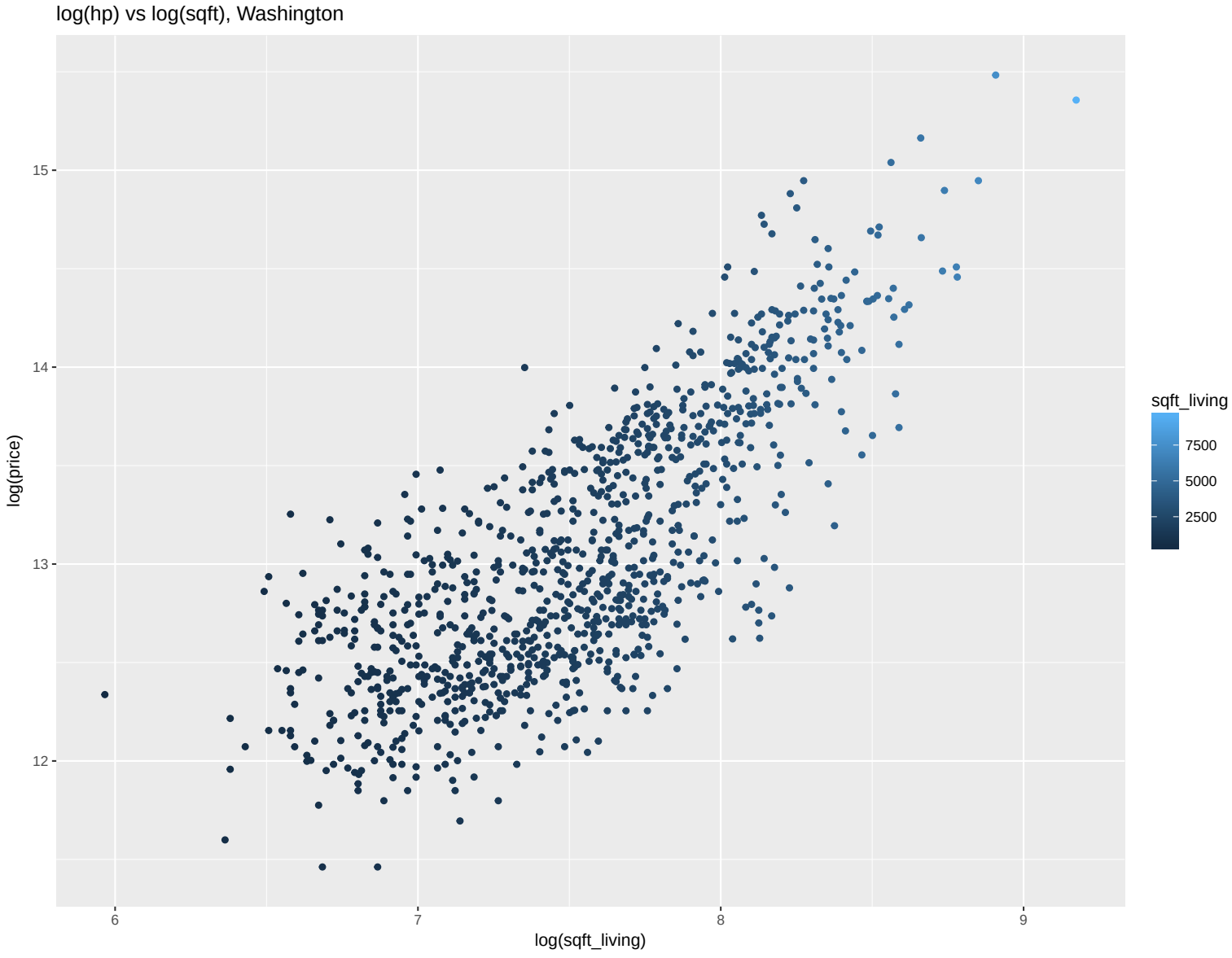

Figure: $R^2 = 1$

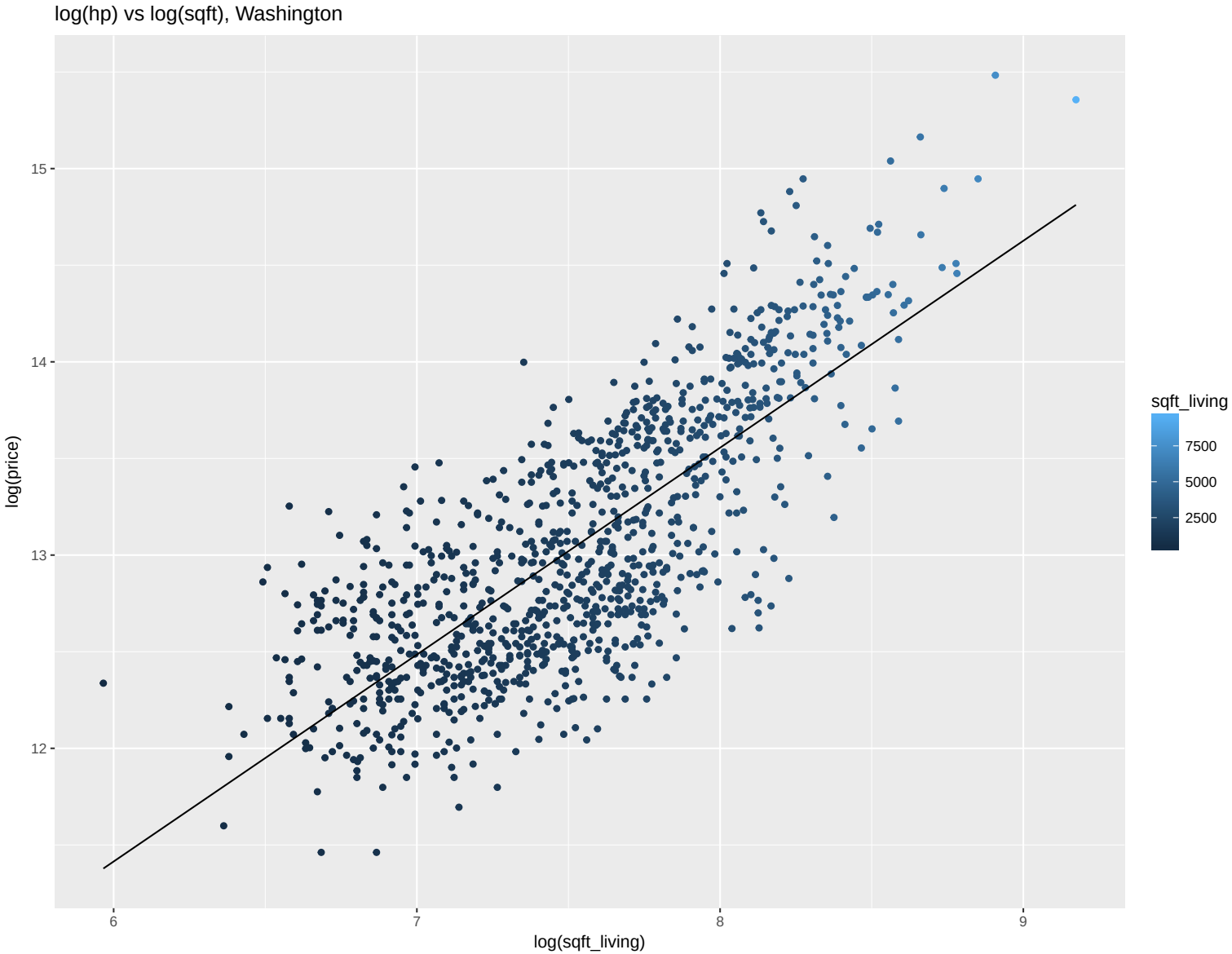
# Lineer Olmayan İlişkiler

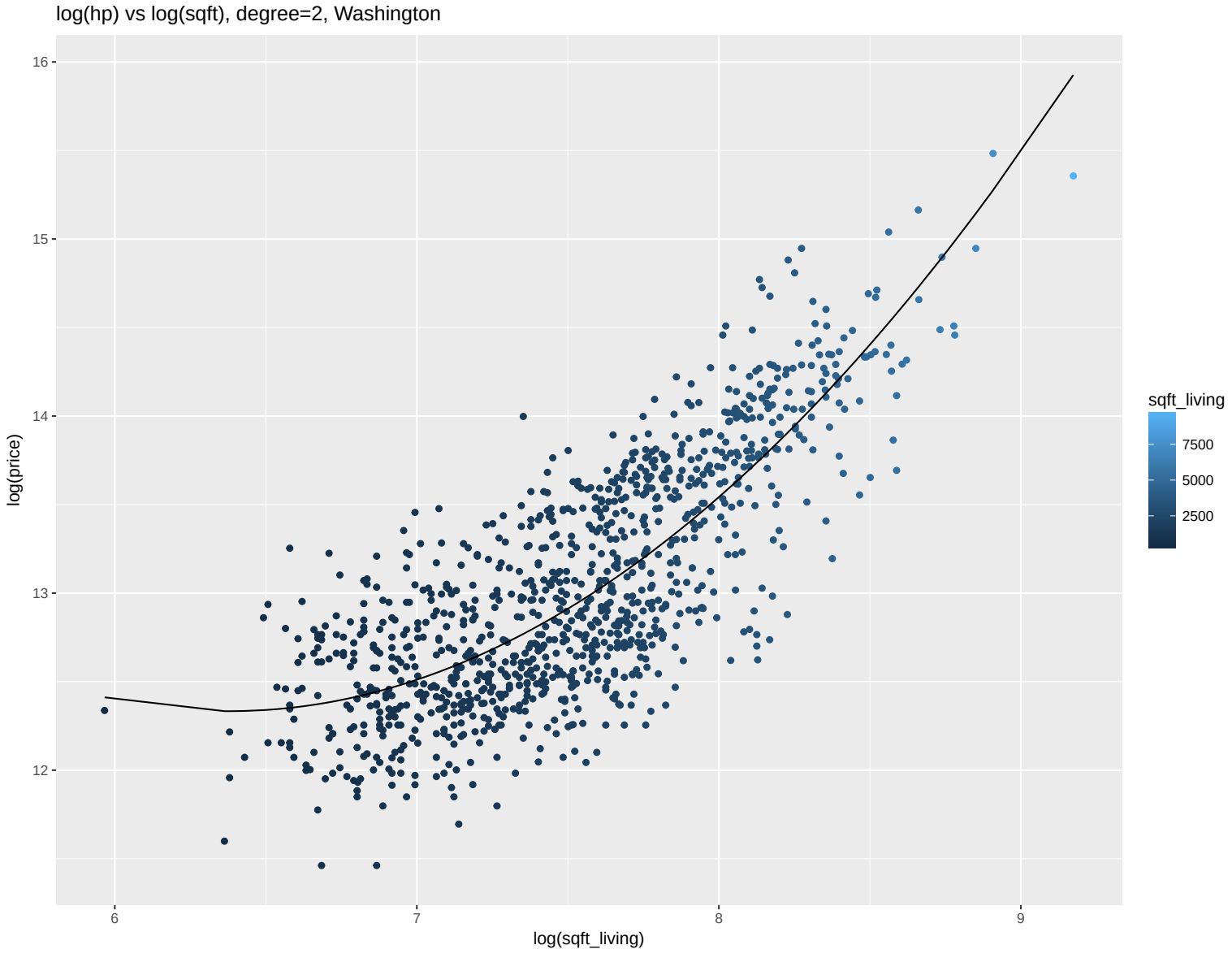
- Bazı durumlarda iki değer arasında lineer dışı ilişki söz konusu olur.
- Bu regresyon varsayımlarının sağlanamamasına neden olacaktır.
- Analizden önce iki değer arasında öncelikle scatter plot yapılması bu açıdan önemlidir.
- Bu gibi durumlarda değerlerden biri ya da ikisinin log değeri alınabilir. Log değerler veriler arasındaki açıklıkları düşürerek normal dağılıma yakınsamalarını sağlayabilir, öte yandan heteroskedasticity sorununu da çözebilir.
- Sorun çözülmediği taktirde polinom modellere başvurulabilir.

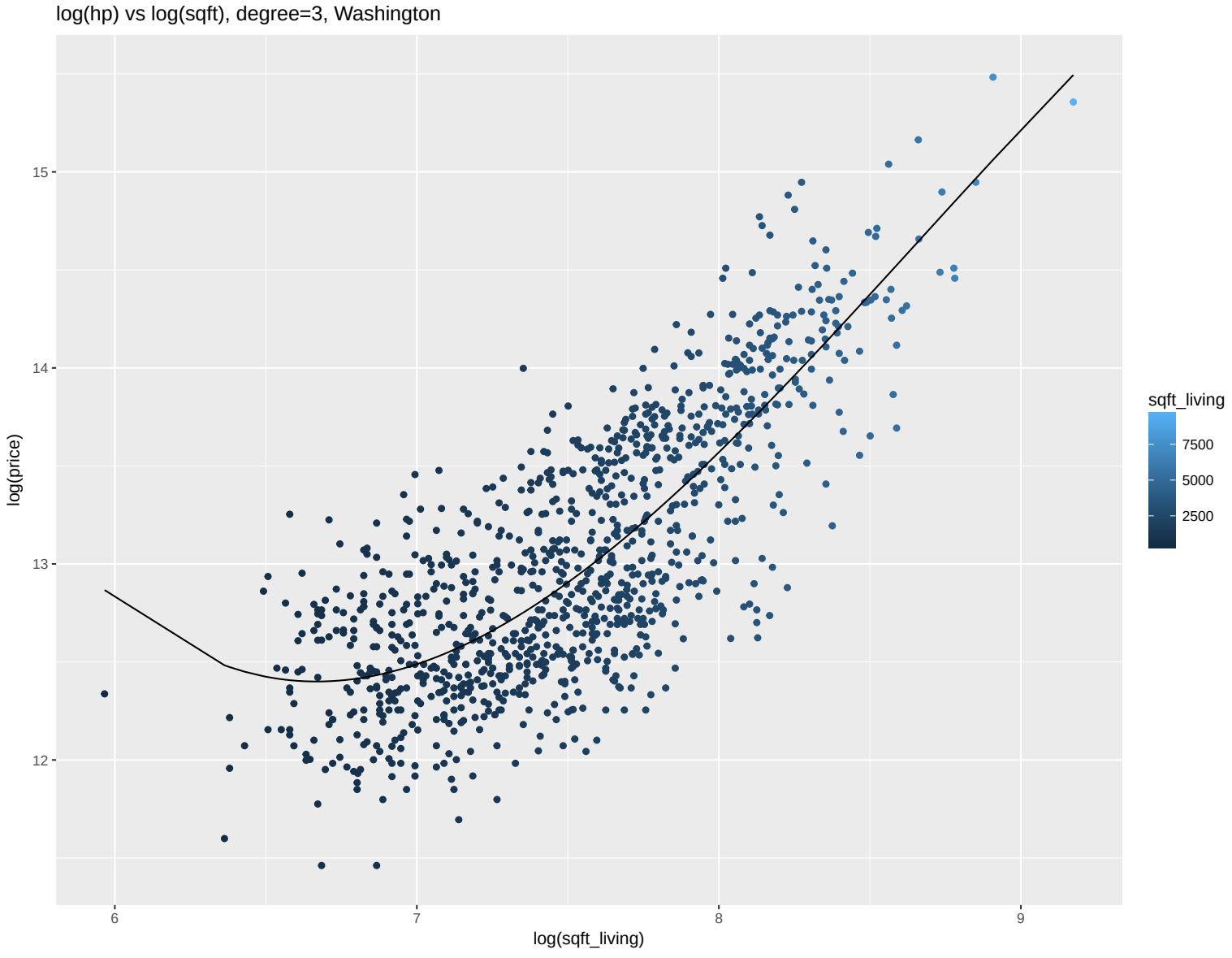

hp vs sqft, Washington
price
4e+06
2e+06
0e+00
0
2500
5000
7500
10000
sqft_living
sqft_living
7500
5000
2500


log(hp) vs sqft, Washington
log(price)
15
14
13
12
0
2500
5000
7500
10000
sqft_living
sqft_living
7500
5000
2500


log(hp) vs log(sqft), Washington
log(price)
15
14
13
12
6
7
8
9
log(sqft_living)
sqft_living
7500
5000
2500

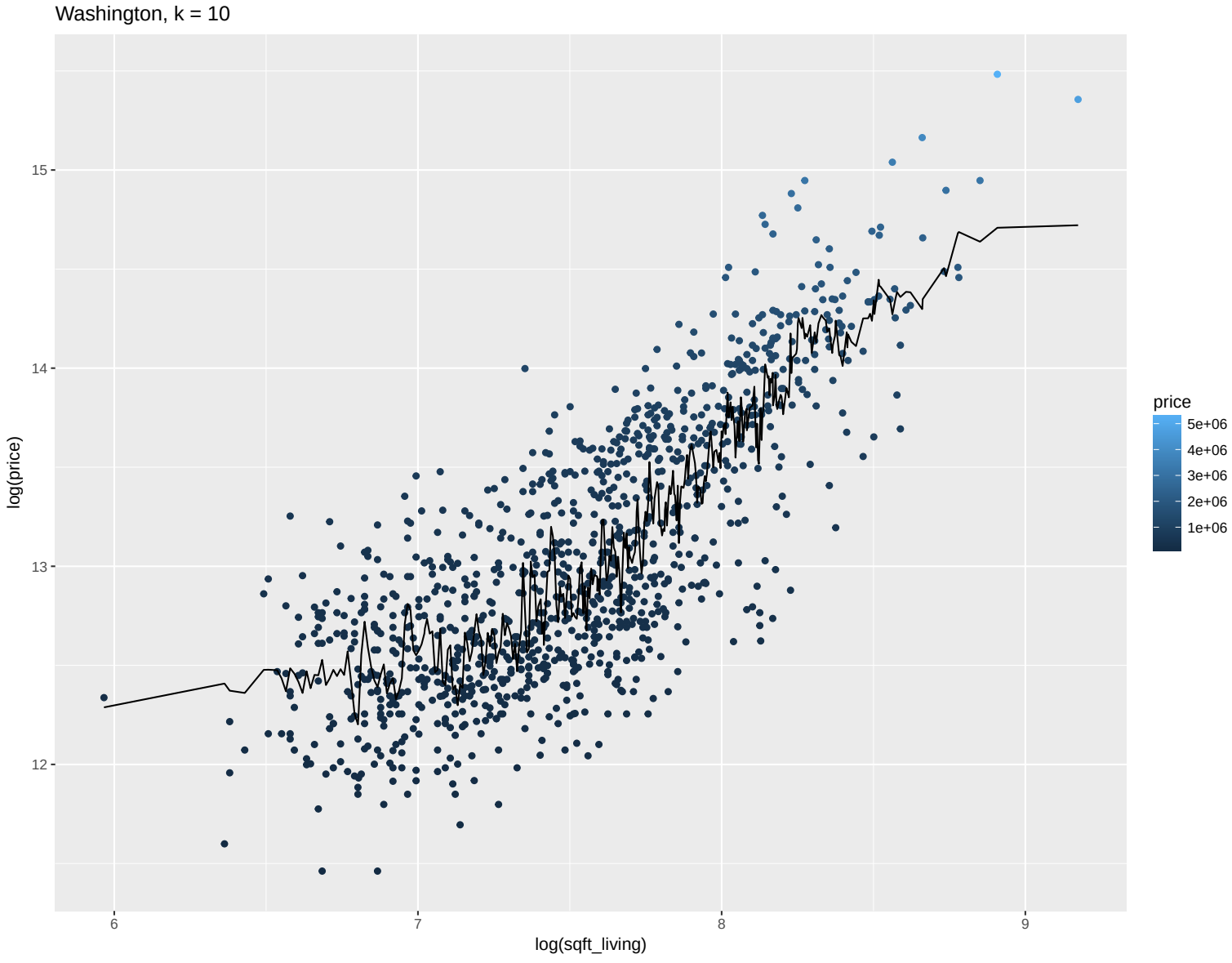

log(hp) vs log(sqft), Washington
log(price)
15
14
13
12
6
7
8
9
log(sqft_living)
sqft_living
7500
5000
2500

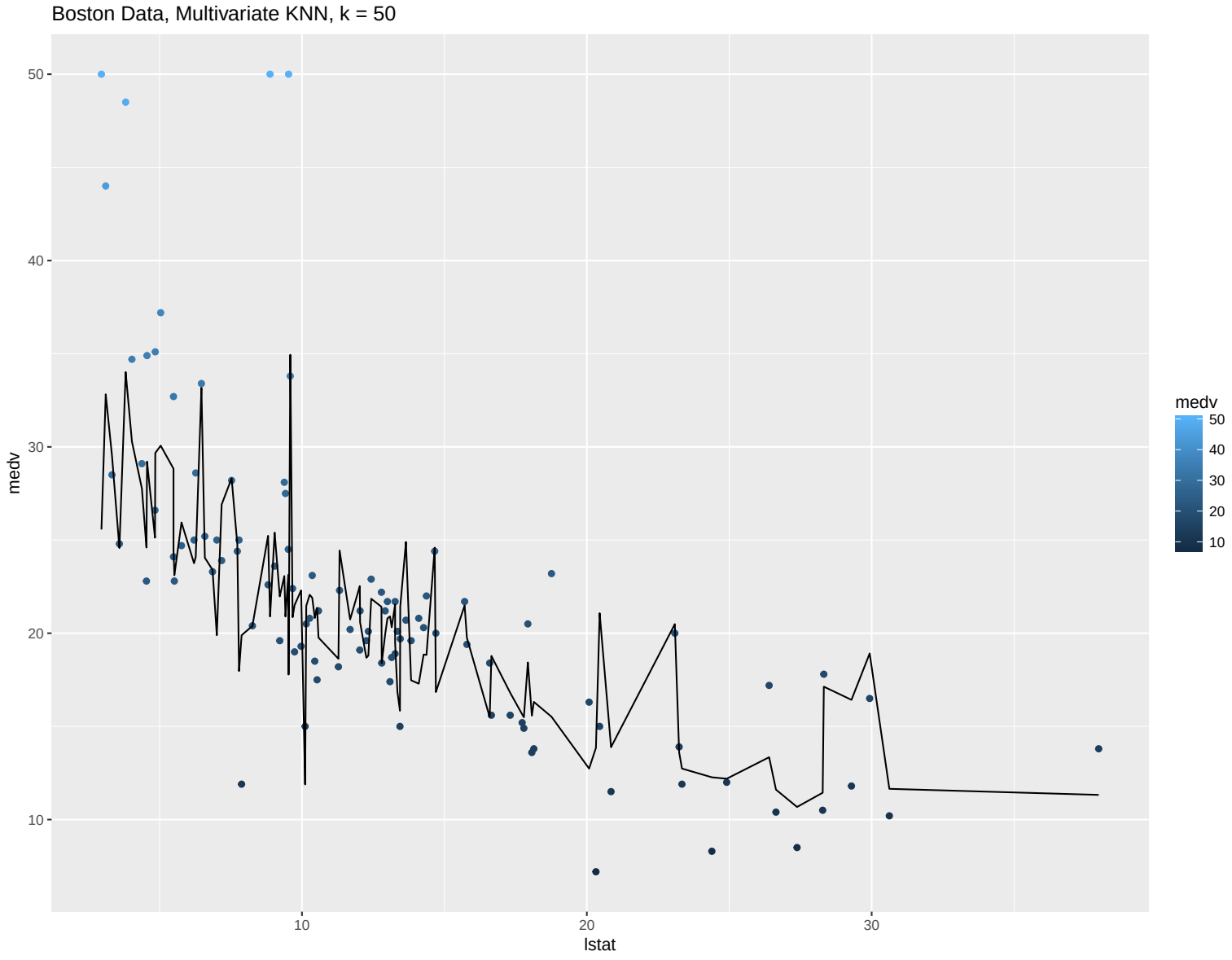

log(hp) vs log(sqft), degree=2, Washington
log(price)
log(sqft_living)
16
15
14
13
12
6
7
8
9
sqft_living
7500
5000
2500


log(hp) vs log(sqft), degree=3, Washington
log(price)
15
14
13
12
6
7
8
9
log(sqft_living)
sqft_living
7500
5000
2500

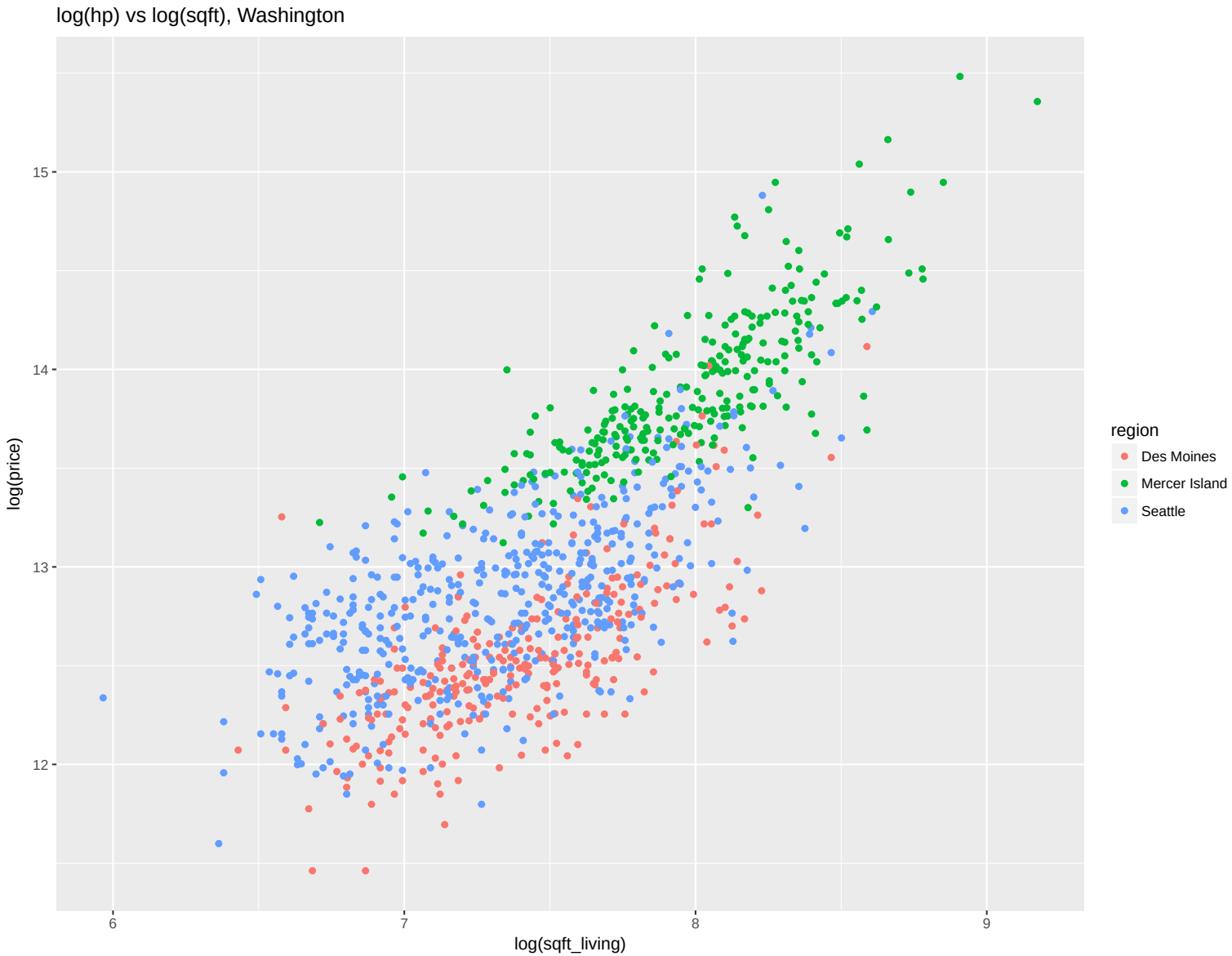
# K-NN Regressions

- Lineer olmayan regresyonlardan biri de k-NN regresyonlardır.
- Arkasında yatan motivasyon, yeni bir değer için tahmin istendiğinde, veri setinde ona en yakın değeri vermektir.
- Başarısı lineer ve polinom modele göre yüksektir.
- Ancak çoklu modeller ile ilgili çok fazla çalışma bulunmamaktadır.


Washington, k = 10
log(price)
log(sqft_living)
price
5e+06
4e+06
3e+06
2e+06
1e+06
15
14
13
12
6
7
8
9


Boston Data, Multivariate KNN, k = 50
medv
lstat
50
40
30
20
10
10
20
30
medv
50
40
30
20
10

# Verinin Yapısı



# Lineer Model için R kodları

| | |
|---|---|
| lm() | Lineer model |
| summary() | Modelin özet istatistikleri |
| coefficients() | Hesaplanan katsayılar |
| confint() | Hesaplanan katsayılar için güven aralıkları |
| fitted() | Her bir gözlem için modelin hesapladığı değerler |
| residuals() | Hesaplanan değer ile gerçekleşen arasındaki farklar |
| AIC() | AIC değeri |
| BIC() | BIC değeri |
| plot() | Modelin başarısı hakkında grafikler |
| predict() | Modeli kullanarak yeni değerler için tahmin |

# Kaynakça I



Kabacoff, Robert I,
*R in Action*,
2010.


Dalpiaz, David
https://daviddalpiaz.github.io/r4sl/k-nearest-neighbors.html,
2017.